Artificial Intelligence in Drug Discovery: The Six Circles of Hell

Andreas Bender, PhD
Professor for Machine Learning in Medicine, Khalifa University, Abu Dhabi, UAE

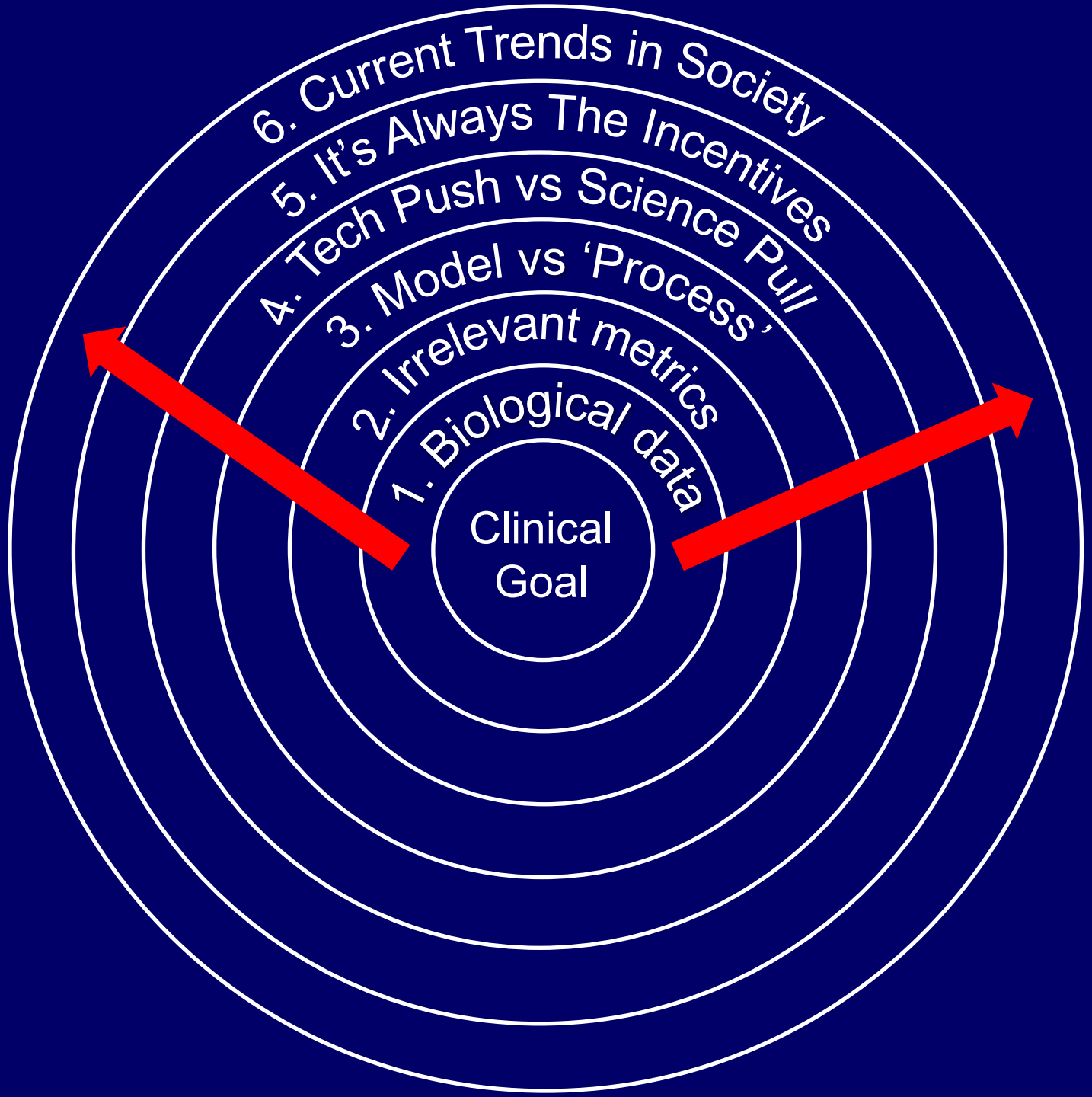
Visiting Professor, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, UK
Research Professor at STAR-UBB Institute, Babeş-Bolyai University, Cluj-Napoca, Romania
Project Leader at Iuliu Hatieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania
Adjunct Faculty at National Institute for Bioprocessing Research and Training (NIBRT), Dublin, Ireland

Co-Founder of Healx, Ltd., PharmEnable Therapeutics, Ltd., Pangea Bio Ltd.

جامعة خليفة
Khalifa University

Six Circles of Hell:

What (often) prevents AI in drug discovery from having impact



This statement is frequently encountered, often celebrated… but ultimately pointless

'Our model achieves 93% Performance on This and that Benchmark, which Outperforms SOTA and revolutionizes drug Discovery, for the 1001st time'

SOTA = ‘State Of The Art’, a term frequently used in machine learning that something is as good as it currently gets

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)

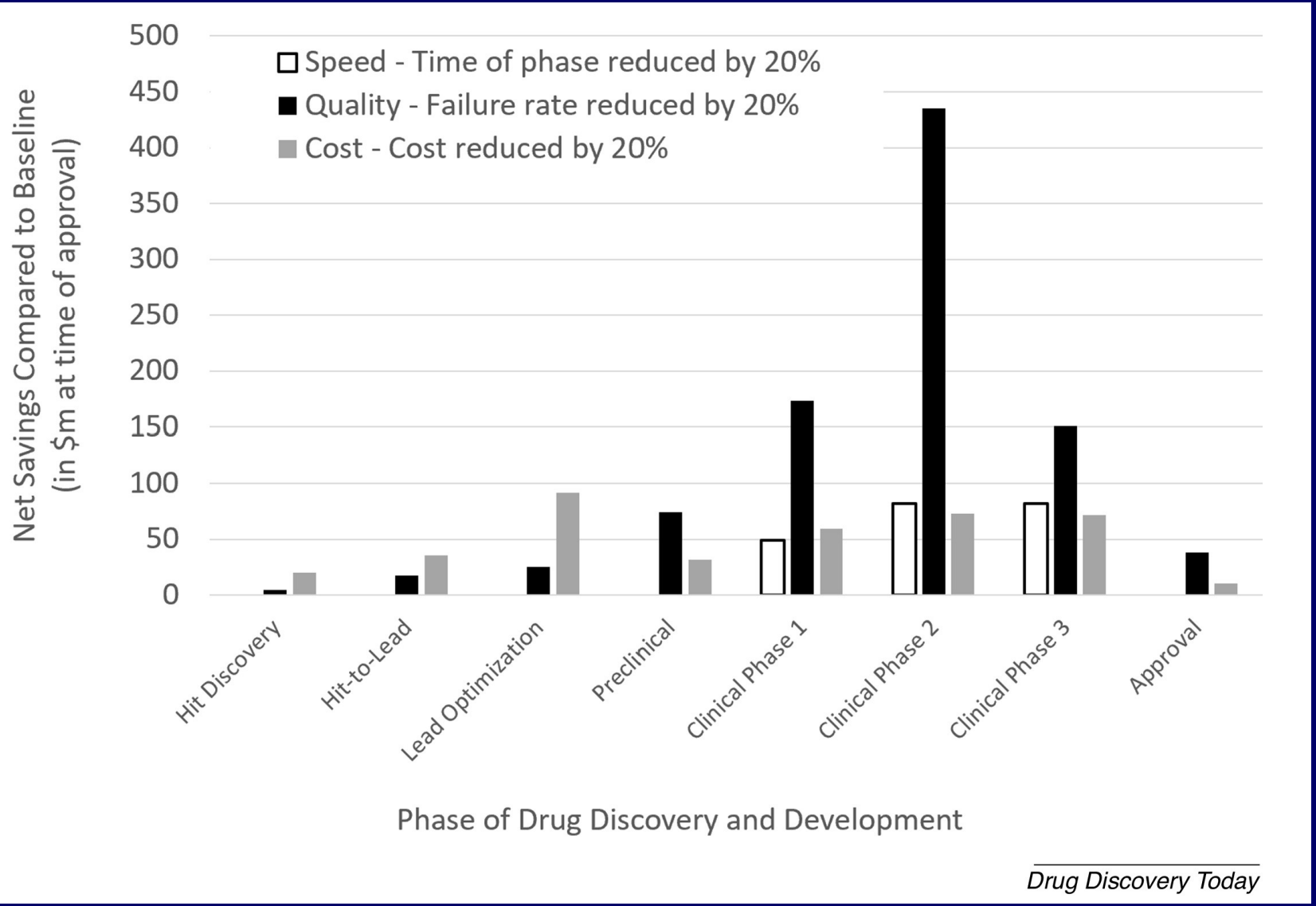
Andreas Bender and Isidro Cortes-Ciriano

Drug Discovery Today 2021

These slides, and new preprint currently under review on '*Artificial intelligence in drug discovery – what does it mean, and where do we really stand?*' available at: www.drugdiscovery.net

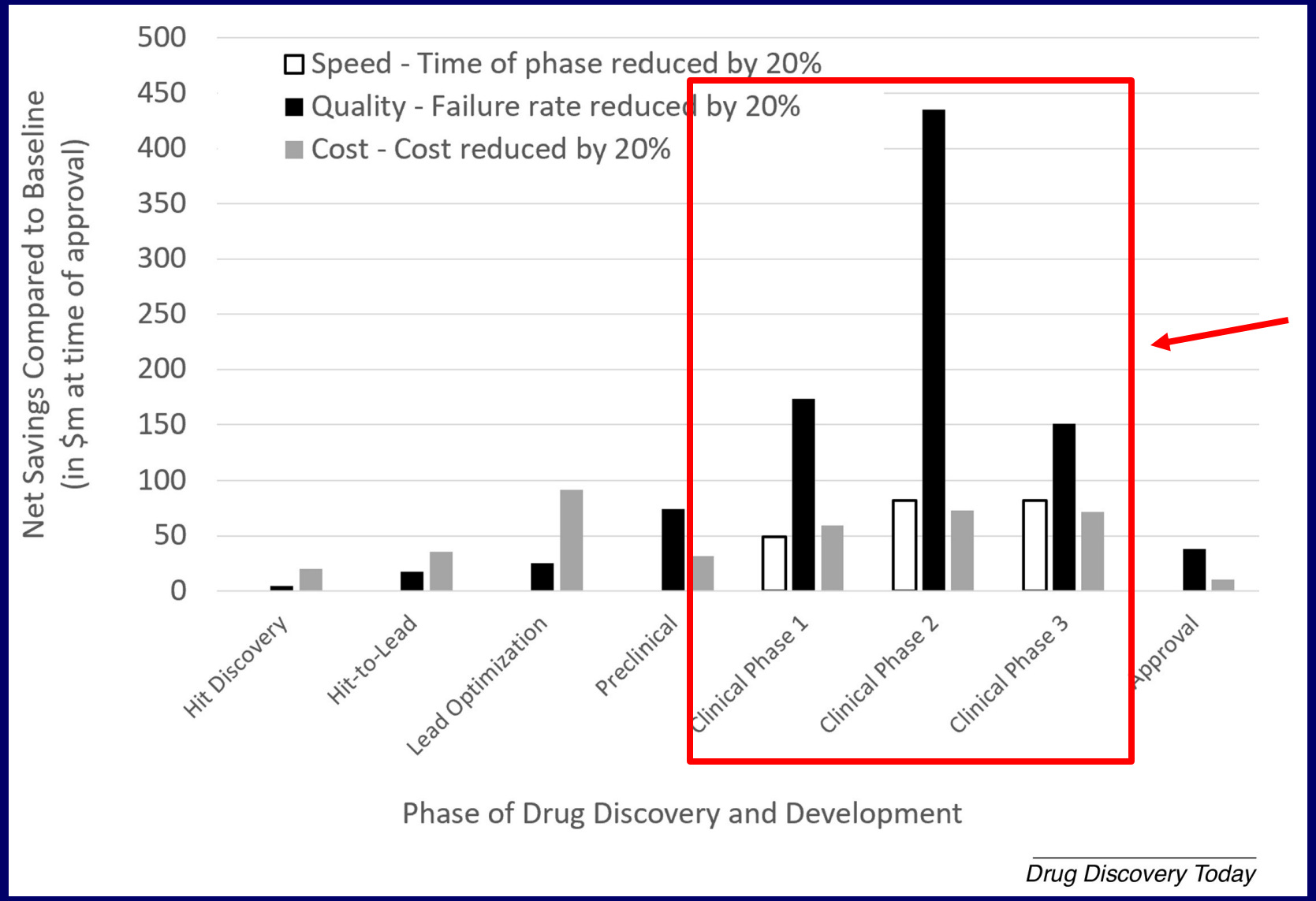
Contents

WHAT MATTERS: Clinically relevant decisions, related to (best) efficacy and (sufficient) safety



Bender and Cortes, Drug Discovery Today 2021

WHAT MATTERS: Clinically relevant decisions, related to (best) efficacy and (sufficient) safety



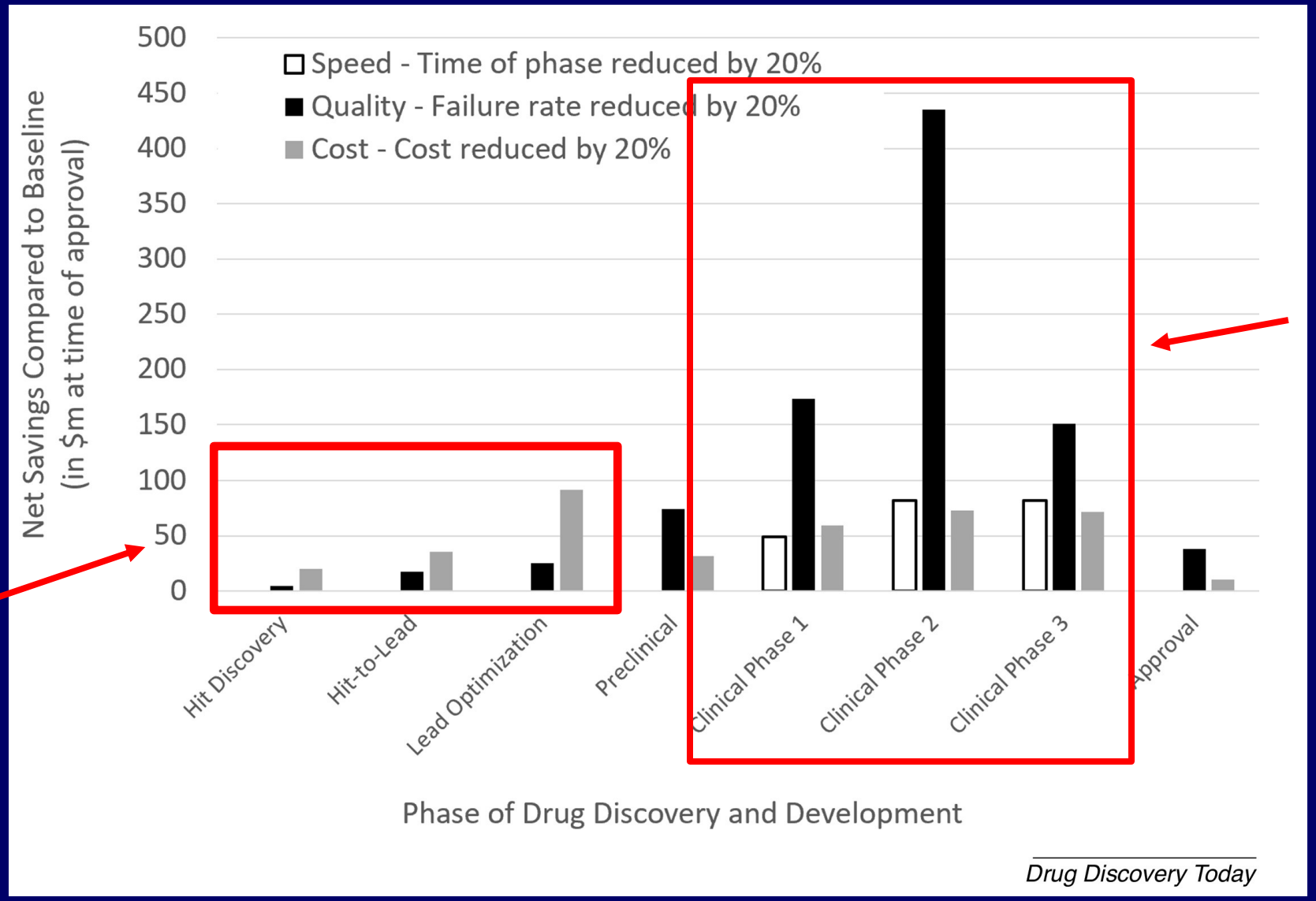
What matters is the right compound in the right patient (dosed in the right way)

Everything else matters less. *Far* less.

Bender and Cortes, Drug Discovery Today 2021

WHAT MATTERS: Clinically relevant decisions, related to (best) efficacy and (sufficient) safety

Many early 'AI in drug discovery' start-ups put their focused here - more data; but (much) less to gain



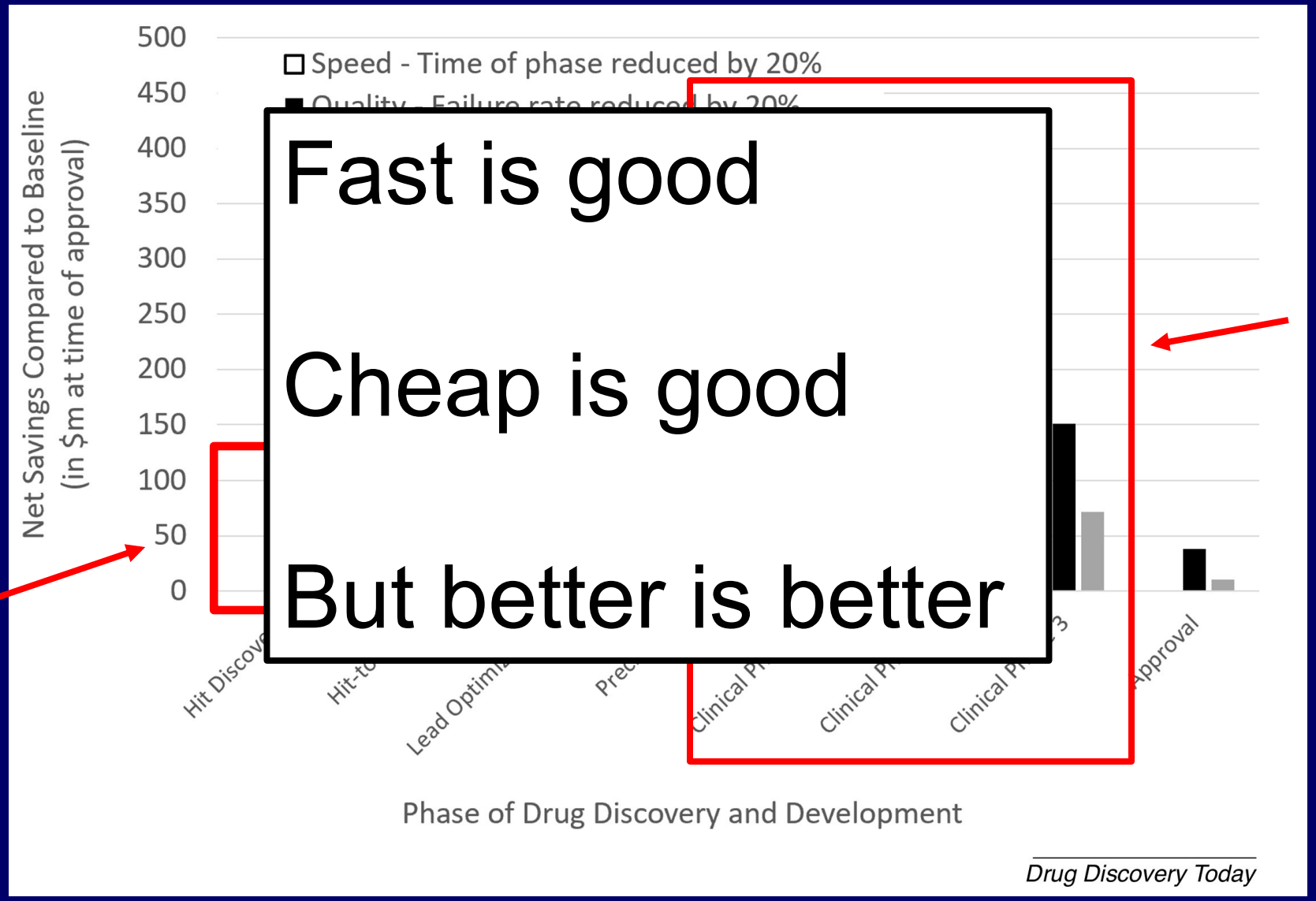
What matters is the right compound in the right patient (dosed in the right way)

Everything else matters less. *Far* less.

Bender and Cortes, Drug Discovery Today 2021

WHAT MATTERS: Clinically relevant decisions, related to (best) efficacy and (sufficient) safety

Many early 'AI in drug discovery' start-ups put their focused here - more data; but (much) less to gain



What matters is the right compound in the right patient (dosed in the right way)

Everything else matters less. *Far* less.

Bender and Cortes, Drug Discovery Today 2021

1. Brief state of affairs: Big headlines (like in the 80s, or 2000s…)

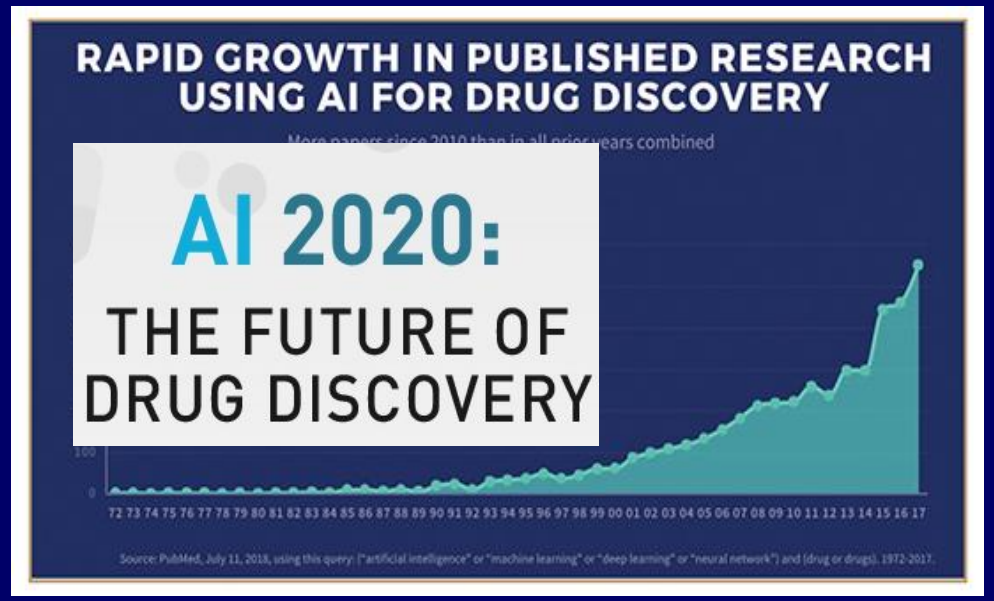
1. Brief state of affairs: Big headlines (like in the 80s, or 2000s…)

Fortune cover 1981

1. Brief state of affairs: Big headlines (like in the 80s, or 2000s…)

Fortune cover 1981

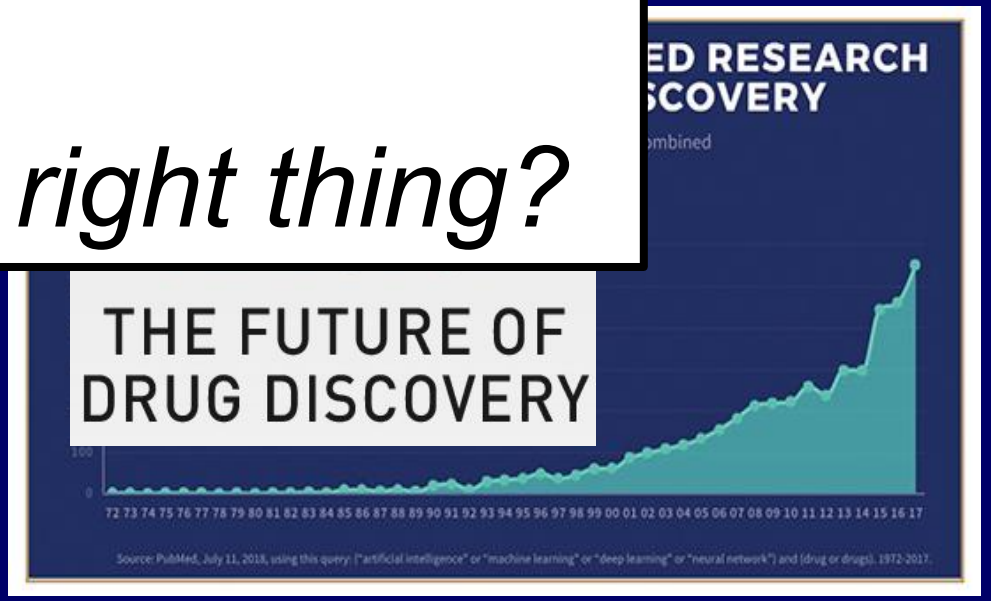
Recent headlines (2018-today)



1. Brief state of affairs: Big headlines (like in the 80s, or 2000s…)

Fortune cover 1981

Recent headlines (2018-today)



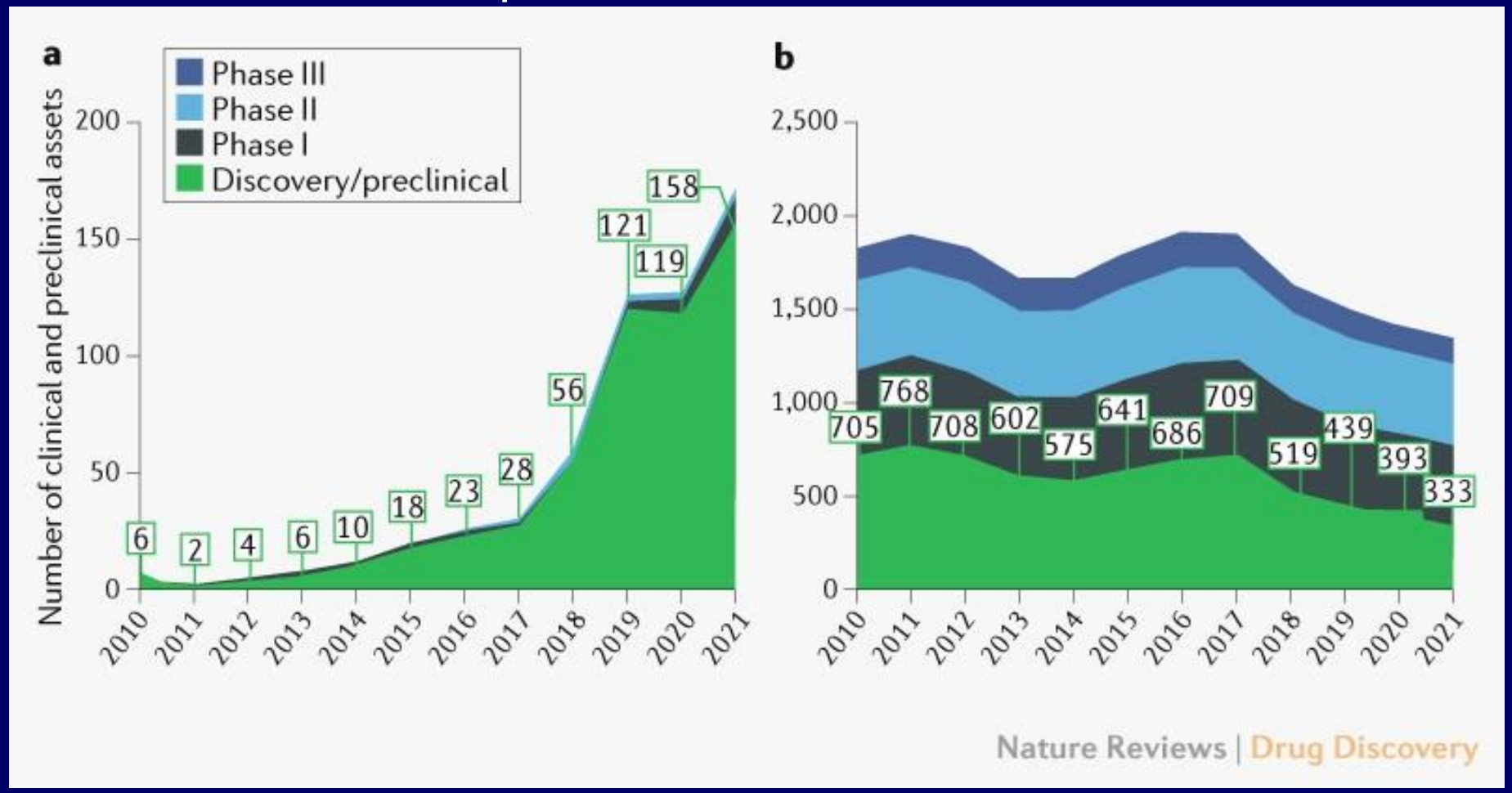
We can do things *right* and *fast* (in many cases)

But do we do *the right thing?*

… little translation into the clinic, and clinical success, yet

Significant *number* of *discovery/ preclinical* programs of AI companies (~160 vs ~330)

Very little Phase 1, less Phase 2, 1 in Phase 3 (2023)



-> Little *in vivo* safety (Phase 1) data *yet*; virtually no *in vivo* efficacy (Phase 2/3) data *yet*

Jayatunga et al., AI in small-molecule drug discovery: a coming wave? *Nature Reviews Drug Discovery* 7 Feb 2022
Most recent: 23 June 2023 Wellcome/BCG Report "Unlocking the Potential of AI in Drug Discovery"

… the great awakening … yes, drug discovery is difficult!

INSIDE THE NASCENT INDUSTRY OF AI-DESIGNED DRUGS

Artificial intelligence tools are beginning to upend the drug discovery pipeline, with several new compounds entering clinical trials. **By Carrie Arnold**

- "**There's no shortcuts to drug discovery**. We can have better informed ideas, but you still have to go through the rest of the [development] process."

- "**These trials are still in their early days** […] **he is confident** that the use of AI is leaving an indelible mark on drug development and promises to make the process better, faster, and cheaper, as well as enabling the development of more first-in-class compounds."

Arnold, Nature Medicine, 1 June 2023, "Inside the nascent industry of AI-designed drugs"

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"

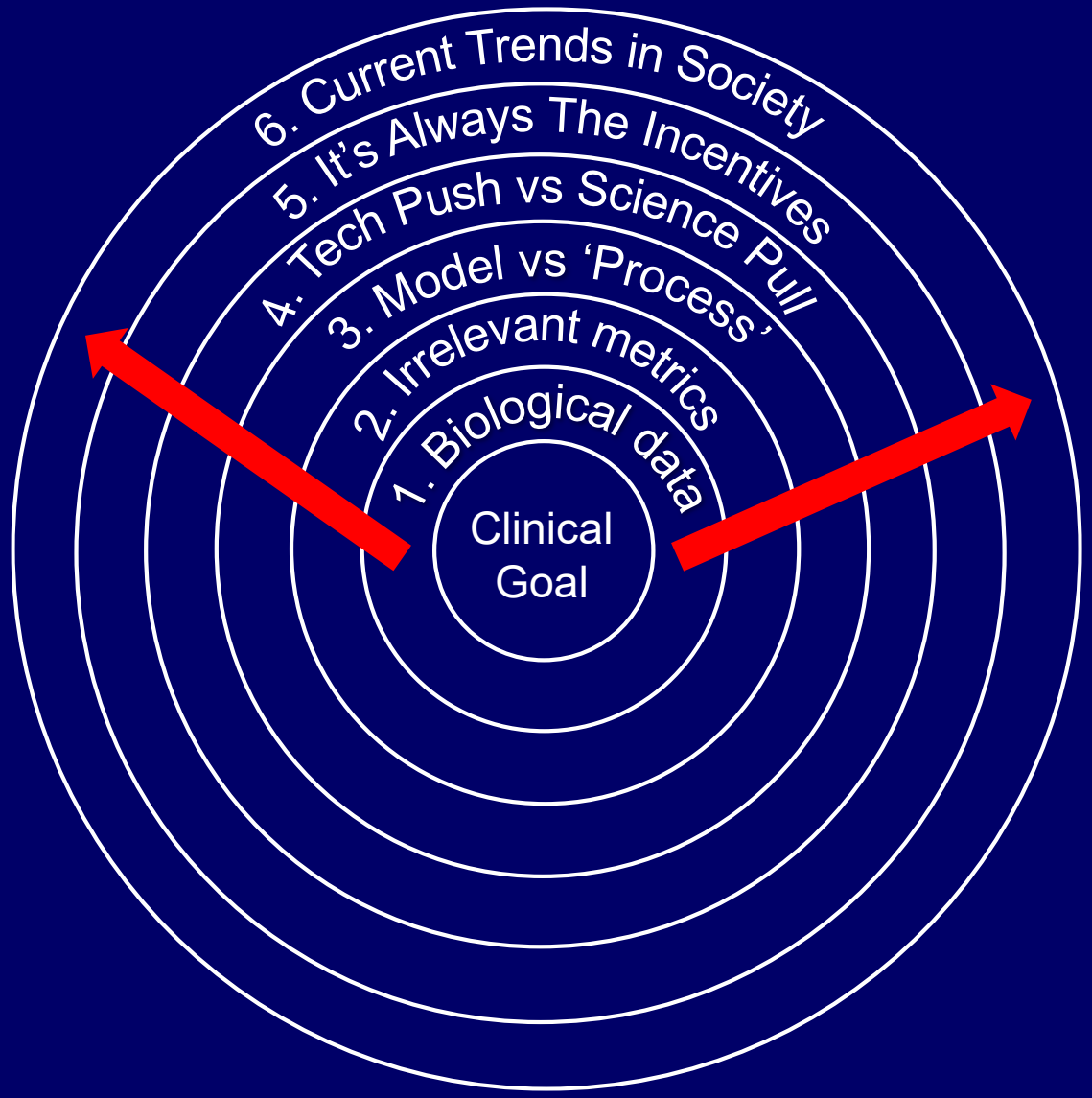
Recent (2017) estimates of drug targets put the number currently at around *667*

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?

Back to the Six Circles of Pleasure



1. Biological data: a) Machine learning needs labels b) Biological data is impossible/very hard to label

1. Biological data: a) Machine learning needs labels b) Biological data is impossible/very hard to label

- *"Does* drug Y cause adverse reaction Z? Yes, or no?"

1. Biological data: a) Machine learning needs labels b) Biological data is impossible/very hard to label

- **“*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
  - *If co-administered with a drug from class C*
  - More frequently in *males* and
  - Only *long-term*
  - (Etc.)

1. Biological data: a) Machine learning needs labels b) Biological data is impossible/very hard to label

- **“*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
  - *If co-administered with a drug from class C*
  - More frequently in *males* and
  - Only *long-term*
  - (Etc.)

- **So – does drug Y cause adverse event Z?**

1. Biological data: a) Machine learning needs labels b) Biological data is impossible/very hard to label

- ***"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
  - *If co-administered with a drug from class C*
  - More frequently in *males* and
  - Only *long-term*
  - (Etc.)

> Mil Med. 2002 May;167(5):432-4.

Death by water intoxication

John W Gardner [1]

Affiliations + expand

PMID: 12053855

- **So – does drug Y cause adverse event Z?**

1. Biological data: a) Machine learning needs labels b) Biological data is impossible/very hard to label

- ***"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
  - *If co-administered with a drug from class C*
  - More frequently in *males* and
  - Only *long-term*
  - (Etc.)

> Mil Med. 2002 May;167(5):432-4.

Death by water intoxication

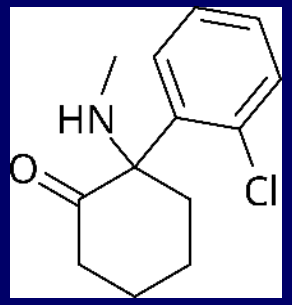
John W Gardner [1]

Affiliations + expand

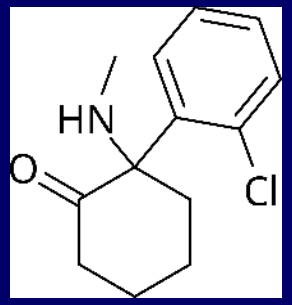
PMID: 12053855

- **So – does drug Y cause adverse event Z? Is water now toxic?**

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



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



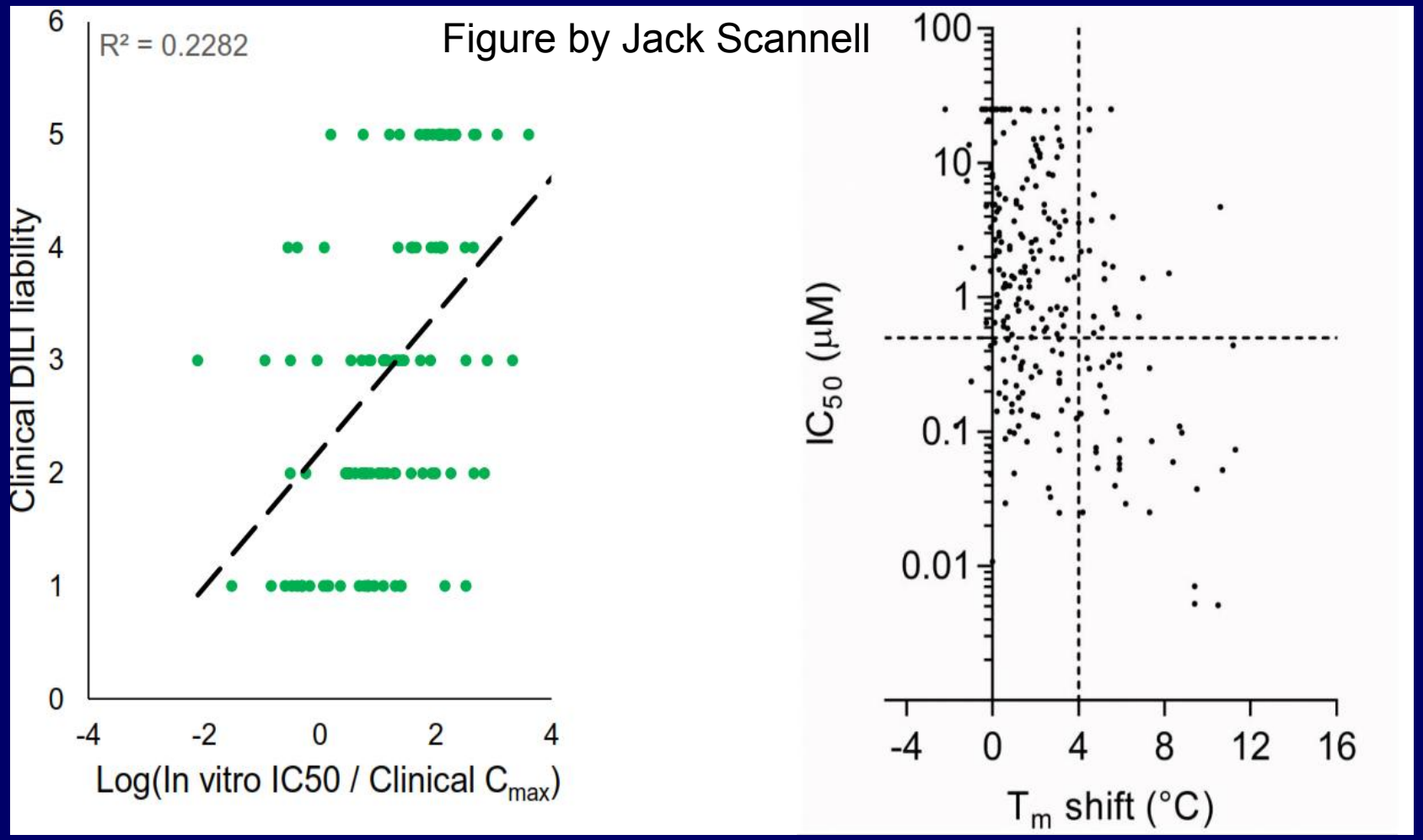
- Ketamine both used as (rather safe) anaesthetic (iv 2mg/kg), approved since 1970, as well as a street drug
- In 2000 effect as antidepressant, when dosed significantly lower, also bronchodilator (acute asthma); iv 0.5mg/kg
- Ketamine long been thought to act via blocking the NMDA receptor - *but* other NMDA blockers such as memantine and lanicemine have not been successful in clinical trials (as antidepressants)
- 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, …)

If it's not in the data (or hidden by conditionality!), it won't be in the model!

Das, J. Repurposing of Drugs–The Ketamine Story. *J. Med. Chem.* 2020

Illustration of low predictivity of much of our data, and hence labels, and hence models



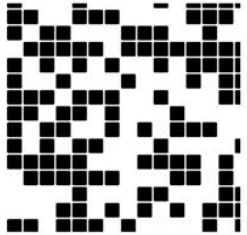
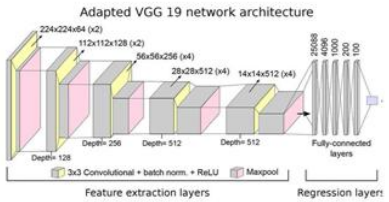
Left: Clinical DILI liability related to Cmax-corrected organoid-derived IC50 values, with low correlation between both values (lower liability index values indicate higher clinical liability)

Right: Low mutual information of enzymatic and thermal-shift derived activity data.

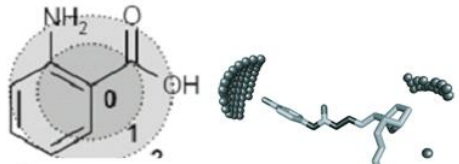
Solely feeding such data with low predictivity into 'AI' models will not lead to better individual decisions, and hence clinical outcomes.

Proctor WR *et al.*. Utility of spherical human liver microtissues for prediction of clinical drug-induced liver injury. Arch Toxicol. 2017 Aug;91(8):2849-2863.

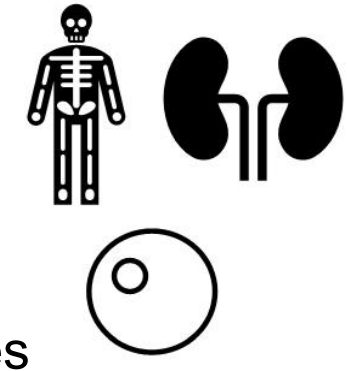
Rudolf AF *et al.*. A comparison of protein kinases inhibitor screening methods using both enzymatic activity and binding affinity determination. PLoS One. 2014 Jun 10;9(6):e98800.

	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

Object | Representation | Model | Object Label

Integration of object, its representation to the ML model, modelling approach, and *unconditional* labels as target annotations -> good for machine learning

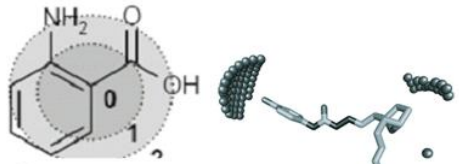
Adapted VGG 19 network architecture

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 = ...

?

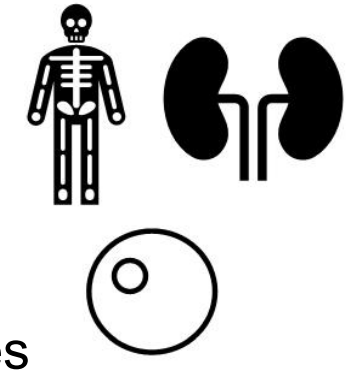
Artificial Neural Network/DNN? Support Vector Machine? Random Forest? Bayesian Classifier?...

Both representation and modelling approach are *largely trial and error* (and *not* intrinsic to the chemical domain)

Property A

Conditional labels (eg dependent on assay system, genotype, dose, endotype, sex, age, comedications, lifestyle, ...) ?

Drug Discovery: Biological Domain



?

Transcript-/proteomics? High-content imaging? Epigenetics? Histopathology? ....

?

Artificial Neural Network? Support Vector Machine? Random Forest?

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)

State/Effect B

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

Bender & Cortes
Drug Discovery Today 2021

Object | Representation | Model | Object Label

Integration of object, its representation to the ML model, modelling approach, and *unconditional* labels as target annotations -> good for machine learning

Adapted VGG 19 network architecture

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

Largely *Unconditional* labels

Drug Discovery: Chemical Domain

Artificial Neural Network/DNN? Random Forest? Bayesian Classifier?...

Molar Refractivity = ...

Property A

Worse integration of object, representation to the ML model, modelling approach, and *conditional* labels as target annotations -> bad for machine learning

Both representation and modelling approach are *largely trial and error* (and *not* intrinsic to the chemical domain)

Drug Discovery: Biological Domain

State/Effect B

Worse integration of object, representation to the ML model, modelling approach, and *conditional* labels as target annotations -> bad for machine learning

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

Bottom line

'Our model achieves 93% Performance on This and that Benchmark, which Outperforms SOTA and revolutionizes drug Discovery, for the 1001st time'

… does not really matter – because if the labels are ‘meaningless’ (their context, **and *in vivo* translation**, is not sufficiently considered) then it also does not matter, in practical terms, if you predict them correctly!

Some further aspects of ‘our data’

- Biology doesn’t have a ground truth
- Biology is practically very difficult
- Chemical data is horrendously biased

‘… but… don’t we have AlphaFold, and didn’t it win the *No-bel-Prize*…?’ I hear you say!

- Sure, and kudos to the developers

But problem in drug discovery (as opposed to *structure prediction*) is:

- Ground truth labels very rarely (never?) exist in drug discovery setting: What matters is *in vivo* relevance (!)
- Finding a *ligand* is only a (very) small part of *drug* discovery (we have ~10^7 ligands, but only ~10^3 drugs)
- Also many problems still largely unsolved – conformational changes/out of domain predictions (new chemistry) *etc.*

‘Data’ isn’t the core problem, it’s *how to get there* … my hands-on experience from a project involving ‘real patient data’

Single cell and spatial transcriptomics in squamous cell lung cancer (LUSC), at University of Medicine Cluj-Napoca, Romania:

- Difficult to get samples of sufficient *quantity* (size), due to tumor (in-)accessibility
- Difficult to be sure of clinical diagnosis (cancer type often not known initially), medical history incomplete, patients from across the country, follow-up difficult
- Quality deterioration of *sample* (difficult to really *understand what happens*!)
- Problems with ‘act of sequencing’ (sample preparation to QC)

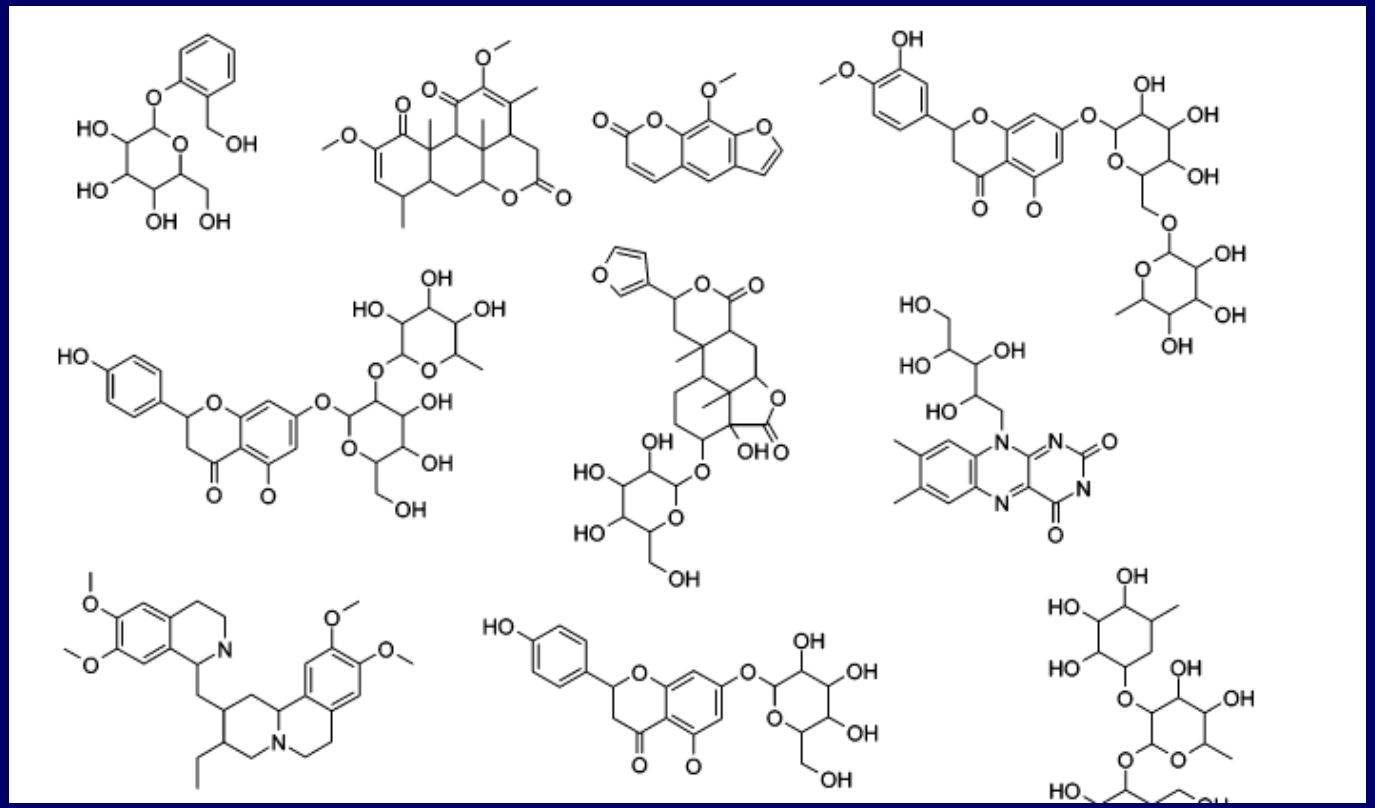
… all this comes *before* any ‘AI’, but it’s the base of all that follows!

… makes me feel that ‘tech/AI discussions’ are a bit detached from reality

…can very much recommend on the cheminformatics path to wisdom to learn about *clinical sample collection, etc.*

Fun fact: So does ‘explainable AI’ actually explain anything? Depends very much on the data, especially in chemistry!

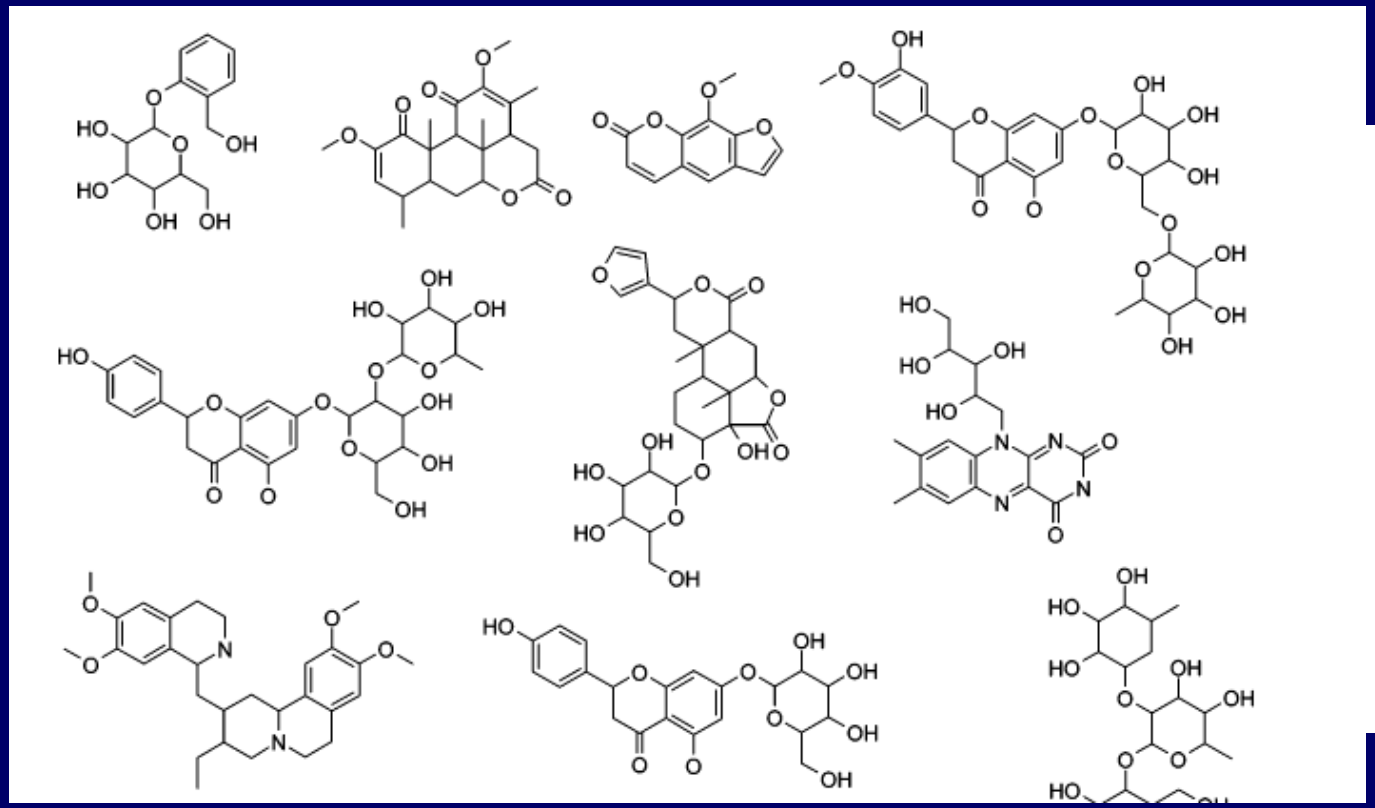
Example from own work: Features of *bitter compounds…*



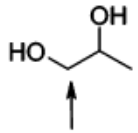
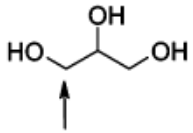
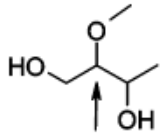
Rodgers et al. JCIM 2006

Fun fact: So does ‘explainable AI’ actually explain anything? Depends very much on the data, especially in chemistry!

Example from own work: Features of *bitter compounds…*



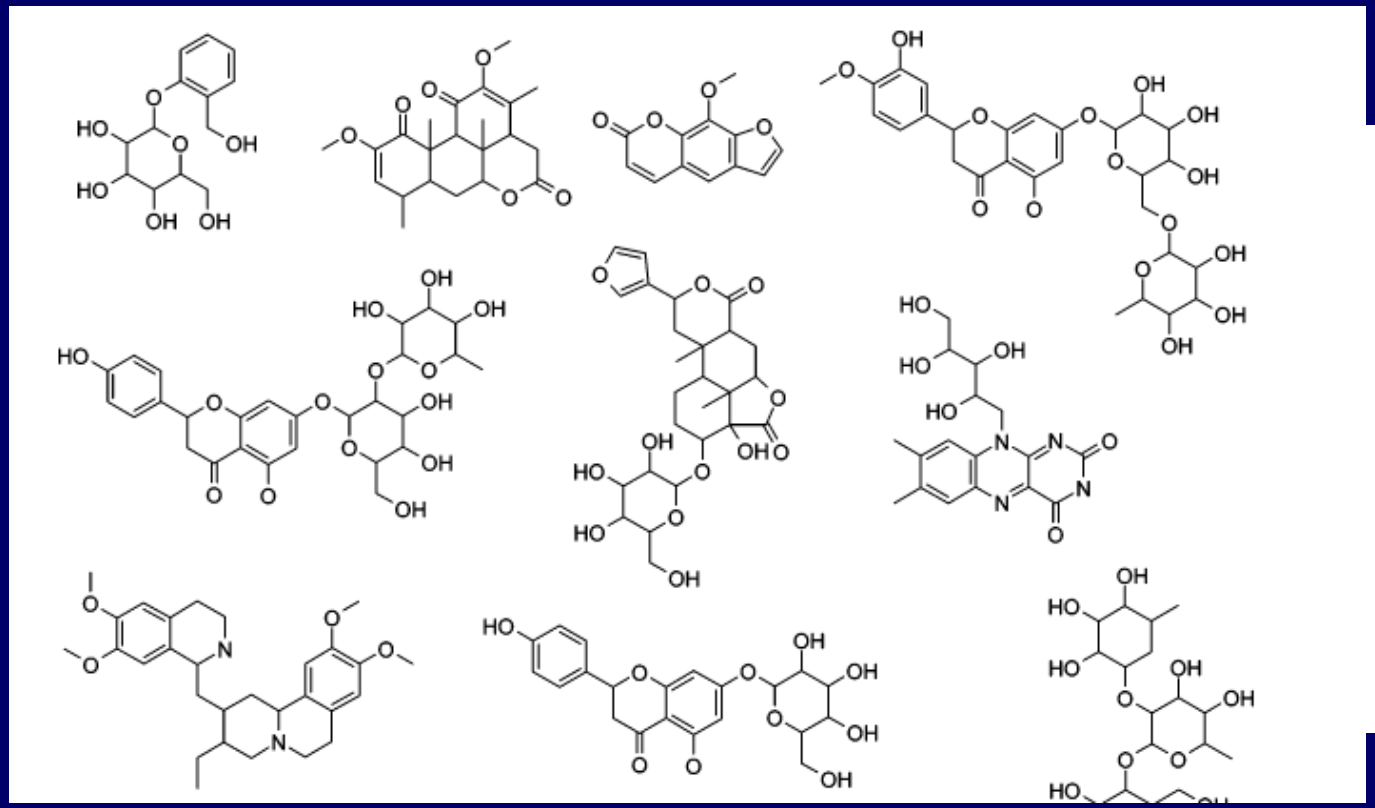
Rodgers et al. JCIM 2006

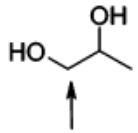
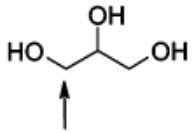
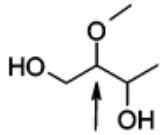
Fragment Number	Fragment	Frequency Bitter Dataset	Frequency Nonbitter Dataset	Information Gain	Relative Frequency Bitter Dataset	Relative Frequency Nonbitter Dataset
1		120	317	0.0139	18.49%	2.34%
2		76	76	0.0139	11.71%	0.56%
3		116	342	0.0125	17.87%	2.53%

… are just glycosylation patterns!

Fun fact: So does ‘explainable AI’ actually explain anything? Depends very much on the data, especially in chemistry!

Example from own work: Features of *bitter compounds…*



Fragment Number	Fragment	Frequency Bitter Dataset	Frequency Nonbitter Dataset	Information Gain	Relative Frequency Bitter Dataset	Relative Frequency Nonbitter Dataset
1		120	317	0.0139	18.49%	2.34%
2		76	76	0.0139	11.71%	0.56%
3		116	342	0.0125	17.87%	2.53%

… are just glycosylation patterns!

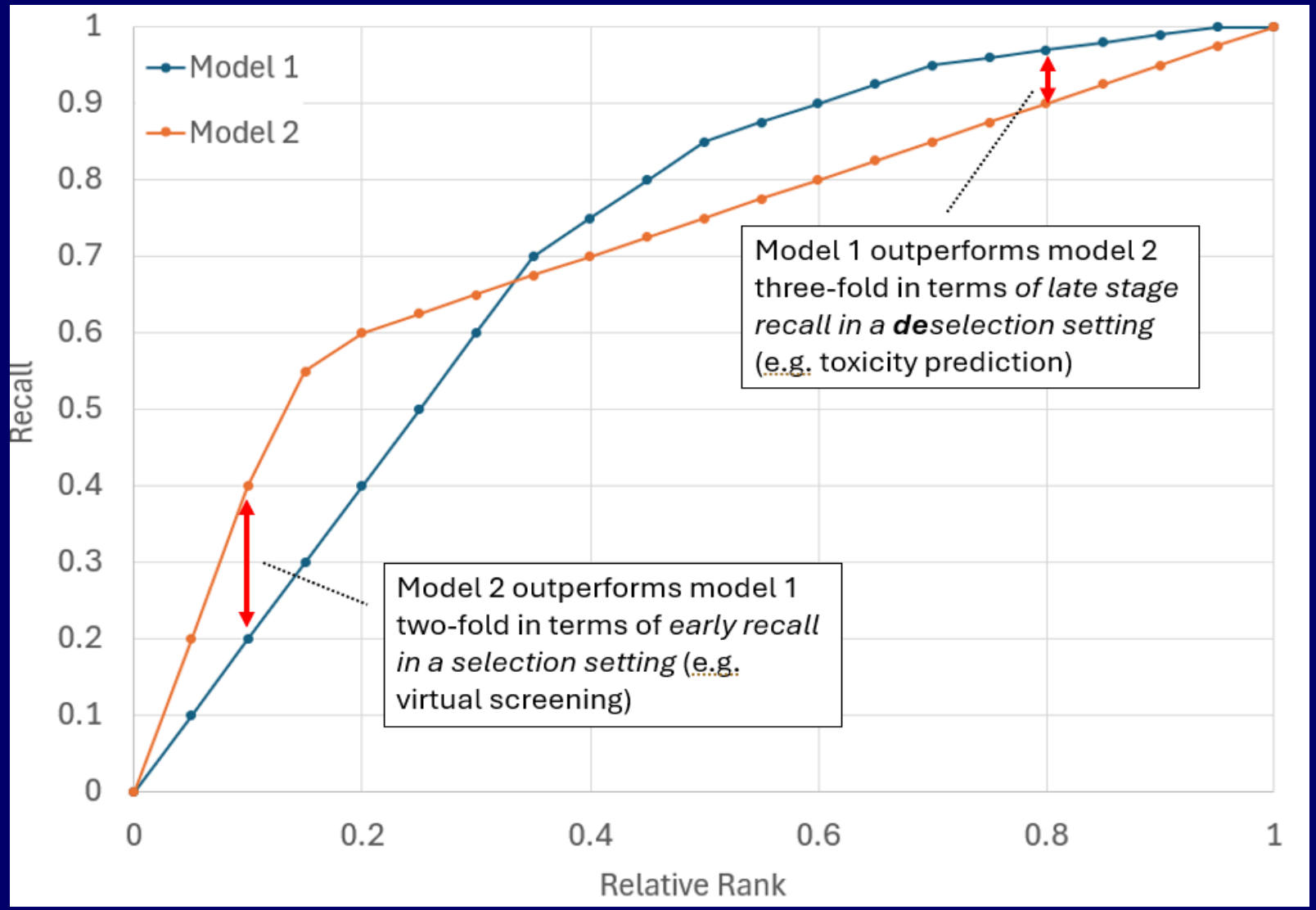
Rodgers et al. JCIM 2006

Problem for all ‘explainable AI’, *in particular very biased chemical data (project bias, synthesis bias, reporting/publication bias, analogue bias, etc.)*

2. Irrelevant metrics: Generic model metrics *never* matter

- Both models have same AUC
- In an early recall setting ('virtual screening') model 2 is 2-fold better than model 1
- In a late-stage deselection setting model 1 is 3-fold better than model 2

- Performance measured without use-case (often AUC, overall accuracy etc) is *generic, and never matters in a use case*
- The 'use case' extends *far beyond* performance metric: Which library is used, which target, etc etc.

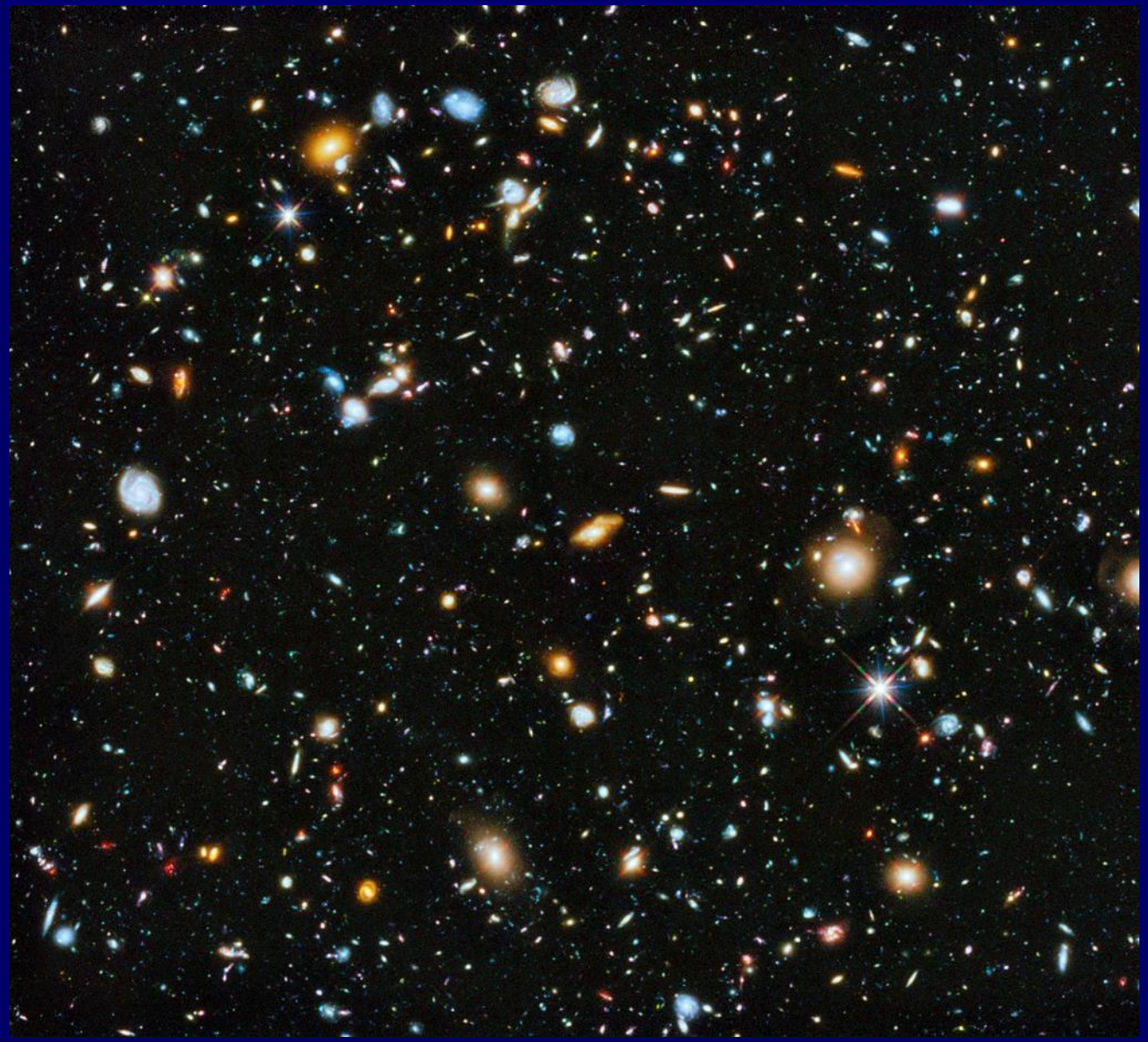


General problem with much published work (and model validation more generally)

- Published models are often ‘models-only’, not presented as deployed and evaluated in production (at least this usually isn’t fully shared)
- Tendency to evaluate models based on distribution-based, not point-based, statistics (e.g. R^2 vs RMSE)
- ‘My AUC is higher than yours’ – ok, and what is your *use case*, and *does AUC matter in your use case?*

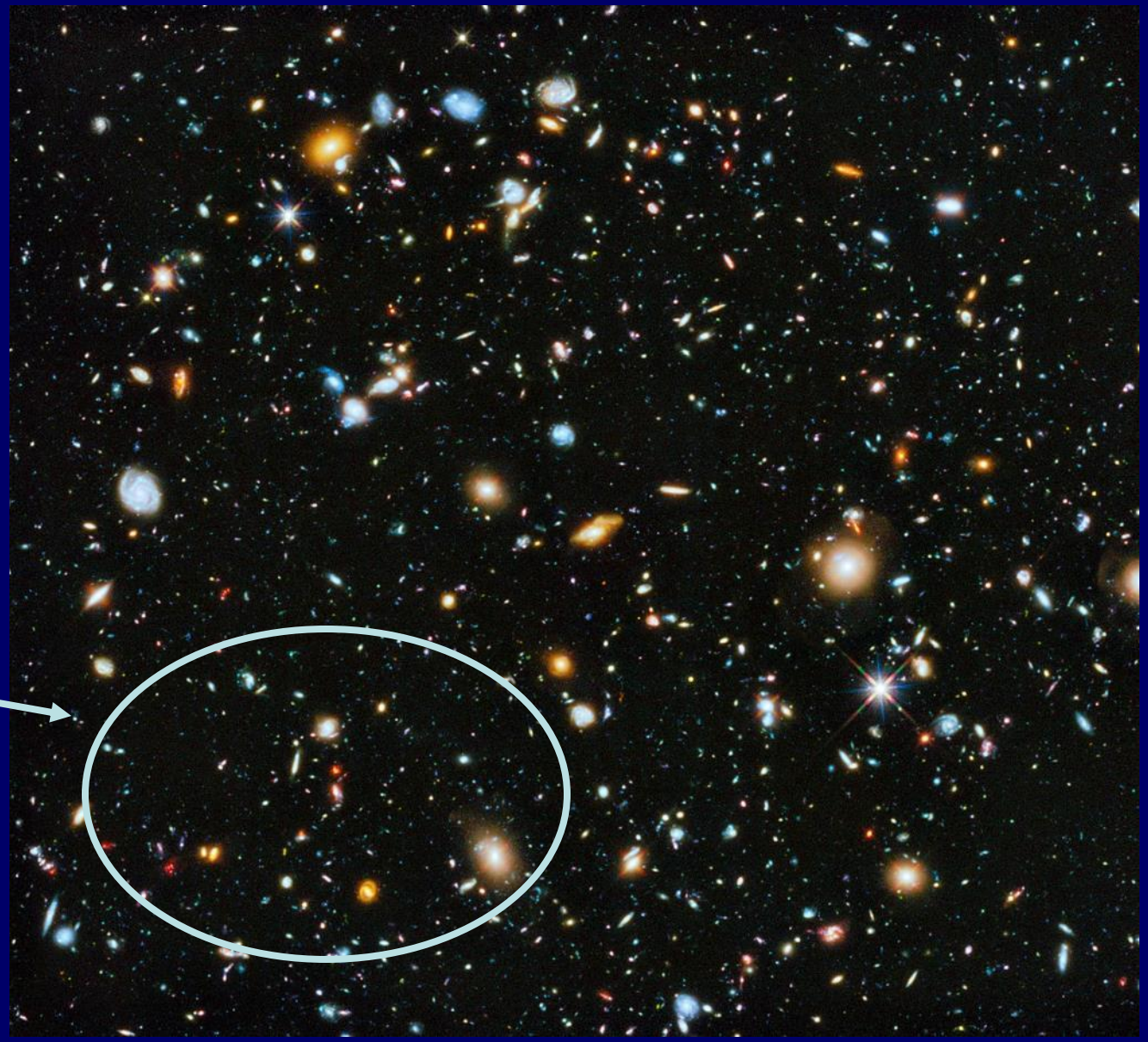
- 90% of what is published doesn’t translate to practice (due to this and many other reasons)

2.2. A numerical performance on *one* dataset does not necessarily (and hardly ever!) predict performance on another dataset (=project)



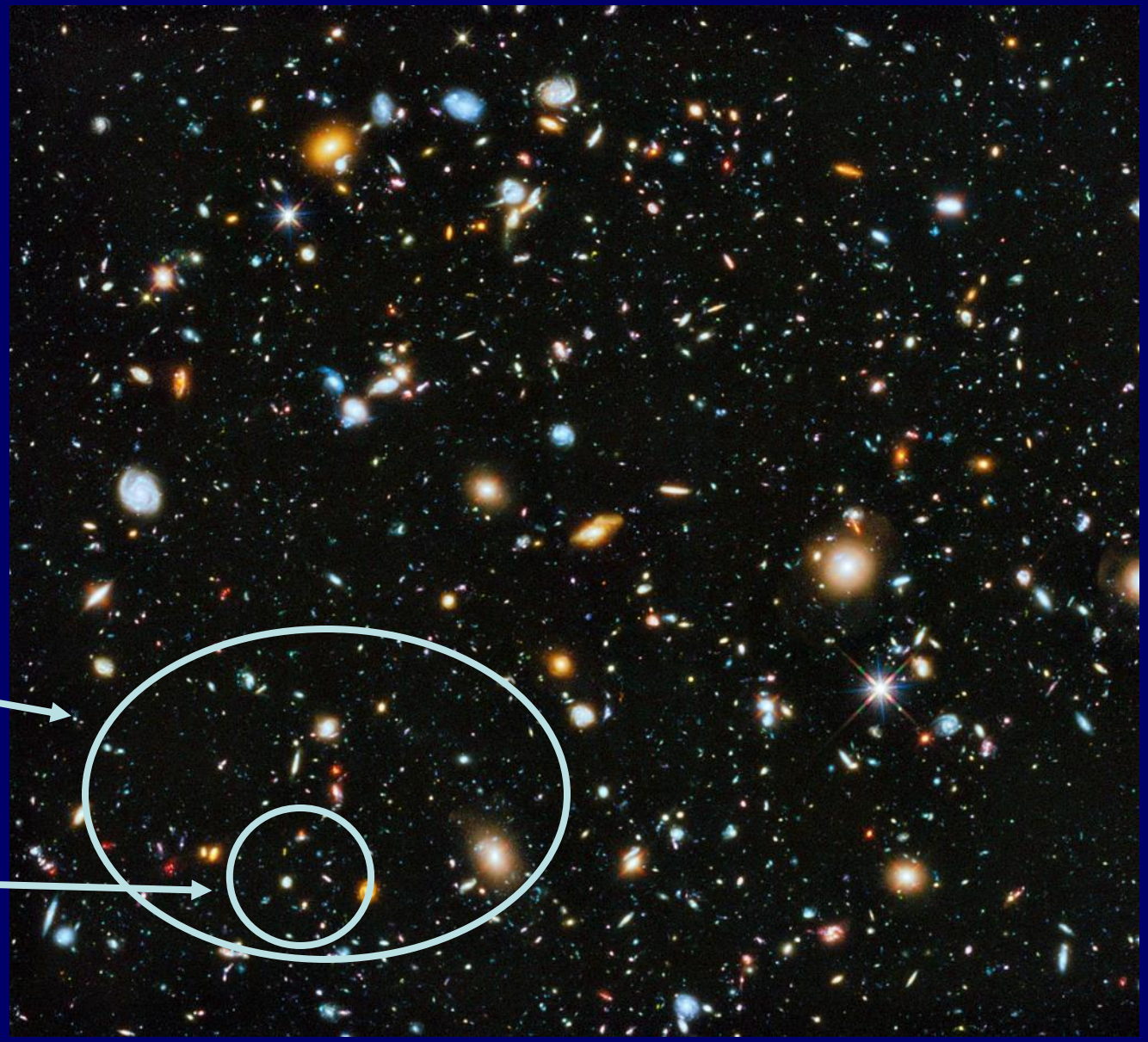
- Chemical space is large; data sets are small
- Retrospective validation, training/test splits… give you performance numbers
- Future projects will *by definition* be outside the training set distribution
- Also time-split doesn't help, it's just different galaxies
- ***Performance measured retrospectively will not hold prospectively (in future projects)***

2.2. A numerical performance on *one* dataset does not necessarily (and hardly ever!) predict performance on another dataset (=project)



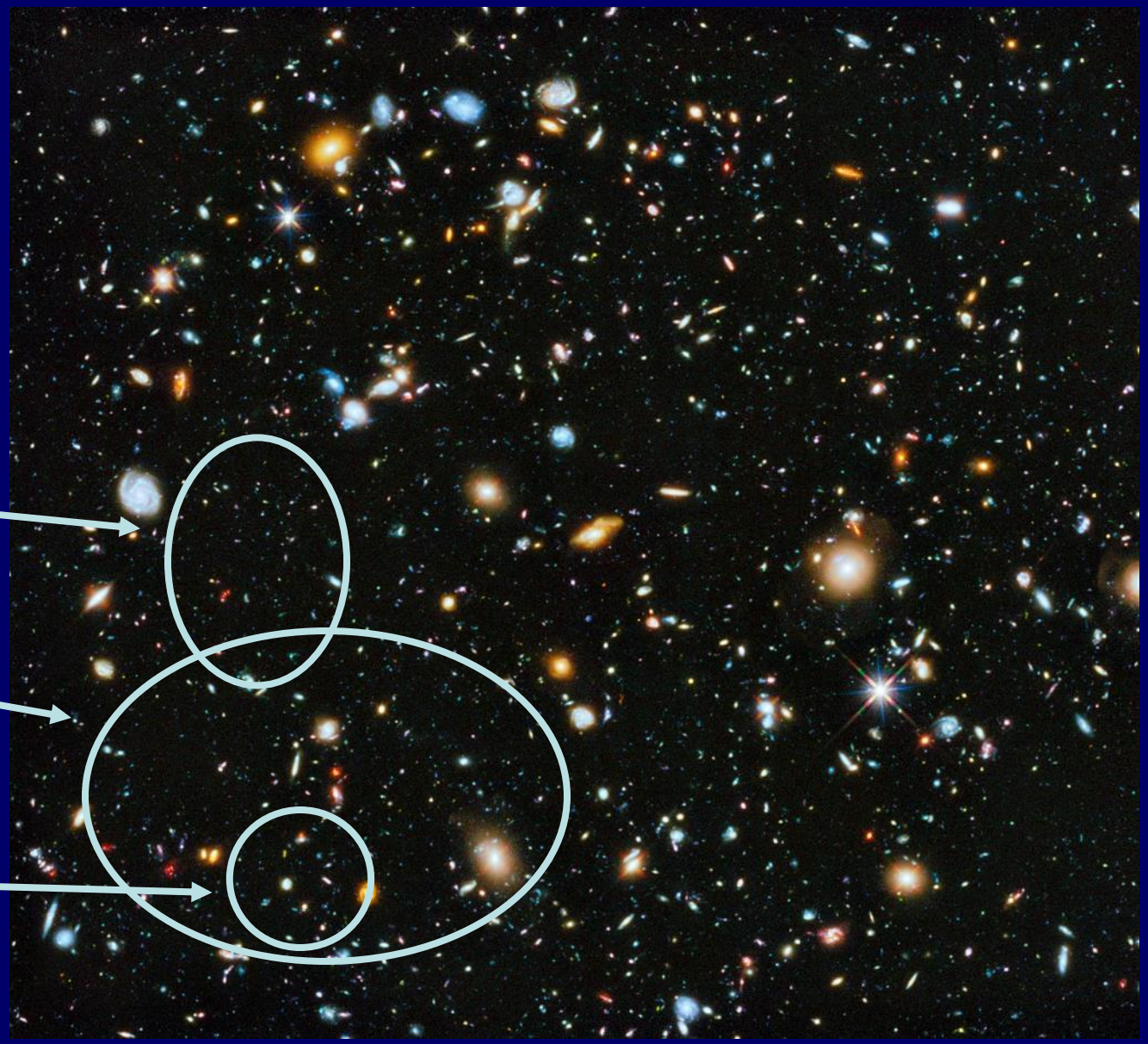
- Chemical space is large; data sets are small
- Retrospective validation, training/test splits… give you performance numbers
- Future projects will *by definition* be outside the training set distribution
- Also time-split doesn't help, it's just different galaxies
- ***Performance measured retrospectively will not hold prospectively (in future projects)***

2.2. A numerical performance on *one* dataset does not necessarily (and hardly ever!) predict performance on another dataset (=project)



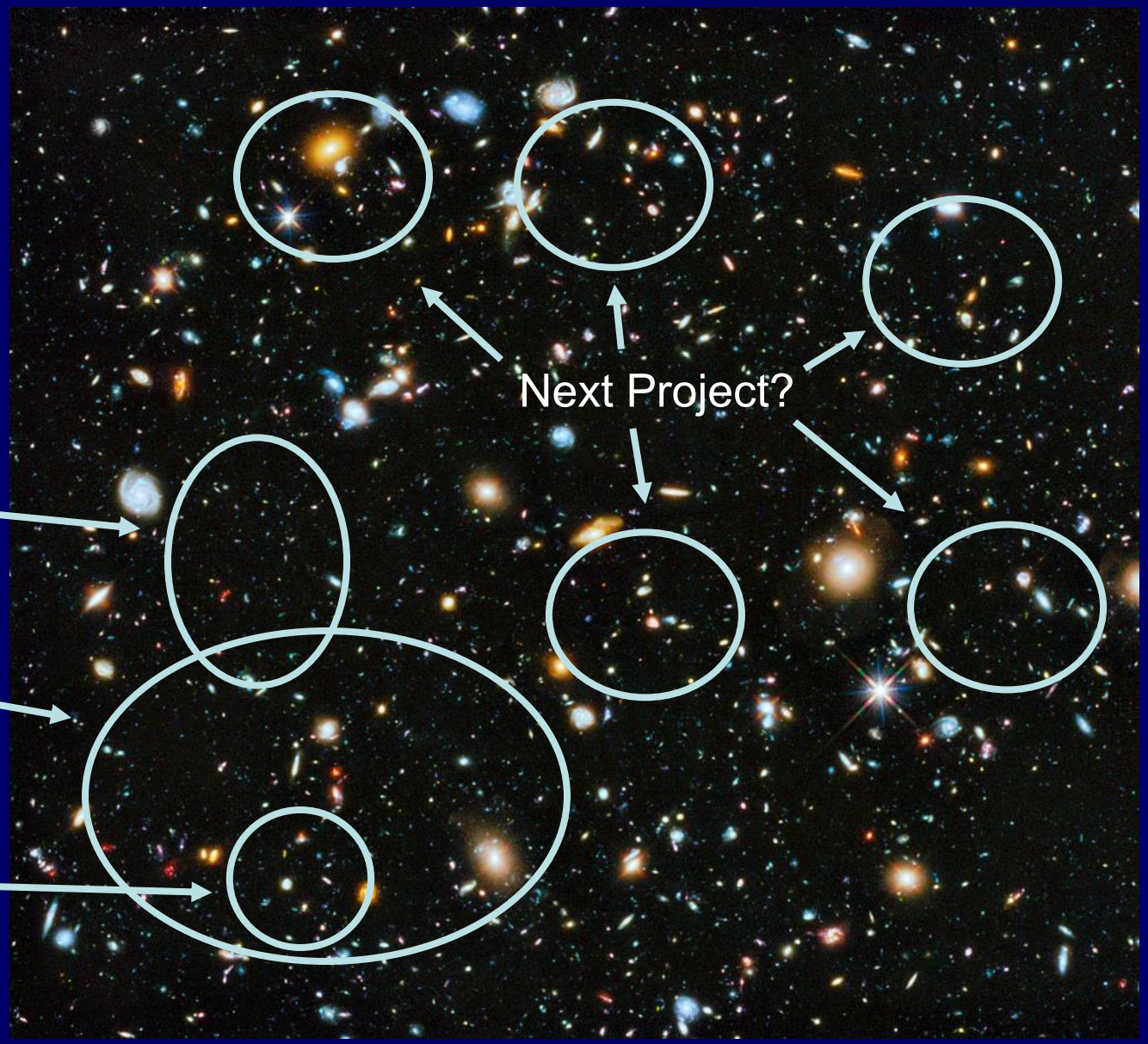
- Chemical space is large; data sets are small
- Retrospective validation, training/test splits… give you performance numbers
- Future projects will *by definition* be outside the training set distribution
- Also time-split doesn't help, it's just different galaxies
- ***Performance measured retrospectively will not hold prospectively (in future projects)***

2.2. A numerical performance on *one* dataset does not necessarily (and hardly ever!) predict performance on another dataset (=project)



- Chemical space is large; data sets are small
- Retrospective validation, training/test splits… give you performance numbers
- Future projects will *by definition* be outside the training set distribution
- Also time-split doesn't help, it's just different galaxies
- ***Performance measured retrospectively will not hold prospectively (in future projects)***

2.2. A numerical performance on *one* dataset does not necessarily (and hardly ever!) predict performance on another dataset (=project)



'External Test Set'

'Training Set'

'Validation Set'

- Chemical space is large; data sets are small
- Retrospective validation, training/test splits… give you performance numbers
- Future projects will *by definition* be outside the training set distribution
- Also time-split doesn't help, it's just different galaxies
- ***Performance measured retrospectively will not hold prospectively (in future projects)***

Major problem: *Absolute* location in chemical space matters, *as does relative change*

- Representations of molecules do not encode all (internal and external) context; mutual dependency of features not covered by data
- Predictions in high-dimensional space always represent out of distribution (OOD) predictions [Balestriero2021]
- Ligand-based prediction models *can* work in some cases:
  - ‘Use-case sufficient’ data; Descriptors ‘use-case sufficiently’ capture underlying trends
  - *Global* models: E.g. logD models with >~10^5 molecules
  - *Local* SAR models where e.g. binding mode is identical

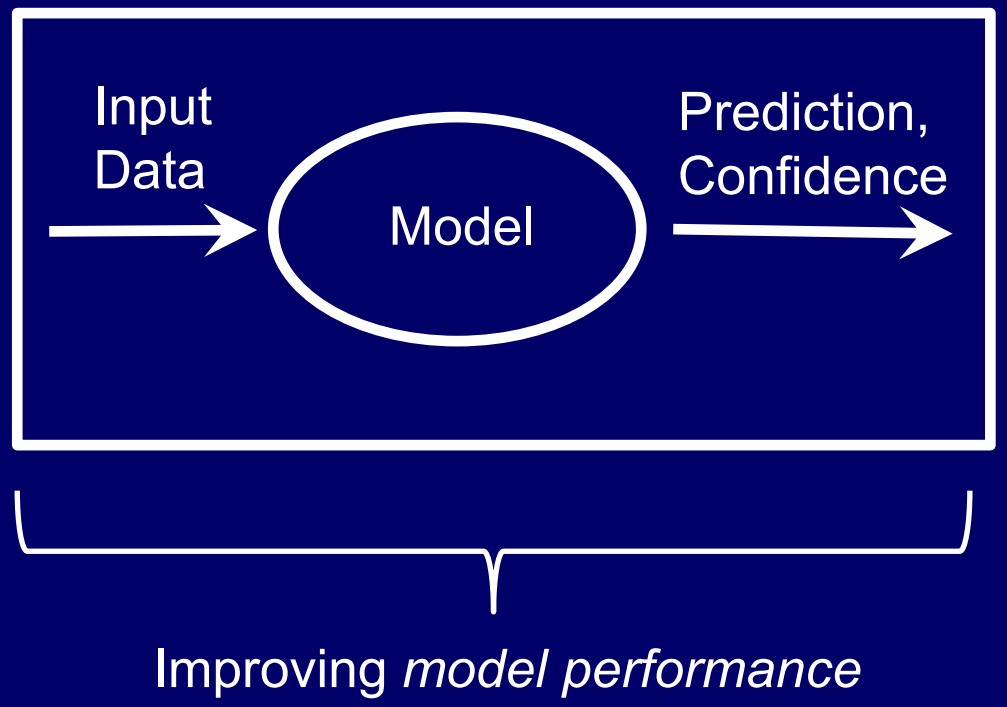
Balestriero et al. Learning in High Dimension Always Amounts to Extrapolation https://arxiv.org/abs/2110.09485 (2021).

Bottom line

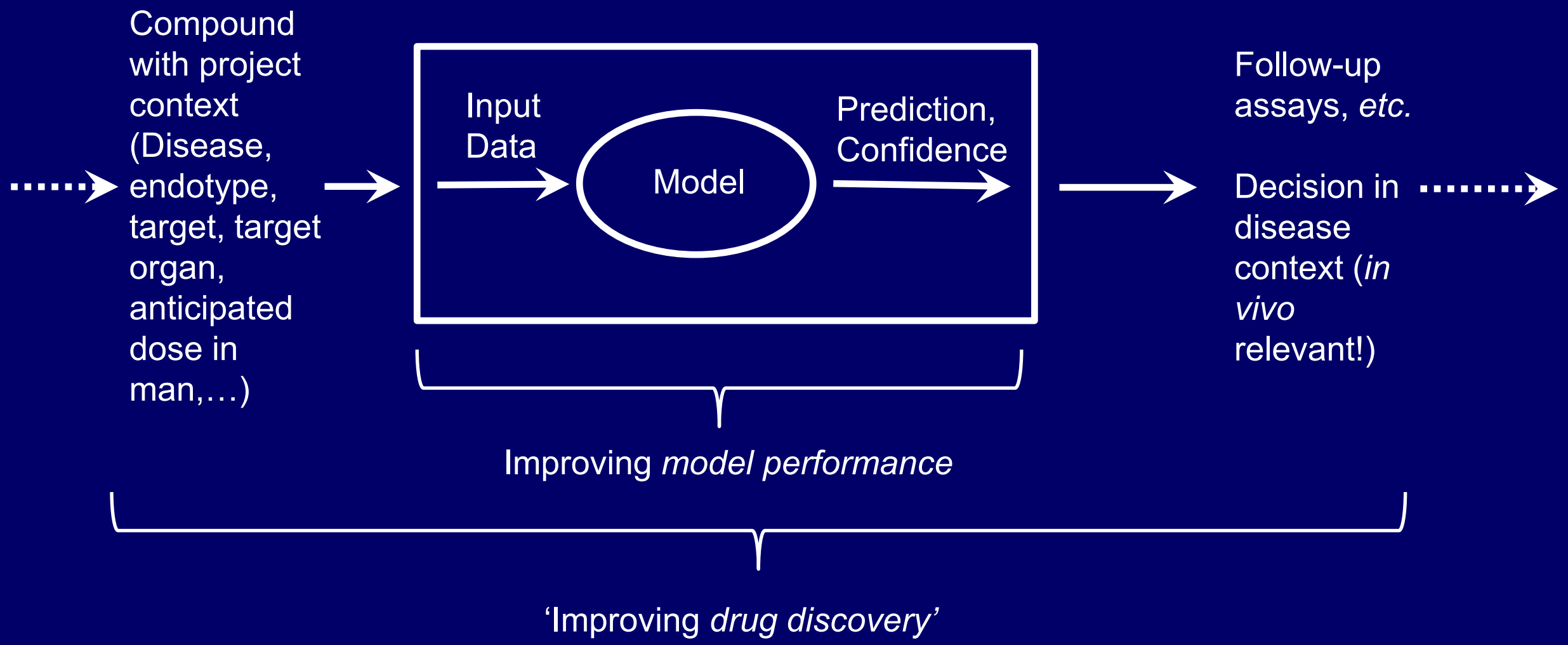
'Our model achieves 93% Performance on This and that Benchmark, which Outperforms SOTA and revolutionizes drug Discovery, for the 1001st time'

… does not really matter – because if the metrics are not fit-for-purpose, then also ‘pumping the numbers’ will not get us there

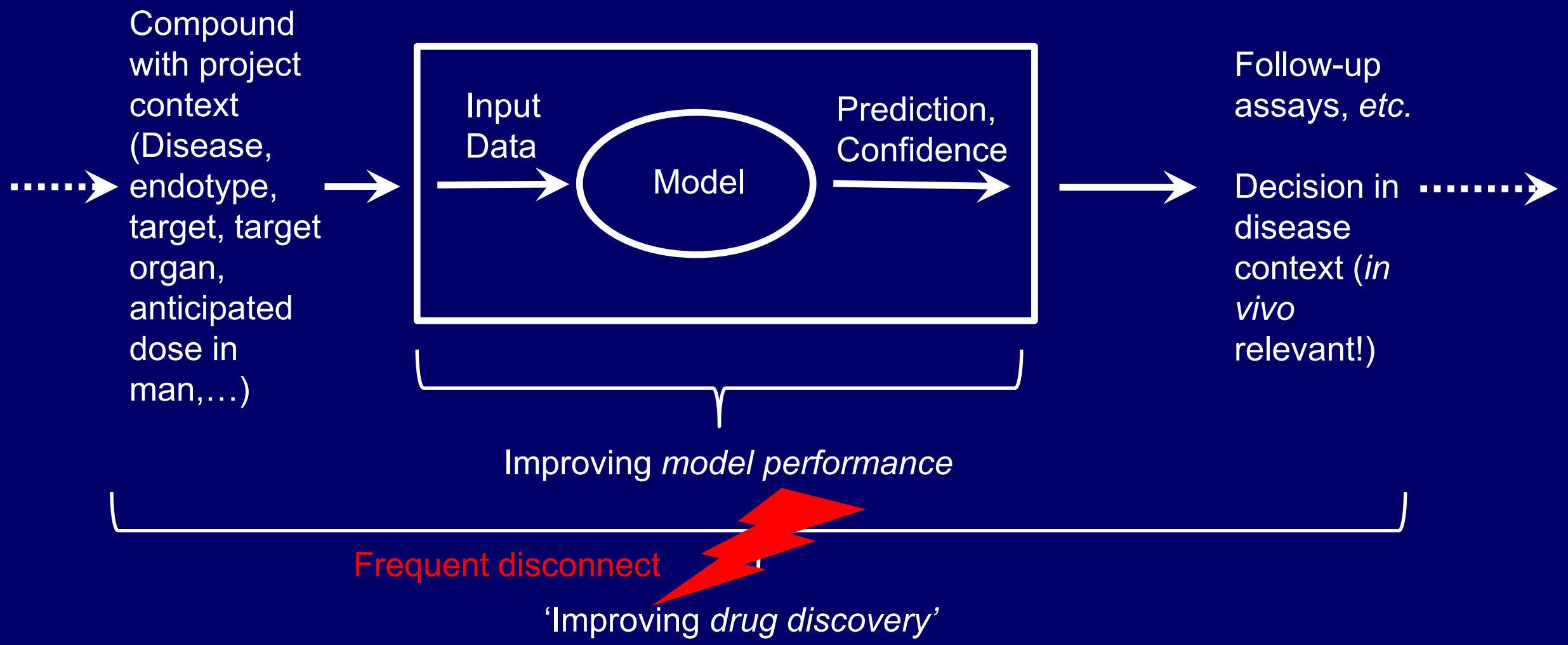
3. Model validation does not matter – it's about the '*process*' (… which only gets validated in the clinic!)



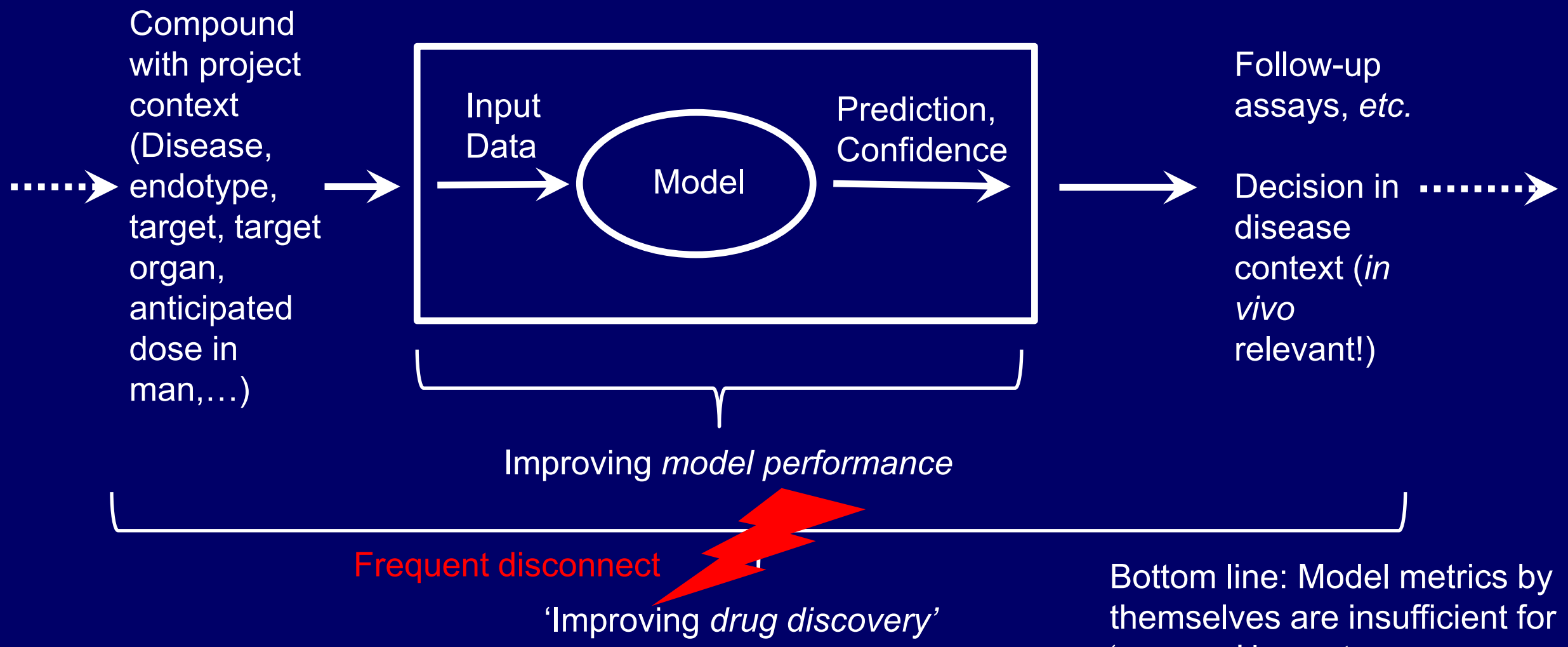
3. Model validation does not matter – it's about the '*process*' (… which only gets validated in the clinic!)



3. Model validation does not matter – it's about the '*process*' (… which only gets validated in the clinic!)



3. Model validation does not matter – it's about the '*process*' (… which only gets validated in the clinic!)

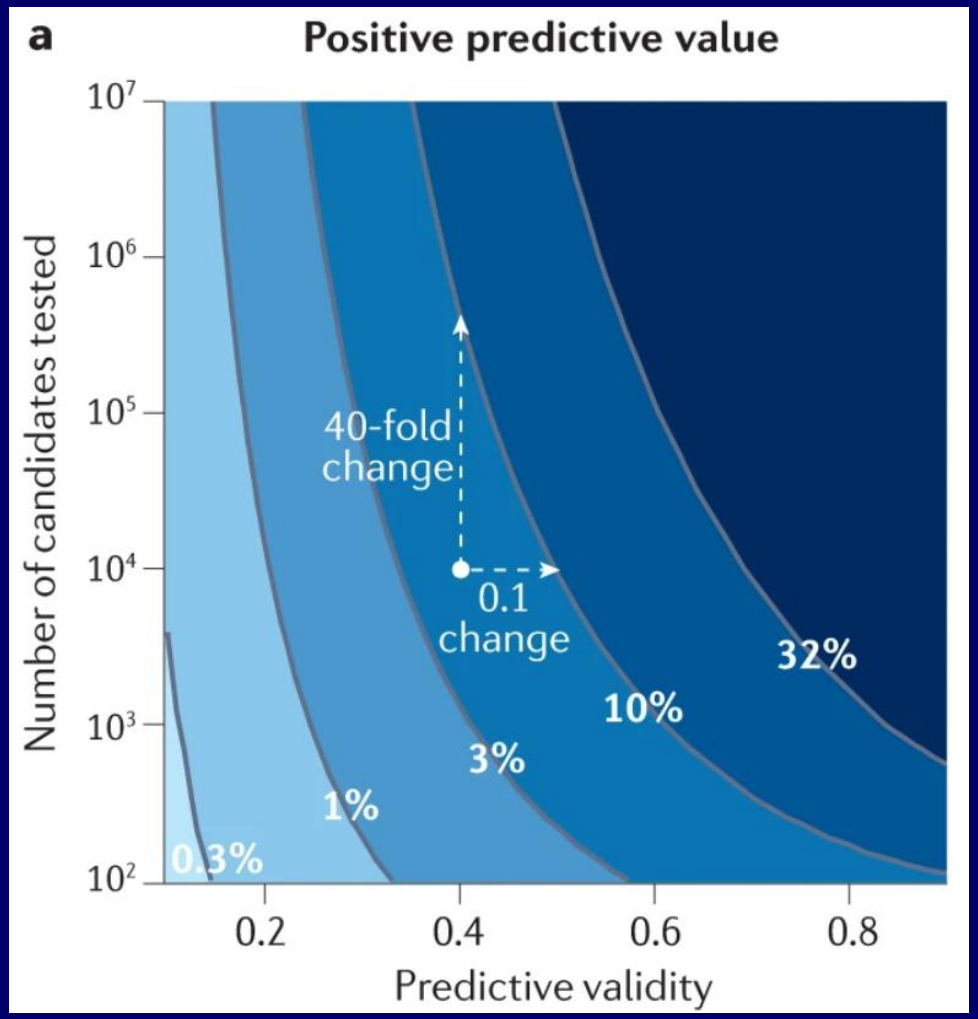


Bottom line: Model metrics by themselves are insufficient for 'process' impact

A 10% better predictive validity is worth ca 10-40x the number of compounds tested (!)

"For much of the parameter space, an absolute 0.1 change in predictive validity (horizontal axis) has a bigger effect on PPV than a 10× change in the number of candidates tested ($\log_{10}$ scale on the vertical axis)."

Note: This does not refer to *model* performance, it refers to predictive validity of the model on the actual endpoint of interest – *process* impact!



Scannell *et al.* **Predictive validity in drug discovery: what it is, why it matters and how to improve it.** Nature Reviews Drug Discovery 2022

Drug discovery is not about ‘tasks’ and ‘leaderboards’… it’s about the end goal, the clinic

- ‘Tasks’ are incomplete representations of the ‘truth’ (‘process’) of drug discovery (and in fact any aspect of life)
- ‘Underspecification’ problem [D’Amour2022] of all ‘ML tasks’
- Tempting to ‘Kaggle a bit’, publish in a ‘high impact journal’ … but tells you nothing about the real world

- Plain theory in cooking, dancing, music, painting ... doesn’t get you there
- Science and the Arts are surprisingly similar here

- *… doing the right thing is more important than doing things right*

D’Amour, A. et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. J. Mach. Learn. Res. 23, 1-61 (2022)

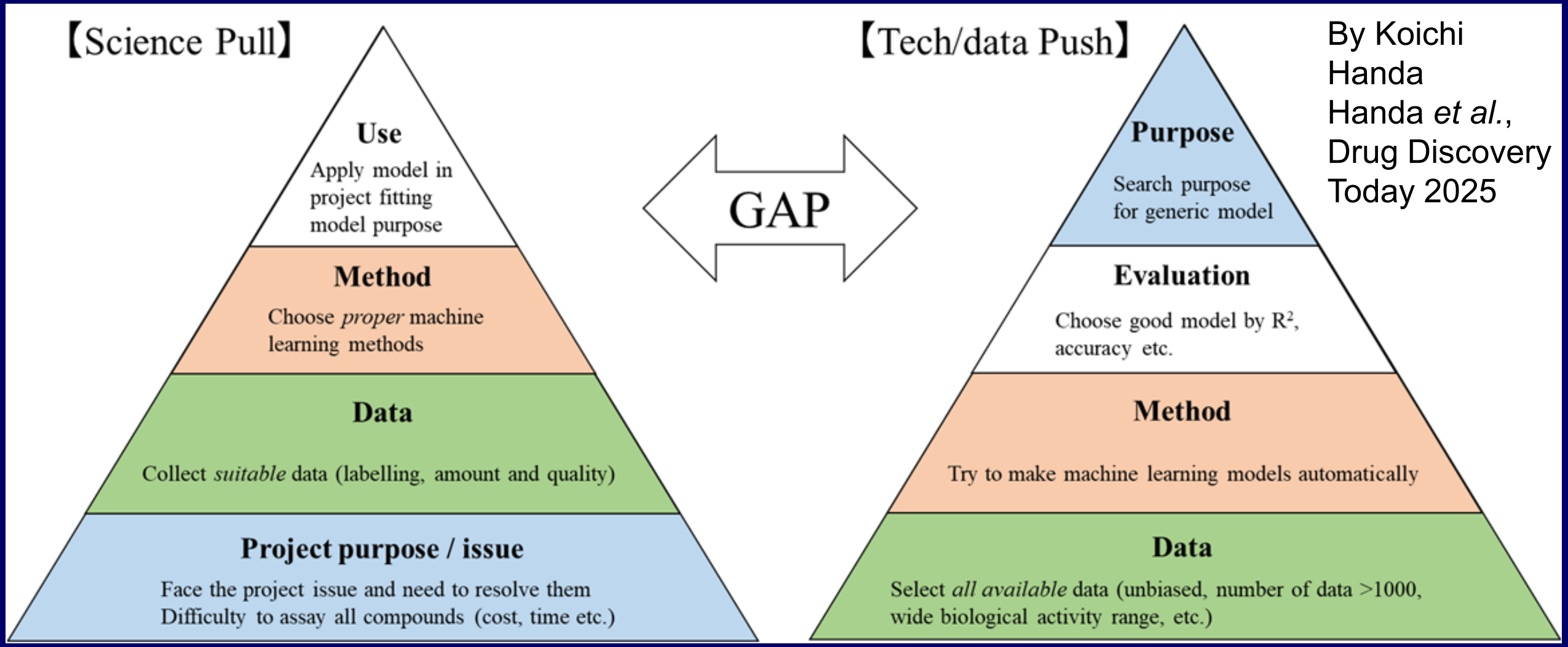
Bottom line

'Our model achieves 93% Performance on This and that Benchmark, which Outperforms SOTA and revolutionizes drug Discovery, for the 1001st time'

… does not really matter – because it is a *model* metric, not a *'process'* metric (… to the extend drug discovery actually is a 'process')

4. The Tech Hammer Looking for the Use Case Nail

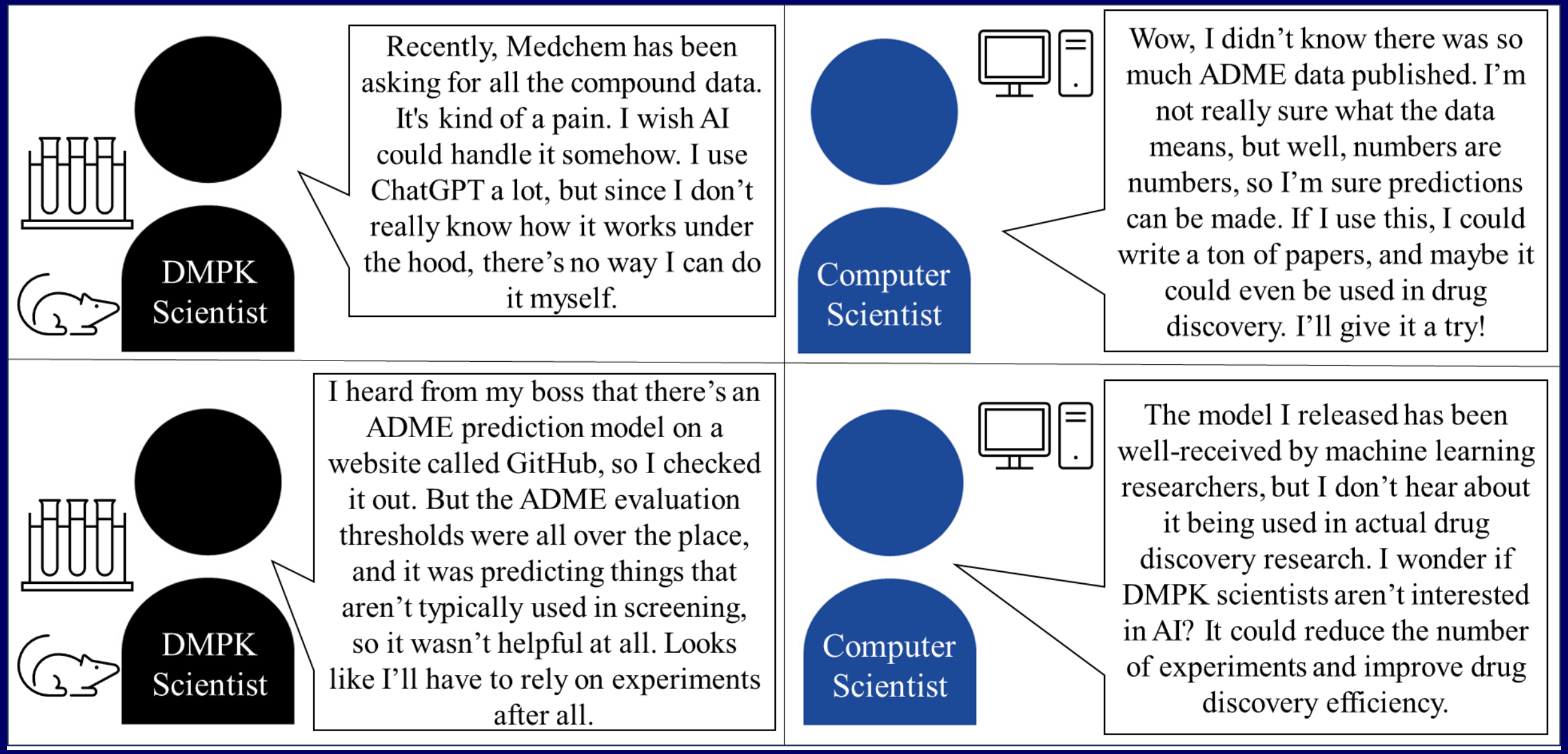
'Tech push' inverts logic of purpose->data->method->use case and can lead to suboptimal results



By Koichi Handa
Handa *et al.*, Drug Discovery Today 2025

The different planets experimentalists and ‘AI-lers’ live on

By Koichi Handa, Handa *et al.*, Drug Discovery Today 2025



5. ‘Our model outperforms…’ It’s Always the Incentives
A (huge) problem in a space without meaningful metrics

Absence of fast feedback on long-term reward function (clinical success), hence optimization on proxies, *e.g.*:

- Big Pharma -> ‘we need a winner’ (we generated TB of data, we now work with DeepLearningAgenticSuper.AI, …)
- Academia -> ‘we published another high-impact paper and improved SOTA, again’ (on entirely irrelevant benchmarks)
- Start-Up Companies -> Stuck in the eternal pain of ‘platform validation’ and pilots
- Grant funding agencies
- Publications

Vicious circle of ‘we fund excellent research’ (overhyped science of the day, published in ‘high-impact journals’, by people who have done it before) and ‘we publish what gets cited’ (as above)

The result of wrong incentives: Lots of hyped Pseudoinnovation

The result of wrong incentives: Lots of hyped Pseudoinnovation

Benchmarking foundation cell models for post-perturbation RNA-seq prediction

Gerold Csendes, Gema Sanz, Kristóf Z. Szalay & Bence Szalai

BMC Genomics **26**, Article number: 393 (2025) | Cite this article

"In this study, we benchmarked two recently published foundation models, scGPT and scFoundation, against baseline models. Surprisingly, we found that **even the simplest baseline model—taking the mean of training examples—outperformed scGPT and scFoundation.**"

The result of wrong incentives: Lots of hyped Pseudoinnovation

Google DeepMind

Millions of new materials discovered with deep learning

Benchmarking foundation cell models for post-perturbation RNA-seq prediction

Gerold Csendes, Gema Sanz, Kristóf Z. Szalay & Bence Szalai

BMC Genomics **26**, Article number: 393 (2025) | Cite this article

"In this study, we benchmarked two recently published foundation models, scGPT and scFoundation, against baseline models. Surprisingly, we found that **even the simplest baseline model—taking the mean of training examples—outperformed scGPT and scFoundation.**"

The result of wrong incentives: Lots of hyped Pseudoinnovation

Google DeepMind

Millions of new materials discovered with deep learning

Benchmarking foundation cell models for post-perturbation RNA-seq prediction

Gerold Csendes, Gema Sanz, Kristóf Z. Szalay & Bence Szalai

BMC Genomics **26**, Article number: 393 (2025) | Cite this article

Robot chemist sparks row with claim it created new materials

Researchers question whether an AI-controlled lab assistant actually made any novel substances.

"In this study, we benchmarked two recently published foundation models, scGPT and scFoundation, against baseline models. Surprisingly, we found that **even the simplest baseline model—taking the mean of training examples—outperformed scGPT and scFoundation.**"

"I don't think the entire work is garbage," says Schoop. "But the analysis of the products clearly failed. Completely."

The result of wrong incentives: Lots of hyped Pseudoinnovation

Google DeepMind

Millions of new materials discovered with deep learning

Benchmarking foundation cell models for post-perturbation RNA-seq prediction

Gerold Csendes, Gema Sanz, Kristóf Z. Szalay & Bence Szalai

BMC Genomics **26**, Article number: 393 (2025) | Cite this article

Robot chemist sparks row with claim it created new materials

Researchers question whether an AI-controlled lab assistant actually made any novel substances.

"In this study, we benchmarked two recently published foundation models, scGPT and scFoundation, against baseline models. Surprisingly, we found that **even the simplest baseline model—taking the mean of training examples—outperformed scGPT and scFoundation**."

"I don't think the entire work is garbage," says Schoop. "But the analysis of the products clearly failed. Completely."

- Problems to evaluate what really works
- 'The bullshit asymmetry principle'; it takes 10 times more energy to refute bad science than to create it

Problems with relevant validation of ‘Co-Scientist’ approaches (generally, not only Google)

Google Research, 19 Feb 2025

Accelerating scientific breakthroughs with an AI co-scientist

We introduce AI co-scientist, a multi-agent AI system built with Gemini 2.0 as a virtual scientific collaborator to help scientists generate novel hypotheses and research proposals, and to accelerate the clock speed of scientific and biomedical discoveries.

https://www.drugdiscovery.net/2025/02/20/the-google-co-scientist-hasnt-yet-lead-to-breakthroughs-a-closer-look-at-its-scientific-validation/

Validations performed, 1.

- ‘Drug repurposing for acute myeloid leukaemia’
  - Biminetinib found to have 7nM IC50 in AML cell lines
- Problems with validation
  - ‘Drug repurposing’ by definition needs *in vivo* validation (so not the right type of experiment for validation)
  - Biminetinib has well-established, *very* related activities in anti-cancer space (very similar activity on other cell lines, incl. leukemia)
  - So only very slight extrapolation, trivial to the even slightly trained human

NCI human tumor cell line growth inhibition assay. Data for the HL-60(TB) Leukemia cell line

Activity Outcome: Active

BioAssay AID: 125

Substance SID: 440808120 Compound CID: 102

Quantitative High-Throughput drug screen in 47 multiple myeloma cell lines against the NCATS MIPE library collection: KMS21BM_JCRB cell viability assay

Activity Outcome: Active Activity Type: Potency Activity Value: 0.001 μM

BioAssay AID: 1918974

Substance SID: 174006430 Compound CID: 10288191

Validations performed, 2.

- ‘Advancing target discovery for liver fibrosis’
  - Claim of discovery of novel targets relevant to disease

- Problems with validation
  - No details of discovered targets given, so novelty etc cannot be assessed
  - Validation seems to be disconnected from claim, ‘drug effects on fibroblast activity’ is shown in plot presented, no e.g. genetic/biological validation of any targets whatsoever
  - *This is not ‘target discovery’*

Validations performed, 3.

- ‘Explaining mechanisms of antimicrobial resistance’
  - ‘Expert researchers instructed the AI co-scientist to explore a topic that had already been subject to novel discovery in their group, but had not yet been revealed in the public domain, namely, to explain how capsid-forming phage-inducible chromosomal islands (cf-PICIs) exist across multiple bacterial species.’

- Problems with validation
  - Algorithm *re-discovers* what has been established in the group experimentally in parallel
  - ‘Successful in predicting yesterday’s weather’

6. The *really* big picture: Trends in Society

- Transition of authority from ‘experts’ and facts; to ‘influencers’ and public opinion/’belief’
- Hence, focus is not on right and wrong, just *opinions*
- Add ‘money push’ (check LinkedIn these days…) ‘AI can do everything and will change the world’ vs everyday pain of ‘my data isn’t clean, my model doesn’t extrapolate’
- Leads to adoption of immature technology (Klarna, Duolingo, …)
- Pressure on pharma to ‘innovate’; in absence of ability to validate this is often (at best) pseudoinnovation
- Gets exploited by tech-first companies (‘*we know how to do this!’… well, usually, no/not yet!*)

6. The *really* big picture: Trends in Society

The Gentle Singularity

Sam Altman, 10 June 2025

We are past the event horizon; the takeoff has started. Humanity is close to building digital superintelligence, and at least so far it's much less weird than it seems like it should be.

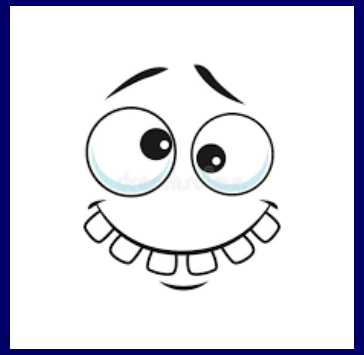
- Transition of authority from 'experts' and facts; to 'influencers' and public opinion/'belief'
- Hence, focus is not on right and wrong, just *opinions*
- Add 'money push' (check LinkedIn these days…) 'AI can do everything and will change the world' vs everyday pain of 'my data isn't clean, my model doesn't extrapolate'
- Leads to adoption of immature technology (Klarna, Duolingo, …)
- Pressure on pharma to 'innovate'; in absence of ability to validate this is often (at best) pseudoinnovation
- Gets exploited by tech-first companies ('*we know how to do this!'… well, usually, no/not yet!*)

6. The *really* big picture: Trends in Society

The Gentle Singularity

Sam Altman, 10 June 2025

We are past the event horizon; the takeoff has started. Humanity is close to building digital superintelligence, and at least so far it's much less weird than it seems like it should be.



- Transition of authority from 'experts' and facts; to 'influencers' and public opinion/'belief'
- Hence, focus is not on right and wrong, just *opinions*
- Add 'money push' (check LinkedIn these days…) 'AI can do everything and will change the world' vs everyday pain of 'my data isn't clean, my model doesn't extrapolate'
- Leads to adoption of immature technology (Klarna, Duolingo, …)
- Pressure on pharma to 'innovate'; in absence of ability to validate this is often (at best) pseudoinnovation
- Gets exploited by tech-first companies ('*we know how to do this!'… well, usually, no/not yet!*)

Is there also a ‘heaven’ inside those circles? Yes:

- Where ***use case, data, methods, and tool/process are aligned***, e.g.

1. Ligand discovery

   Labels are largely unconditional, and lots of data available – *but, in vivo* relevance not necessarily a given

2. Compound (de-)selection for very high clearance

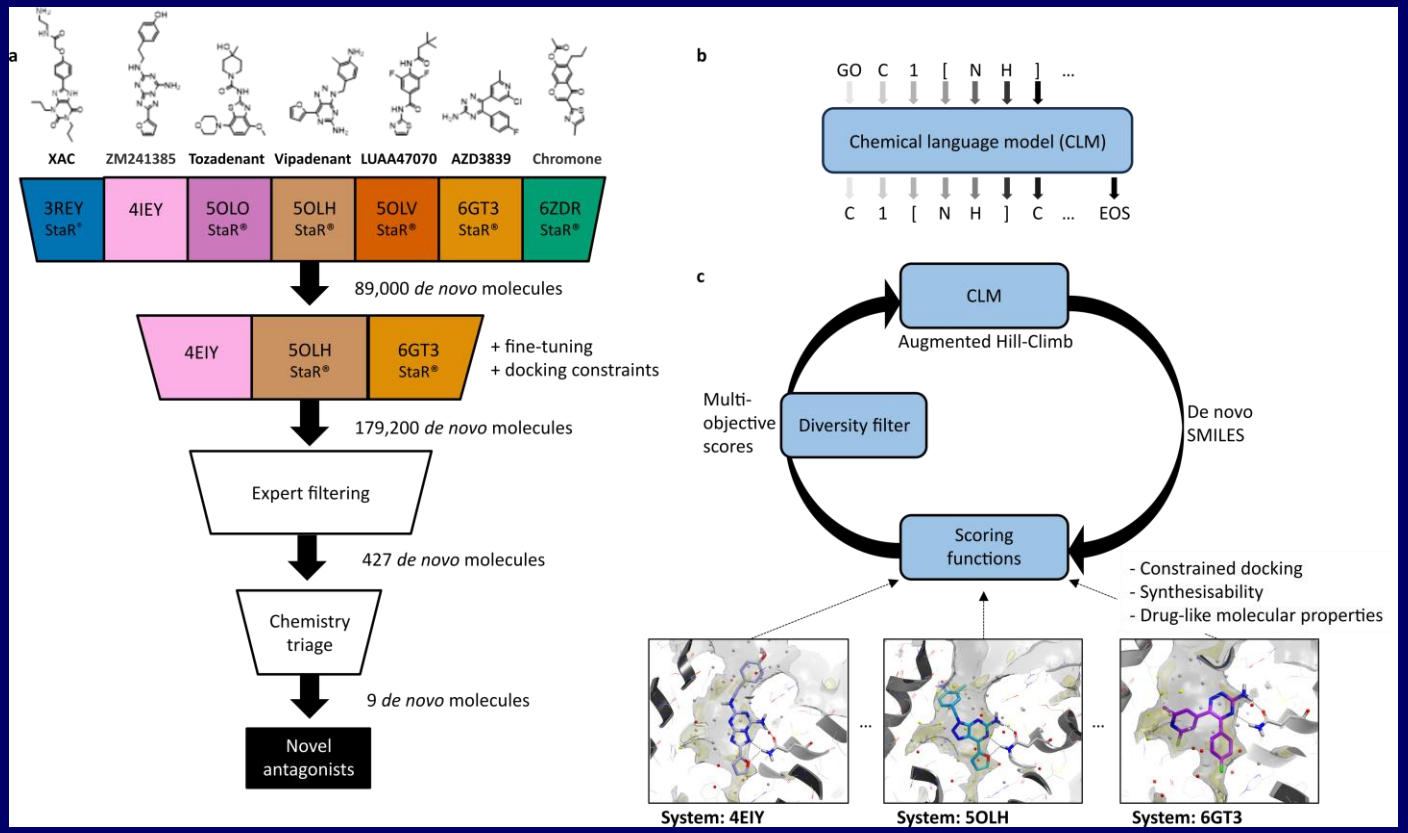
   Labels used for model generation are *in vivo* relevant – ‘doing the right thing’, even if model performance is not numerically perfect

3. Selecting compounds to influence cell fate

   Impact of compound treatment on gene expression often sufficiently retained between systems, hence extrapolation ‘sufficiently possible’

1. Example study: Identification of novel nanomolar adenosine A2A receptor ligands using reinforcement learning

- Work of Morgan Thomas with SoseiHeptares; using chemical language models for GPCR ligand design, against A2A, involving synthesis
- 5 out of 9 novel scaffolds for receptor identified, including nanomolar actives with functional activity
- Co-crystals partially confirm computationally established binding mode



Thomas, M., Matricon, P.G., Gillespie, R.J. *et al.* Identification of nanomolar adenosine A_{2A} receptor ligands using reinforcement learning and structure-based drug design. *Nat Commun* **16**, 5485 (2025).

2. Example of *in vivo* relevant data modelled directly: PK models based on chemical structure (human, rat)

molecular pharmaceutics
pubs.acs.org/molecularpharmaceutics
Article

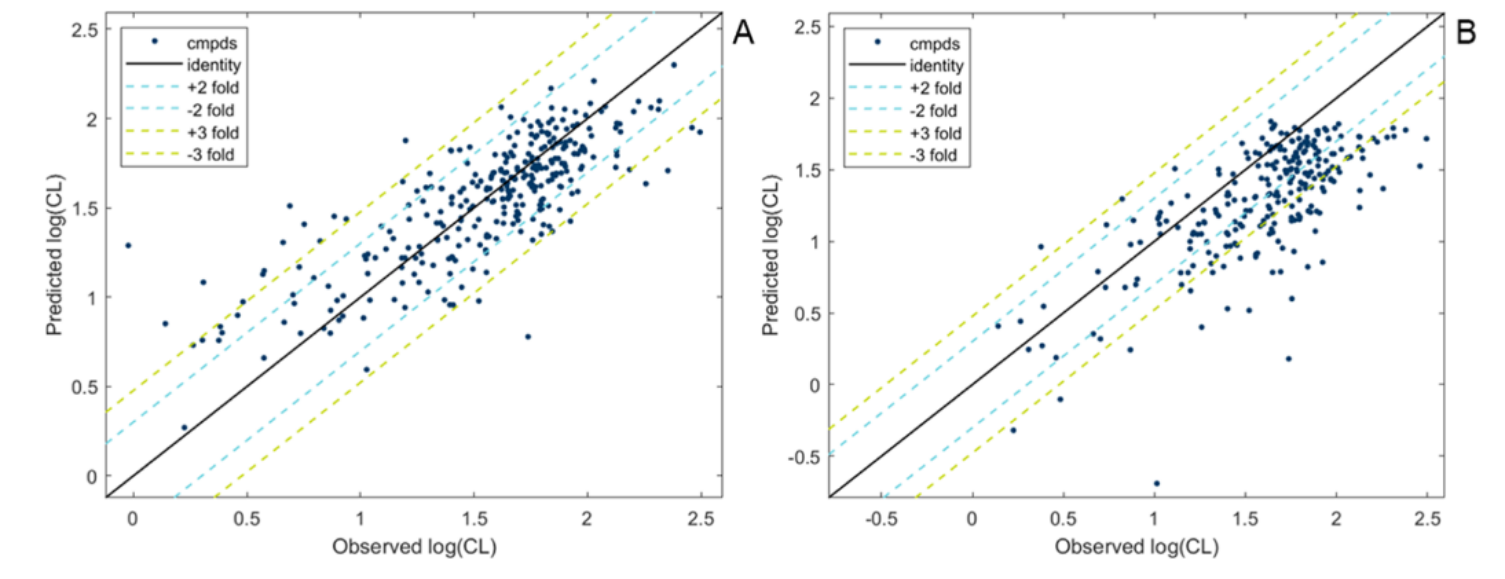
Prediction of In Vivo Pharmacokinetic Parameters and Time–Exposure Curves in Rats Using Machine Learning from the Chemical Structure

Olga Obrezanova,* Anton Martinsson, Tom Whitehead, Samar Mahmoud, Andreas Bender, Filip Miljković, Piotr Grabowski, Ben Irwin, Ioana Oprisiu, Gareth Conduit, Matthew Segall, Graham F. Smith, Beth Williamson, Susanne Winiwarter, and Nigel Greene

molecular pharmaceutics
pubs.acs.org/molecularpharmaceutics
Article

Machine Learning Models for Human *In Vivo* Pharmacokinetic Parameters with In-House Validation

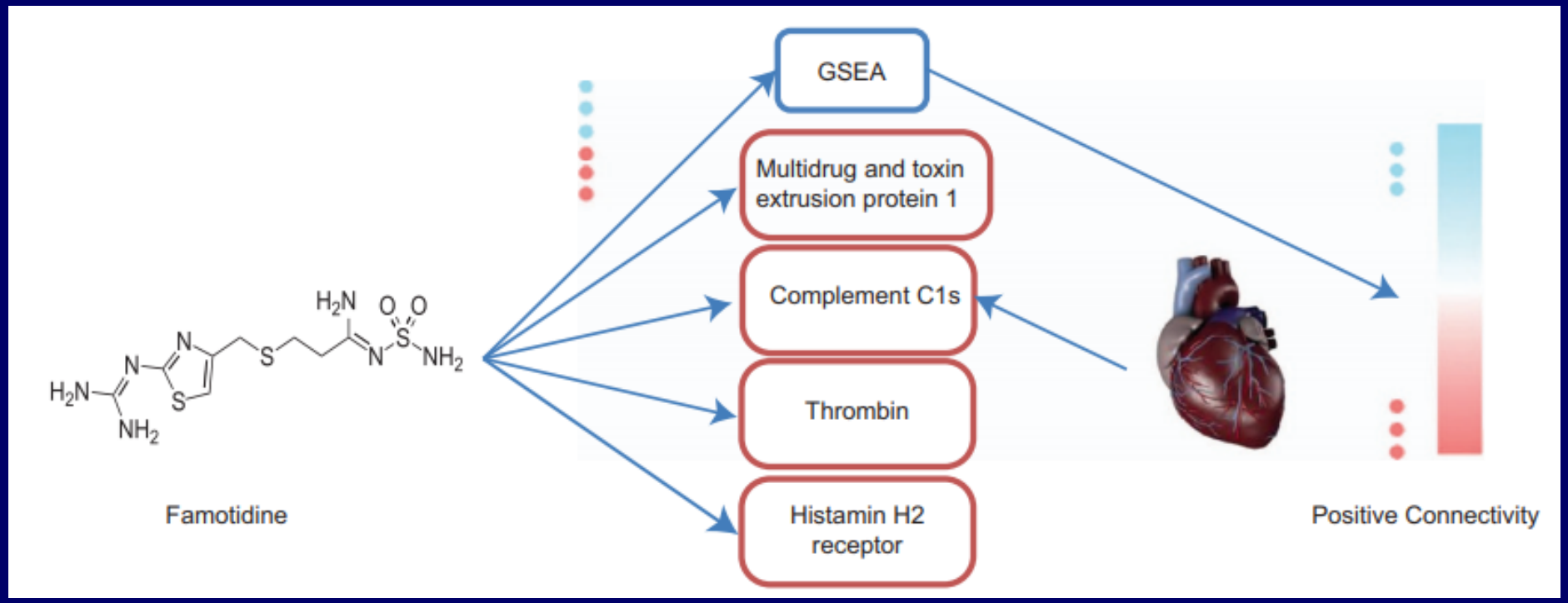
Filip Miljković,*,▽ Anton Martinsson,▽ Olga Obrezanova, Beth Williamson, Martin Johnson, Andy Sykes, Andreas Bender, and Nigel Greene



- ‘Tell me what’s bad and remove’ (e.g. compounds which are very metabolically labile; CL > liver blood flow); ***deselection of ~20% of worst compounds, in an idea-rich environment*** (as opposed to *selection* setting in an *idea-poor* environment): **Always understand how you use a model**
- Can be used fast at point of design, e.g. im DMTA cycles

3. Cellular Reprogramming: From ‚inhibiting/modulating one target‘ to *pushing the cell/system into a different state*

- For regenerative medicine (differentiating stem cells in different organs), cancer (‘converting’ cancer cells into other cell types, e.g. for recognition by immune system)
- E.g. Y KalantarMotamedi et al. Cell Death Discovery (2016) 2, 16007

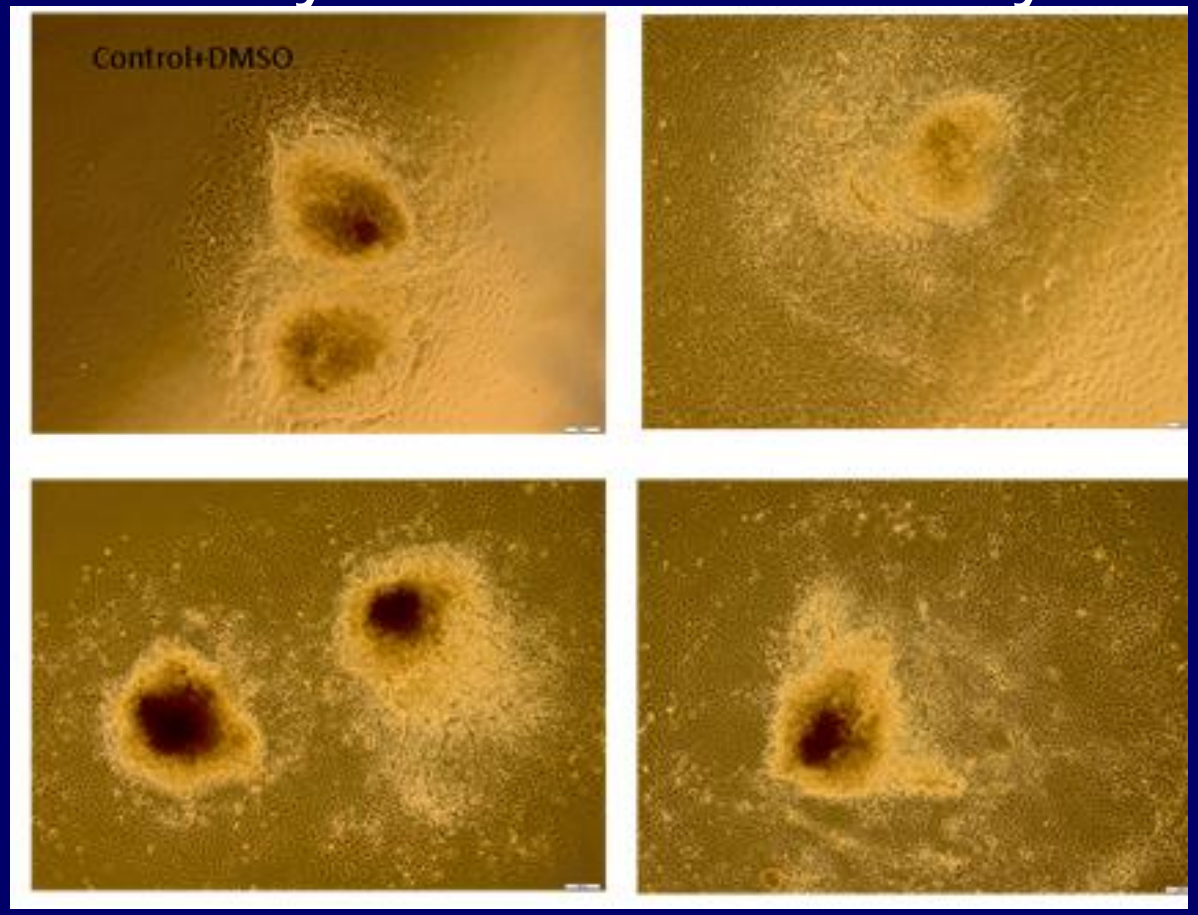


Selected compound induces differentiation of stem cells into cardiac myocytes (validated by RT-PCR and on proteomic level; work with Dr Nasr, Royan Institute, Isfahan)

3 days 5 days

Control

Compound



Wide use case – regenerative medicine (pancreatic beta cells, macular degeneration, …)

Even more widely *influencing cell fate*

KalantarMotamedi *et al. Cell Death Discovery* **2016**

EIC Pathfinder Project

RESYNC

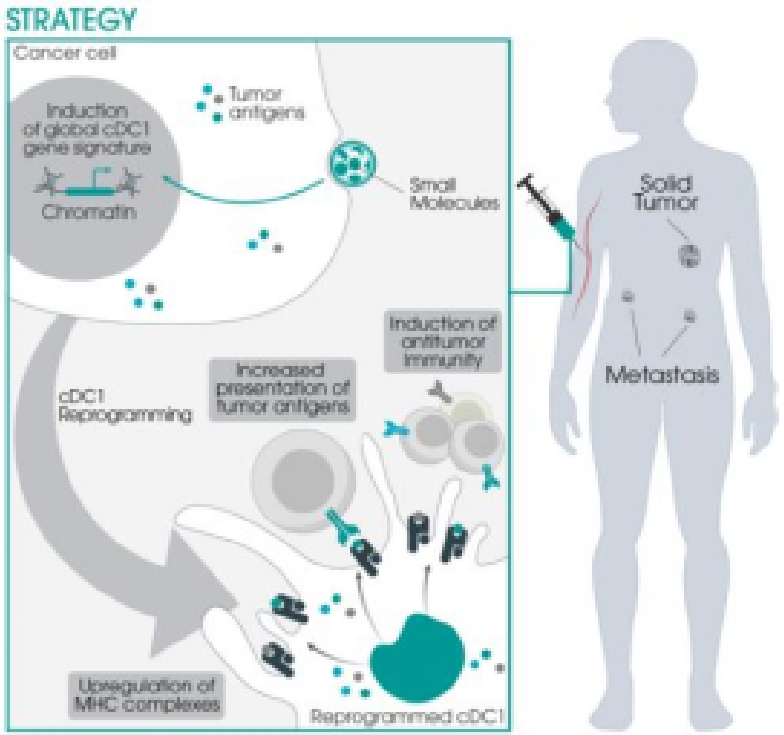
Functional chemical reprogramming of cancer cells to induce antitumor immunity.

Participant No	Participant organisation name	Short Name	Country
1 (Coordinator)	Lund University	LU	SE
2	Asgard Therapeutics	ASG	SE
3	Politecnico di Torino	POL	IT
4	Universitatea Babeș-Bolyai Cluj	UBB	RO
5	IOCB Prague	IOCB	CZ
6	Karolinska Institutet	KI	SE

1. EXCELLENCE

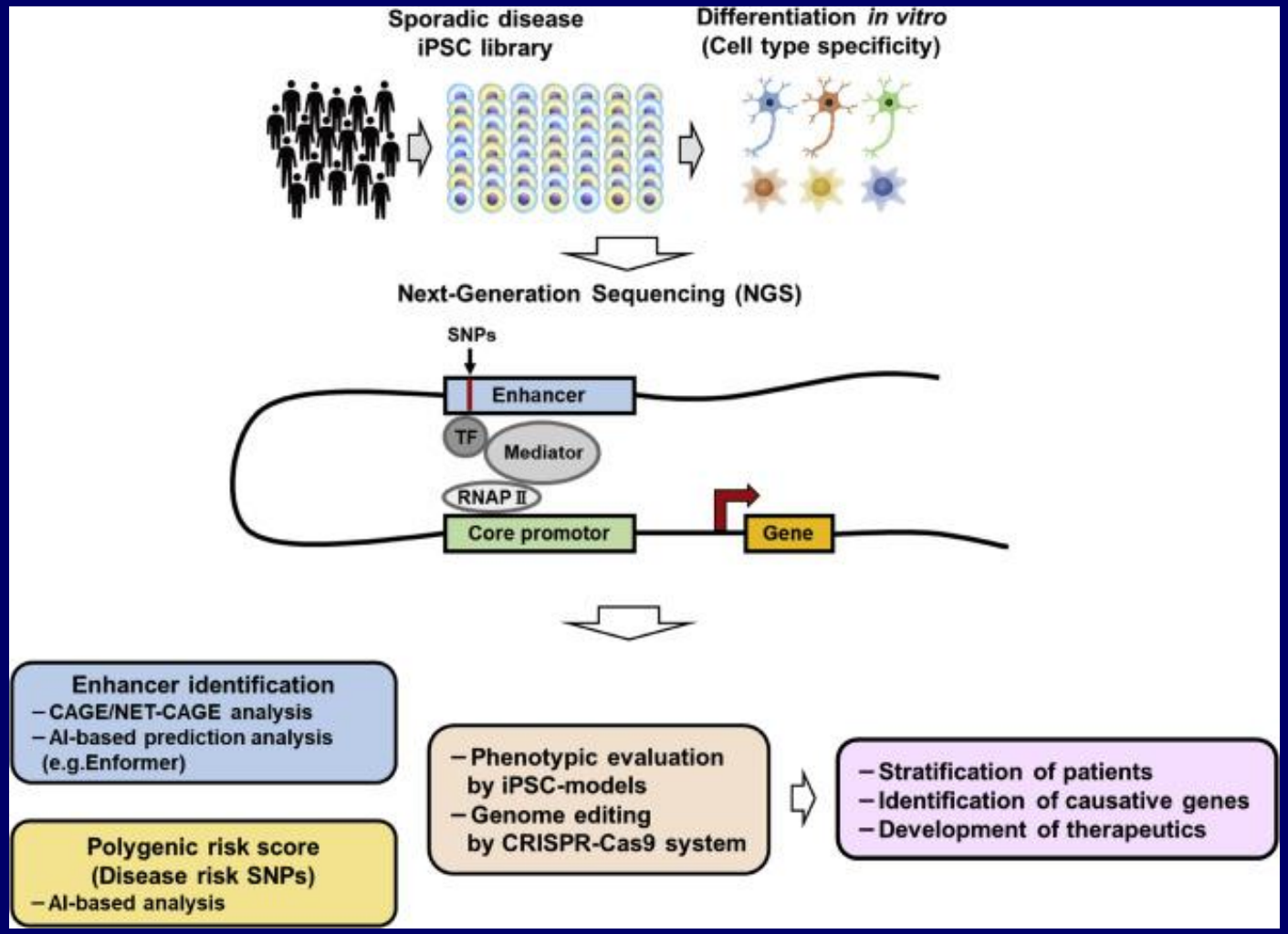
1.1 Long-term vision

The radical vision of the RESYNC consortium is to revolutionize cancer immunotherapy through small-molecule (SM)-based reprogramming of cancer cells into immunogenic cancer antigen-presenting type 1 conventional dendritic cells (cDC1) to elicit a personalized anti-tumor immunity. Ultimately, this technology has the potential to overcome the barriers of available immunotherapies, resulting in an off-the-shelf, systemic SM cocktail to treat patients more effectively, safer and at lower costs.



General possible path currently/in the future: Biological models with sufficient complexity, but still feasible, e.g. iPSC systems

- Sweet spot: Representative/predictive, accessible, testable/scalable
- E.g. CNS: Primary samples inaccessible, simplistic models non-predictive (also animal models often not sufficient)
- ALS example: Ropinirole and bosutinib identified via iPSC models, currently clinical candidates
- *Predictive; yet feasible*

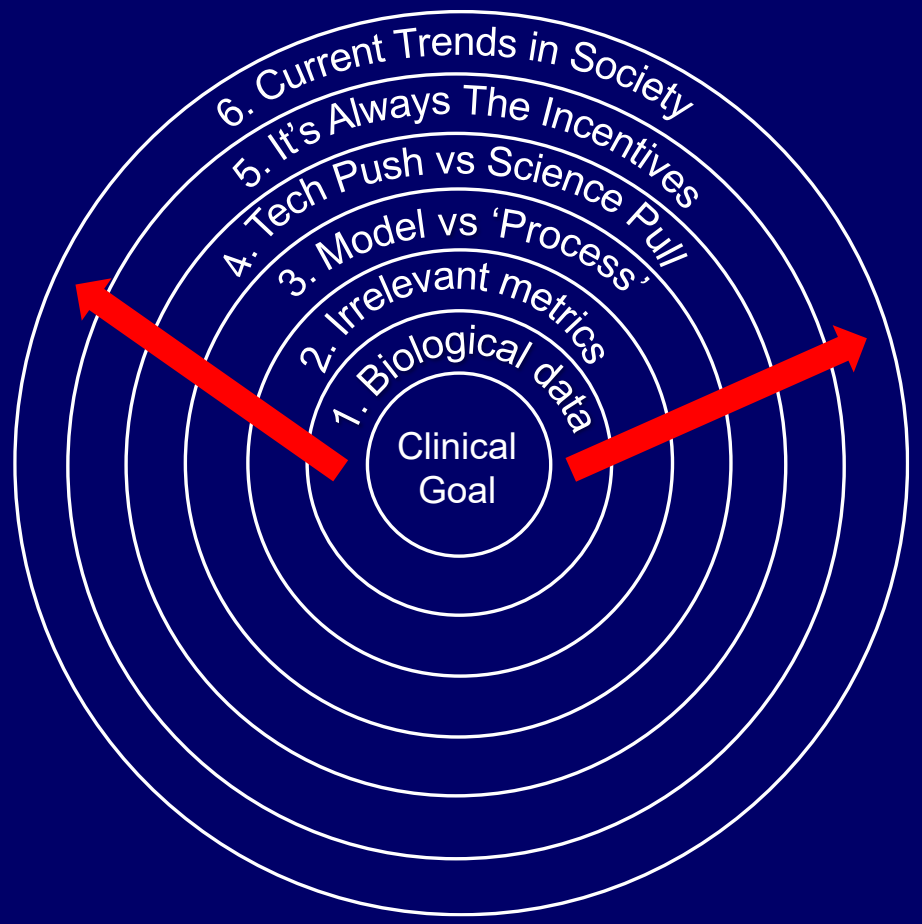


Okano *et al.*, *Cell Stem Cell*, 2022

Summary: Where does AI in drug discovery stand?

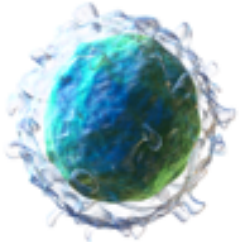
- At least ‘six circles of hell’ need to be overcome to get to greener pastures in AI in drug discovery

- We need to align aims, data, methods, and validation better to go beyond optimization of irrelevant metrics in proxy spaces, and towards real-world impact



Contact: andreas.bender@ku.ac.ae, andreas@bio.bi

UAE Drug Discovery and Biotech Network WWW.GENETIC.AE



First meeting: 4 June 2025, 1-4.30pm GST, on Zoom

Free and Open to All – Join and Circulate!

The program for the first meeting of the UAE Drug Discovery and Biotech Network has just been announced, to give an overview of the local and regional research landscape – the event is open to all and will be held via Zoom.

The aim is to create an informal network of translationally interested scientists to work together, across all organizations in the UAE and beyond and open to all, to advance drug discovery and biotech, from research to its applications.

Please join in and circulate to your friends and colleagues!

Khalifa University Experimental Facilities – open to academic collaboration and commercial services

- State-of-the-art Cryo-EM Structural Biology Facilities for drug design
- Krios G4 (300kV) etc.

- Animal house (largest in the Middle East; 30,000 rodents) and microbiology facilities
- … Let me know if you wish to work together (flexible arrangements possible)

Cryogenic Electron Microscopy (Cryo-EM) Facility Equipment

- Cryo-EM TEM equipment under installation on Main Campus (B Building B0-0061)

Talos L120C: Cryo TEM 120 kV
- Designed for cryo and room-temperature imaging of biological and nanoscale materials.
- Nominal TEM resolution ~0.38 nm
- Easy operation and fast alignment—ideal for training and screening
- Gentle imaging of beam-sensitive samples

Glacios 2: Cryo TEM 200 kV
- Advanced 200 kV TEM designed for high-throughput cryo-EM applications, including single particle analysis (SPA) and tomography.
- Resolution for SPA ~3 Å.
- Higher throughput than 120 kV systems, ideal for intermediate to high-resolution studies.

Krios G4 – Cryo TEM 300 kV
- State-of-the-art 300 kV TEM for ultra-high-resolution cryo-electron microscopy.
- Gold standard platform for structural biology and high-throughput single particle analysis (SPA).
- Better than 2.0 Å routinely for SPA
- Sub-nanometer resolution for CryoET

Animal Research Facility (Vivarium)

Scope: Devoted to human modelling and drug discovery science and supports life science, bacteriology, genetics, medicine, and bio/chemistry pillars to provide applied R&D and testing solutions.

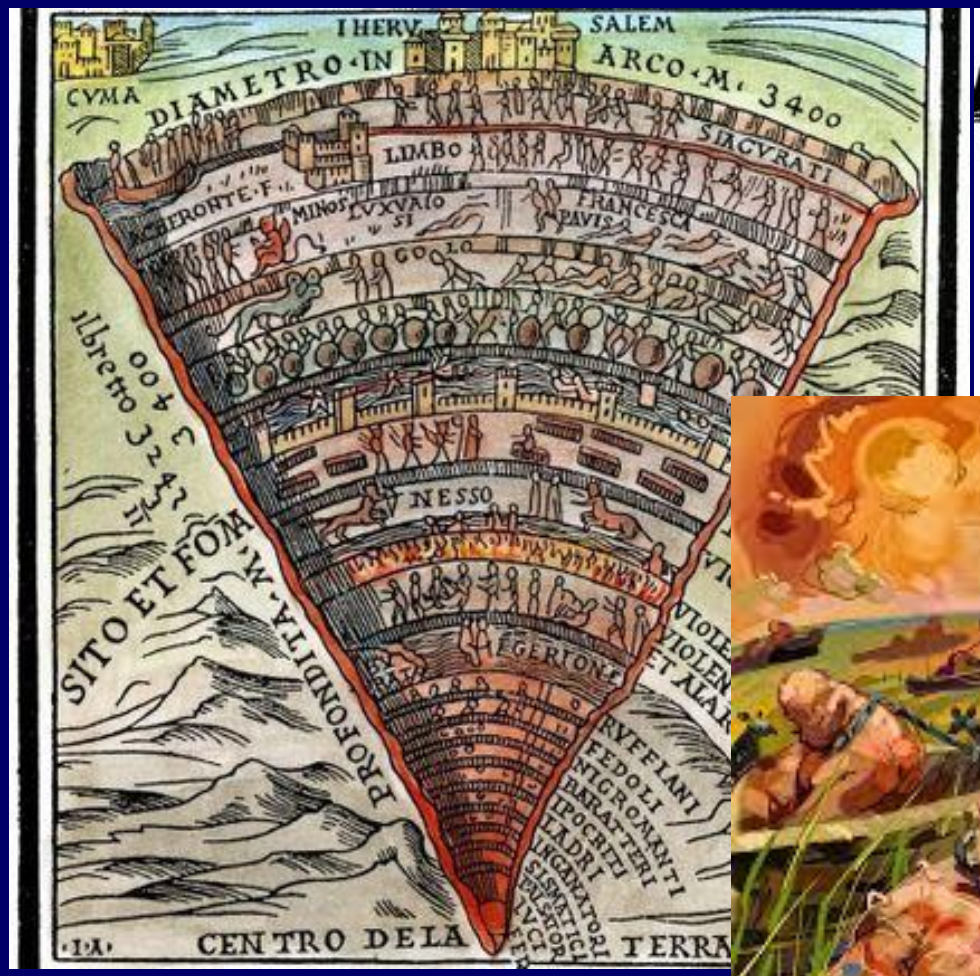
Key Capabilities:
- Specific pathogen free (SPF) rodent housing
- Behavioural studies
- Human model studies
- Preclinical imaging

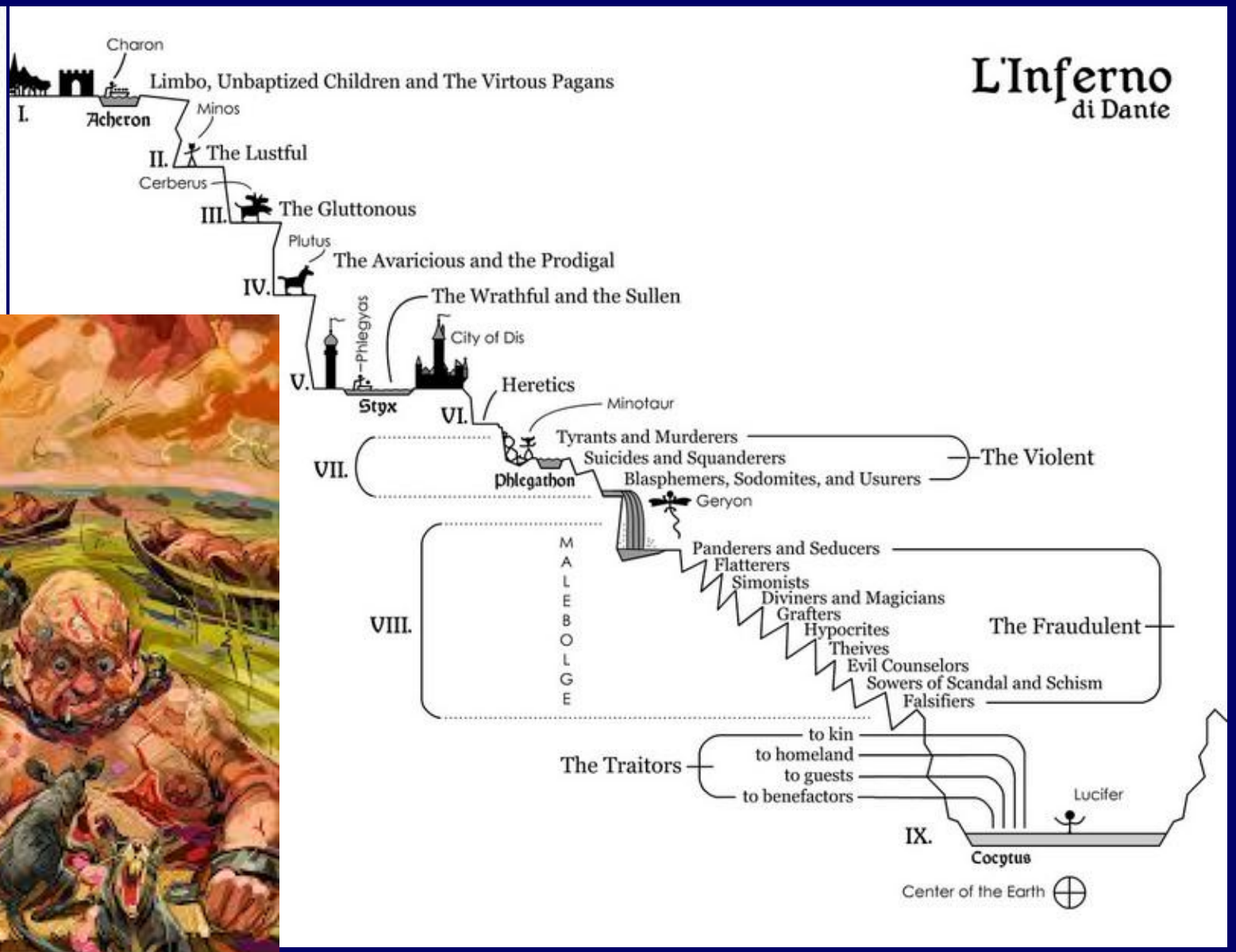
Clinical Microbiology and Immunology Laboratory

Scope: (i) Devoted to the culturing, examination, and identification of microorganisms (bacteria, fungi, yeasts)
(ii) Supports life science, bacteriology, molecular genetics, medicine, and chemistry pillars to provide applied R&D and testing solutions.
(iii) Flow cytometry techniques for cell analyzing and sorting.

Key Capabilities:
- Equipped with state-of-the-art technologies, including flow cytometry and cell sorter, real time PCR, cell and tissue culture, biochemical analysis of gene expression.
- Functions at an enhanced Biological Safety Level (BSL) 2 Laboratory while meeting Good Laboratory Practice (GLP) regulations and includes Class II, Type A2 biological safety cabinets.
- Identification of pathogens, relative quantification of identified pathogens and profile of antibiotic sensitivity testing.

With Special Thanks to Dante's 'Nice Circles of Hell'





From Kozachok's Inferno: 3rd Circle of Hell: GLUTTONY