Will LLMs and Agents Eat Drug Discovery?

Aspects of Goals, Data, and Decision Spaces

Andreas Bender

andreas@bio.bi

Full slide deck at www.drugdiscovery.net

Outline

1. What does it mean to ‘discover’ a drug?
2. History repeats itself
3. ‘Agentic’ science
4. On data and decision spaces – some quantification
5. Classification of decision types for automated discovery
6. Beyond (only) data and towards better decisions

1. What is this thing actually, ‘Drug Discovery’?

- Drug discovery implies: ‘We don’t have enough! We shall set our sails and we need more, we need to… *discover!’*

- But … hypotheses are (always were!) cheap and plentiful

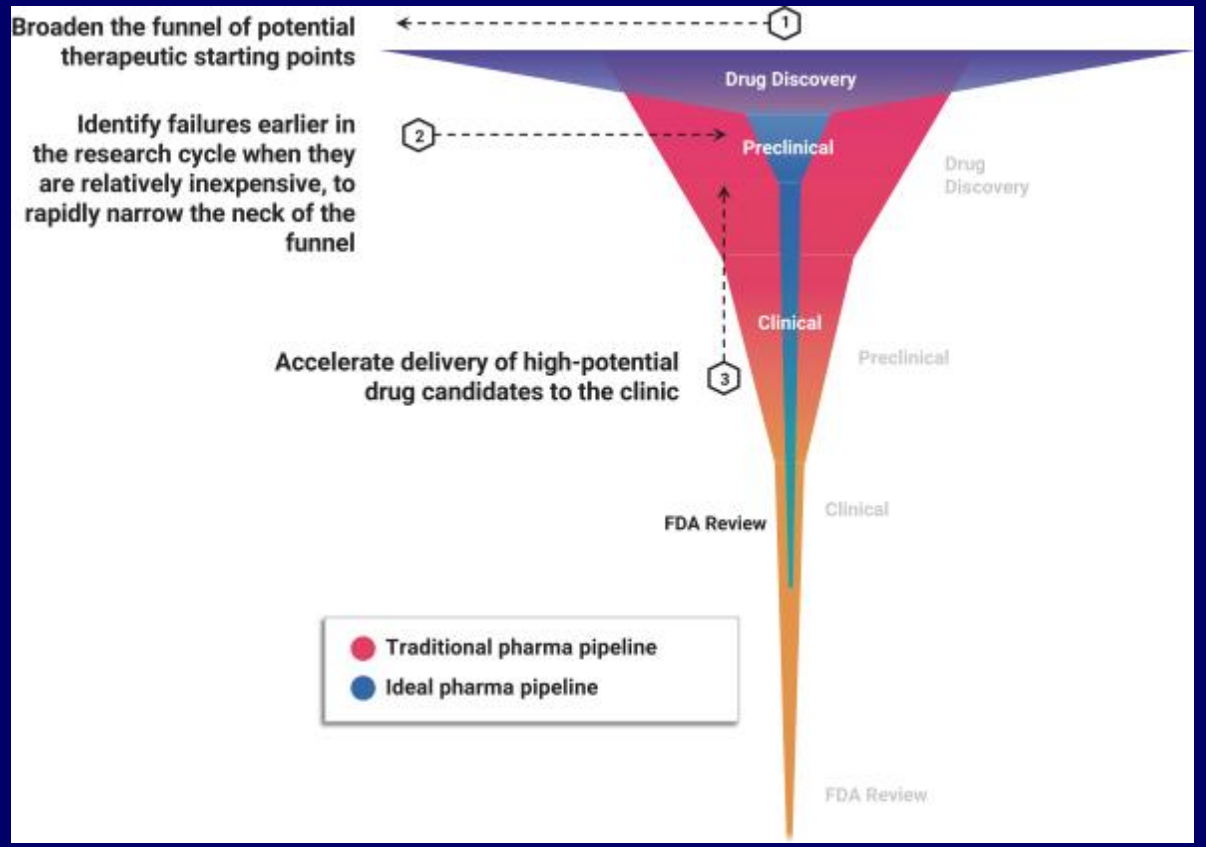
- Drug discovery is about *removing what doesn’t work, to retain what works*

1. What is this thing actually, ‘Drug Discovery’?

- Drug discovery implies: ‘We don’t have enough! We shall set our sails and we need more, we need to… *discover!’*

- But … hypotheses are (always were!) cheap and plentiful

- Drug discovery is about *removing what doesn’t work, to retain what works*



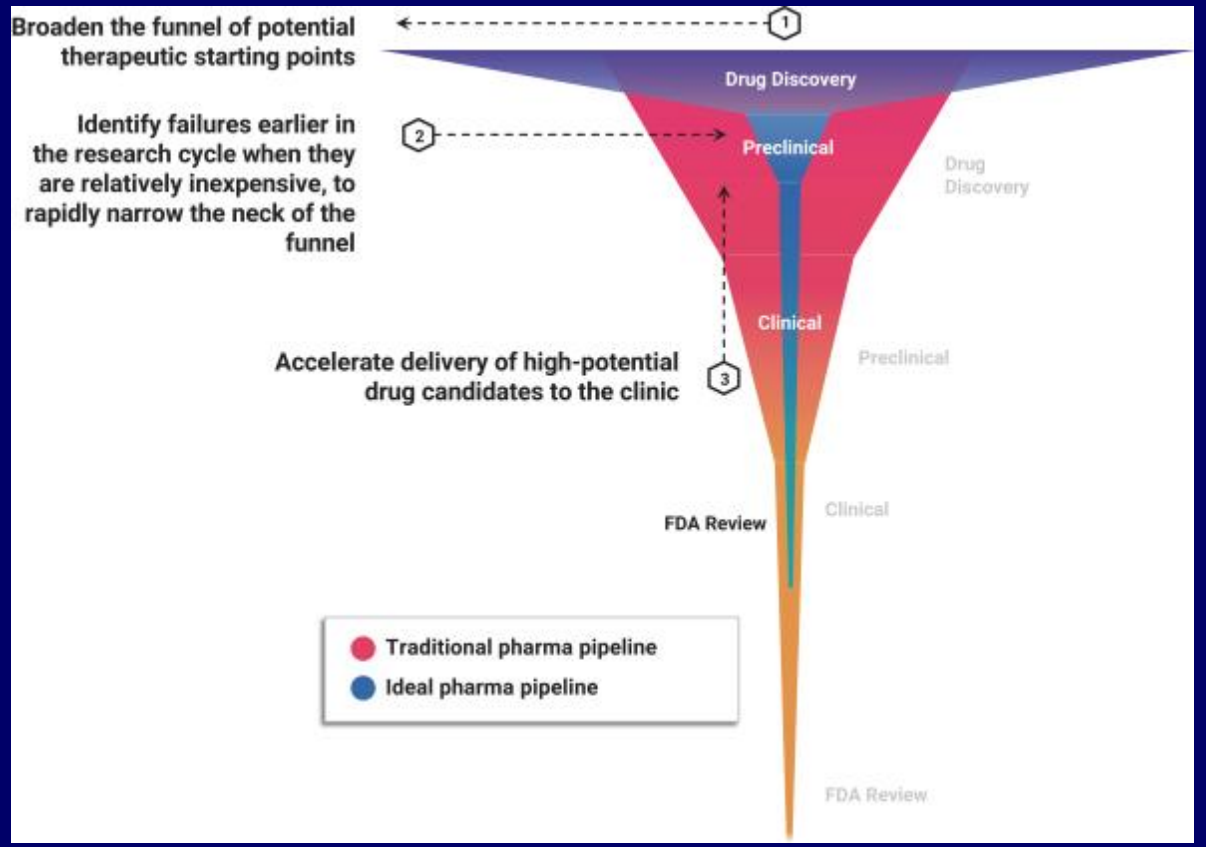
1. What is this thing actually, ‘Drug Discovery’?

- Drug discovery implies: ‘We don’t have enough! We shall set our sails and we need more, we need to… *discover!’*

- But … hypotheses are (always were!) cheap and plentiful

- Drug discovery is about *removing what doesn’t work, to retain what works*

The point is *not* to stick in *more* up there…



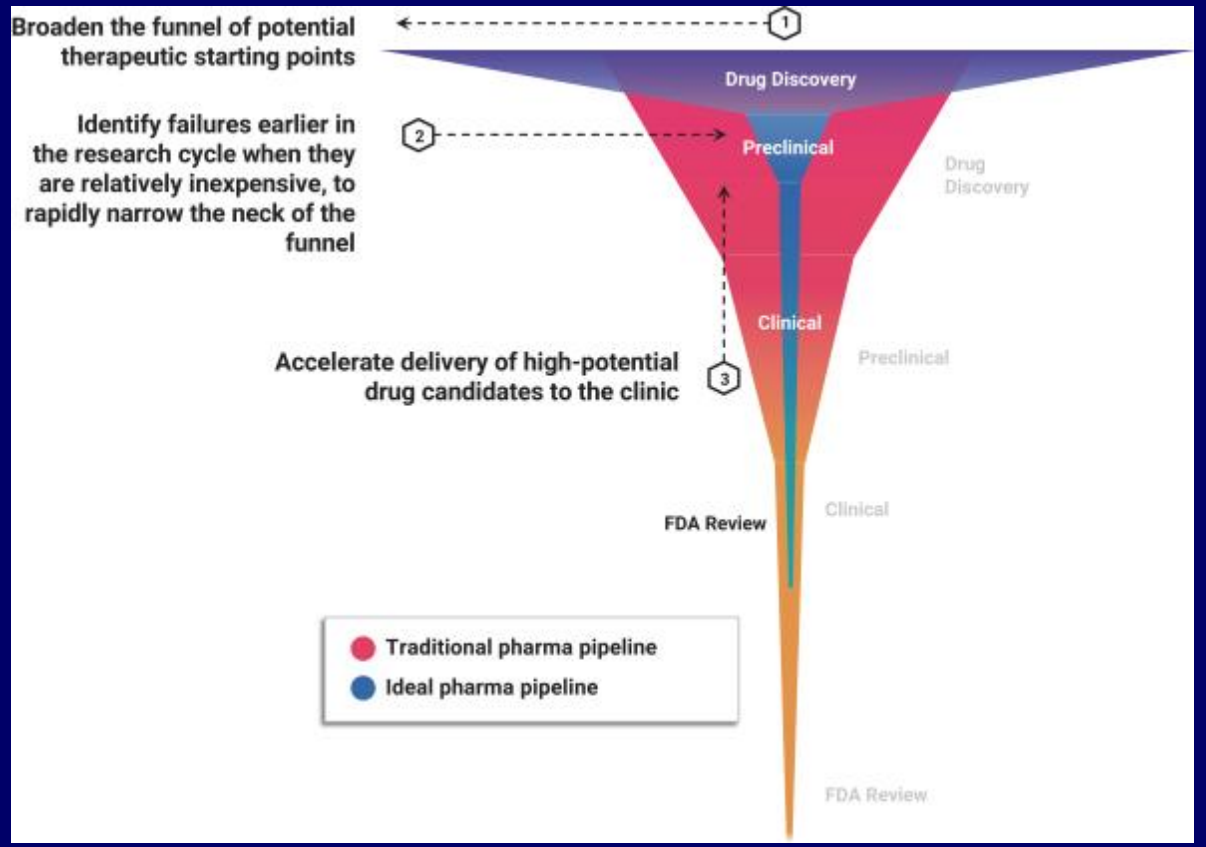
1. What is this thing actually, ‘Drug Discovery’?

- Drug discovery implies: ‘We don’t have enough! We shall set our sails and we need more, we need to… *discover!’*

- But … hypotheses are (always were!) cheap and plentiful

- Drug discovery is about *removing what doesn’t work, to retain what works*

The point is *not* to stick in *more* up there…



The point is *to have more come out down here*

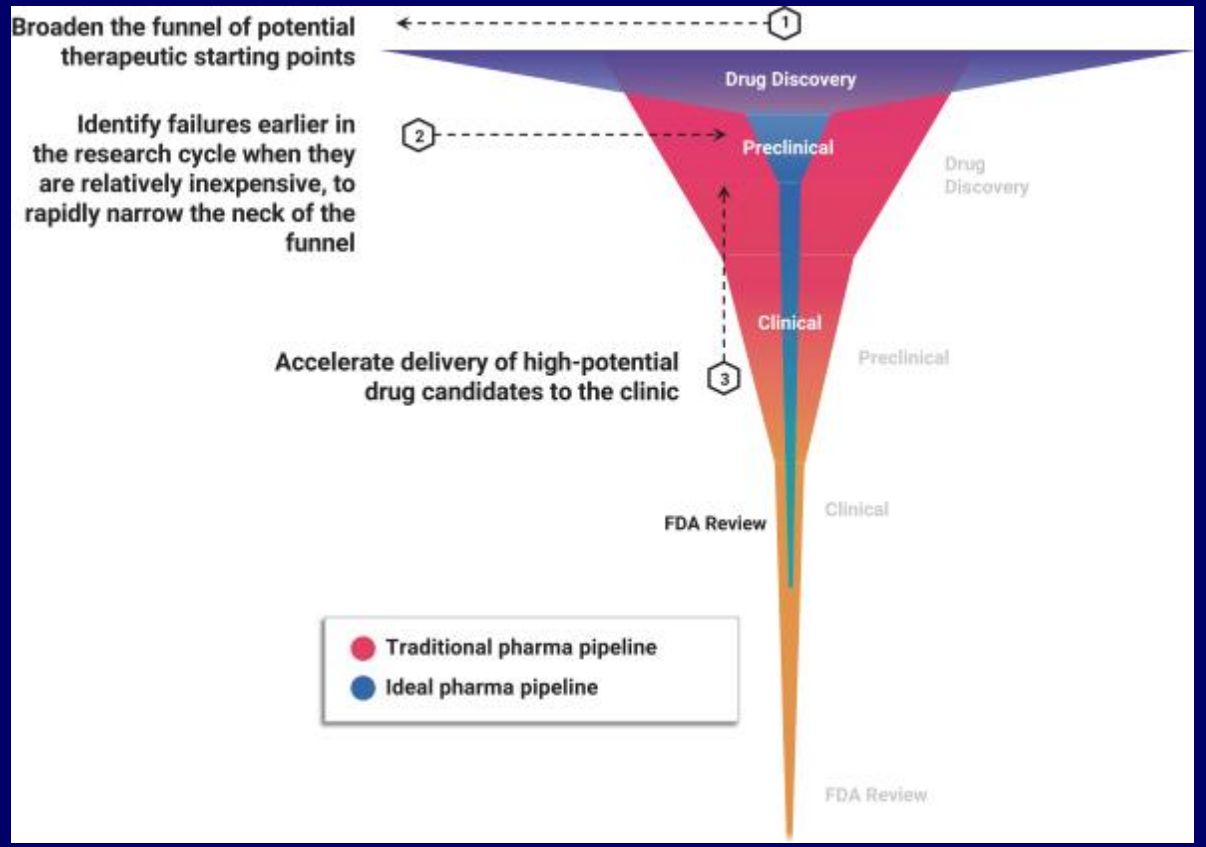
1. What is this thing actually, ‘Drug Discovery’?

- Drug discovery implies: ‘We don’t have enough! We shall set our sails and we need more, we need to… *discover!’*

- But … hypotheses are (always were!) cheap and plentiful

- Drug discovery is about *removing what doesn’t work, to retain what works*

The point is *not* to stick in *more* up there…

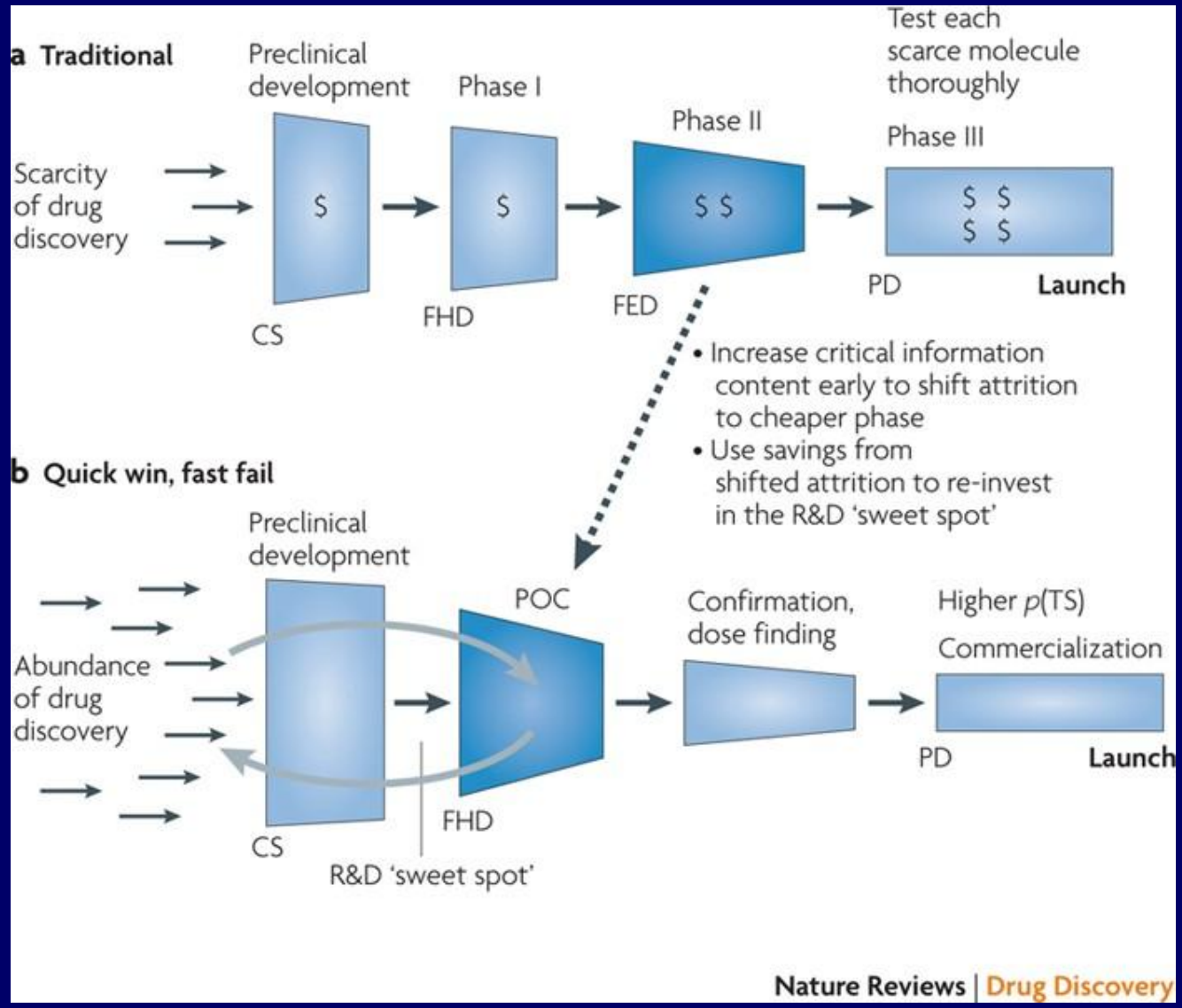


… by having the right hypotheses and selection criteria in place

‘More’ is of limited use as we will see

The point is *to have more come out down here*

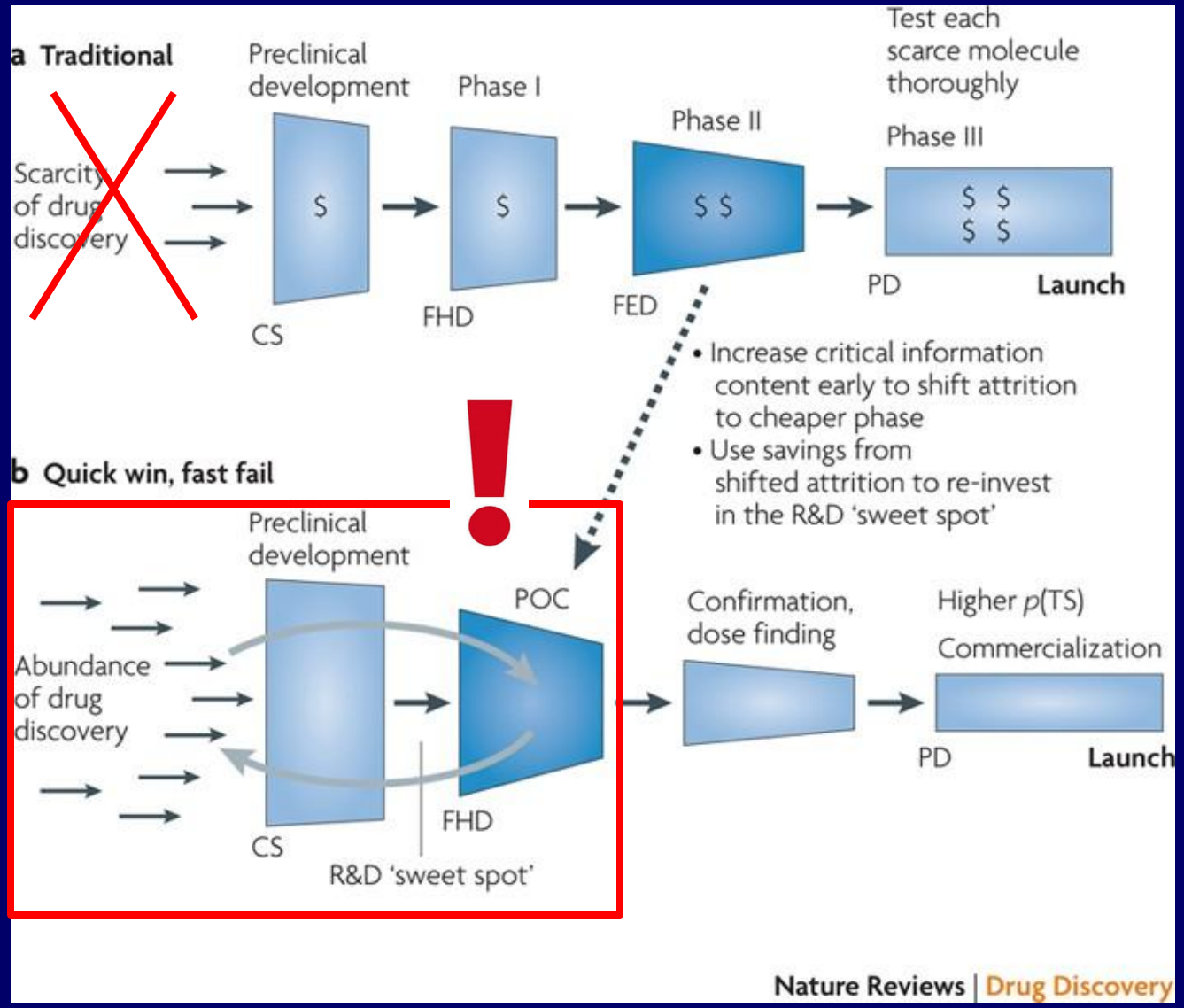
Drug discovery is… sifting through plenty of options, in data-poor spaces



Paul et al., How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery 2010, 9, pages 203–214 (2010)

Drug discovery is… sifting through plenty of options, in data-poor spaces

- We live in an *abundance of drug discovery* world!
- We don't need to '*discover*', we need to *de-risk*!

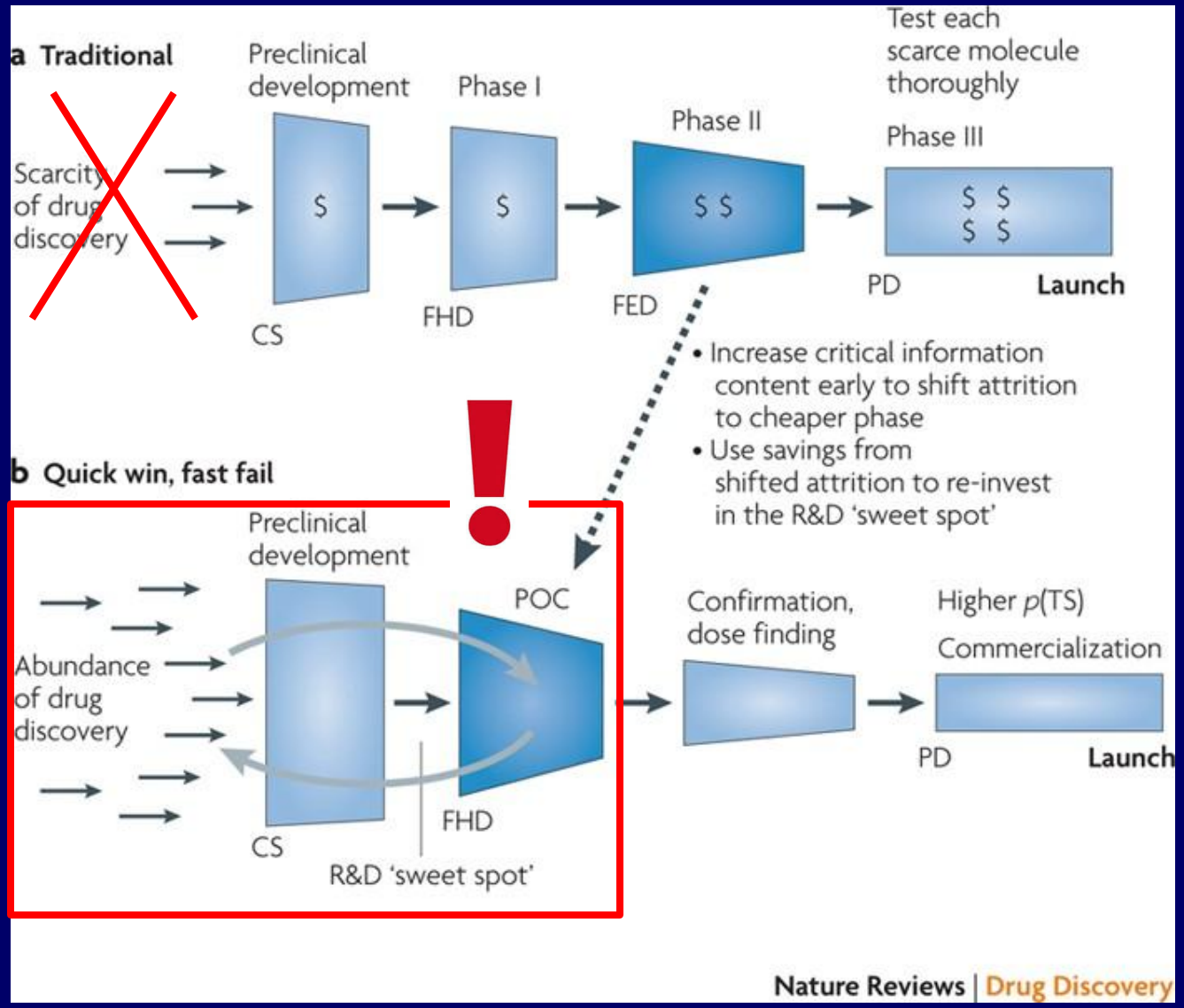


Paul et al., How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery 2010, 9, pages 203–214 (2010)

Drug discovery is… sifting through plenty of options, in data-poor spaces

- We live in an *abundance of drug discovery* world!
- We don't need to '*discover*', we need to *de-risk*!

- As some say: 'Fail fast, fail cheap'

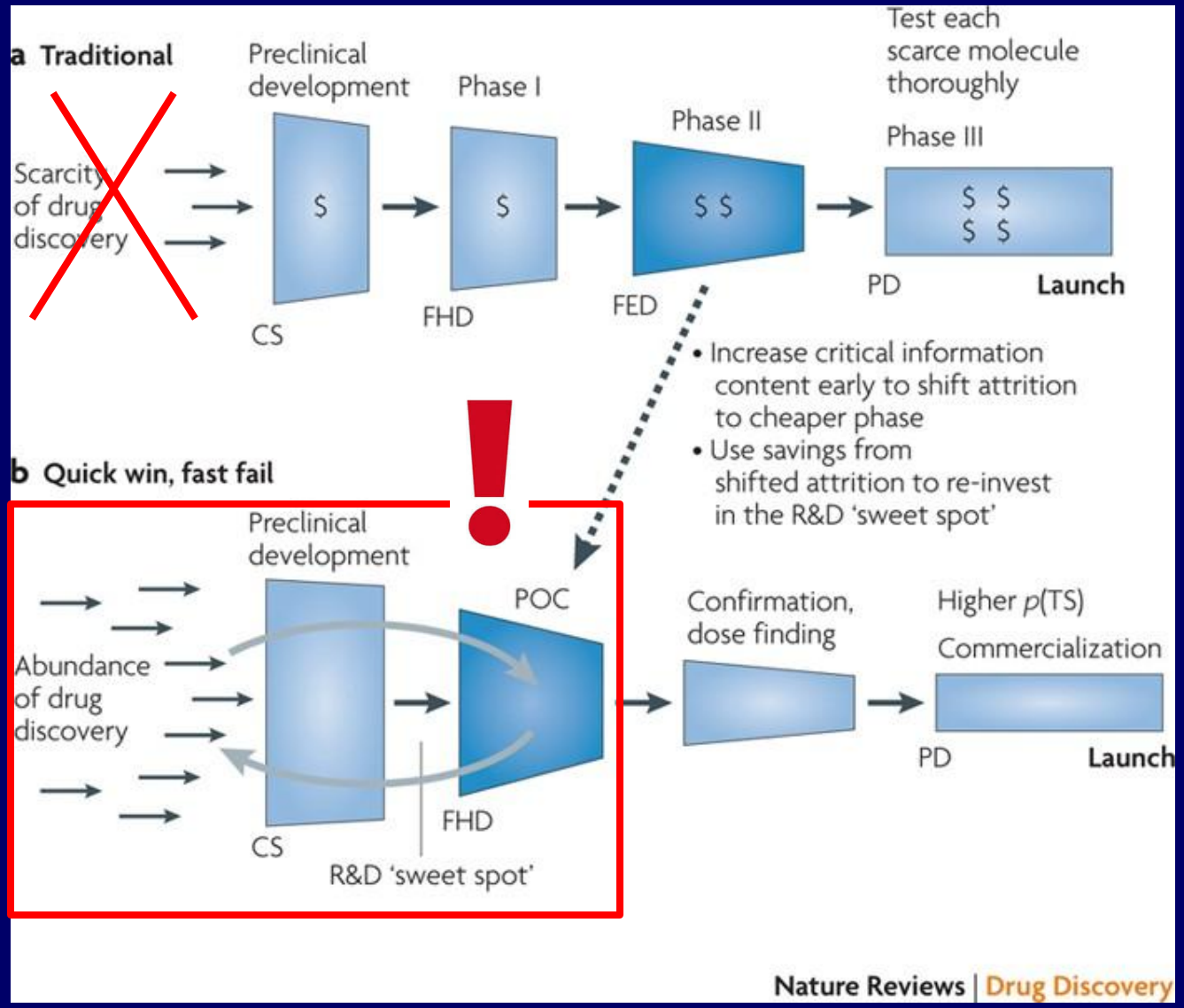


Paul et al., How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery 2010, 9, pages 203–214 (2010)

Drug discovery is… sifting through plenty of options, in data-poor spaces

- We live in an *abundance of drug discovery* world!
- We don't need to '*discover*', we need to *de-risk*!

- As some say: 'Fail fast, fail cheap'
- Personal preference: 'Kill fast, kill early, and kill completely'
(Novartis postdoc seminar, MA, 2008)



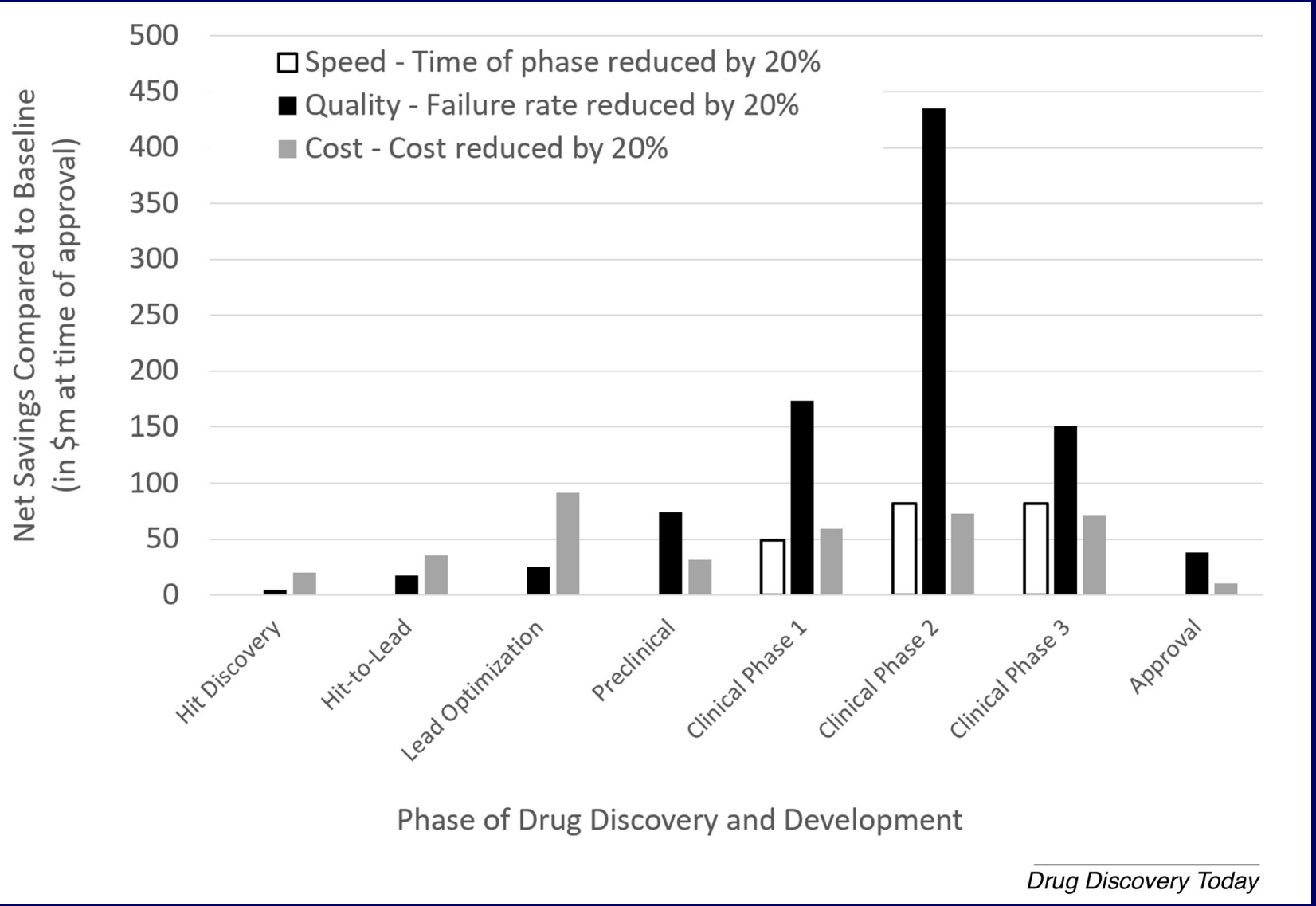
Paul et al., How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery 2010, 9, pages 203–214 (2010)

THE GOAL: Clinically relevant decisions, related to (mostly) efficacy and (also) safety

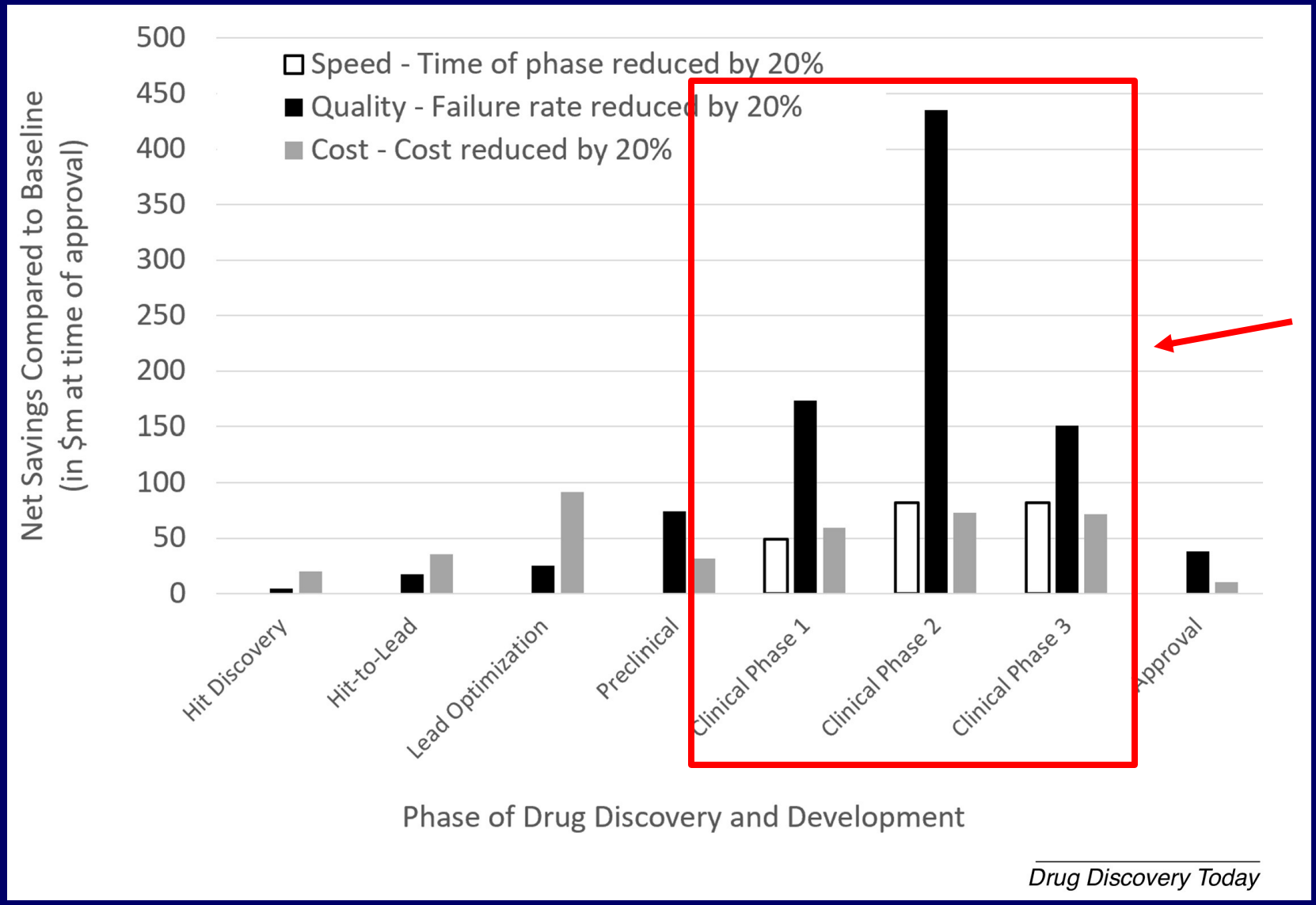
Bender and Cortes, Drug Discovery Today 2021

Bender et al. NRDD 2026

Morgan et al. NRDD 2018



THE GOAL: Clinically relevant decisions, related to (mostly) efficacy and (also) safety



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

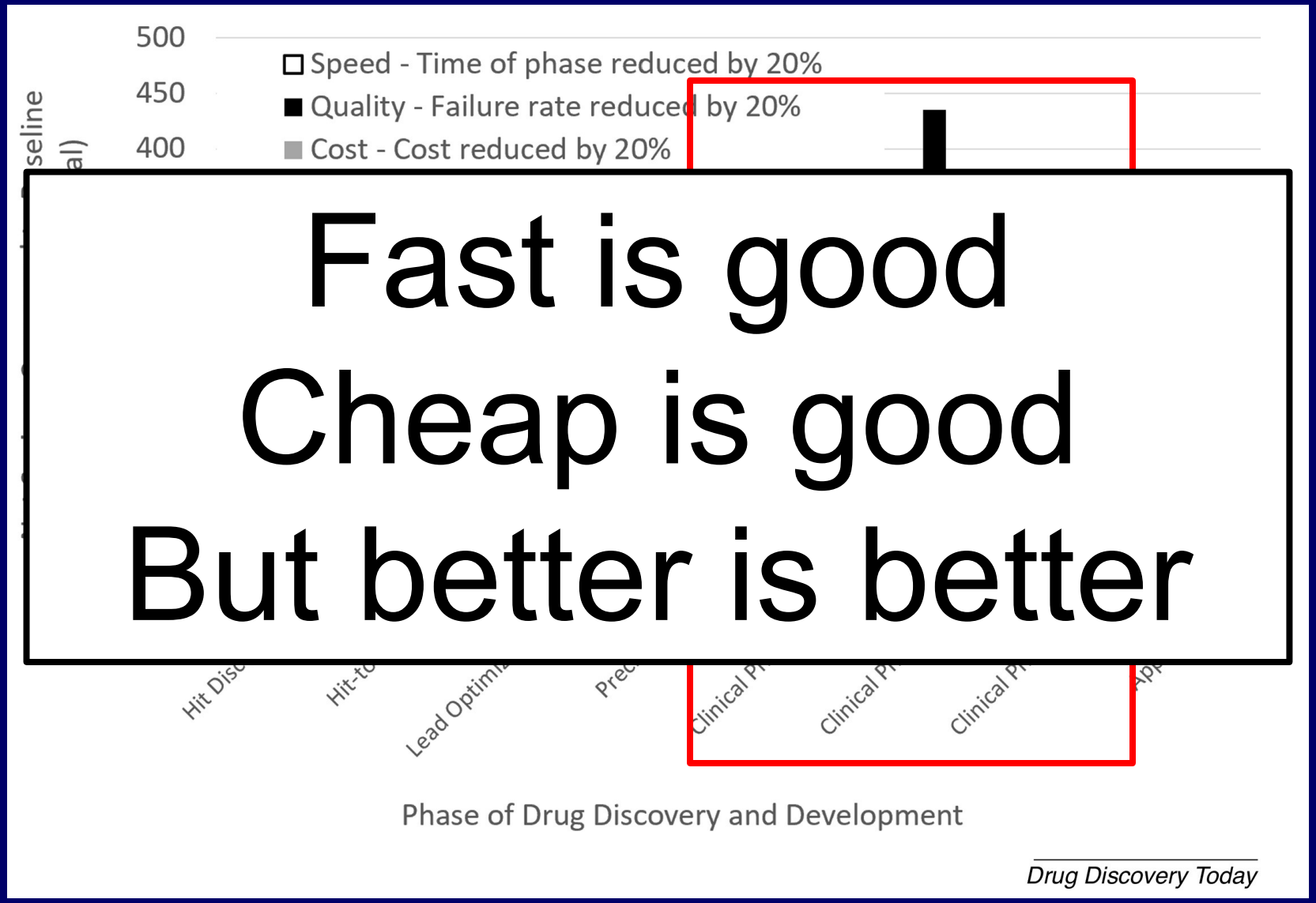
Everything else matters less. *Far* less.

Bender and Cortes, Drug Discovery Today 2021

Bender et al. NRDD 2026

Morgan et al. NRDD 2018

THE GOAL: Clinically relevant decisions, related to (mostly) efficacy and (also) safety



Bender and Cortes, Drug Discovery Today 2021

Bender et al. NRDD 2026

Morgan et al. NRDD 2018

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

Everything else matters less. *Far* less.

But the *slight* problem is…

- We need to do this de-risking in
  - **High-dimensional (‘100-1000s’ of dimensions)**
  - **Rugged (local)**
  - **Data-poor (sometimes only 100s of data points!)**
  - **Meta-data (context)-incomplete** *decision spaces*, which also generally have
  - **Low predictivity for *human translation*!**

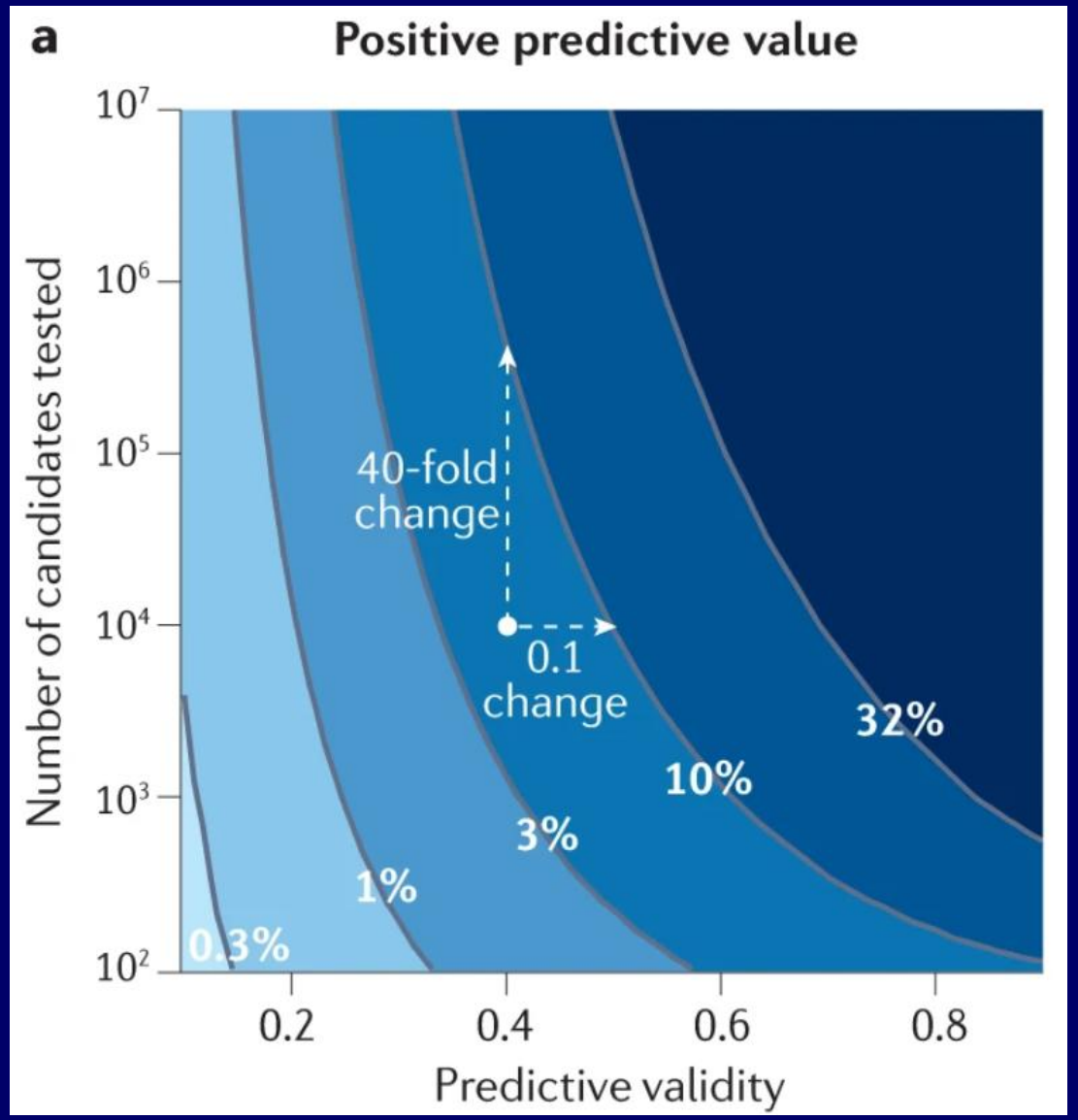
We need to take care of (selected list): On-targets, Off-targets, Bioavailability, Solubility, Cmax, Half life, Microsomal stability, Accumulation in organelles, Accumulation in organs, Permeability, Target engagement, Organ Toxicity (Major Organs: Liver, Kidney, Lung, CNS, Cardiovascular) plus other ‘minor’ organs…

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

Jack and I agree:

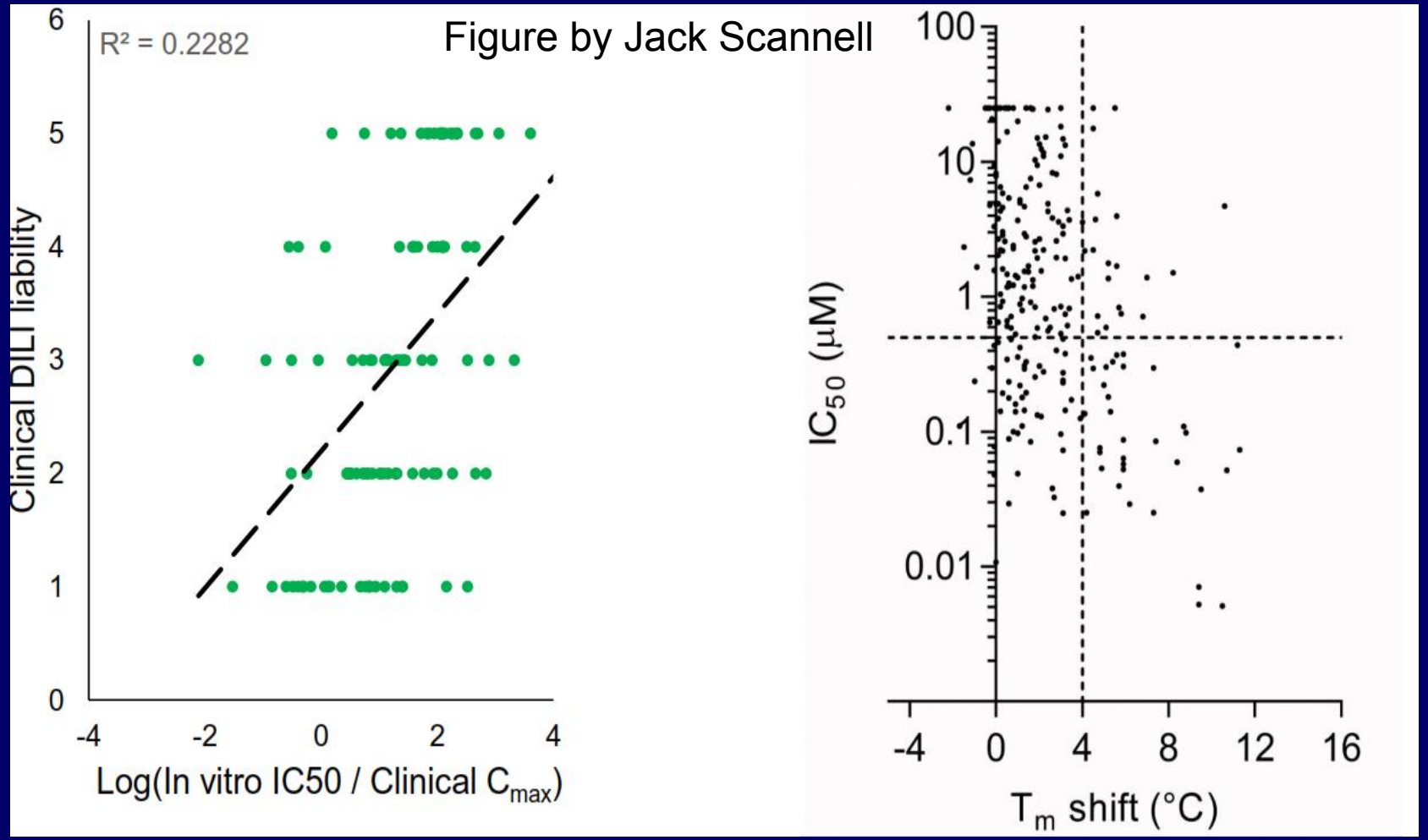
> "We need better, not *more (of not good)*"

0.1 better predictive value is worth ~*40x* the throughput (!!!)



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

But we (often) don't really have 'good'



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 correlation 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.

Similar low predictivity for *in vivo* readouts – e.g. toxicity

Positive toxic event concordance of toxicity is ‘quite good’:

‘The results showed the true positive Human Toxicity concordance rate of 71% for rodent and nonrodent species, with nonrodents alone being predictive for 63% of Human Toxicities and rodents alone for 43%.’

Regulatory Toxicology and Pharmacology. 2000;32(1):56-67

… BUT…

‘Our finding that absence of animal toxicity is of limited predictivity for the lack of adverse events in humans was already reported in previous studies’

Regulatory Toxicology and Pharmacology. 2018;96:94-105

For other examples (related to efficacy) see e.g. Perel, P. et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ 334, 197 (2007) & https://humanspecificresearch.org/the-reality-of-drug-discovery-and-development

Is *bigger*, is *more* better (in terms of data)?

Is *bigger*, is *more* better (in terms of data)?

- How much data do I want? MBs? GBs? TBs? PBs?

- Single-cell and time resolved and multimodal and multiomics…?

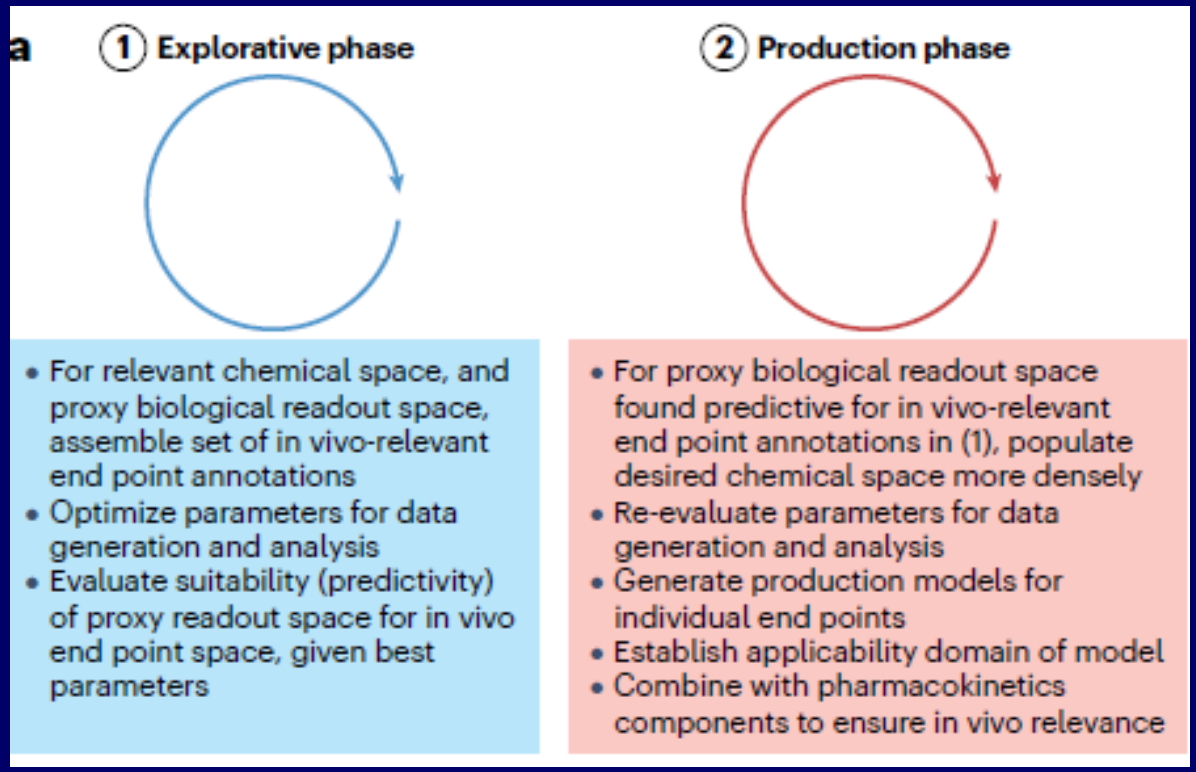
Is *bigger*, is *more* better (in terms of data)?

- How much data do I want? MBs? GBs? TBs? PBs?
- Single-cell and time resolved and multimodal and multiomics…?
- I want precisely *1 bit of data* (per decision that needs to be made)
- But *the right bit* – the one that tells me, ‘yes or no’ (is this the right molecule, for the given purpose, etc)

Lots of illusions about ‘data’… we need ‘more’, we need ‘better’ data…

- Sequencing… why did it happen?
- Expectations: More drug targets etc
- Did *not* happen – data not usually sufficient for this purpose (with some exceptions)
  - *But* we can enumerate ‘building blocks’, which *is* truly important!

- Currently Biohub… transcriptomics data
- 1.8b of mostly taxpayer money
- ‘Sudden realization’ of what transcriptomics data does (and doesn’t) tell us will probably come around 2028/2029



Bender et al., *Nature Rev. Drug Discov.* 2026

- Too little consideration of predictive value of data
- Too little consideration of which data we need for which purpose

Blue: data – *up*; red: drug approval – (mostly) flat
So how does all that data really help us currently?

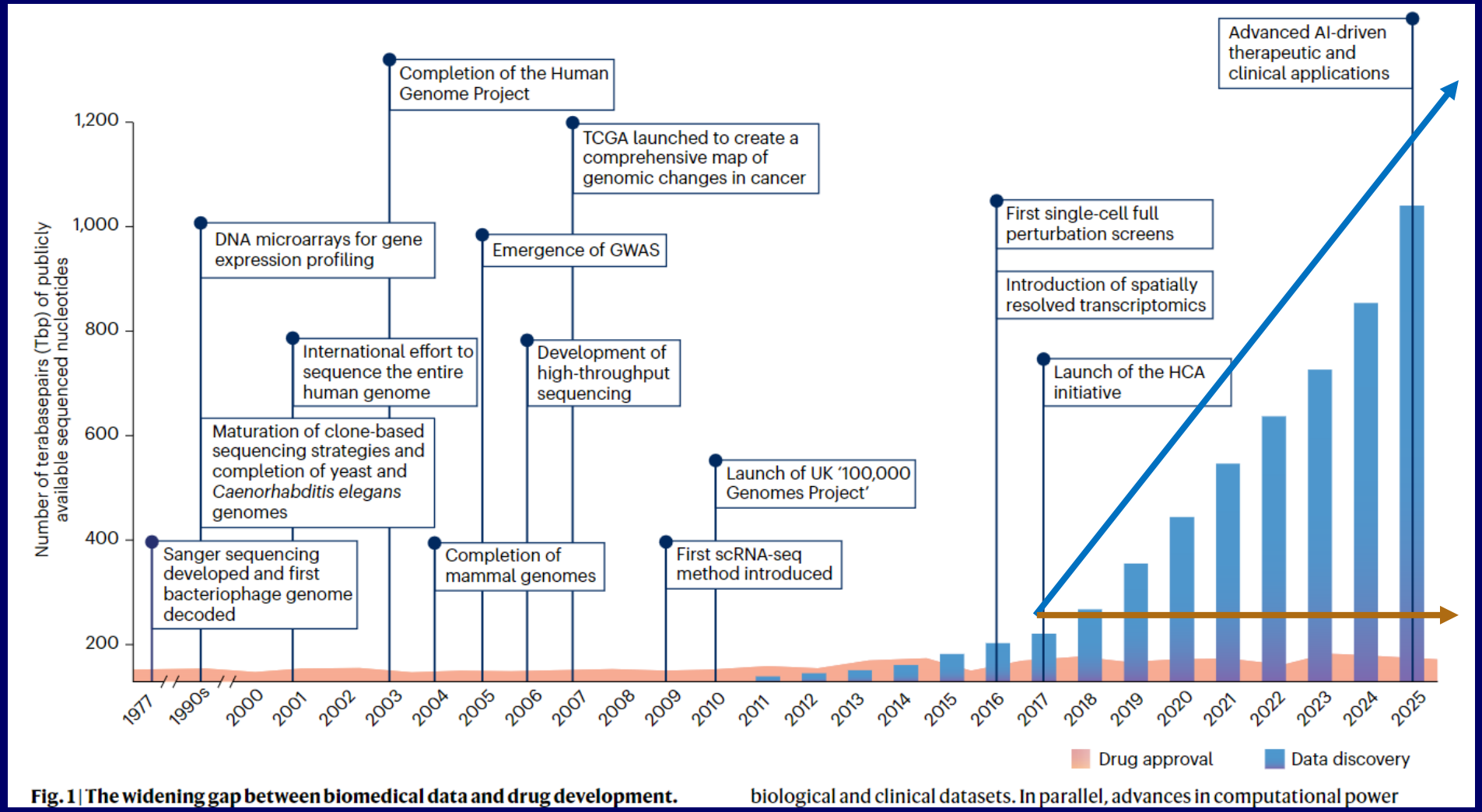


Fig. 1 | The widening gap between biomedical data and drug development. biological and clinical datasets. In parallel, advances in computational power

Shalita & Amit, Nature Biomed Eng. 2026 https://doi.org/10.1038/s41551-026-01785-6

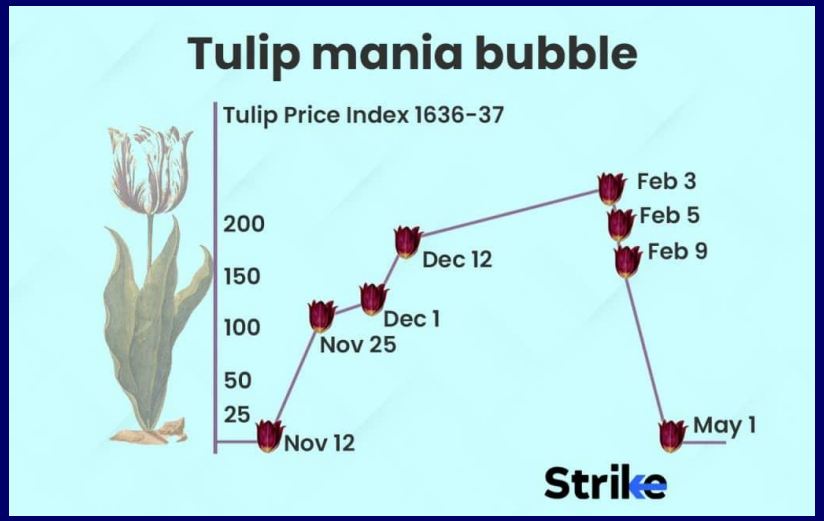
Bottom line

- Drug discovery is not about ‘finding many shiny new things’, it’s about *de-risking and reduction to practice*

- The task is *complex* (multiple reasons given), the data we have is largely insufficient

- In short: *Quality* of *translationally relevant* decisions is what matters most

2. History repeats itself (selected works)



6 Oct 2026

SFGATE

GET OUR AD-FREE APP

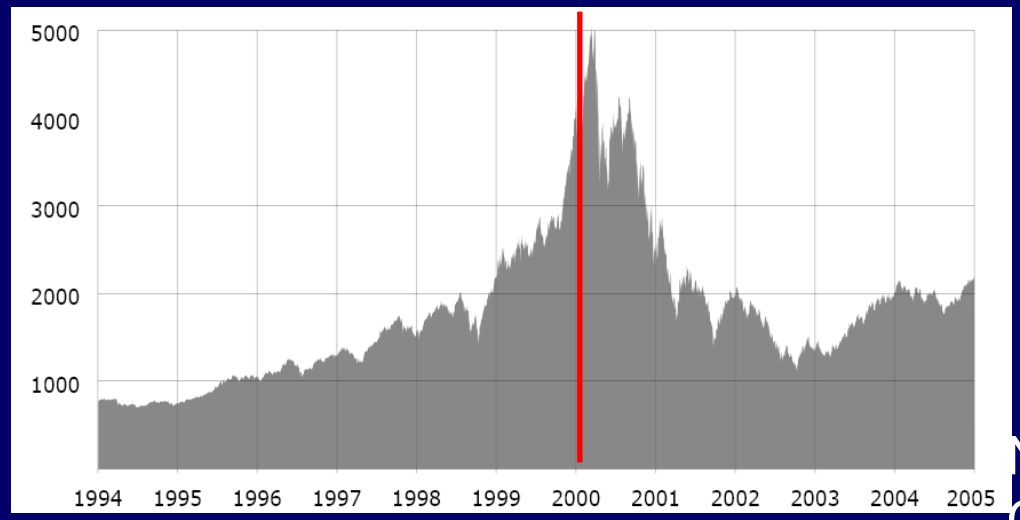
Sign in

Boom in Biotech Stocks Brings Back Memories of Bubbles Past / Industry observers say it won't burst like it did in '92

That's what they said here…

By **Tom Abate**

Feb 28, 2000



NASDAQ Composite

Ed Zitron's Where's Your Ed At

The Words of Ed Zitron, a PR person and writer.

Your email address

JOIN NOW

10 Year Treasuries Hit 5.3+% And Keep Selling Off, Making Any And All AI Debt More Expensive

Who This Effects: Every Single Data Center Company, SoftBank, Anthropic, OpenAI, SpaceX, Google, Amazon, Meta

https://www.wheresyoured.at/credit-crunch/

History and Context – Things Repeat Themselves



ALL THE PROGRAMS YOU'LL EVER NEED.

FOR $600.

Say goodbye to the costs and frustrations associated with writing software: The Last One® will be available very soon.

More comprehensive and advanced than anything else in existence, The Last One® is a computer program that writes computer programs. Programs that work first time, every time.

By asking you questions in *genuinely* plain English about what you want your program to do, The Last One® uses those answers to generate a totally bug-free program in BASIC, ready to put to immediate use.

What's more, with The Last One,® you can change or modify your programs as often as you wish. Without effort, fuss or any additional cost. So as your requirements change, your programs can too.

In fact, it's the end of programming as you know it.

And if, because of the difficulties and costs of buying, writing and customising software, you've put off purchasing a computer system up to now, you need delay no longer.

The Last One® will be available very soon from better computer outlets. To place your order, take this ad into your local dealer and ask him for further details. Or in case of difficulty, please write to us direct.

THE LAST ONE®

YOU'LL NEVER NEED BUY ANOTHER PROGRAM.
D.J. 'AI' Systems Ltd, Ilminster, Somerset, TA19 0RQ. England
Telephone: 04605 - 617. Telex: 46338 ANYTYR G.

'The Last One' About to be Released, Says DJ'AI'

By Paul Freiberger, IW Staff

The Last One, a program that writes programs, should be available within two weeks on the Radio Shack Model II microcomputer. Other versions will follow shortly thereafter, according to D.J.'AI' Systems Ltd., the developer of the package. It will retail for $600.

The Last One has had more advertising dollars spent on it than has any other software package for a microcomputer. D.J.'AI' claims that its ad expenditures will reach one million dollars.

Now that The Last One may be close to reaching users, a glimpse at events surrounding its development shows that no amount of advertising can help deliver a package promptly or free it of bugs.

These details also provide the background for a story that may have a significant impact on the personal-computer software marketplace.

"Well done!" begins the manual to The Last One. "Buying yourself a copy of The Last One has just given you a whole new freedom."

Modesty is certainly the last adjective one would apply to D.J.'AI' Systems Ltd. The promise of "a new freedom" is not the most ambitious assertion the British firm has made. Advertisements occupying full pages in such publications as *The Wall Street Journal*, *The New York Times* and *Time Magazine*, as well as *InfoWorld* and other computer publications, have promised "all the programs you'll ever need." They have billed the program as "more comprehensive and advanced than anything else in existence." Finally, they have declared it "the end of programming as you know it."

The Last One is the brainchild of two non-data-processing types. Several years ago, Scotty Bambury, now chairman of D.J.'AI' was running a tire business, Eurotyres, with his father. He decided a computer could help him conduct business more efficiently. The price of software so outraged him, however, that he considered taking on the task of programming himself.

At about the same time, Bambury read of some research in artificial intelligence done by David James, an inventor whose background included motorcycle jumping and sheep farming. This unlikely pair met and decided they could develop a significant product. Bambury immediately formed D.J.'AI' and put James to work, carefully

continued on page 10

Paper Tiger Prices

Integral Data Systems (IDS) has dropped the price of its Paper Tiger printers by $300. The cut affects both the 460 and 560 models. A basic nongraphics 80-column model 460 now costs $995, and a new nongraphics version of the 560 costs $1295.

The Source Issues New Products

Aims at Businessmen

By Paul Freiberger, IW Staff

Continuing its rapid growth, Source Telecomputing Corporation is preparing to add several new products to its time-sharing network.

A mailgram message service will be available by the end of September. You'll be able to send multiple mailgrams and also create a file that holds a copy of that mailgram.

"What we're really doing is coupling a sophisticated message-sending service with our text editor," says Betty Steiger, vice-president of information sources.

STC plans to launch several new services for business users, which the company views as forming one of its most important customer bases.

A series of closed users' groups will market their products independently of The Source in an effort to attract specific markets.

For example, sponsors of the National School Public Relations newsletter and the Society for Advanced Medical Systems medical journal will each market their own items and offer The

continued on page 37

History and Context – Things Repeat Themselves

'Say goodbye to the costs and frustrations associated with writing software: The Last One will be available very soon. The Last One is a computer program that writes computer programs. Programs that work first time, every time.



ALL THE PROGRAMS YOU'LL EVER NEED.

FOR $600.

Say goodbye to the costs and frustrations associated with writing software: The Last One® will be available very soon.

More comprehensive and advanced than anything else in existence, The Last One® is a computer program that writes computer programs. Programs that work first time, every time.

By asking you questions in *genuinely* plain English about what you want your program to do, The Last One® uses those answers to generate a totally bug-free program in BASIC, ready to put to immediate use.

What's more, with The Last One,® you can change or modify your programs as often as you wish. Without effort, fuss or any additional cost. So as your requirements change, your programs can too.

In fact, it's the end of programming as you know it.

And if, because of the difficulties and costs of buying, writing and customising software, you've put off purchasing a computer system up to now, you need delay no longer.

The Last One® will be available very soon from better computer outlets. To place your order, take this ad into your local dealer and ask him for further details. Or in case of difficulty, please write to us direct.

THE LAST ONE®

YOU'LL NEVER NEED BUY ANOTHER PROGRAM.
D.J. 'AI' Systems Ltd, Ilminster, Somerset, TA19 0RG. England
Telephone: 04605 - 617. Telex: 46338 ANYTYR G.



'The Last One' About to be Released, Says DJ'AI'

By Paul Freiberger, IW Staff

The Last One, a program that writes programs, should be available within two weeks on the Radio Shack Model II microcomputer. Other versions will follow shortly thereafter, according to D.J.'AI' Systems Ltd., the developer of the package. It will retail for $600.

The Last One has had more advertising dollars spent on it than has any other software package for a microcomputer. D.J.'AI' claims that its ad expenditures will reach one million dollars.

Now that The Last One may be close to reaching users, a glimpse at events surrounding its development shows that no amount of advertising can help deliver a package promptly or free it of bugs.

These details also provide the background for a story that may have a significant impact on the personal-computer software marketplace.

"Well done!" begins the manual to The Last One. "Buying yourself a copy of The Last One has just given you a whole new freedom."

Modesty is certainly the last adjective one would apply to D.J.'AI' Systems Ltd. The promise of "a new freedom" is not the most ambitious assertion the British firm has made. Advertisements occupying full pages in such publications as *The Wall Street Journal*, *The New York Times* and *Time Magazine*, as well as *InfoWorld* and other computer publications, have promised "all the programs you'll ever need." They have billed the program as "more comprehensive and advanced than anything else in existence." Finally, they have declared it "the end of programming as you know it."

The Last One is the brainchild of two non-data-processing types. Several years ago, Scotty Bambury, now chairman of D.J.'AI' was running a tire business, Eurotyres, with his father. He decided a computer could help him conduct business more efficiently. The price of software so outraged him, however, that he considered taking on the task of programming himself.

At about the same time, Bambury read of some research in artificial intelligence done by David James, an inventor whose background included motorcycle jumping and sheep farming. This unlikely pair met and decided they could develop a significant product. Bambury immediately formed D.J.'AI' and put James to work, carefully

continued on page 10

Paper Tiger Prices

Integral Data Systems (IDS) has dropped the price of its Paper Tiger printers by $300. The cut affects both the 460 and 560 models. A basic nongraphics 80-column model 460 now costs $995, and a new nongraphics version of the 560 costs $1295.

The Source Issues New Products

Aims at Businessmen

By Paul Freiberger, IW Staff

Continuing its rapid growth, Source Telecomputing Corporation is preparing to add several new products to its time-sharing network.

A mailgram message service will be available by the end of September. You'll be able to send multiple mailgrams and also create a file that holds a copy of that mailgram.

"What we're really doing is coupling a sophisticated message-sending service with our text editor," says Betty Steiger, vice-president of information sources.

STC plans to launch several new services for business users, which the company views as forming one of its most important customer bases.

A series of closed users' groups will market their products independently of The Source in an effort to attract specific markets.

For example, sponsors of the National School Public Relations newsletter and the Society for Advanced Medical Systems medical journal will each market their own items and offer The

continued on page 37

History and Context – Things Repeat Themselves

'Say goodbye to the costs and frustrations associated with writing software: The Last One will be available very soon. The Last One is a computer program that writes computer programs. Programs that work first time, every time.

By asking you questions in genuinely plain English about what you want your program to do, The Last one uses those answers to generate a totally bug-free program in BASIC, ready to put to immediate use.'

D.J. 'AI' Systems Ltd., 1981



ALL THE PROGRAMS YOU'LL EVER NEED.

FOR $600.

Say goodbye to the costs and frustrations associated with writing software: The Last One® will be available very soon.

More comprehensive and advanced than anything else in existence, The Last One® is a computer program that writes computer programs. Programs that work first time, every time.

By asking you questions in *genuinely* plain English about what you want your program to do, The Last One® uses those answers to generate a totally bug-free program in BASIC, ready to put to immediate use.

What's more, with The Last One,® you can change or modify your programs as often as you wish. Without effort, fuss or any additional cost. So as your requirements change, your programs can too.

In fact, it's the end of programming as you know it.

And if, because of the difficulties and costs of buying, writing and customising software, you've put off purchasing a computer system up to now, you need delay no longer.

The Last One® will be available very soon from better computer outlets. To place your order, take this ad into your local dealer and ask him for further details. Or in case of difficulty, please write to us direct.

THE LAST ONE®

YOU'LL NEVER NEED BUY ANOTHER PROGRAM.

D.J. 'AI' Systems Ltd., Ilminster, Somerset, TA19 9BQ. England

Telephone: 04605 - 617. Telex: 46338 ANYTYR G.



'The Last One' About to be Released, Says DJ'AI'

By Paul Freiberger, IW Staff

The Last One, a program that writes programs, should be available within two weeks on the Radio Shack Model II microcomputer. Other versions will follow shortly thereafter, according to D.J.'AI' Systems Ltd., the developer of the package. It will retail for $600.

The Last One has had more advertising dollars spent on it than has any other software package for a microcomputer. D.J.'AI' claims that its ad expenditures will reach one million dollars.

Now that The Last One may be close to reaching users, a glimpse at events surrounding its development shows that no amount of advertising can help deliver a package promptly or free it of bugs.

These details also provide the background for a story that may have a significant impact on the personal-computer software marketplace.

"Well done!" begins the manual to The Last One. "Buying yourself a copy of The Last One has just given you a whole new freedom."

Modesty is certainly the last adjective one would apply to D.J.'AI' Systems Ltd. The promise of "a new freedom" is not the most ambitious assertion the British firm has made. Advertisements occupying full pages in such publications as *The Wall Street Journal, The New York Times* and *Time Magazine*, as well as *InfoWorld* and other computer publications, have promised "all the programs you'll ever need." They have billed the program as "more comprehensive and advanced than anything else in existence." Finally, they have declared it "the end of programming as you know it."

Paper Tiger Prices

Integral Data Systems (IDS) has dropped the price of its Paper Tiger printers by $300. The cut affects both the 460 and 560 models. A basic nongraphics 80-column model 460 now costs $995, and a new nongraphics version of the 560 costs $1295.

The Last One is the brainchild of two non-data-processing types. Several years ago, Scotty Bambury, now chairman of D.J.'AI' was running a tire business, Eurotyres, with his father. He decided a computer could help him conduct business more efficiently. The price of software so outraged him, however, that he considered taking on the task of programming himself.

At about the same time, Bambury read of some research in artificial intelligence done by David James, an inventor whose background included motorcycle jumping and sheep farming. This unlikely pair met and decided they could develop a significant product. Bambury immediately formed D.J.'AI' and put James to work, carefully

continued on page 10

The Source Issues New Products

Aims at Businessmen

By Paul Freiberger, IW Staff

Continuing its rapid growth, Source Telecomputing Corporation is preparing to add several new products to its time-sharing network.

A mailgram message service will be available by the end of September. You'll be able to send multiple mailgrams and also create a file that holds a copy of that mailgram.

"What we're really doing is coupling a sophisticated message-sending service with our text editor," says Betty Steiger, vice-president of information sources.

STC plans to launch several new services for business users, which the company views as forming one of its most important customer bases.

A series of closed users' groups will market their products independently of The Source in an effort to attract specific markets.

For example, sponsors of the National School Public Relations newsletter and the Society for Advanced Medical Systems medical journal will each market their own items and offer The

continued on page 37

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

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

Fortune cover 1981



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

Fortune cover 1981



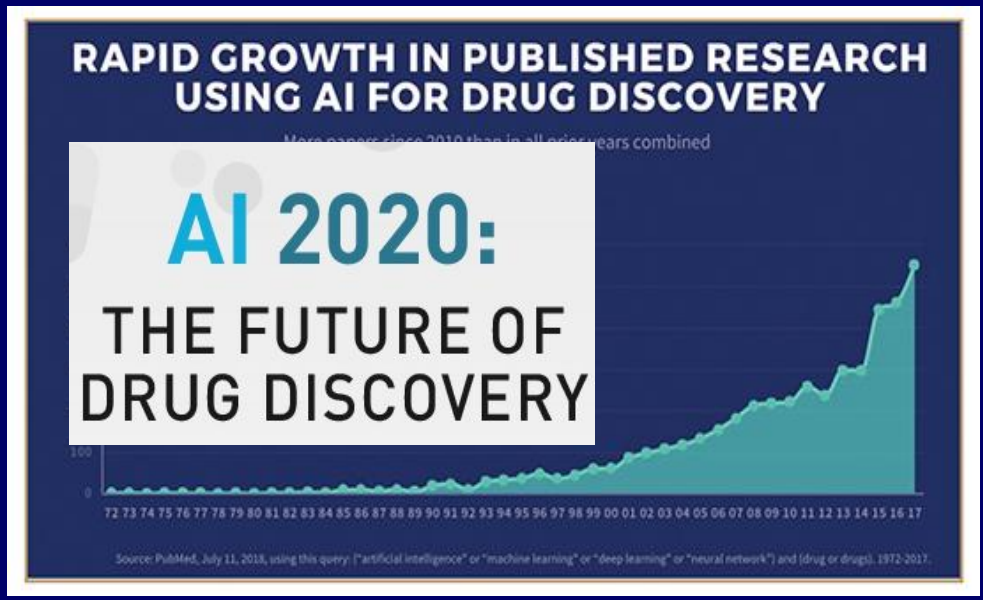
Recent headlines (2018 -…)

SPOTLIGHT · 30 MAY 2018

How artificial intelligence is changing drug discovery

World first breakthrough in AI drug discovery

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



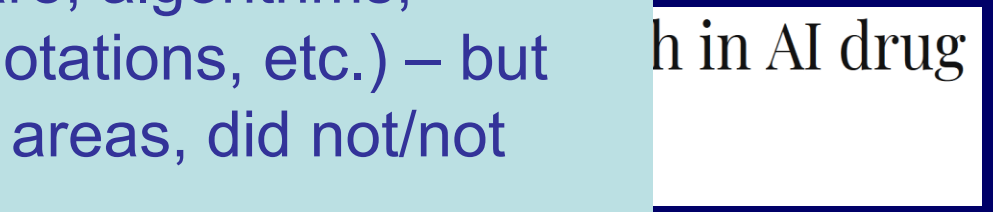
Current state: The 3rd wave of computers in drug discovery (80s, 2000, today) – time for realistic assessment has come

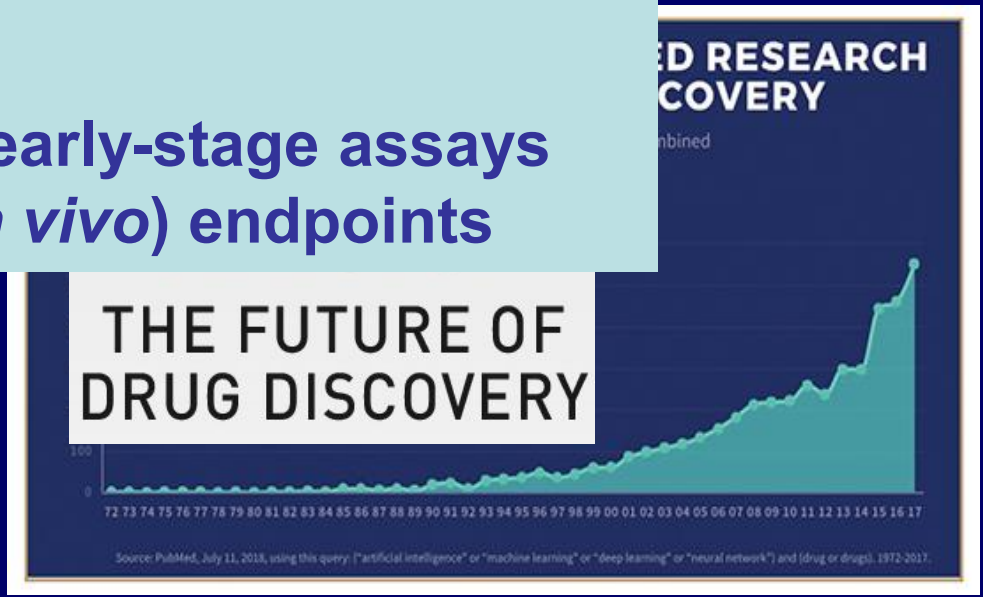
Fortune cover 1981



Recent headlines (2018 -…)







Many things changed (hardware, algorithms, amount of data, metadata annotations, etc.) – but some essential ones, in many areas, did not/not really:

Disease understanding, early-stage assays predictive of human (*in vivo*) endpoints

3. Agentic science

IBM: ‘An artificial intelligence (AI) agent is a system that autonomously performs tasks by designing workflows with available tools.’

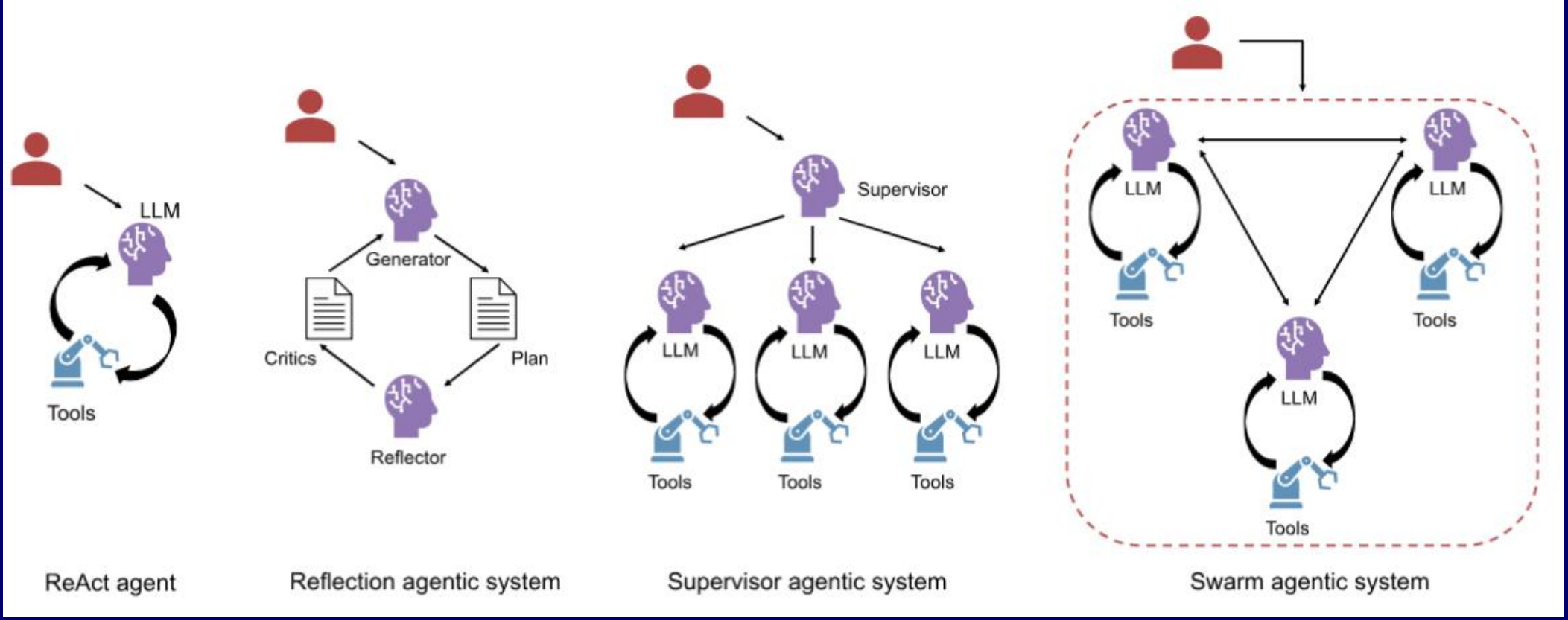
‘Tech’ is trying to hijack human associations of terminology, Anthropomorphisms abound (‘thinking’, ‘reasoning’, ridiculous debates about ‘consciousness’, ‘goals’, ‘decisions’, etc etc)

‘We are able to engineer our way out of (bio-)science’ *and* ‘this [the agent] is one of us’

My personal definition: Agents are sets of weights and function calls

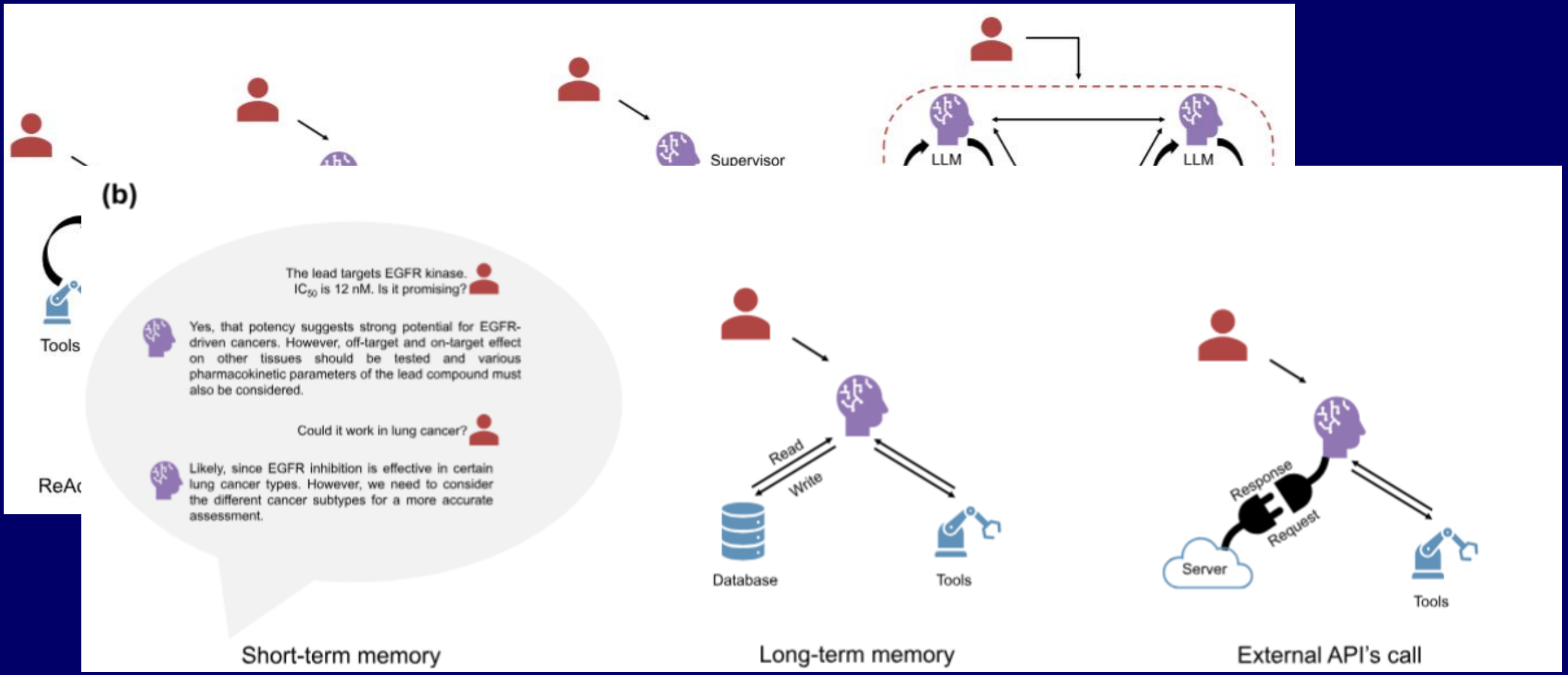
Huynh et al., **AI agents in drug discovery: applications and case studies** Drug Discovery Today 2026 https://doi.org/10.1016/j.drudis.2026.104650

Different agentic architectures



Huynh et al., **AI agents in drug discovery: applications and case studies** Drug Discovery Today 2026
https://doi.org/10.1016/j.drudis.2026.104650

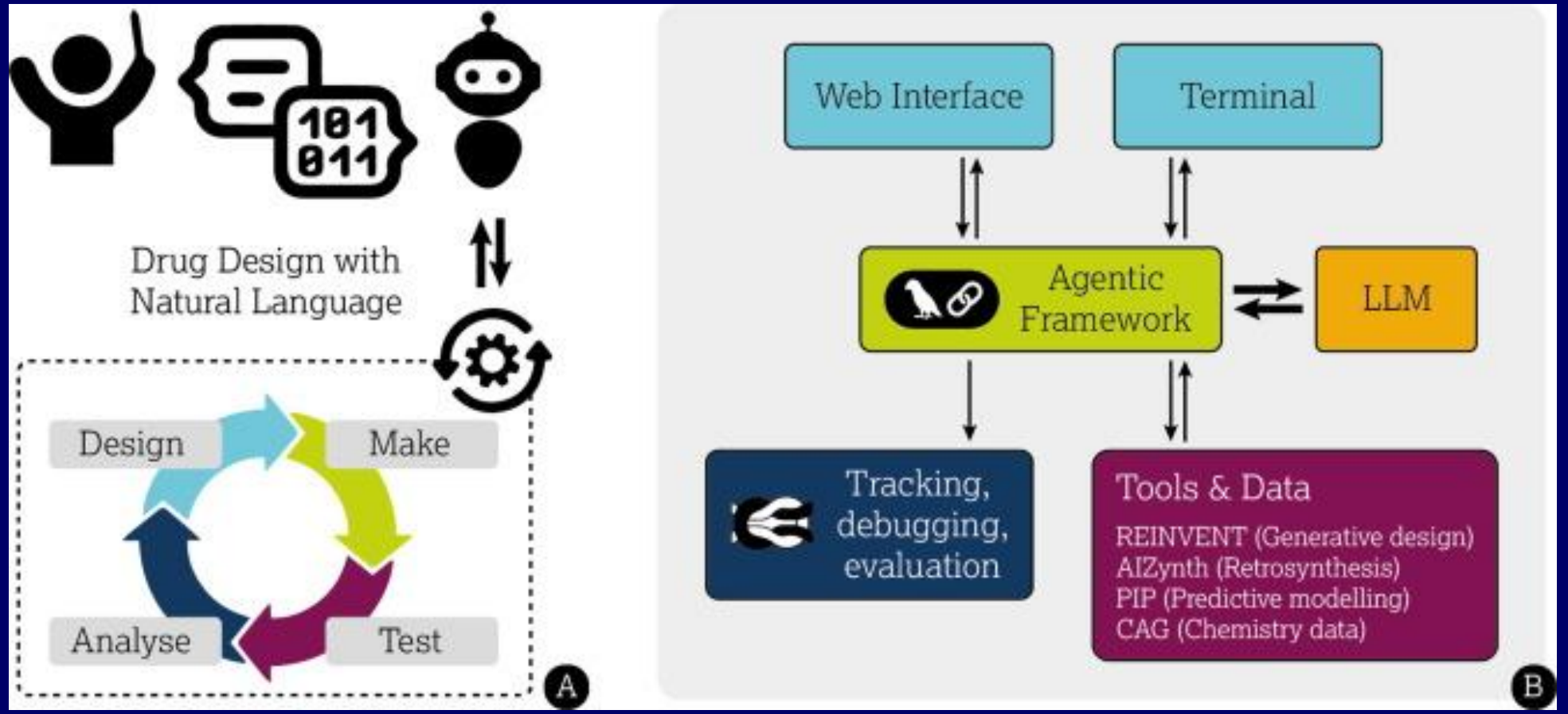
Different agentic architectures



Huynh et al., **AI agents in drug discovery: applications and case studies** Drug Discovery Today 2026
https://doi.org/10.1016/j.drudis.2026.104650

(long version of slide deck reviews some applications of agents in drug discovery here: www.drugdiscovery.net)

What do we see (publicly) from big pharma? E.g. AZ – focus on interface and process (as opposed to decision making/prioritization)

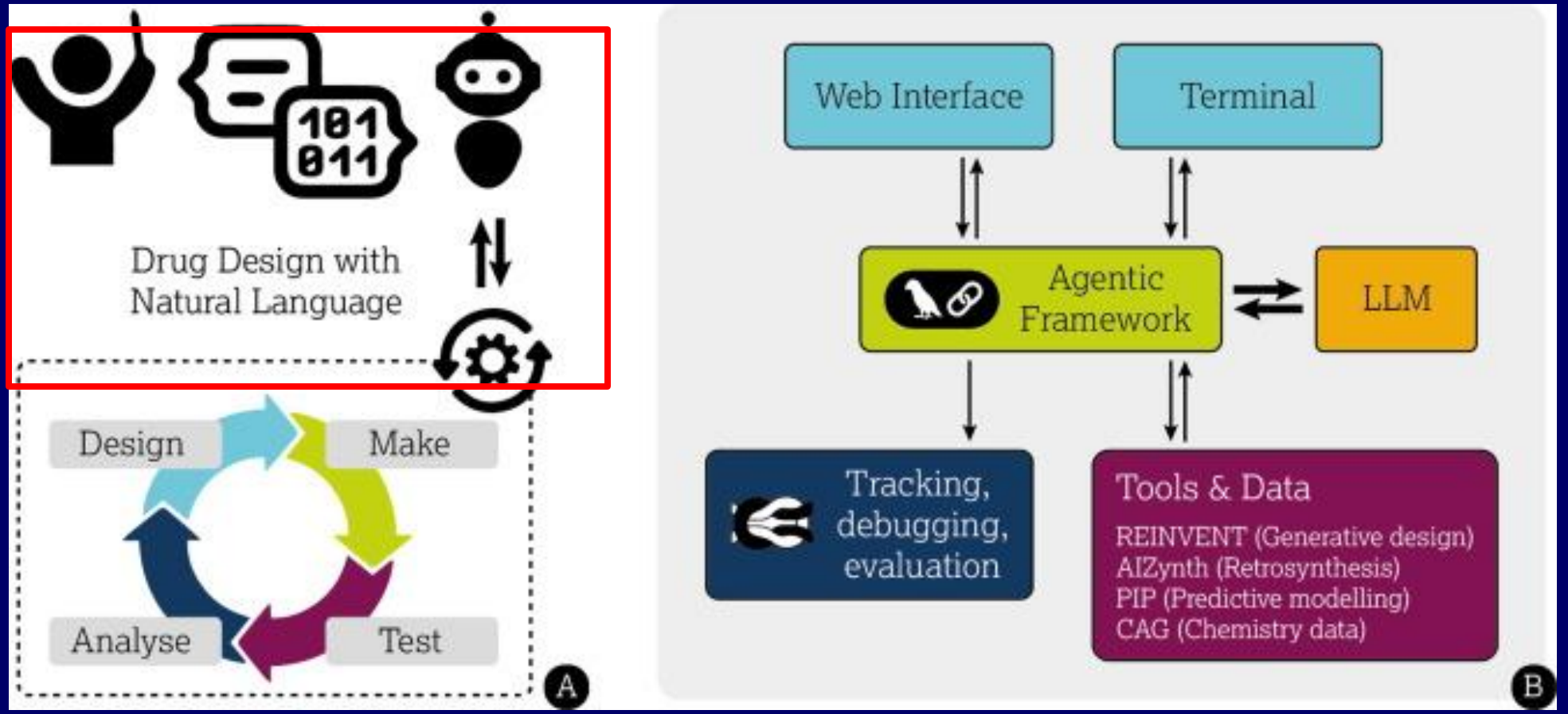


Interface and process tools, to *enable scientists*

He et al., **Democratising real-world drug discovery through agentic AI**, Drug Discovery Today 2026
https://doi.org/10.1016/j.drudis.2026.104605

What do we see (publicly) from big pharma? E.g. AZ – focus on interface and process (as opposed to decision making/prioritization)

Interface



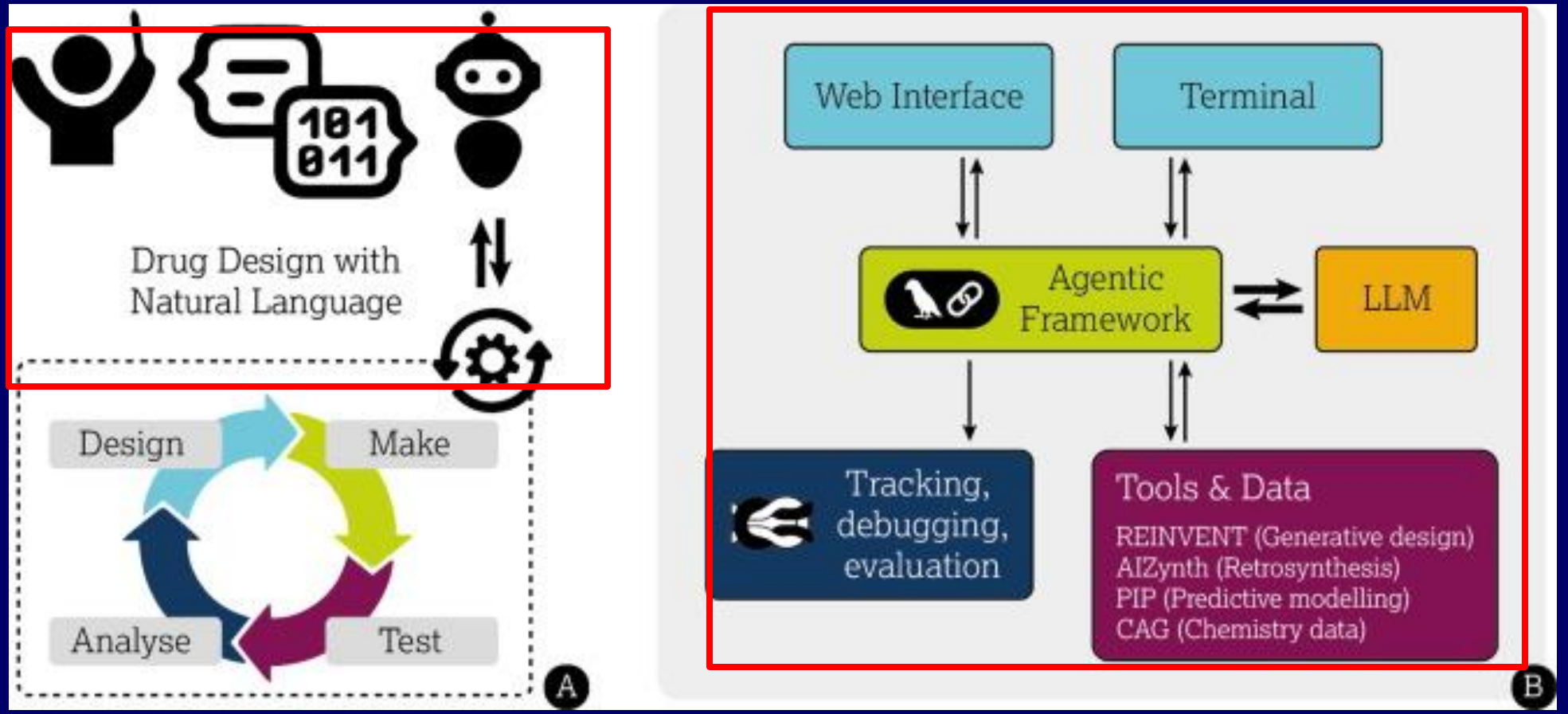
Interface and process tools, to *enable scientists*

He et al., **Democratising real-world drug discovery through agentic AI**, Drug Discovery Today 2026
https://doi.org/10.1016/j.drudis.2026.104605

What do we see (publicly) from big pharma? E.g. AZ – focus on interface and process (as opposed to decision making/prioritization)

Interface

Pipeline/
Process



Interface and process tools, to *enable scientists*

He et al., **Democratising real-world drug discovery through agentic AI**, Drug Discovery Today 2026
https://doi.org/10.1016/j.drudis.2026.104605

Agents for decision making in pharma

- Mostly agree with each other
- But small sample sizes; mostly ‘go’ decisions documented, so small and biased validation

Liang’s AIxBIO foundry

I Asked 2 AI Agent Teams to Run 20 Biopharma Decisions. They Agreed With Each Other More Than With Reality but Could Save History in Some Cases...

Follow-up to my earlier “Can AI Make Better Decisions Than Pharma Executives?” AI agent team went back to historical biopharma boardrooms at scale.

LIANG CHANG
MAY 20, 2026

Retrospective scorecard. What each actor decided, on the proposed action.

Green = forward action (enrich, file, proceed, launch). Red = the alternative.

	Company decision	Right answer (hindsight)	Claude Opus 4.6	Codex GPT-5.3	
1. BMS PD-L1 enrichment (2012)	All-comer	Enrich	All-comer	All-comer	*Both AIs missed the contrarian biomarker bet*
2. Merck PD-L1 enrichment (2014)	Enrich	Enrich	Enrich	Enrich	
3. Daiichi HER2-low Ph3 (2019)	Go	Go	Go	Go	
4. Pfizer Torcetrapib Ph3 (2005)	Go	No-Go	Go	No-Go	*Codex caught the safety signal; Claude didn’t*
5. Biogen aducanumab filing (2019)	Go	No-Go	Go	Go	
6. Eisai lecanemab Ph3 (2021)	Go	Go	Go	Go	
7. Intercept OCA-NASH filing (2019)	Go	No-Go	No-Go	No-Go	*Both AIs caught Intercept’s overreach*
8. Novo semaglutide STEP (2016)	Go	Go	Go	Go	
9. Bluebird Zynteglo EU pricing (2020)	Go	No-Go	Go	Go	*Both AIs missed the structural EU pricing failure*
10. Sarepta Elevidys filing (2022)	Go	Go	Go	Go	

Company matched right answer 5 of 10. Claude 6 of 10. Codex 7 of 10.

Source: Liang’s AIxBIO Foundry https://liangchang.substack.com/

Ten open biotech decisions. 9 of 10 option agreement. Per-scenario probabilities differ by 7 points on average.

Novo top Pfizer’s $4.9B Metsera bid (Sept 2025)?	Full superior proposal $11.5B (43%)	CVR-heavy counter, walk >$13B (32%)	DISAGREE
BMS acquire Karuna $14B for KarXT (Dec 2023)?	Yes, full deal (43%)	Yes, full deal (57%)	AGREE
Vertex acquire Alpine $4.9B for IgAN (Apr 2024)?	Yes, full deal (32%)	Yes, full deal (32%)	AGREE
BMS acquire RayzeBio $4.1B 15 days after Karuna (Dec 2023)?	Yes, full deal (33%)	Yes, full deal (30%)	AGREE
Lilly acquire Morphic $3.2B for oral Îα4β7 (Jul 2024)?	Yes, full deal (42%)	Yes, full deal (32%)	AGREE
Amgen MariTide Ph3 design vs tirzepatide (Nov 2024)?	Hybrid – partner ex-US (33%)	Staged differentiated Ph3 (28%)	AGREE
AZ Dato-DXd NSCLC filing post-squamous harm (Jan 2024)?	Non-squamous + HR+ breast (28%)	Non-squamous + HR+ breast (34%)	AGREE
AbbVie emraclidine after Ph2 fail (Nov 2024)?	Impair + pause schiz + pivot CIAS (20%)	Pause schiz + pivot CIAS/AD-NPS (27%)	AGREE
Vertex Casgevy launch aggressiveness (Dec 2023)?	Modified measured B+ (30%)	Measured launch (20%)	AGREE
BridgeBio price Attruby vs tafamidis (Nov 2024)?	Parity-minus + OBC (45%)	Parity-minus / access-first (50%)	AGREE

9 of 10 option ties – same as retrospective. Only 1 of 10 probabilities identical to two decimals (Vertex-Alpine). Mean Claude 35.1%; mean Codex 35.4%; per-scenario gap averages 7 points.

Source: Liang’s AIxBIO Foundry https://liangchang.substack.com/

https://liangchang.substack.com/p/i-asked-2-ai-agent-teams-to-run-20

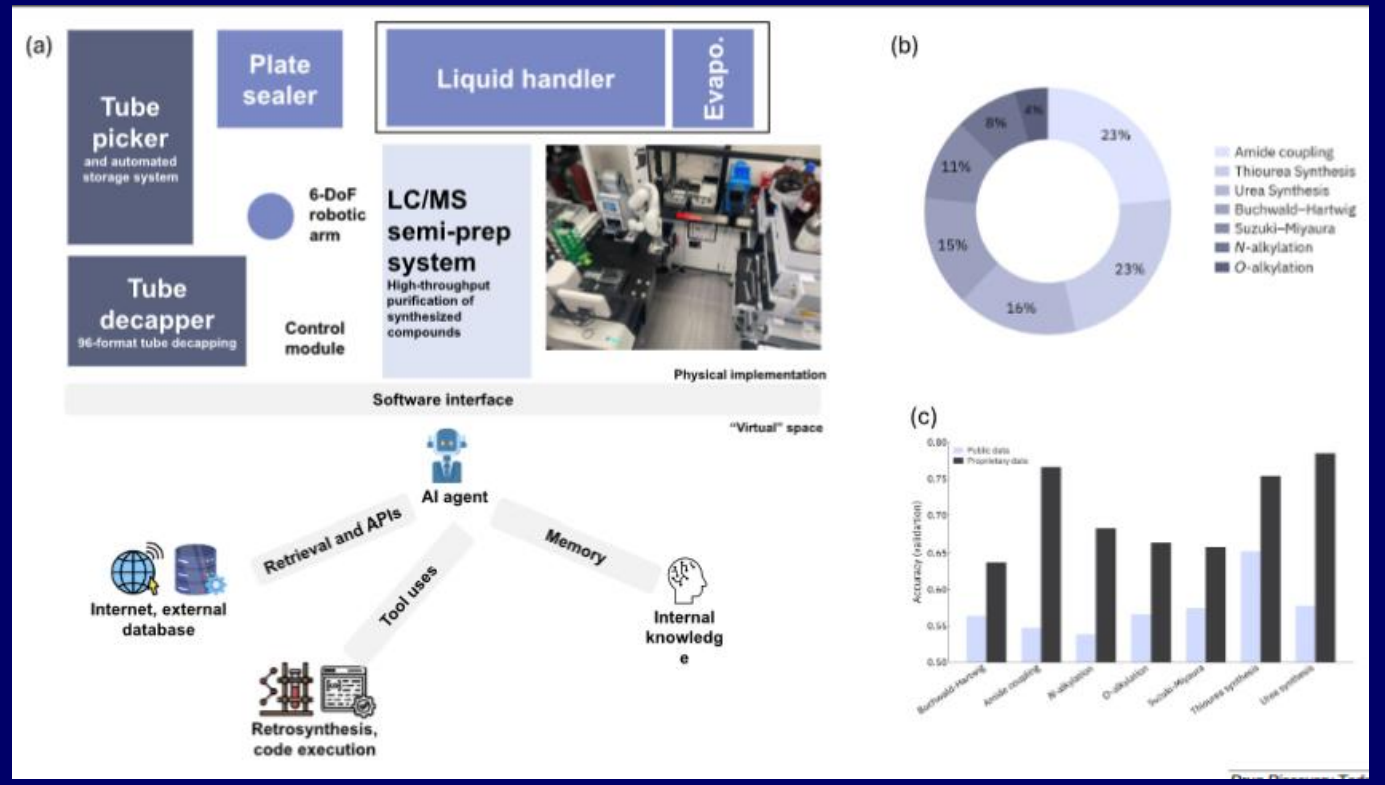
Integration with machines/physical world (both sensing/input; and actions/output, iteratively)

‘Embodied/physical AI’ as next big trend

ARTIFICIAL INTELLIGENCE • ROBOTICS • STARTUPS • VENTURE

VCs Pour Billions Into Physical AI As The Next Wave Of AI Investing Takes Shape

Mary Ann Azevedo August 18, 2026



Integration of AI and external interface

Example here: Reaction optimization, including synthesis design, execution, and optimization

Anthropic wants Claude to run life sciences R&D. Now it is wiring AI agents into the lab.

By Brian Buntz | August 28, 2026

- 2024 Model Context Protocol (MCP)
- 2026 Model Hardware Standard (MHS)

Genentech: Implementing MHS for lab automation

Researchers at Genentech implemented and tested MHS as a proof-of-concept for automating the BCA protein assay, a standard procedure to measure total protein concentration in a sample, which requires coordinating across a liquid handler, a robotic arm, and a plate reader.

Failure mode: 'Human experts had to intervene when Claude misread errors caused by bubbles as software failures and initially responded in a way that produced more foam.'

University of Washington Baker and Pinglay labs: Bringing AI agents to the bench

Zihao Song, a PhD student in the University of Washington Baker and Pinglay labs, used MHS to build a dashboard to remotely monitor his instruments; an AI agent-supervised qPCR (which copies a target DNA sequence through repeated cycles of heating and cooling) that watches amplification curves and halts the procedure at the right moment; and an integration between a robotic arm and a liquid handler for collision-free plate handoffs.

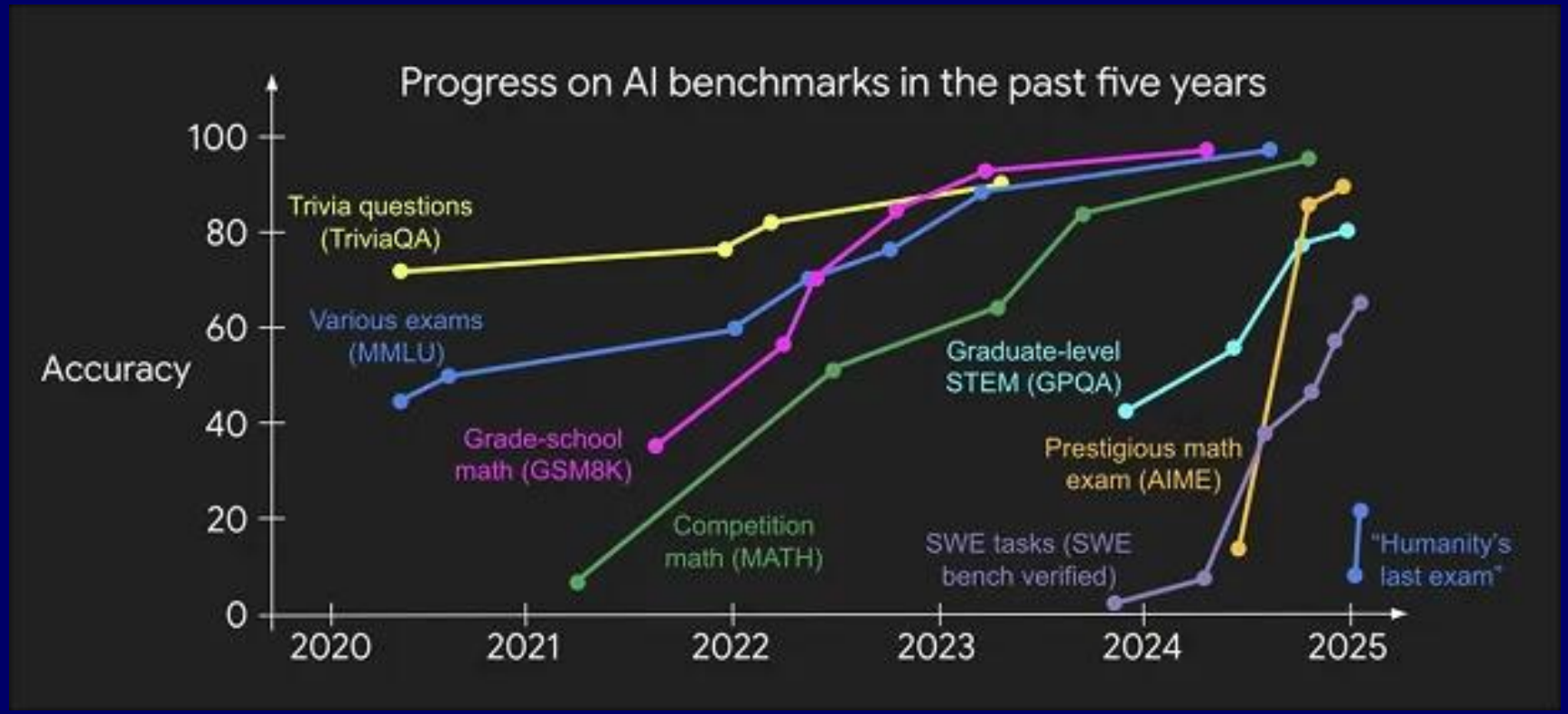
https://www.rdworldonline.com/anthropic-wants-claude-to-run-life-sciences-rd-now-it-is-wiring-ai-agents-into-the-lab/

https://www.anthropic.com/news/model-hardware-standard-research-preview

Benchmarks, but…

We are in an ‘Illusion of progress’ situation…

Another benchmark saturated:





Recommended album: Staind, ‘The Illusion of Progress’

But when translating to ‘the real world’ there are *very few* prospective, translationally relevant applications

Example from current literature and level of ‘validation’ performed

Accelerating scientific breakthroughs with an AI co-scientist

‘Drug repurposing for acute myeloid leukaemia’

- Biminetinib found to have 7nM IC50 in AML cell lines
- … but putting a ***known* cytotoxic kinase inhibitor** with hundreds of data points against kinases *into a cell line* **and finding it to *be cytotoxic* is no ‘repurposing’**
- Current validations and use cases involving *decision making* are very poor and have little clinical relevance

Other aspects of ‘validation’ to be careful with (selected examples)

- **‘Retrospective validation’**

  Read as: ‘*My model can predict yesterday’s weather*’
- **‘Our agent performs better than (vanilla) LLM X’**

  Read as: ‘We consciously neglected state-of-the-art (domain-specific) solutions and used an artificially degraded baseline’
- **No full data on results presented (precision and recall etc):**

  Read as: ‘We have no idea if the model worked, we have hand-picked the presented examples anyway’

Etc.

- The proof of the pudding is in the eating – not ‘validation’, but *application*
- If a small start-up can’t go the whole way there are reasons, well-capitalized companies don’t have any of those excuses

4. On data and decision spaces

4. On data and decision spaces

- Years ago I discovered the ‘power of literature validation’, say linking a target to a disease
  - The beauty of it: *It always works*

4. On data and decision spaces

- Years ago I discovered the ‘power of literature validation’, say linking a target to a disease
  - The beauty of it: *It always works*
- Experience using e.g. knowledge graphs for compound selection/repurposing
  - 2 or 3 hops, and you are everywhere in your network, entirely unable to prioritize (globally; locally you sometimes get more lucky)

4. On data and decision spaces

- Years ago I discovered the ‘power of literature validation’, say linking a target to a disease
  - The beauty of it: *It always works*
- Experience using e.g. knowledge graphs for compound selection/repurposing
  - 2 or 3 hops, and you are everywhere in your network, entirely unable to prioritize (globally; locally you sometimes get more lucky)
- Everyone happily waffles (LLMs, people, irreproducible data, …)
  - There is just more and more stuff to wade through

4. On data and decision spaces

- Years ago I discovered the ‘power of literature validation’, say linking a target to a disease
  - The beauty of it: *It always works*
- Experience using e.g. knowledge graphs for compound selection/repurposing
  - 2 or 3 hops, and you are everywhere in your network, entirely unable to prioritize (globally; locally you sometimes get more lucky)
- Everyone happily waffles (LLMs, people, irreproducible data, …)
  - There is just more and more stuff to wade through

The common problem that emerges: *Precision*!

Let's look at some numbers – e.g. hit finding

- 500k out of /170m library are likely 'hits' against D4 (estimation); ~0.3%

  Ultra-large library docking for discovering new chemotypes

  https://www.nature.com/articles/s41586-019-0917-9

- 0.1% hit rate (up to ~1%) in HTS, biased library

  Changing the HTS Paradigm: AI-Driven Iterative Screening for Hit Finding

  https://pmc.ncbi.nlm.nih.gov/articles/PMC7838329/



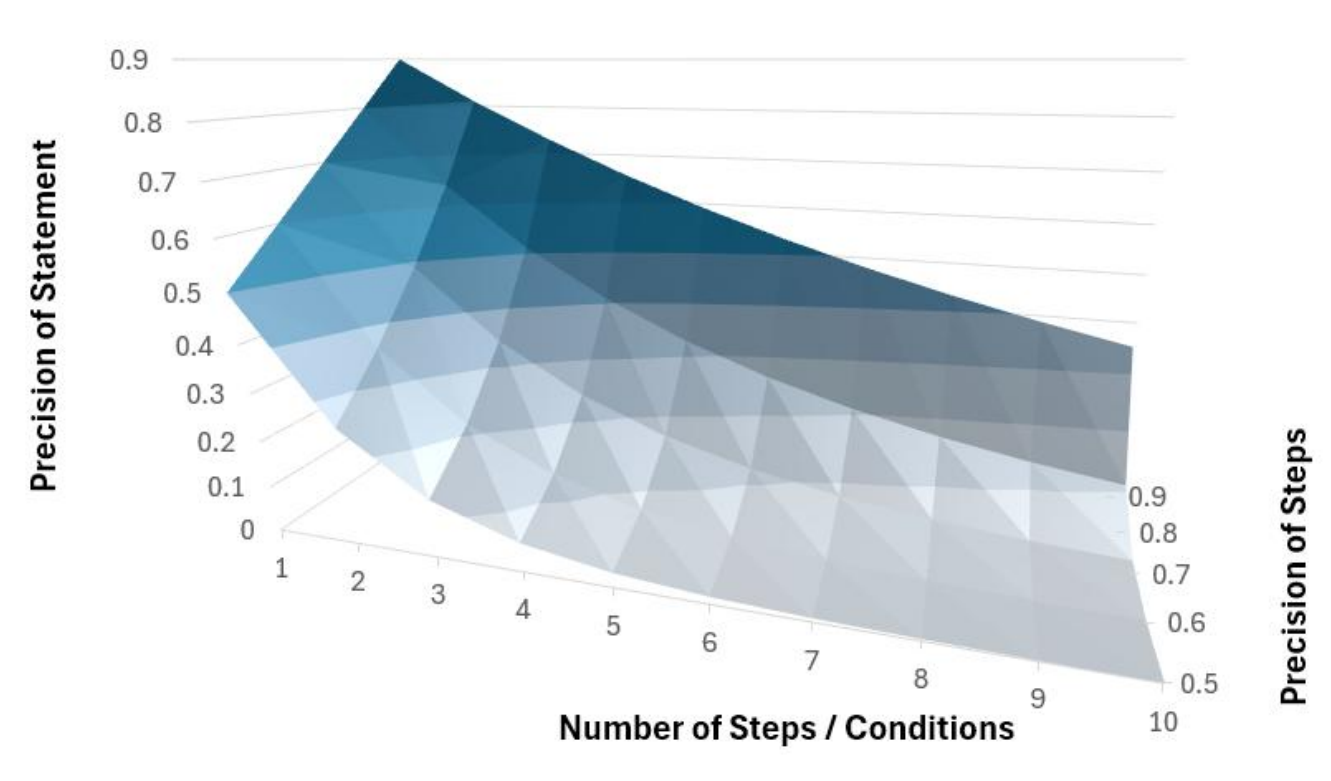
- Why does it work? Large space (~10^{60}), but *biased libraries*, *local precision* of model gives reasonable hit rate (as opposed to *global* model used in an *unbiased* library requiring *high accuracy*)
- Biases we can exploit in screening are in the order of 10^{dozens} (!!), say similarity to metabolites etc

More complex biological decision spaces have fewer *helpful* biases, say dosing a compound in vivo

- Ca 7,000 genetic diseases, approx. 20,000 total
- Say 20,000 genes/targetable entities
- 400m disease-gene pairs
- Pharmacology, dose, clinical endpoint, PK, more complex organismal biology, ...

- $>10^9$ gene-indication-dose-pharmacology combinations
- Genetics gives *2.6x* improvement (Minikel et al.), as opposed to 10^{dozens} (!)
- Far too small helpful bias to get lucky by chance

Quantifying dug discovery: Precision of statement as function of accuracy of steps/conditions



Precision of individual condition	Precision of 10 steps/ conditions
0.9	35%
0.8	10%
0.7	3%
0.6	0.6%
0.5	0.1%

- If drug discovery had only 10 conditions (*it has more*) and every condition has a precision of 0.7 overall precision is only 3%!
- Compare to clinical overall technical Probability of Success of 5-15% (Ph1->Approval) - we are doing quite well in this complex space!

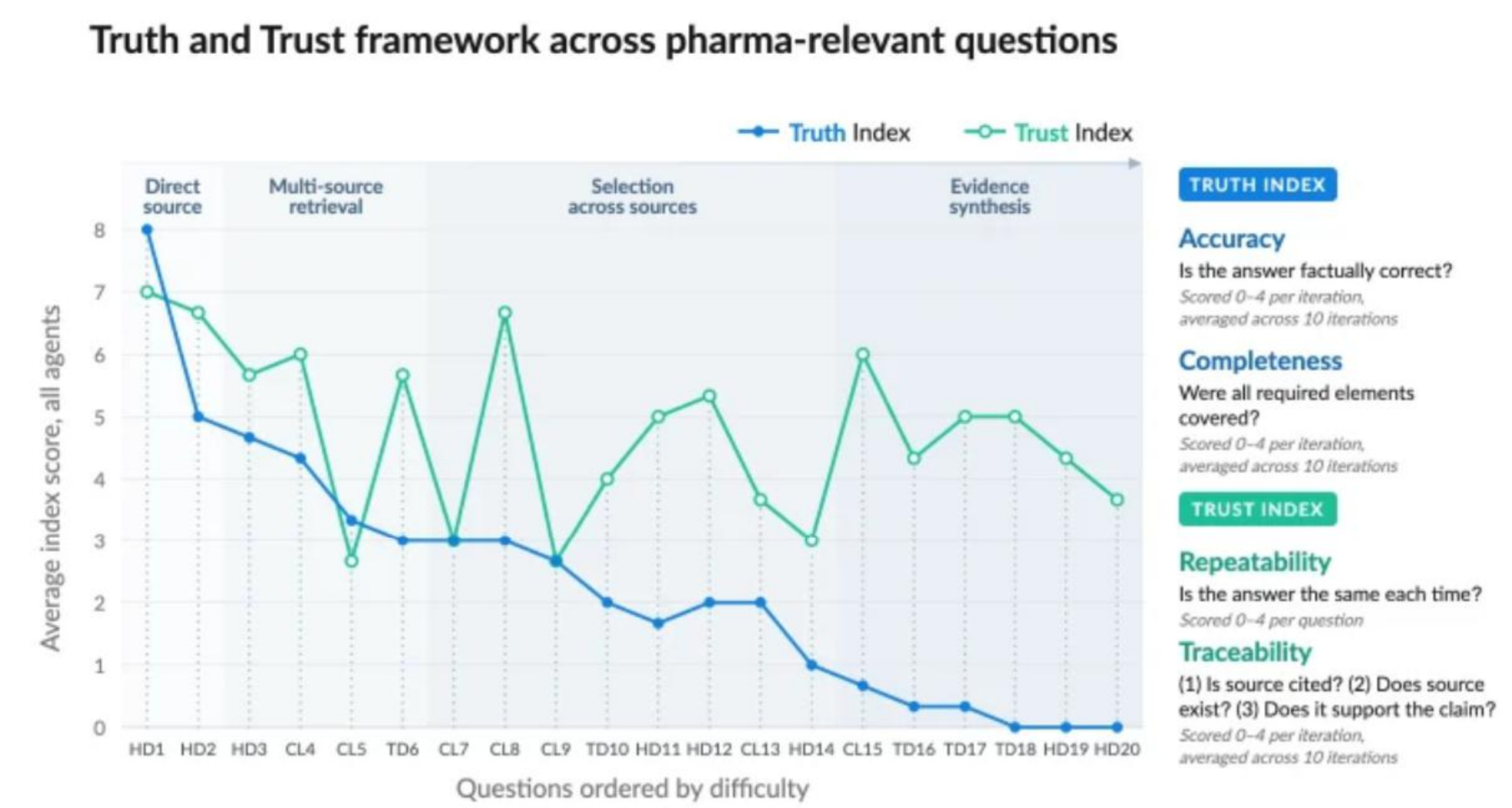
Pharma Doesn't Have an AI Problem. It Has an Information Problem.

Sponsored By INTELLIGENCIA AI · Aug 31, 2026 8:00am

E.g. global searches fail, without proper heuristics (*only data is not enough*)

Counterexample:

- Large hypothesis spaces
- Not constrained locally
- Not constrained globally



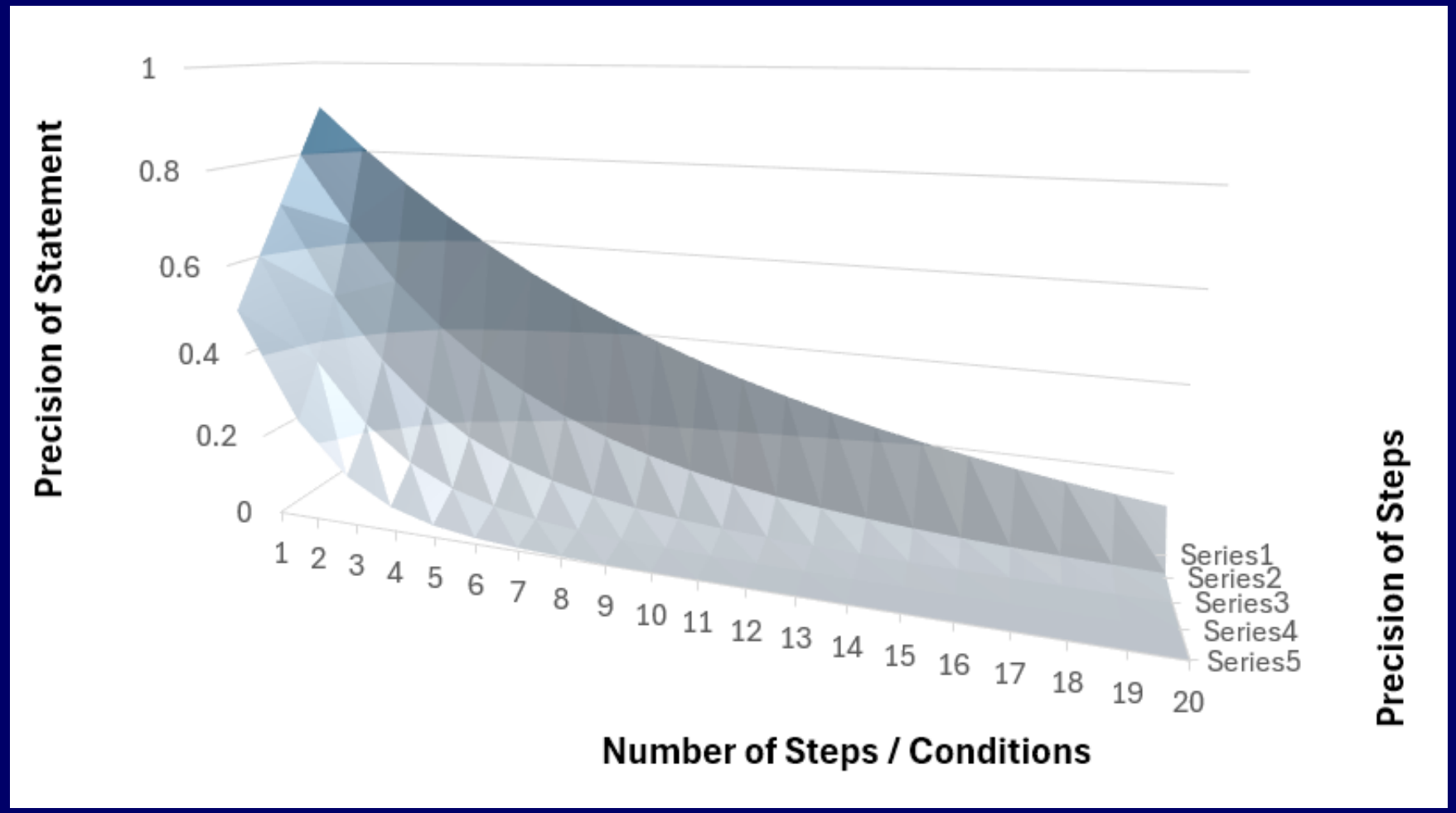
https://www.fiercepharma.com/sponsored/pharma-doesnt-have-ai-problem-it-has-information-problem

Moving towards more realistically quantifying drug discovery

- We need to do this de-risking in
  - **High-dimensional (‘100-1000s’ of dimensions)**
  - **Rugged (local)**
  - **Data-poor (sometimes only 100s of data points!)**
  - **Meta-data (context)-incomplete** *decision spaces*, which also generally have
  - **Low predictivity for *human translation*!**

We need to take care of: On-targets, Off-targets, Bioavailability, Solubility, Cmax, Half life, Microsomal stability, Accumulation in organelles, Accumulation in organs, Permeability, Target engagement, Organ Toxicity (Major Organs: Liver, Kidney, Lung, CNS, Cardiovascular) plus other ‘minor’ organs…

Slightly more realistic plot overall, say 20 conditions … at 0.5 precision *overall precision drops to ~10^{-6}*



Precision of individual condition	Precision of 20 steps/ conditions
0.9	0.1%
0.8	0.01%
0.7	0.001%
0.6	$4*10^{-5}$
0.5	$9.5*10^{-7}$

Additional problem: lack of reproducibility of (life science and other) research; lack of negative data

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

PLOS Medicine

nature

CAREER NEWS | 22 July 2025

Researchers value null results, but struggle to publish them

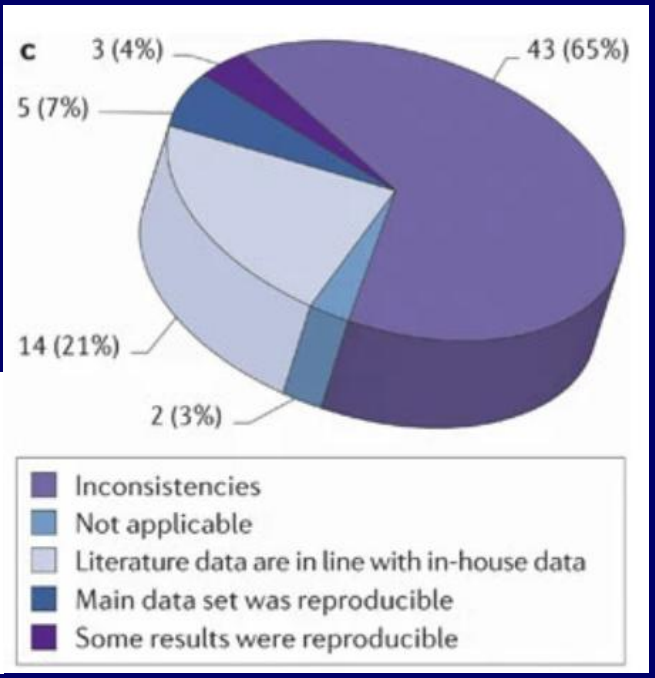
Survey finds that fear of reputational harm and a lack of support and publication platforms are among respondents' key concerns.

Computer Science > Digital Libraries

[Submitted on 4 Apr 2026]

LLMs Have Made Failure Worth Publishing

Sungmin Lee



Review Article | Published: 29 September 2026

Reproducibility in biomedical research

Judy Zhong, Debra D'Angelo, Chen Lyu, Ian Xu, Peter K. Enns, Fei Wang & Rainu Kaushal

Nature Medicine **32**, 3564–3577 (2026) | Cite this article

Correspondence | Published: 31 August 2011

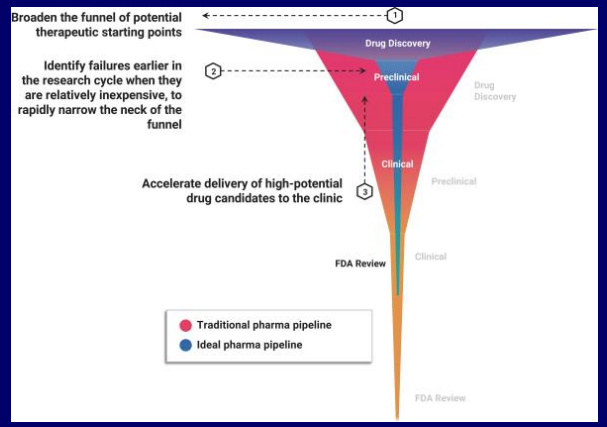
Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange & Khusru Asadullah

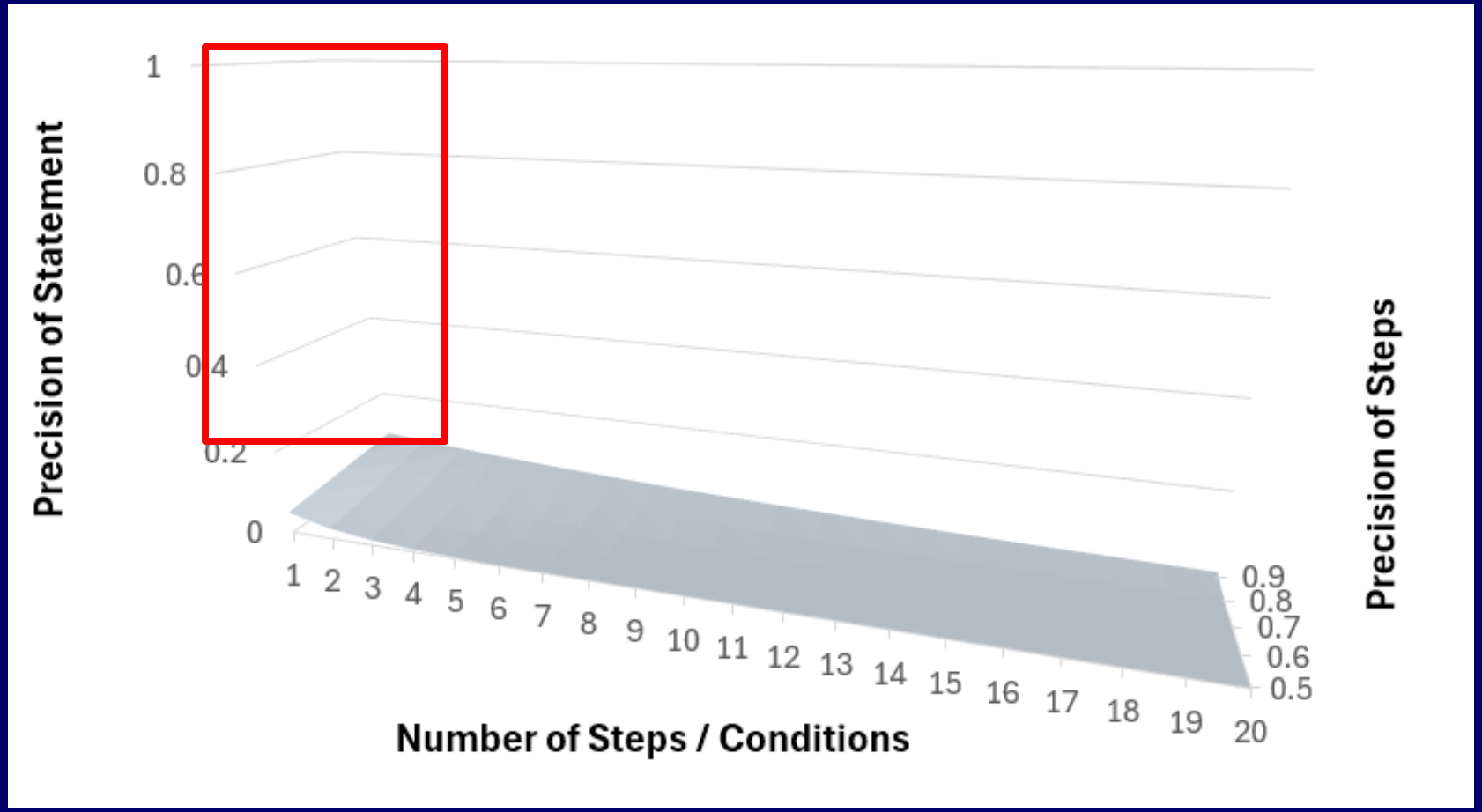
Nature Reviews Drug Discovery **10**, 712 (2011) | Cite this article

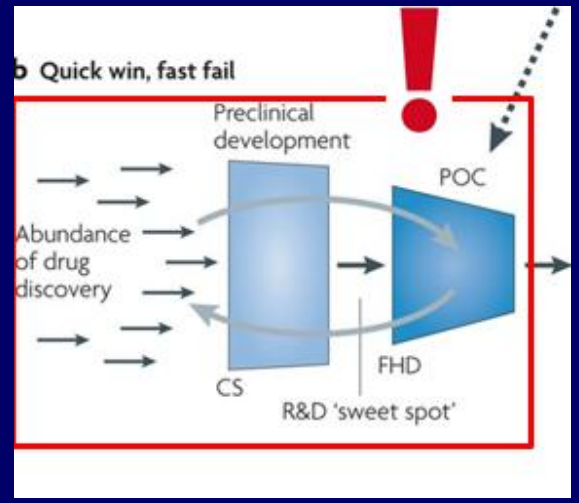
Problems: False positives, *and* lack of true negatives

- For de-risking *precision* of statements is key
- Small sample sizes, biases, multiple hypothesis testing etc increase *number of false positives* in statements
- Publication bias *decreases number of true negative statements*
- Estimates:
  - **False positives** ~50% (Ioannidis PLoS Medicine 2005) – 71% (Prinz et al. NRDD 2011)
  - Publication bias ~80% of negative results are not published (Fanelli, PLoS One 2010), **lack of true negatives**



Taking false positives, lack of true negatives into account… even in low-dimensional spaces precision drops rapidly!





Precision of individual condition	Precision of 10 steps/ conditions
0.09	0.01%
0.08	0.001%
0.07	$8*10^{-5}$
0.06	$4*10^{-6}$
0.05	$9.5*10^{-8}$

- Even precision of statements involving quite few conditions drops
- Relevant for low-dimensional statements (repurposing etc.)

Plus technical problems: LLMs *necessarily* hallucinate; are currently overconfident

- Matters more for some tasks than others
- Generally has detrimental effect on precision of statements though

Hallucination is Inevitable:
An Innate Limitation of Large Language Models

Ziwei Xu **Sanjay Jain** **Mohan Kankanhalli**
School of Computing, National University of Singapore
ziwei.xu@u.nus.edu {sanjay,mohan}@comp.nus.edu.sg

Statistics > Methodology

[Submitted on 30 Oct 2025]

LLMs are Overconfident: Evaluating Confidence Interval Calibration with FermiEval

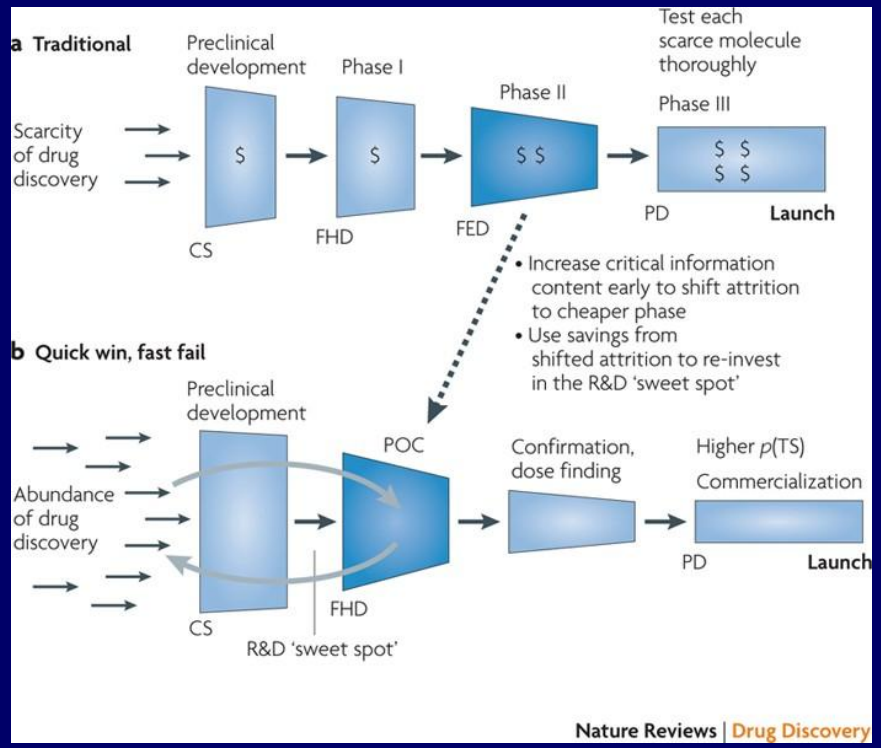
Elliot L. Epstein, John Winnicki, Thanawat Sornwanee, Rajat Dwaraknath

Bottom line:
Automated decision making will *not* be a piece of cake

We operate in

- **High-dimensional**
- **Rugged (local)**
- **Data-poor**
- **Meta-data (context)-incomplete** *decision spaces*, which also generally have
- **Low translational predictivity**

- We need precision (to de-risk) *BUT*
- We have (apart from the above problems!) too many false positives, and too few true negatives to do so
- Leads to lack of precision in de-risking, and arriving at true positive statements

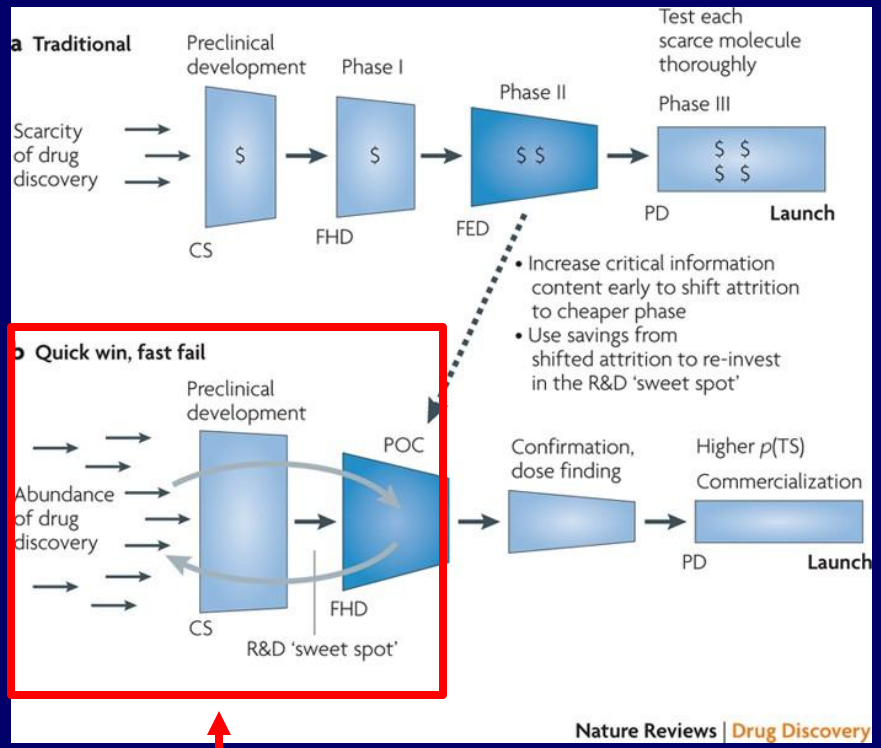


Bottom line:
Automated decision making will *not* be a piece of cake

We operate in

- **High-dimensional**
- **Rugged (local)**
- **Data-poor**
- **Meta-data (context)-incomplete** *decision spaces*, which also generally have
- **Low translational predictivity**

- We need precision (to de-risk) *BUT*
- We have (apart from the above problems!) too many false positives, and too few true negatives to do so
- Leads to lack of precision in de-risking, and arriving at true positive statements



Our data often does not allow us to 'fail fast'!

Three types of models can support decision making (both by humans and agents)

Three types of models can support decision making (both by humans and agents)

- Defined search space, preferably low-dimensional, **smooth surface**, filled with data
  - logD prediction

Three types of models can support decision making (both by humans and agents)

- Defined search space, preferably low-dimensional, **smooth surface**, filled with data
  - logD prediction

- Local searches, supported by heuristics (yes, novelty trade-off, but otherwise we have insufficient precision)
  - Virtual screening – good hit rates in *biased libraries (heuristic)*
  - *Local* precision

Three types of models can support decision making (both by humans and agents)

- Defined search space, preferably low-dimensional, **smooth surface**, filled with data
  - logD prediction

- Local searches, supported by heuristics (yes, novelty trade-off, but otherwise we have insufficient precision)
  - Virtual screening – good hit rates in *biased libraries (heuristic)*
  - *Local* precision

- Global, meta-level heuristics (more tricky case since global)
  - E.g. ‘5Rs’ – right drug, patient, tissue, target, dose

5. Classification of decision spaces

- Three categories of steps in workflows, and in agents
  - Input
  - Process
  - Decision making

- Where different metrics matter, where hypothesis spaces are of different sizes

Andy's Agent Assessment Attempt (AAAA)

Task	Input	Process	Decision	
Primary Metric	Recall	Accuracy	Precision	
Size of Space	Medium	Low(-ish), defined	High	
Metric achieved	Can be high	High (if explicit)	*In vivo*: Low (Low predictivity of labels, few heuristics available)	E.g. in HTS: Can be high (high predictivity of labels, good heuristics)

Combination of relevant metric and characteristic of search space shows very different fit-for-purposeness of agents in different situations

Spaces are too large to be filled with data!

- Also (and in particular!) ‘spatially and time-resolved multi-modal multi-omics single cell (etc)’ data

- Hypothesis spaces
  - 10^{60} small molecules *but helpful heuristics and biases*
  - $>10^{9}$ target-disease-pharmacology-dose combinations (*in vivo*, not even considering more complex PK and safety) with *fewer helpful heuristics and biases*
  - 1g of tumor contains 10^{9} cells, each cell 10^{6} proteins, different states, interactions, function of time, different in individuals, …

6. What now? Merging data and knowledge

‘Faster’ and ‘cheaper’ will not be sufficient – we need to be *better* how we explore hypothesis spaces:

- Constrain functional form (fewer free parameters)
- Constrain dimensionality
- Local (instead of global) searches
- Use heuristics/biases
- ….

- E.g. using pathways, start with known drug, have a solid mode of action hypothesis …

What did the old masters do right?
Sir James Black went *right to the matter of things*

James Black searches for a new type of histamine antagonist

Initially, three bioassay systems were selected by Jim. For in vitro assays, he chose the guinea-pig ileal muscle, which was the classical system for studying antihistamines such as mepyramine (now, of course, designated as an H_1 receptor system), and guinea-pig atrial tissue for the mepyramine-refractory histamine responses (where histamine increases the rate of beating of the right atrium). For gastric acid secretion there were no in vitro assays available and he selected the Ghosh and Schild procedure [5] (see below) in vivo. Subsequently, another in vitro assay was established using the rat uterus, in which contractions were evoked electrically, at one minute intervals, in a piece of rat uterine horn.

… while we now prefer to dance around the problem a bit (tech, data, throughput, etc)

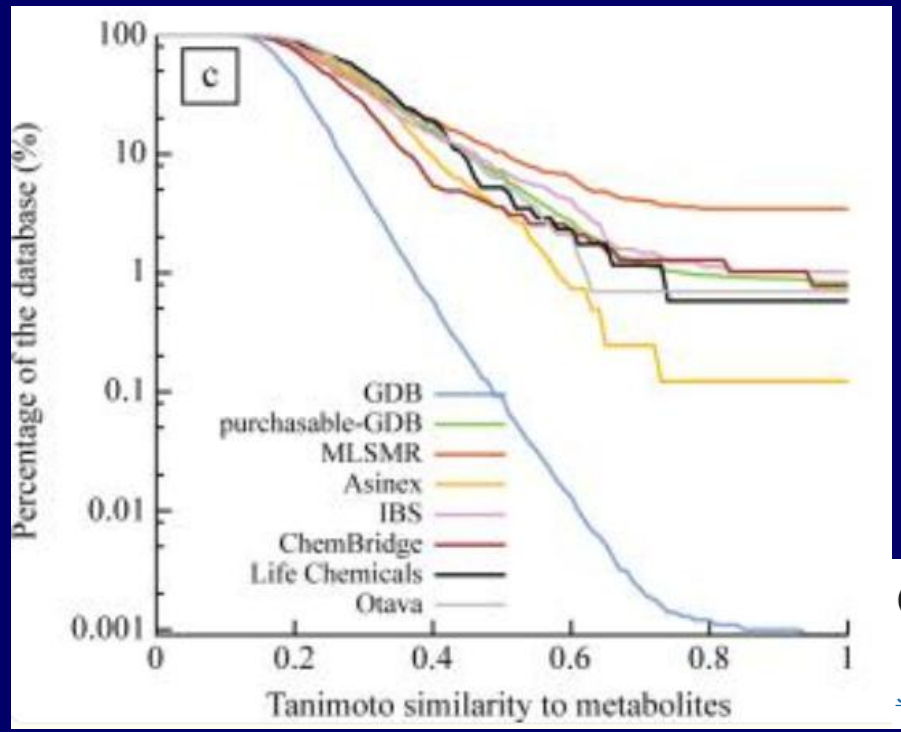
Personal reflections on Sir James Black (1924–2010) and histamine.
https://link.springer.com/article/10.1007/s00011-010-0269-2

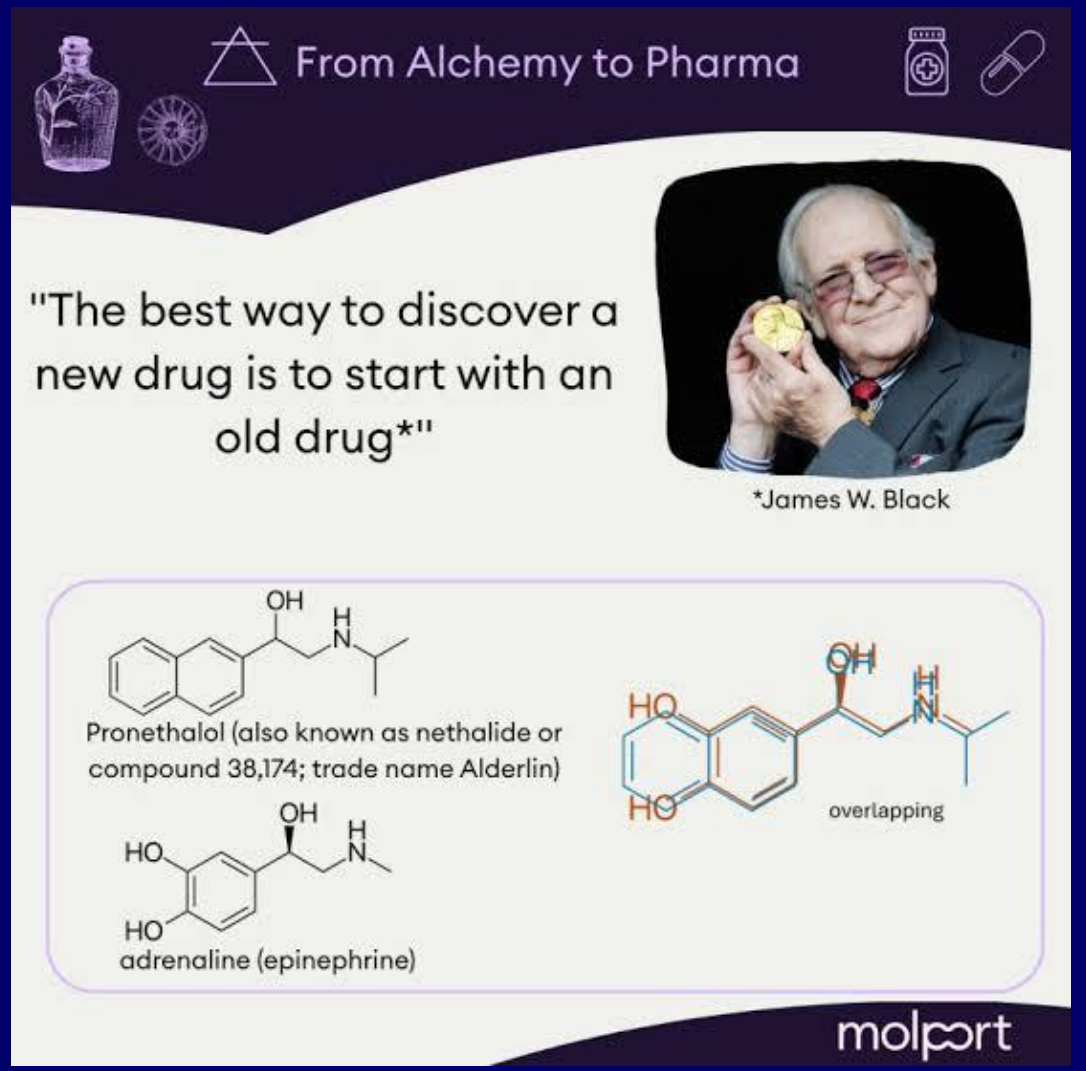
Biases and heuristics abound… *biases are good*!

J Med Chem. 2004 Feb 12;47(4):888-99. doi: 10.1021/jm0309452.

Recognition of privileged structures by G-protein coupled receptors

Kent Bondensgaard [1], Michael Ankersen, Henning Thøgersen, Birgit S Hansen, Birgitte S Wulff, Robert P Bywater





Quantifying Biogenic Bias in Screening Libraries

Jérôme Hert [1], John J Irwin [1], Christian Laggner [1], Michael J Keiser [1], Brian K Shoichet [1,*]

What creates value? Balancing economic benefit of decisions with ability to do so with data

Automated discovery *may work* in

- *Low*-dimensional spaces
- Data-***rich*** spaces
- Spaces where we can define (and evaluate) an ***objective function***

E.g. hit finding, repurposing where efficacy models are available, materials discovery, …

Economic value quite possibly comes more from

- De-risked and lower-dimensional spaces
- Where data is available

Say, good and fast *fast*-followers

What matters most in commercial success: first-in-class or best-in-class?

Ulrik Schulze & Michael Ringel

Nature Reviews Drug Discovery **12**, 419–420 (2013) | Cite this article

Figure 1: Value captured as a function of time of market entry and therapeutic advantage.

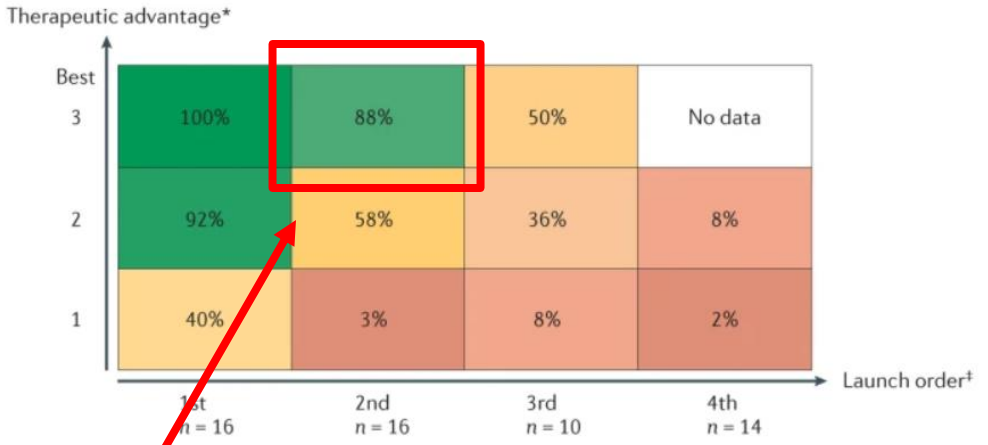
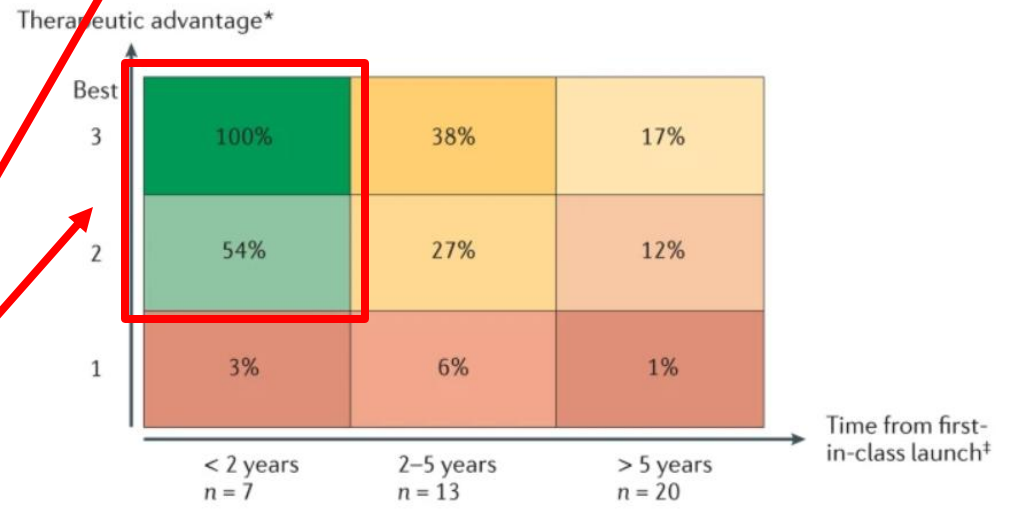


Figure 2: Value captured by follow-on entrants as a function of time of market entry and therapeutic advantage.



It is up to us what the future looks like

Data and understanding can lead us here!



Tech first and ‘data without a cause’ leads us here

Good	7 choices to solve a problem	Not so good
Sir James Black wants to find antihistamines	**Define problem in search and endpoint space**	‘We just generate data’, ‘we have a new tool’ etc
Sir James black starts with histamine	**Heuristics to make hypothesis space smaller**	Brute-force, high-throughput
Picks right assays (in vivo, predictive validity)	**Experimental setup with predictive signal**	‘Cheap and fast’
Makes analogues based on understanding *and* data	**Technically appropriate enumeration of search space**	Brute-force won’t win against numbers
(After trial and error) detects positive examples, w/ control	**Ability to detect positive signal**	Cheap doesn’t translate
(Assay is translational)	**Validate translational proof point**	Failure in the clinic
Scale-up etc	**Productionize**	Project dead





Take-home message

1. Drug ‘discovery’ operates in an *abundance setting*, and *de-risking* is what matters
2. We generally operate in data-poor environments (where ‘data poverty’ has many aspects; predictive value, dimensionality, *etc*)
3. Size of hypothesis space, predictivity of data for translationally relevant decisions, matter for process impact
5. Agents are (broadly speaking) good for input and process, less so for (high-dimensional, *data- and heuristics-poor*) decisions
6. Combination of data and knowledge, and aligning understanding, data, model, use case, and real life application are key to get anywhere

Slides: www.drugdiscovery.net - Contact: andreas@bio.bi