Recursion

Thanks, I hate it a little less!

(an update from the world of machine learning)

Imran S. Haque, PhD

Twitter:	*@ImranSHaque*
Mastodon:	*@ihaque@genomic.social*

CUP XXII

2023-03-16

Takeaways before you tune out

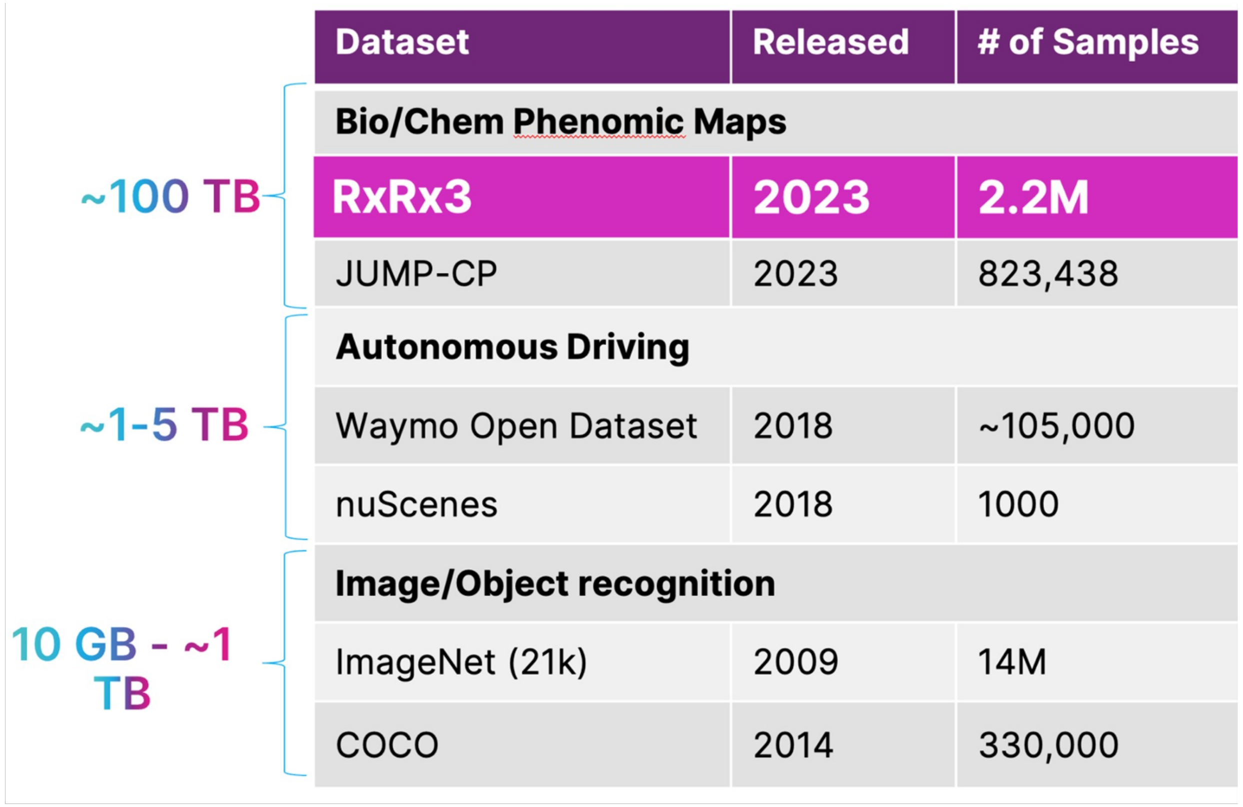
- What makes (bio)chemistry a hard ML domain?
- Advances in self-supervision and modeling improving chemistry **representations**.
- High-dimensional data from biology improving dense chemical **data** for ML.

RxRx3: Leading the field in open science at rxrx.ai/rxrx3

Genetics and chemistry: two great flavors that go great together

RxRx3:

- Images, metadata, and DL embeddings of knockouts of ~17K genes* + ~1700 SMs @ multiple concentrations .
- The largest publicly-released data set of perturbative cellular imaging, all generated at a single site with a consistent protocol.
- Genetic and chemical perturbations in a shared embedding space enabling inference of mechanism and phenotype.



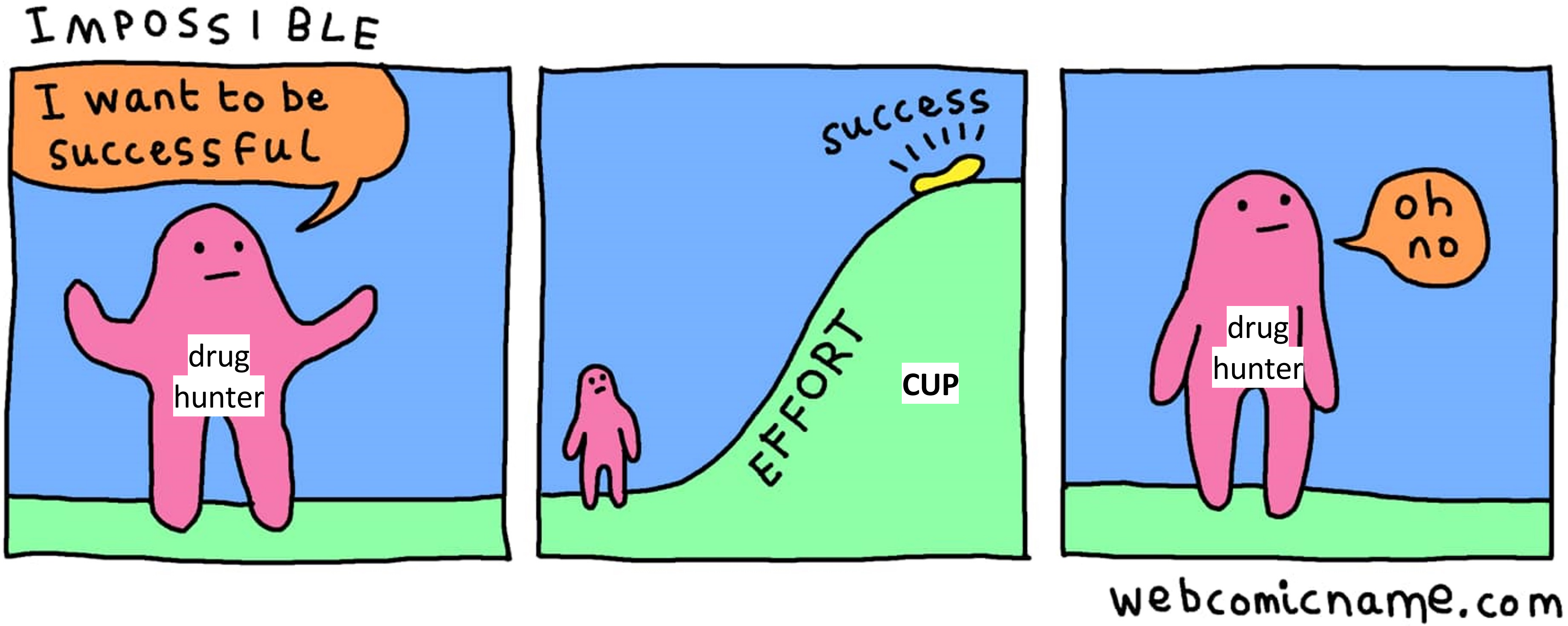
Dataset	Released	# of Samples
Bio/Chem Phenomic Maps		
RxRx3	**2023**	**2.2M**
JUMP-CP	2023	823,438
Autonomous Driving		
Waymo Open Dataset	2018	~105,000
nuScenes	2018	1000
Image/Object recognition		
ImageNet (21k)	2009	14M
COCO	2014	330,000

~100 TB
~1-5 TB
10 GB - ~1 TB

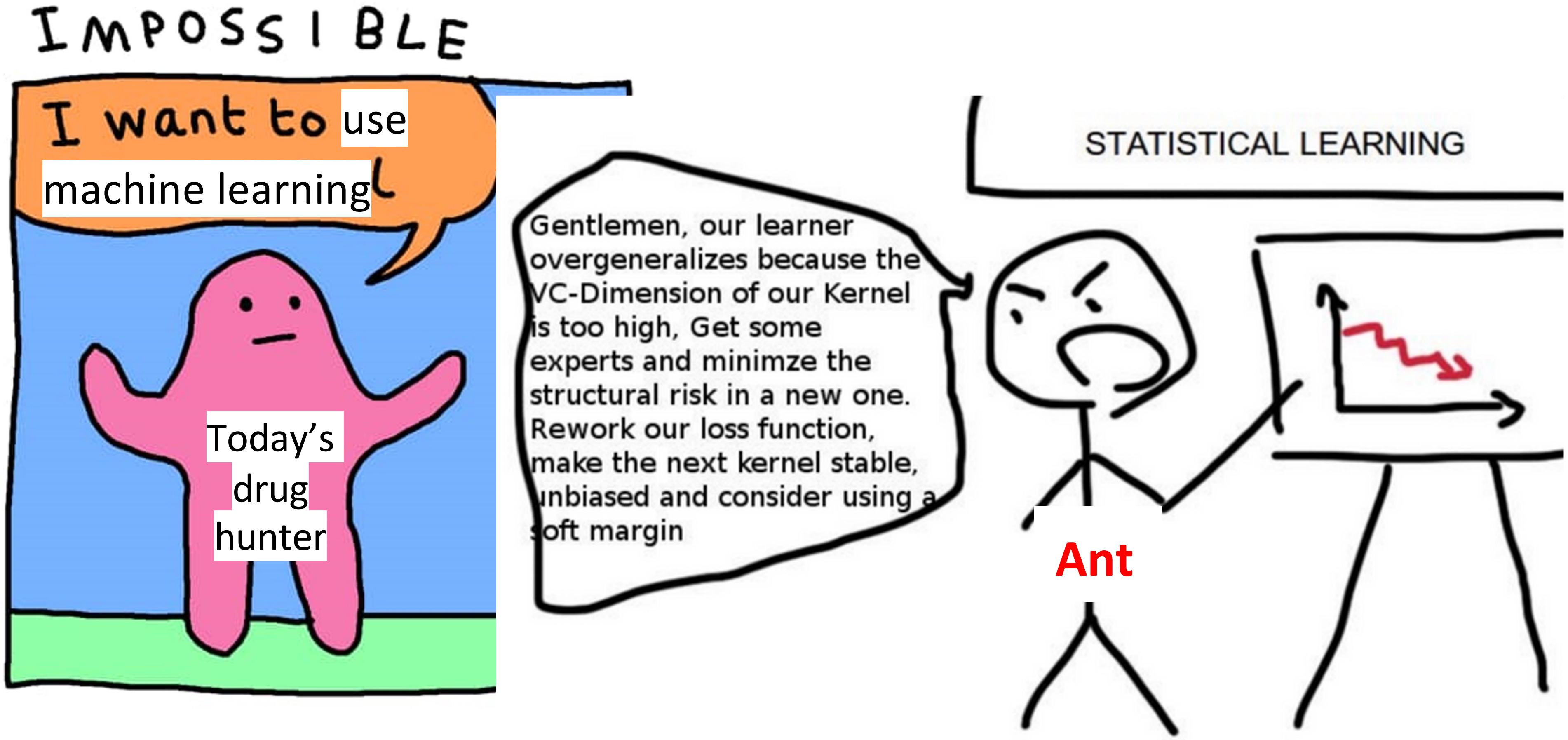
* **Mostly blinded, but let's talk if you're interested…**

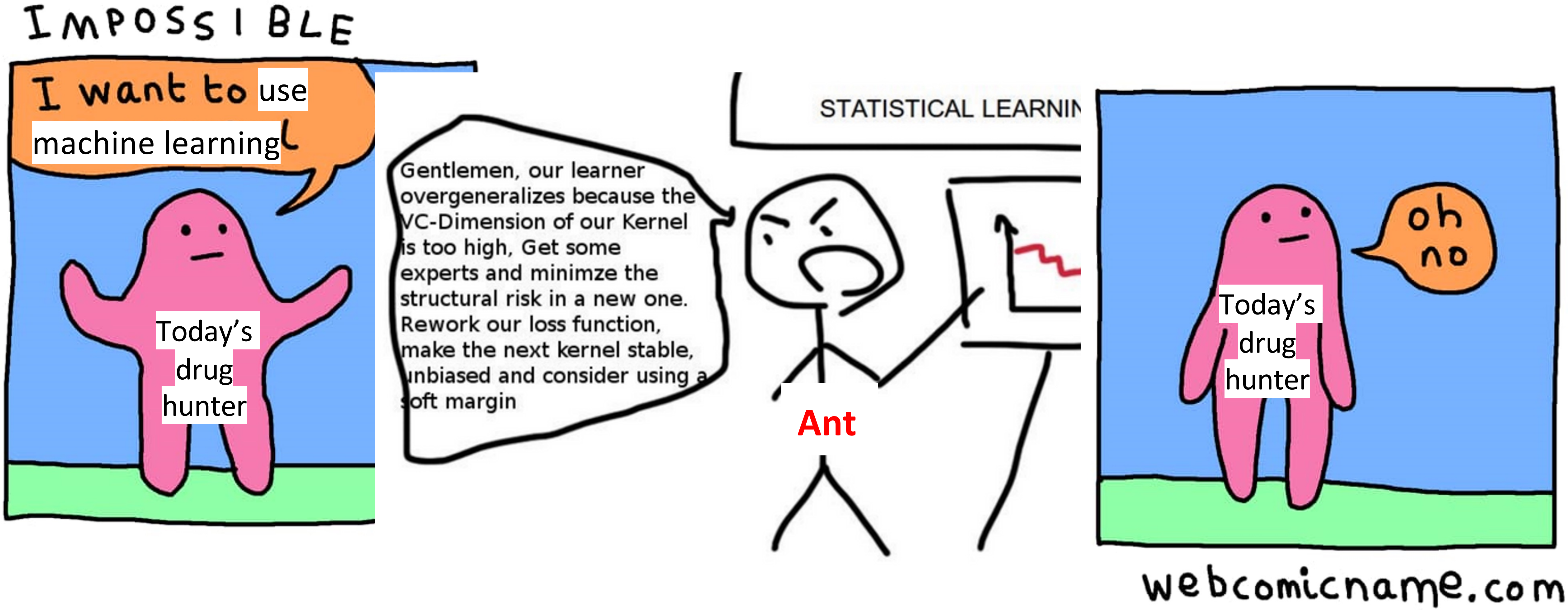
Let’s turn the clock back to 2019.

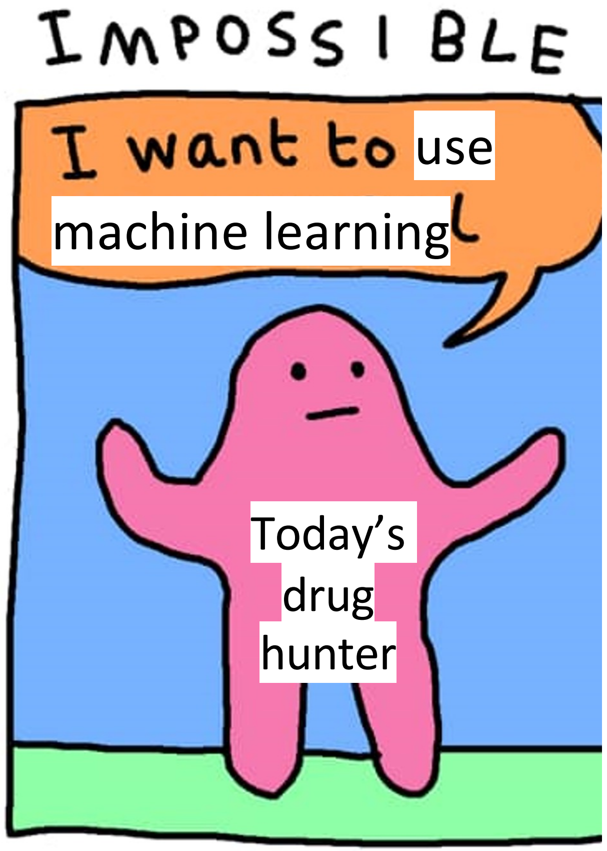
The last time I spoke at CUP was in 2019, explaining
why ML in chemistry (and biology) is challenging…

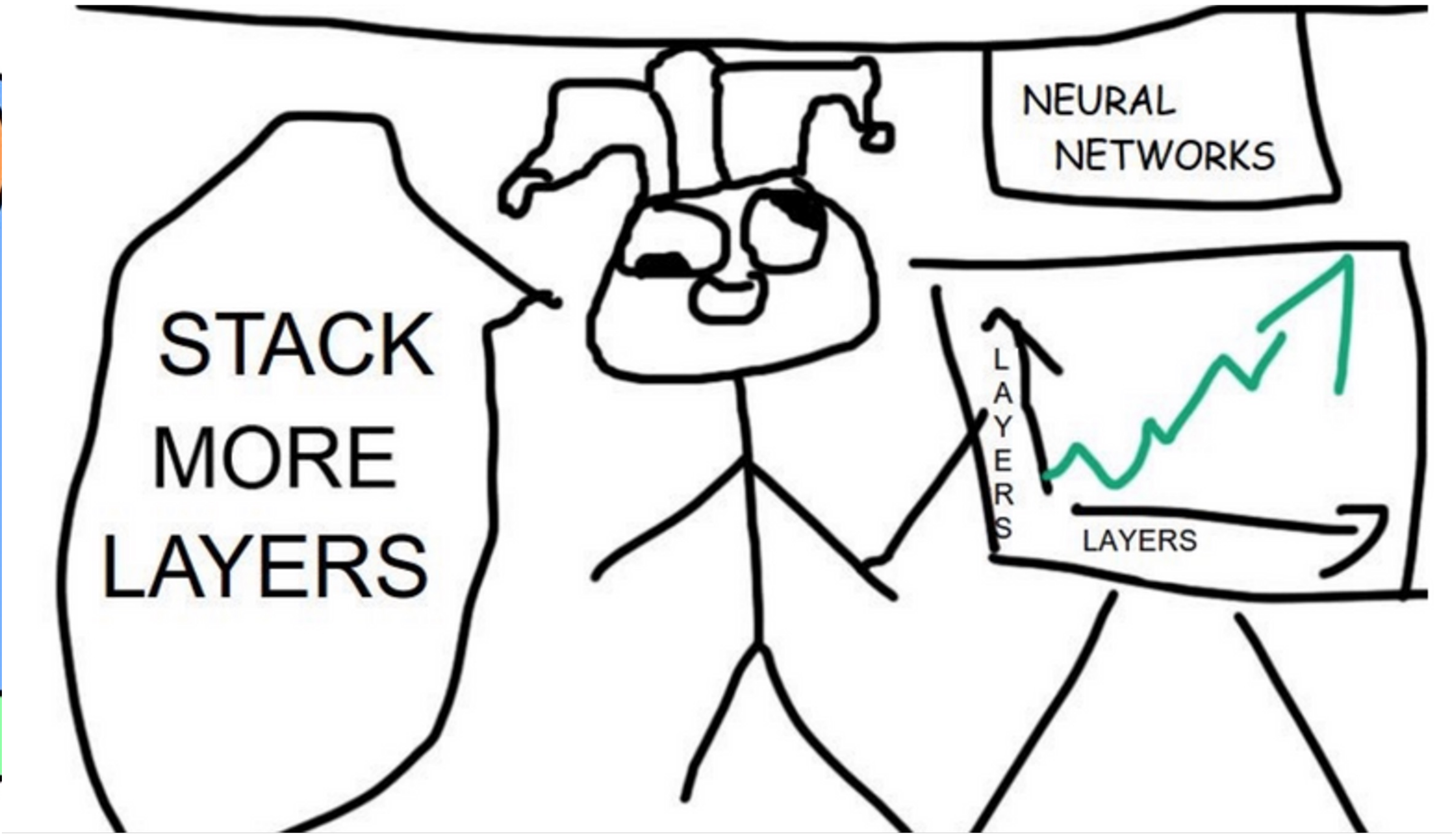

IMPOSSIBLE
I want to be successful
drug hunter
success
EFFORT
CUP
oh no
drug hunter
webcomicname.com

Recursion


IMPOSSIBLE
I want to use machine learning
Today's drug hunter
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high, Get some experts and minimze the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin
STATISTICAL LEARNING
Ant


IMPOSSIBLE
I want to use machine learning
Today's drug hunter
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high, Get some experts and minimze the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin
STATISTICAL LEARNIN
Ant
oh no
Today's drug hunter
webcomicname.com


IMPOSSIBLE
I want to use machine learning
Today's drug hunter


STACK MORE LAYERS
NEURAL NETWORKS
LAYERS
LAYERS

IMPOSSIBLE
I want to use machine learning
Today's drug hunter
STACK MORE LAYERS
OH, YEAAHH!
Today's drug hunter

YOU WILL NOT GO TO SPACE TODAY

n

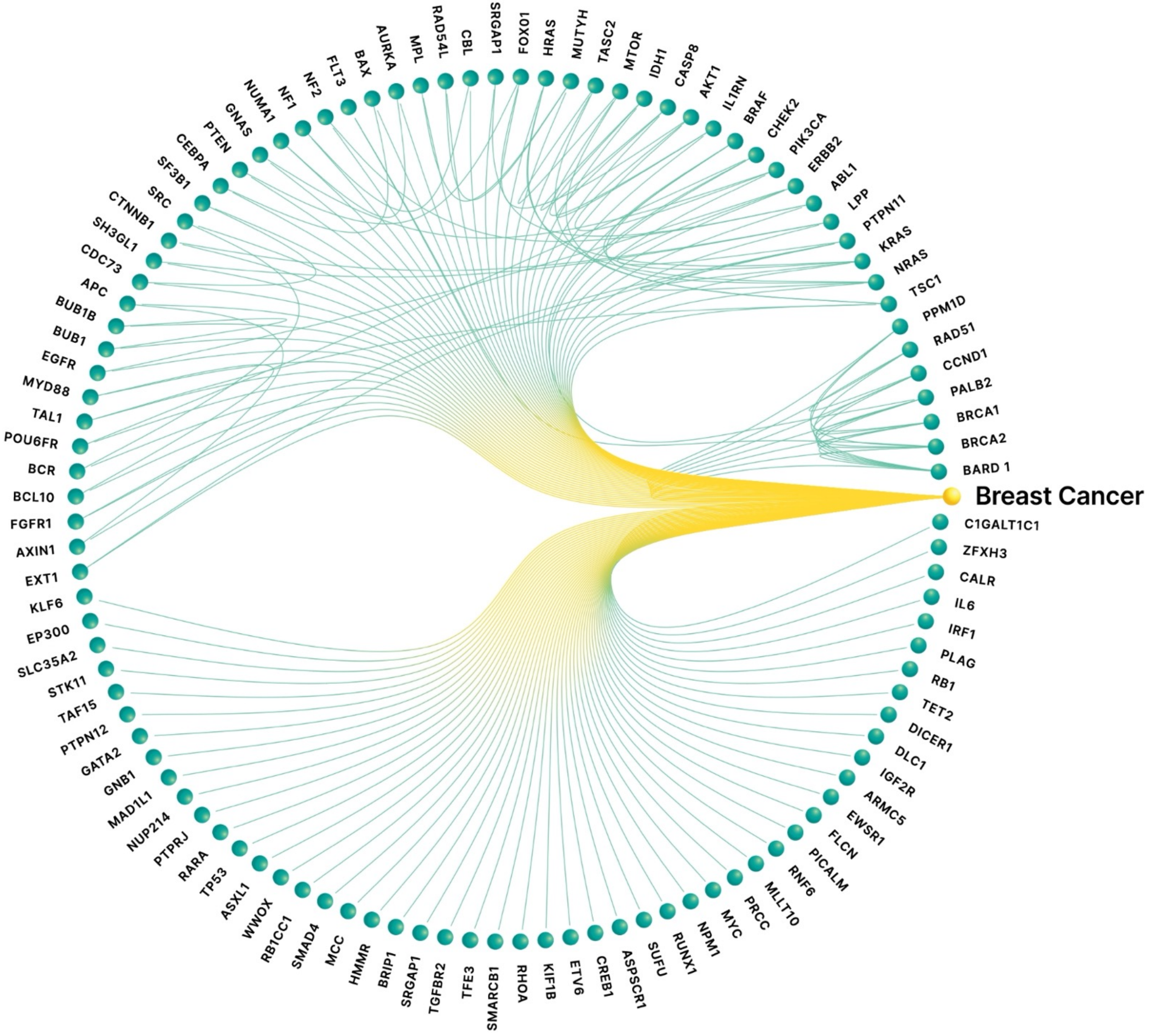
Why is ML in chemistry difficult?

- All we want in ML is, given **x** and **y** input pairs, to identify a function **f** such that **f**(**x**) is “usually close to” **y.**
- Tricks in machine learning are usually of the form:
  - *Find a new family of **f** that is predictive.*
  - *Find a new optimization algorithm over **f***
  - *Get “better” forms of **x** (ie, ones more suited to your **f**)*
  - *Get more (**x**,**y**) pairs.*
- Chemistry ML is hard because:
  - it is hard to define chemical representations that are highly informative and suited for particular predictive functions.
  - The available data are sparse (over molecules, and over targets).

Could we solve the sparse-data-matrix problem in chemistry by taking cells, knocking out each gene in the genome, and individually doing 100,000s of molecules at multiple concentrations, and taking some pictures?

…no, really. What if we did?

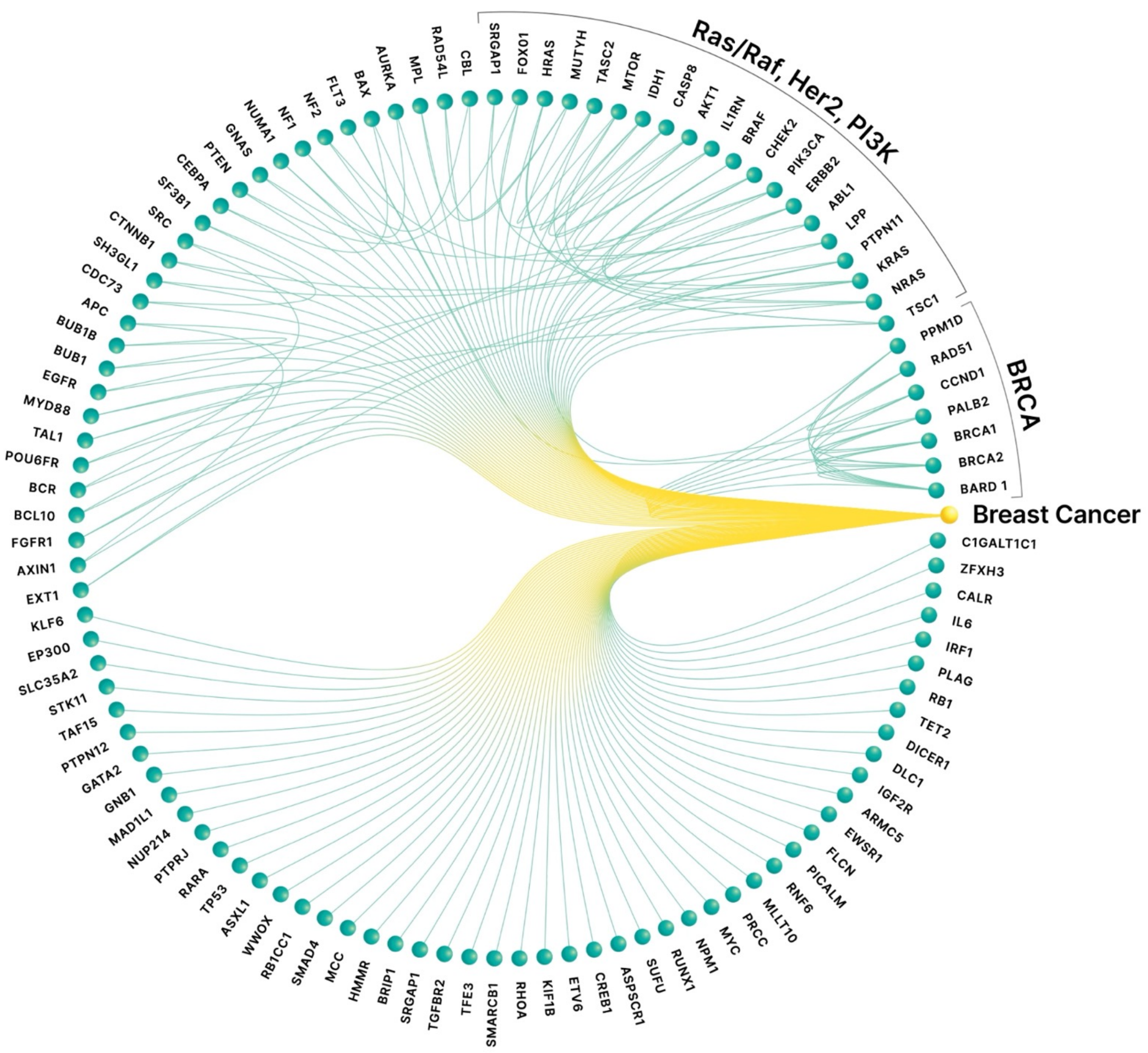
At Recursion, we have individually knocked out (almost) every gene in the genome in primary human cells, imaged them, and transformed those images through proprietary deep learning models to put them in a “relatable”, biologically meaningful space in which we can identify similar or dissimilar perturbations. This is what we call a “map of biology”.



In this illustration, I’ve queried the Human Phenotype Ontology for genes related to “breast cancer”, giving us the genes in green arrayed around the circle. Yellow edges represent those associations in HPO.

Green edges between genes represent significant similarity associations that we observe between pairs of genes in our maps.

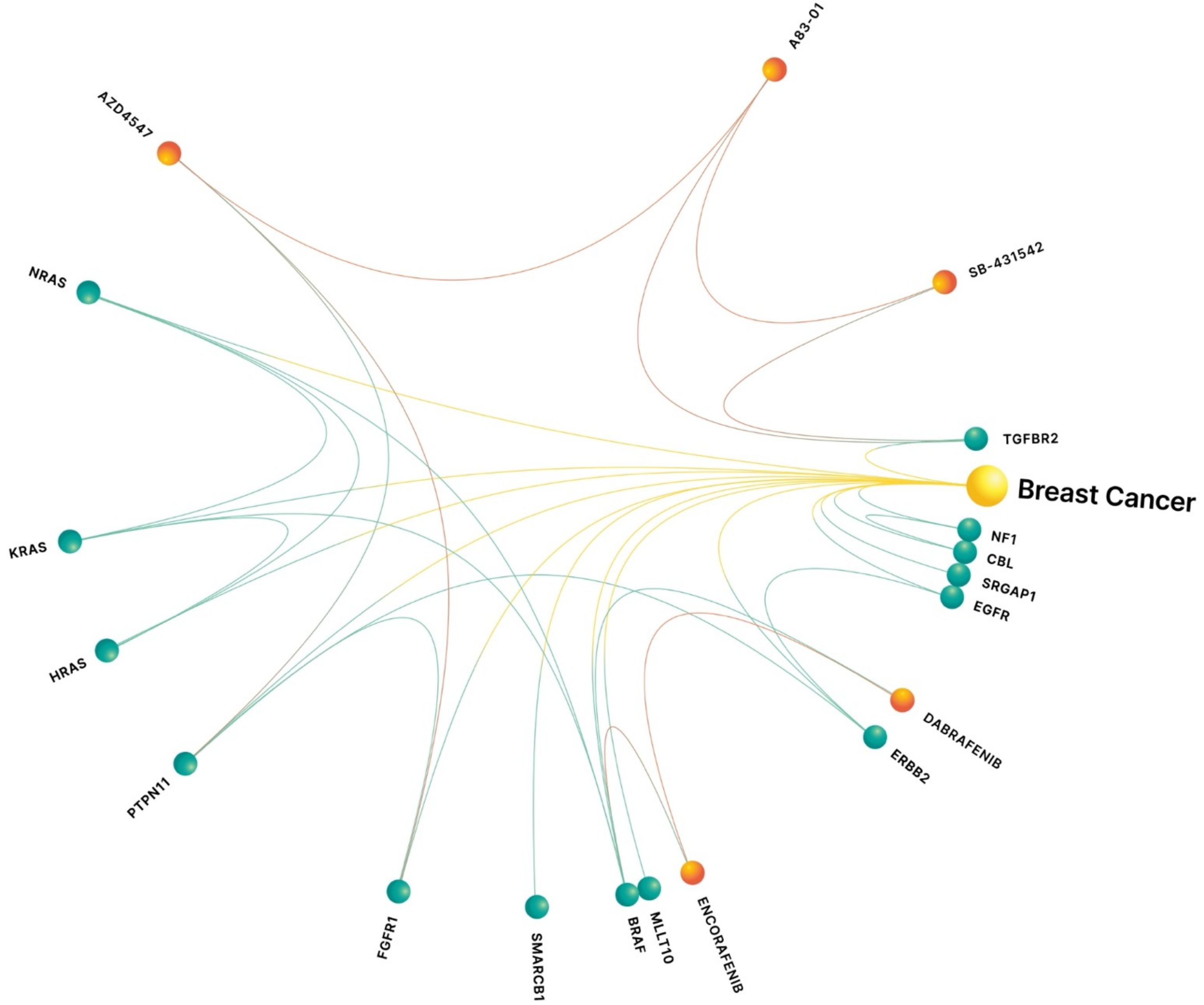
The associations seen in Recursion’s maps immediately recapitulate decades of cancer biology, with the genes in the BRCA complex immediately visible as a cluster, as well as the Ras/Raf, Her2, and PI3K pathways.



Crucially, **the models processing images to derive embeddings for maps, and therefore the green edges between genes, have no prior information about what those genes are, or anything in the literature.** These edges are purely derived from large-scale experimental data and recapitulate huge amounts of ground truth biology.

Recursion

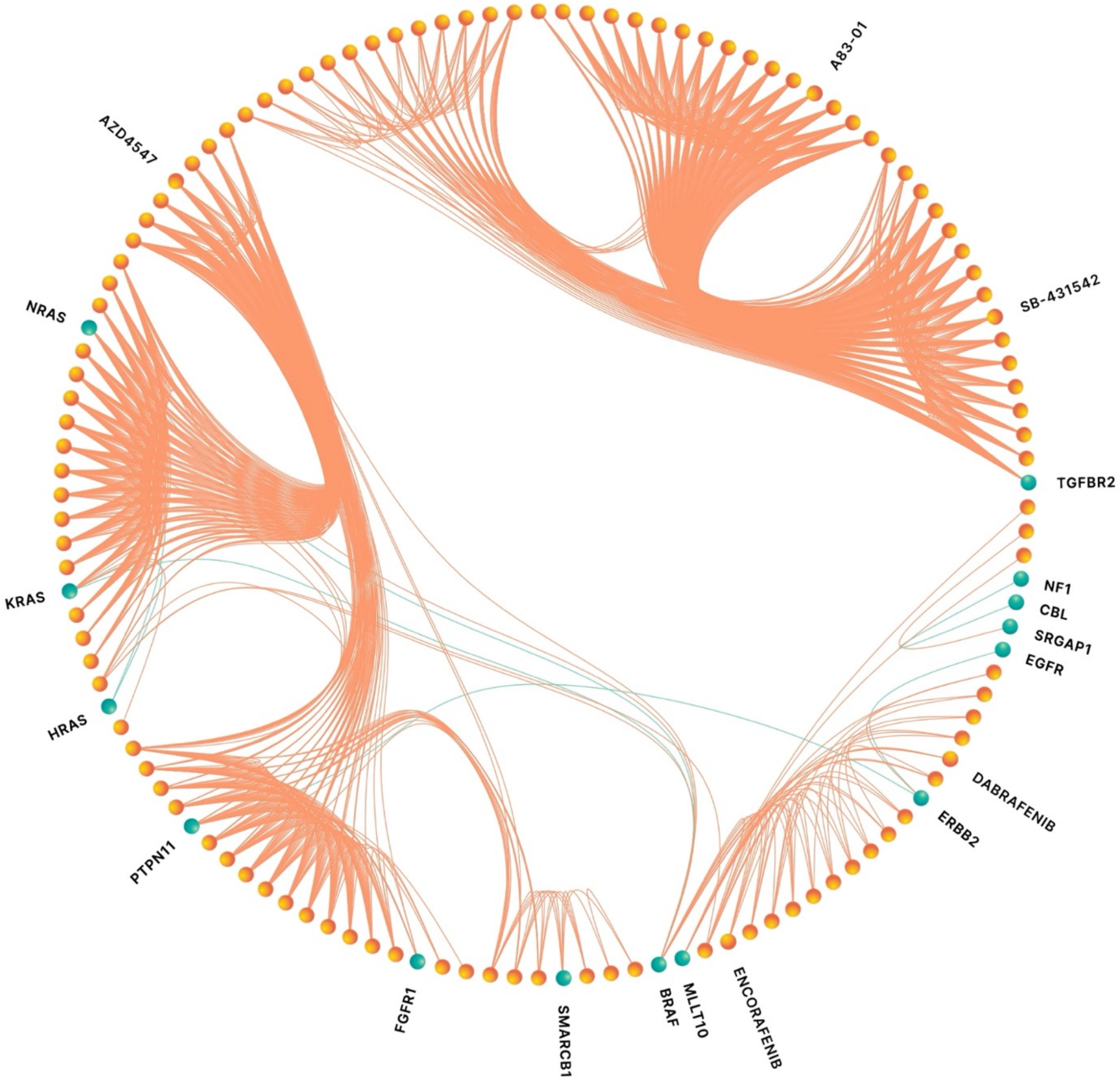
An important feature of the Recursion platform and our maps is that we can profile not just genetic knockouts, but also pharmacological perturbations (small and large molecules) in concentration-response.



Here, I've zoomed into the Ras/Raf cluster, and added orange nodes representing ground truth small molecule inhibitors of a number of genes in the pathway. We see the relationships that we would expect to see show up – inhibitors whose effects look phenotypically similar to knockout of their corresponding genes – giving us confidence that we can relate genetic and chemical perturbations to one another.

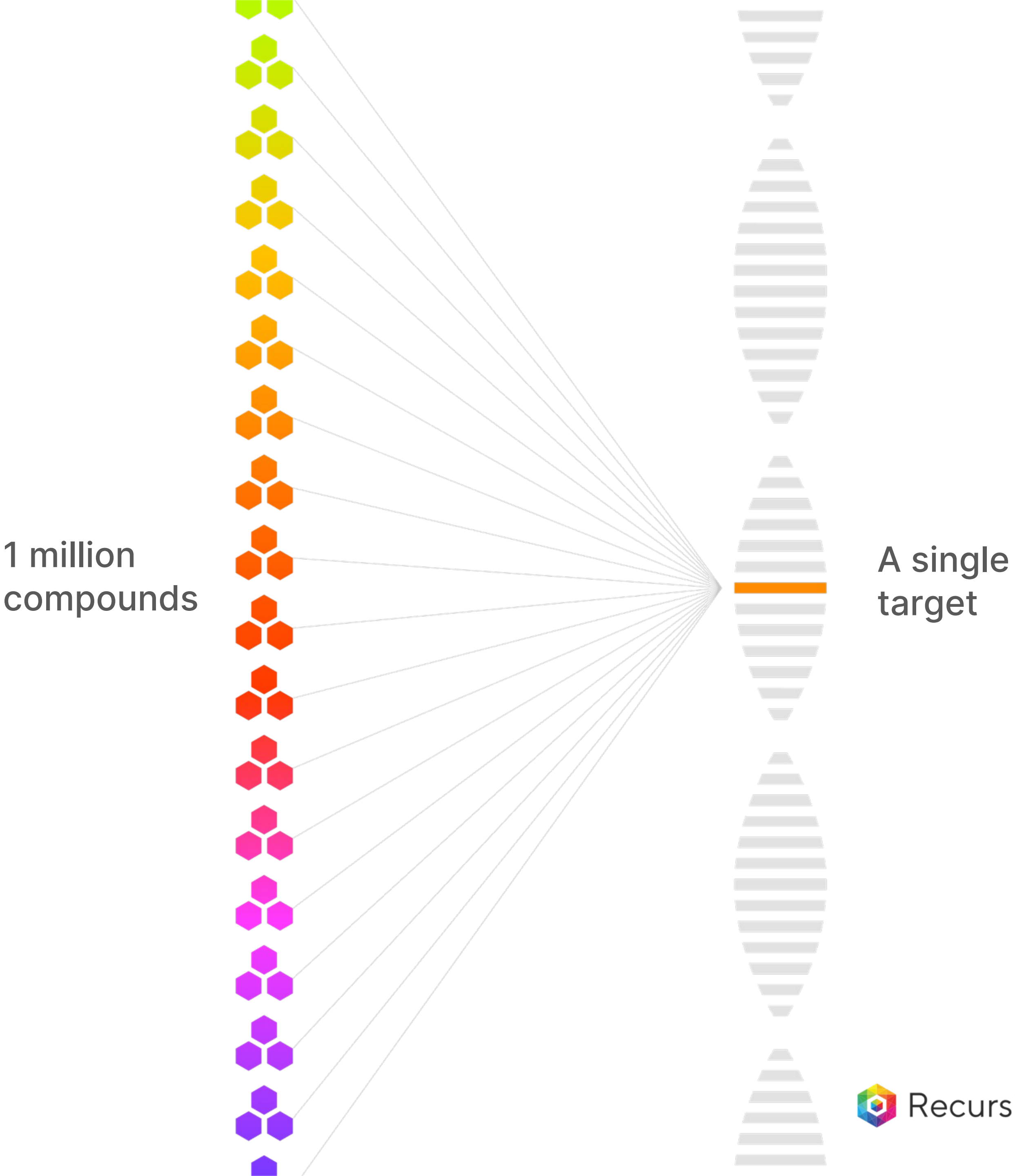
Recursion

Of course, the platform would not be so interesting if all it could do was recapitulate known biology. Here, I've added in a number of unlabeled orange nodes corresponding to NCE starting points from our screening library – demonstrating that even in the absence of known ground truth, the Recursion "mapping and navigating" approach can find starting material to initiate programs against a variety of interesting potential targets.



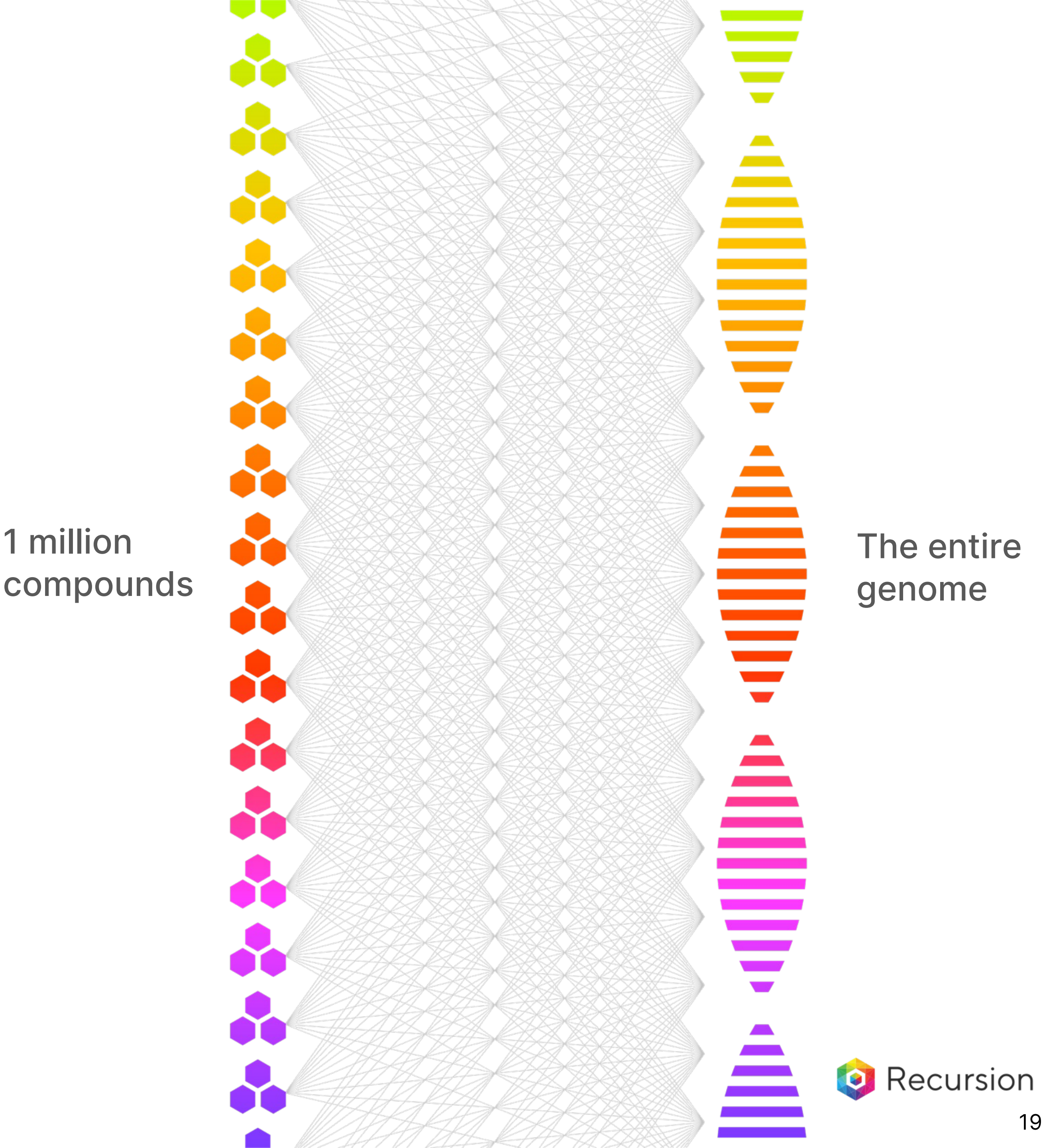
What I've shown is a very tiny slice of the entire content of our maps. There are not enough pixels on your screen to represent all of the trillions of inferred relationships in our maps, but at right I show a schematic view of the breadth of what we are able to see across the genome.

Traditional target-based screens



In a typical target-based screen, you may run an entire HTS deck against a single target, getting information for all compounds against that one protein…but no information at all about other targets. For those, you would have to explicitly set up counterscreens, increasing the time and cost of screening, and requiring advance knowledge of what to counterscreen. Furthermore, negatives from this screen are essentially “exhaust” – useless waste of the screen.

Recursion is generating exponentially more insights



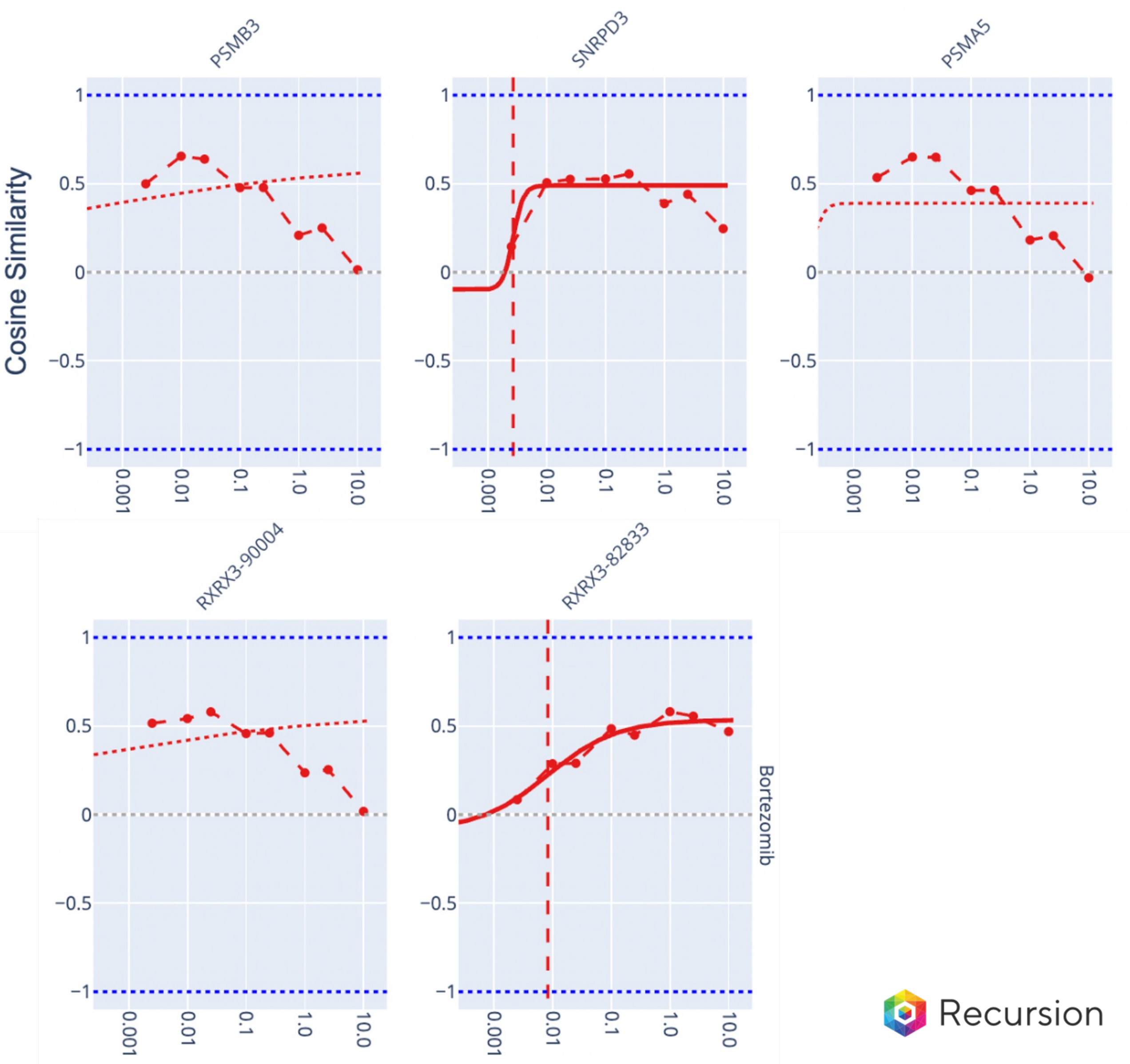
By contrast, in our approach, when we screen a single compound, we are able to relate it to each gene in the human genome as well as every other compound we have screened – turning past screening data from “exhaust” into fuel for our discovery engine, and enabling us to see off-target or polypharmacological activity of compounds immediately and continuously through hit-to-lead and LO.

Maps of Biology: **high-dimensional genome-wide screening**

One compound, biological similarity to all targets

- Recursion maps of biology allow the evaluation of concentration-response activity of each compound against *all* gene knockouts in one assay, rather than one assay *per target*.
- CRCs for bortezomib show similarity not just to proteasome subunit KOs, but also to splicing factor *SNRPD3* (known to regulate proteasomal RNAs) and masked potential targets.

(Check out MolRec™ at rxrx3.rxrx.ai!)

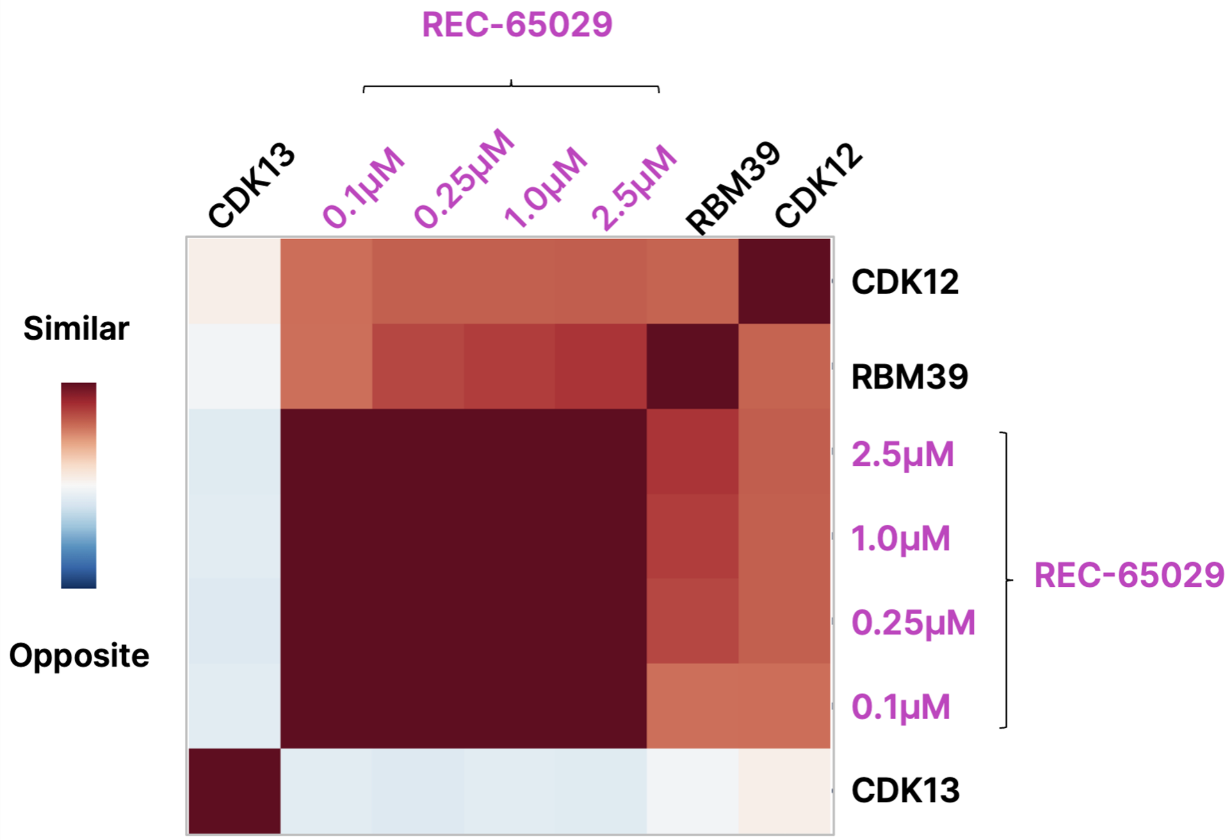


Recursion

Maps of Biology: **primary and alternative target selection**

Recursion “Target Gamma” program for HR-proficient ovarian cancer

- CDK12 has been advanced as a target to improve response in the HR-proficient setting.
- Selective inhibition of CDK12 over other CDKs, especially CDK13, is very challenging.
- Recursion maps of biology show that Inhibition of target RBM39 (e.g., with REC-65029) may mimic inhibition of CDK12 while mitigating toxicity due to CDK13 inhibition.

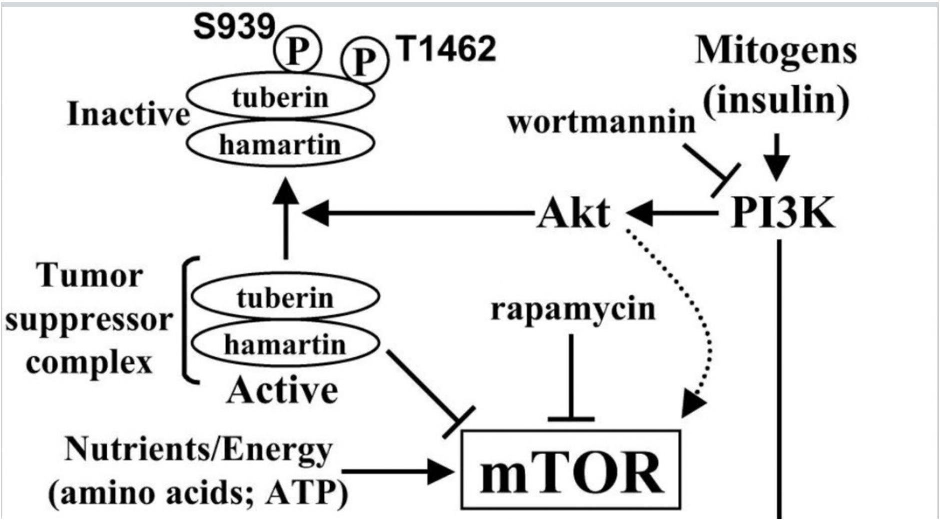


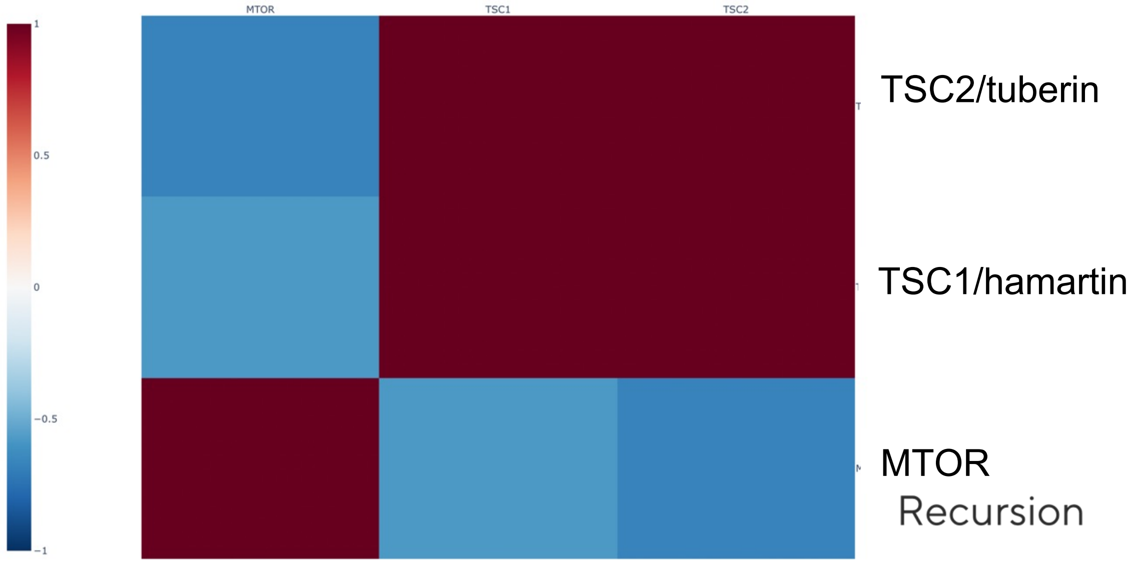
Enabling image-based maps of biology with ML/AI

Similarity is the fundamental property of maps

In order to build a map, we want to know how **similar** or **dissimilar** two biological states are.

We can intuitively think of this in a perturbative context: two perturbations are similar if they make the cell “do the same thing”, and opposite if one reverses the effect of the other.

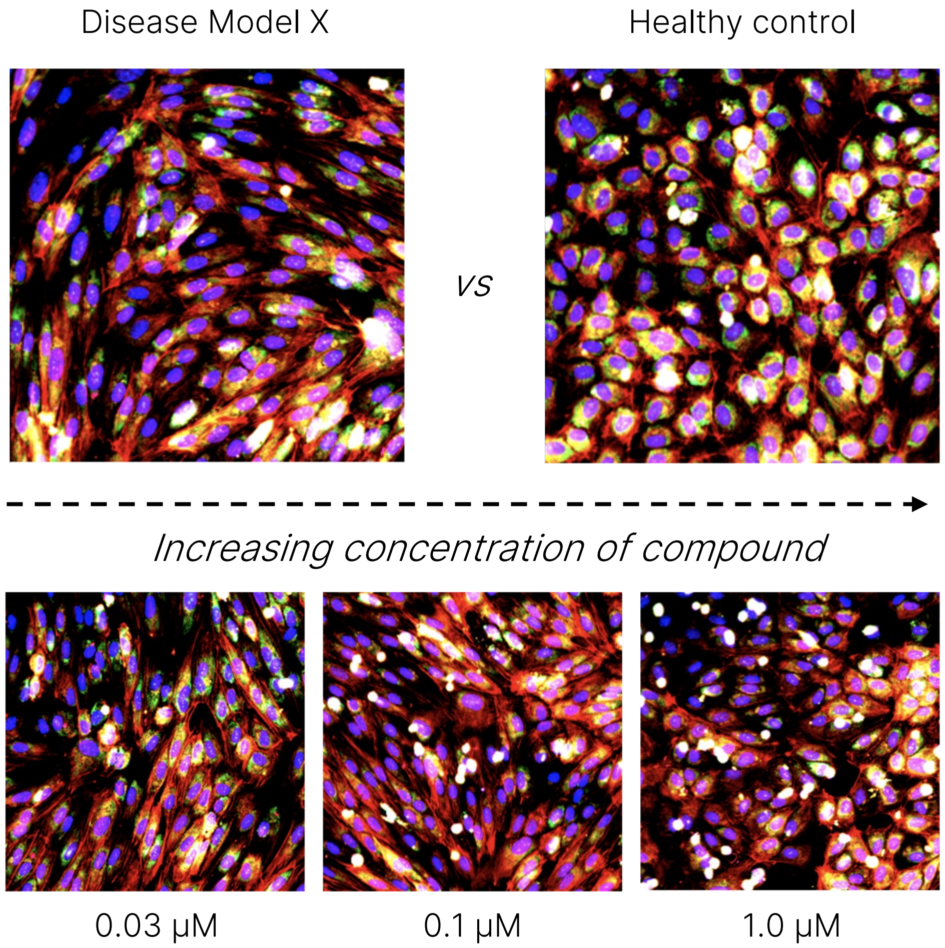




Tee AR et al. PNAS 2002

Imaging is distal, data-rich, and cheap

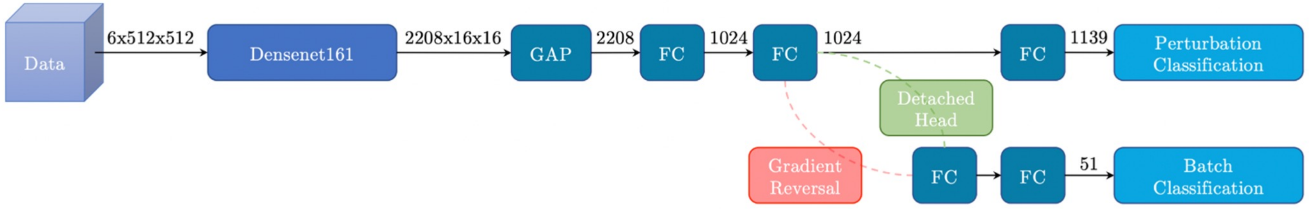
- Morphology is downstream of RNA and protein activity – and may observe effects molecular assays would not.
- Images can be **super cheap**.
- Recursion uses a standardized assay we call “**phenomics**”, staining six common cellular substructures.

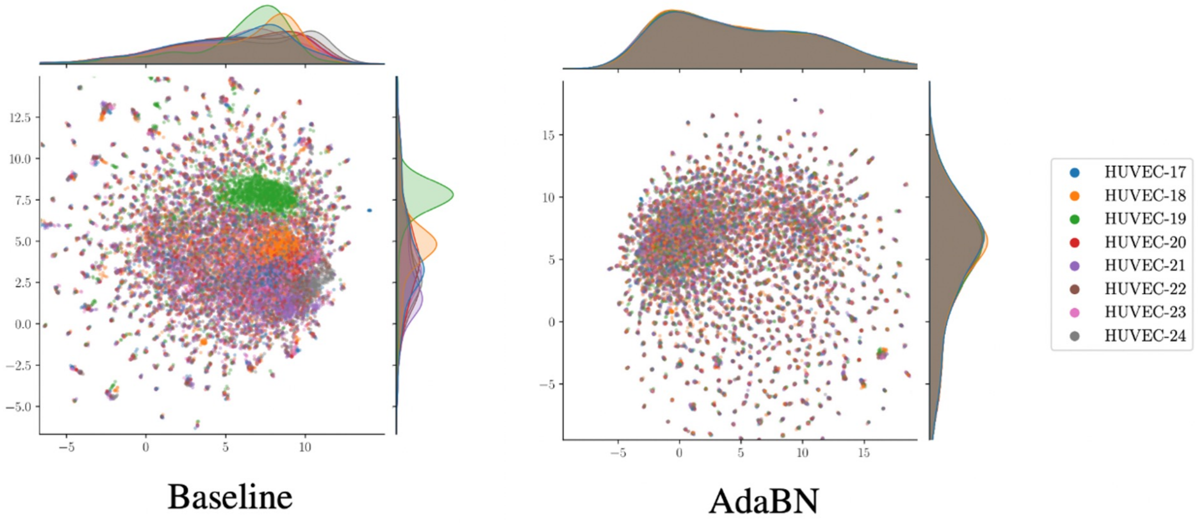


*Note: images shown above depict a disease model with visible phenotype for illustrative purposes only; primary utility of Recursion platform is to readily distinguish non-visible phenotypes

AI/ML turns unstructured images into computable data

- Computability is the fundamental challenge of mapping with images.
- DL algorithms can extract *biologically meaningful* representations of images and automatically correct issues like batch effect.
- These models have shown the power to accelerate development across cell types, and get better with more data.





Baseline

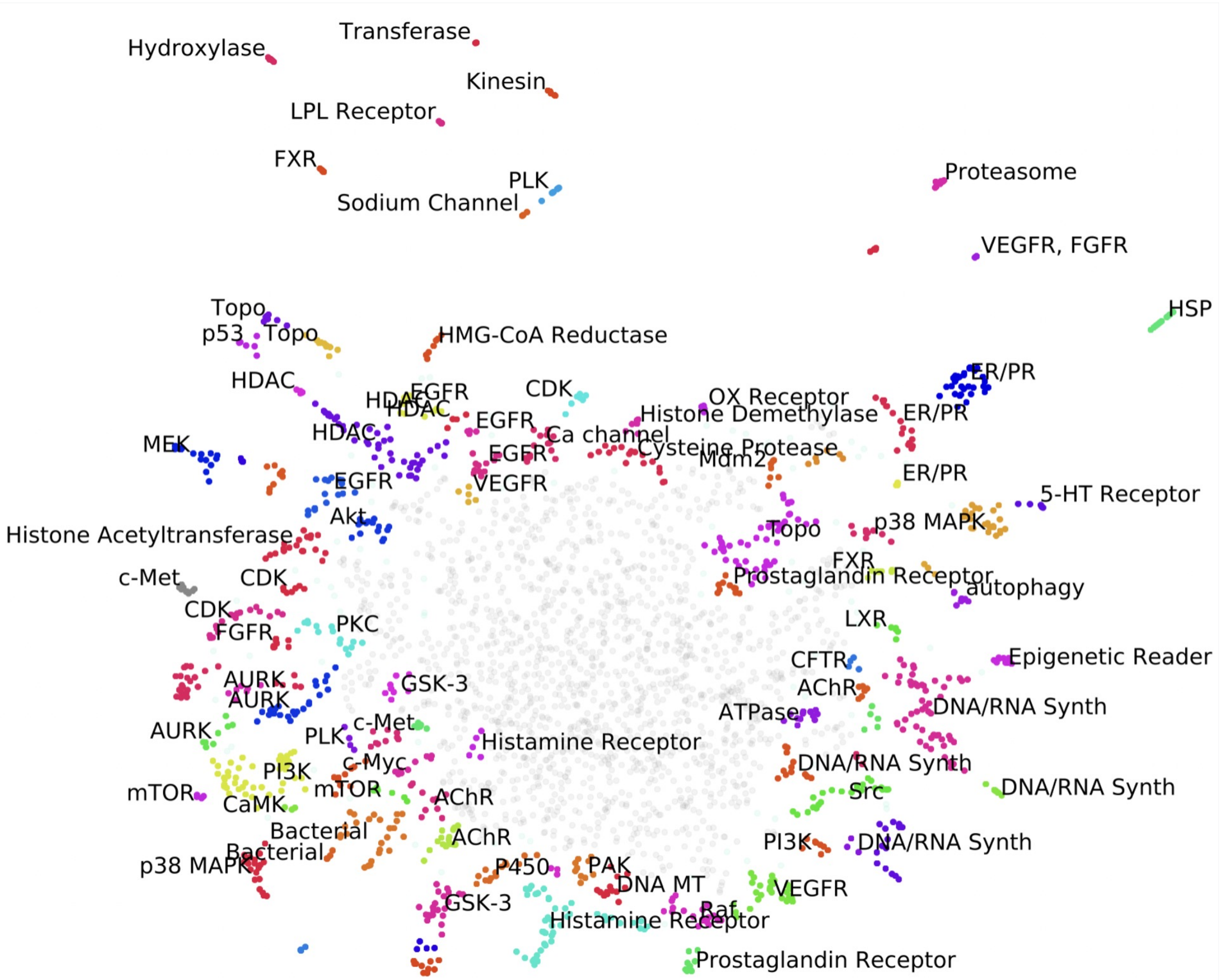
AdaBN

Recursion

Sypetkowski et al. https://arxiv.org/abs/2301.05768

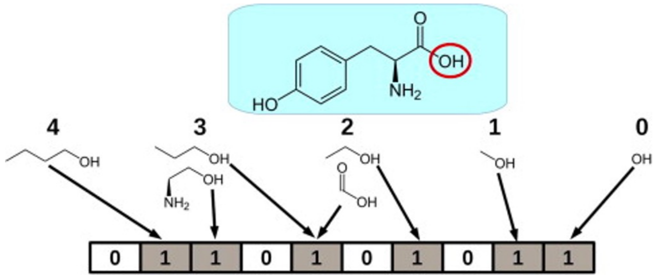
Standardized imaging assays capture broad swaths of biology

- Phenomics is an unexpectedly powerful standard assay capable of sensitive detection and quantification across 100s-1000s of mechanisms.



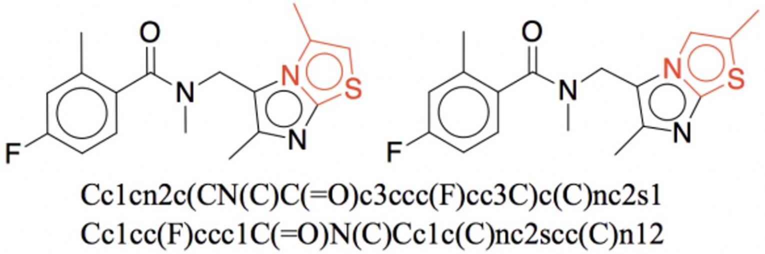
Molecular Foundation Models for Chemical Representations

Current chemical representations are unsatisfactory



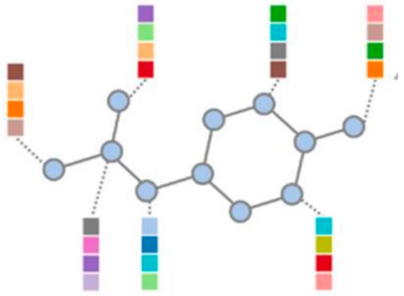
Fingerprints
Fixed-length, hand-tuned vector representations.

Strong baselines, but interpolate poorly outside training, and incompletely represent molecules.



Cc1cn2c(CN(C)C(=O)c3ccc(F)cc3C)c(C)nc2s1
Cc1cc(F)ccc1C(=O)N(C)Cc1c(C)nc2scc(C)n12

SMILES
String representations convenient for language models, but nonunique: many SMILES for the same molecule, and similar molecules can have very different SMILES.



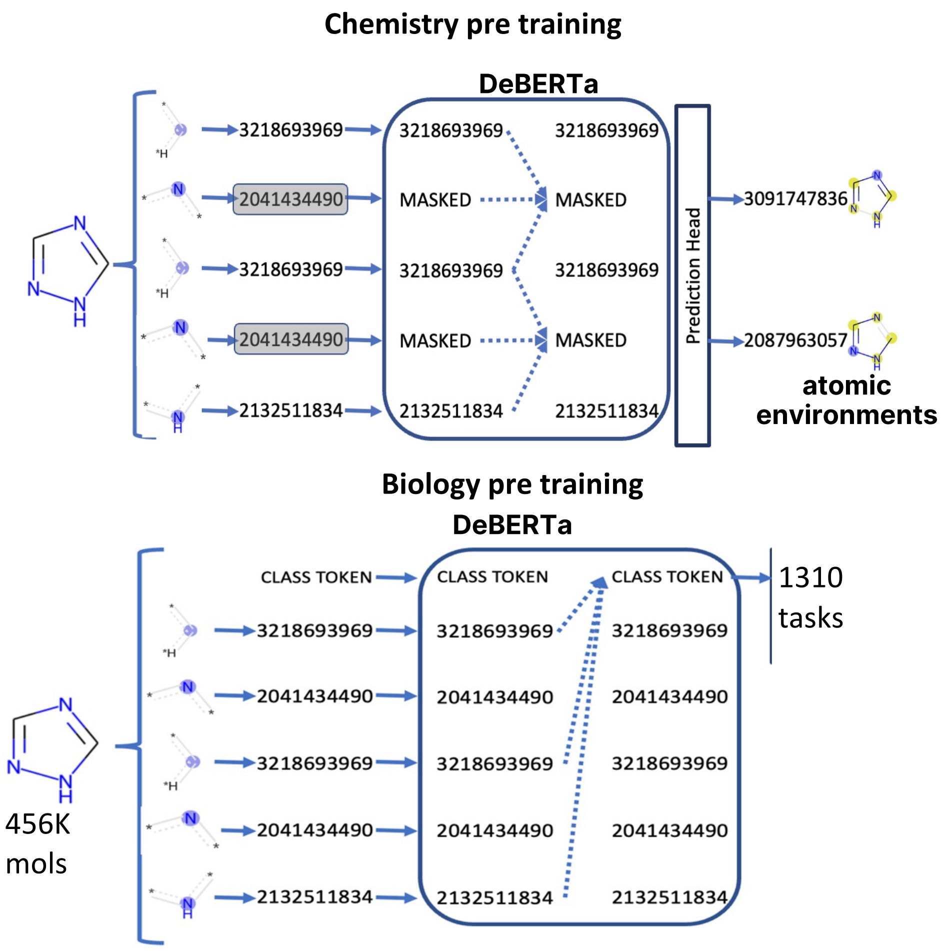
Graphs
Natural representations of molecules useful for graph neural networks, but typical graph models fail to aggregate information from distant parts of a molecule.

MolE adapts modern large language models to train a superior molecular representation space.

Méndez-Lucio O, Nicolaou C, Earnshaw B. arXiv:2211.02657v1

MolE: A Molecular Foundation Model for Drug Discovery

- Chemistry problems are categorically poor in labeled data.
- MolE adapts the DeBERTa Transformer architecture to represent *both content and relative graph position* to pre-train models that both “understand chemistry” and “understand biology”.
- Combining self-supervision and fully–supervised learning enables effective few-shot learning -- only limited fine-tuning needed on particular target problems.



Méndez-Lucio O, Nicolaou C, Earnshaw B. arXiv:2211.02657v1

MolE is a top performer on the TDC ADMET benchmarks

- On the 22 Therapeutic Data Commons ADMET prediction tasks: MolE achieves #1 or #2 performance on the leaderboard on 14/22 tasks, including all distribution and metabolism (CYP2C9/2D6/3A4) tasks, and #1 in 9/22.

	MolE TDC Rank			MolE TDC Rank	
A	Lipophil	1	D	Plasma binding	1
	Caco2	2		VDss	1
	Solubility	3		BBB	2
	Bioavail.	5	E	t1/2	1
	Pgp	5		Cl_mic	1
	HIA	6		Cl_hep	6
M	2C9 inhibition	1	T	LD50	1
	2C9 substrate	1		hERG	4
	2D6 inhibition	2		DILI	5
	2D6 substrate	1		Ames	8
	3A4 inhibition	2			
	3A4 substrate	2			

Méndez-Lucio O, Nicolaou C, Earnshaw B. arXiv:2211.02657v1

RxRx3: Enabling ML research in phenomics

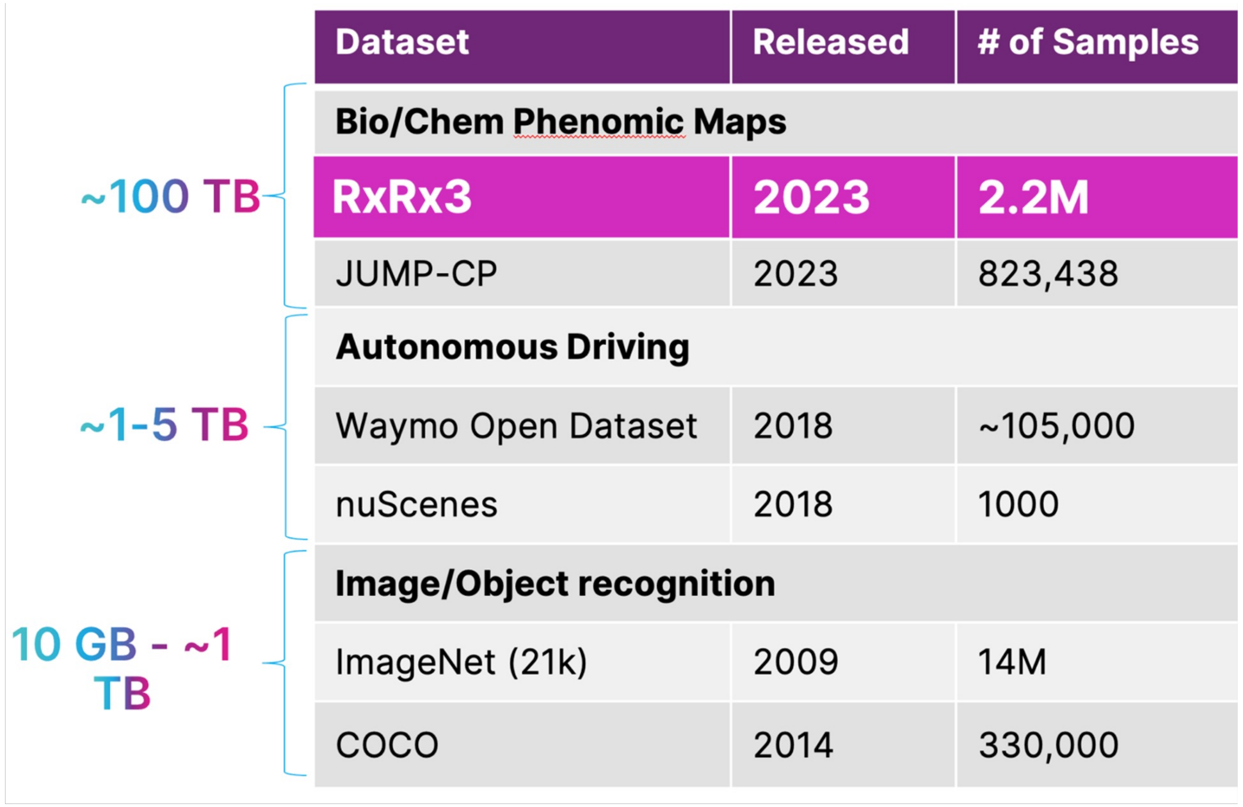
RxRx3: Leading the field in open science

rxrx.ai/rxrx3

RxRx3:

- Images, metadata, and DL embeddings of knockouts of ~17K genes + ~1700 SMs @ multiple concentrations .
- The largest publicly-released data set of perturbative cellular imaging, all generated at a single site with a consistent protocol.

Historically, datasets have driven advances in machine learning technology.

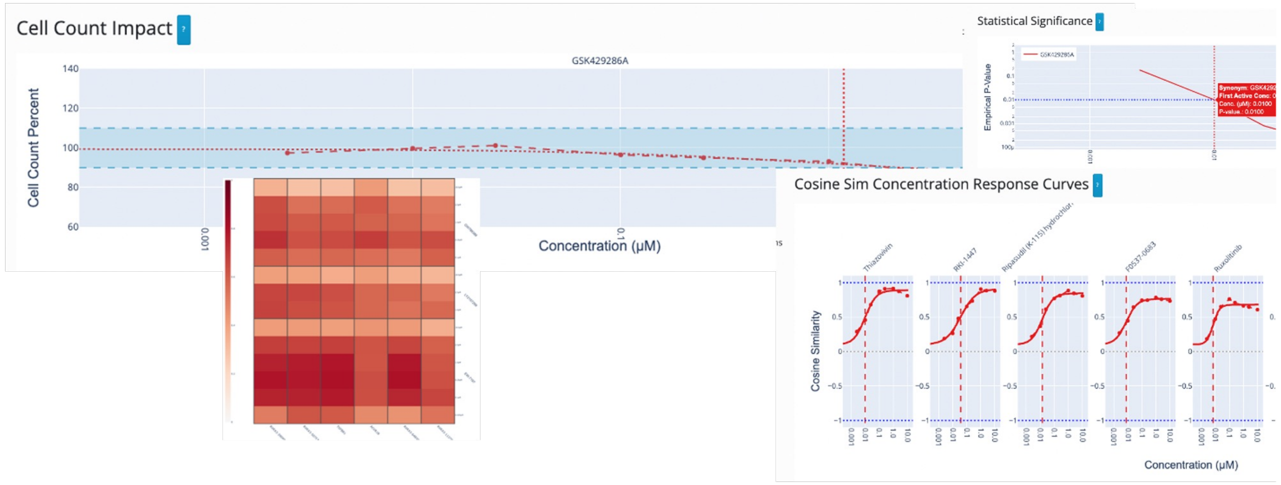


	Dataset	Released	# of Samples
~100 TB	**Bio/Chem Phenomic Maps**		
	RxRx3	**2023**	**2.2M**
	JUMP-CP	2023	823,438
~1-5 TB	**Autonomous Driving**		
	Waymo Open Dataset	2018	~105,000
	nuScenes	2018	1000
10 GB - ~1 TB	**Image/Object recognition**		
	ImageNet (21k)	2009	14M
	COCO	2014	330,000

MolRec: A keyhole view into the Recursion Map of Biology

rxrx.ai/rxrx3

MolRec shows off the ability to relate small molecules to each other and to gene knockouts, with the suite of analyses you might expect to want in driving a discovery program – cellular toxicity, on- and off-target similarity, and compound similarity.

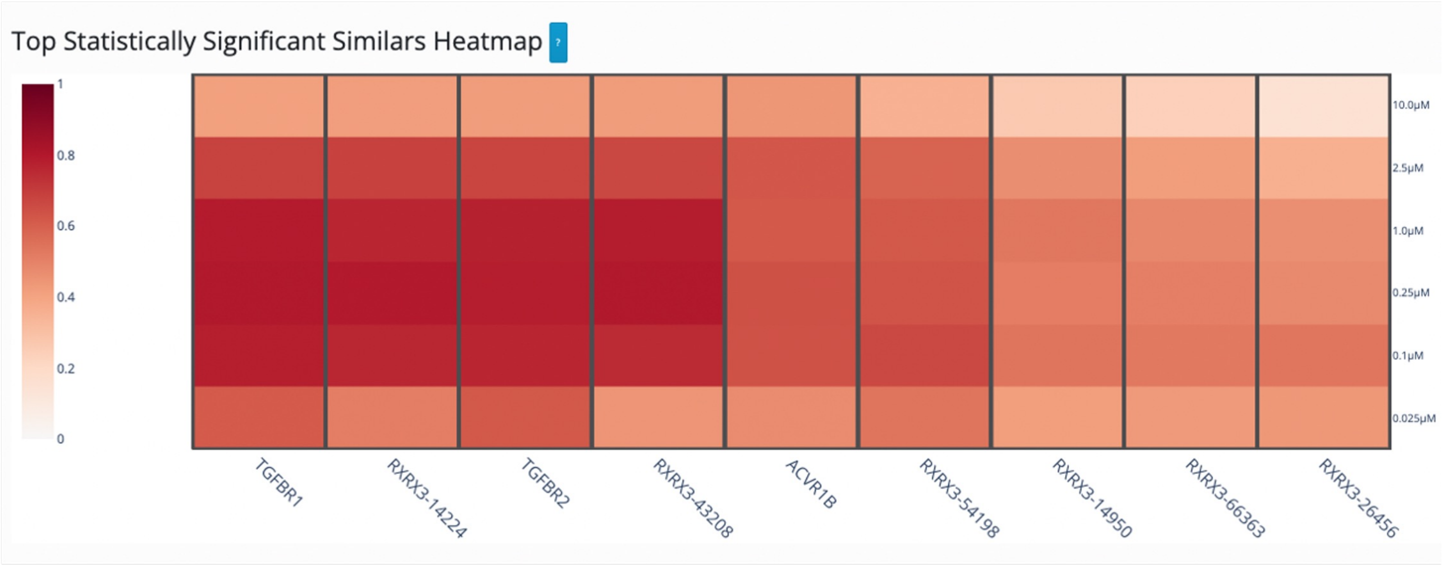


Recursion

RxRx3 and MolRec: blinded research data sets

rxrx.ai/rxrx3

RxRx3 and MolRec are partially blinded. We expect to unblind more of these over time.



Top Statistically Significant Similars Heatmap

RxRx3 and MolRec: blinded research data sets

rxrx.ai/rxrx3

RxRx3 and MolRec are partially blinded. We expect to unblind more of these over time.

But you know, if you want to pay us some money to unblind it, let's talk.

For just $0.99 more, I'll personally hand-deliver the hard drives.

Conclusion

Biology is the next frontier for chemistry (learning)

- Standardized cellular imaging (**phenomics**) is incredibly data-rich and scalable; DL-learned features enable high-dimensional, **dense data matrices** for chemistry
- Self-supervision and biological pretraining enables DL methods to create chemical representations that power **few-shot learners** for challenging biological properties
- Recursion has released **MolRec** to provide a view into the power of *mapping and navigating*, and **RxRx3** to advance research on machine learning and cellular imaging.

Questions?

info@rxrx.ai for questions on RxRx3 and MolRec

@ImranSHaque on Twitter or @ihaque@genomic.social on Mastodon

Questions?

info@rxrx.ai for questions on RxRx3 and MolRec

@ImranSHaque on Twitter or @ihaque@genomic.social on Mastodon

Recursion

We are hiring!

recursion.com/careers

Computational chemistry, computational biology, machine learning, software engineering, and more!