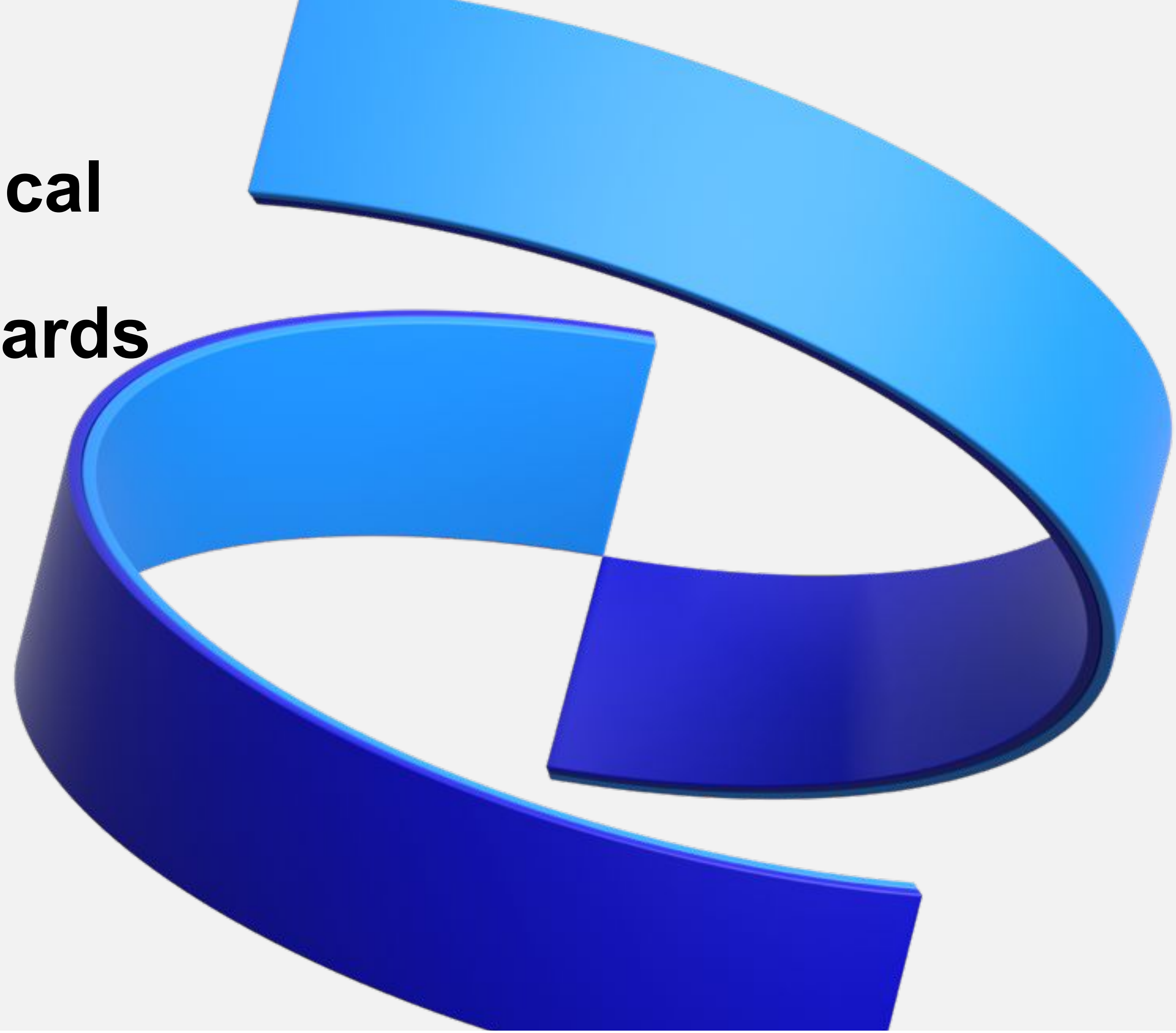
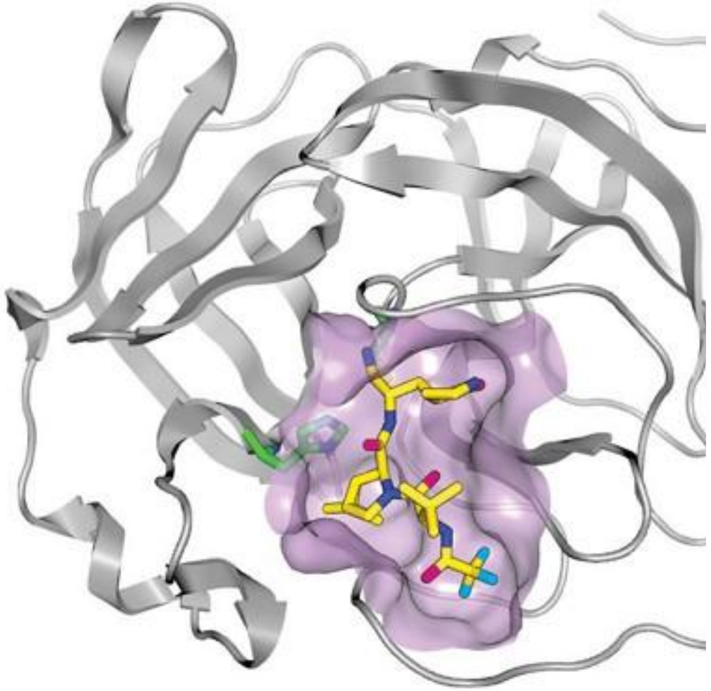
Applications of classical and contemporary machine learning towards drug discovery

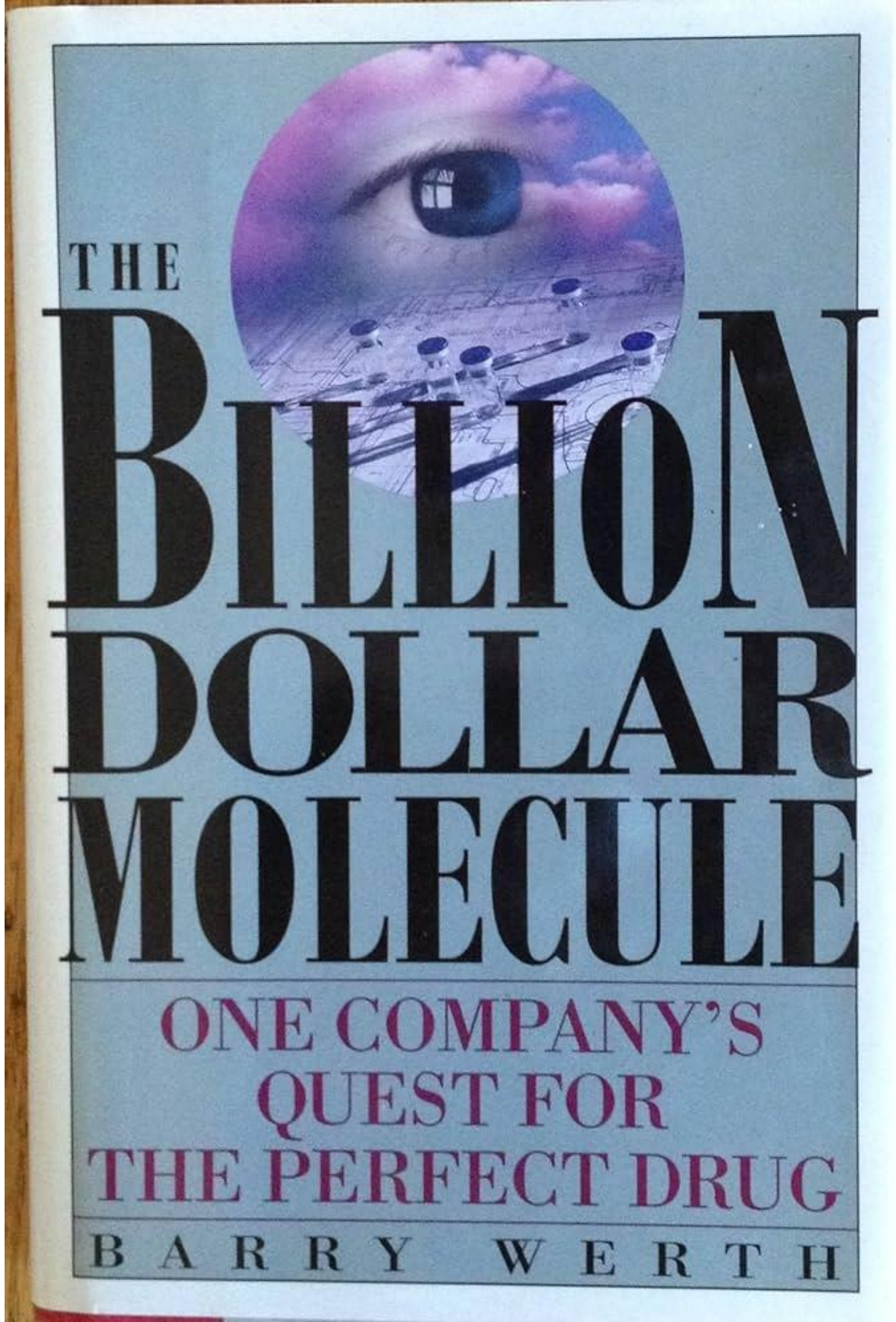


Enoch Huang

15 November 2023

- Elegance





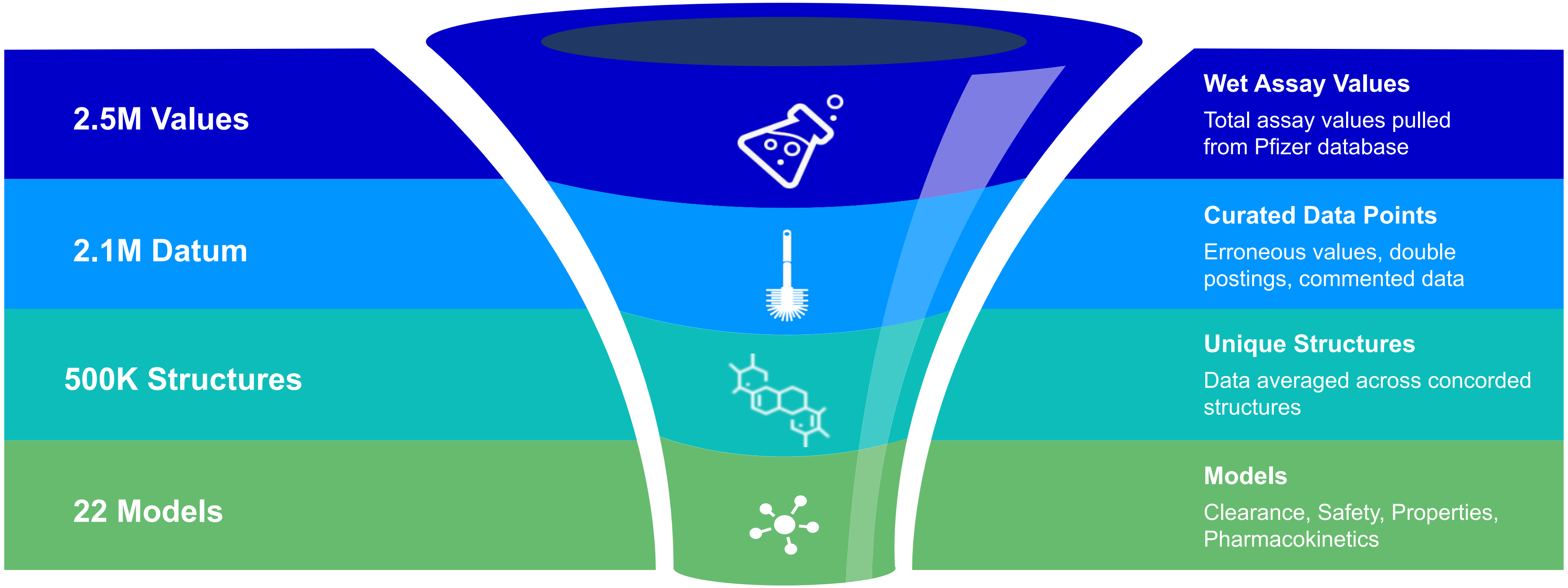
THE BILLION DOLLAR MOLECULE

ONE COMPANY'S QUEST FOR THE PERFECT DRUG

BARRY WERTH

- Reality

Machine Learning Tools to Expedite Small Molecule Drug Design



2.5M Values

Wet Assay Values
Total assay values pulled from Pfizer database

2.1M Datum

Curated Data Points
Erroneous values, double postings, commented data

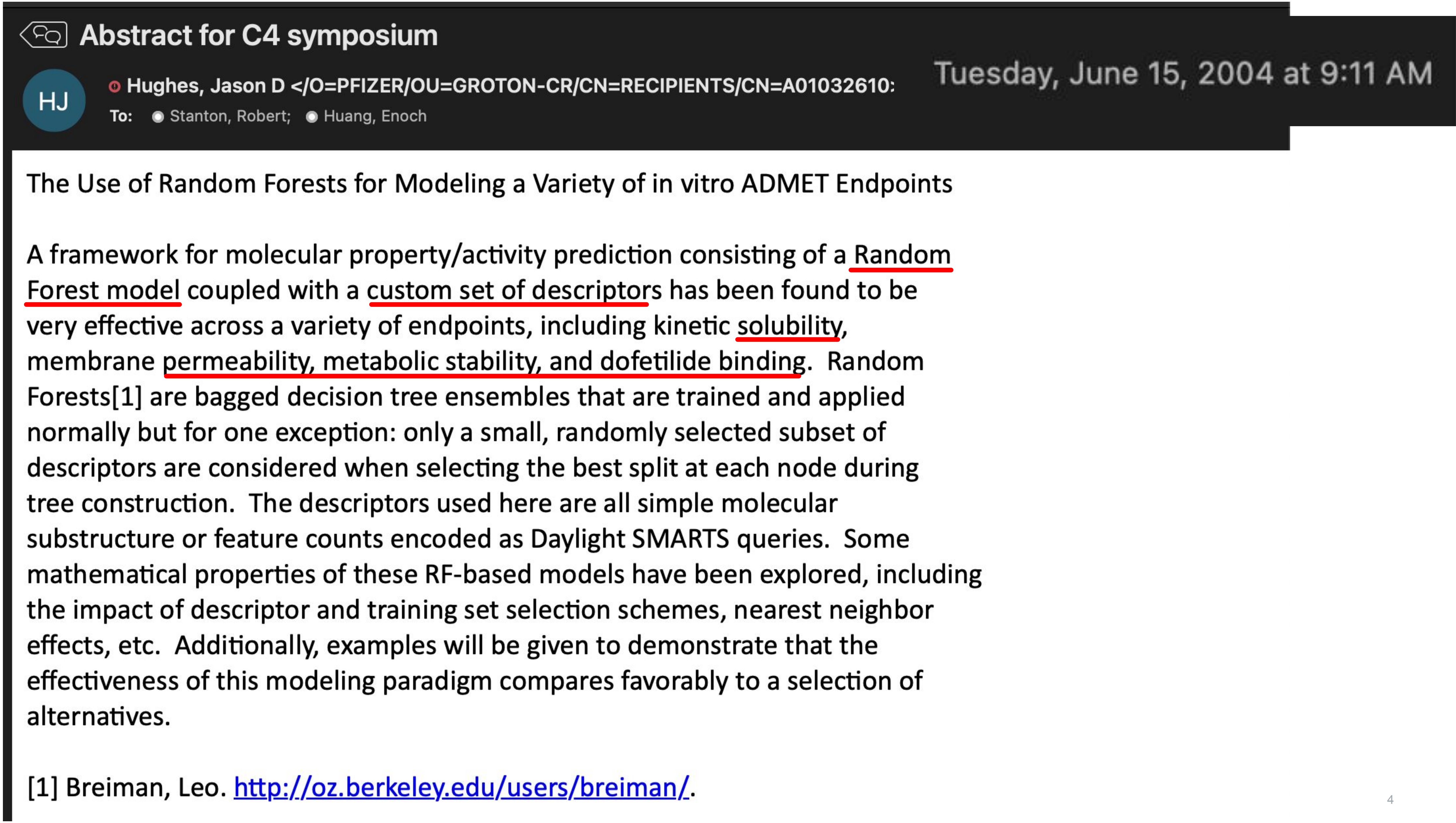
500K Structures

Unique Structures
Data averaged across concorded structures

22 Models

Models
Clearance, Safety, Properties, Pharmacokinetics

Models are rebuilt to include new lab data every 2 weeks □ ~80 % of predictions within 2-fold



Abstract for C4 symposium

HJ Hughes, Jason D </O=PFIZER/OU=GROTON-CR/CN=RECIPIENTS/CN=A01032610:

To: Stanton, Robert; Huang, Enoch

Tuesday, June 15, 2004 at 9:11 AM

The Use of Random Forests for Modeling a Variety of in vitro ADMET Endpoints

A framework for molecular property/activity prediction consisting of a Random Forest model coupled with a custom set of descriptors has been found to be very effective across a variety of endpoints, including kinetic solubility, membrane permeability, metabolic stability, and dofetilide binding. Random Forests[1] are bagged decision tree ensembles that are trained and applied normally but for one exception: only a small, randomly selected subset of descriptors are considered when selecting the best split at each node during tree construction. The descriptors used here are all simple molecular substructure or feature counts encoded as Daylight SMARTS queries. Some mathematical properties of these RF-based models have been explored, including the impact of descriptor and training set selection schemes, nearest neighbor effects, etc. Additionally, examples will be given to demonstrate that the effectiveness of this modeling paradigm compares favorably to a selection of alternatives.

[1] Breiman, Leo. http://oz.berkeley.edu/users/breiman/.

Our Global ADME/T Machine Learning Models are used ~ 6M / day

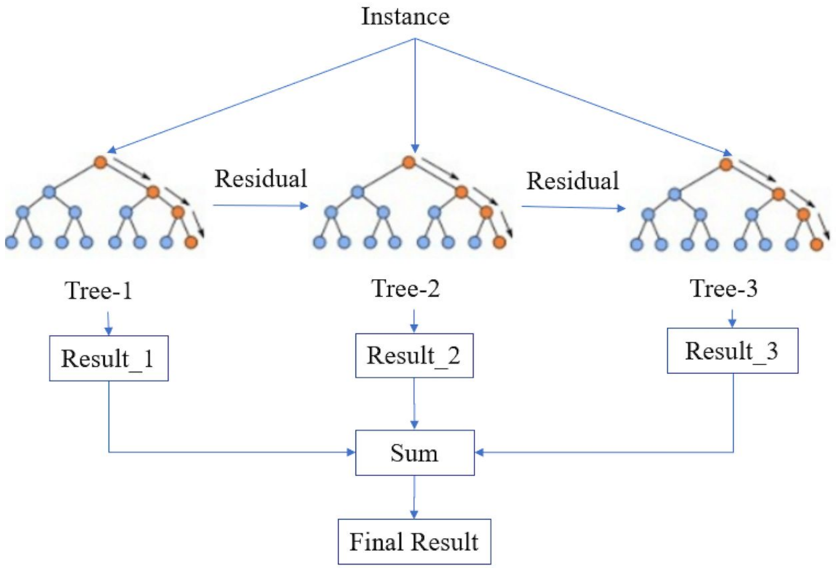
Model	Row Count	% Within 2-Fold
HHEP Clearance	92,944	74
HLM Clearance	393,826	76
RLM Clearance	118,201	75
RRCK (Pass. Perm.)	265,074	77
NIH MDR (Pgp) ER	32,598	78
BCRP ER	28,992	80
Fu, microsomes	7,845	87
Human Fu, plasma	10,215	69
Rat Fu, plasma	8,030	69
Mouse Fu, plasma	4,013	67
Brain Fu, tissue	3,216	70
Human Blood/Plasma	2,948	R^2=0.45
Rat Blood/Plasma	1,742	R^2=0.68
Human Vdss	1,271	62
Rat Vdss	2,341	61
SFLogD	212,234	R^2=0.78
ELogD	83,277	R^2=0.86
Kinetic Solubility	82,996	64
Dofetilide Ki	224,486	66
Herg IC50	12,963	60
THLE IC50	101,201	77
OATP1B1 Inh	11,450	R^2=0.67

- Design idea prioritization
- Monomer selection in Parallel Medicinal Chemistry (PMC)
- Calculation of PK and Dose

Most of the effort for a new model is the curation of input data

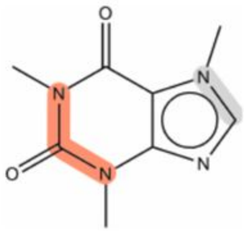
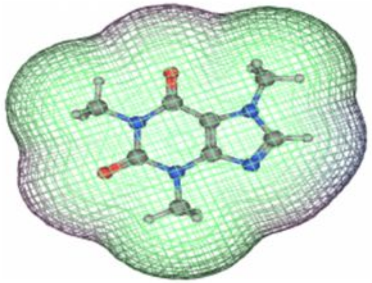
- What assay data is available in the database?
- Are the data suitable?
  - Replicate variability
  - Comment Fields
  - Posting errors
  - Unit errors
- If there are different assays for the same endpoint, can they be combined?
  - Normalization of units of measurement
  - Overlap
  - Correlation
- Is the assay updated regularly with new data?

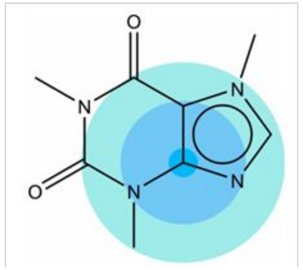
'Classical’ machine learning methods: tree-based with descriptors



- **XGBoost**
- **Cubist**

HHEP Metrics	Cubist	DNN	XGBoost
%within 2-Fold	64	60	63
Mean Fold Error	2.2	2.3	2.1
Pearson’s R	0.56	0.50	0.59
Spearman’s R	0.66	0.61	0.70

Representation		Descriptors
$C_8H_{10}N_4O_2$	→	molecular weight: 194.2 u number of heavy atoms: 14
	→	number of rings: 2 logP(o/w): -0.604 MACCS keys: 65, 77
	→	van der Waals volume: 175 Å^3 van der Waals surface area: 203 Å^2

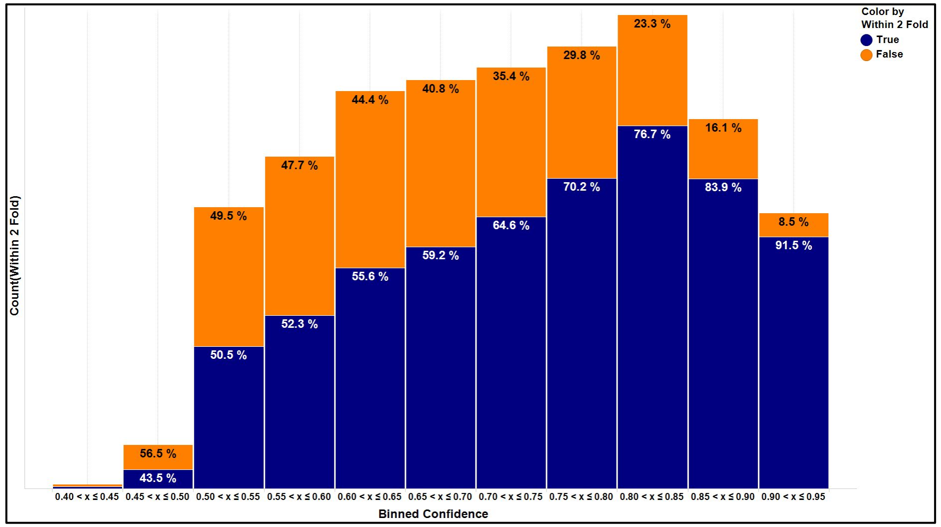


→

Layer	Atom types
0	C.ar
1	C.ar, N.2, N.am
2	C.2, N.pl3, C.2, C.3, C.2

Confidence metrics significantly increased adoption of *in silico* models

- We generate an interpretable probability-based confidence metric
- The score is calibrated via cross-validation to a confidence metric that represents an expected error probability
- The confidence metric captures how close the test compound is to its nearest neighbors in both descriptor space and activity space



Prospective Confidence Metric Performance

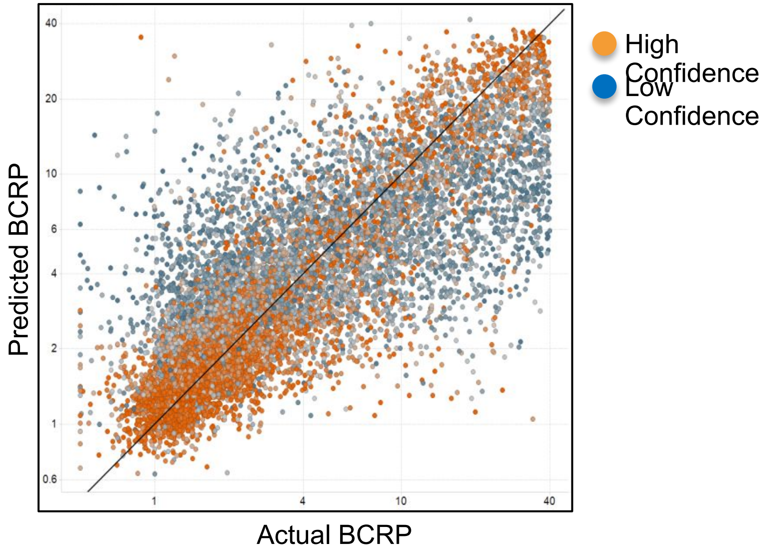
$$wRMSD = \sqrt{\frac{\sum_{i=1}^{N} w_i^2 (\hat{y} - y_i)^2}{\sum_{i=1}^{N} w_i^2}}$$

$\hat{y} = \textit{predicted value of test compound}$

$y_i = \textit{actual value of ith neighbor in training set}$

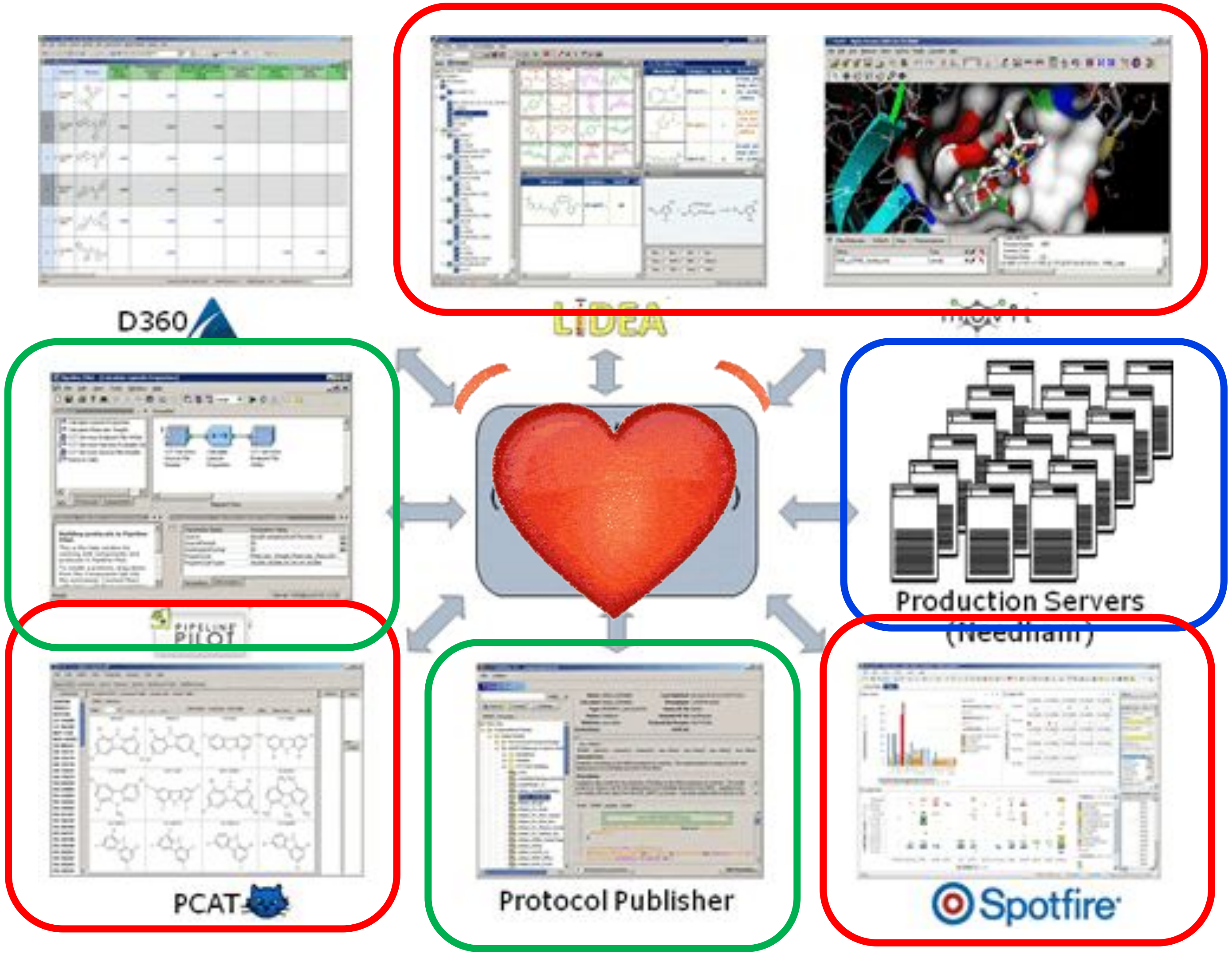
$w_i = \frac{1}{D+0.5}$

$\mathrm{D} = \text{Manhattan Distance between test compound and ith neighbor}$



Keefer et al (2013) dx.doi.org/10.1021/ci300554t

ML models, a computational ecosystem, and culture

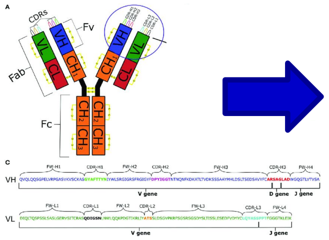


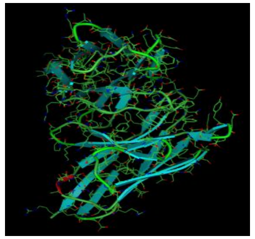
Keys to Success

- Talent, expertise, and remit
- Global, authoritative, standardized data repository
- Infrastructure for publishing, executing, and deploying models
- Confidence scores
- Sophisticated design culture

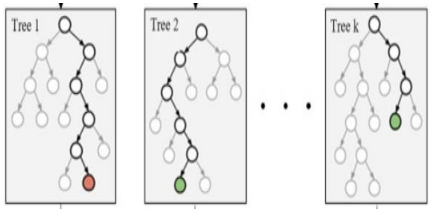
Attempting the same strategy for large molecules (mAbs) was unsuccessful

Primary sequence

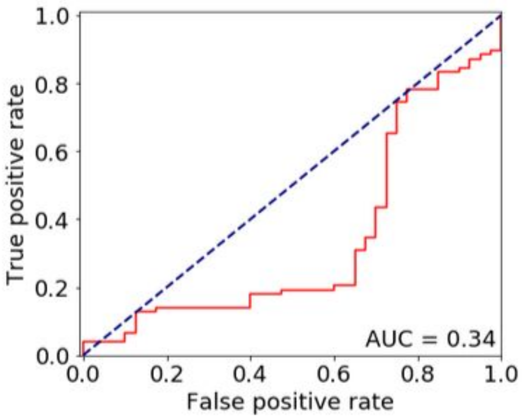




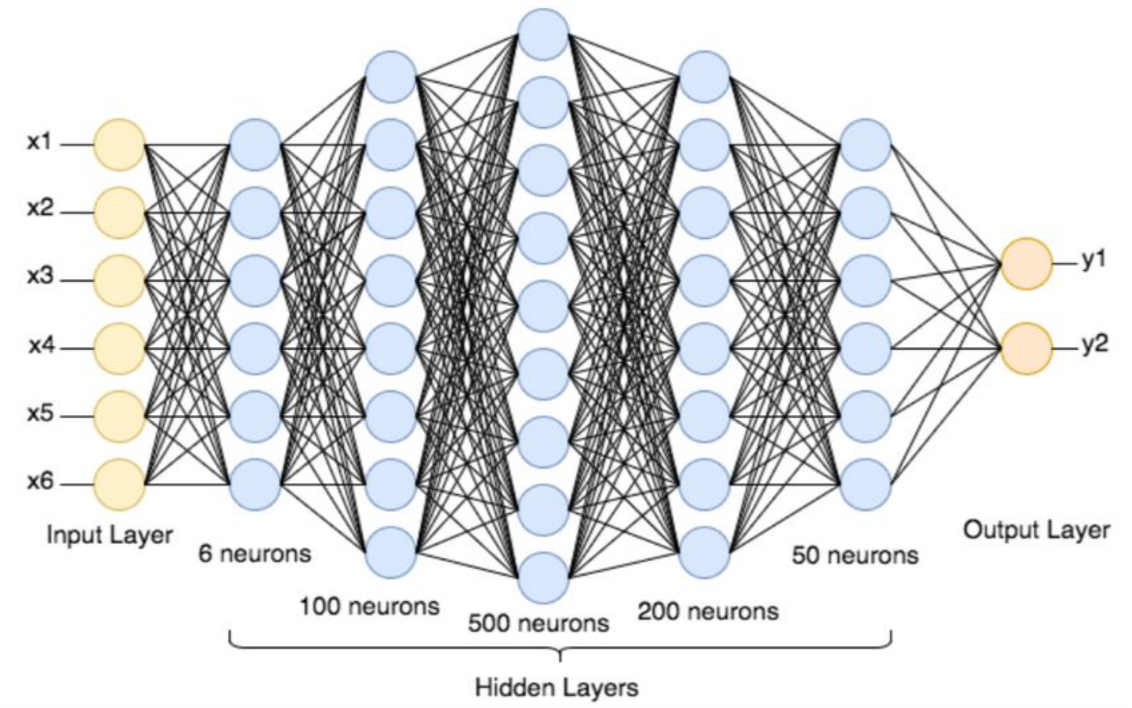
Generate features
Pre-specified set of physchem properties



Leave-group-out validation

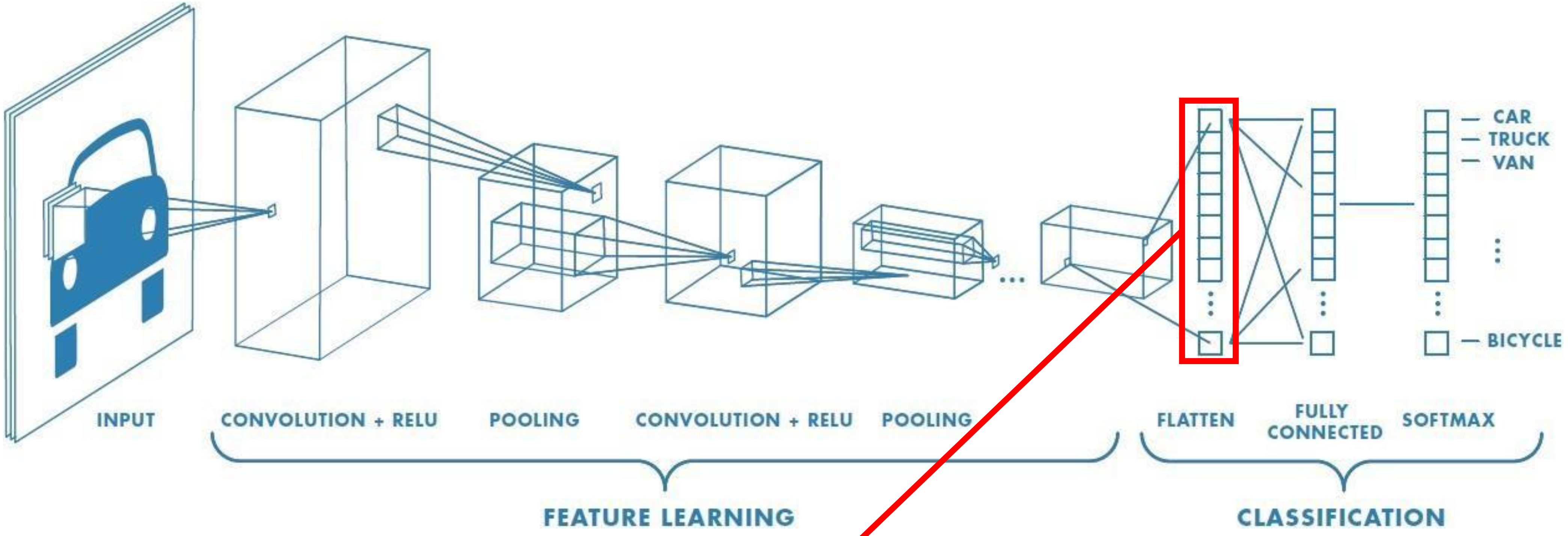


Traditional ML techniques *failed* to produce generalizable models


x1
x2
x3
x4
x5
x6
y1
y2
Input Layer
6 neurons
100 neurons
500 neurons
200 neurons
50 neurons
Output Layer
Hidden Layers

IMAGENET

We convert images into lower dimensional vector embeddings

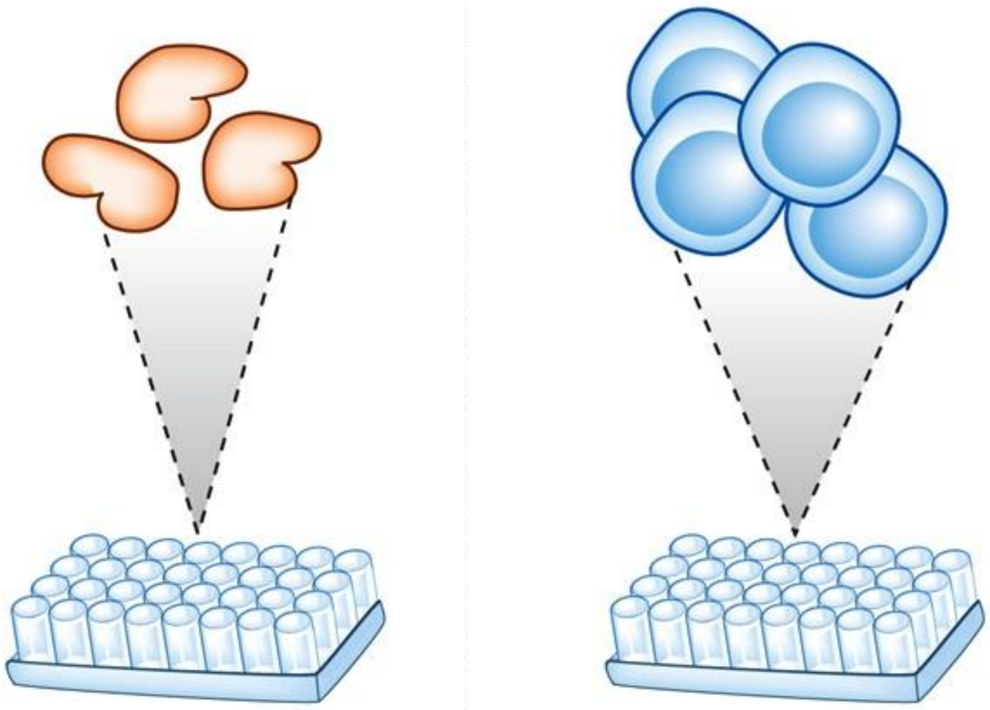


Extract intermediate hidden layer as feature embedding

Image source: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a5 3

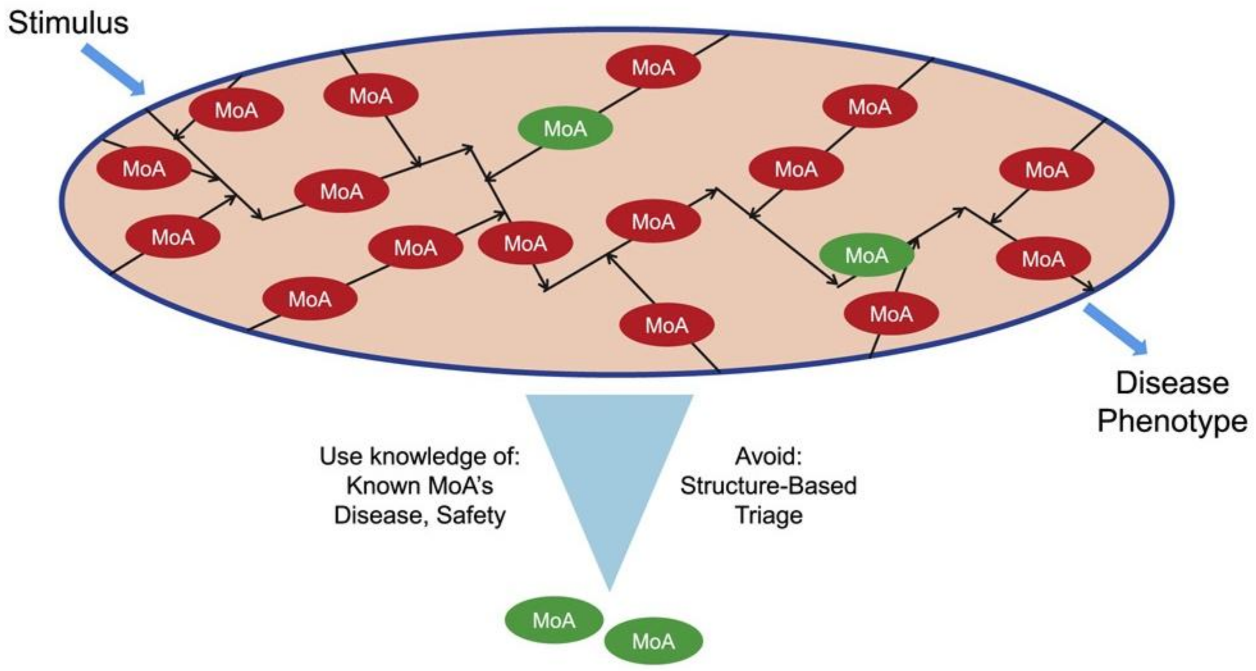
Target-based assays

Cell-based assays



Target-centric
Reductionist view
Validation in cell-based assays needed

Target-agnostic
Holistic view
More physiologically relevant
Target identification and validation needed

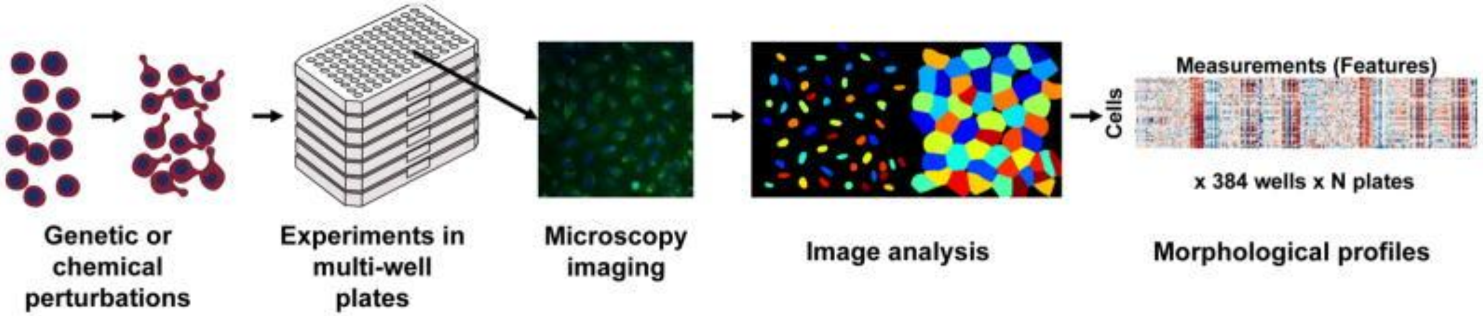


https://www.nature.com/articles/nchembio.2383

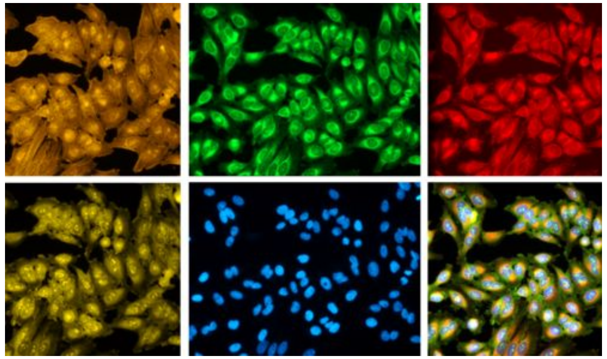
Vincent et al (2020) https://doi.org/10.1016/j.chembiol.2020.08.009

Can phenotypic changes of cellular components from compound treatment be learned and associated with specific mechanisms of action (MOAs) via deep learning?

- Screen compounds, use computer vision to determine targets/MOAs

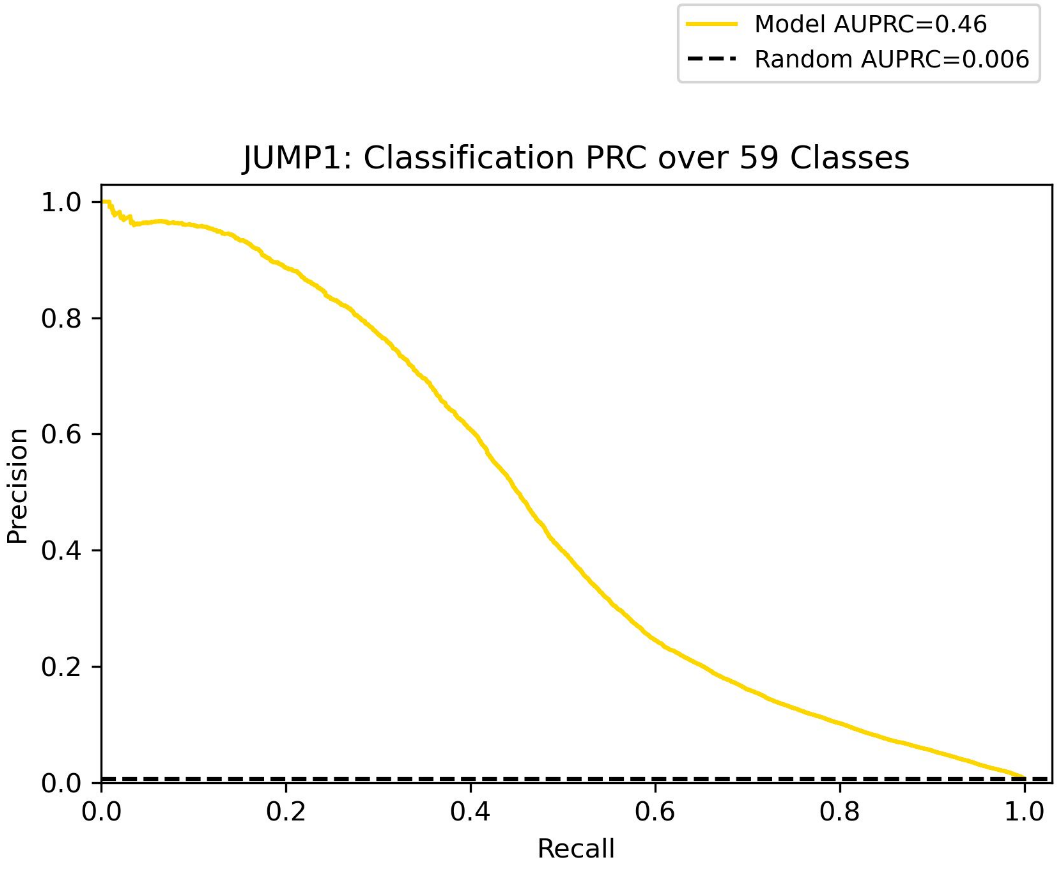


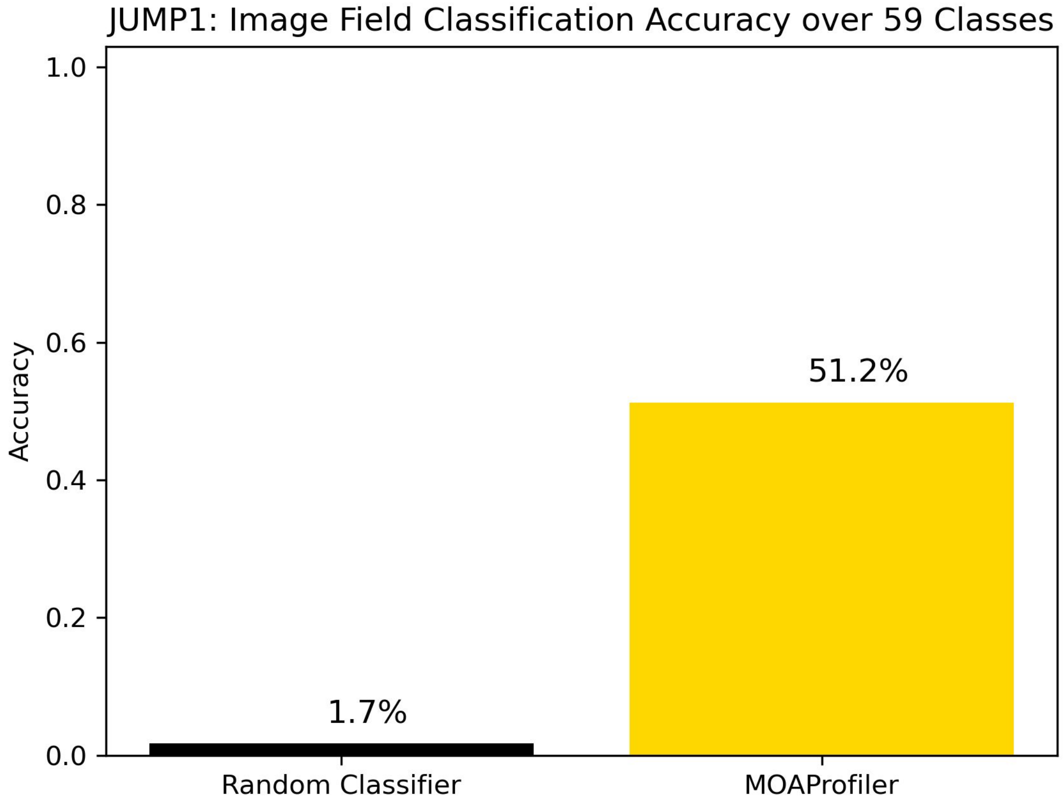
- Cell Painting assay (Bray *et al*, 2016)
  - Reveals 8 broadly relevant cellular components or organelles using 6 fluorescent dyes



Chandrasekaran et al. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nature Reviews Drug Discovery (2020).

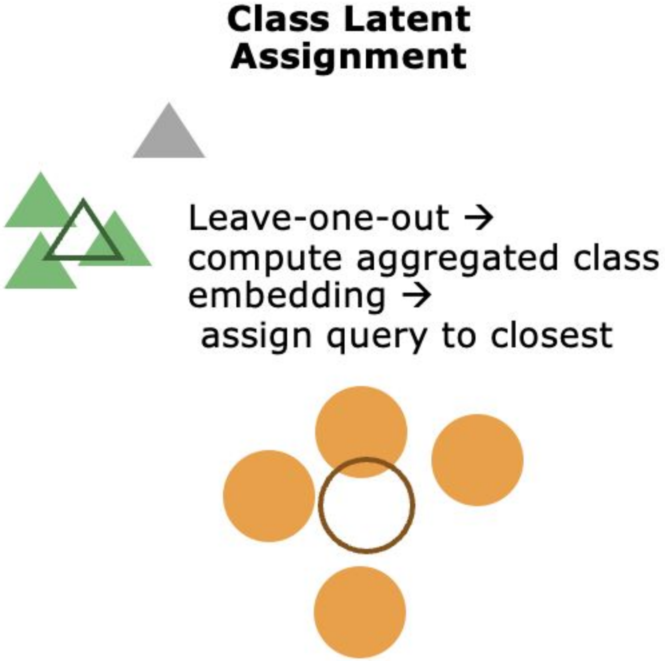
Deep Learning can accurately classify 59 different multi-compound MOAs





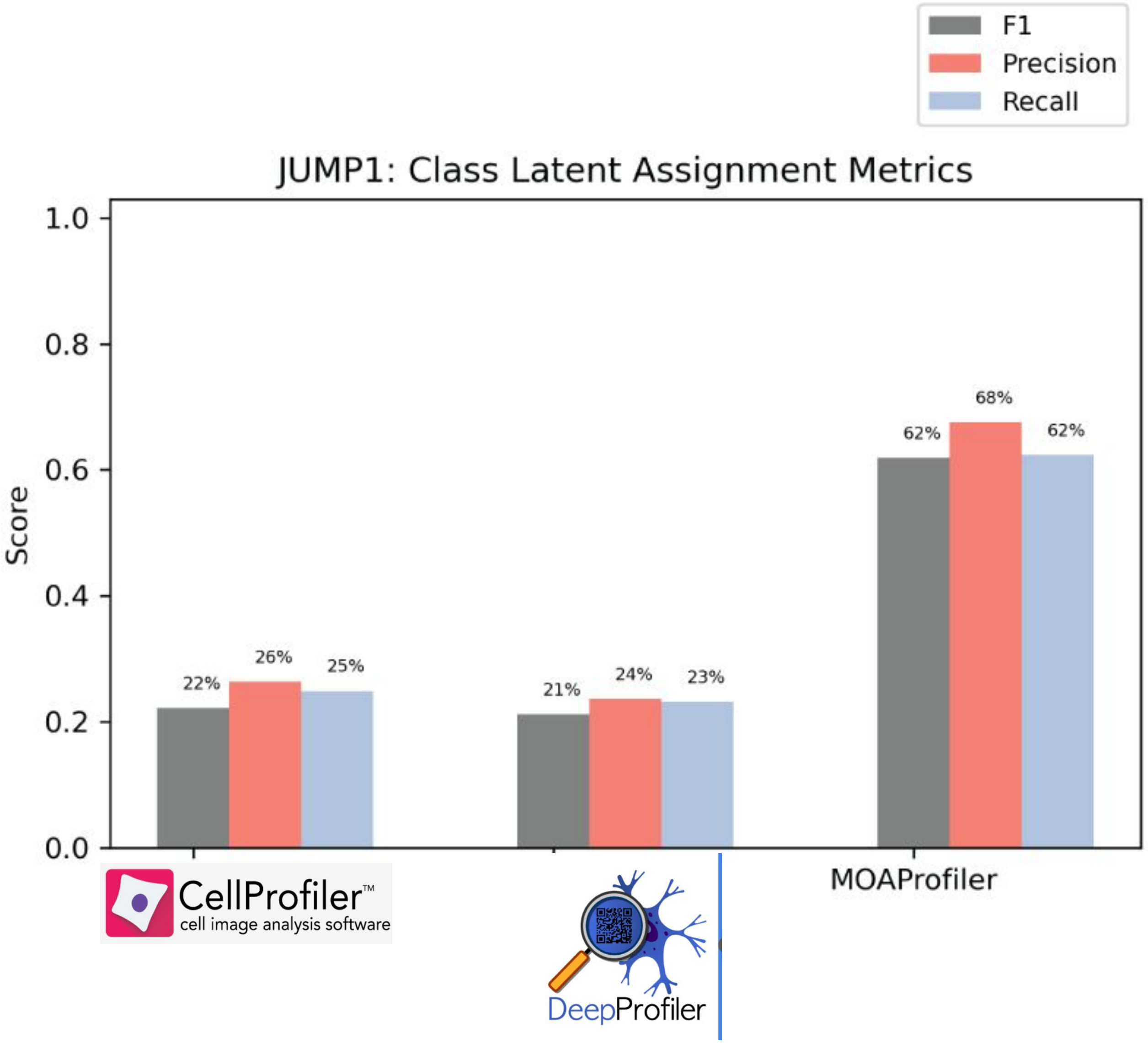
Wong et al 2023 DOI: 10.1039/d3dd00060e

Our embeddings outperform existing methods for MOA class assignment



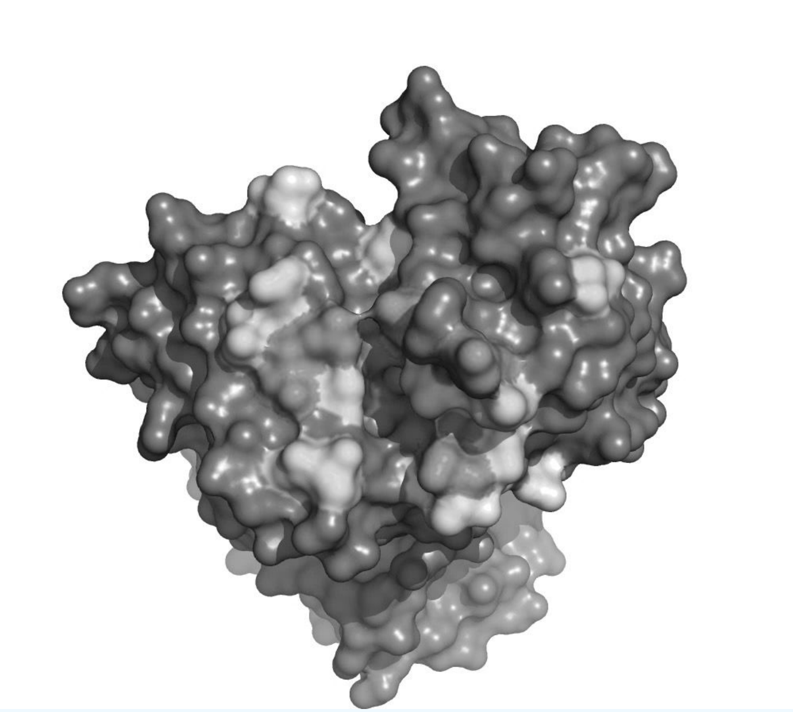
Class Latent Assignment

Leave-one-out → compute aggregated class embedding → assign query to closest



Case Study: Optimize a Domain of Trispecific Antibody

Internally developed AI tool delivers key physical property with speed



Situation & Challenges Targeted

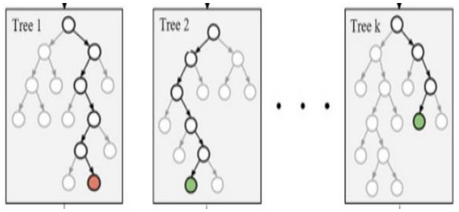
- Low antibody viscosity is critical for high dose, low volume subcutaneous delivery and is easier to manufacture
- Traditional viscosity optimization typically requires multiple production / screening cycles
- Scarcity of training data prevented prior AI methods from making accurate antibody viscosity predictions

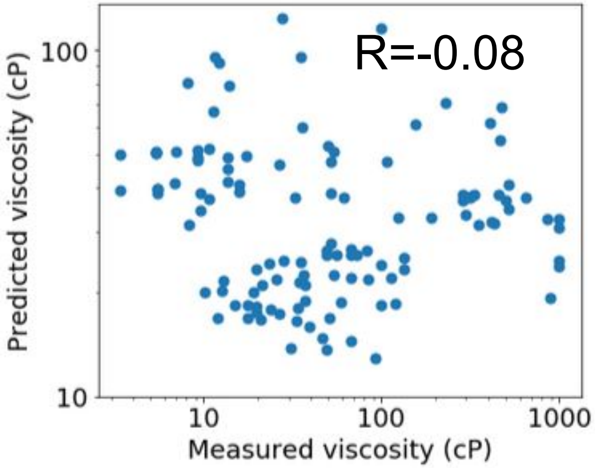
Using electrostatic potential surface map as the only input to the 3D-CNN prevents overfitting and enables these models to generalize

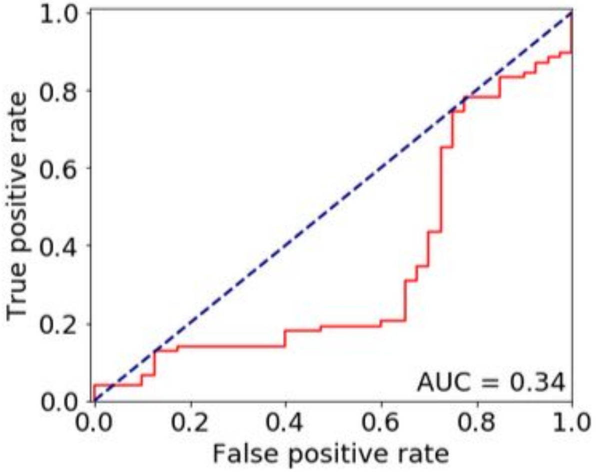
Generate features

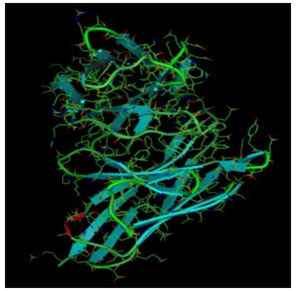
Pre-specified set of physchem properties

Traditional ML model failed to generalize

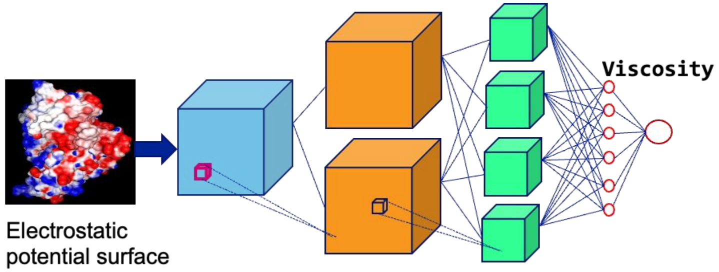


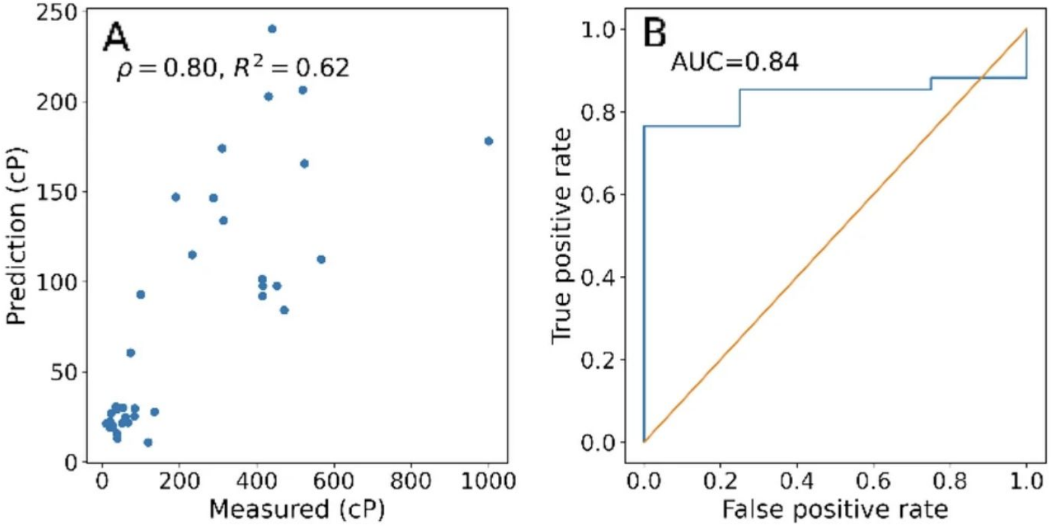






Homology model / crystal structure



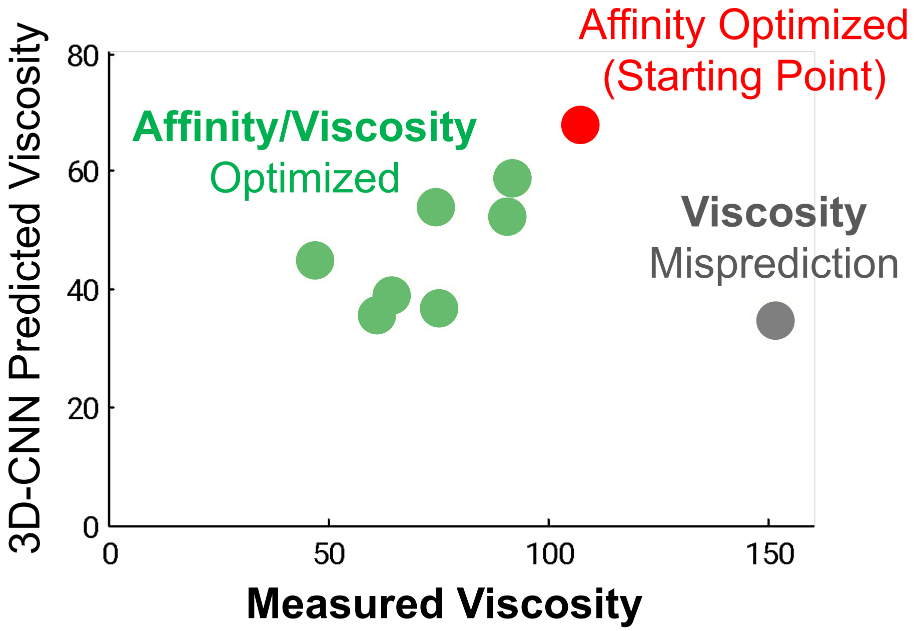


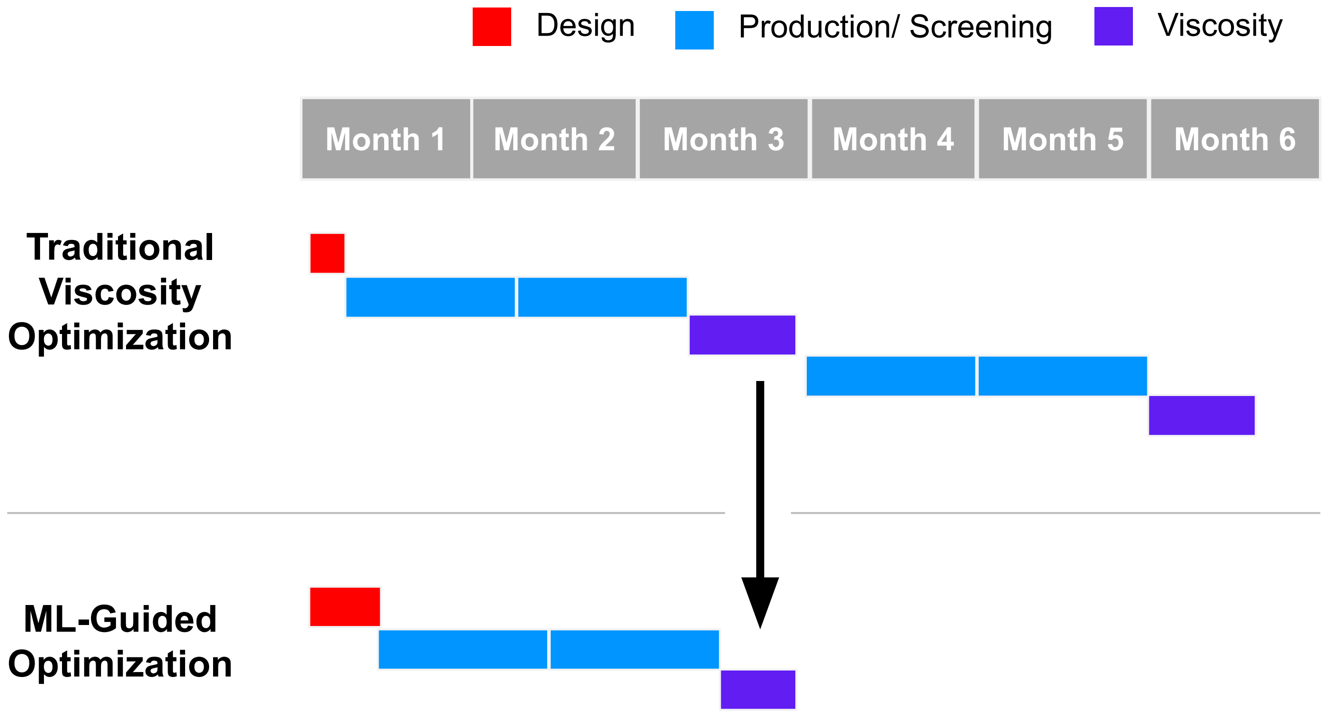
Worldwide Research, Development and Medical

Rai et al. (2023) https://doi.org/10.1038/s41598-023-28841-4

ML-Guided Antibody Viscosity Optimization was at Least 50% Faster*

AI prediction correlated strongly with measured viscosity of optimized mutants



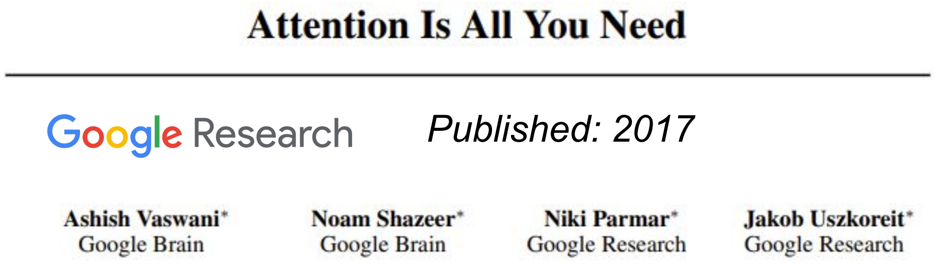


Prioritization of Antibody Mutants for Testing Eliminated Need for Multiple Production / Screening Cycles

*Compared to traditional approach

We are Leveraging Recent Advances in Language Modeling Techniques to Support Large Molecule Discovery Efforts

- Recent advances in AI can be attributed to one methodological breakthrough in deep learning: *Transformers*



Attention Is All You Need

Google Research *Published: 2017*

Ashish Vaswani* Google Brain

Noam Shazeer* Google Brain

Niki Parmar* Google Research

Jakob Uszkoreit* Google Research

<u>Prominent Transformer-based AI models</u>

- Generative Pre-trained Transformer (GPT)

OpenAI ChatGPT

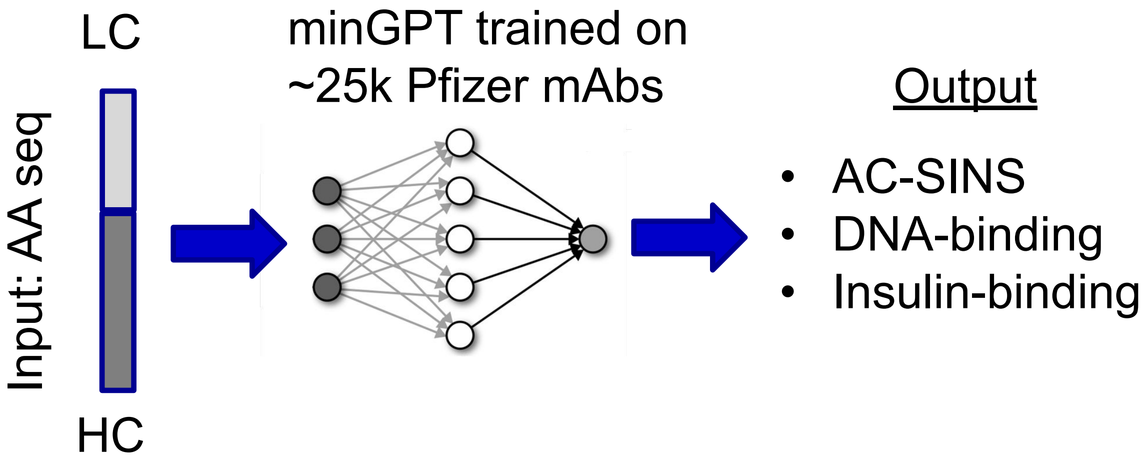
- AlphaFold2

We use Transformers in predictive modeling efforts for three tasks:

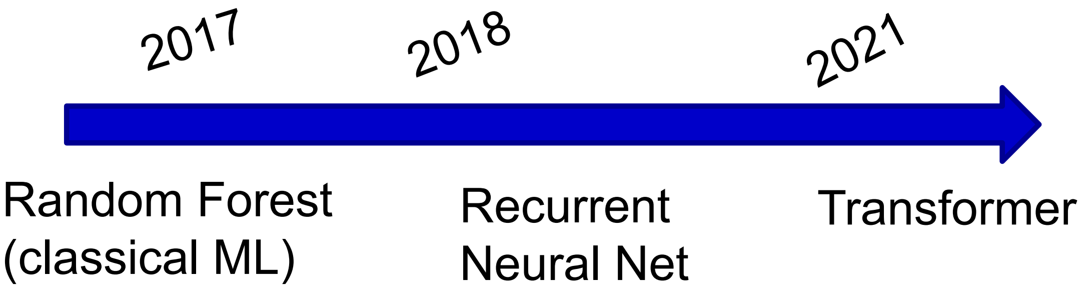
1. Antibody clearance
2. mRNA design
3. mAb immunogenicity risk assessment

Non-specificity Predictions from *minGPT*[*] based Models are being used to Reduce mAb PK Risk in the Early Discovery Stage

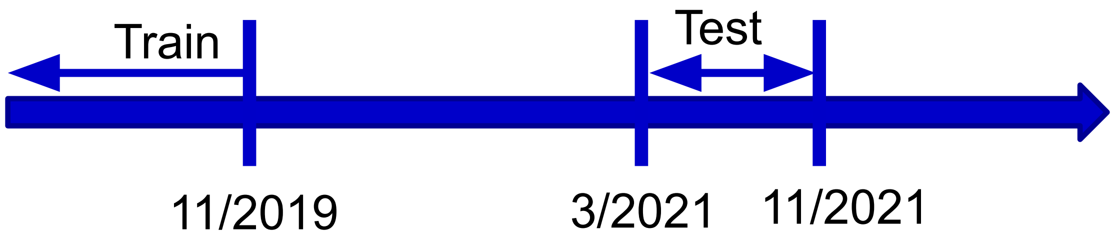
- *In vitro* non-specificity endpoints correlate well with *in vivo* clearance (Avery et al., *mAbs* 2018)



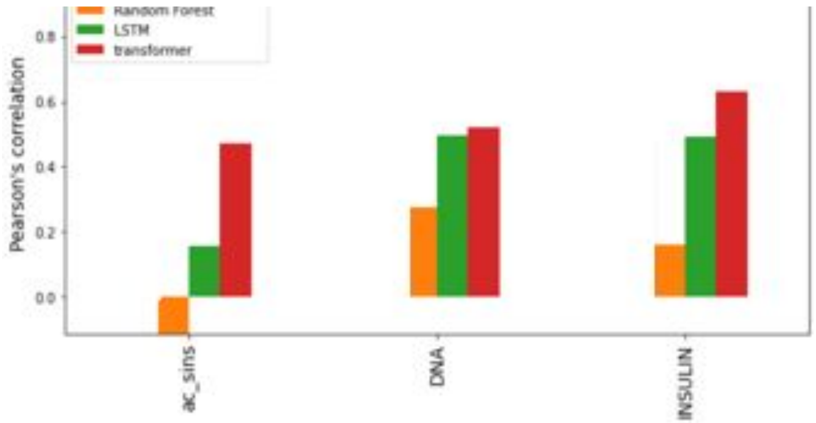
- Our choice of ML techniques over time



- Adoption of advanced ML techniques have led to better prediction performance



Test on antibodies from new portfolio projects

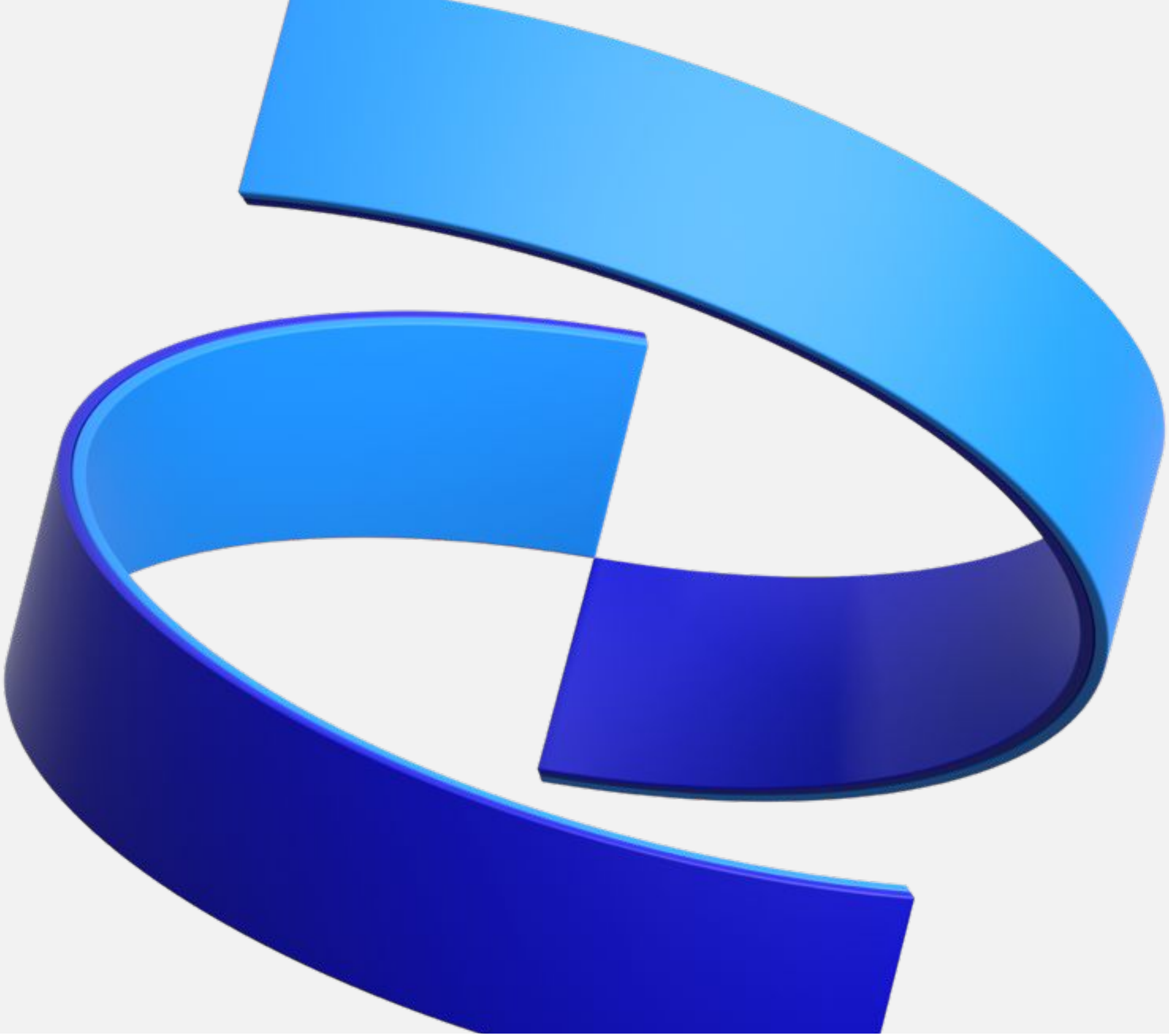


- Models have been integrated into the Pfizer developability assessment workflow

[*] An open-source implementation of GPT

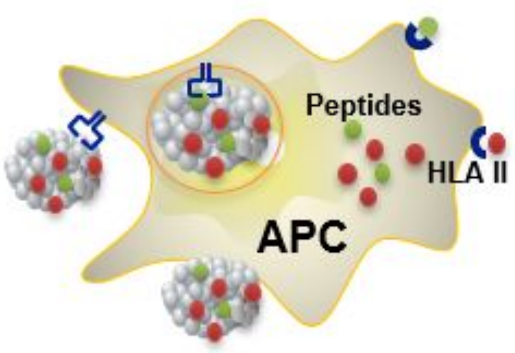
Thanks to my Pfizer colleagues

Chris Keefer
Daniel Wong
Brajesh Rai



Transformer Models have been Developed to Identify Potential Epitopes on Antibody Sequences

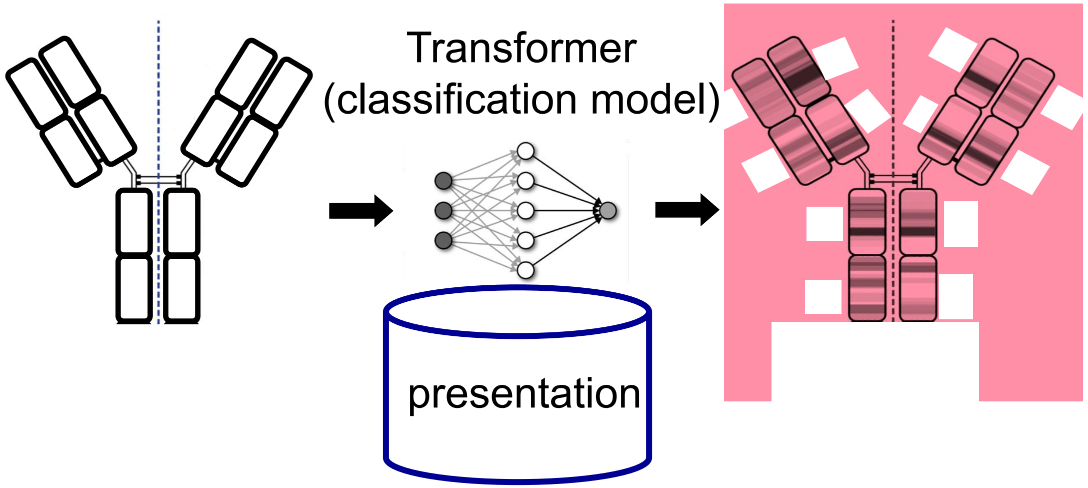
- Therapeutic antibodies run the risk of being recognized as foreign by a host immune system



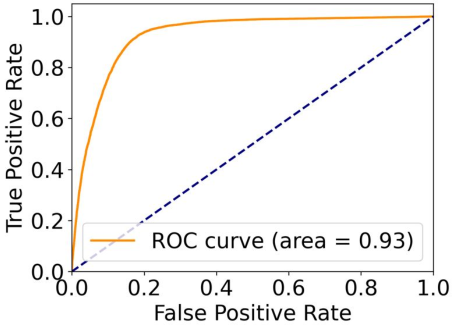
- Current immunogenicity risk assessment relies on peptide-HLA II binding predictions
    - Trained on *in vitro* binding affinities

- Recent publications have shown that peptide-HLA II presentation is a better predictor of immunogenicity
    - Trained on MS immunopeptidomic data
    - Chen, Nat Biotechnol 2019; MARIA, Stanford

Prediction of HLA II allele-specific presentation

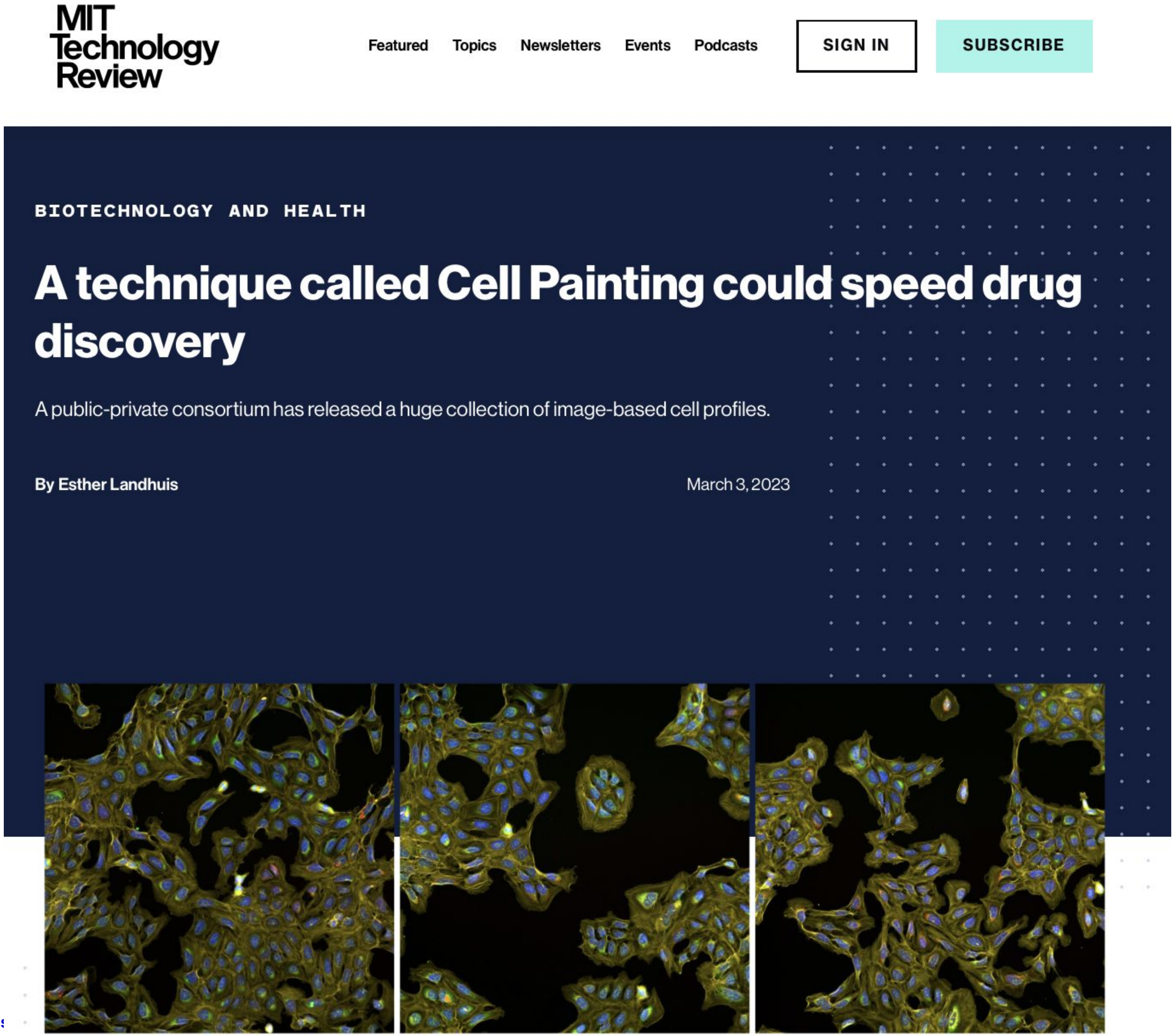


Retrieval of true HLA2 presented peptides



Epitope prediction accuracy (96 mAbs): 97%

Binding-based (current): 78%



BIOTECHNOLOGY AND HEALTH

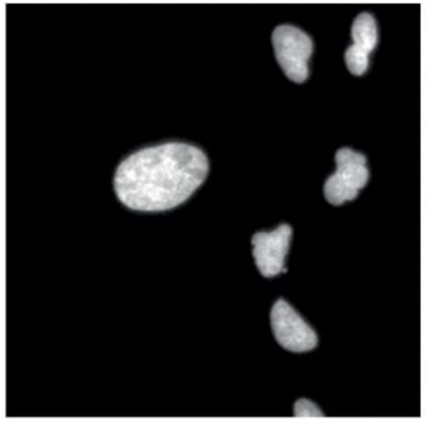
A technique called Cell Painting could speed drug discovery

A public-private consortium has released a huge collection of image-based cell profiles.

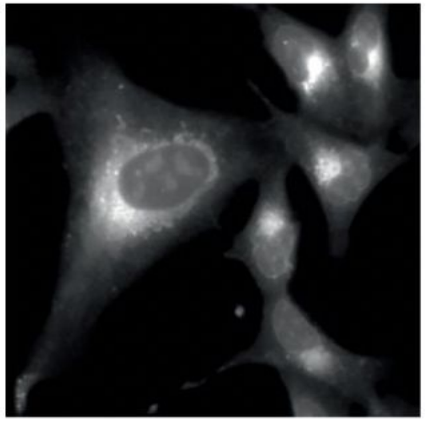
By Esther Landhuis

March 3, 2023

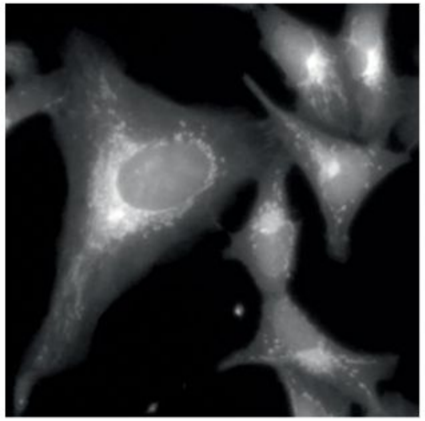
Cell Painting provides phenotypic information about key cellular structures



Hoechst
33342
Nucleus

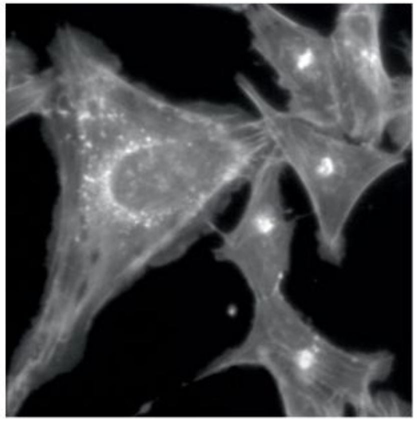


Alexa 488
Endoplasmic
Reticulum



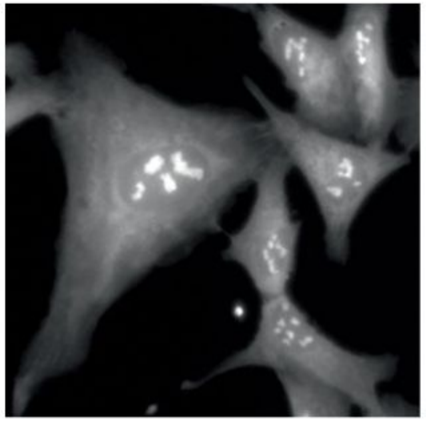
Alexa 647

Mitochondria

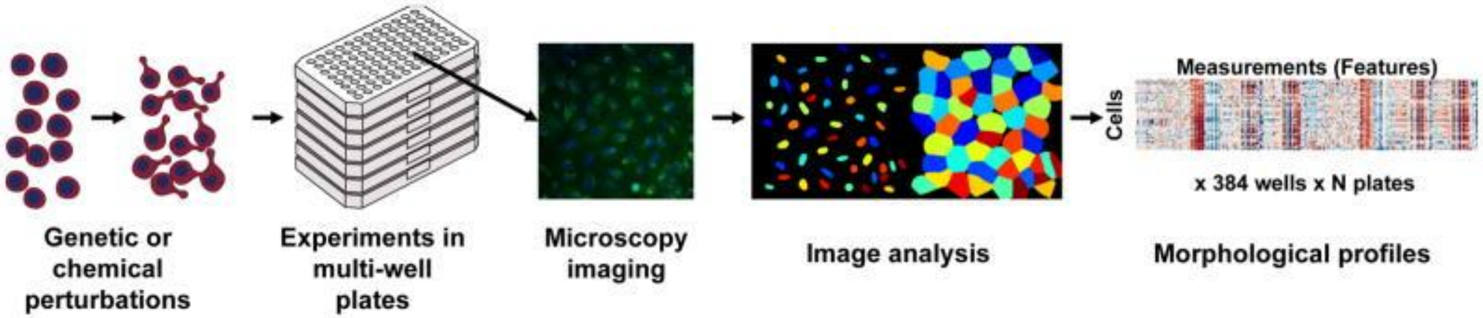


Alexa 568

Actin, Golgi Apparatus,
Plasma Membrane

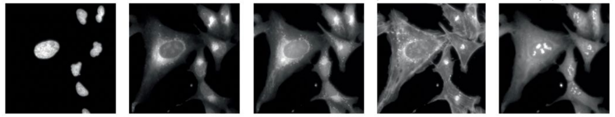


Alexa long
Cytoplasmic
RNA

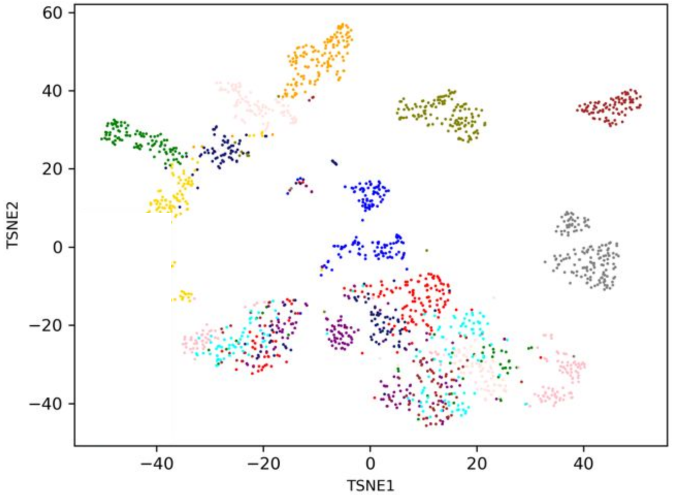


Chandrasekaran et al. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nature Reviews Drug Discovery (2020).
Bray et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat Protocol* (2016).

We aggregate individual image embeddings into well-level embeddings to compare with CellProfiler and DeepProfiler

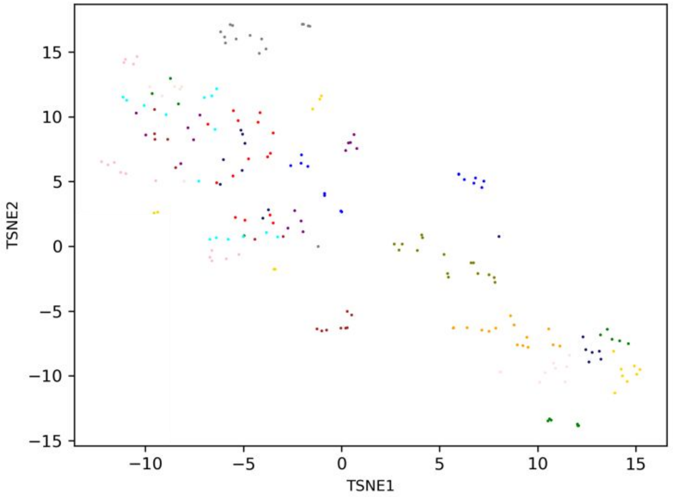


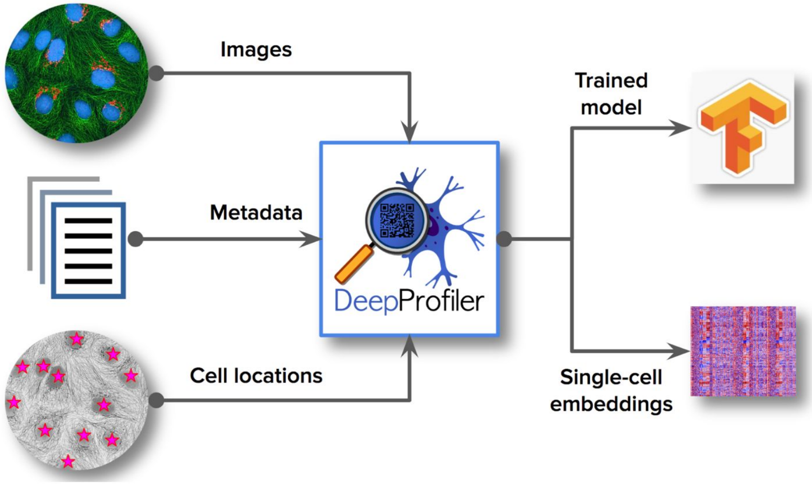
Encode image fields via neural network



Visualization of embeddings at the image level

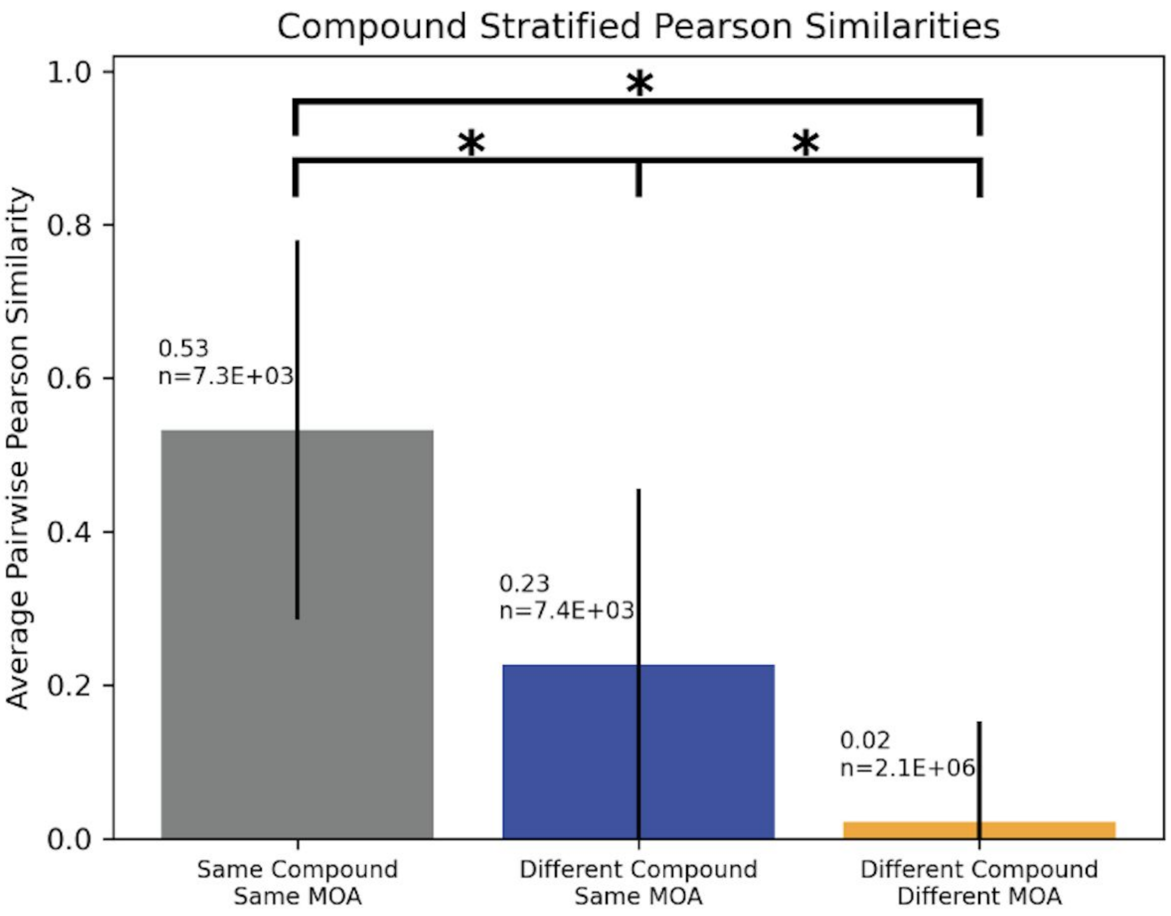
aggregate into well-level embeddings

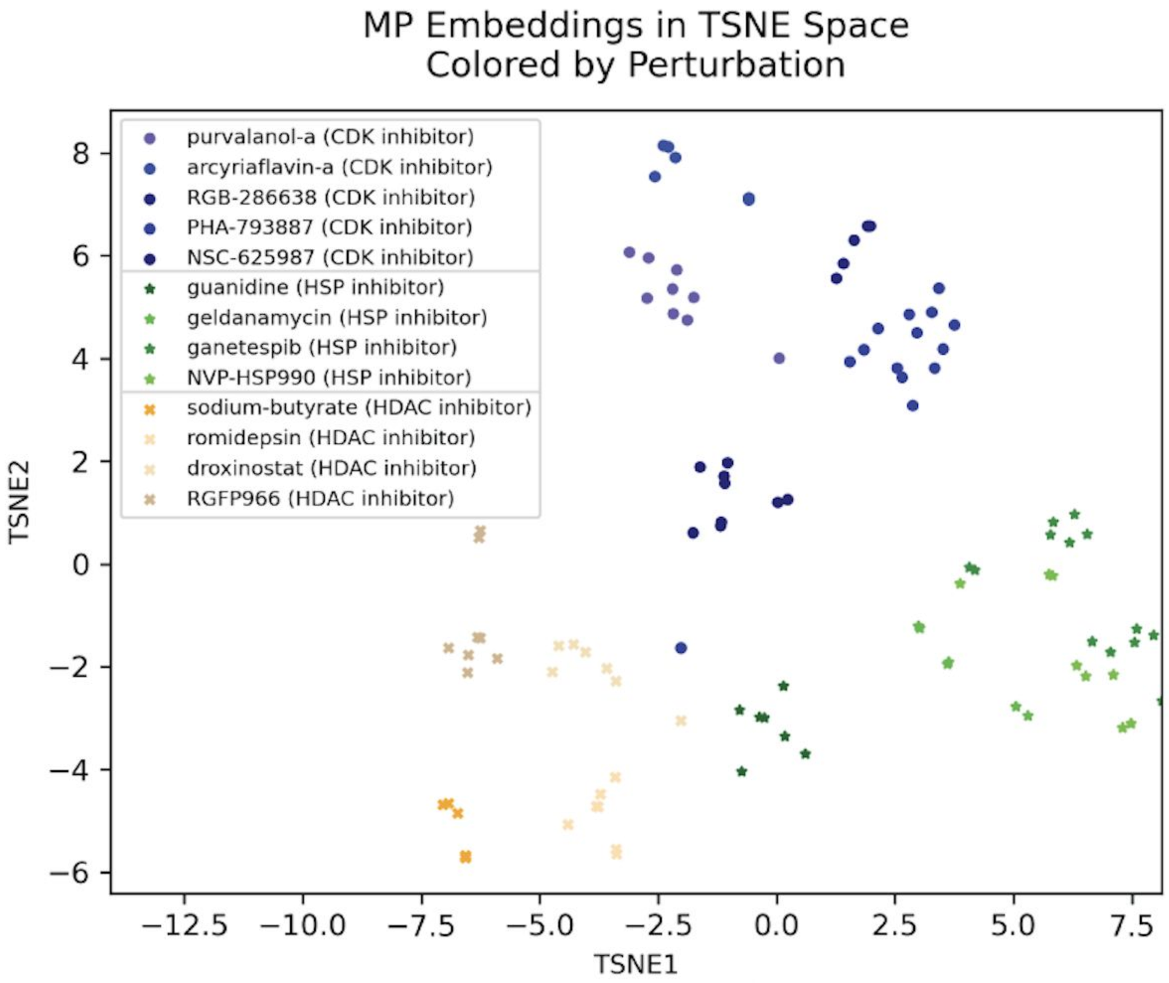




Carpenter et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biology (2006).
https://github.com/cytomining/DeepProfiler

Model learns MOA-specific phenotypes amidst a diverse compound space

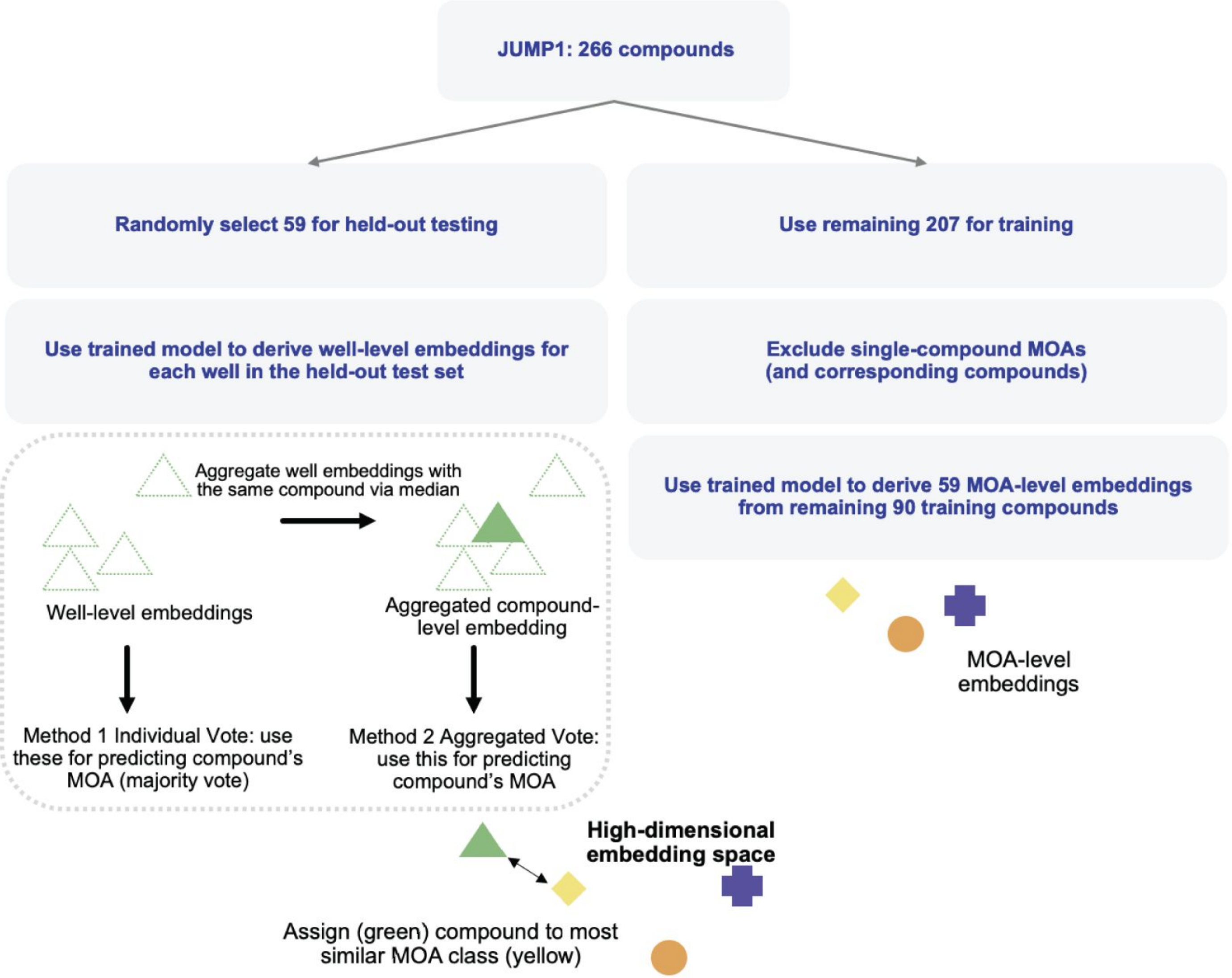




ical

Can we predict the MOA of held-out compounds never exposed to model training?

Same set up but hold out *compounds* instead of wells!



JUMP1: 266 compounds

Randomly select 59 for held-out testing

Use remaining 207 for training

Use trained model to derive well-level embeddings for each well in the held-out test set

Exclude single-compound MOAs (and corresponding compounds)

Use trained model to derive 59 MOA-level embeddings from remaining 90 training compounds

Aggregate well embeddings with the same compound via median

Well-level embeddings

Aggregated compound-level embedding

Method 1 Individual Vote: use these for predicting compound's MOA (majority vote)

Method 2 Aggregated Vote: use this for predicting compound's MOA

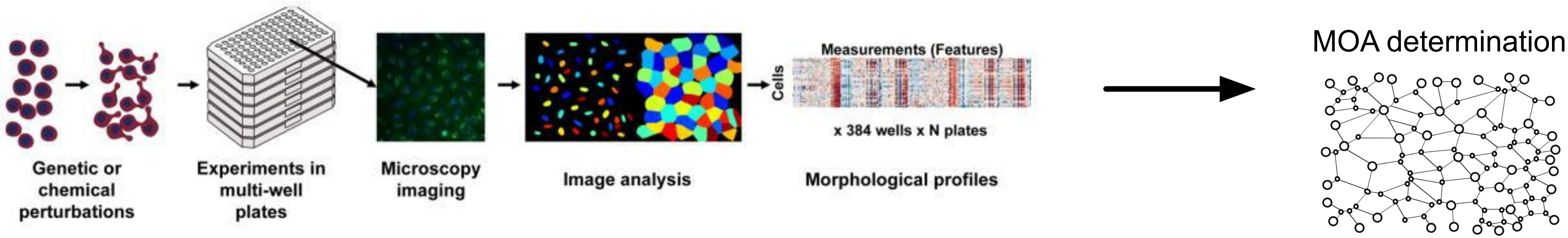
MOA-level embeddings

High-dimensional embedding space

Assign (green) compound to most similar MOA class (yellow)

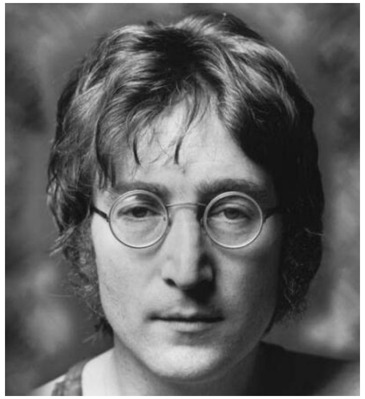
Conclusions

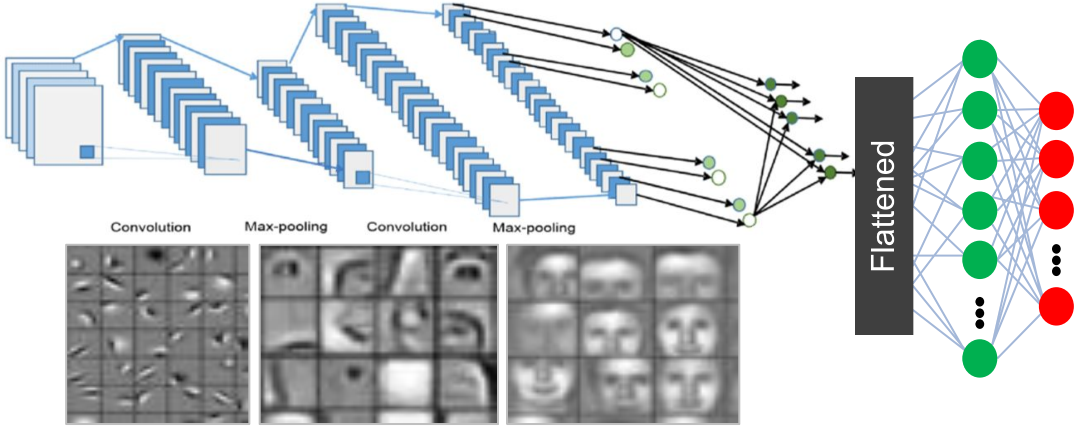
- Drug MOA determination via CellPainting phenotypes is possible!
- MOAProfiler is more performant for MOA determination than both the gold-standard CellProfiler and DeepProfiler
- Approach generalizes to two different datasets and predicts the MOA of *held-out* compounds



PfAbNet: Pfizer's internally developed 3D CNN that predicts antibody viscosity

Traditional 2D CNN





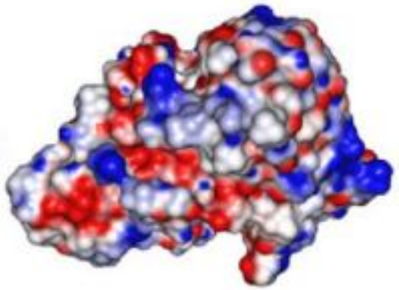
"John Lennon"

INPUT

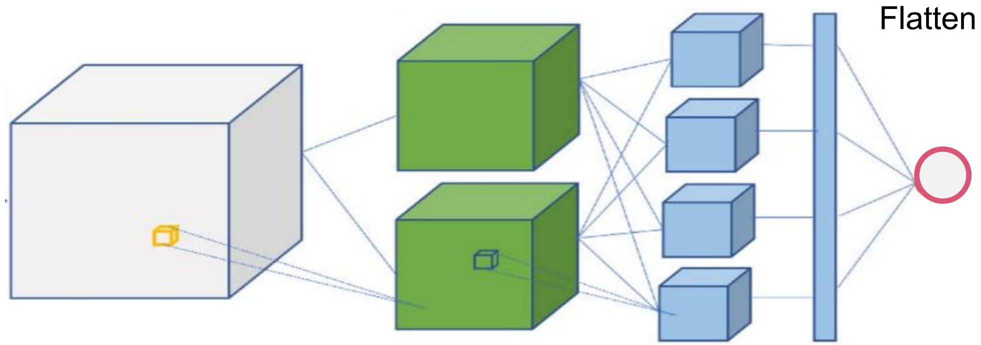
FEATURE EXTRACTION

OUTPUT

PfAbNet



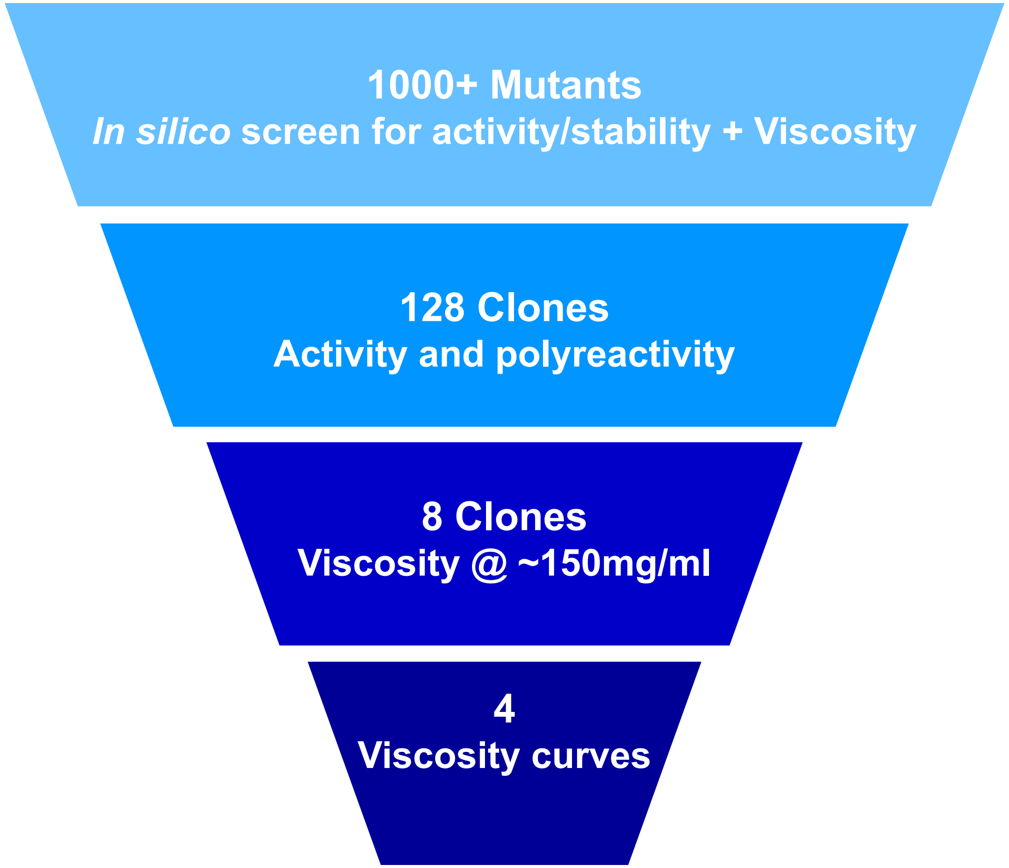
Electrostatic Potential



Viscosity Prediction

Rai et al. (2023) https://doi.org/10.1038/s41598-023-28841-4

AI Guided Optimization Delivering Antibodies in Less than Half the Time



1000+ Mutants
In silico screen for activity/stability + Viscosity
128 Clones
Activity and polyreactivity
8 Clones
Viscosity @ ~150mg/ml
4
Viscosity curves

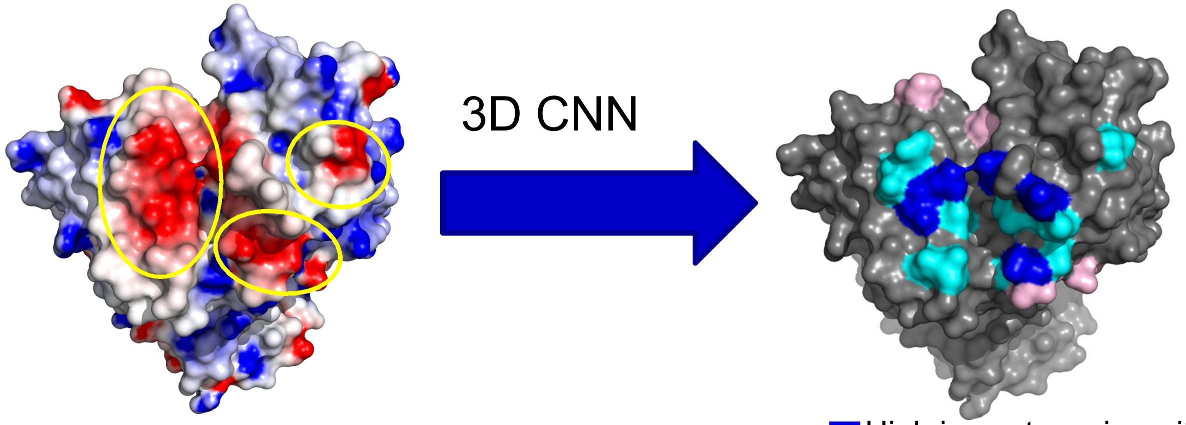
FINAL LEAD

INPUT

Targeted 3 charge patches for mutants

OUTPUT

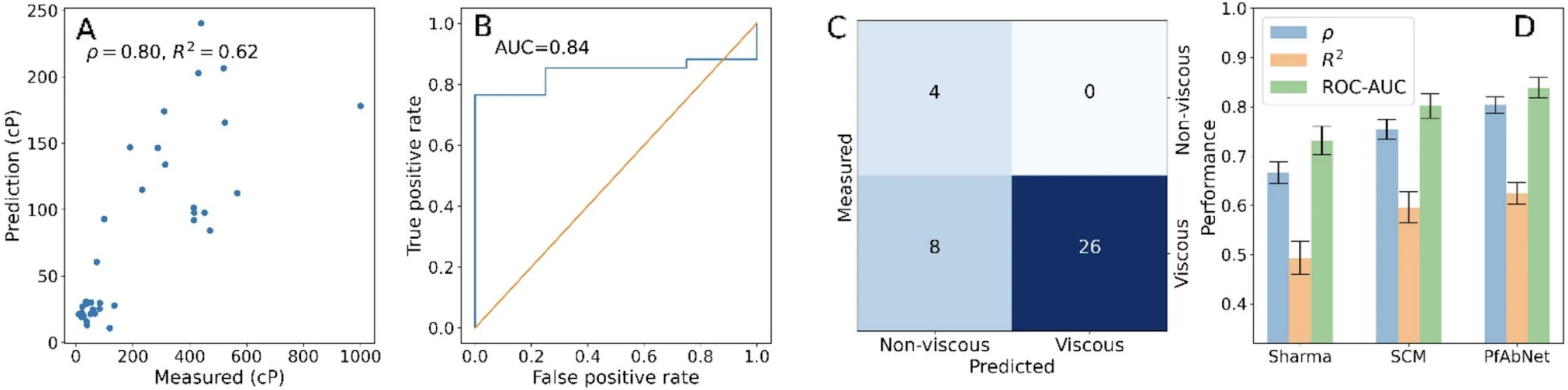
CNN
Viscosity Prediction

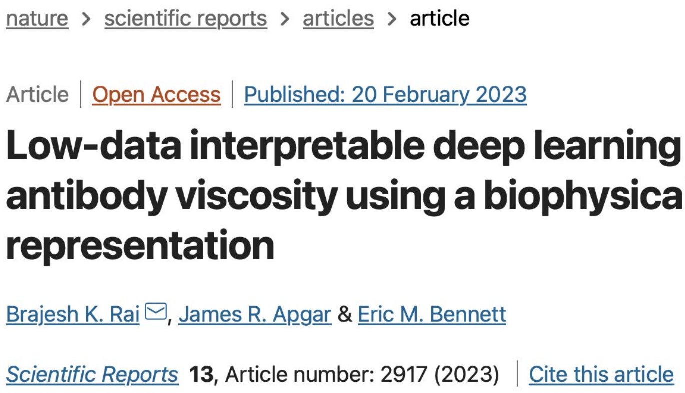


🟦 High impact on viscosity
🟩 Med. impact on viscosity
🟪 Low impact on viscosity

3D-CNN model outperforms previous methods[14,15] in viscosity prediction

From: Low-data interpretable deep learning prediction of antibody viscosity using a biophysically meaningful representation



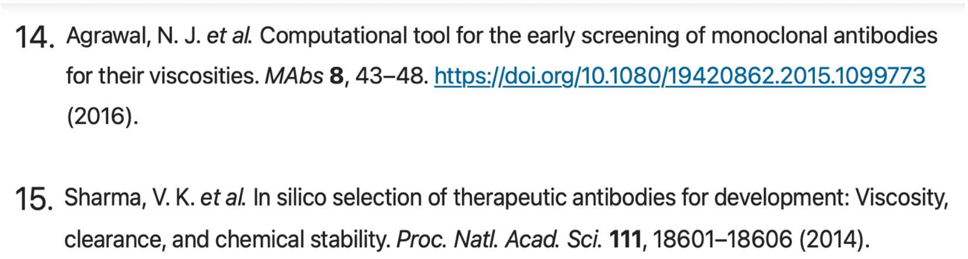

nature > scientific reports > articles > article

Article | Open Access | Published: 20 February 2023

Low-data interpretable deep learning prediction of antibody viscosity using a biophysically meaningful representation

Brajesh K. Rai, James R. Apgar & Eric M. Bennett

Scientific Reports **13**, Article number: 2917 (2023) | Cite this article


14. Agrawal, N. J. *et al.* Computational tool for the early screening of monoclonal antibodies for their viscosities. *MAbs* **8**, 43–48. https://doi.org/10.1080/19420862.2015.1099773 (2016).

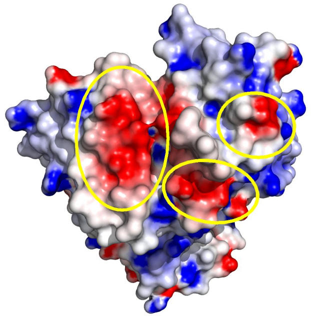
15. Sharma, V. K. *et al.* In silico selection of therapeutic antibodies for development: Viscosity, clearance, and chemical stability. *Proc. Natl. Acad. Sci.* **111**, 18601–18606 (2014).

Application of PfAbNet to a Trispecific Antibody

AI prediction correlated strongly with measured viscosity of optimized mutants

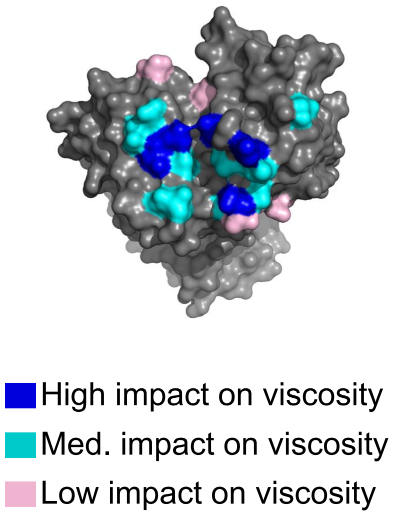
INPUT

Targeted 3 charge patches for mutants



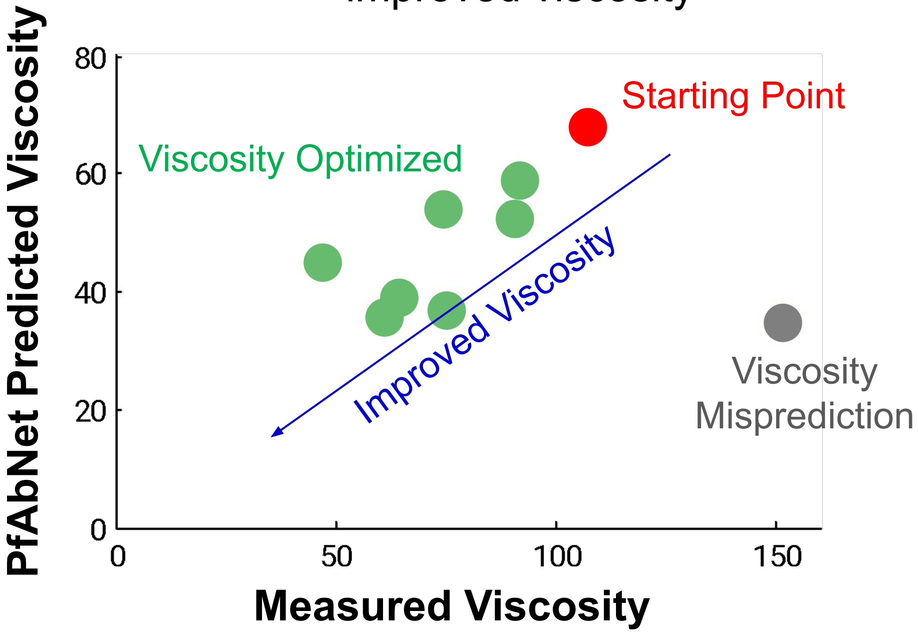
OUTPUT

PfAbNet Viscosity Prediction



VALIDATION:

One round of design delivered improved viscosity


PfAbNet Predicted Viscosity
Measured Viscosity
Starting Point
Viscosity Optimized
Improved Viscosity
Viscosity Misprediction