Applying Data Science for Environment and Health Assessment

Dr. Sean Ekins, Ph.D., D.Sc.

CEO,

Collaborations Pharmaceuticals, Inc.

COLLABORATIONS PHARMACEUTICALS, INC.

AI is increasingly in the news

Dec 26, 2020, 04:59pm EST | 2,894 views

The Increasing Use Of AI In The Pharmaceutical Industry

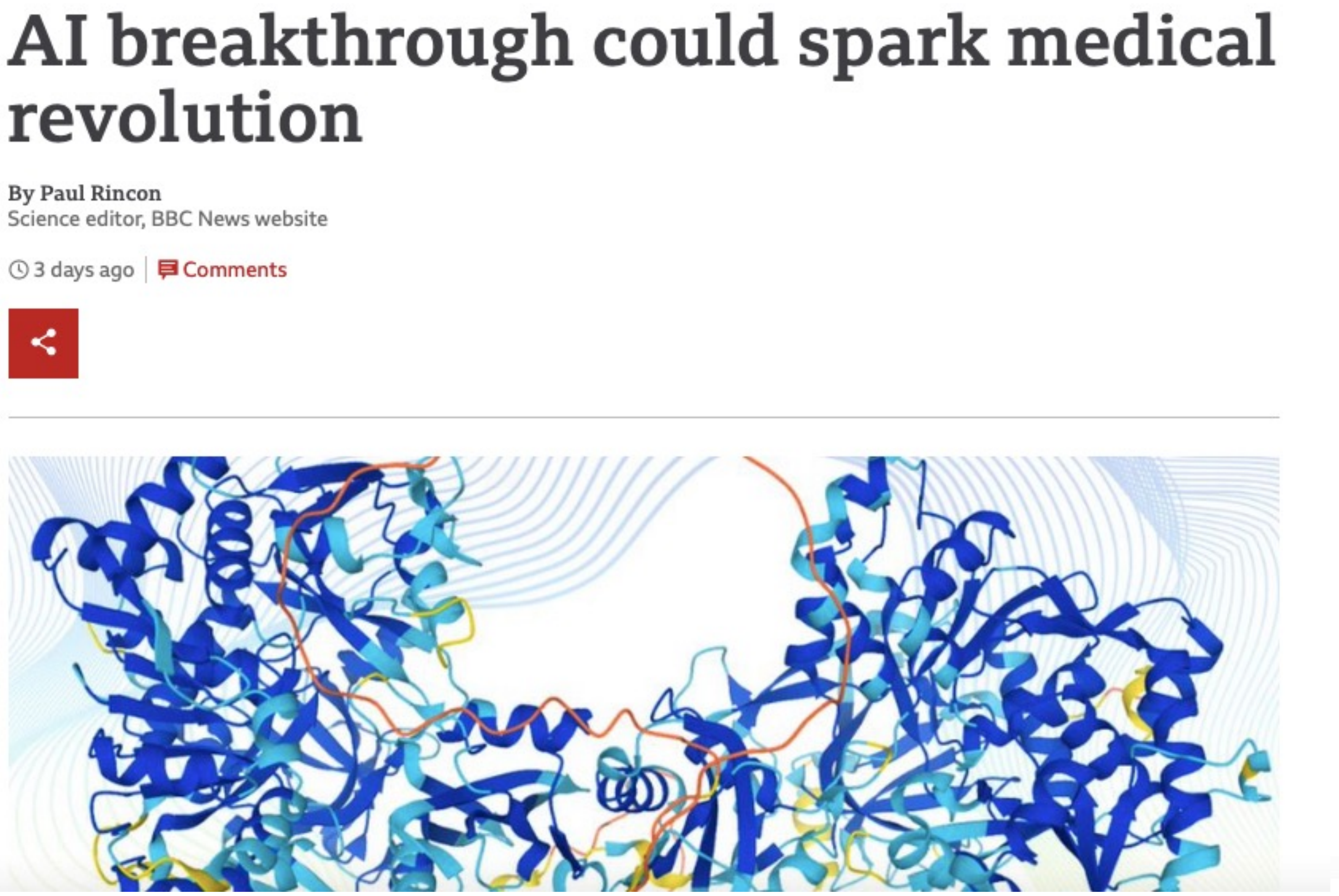
Forbes

Kathleen Walch Contributor
COGNITIVE WORLD Contributor Group
AI

PHARMA INDUSTRY HAS A NEW DRUG: ARTIFICIAL INTELLIGENCE

Analytics Insight

ARTIFICIAL INTELLIGENCE HEALTHCARE LATEST NEWS
by Priya Dialani / December 11, 2020 / 0 comments

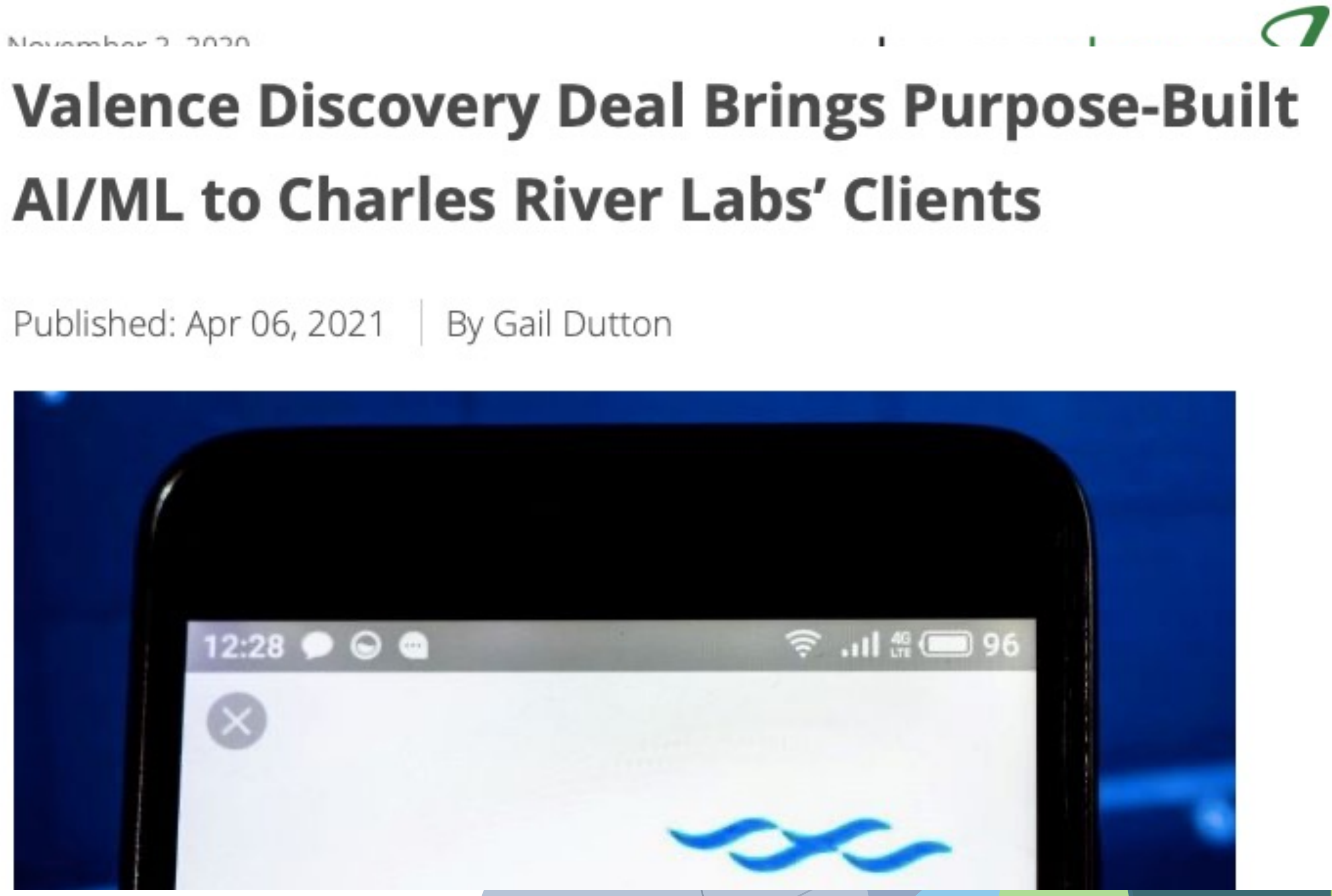

AI breakthrough could spark medical revolution

By Paul Rincon
Science editor, BBC News website

3 days ago | Comments

AI's infiltration of pharma: How COVID-19 accelerated change

November 2, 2020


Valence Discovery Deal Brings Purpose-Built AI/ML to Charles River Labs' Clients

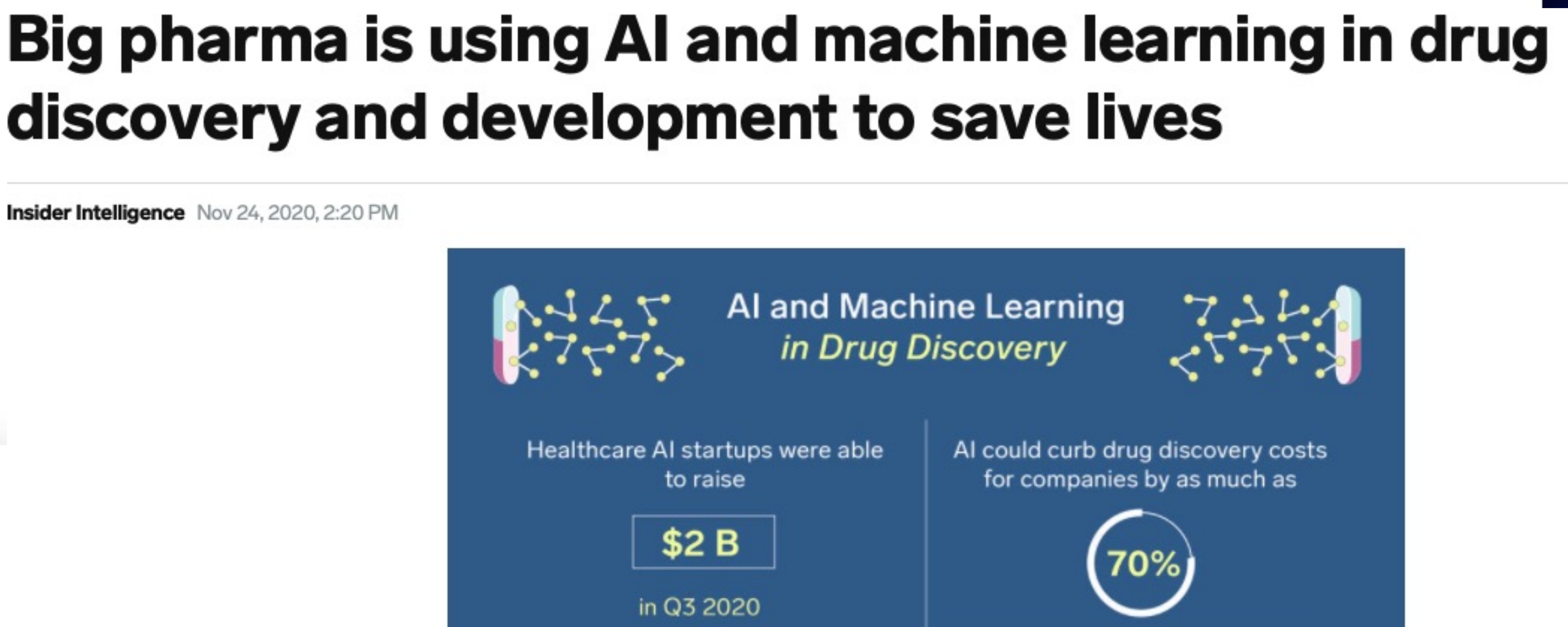
Published: Apr 06, 2021 | By Gail Dutton

thepharmaletter
* Up to date news for the Pharmaceutical and Biotechnology industries

HOME M&A NEWS ▾ INSIGHTS ▾ PRICING, POLICY, REGULATION ▾ THERAPY AREAS ▾

YOU ARE HERE HOME > PHARMACEUTICAL

The FDA and artificial intelligence


Big pharma is using AI and machine learning in drug discovery and development to save lives

Insider Intelligence Nov 24, 2020, 2:20 PM

AI and Machine Learning *in Drug Discovery*

Healthcare AI startups were able to raise $2 B in Q3 2020

AI could curb drug discovery costs for companies by as much as 70%

ARTIFICIAL INTELLIGENCE, BIOPHARMA

AI offers promise but faces barriers in drug development

Inertia is a barrier as is the traditional split between the clinical a the data-driven spheres of drug development. While smaller firm have an edge in bridging the gap, big pharma will eventually ge there, said panelists at the INVEST conference session.

By JOEL BERG

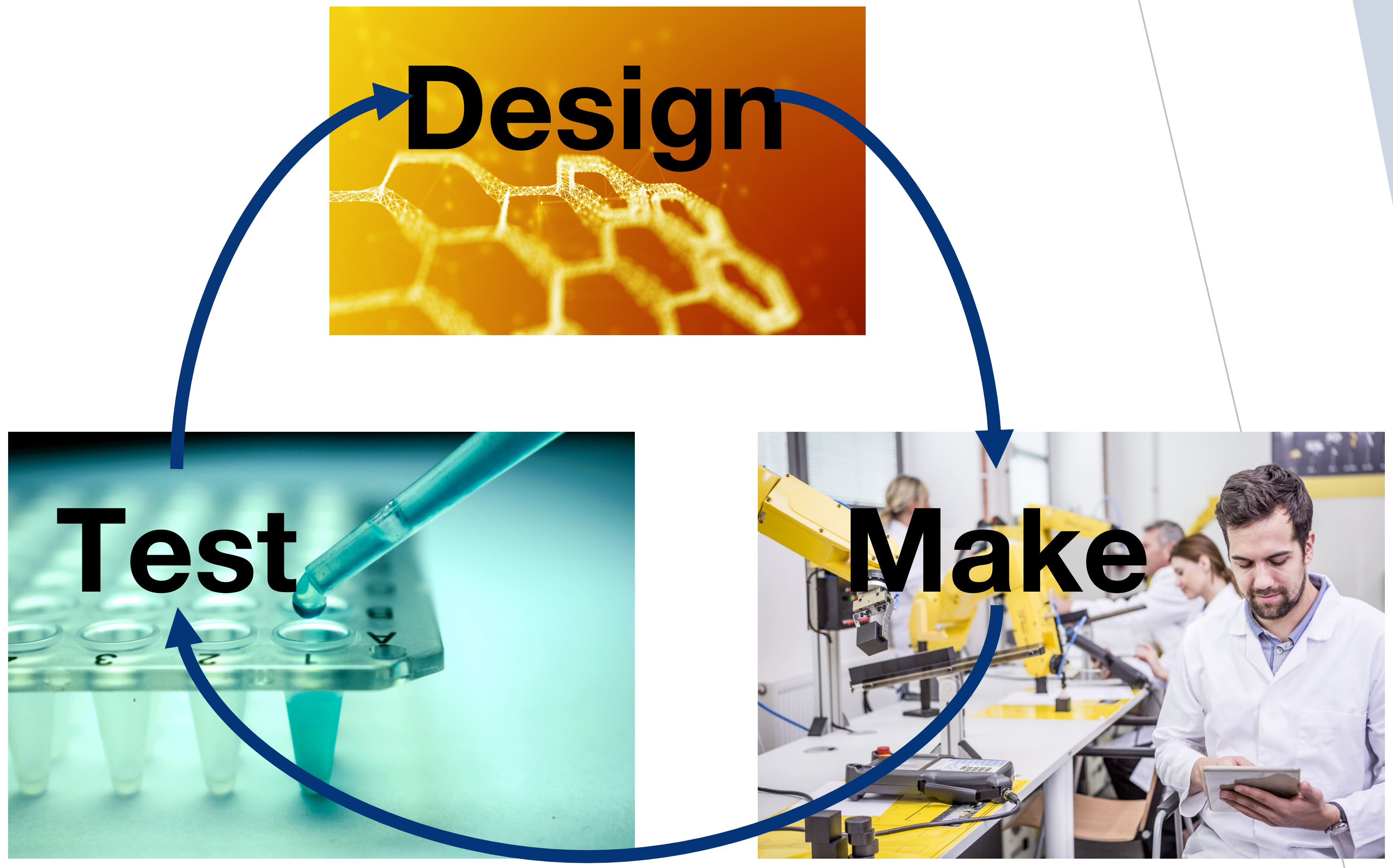
MedCityNews

Post a comment / Dec 10, 2020 at 12:54 PM

AI In Industry

COLLABORATIONS PHARMACEUTICALS, INC.

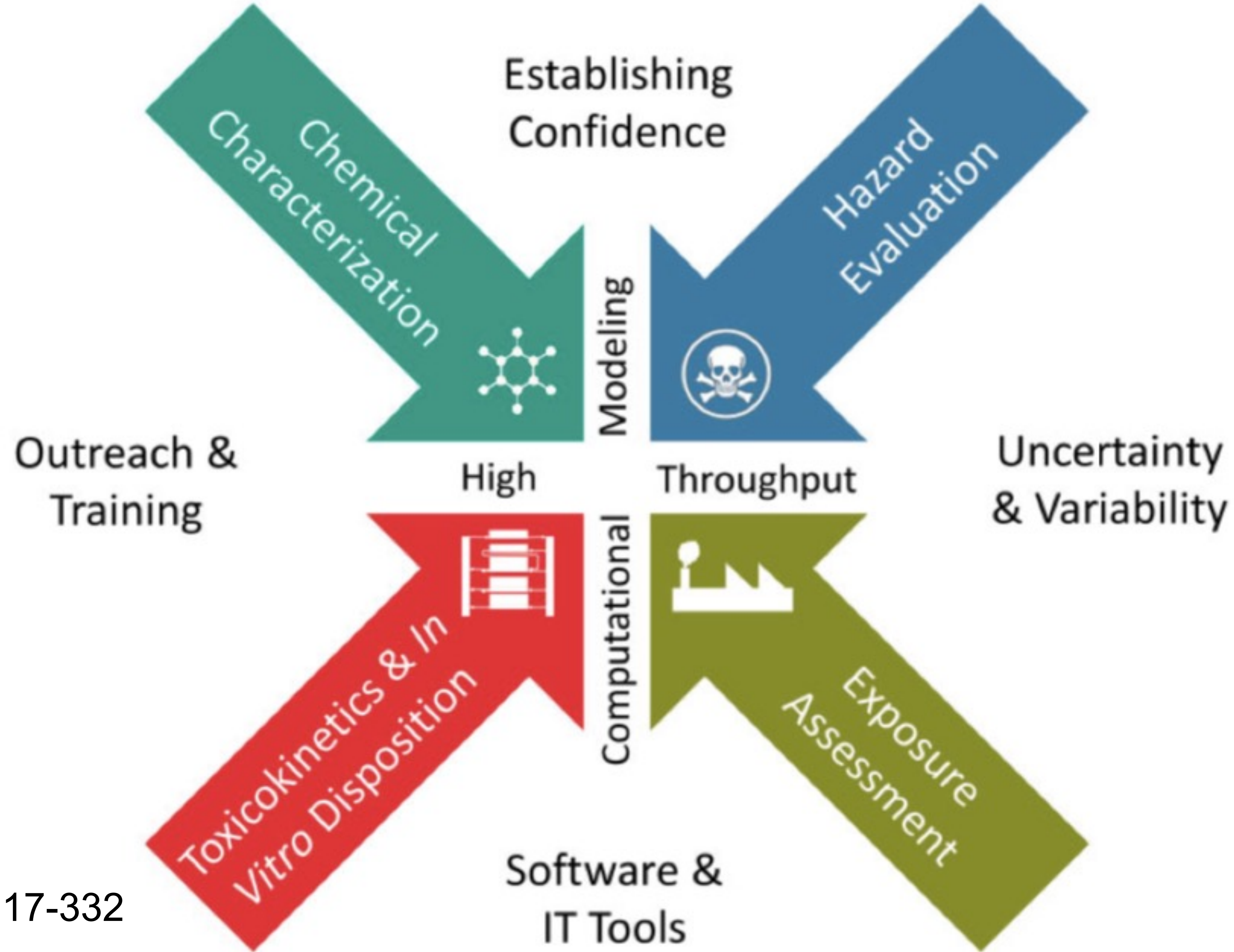
- Pharmaceutical
  - Design new molecules
  - Repurpose drugs
  - Predict Toxicity & Drug-drug interactions
- Consumer products
  - Cleaning – prioritize endocrine disruption
  - Cosmetics – non-animal testing options
  - Environmental impact
- Agrochemical
  - Biodegradation
  - Toxicity to non-target species
- Environmental
  - Predict impact of chemicals
- Animal health
  - Cost-effectively develop new treatments



Design

Make

Test



Establishing Confidence

Chemical Characterization

Hazard Evaluation

Modeling

High

Throughput

Computational

Outreach & Training

Uncertainty & Variability

Toxicokinetics & In Vitro Disposition

Exposure Assessment

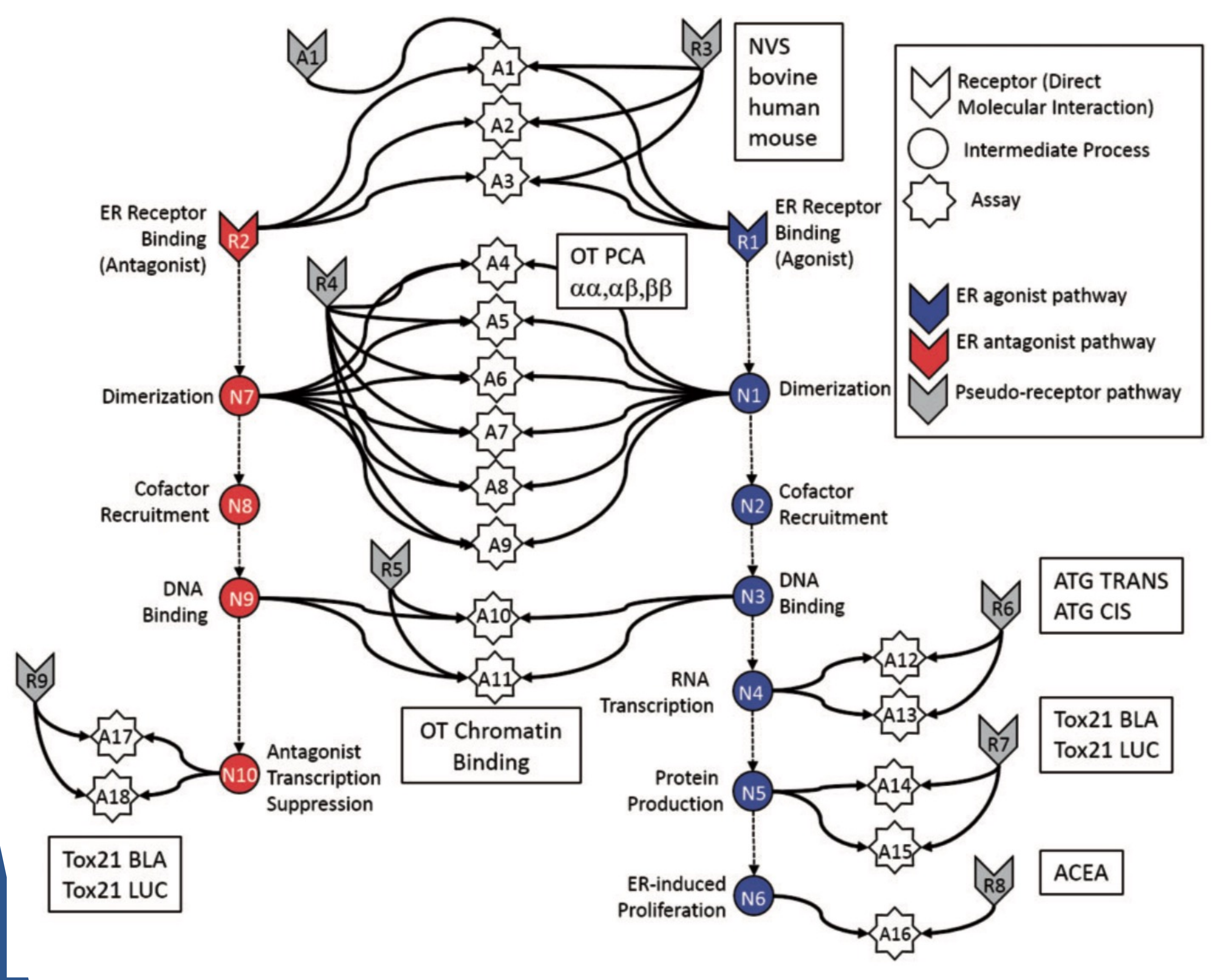
Software & IT Tools

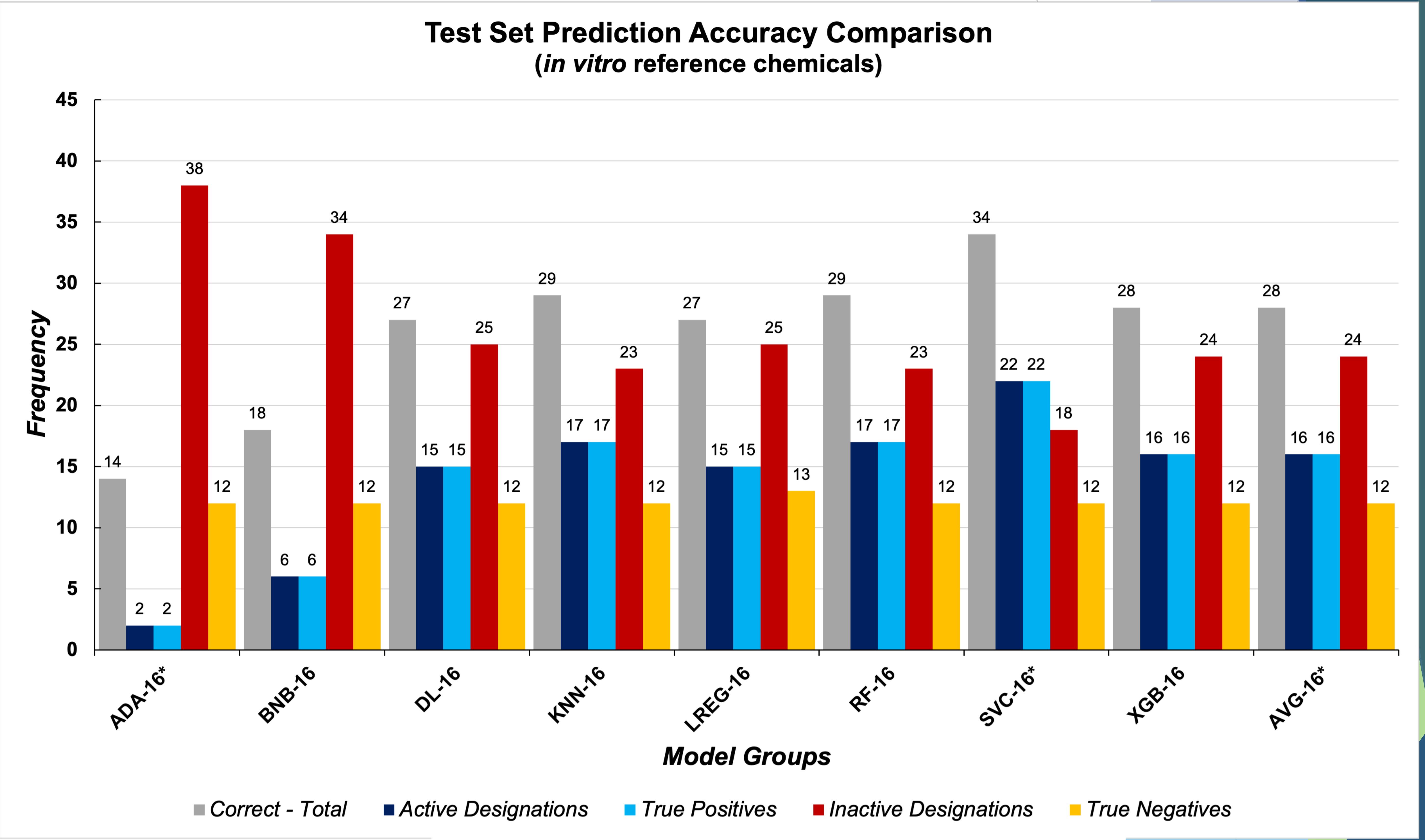
Toxicol Sci. 2019;169: 317-332

Modeling Targets in Pathways

Collaborations Pharmaceuticals, Inc.

- 16 assays for ER with 1000s molecules published by the EPA
- Evaluate algorithm performances and identify which is best-suited for predicting ER agonism
- *in vitro* reference chemicals – 40 total, 28 active/12 inactive



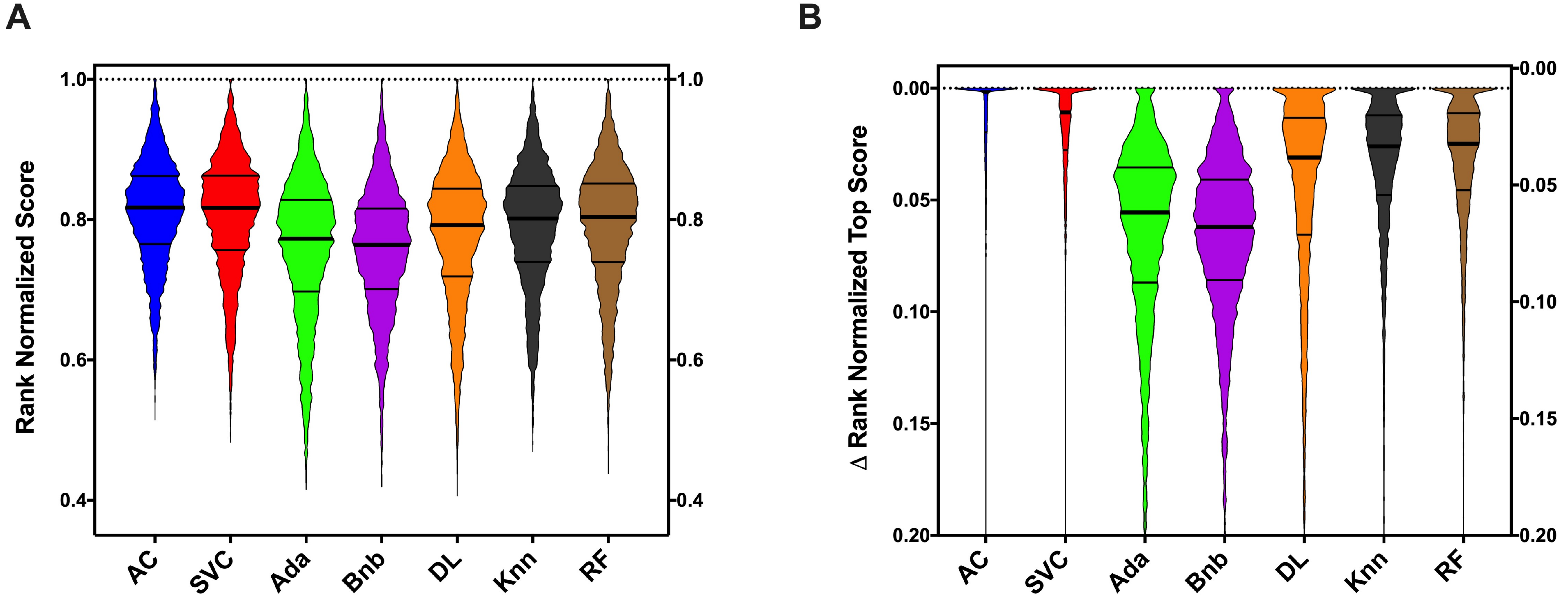


Model Groups	Correct - Total	Active Designations	True Positives	Inactive Designations	True Negatives
ADA-16*	14	2	2	38	12
BNB-16	18	6	6	34	12
DL-16	27	15	15	25	12
KNN-16	29	17	17	23	12
LREG-16	27	15	15	25	13
RF-16	29	17	17	23	12
SVC-16*	34	22	22	18	12
XGB-16	28	16	16	24	12
AVG-16*	28	16	16	24	12

Toxicol Sci. 2015;148(1):137-154.
Environ Sci Technol. 2015;49:8804−8814

Mol Pharmaceutics. 2018;15:4361-4370.
Environ Sci Technol. 2020;54(19):12202-12213.
Environ Sci Technol. 2020;54(21):13690-13700
Environ Sci Technol. 2020;54(23):15546-15555

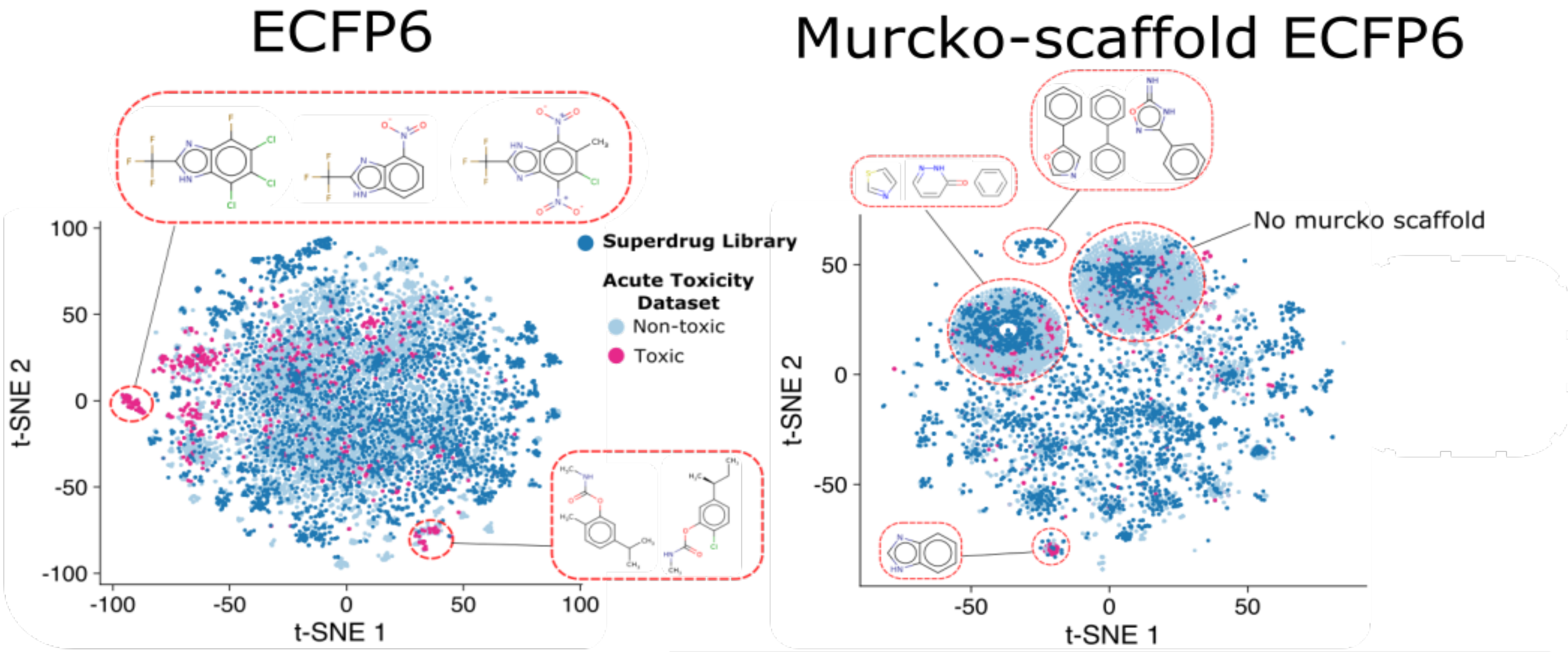
Model Human Diseases and Biology / Targets

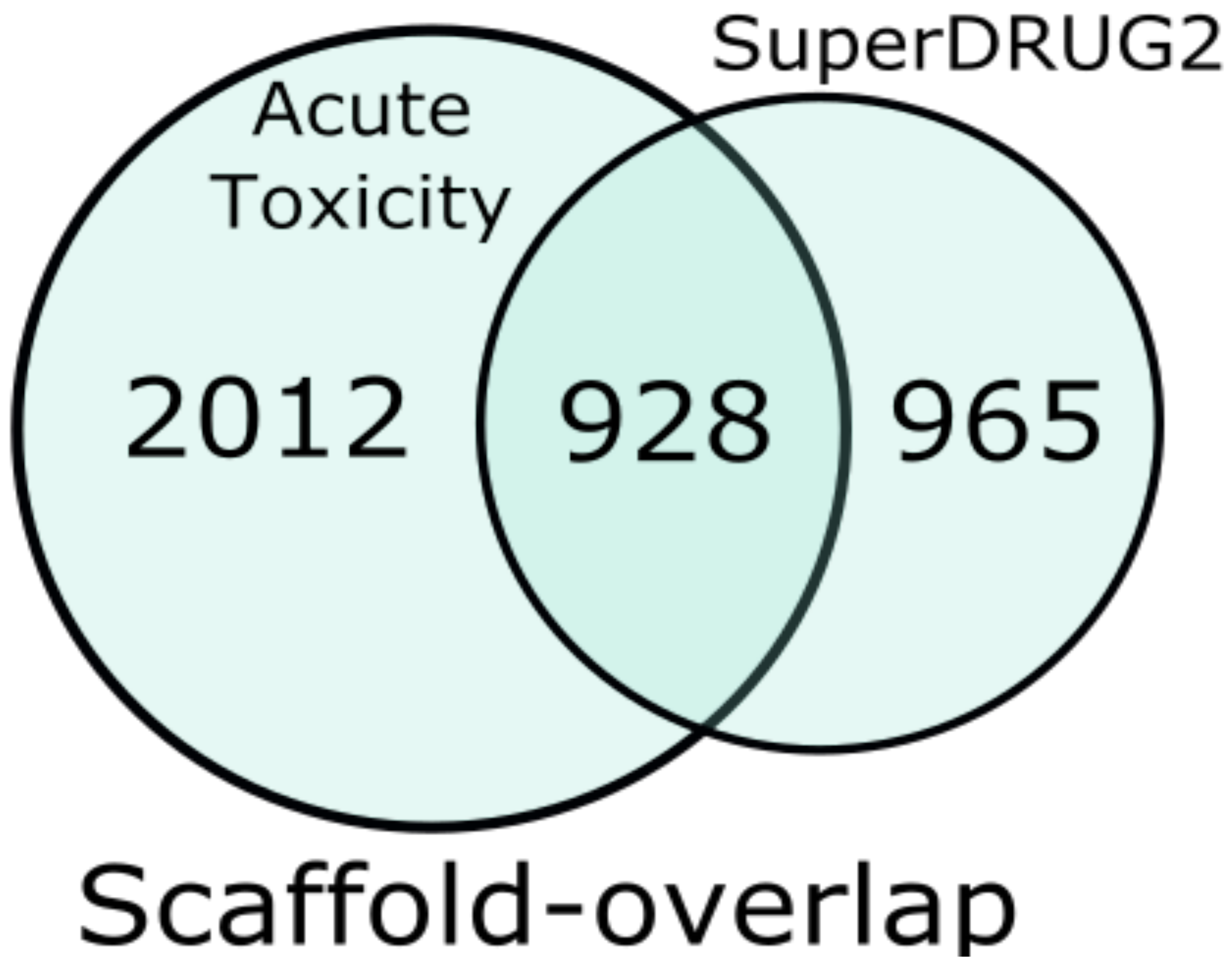


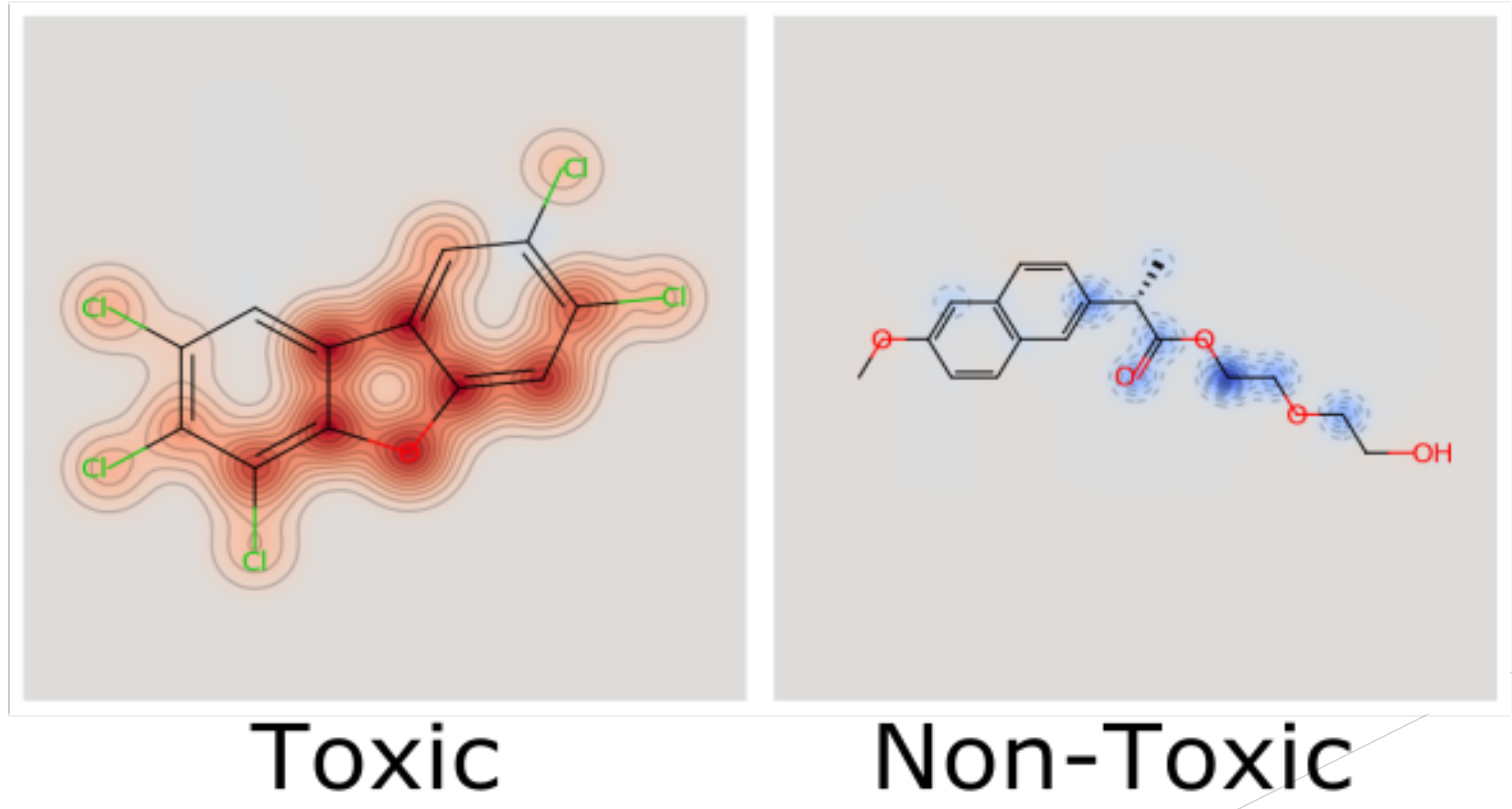
Lane et al., Mol Pharm. 2021;18(1)403-415.

- >5000 ChEMBL datasets, >100 compounds in each
- Compared support vector classification, AdaBoosted decision trees, multiple Bayesian methods, deep learning, K nearest-neighbors, and random forests
- Assessed five-fold cross-validation statistics
- External testing on various ADME/Tox datasets
- www.assaycentral.org

Exploring Toxicity Property Space

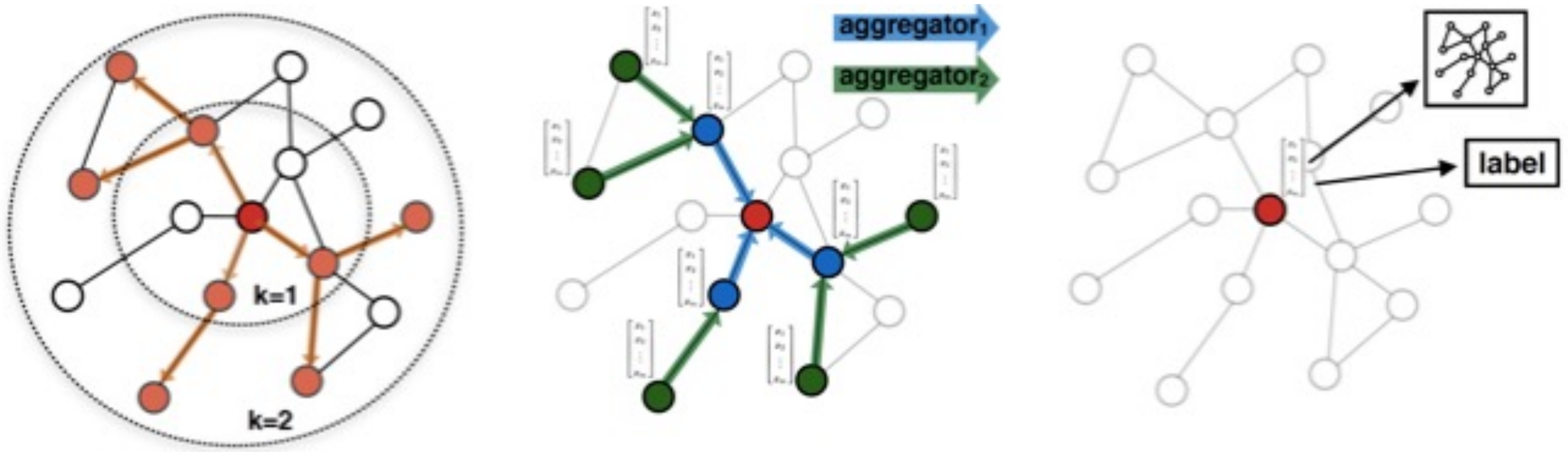






Environ Health Perspect. 2021 Apr;129(4):47013
ACS Sustainable Chem. Eng. 2020, 8, 42, 16020–16027

Model Protein Networks With Graphs



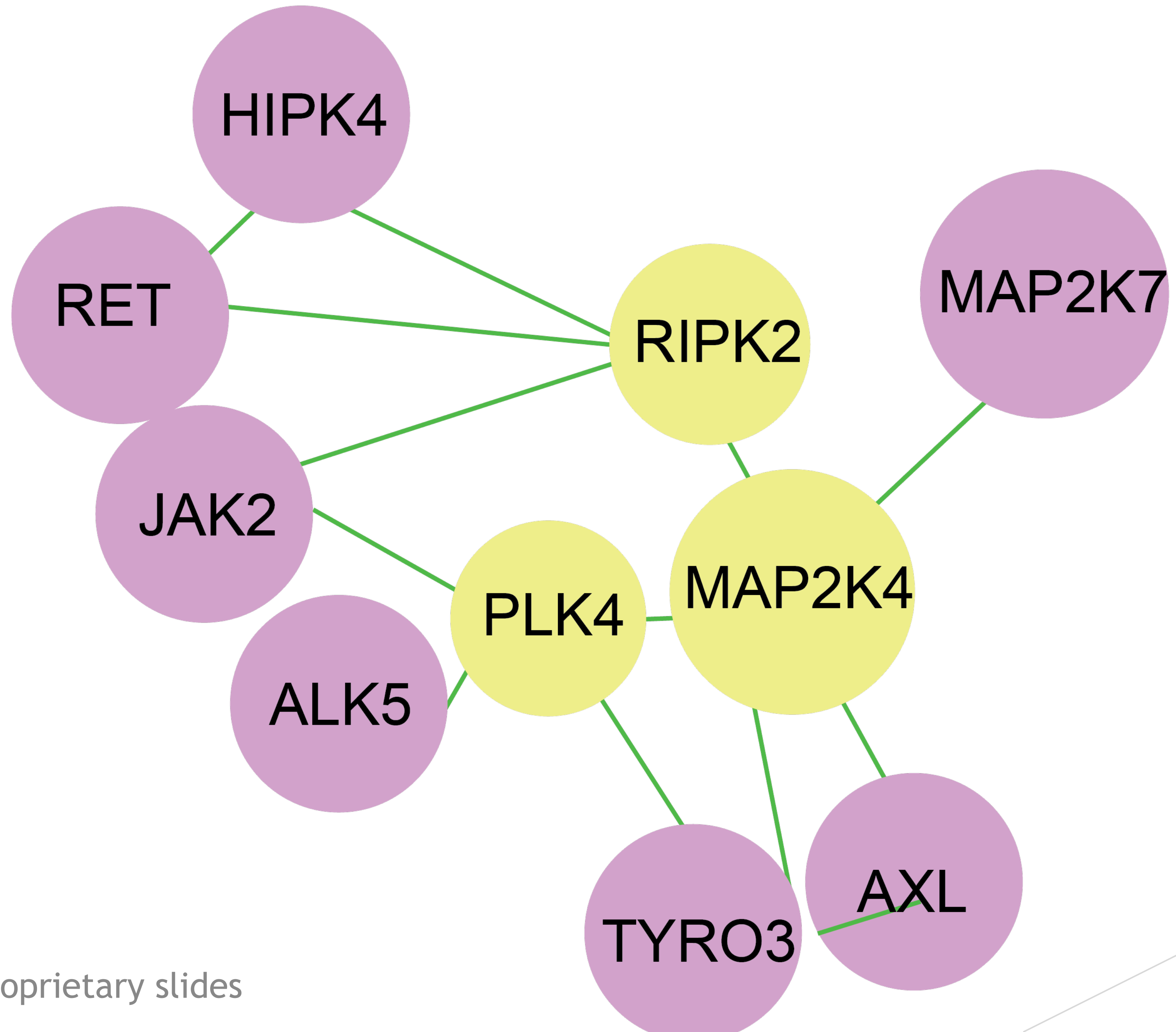
1. Sample neighborhood

2. Aggregate feature information from neighbors

3. Predict graph context and label using aggregated information

Inductive Representation Learning on Large Graphs. W.L. Hamilton, R. Ying, and J. Leskovec *arXiv:1706.02216 [cs.SI]*, 2017.

- By including drug-target interactions along with target kinase features in a graph-based model, we can use “transfer learning” to make better drug-target predictions, including kinases with little data.



● robust drug-target interaction datasets

● sparse drug-target interaction datasets

Graph-based Kinase Model : EGFR

Collaborations Pharmaceuticals, Inc.

graphSAGE can scale to hundreds or thousands of targets

MegaKinase: 475 human kinase targets (all ChEMBL human kinase data to date). Activity threshold: 100nM for any target.

ROC of the full 475 human kinase model on a 15% test set is 0.86 (predicting a heterogenous mix of activities on each of the targets)

ROC on EGFR validation: 0.83

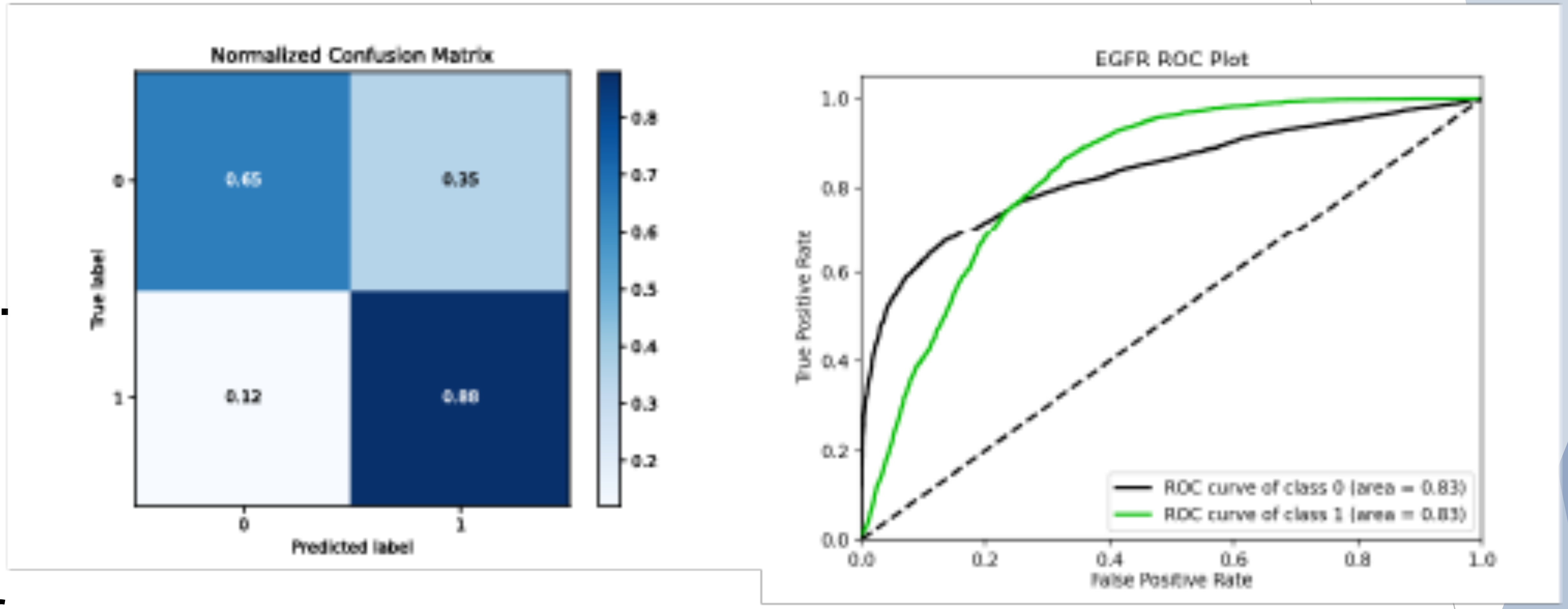
ROC on EGFR, when the model has not seen any examples on EGFR itself: 0.67

Useful for dark kinases

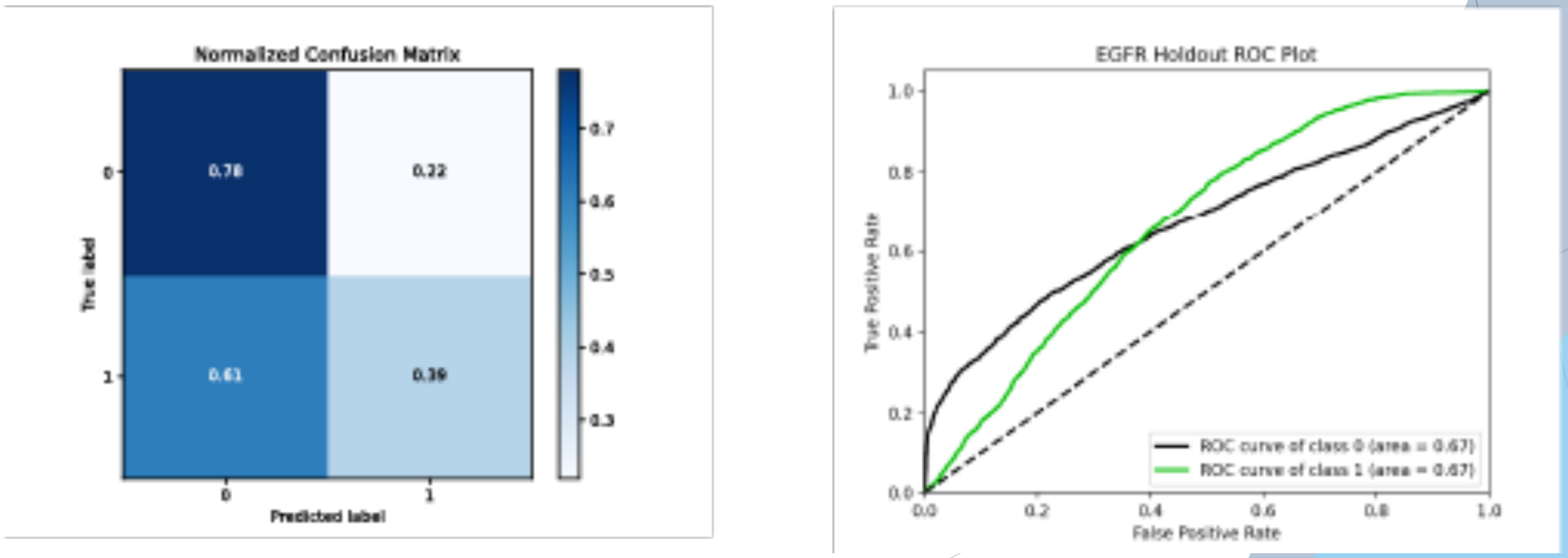
Other proteins with limited data

Apply in toxicology modeling

EGFR ROC = 0.83 with examples in training data

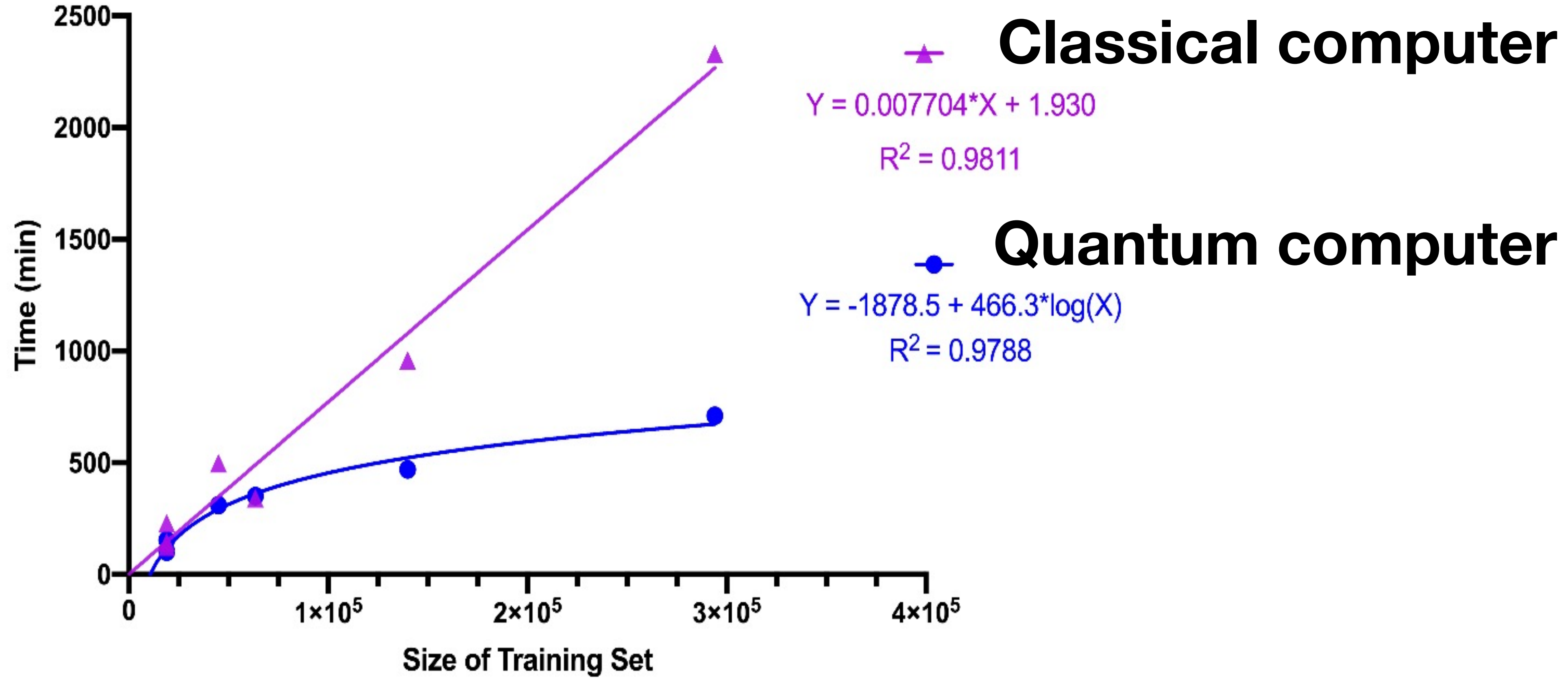


EGFR ROC = 0.67 with no examples in training data



Quantum Machine Learning (QML)

- As the dataset size increases, we need faster methods to build complex models (SVM etc)
- QC simulation outperforms classical computers with increased data set size
- Comparing accuracy and run time results for *M. tuberculosis* inhibition datasets (18,886 compounds) using data re-uploading classifier on classical vs quantum computer with 5-fold cross validation. On 54Qubit IBM machine - QML Faster with trade off in accuracy



Dataset threshold (number of actives)	Time on CC (min)	CC Accuracy (%)	Time on QC (min)	QC Accuracy (%)
100 nM (645)	125	97.1	104	90.5
1 µM (2351)	144	90.4	101	81.4
10 µM (7762)	229	75.6	153	54.9

Batra et al., J Chem Inf Model. 2021 Jun 28;61(6):2641-2647

Predicting Billions of Molecules Bottleneck: DNA Encoded Libraries

Collaborations Pharmaceuticals, Inc.

Scaling up: DNA encoded libraries often require scoring >billion compounds

Problem- ECFP6 algorithms represent a costly bottleneck: too slow

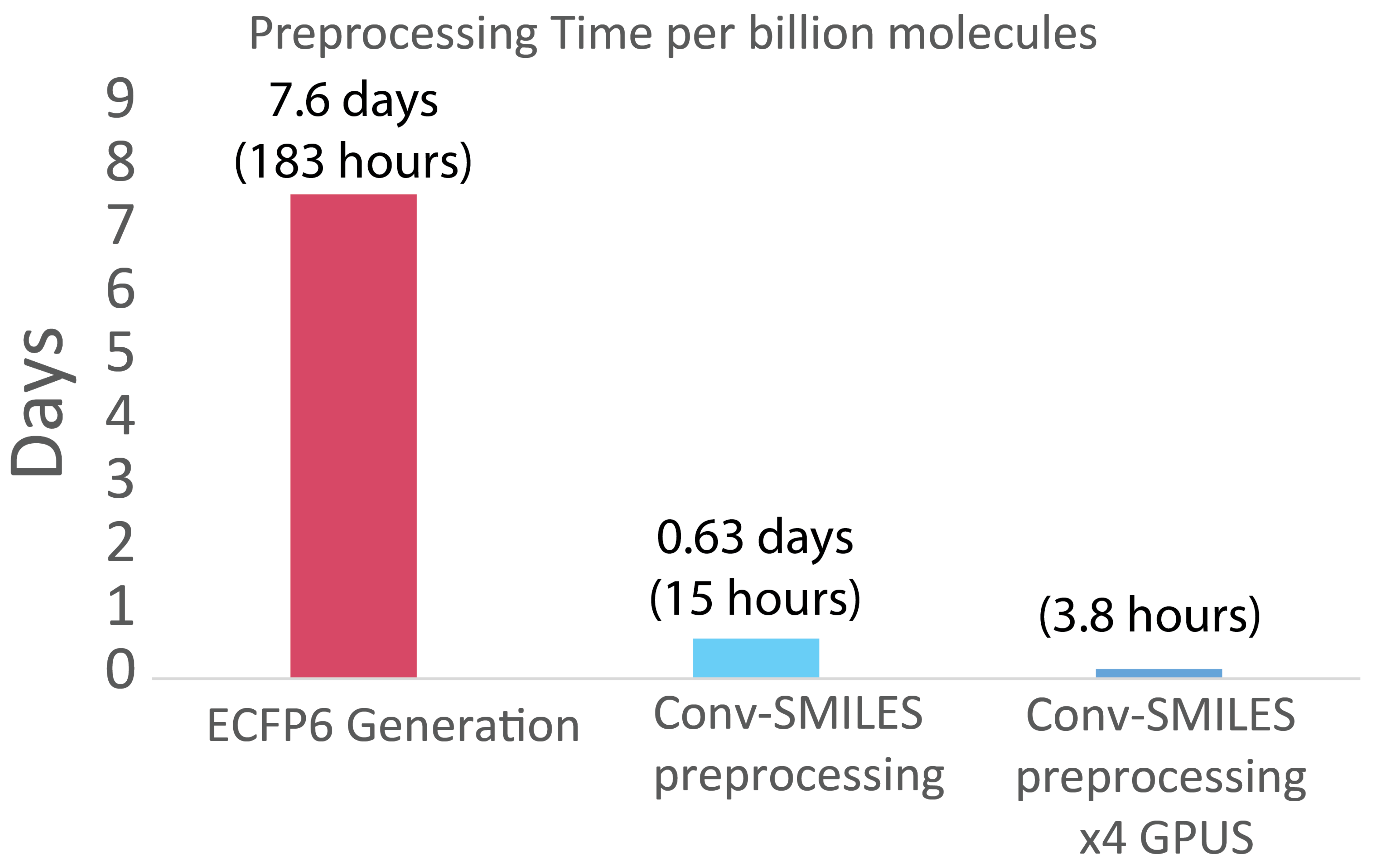
Solution: SMILES based end-to-end Convolution-LSTM model

The model uses encoded SMILES as input to perform classification

~12-15x increase in processing speed on a 1080ti: from a week of preprocessing to hours

GPU enabled: All calculations take place on the GPU, allowing parallel model prediction/preprocessing: 4x GPUs = ~50x speedup on predictions:

No secondary preprocessing storage necessary: SMILES only input



The Need For Speed: Faster End to End Models

Collaborations Pharmaceuticals, Inc.

SMILES based end-to-end Convolution-LSTM model have similar or better predictive power compared to ECFP6-based classification models

$$\text{F1} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}.$$

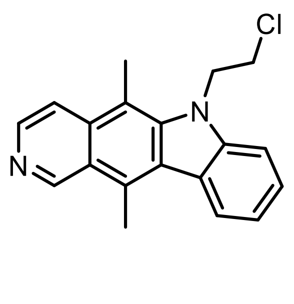
5x cross-validation F1 Score of multiple models and datasets vs. Conv-LSTM

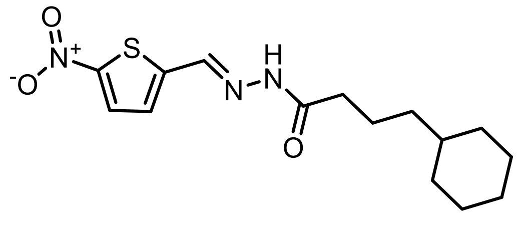
Model Datasets	Adaboost	Bayes	Xgboost	K-NN	Linear Regression	Random Forest	SVC	Conv-LSTM
Water Solubility	0.24	0.38	0.48	0.48	0.25	0.27	0.30	**0.49**
Ames Mutagenesis	NA	0.70	NA	NA	0.75	NA	NA	**0.78**
Blood-Brain Barrier	0.93	0.93	0.95	0.91	0.91	0.95	**0.96**	0.93
CHO Cytotoxicity Assay	0.68	0.71	0.70	0.70	0.66	0.72	0.69	**0.74**
CYP3A4 Inhibition	0.84	0.83	0.85	0.82	0.80	0.83	**0.85**	0.80
hERG Ki	0.85	0.87	0.86	0.81	0.84	0.86		0.85
Plasma Protein Binding	0.85	0.84	0.87	0.86	0.85	0.87	**0.88**	0.82

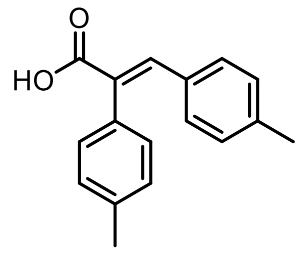
UV adVISor

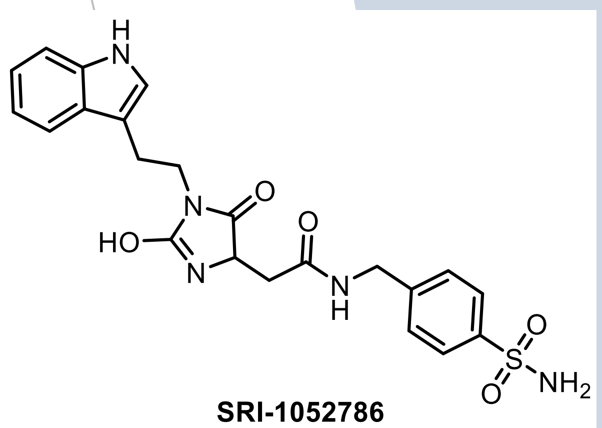
Predicting UV-Vis Spectra For Molecules Without Physical Samples

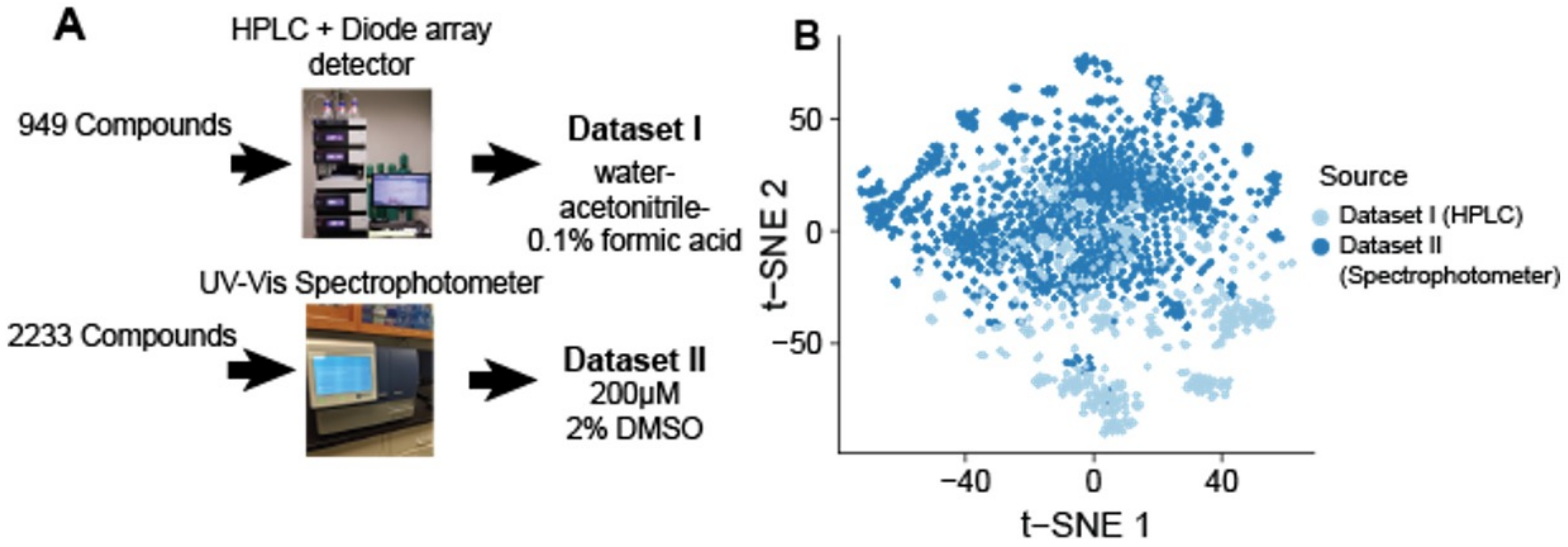
Collaborations Pharmaceuticals, Inc.


SRI-0000202


SRI-0000449


SRI-0000497

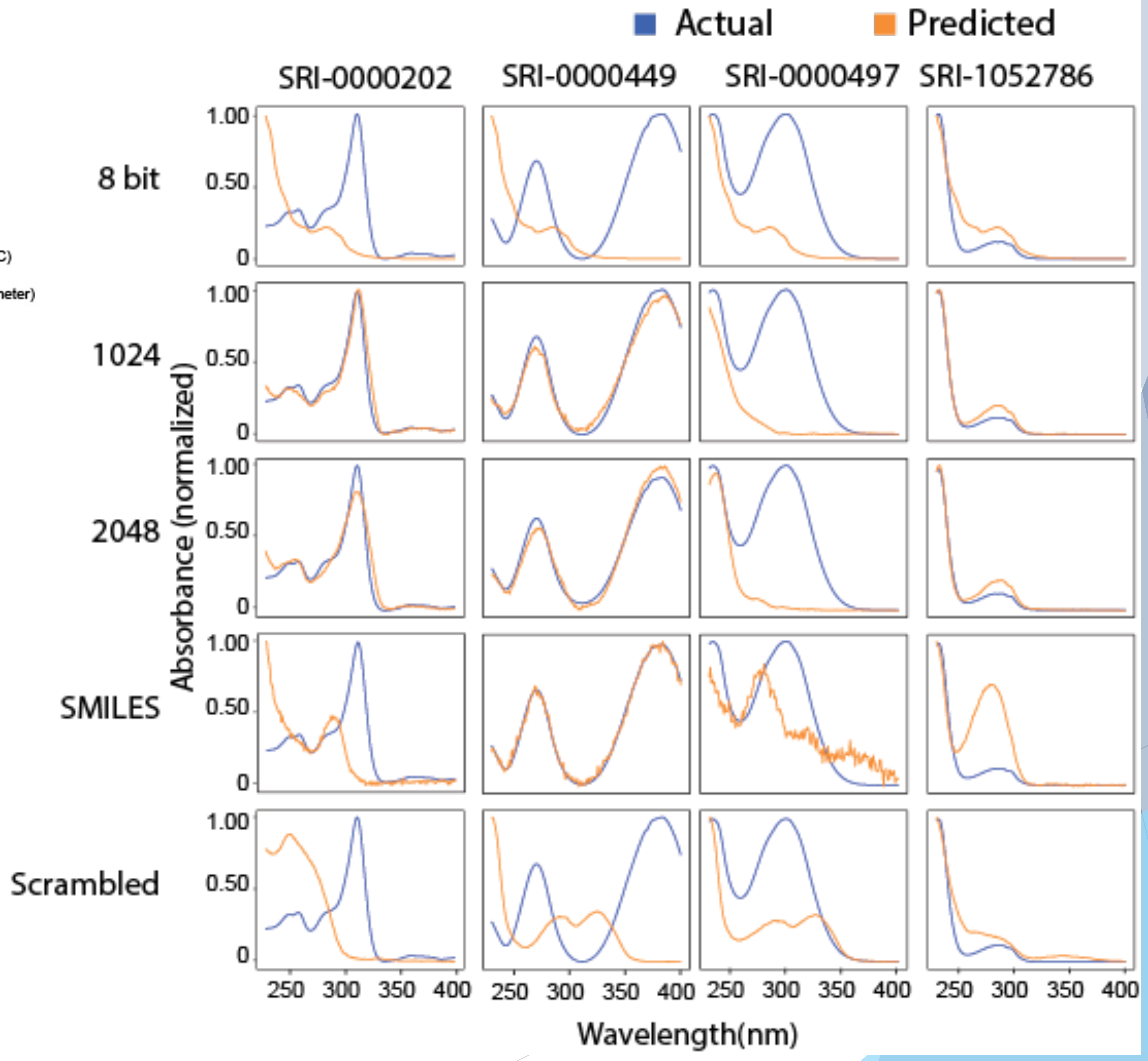

SRI-1052786



- Potential use in predicting:
- Chemistry-in-a-box analytical
- Assay interference for novel libraries
- Predicting color
- Predicting phototoxicity
- **70:15:15 (train: test: validation)**

- SMILES Median RMSE = 0.166
- predictions better than DFT RMSE ~0.3-0.4

- **https://spectra.collaborationspharma.com**



© 2021 Collaborations Pharmaceuticals Inc. Non-Proprietary slides.

Funding: DARPA (HR0011-19-C-0108; PI: P. Madrid)

The Future of Computational Toxicology

COLLABORATIONS PHARMACEUTICALS, INC.

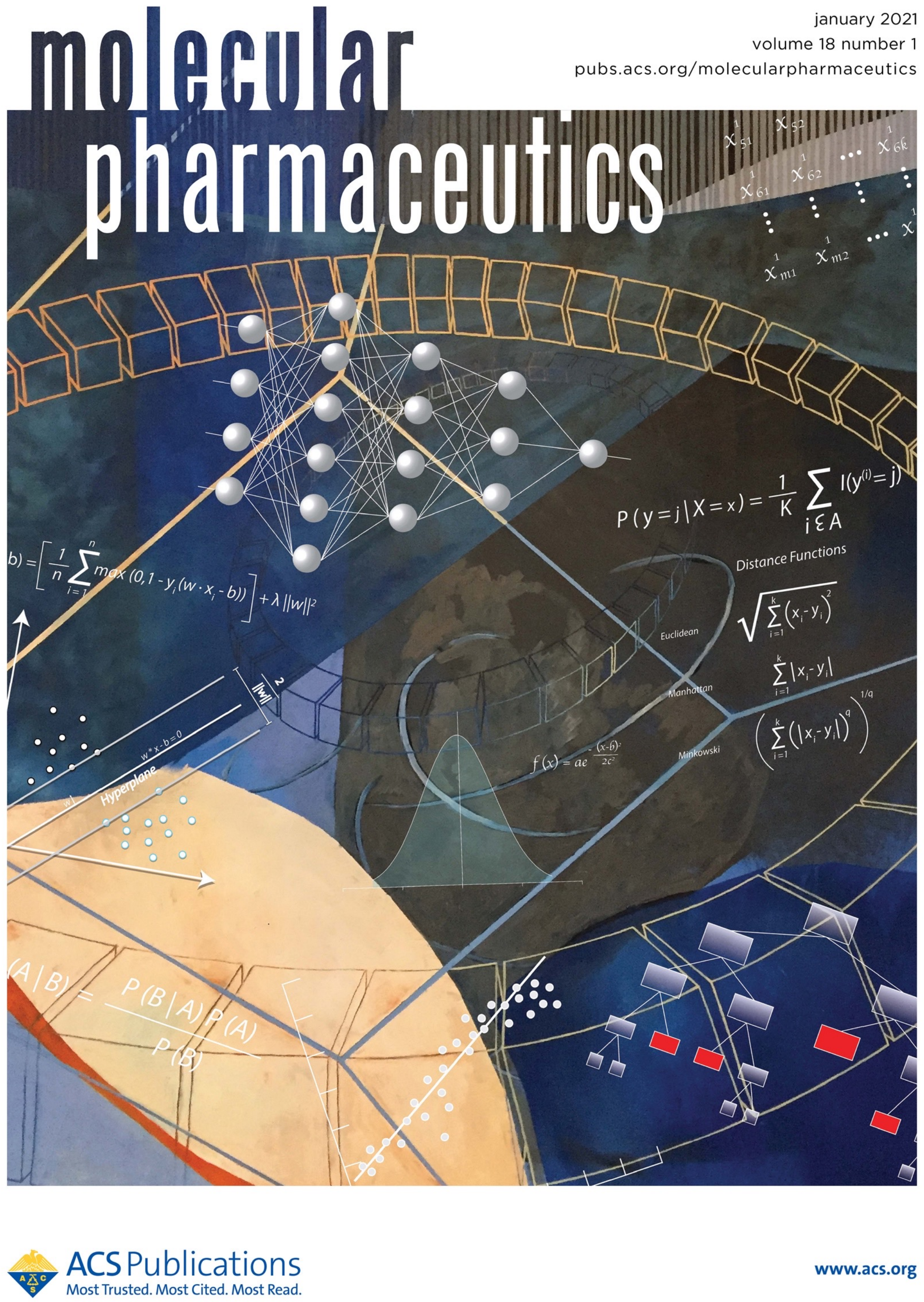
We have the technology to create massive numbers of molecules

Molecule design becomes autonomous

We have the tools to predict toxicology and physicochemical properties faster

Integrated design-make-test cycles becomes a reality

AI can help us learn from the data we have for predicting impact on human and other targets



molecular pharmaceutics

january 2021
volume 18 number 1
pubs.acs.org/molecularpharmaceutics

ACS Publications
Most Trusted. Most Cited. Most Read.

www.acs.org

NIH National Center for Complementary and Integrative Health

1R43AT010585-01

NIH National Institute of General Medical Sciences

R43GM122196
R44GM122196-02A1
1R41GM131433-01A1
3R44GM122196-03S1

NIH National Institute of Environmental Health Sciences
Your Environment. Your Health.

1R43ES031038-01
3R43ES031038-01S1