ATOM

Data-driven modeling approaches in computational drug discovery

Jonathan Allen, Ph.D.
Informatics Scientist
Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

LLNL-PRES-821879

Roadmap of Talk

- Introduction to ATOM Consortium
- Introduction to small molecule drug discovery data
- Target specific drug modeling approaches
- Structure based multi-target modeling

ATOM is an open public-private partnership for accelerating drug discovery

Goals

- Accelerate the drug discovery process
- Improve success rate in translation to patients

Approach

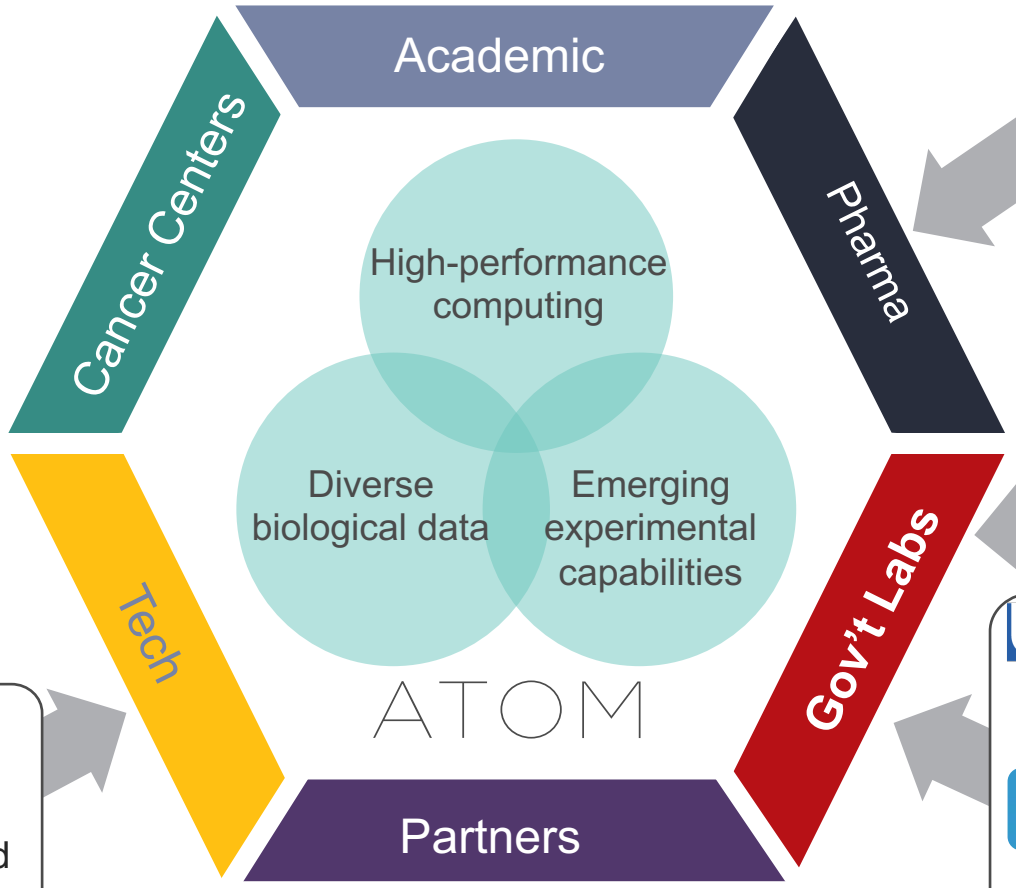
- Computation-driven drug design, supported and validated by targeted experiments
- Data-sharing to build models using everyone's data

Product

- An open-source platform for active-learning based molecular design and optimization

Status

- Shared collaboration space at Mission Bay, SF
- 25 FTEs engaged across the partners
- R&D started February 2018



gsk

Drug discovery, chemistry, and dark data

Frederick National Laboratory for Cancer Research

Experimental biology and data analytics

Lawrence Livermore National Laboratory
Argonne National Laboratory
Oak Ridge National Laboratory
Brookhaven National Laboratory

High performance computing, data science

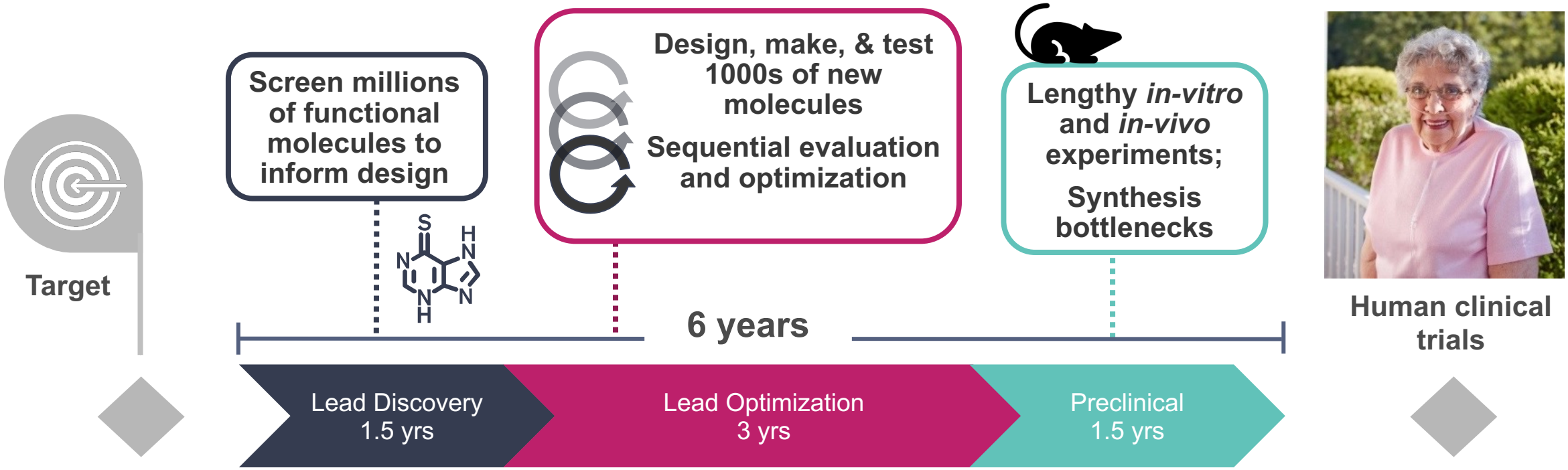
UCSF

Cancer biology and new assay development

Example member contributions

Current drug discovery: long, costly, high failure

Goal: transform early drug discovery to get drugs to patients faster



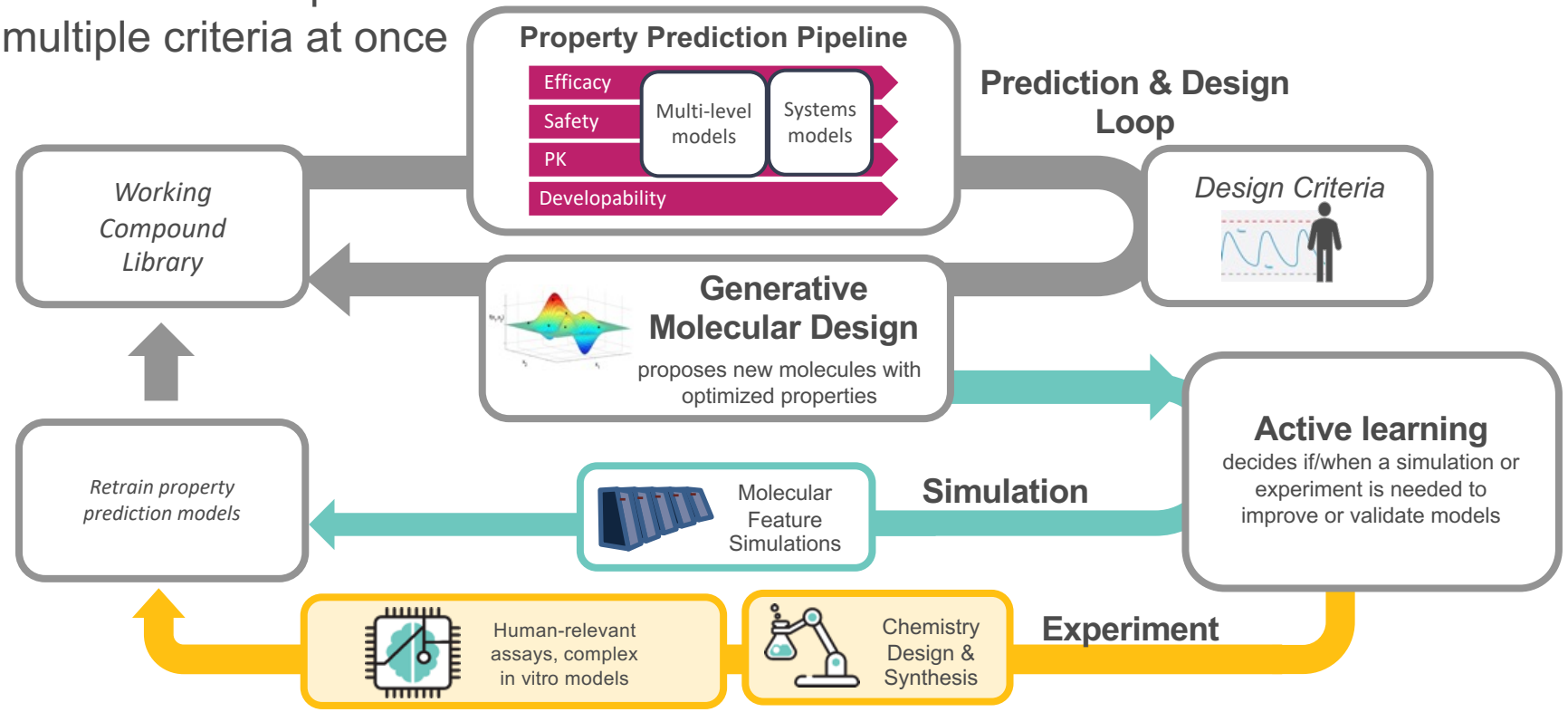
- **33% of total cost of medicine development**
- **Clinical success only ~12%, indicating poor translation in patients**

Source: http://www.nature.com/nrd/journal/v9/n3/pdf/nrd3078.pdf

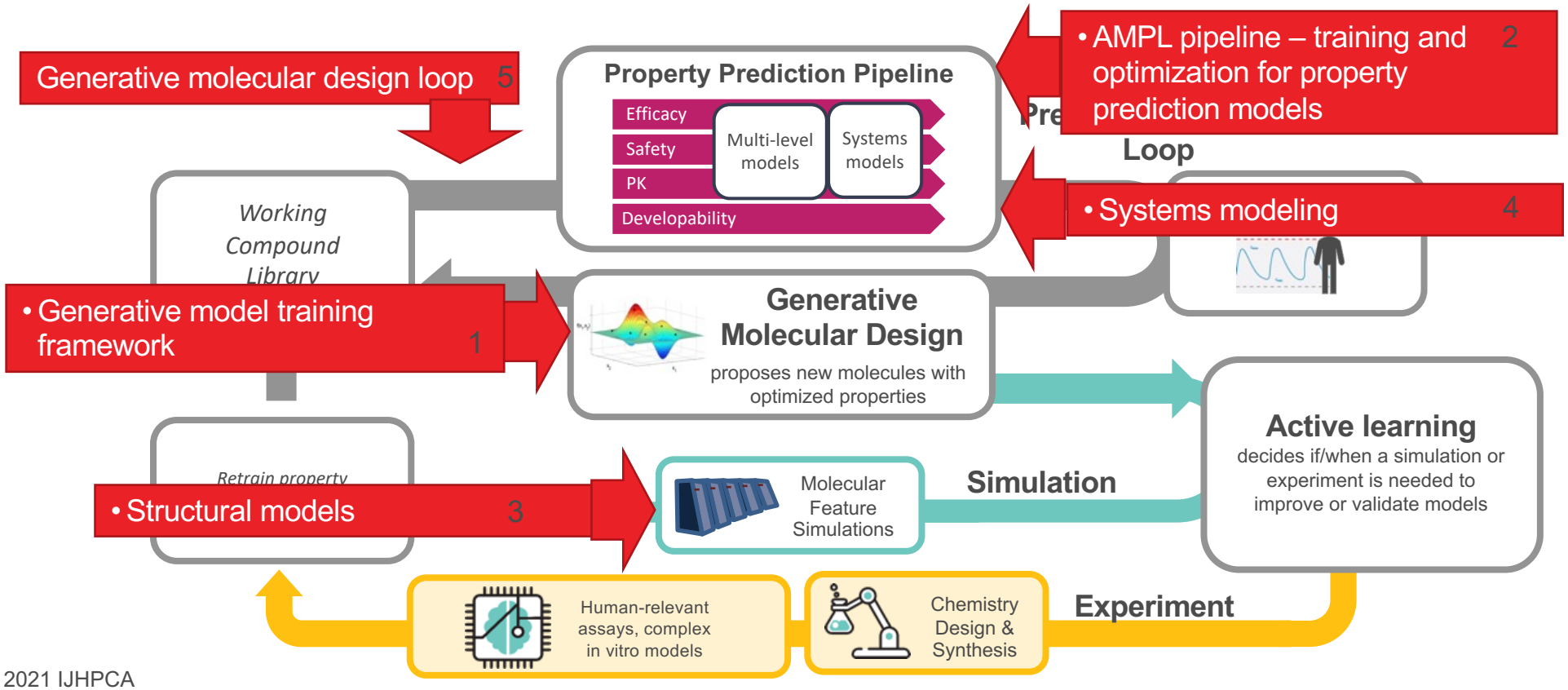
The ATOM Platform
Active Learning Drug Discovery Framework

Two computational challenges to highlight

1. Efficient exploration of chemical space
2. Optimize against multiple criteria at once



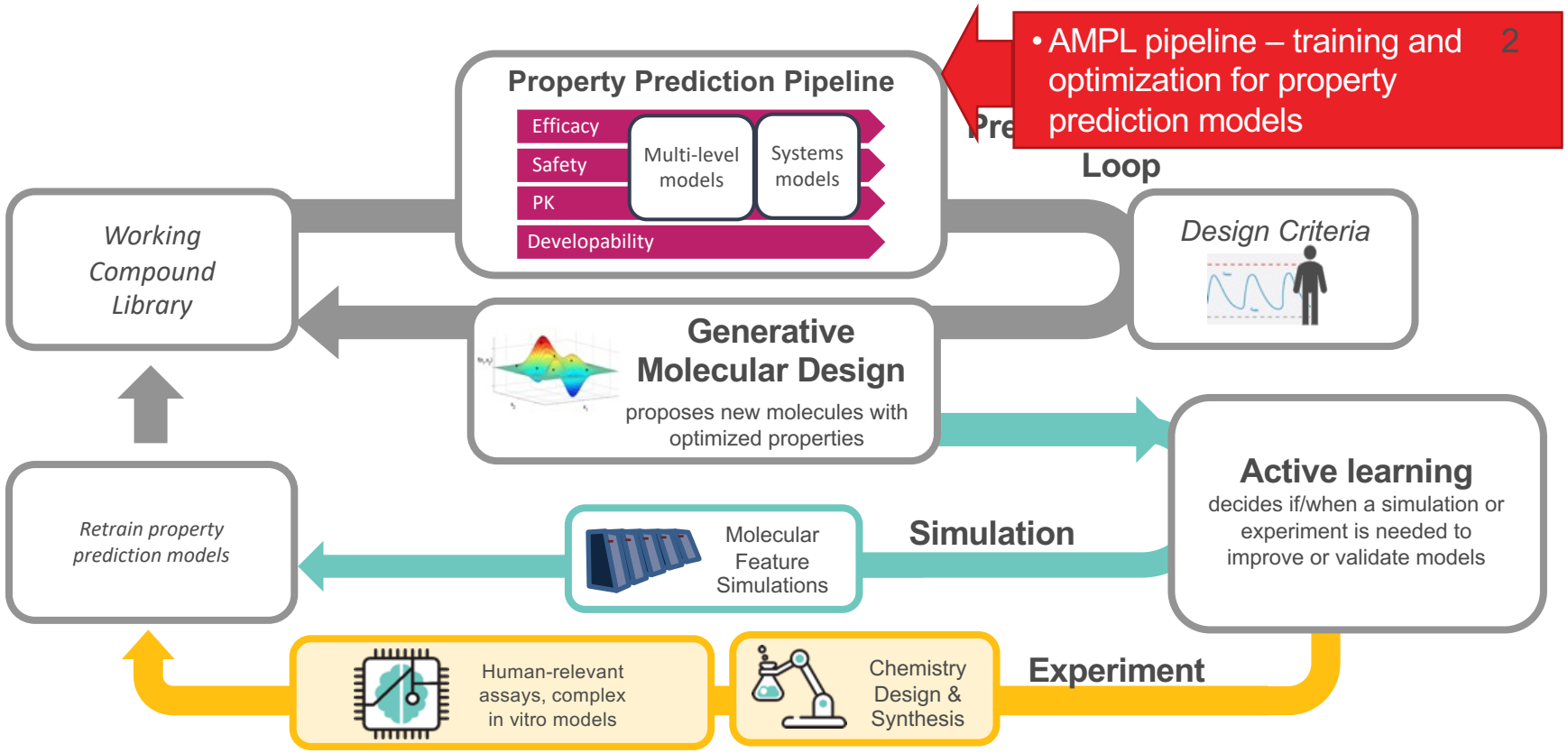
The ATOM Platform
Active Learning Drug Discovery Framework



1) Jacobs et al., 2021 IJHPCA
2) Minnich et al., 2020 JCIM ; McLoughlin et al., 2021 JCIM
3) Zhang et al., 2017 CTMC, Jones et al., 2021 JCIM
4) Murad et al., 2021 DMD
5) Code approved for release, manuscript forthcoming

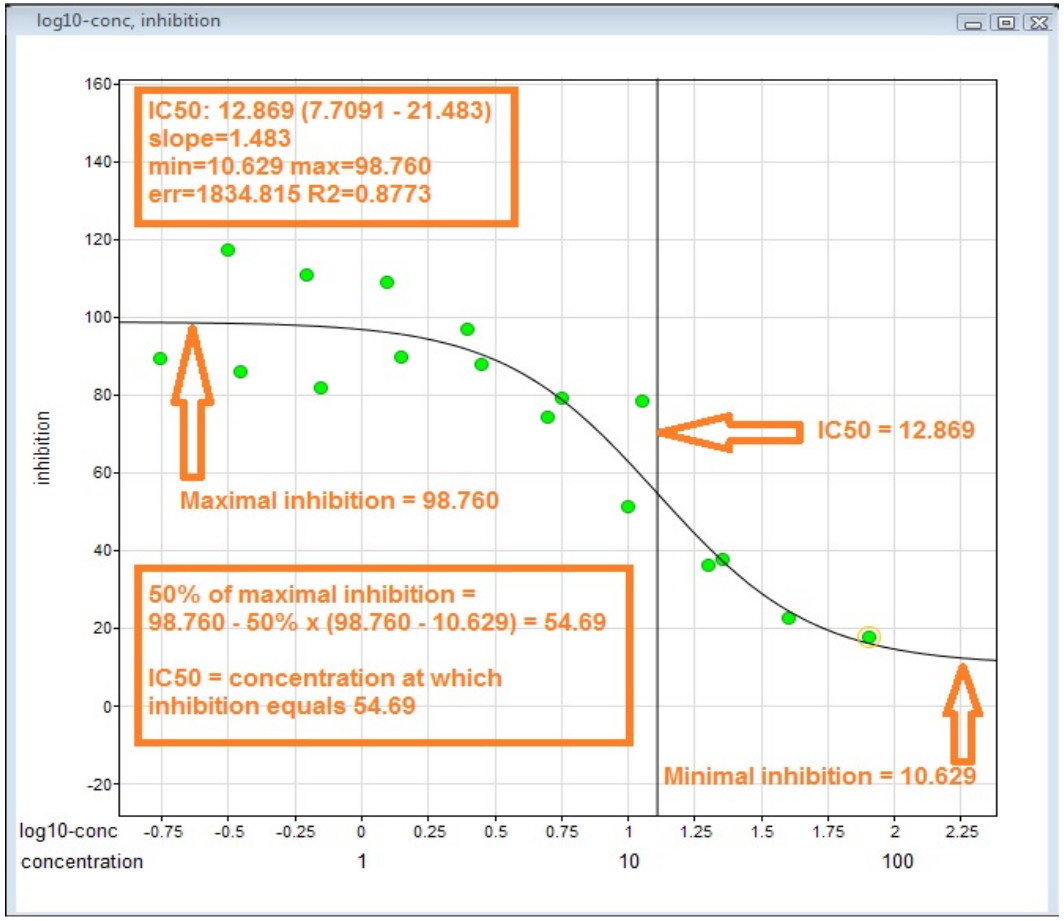
Targeting GMD platform for opensource release after partner testing (currently Frederick National Lab, LBNL)

The ATOM Platform
Active Learning Drug Discovery Framework



Types of drug discovery screening assays

Example IC50 curve



- Cell-based assays
- Immunoassays
- Enzyme activity assays
- Phenotypic assays
  - Cytotox assays

Measurements:

- IC50, AC50, Ki, Kd
- Single concentration % inhibition

Benefits:

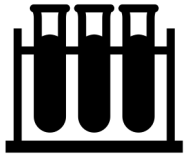
- Limited knowledge of the precise molecular mechanisms of action

Drawbacks:

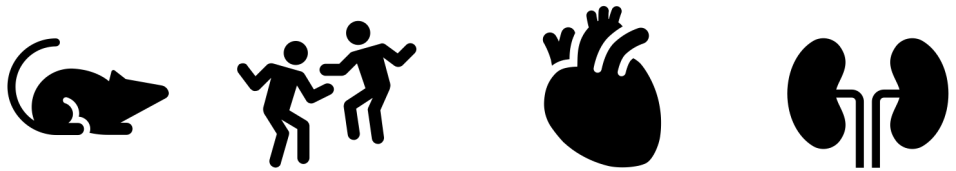
- The same molecule may yield very different results depending on the assay technology

Types of medicinal chemistry, pharmacokinetics properties

- Med. Chem
  - Solubility
  - Dissociation constants (pKa, pKb)
  - Octanol water partitioning (logP, logD)
  - Permeability through biological membranes (Papp)
  - Transporter substrates and inhibition



- Pharmacokinetics and toxicity properties
  - Fraction unbound to plasma proteins (fup)
  - Ratio of blood to plasma (RBP)
  - Fraction unbound in liver microsomes (fumic)
  - Volume of distribution at steady state (VDss)
  - Clearance (CL)
  - Metabolic enzyme substrates and inhibitors (CYP, UGT)
  - Liver toxicity (BSEP, MRP3, …)
  - Cardiac toxicity (KCNH2, …)



Common data sources to build model ready datasets

ChEMBL – Manually curated repository of bioactive molecules (updated)

- Sponsored by European Bioinformatics Institute (EMBL-EBI)
- 1.9M compounds, 11K targets

Excape-DB – Exascale Compound Activity Prediction

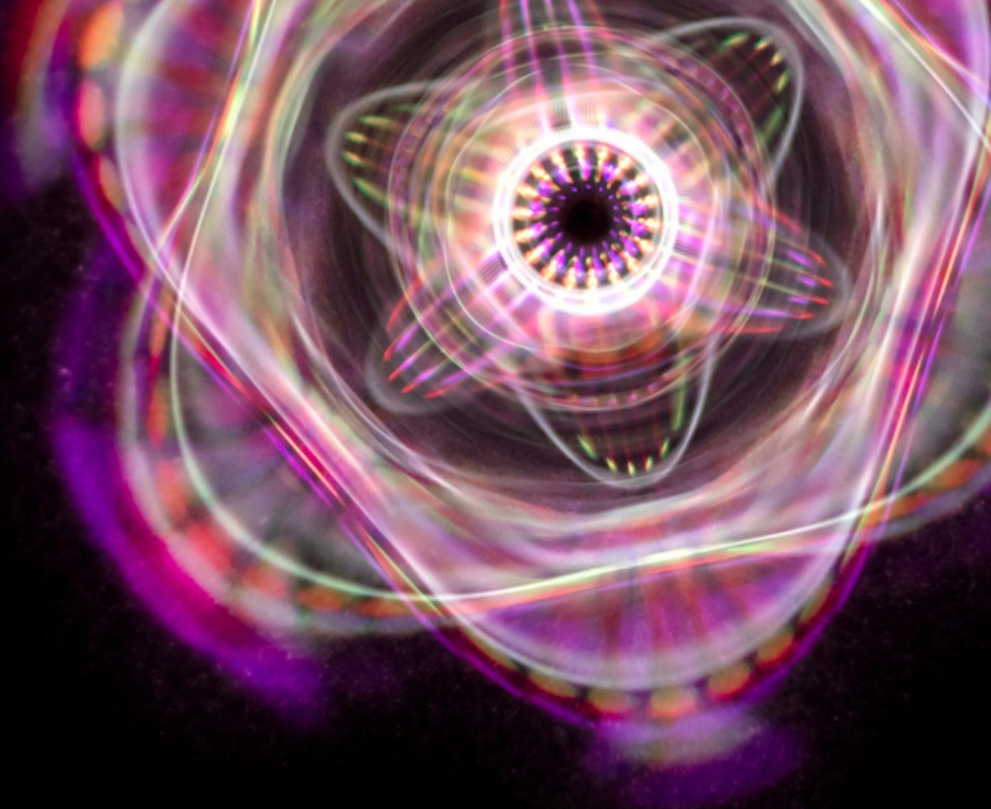
- EU program on predictive modeling for compound activities
- 1M compounds, 1.7K targets

Drug Target Commons – An open multi-database platform for curation with common ontology

- *Sponsored by University of Helsinki*
- *Largest source is CHEMBL*
- *1.7M compounds, 13K targets*

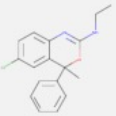
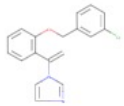
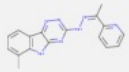
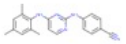
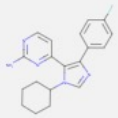
Excelra GoSTAR (updated)

- Commercial database
- 7.8M compounds, 9.3K targets
- Derived data products (e.g. models) are open

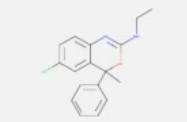
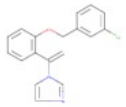
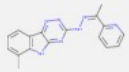
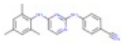
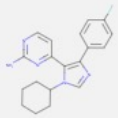


Introduction to Quantitative Structure Activity Relationships (QSAR)

Cheminformatics datasets

Compound ID	Structure	MW	AlogP	Target	Active	IC50 (uM)
CHEMBL2106227	 CHEMBL2106227	300.79	4.23	Aurora kinase B	False	1.5
CHEMBL27289	 CHEMBL27289	310.78	4.63	Aurora kinase B	False	3
CHEMBL2094620	 CHEMBL2094620	317.36	3.05	Aurora kinase B	True	0.10
CHEMBL70633	 CHEMBL70663	329.41	4.76	Aurora kinase B	False	> 100
CHEMBL1951415	 CHEMBL1951415	337.40	4.23	Aurora kinase B	False	> 100

Types of machine learning tasks

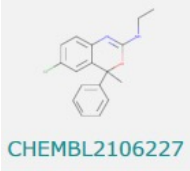
Compound ID	Structure	Active	IC50 (uM)
CHEMBL2106227	 CHEMBL2106227	False	1.5
CHEMBL27289	 CHEMBL27289	False	3
CHEMBL2094620	 CHEMBL2094620	True	0.10
CHEMBL70633	 CHEMBL70663	False	> 100
CHEMBL1951415	 CHEMBL1951415	False	> 100

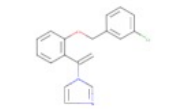
Classification task (with classes)

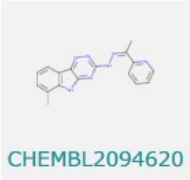
Regression task (with numerical values)

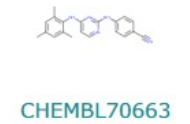
How do we encode molecules?

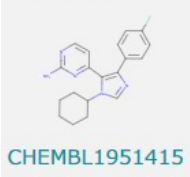
- There are many **featurization** approaches





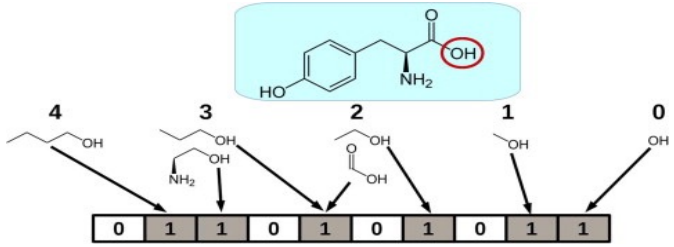






SMILES
CC(=O)Nc1ccc(O)cc1

Fingerprints



Molecular descriptors

clogD	TPSA	QM1	…
3.1	10.2	4.1	…

?

1.5

3

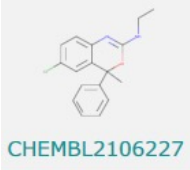
0.10

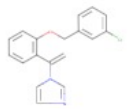
> 100

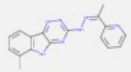
> 100

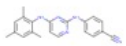
How do we predict a property?

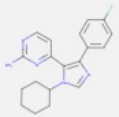
- Fit **machine learning models** and parameters to predict properties



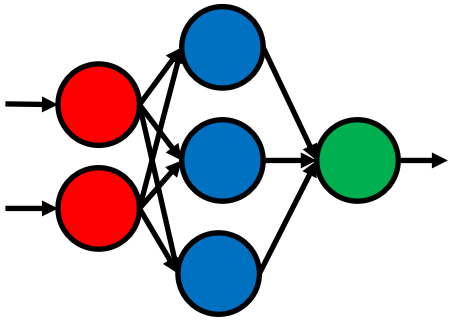




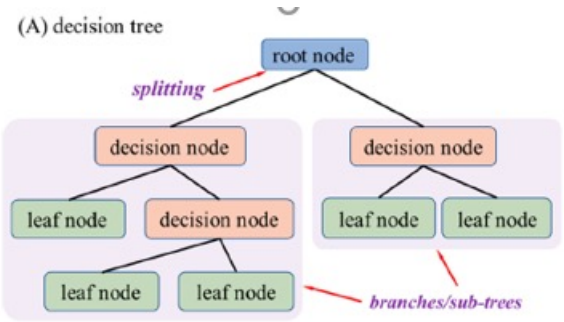




Featurization



Neural network
- Mathematical functions
- Parameters for functions



Random Forest
- Decision trees based on features
- Forest: Collection of decision trees

1.5

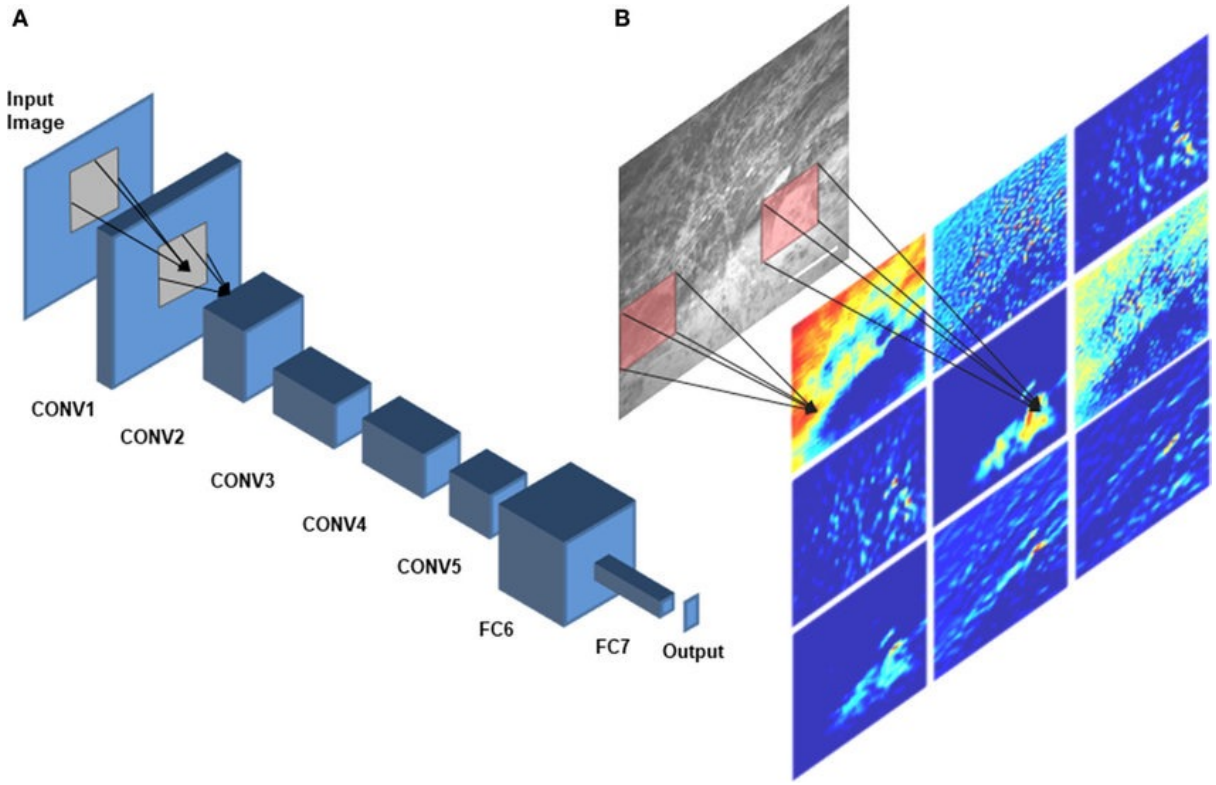
3

0.10

> 100

> 100

What about Deep Learning?

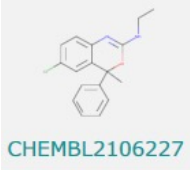


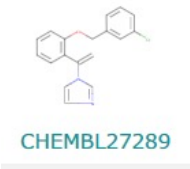
- Deep convolutional neural networks (DCNN) have been successful in a variety of tasks
  - Image recognition
  - Natural language processing
  - AlphaGo
- Two key cheminformatics applications:
  - Representation learning
  - Multi-task and transfer learning

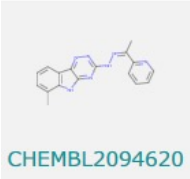
Izadyyazdanabadi Frontiers Oncology 2018

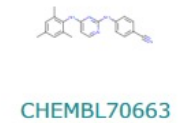
Deep learning for QSAR

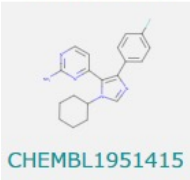
- Yes, there are now several deep models for chemistry applications

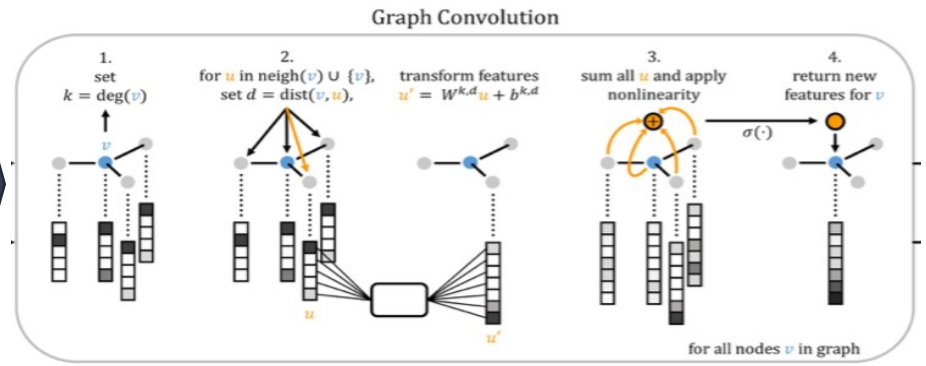












- **Message Passing Neural Networks**
- **Spatial graph**
- **Weave model**
- **Others**

Machine learning model

1.5

3

0.10

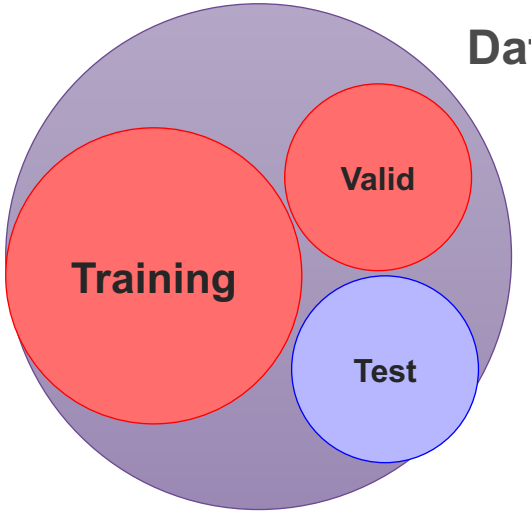
> 100

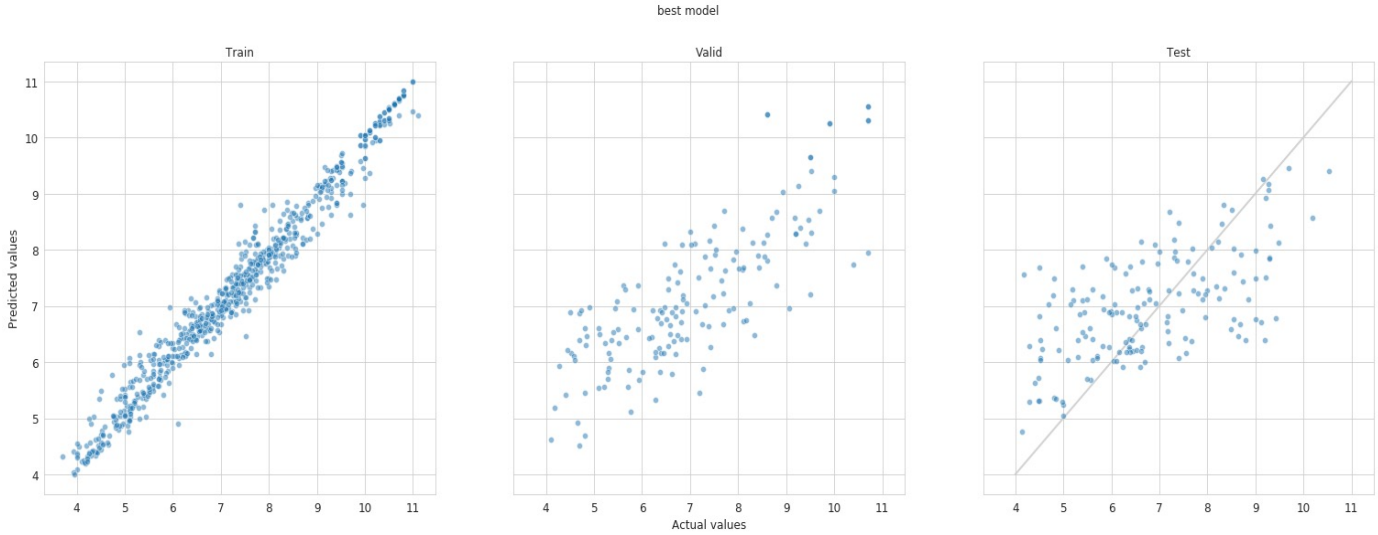
> 100

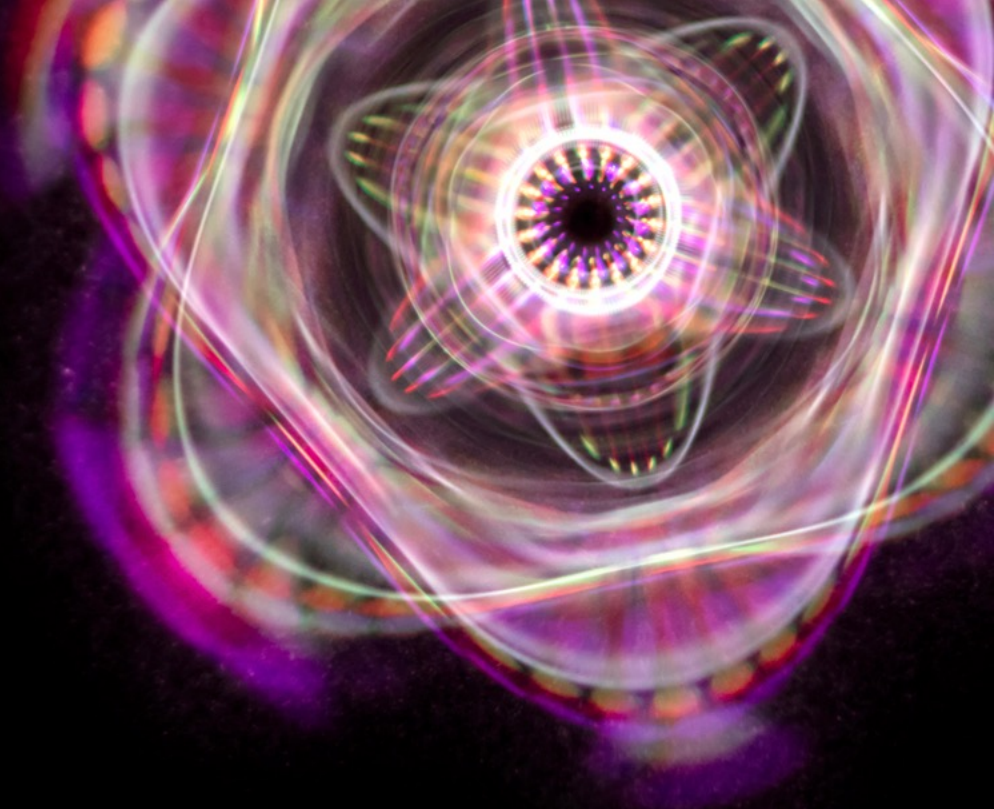
Wu arXiv:1703.00564v3

How can we test our model?

- Test model predictions prospectively on new compounds to be measured
- Artificially split historic data into sets
  - Training
  - Validation
  - Test
- Test set becomes the simulated prospectively tested compounds





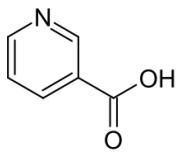


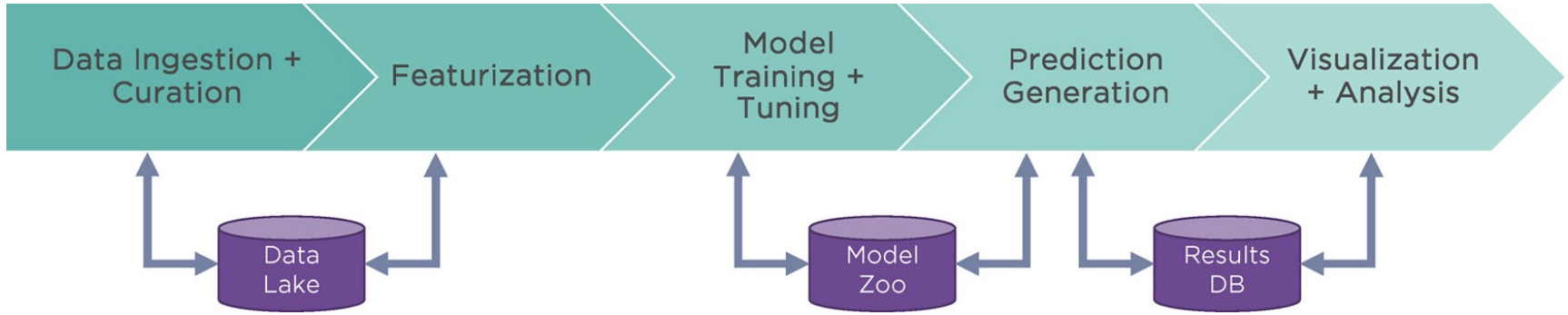
The ATOM Modeling PipeLine

AMPL: The ATOM Modeling PipeLine

From chemical structure and bioassay/property data to model to prediction

Molecular structure





Predicted activities and properties

IC50
AC50
LogP
fup
CL

- Many ML algorithms exist but they are not “one size fits all”
- Building state-of-the-art reproducible models remains challenging
- Goal of AMPL: an open source tool to automate QSAR model fitting

AMPL: The ATOM Modeling PipeLine

Data Ingestion + Curation → Featurization → Model Training + Tuning → Prediction Generation → Visualization + Analysis

1. Data curation	• Generation of RDKit molecular SMILES structures • Processing of qualified or censored data processing • Curation of activity and property values
2. Featurization	• Extended connectivity fingerprints (ECFP) • Graph convolution latent vectors from DeepChem • Chemical descriptors from Mordred package • Descriptors generated by MOE (requires MOE license)
3. Model training and tuning	• Test set selection • Cross-validation • Uncertainty quantification • Hyperparameter optimization
4. Supported models	• scikit-learn random forest models • XGBoost models • Fully connected neural networks • Graph convolution models (DeepChem)
5. Visualization and analysis	• UMAP • Chemical diversity analysis

RDKit
Open-Source Cheminformatics and Machine Learning

deepchem

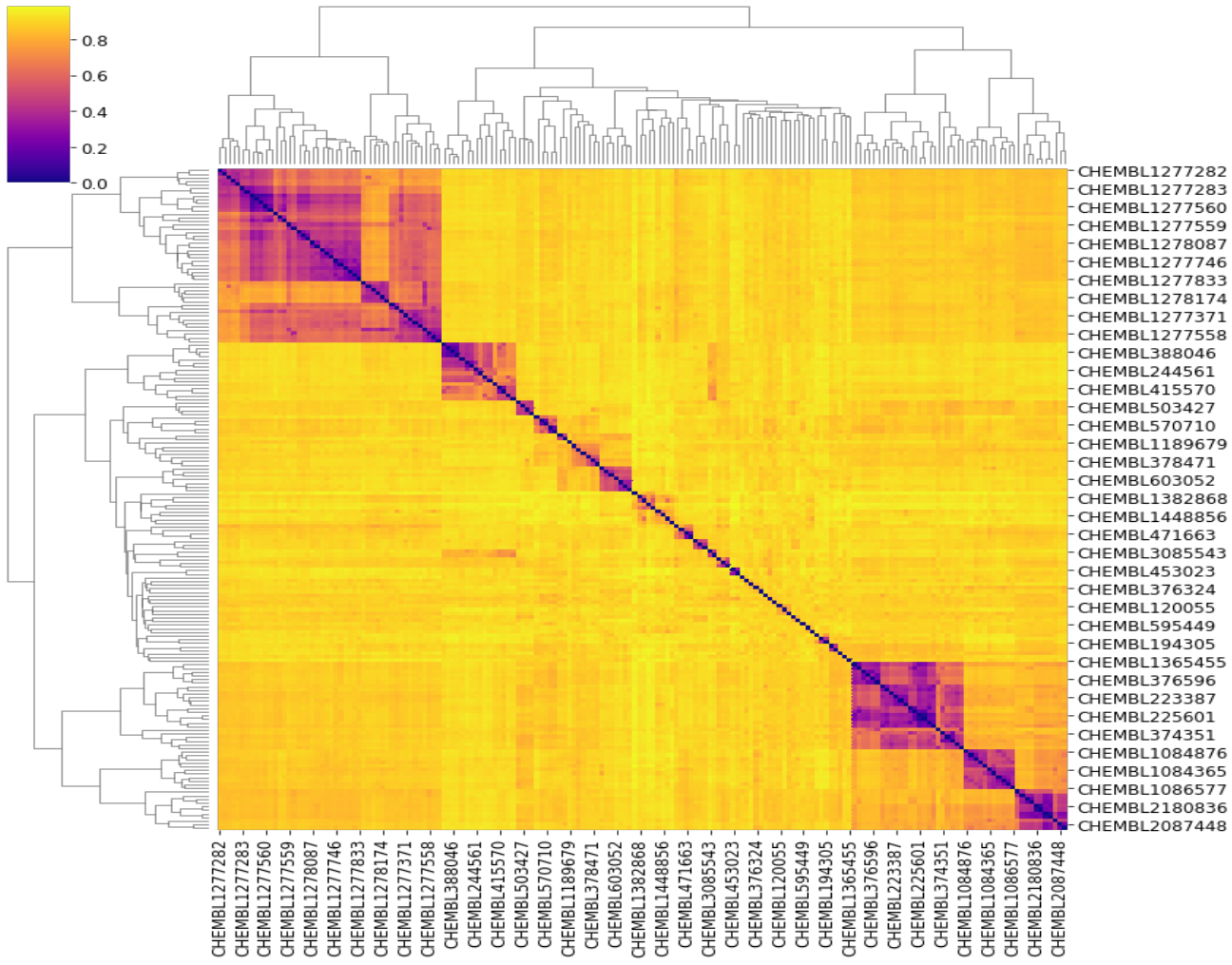
TensorFlow

dmlc XGBoost

scikit learn

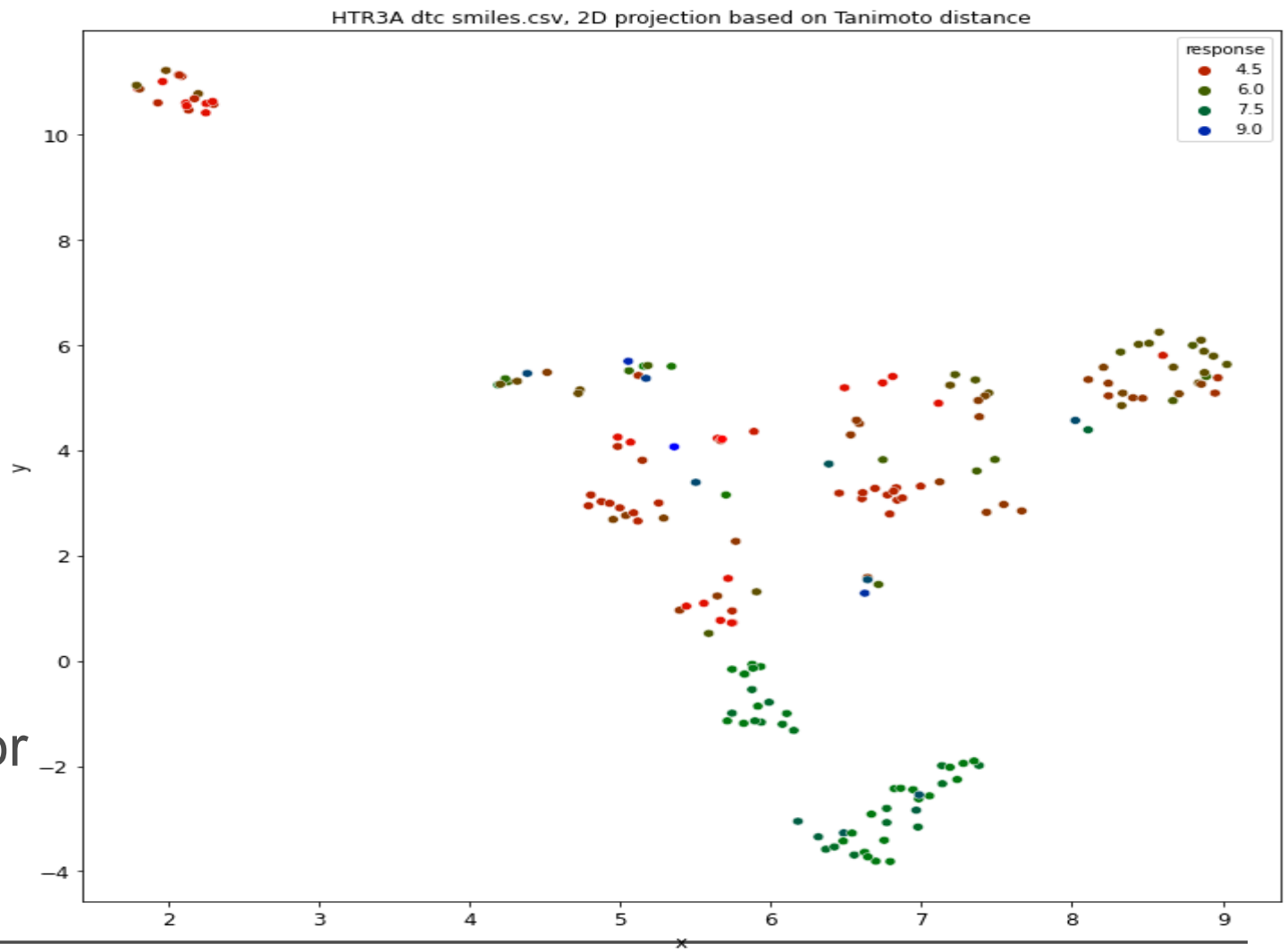
Using AMPL: Curation

- Remove inaccurate or incorrect structures
- Clean structures
  - Generate canonical representation
- Analyze duplicates
  - Average measurements
- Analyze properties
  - Characterize structures and features
  - Examine predicted property or activity distributions



Using AMPL: Curation

- Remove inaccurate or incorrect structures
- Clean structures
  - Generate canonical representation
- Analyze duplicates
  - Average measurements
- Analyze properties
  - Characterize structures and features
  - Examine predicted property or activity distributions



Using AMPL: Curation – track measurement variability

```
old_compound_id='rdkit_smiles'
new_compound_id='rdkit_smiles'
reject=data[~data[old_compound_id].isin(check_df[new_compound_id])]
reject
```

assay_description	title	journal	doc_type	annotation_comments	PXC50	Unnamed: 0	CID	smiles
Displacement of [3H]granisetron from human rec...	Novel antagonists of serotonin-4 receptors: sy...	Bioorg. Med. Chem.	PUBLICATION	NaN	8.159894	41	42636941	CCCN1CCC(CC1)COC2=NC3=C(C(=CS3)C)N4C2=CC=C4
Antagonist activity at 5HT3 receptor in hybrid...	Novel antagonists of serotonin-4 receptors: sy...	Bioorg. Med. Chem.	PUBLICATION	NaN	6.414539	41	42636941	CCCN1CCC(CC1)COC2=NC3=C(C(=CS3)C)N4C2=CC=C4
Binding activity radioligand.	Synthesis and serotonergic activity of N,N-dim...	J. Med. Chem.	PUBLICATION	NaN	5.000000	72	5358	CNS(=O)(=O)CC1=CC2=C(C=C1)NC=C2CCN(C)C
Compound was evaluated for the affinity at 5-h...	Selective, orally active 5-HT1D receptor agoni...	J. Med. Chem.	PUBLICATION	NaN	8.031517	72	5358	CNS(=O)(=O)CC1=CC2=C(C=C1)NC=C2CCN(C)C

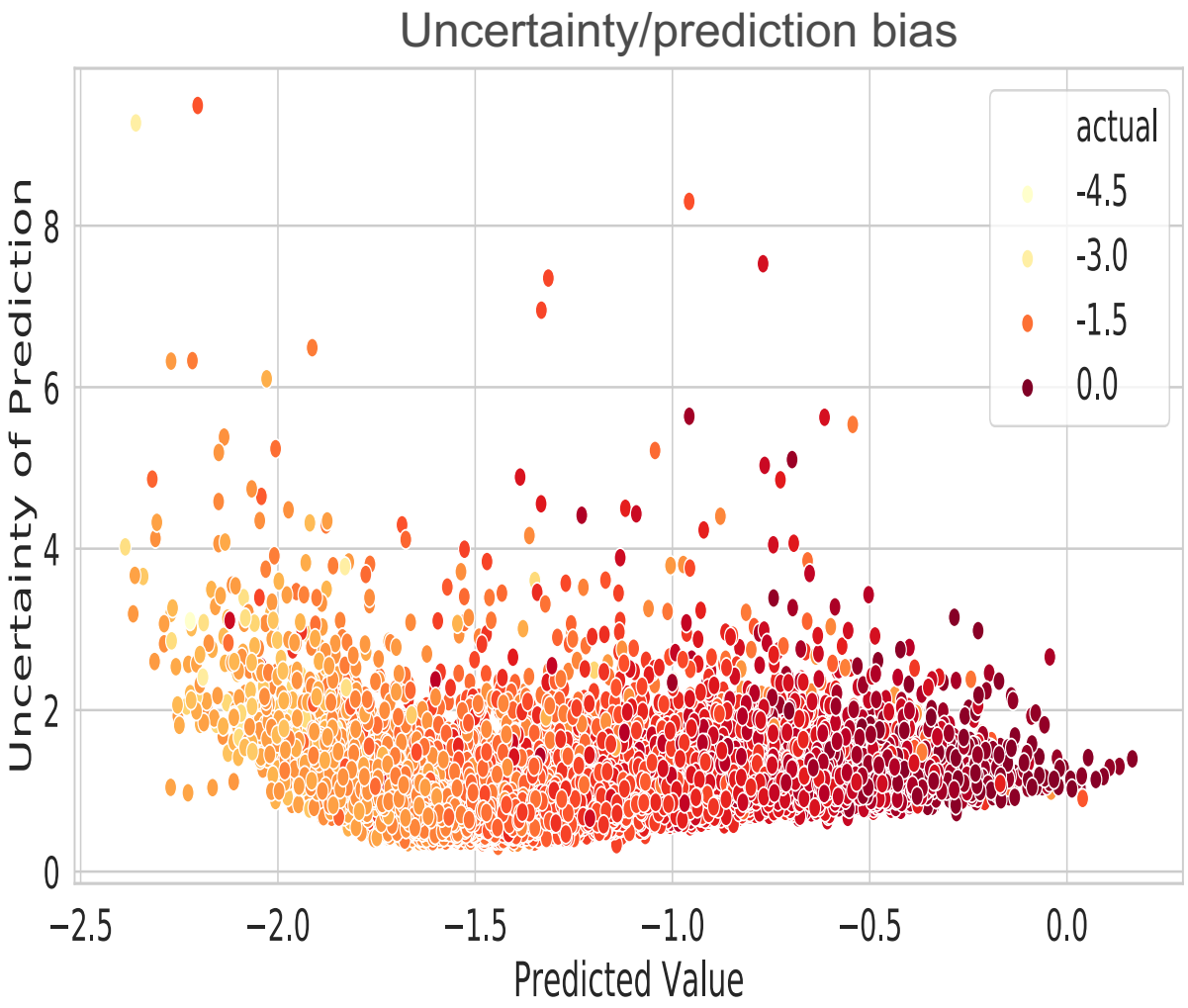
Important to have searchable, sharable, reusable datasets

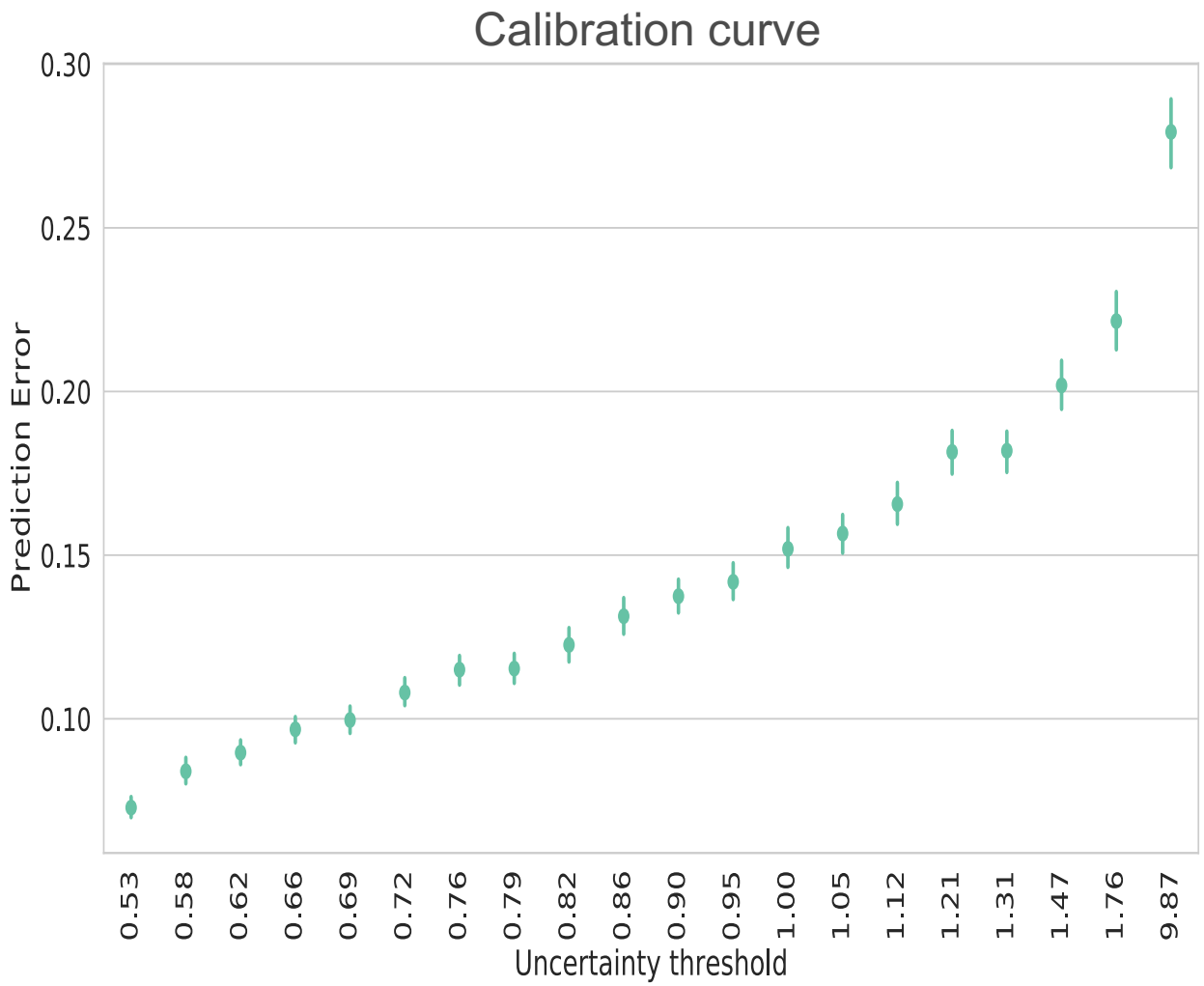
- Store raw data and final machine learning ready dataset
  - 'file_category': 'experimental',
  - 'assay_category': 'safety',
  - 'assay_endpoint' : 'pic50',
  - 'curation_level': 'ml_ready',
  - 'data_origin' : 'ExcapeDB',
  - 'functional_area' : 'Liability screen',
  - 'matrix' : 'multiple values',
  - 'journal_doi' : https://doi.org/10.1016/j.chembiol.2017.11.009',
  - 'sample_type' : 'in_vitro',
  - 'species' : 'human',
  - 'target' : 'CYP2D6',
  - 'target_type' : 'protein',
  - 'id_col' : 'compound_id',
  - 'response_col' : 'VALUE_NUM_mean',
  - 'prediction_type' : 'regression',
  - 'smiles_col' : 'rdkit_smiles',
  - 'units' : -log10 molar',
  - 'source_file_id' : source_of_raw_data,
  - 'user' : 'user99'

Modeling uncertainty

- Random Forest
  - Calculate the standard deviation of predictions from individual trees
- Neural Networks
  - Use DeepChem's method, which combines aleatoric (sensing uncertainty) and epistemic (model uncertainty) values (*Kendall and Gall 2017*)
  - Aleatoric: Modify loss function and train model to predict both response variable and input variance
  - Epistemic: Apply dropout masks during prediction and quantify variability in predictions
  - Then $\sigma_{total} = \sqrt{{\sigma_{aleatoric}}^2 + {\sigma_{epistemic}}^2}$

Model uncertainty is critical to active learning and remains an open challenge





Blood plasma binding (HSA)

Domain of applicability

K nearest neighbors mean distance local density

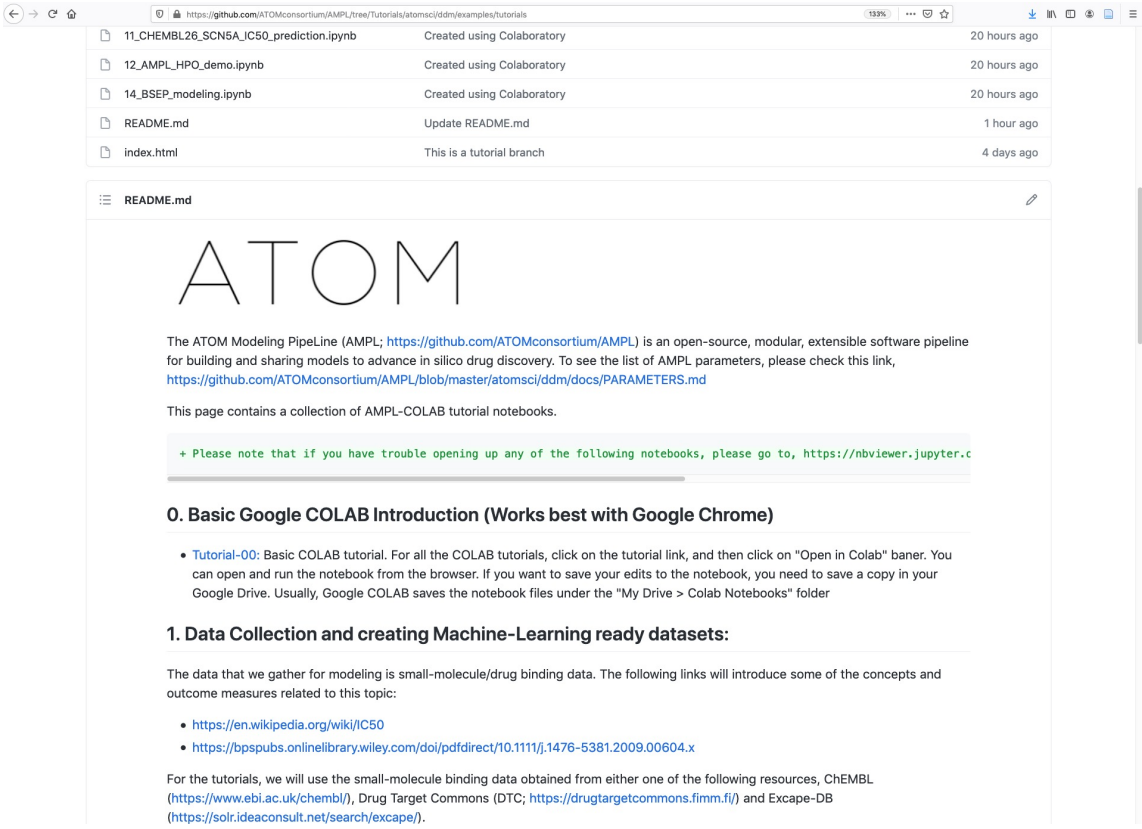
- Euclidean distance
- Calculate pairwise distance between points in the training set
- For each point in the prediction set, calculate the mean distance of the point to its K nearest neighbors in the training set.
- For the K nearest neighbors of each prediction point, calculate the mean distance of their K nearest neighbors in the training set.
- Calculate the ratio as below.

$$\rho(\mathbf{x}_u) = \frac{\frac{1}{k} \cdot \sum_{i=1}^{k} \left\| \mathbf{x}_u - NN_i^{tr}(\mathbf{x}_u) \right\|}{\frac{1}{k^2} \cdot \sum_{i=1}^{k} \sum_{j=1}^{k} \left\| NN_i^{tr}(\mathbf{x}_u) - NN_j^{tr}(NN_i^{tr}(\mathbf{x}_u)) \right\|}.$$

Tax, David MJ, and Robert PW Duin. Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR). Springer, Berlin, Heidelberg, 1998.

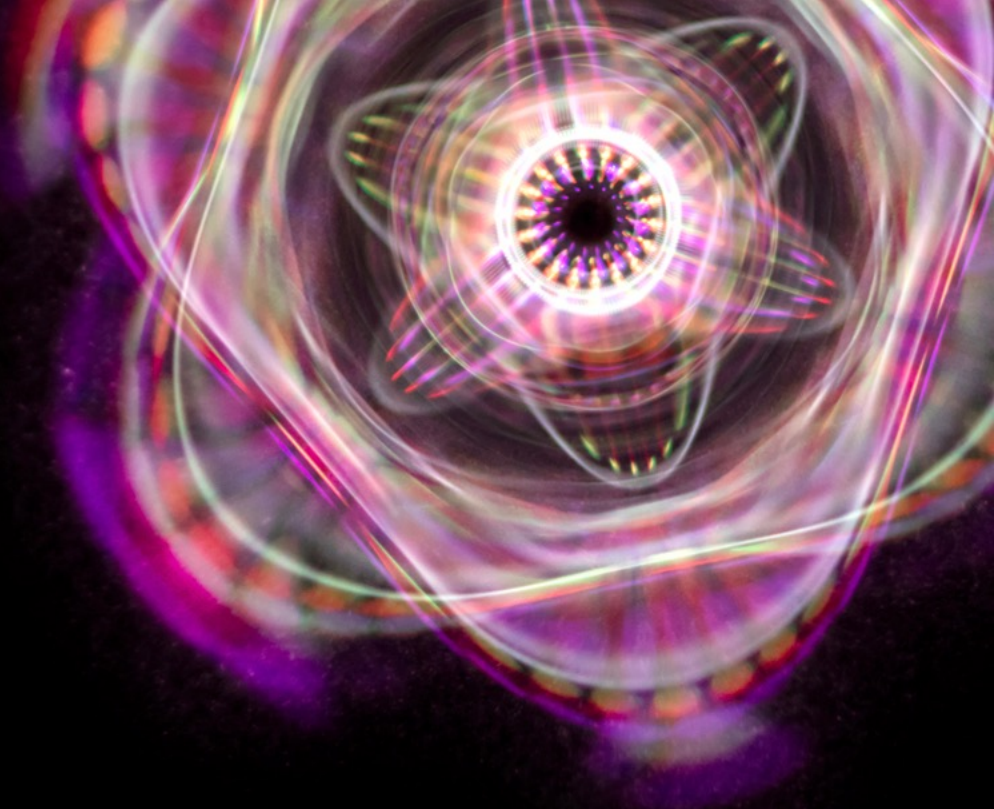
AMPL Tutorials available to run with Collab

1. Data Collection and creating Machine-Learning ready datasets
2. Model training and tuning
3. Hyper-parameter Optimization (HPO), Uncertainty Quantification (UQ), and using metrics for analyzing model performance.
4. Creating high-quality models
5. Model Inference



https://github.com/ATOMconsortium/AMPL/tree/Tutorials/atomsci/ddm/examples/tutorials

Thanks to Ravichandran Sarangan (FNL)

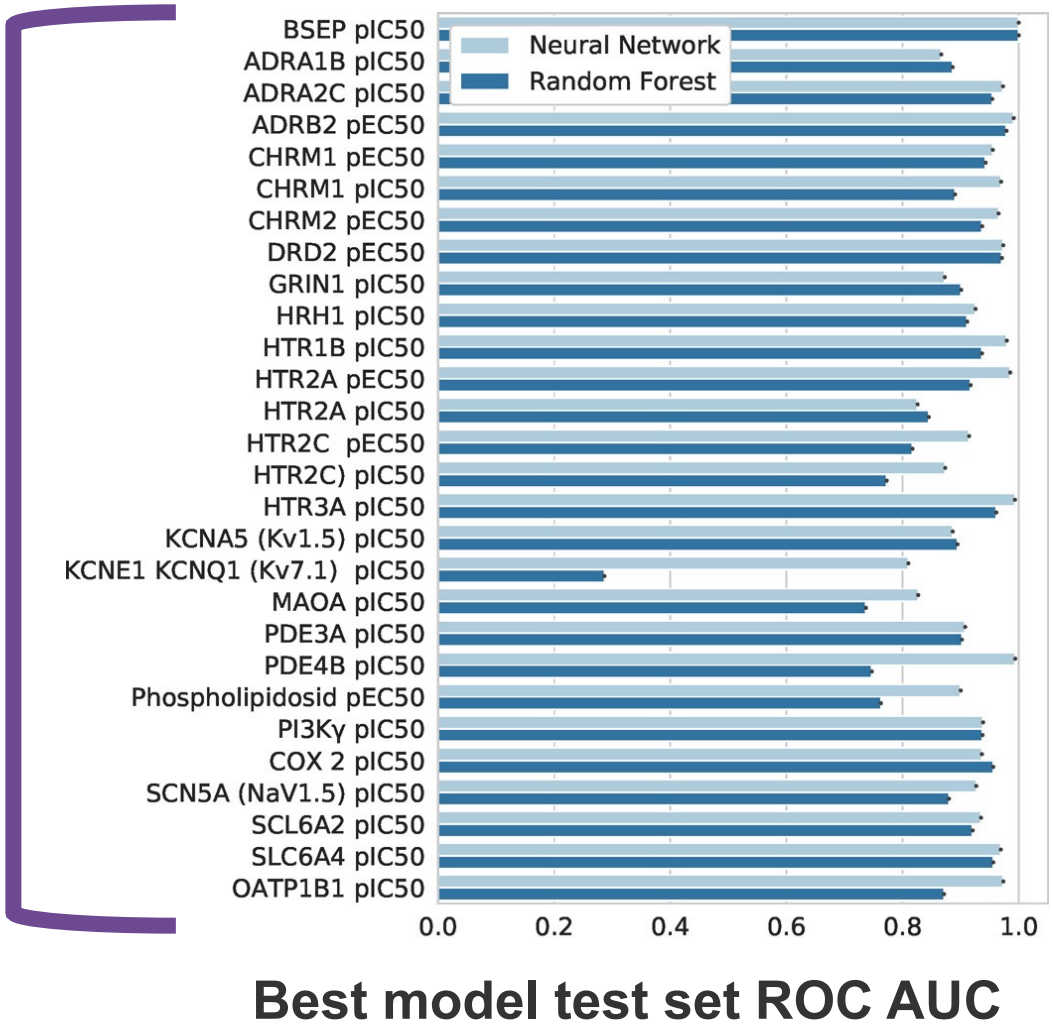


ATOM Modeling PipeLine validation

Safety validation classification models

Splitting method	Scaffold based
Model types	Neural network, random forest

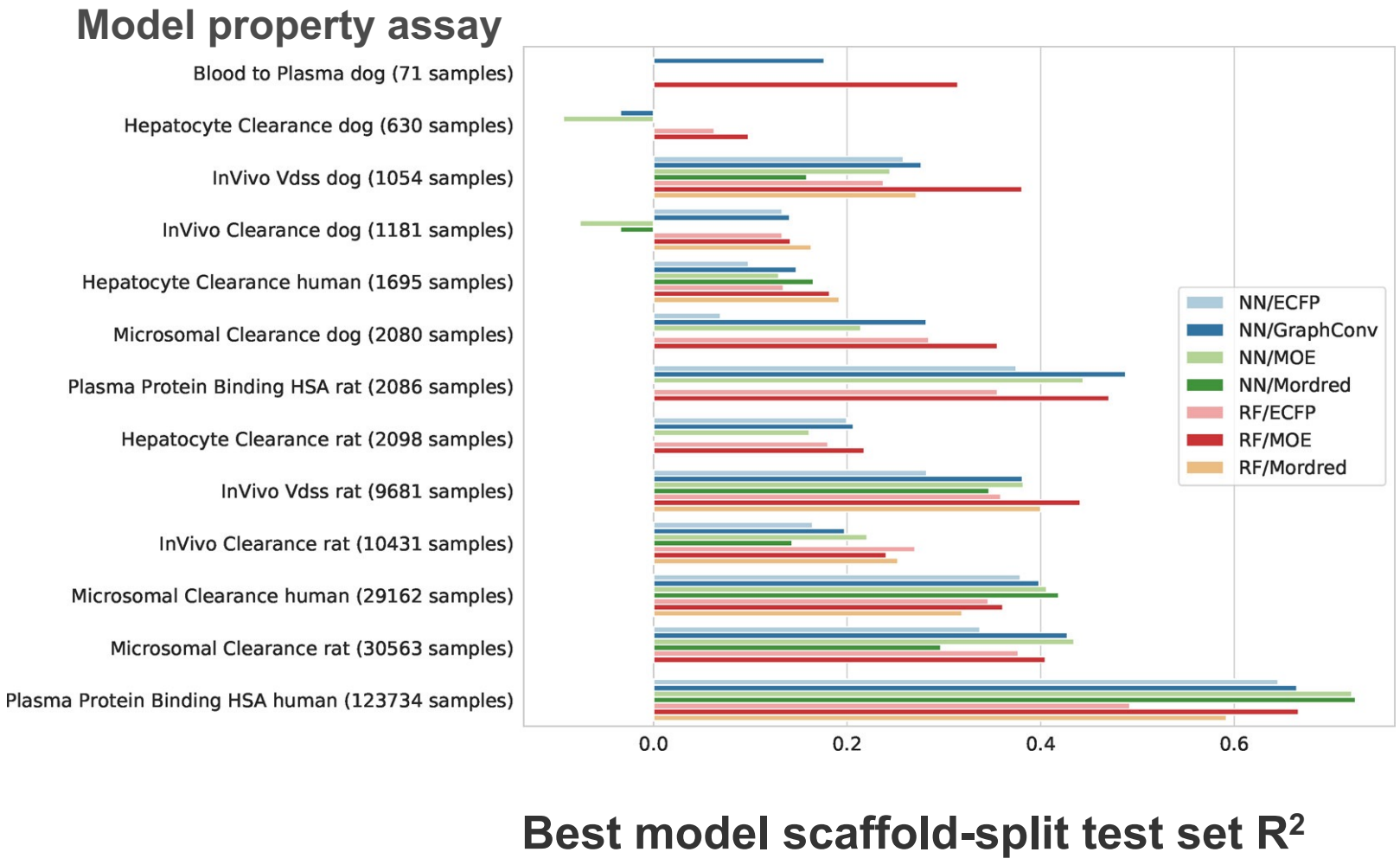
- Neural network and random forest models were able to differentiate between active and inactive test compounds on these tasks



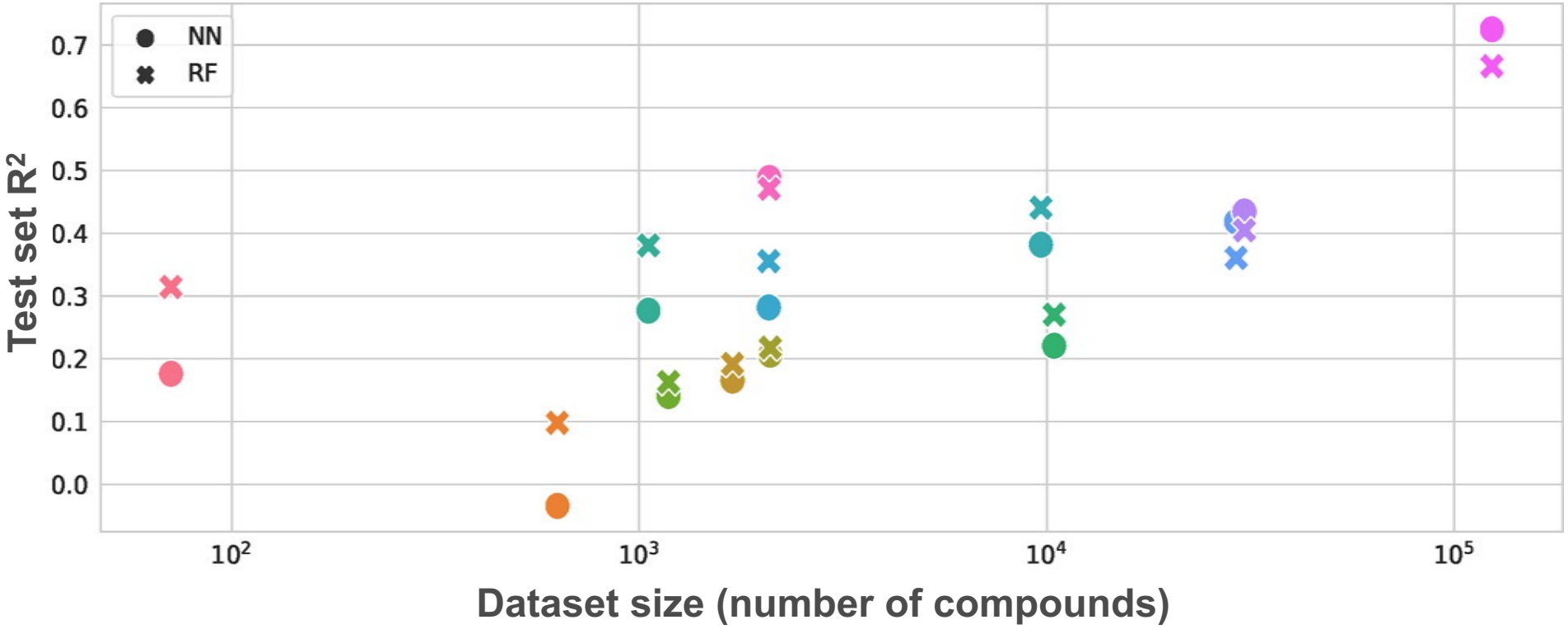
Pharmacokinetics validation regression models

Splitting method	Scaffold based
Features	Extended connectivity fingerprint (ECFP), graph convolution (DeepChem), Mordred, MOE
Model types	Neural network, random forest

- Neural network and random forest models were able to predict many PK properties

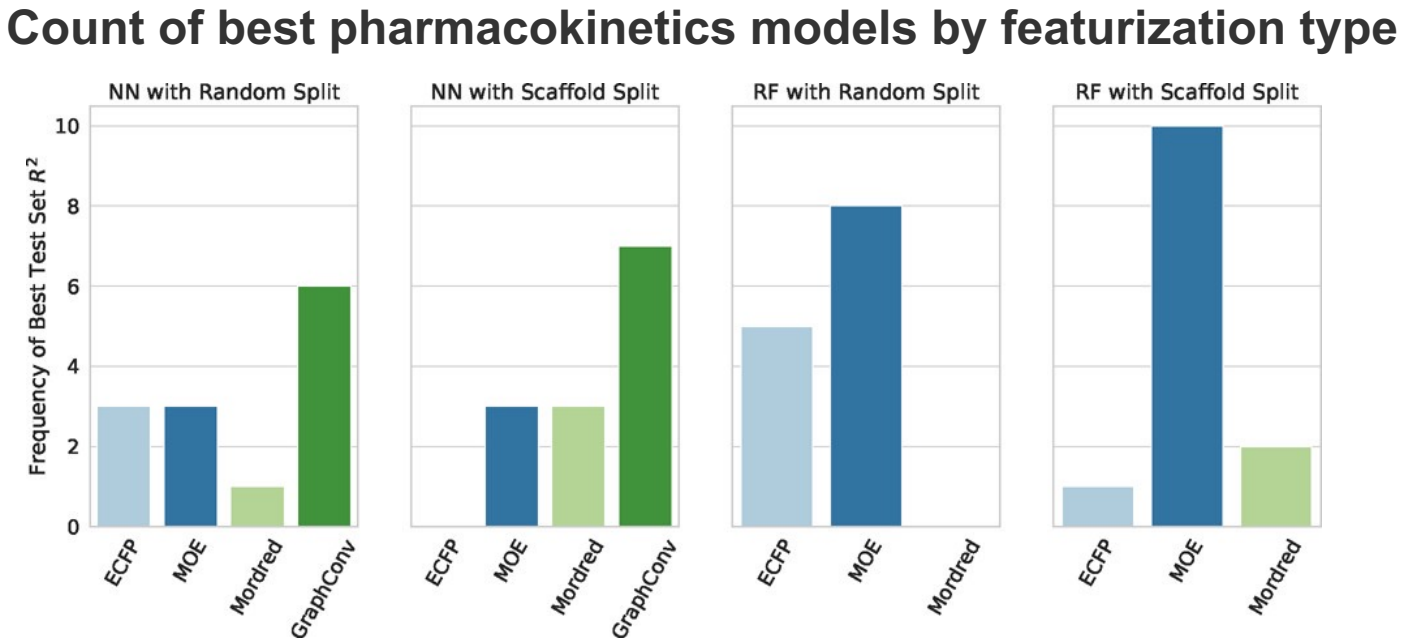


Effect of dataset size: Big data



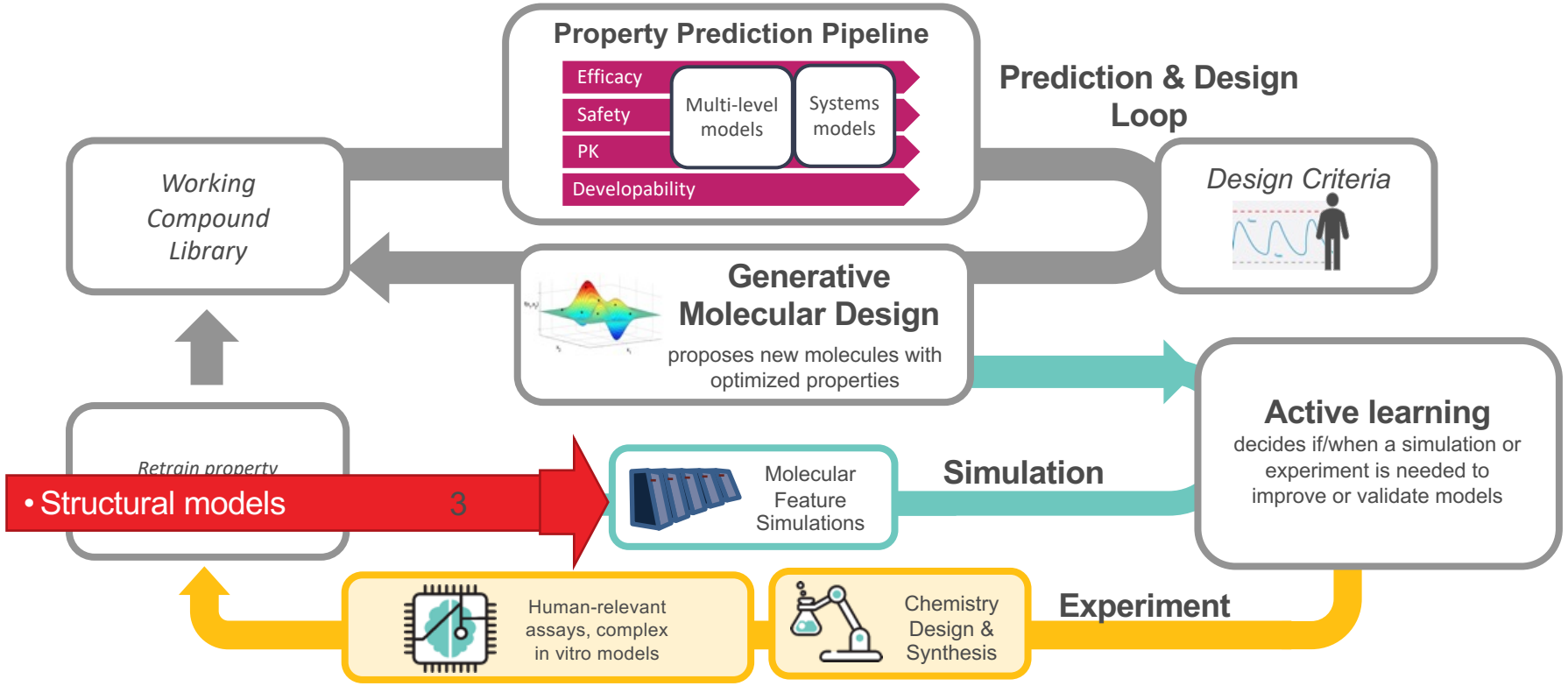
- Larger datasets were beneficial for fitting models for pharmacokinetics properties

Effect of feature and model types: Which is better?



- Graph convolutional DeepChem featurization worked well with neural network models
- MOE descriptors worked well with random forest models

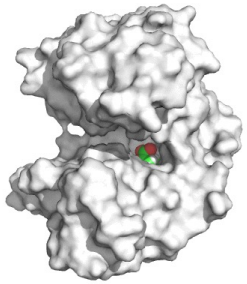
The ATOM Platform
Active Learning Drug Discovery Framework



Calculated protein interactions with new molecules presents scaling challenges for virtual screens

In a virtual screen we want to evaluate billions of virtual molecules:
a “needle in the haystack” problem

- Vina – speed=***moderate*** fast (1-2 minutes)
- Molecular Mechanics – Generalized Born / Surface Area (MM/GBSA) -- speed=moderate (62 minutes)
- Implicit solvent Molecular Dynamics (MD) = slower (7.2 hrs/GPU)
- Explicit solvent MD = slower (at least 7.2 hrs)

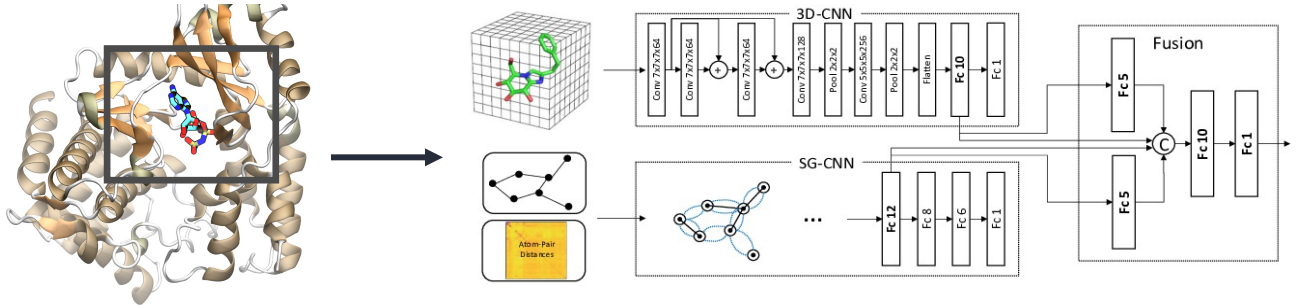


Physics based protein-ligand binding affinity does not scale to modeling billions of interactions

Two machine learning strategies currently employed

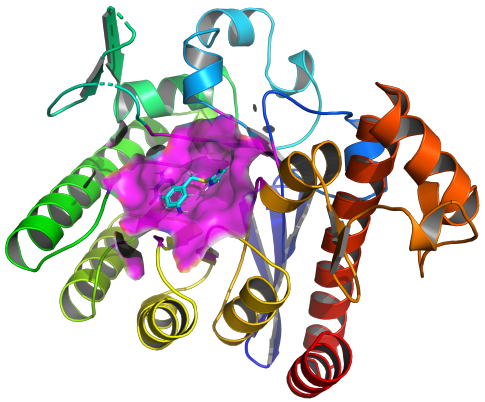
- Generate target specific scoring data using MM/GBSA
  - Use ML model to learn scoring function (surrogate model)
    - Pros: Develop a faster scoring function that could match MM/GBSA accuracy
    - Cons: MM/GBSA scores still have limitations in accuracy
- Use 3D structure based spatial information to learn across multiple targets
  - Pros: Train on experimental binding data, apply to any new target (within reason-relative to training data)
  - Cons: Requires some 3D structure of the protein and a pocket

Fusion models for Atomic and molecular STructures (FAST)



- 3D-CNNs have been used by numerous teams starting with AtomNet in 2015. (AtomWise)
- 3D Spatial Graphs were introduced with PotentialNet in 2018. (Genesis Therapeutics)
- No publications comparing the approaches directly
- Our results suggest potential benefits for combining two approaches
- Open Source: https://github.com/llnl/fast
- Paper: Jones, D., Kim, H et al., 2021 JCIM (https://pubs.acs.org/doi/full/10.1021/acs.jcim.0c01306)

Extract atomic features that generalize across multiple targets



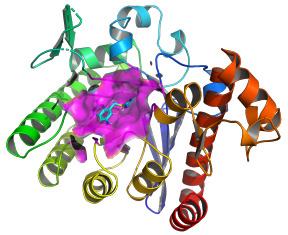
- Element type: one-hot encoding of B, C, N, O, P, S, Se, halogen or metal
- Atom hybridization (1, 2, or 3)
- Number of heavy atom bonds (i.e., heavy valence)
- Number of bonds with other heteroatoms
- Structural properties: bit vector (1 where present) encoding of hydrophobic, aromatic, acceptor, donor, ring
- Partial charge
- Molecule type to indicate protein atom versus ligand atom (-1 for protein, 1 for ligand)
- Van der Waals radius

Model is trained on existing experimentally solved structures

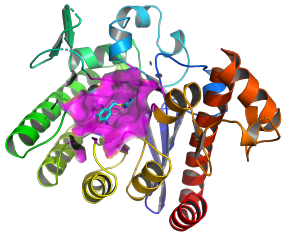
Models trained on a dataset called 2016 version of PDBBind http://www.pdbbind.org.cn/
Training size = 13,308 complexes

Current training size (2019):
17,679 samples

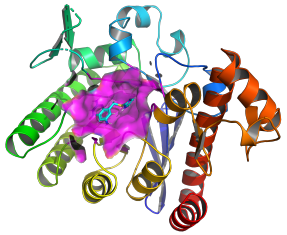
Ligand A
Protein A + Ki



Ligand B
Protein B + Ki

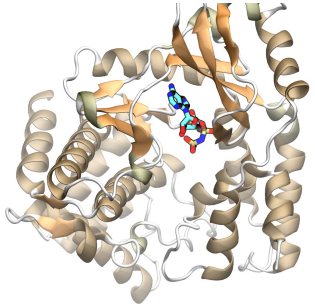


Ligand C
Protein C + Ki

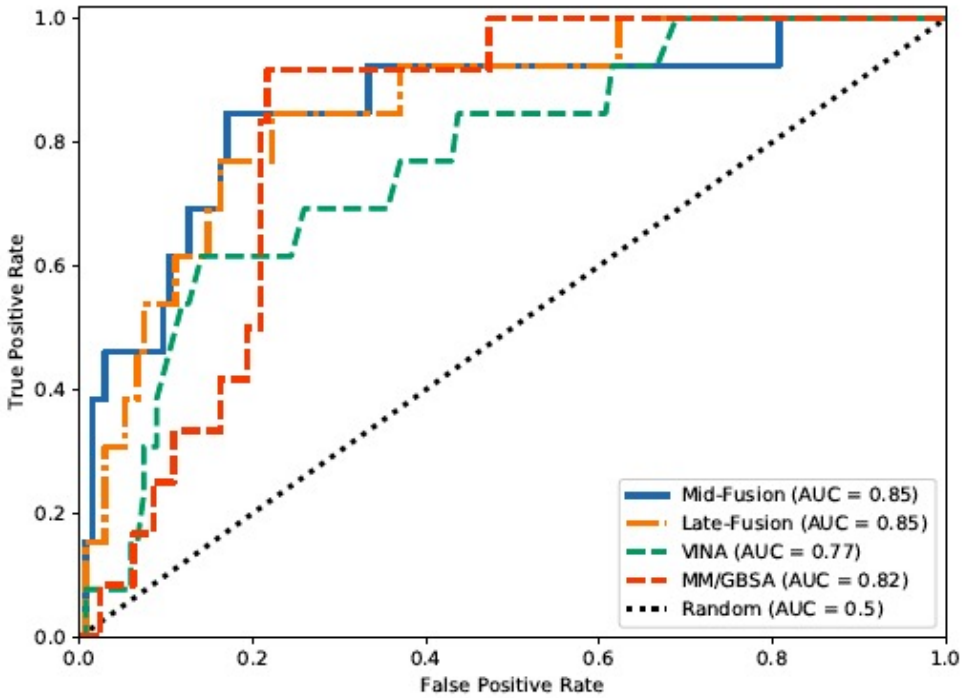


…..

Created a special hold out set – structures taken from 2019 with a detailed analysis to find structurally novel pockets and novel ligands – 222 complexes.



Fusion model performs well compared to other costly methods



Fusion model provides a more scalable alternative or compliment to more expensive scoring functions
Fusion model scores 108 poses per second (with 4 compute nodes) and is 403 times faster than MM/GBSA

May still have model uncertainty with new parts of chemical space

Conclusion

- There is no universal optimal model that can be applied to every new dataset. General heuristics:
  - Smaller datasets: Random Forests with MOE descriptors
  - Larger datasets: Neural Networks with descriptors or graph learned features
  - Static fingerprint methods tend to be less competitive
- Quantifying model uncertainty remains an open but important challenge
- In the absence of data, machine learning models can still be used by exploiting cross-target learning
- Presented tools are open-source software to support computational drug discovery in non-commercial settings

Acknowledgements

Current Computational Tech Team

- Kevin McLoughlin (GMD/DM)
- Amanda Paulson (SM)
- Jeff Mast (GMD)
- Ravichandran Sarangan (DM)
  - **(Organizing tutorials)**
- Derek Jones (GMD)
- Marisa Torres (DM)
- Sergio Wong (MM)
- Dan Kirshner (MM)
- Brian Bennion (MM)
- Hyojin Kim (MM)
- Garrett Stevenson (MM)
- Da Shi (GMD/DM)
- Jessica Mauvais (DM)
- Xiaohua Zhang (MM)
- Sam Jacobs (GMD)
- Brian Van Essen (GMD)

Past Team Members

- Jason Deng (GMD)
- Amanda Minnich (DM)
- Tom Sweitzer (GMD)
- Juliet McComas (GMD)
- Margaret Tse (SM/DM)
- Michael Gunshenan (DM)
- Andrew Weber (GMD/SM)
- Stacie Calad-Thomson (JRC)
- Kishore Pasikanti (SM)
- Neha Murad (SM)
- Benjamin Madej (SM)
  - **(Special thanks for contributing background slides)**

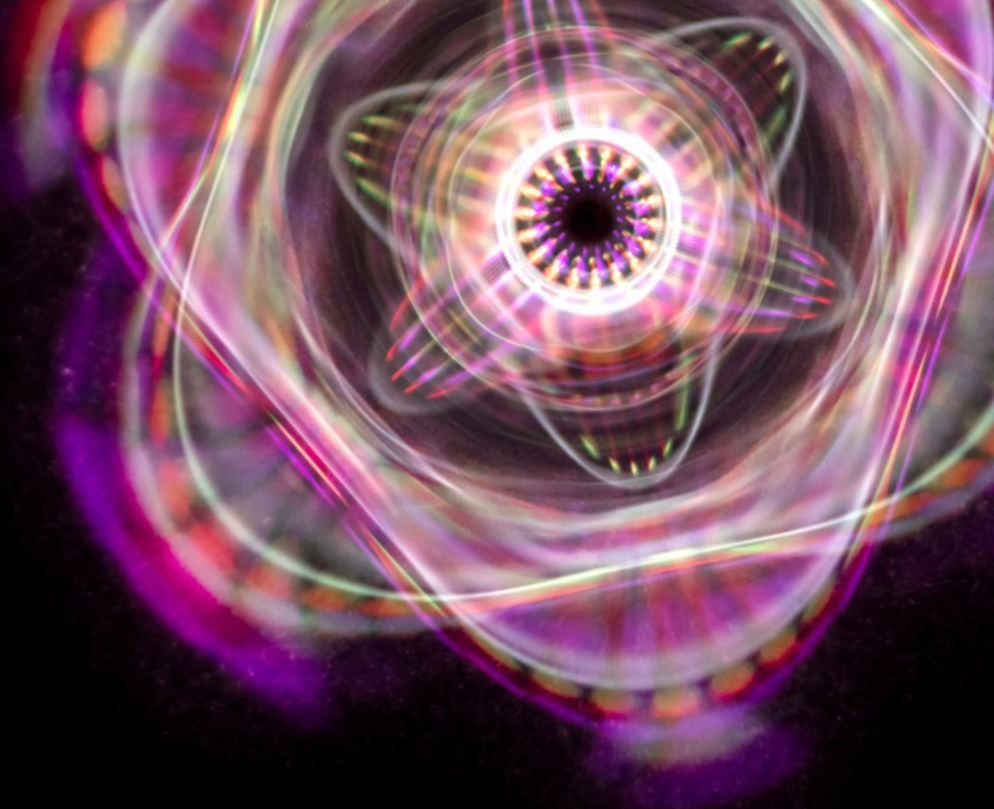
ATOM Joint Research Committee (JRC)

- Eric Stahlberg (FNL)
- Jim Brase (LLNL)
- Michelle Arkin (UCSF)
- Dwight Nissley (FNL)
- Marti Head (ORNL)
- Tom Brettin (ANL)
- Rick Stevens (ANL)

FUNDING

- DoD – DTRA
- ATOM
  - NNSA-DOE, GSK, UC, NCI
- American Heart Association
- LLNL Laboratory Directed Research

SM=systems modeling DM=data modeling, GMD=design loop



Questions?