ATOM

Using GPUs to Generate Reproducible Workflows to Accelerate Drug Discovery

Amanda J. Minnich

Staff Research Scientist

Lawrence Livermore National Laboratory

GPU Technology Conference | March 21, 2019

LLNL-PRES-769348

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

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

ATOM:
Accelerating Therapeutics for Opportunities in Medicine

Founding Members

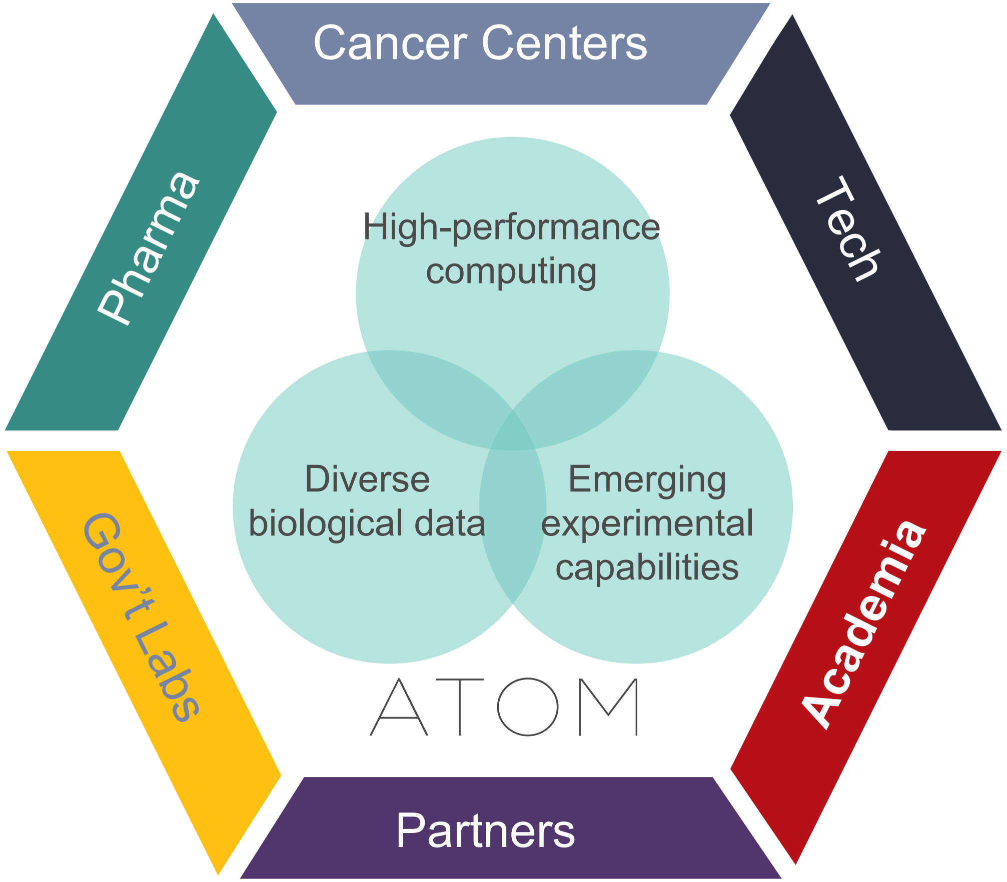
gsk

Lawrence Livermore National Laboratory

UCSF

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute



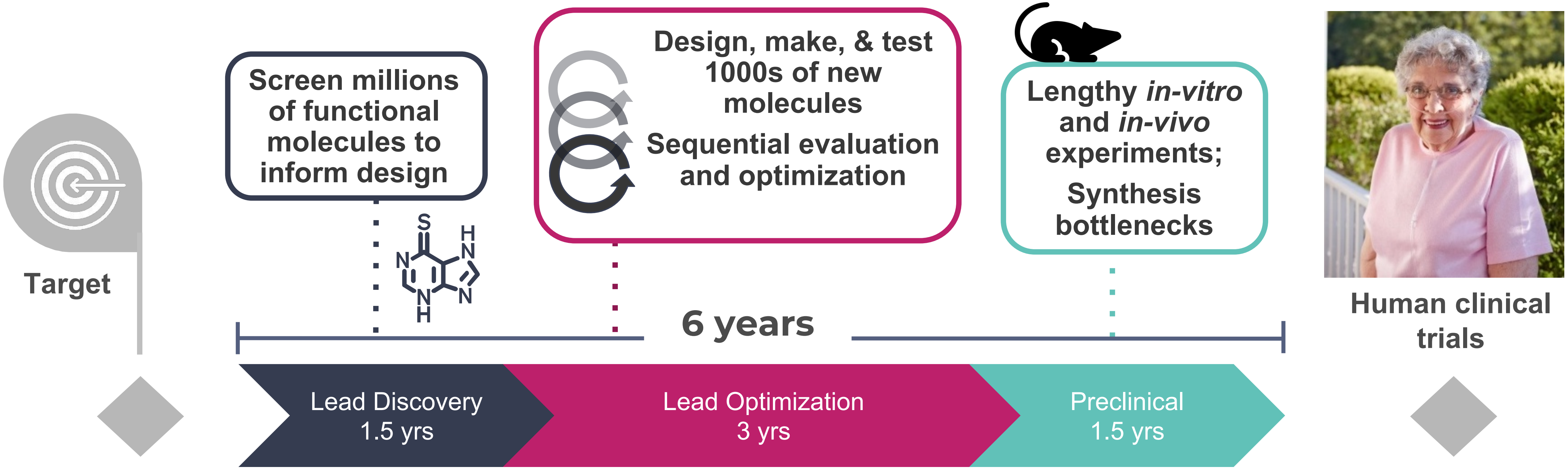
What is ATOM?

- **Approach**: An open public-private partnership
  - Lead with computation supported by targeted experiments
  - Data-sharing to build models using everyone’s data
  - Build an open-source framework of tools and capabilities

- **Status**:
  - Shared collaboration space at Mission Bay, SF
  - 25 FTE’s engaged across the partners
  - **R&D started March 2018**
  - In the process of engaging new partners

Current drug discovery: long, costly, high failure

Is there a better way to get medicines to patients?



Target

Screen millions of functional molecules to inform design

Design, make, & test 1000s of new molecules

Sequential evaluation and optimization

Lengthy *in-vitro* and *in-vivo* experiments;

Synthesis bottlenecks

Human clinical trials

6 years

Lead Discovery
1.5 yrs

Lead Optimization
3 yrs

Preclinical
1.5 yrs

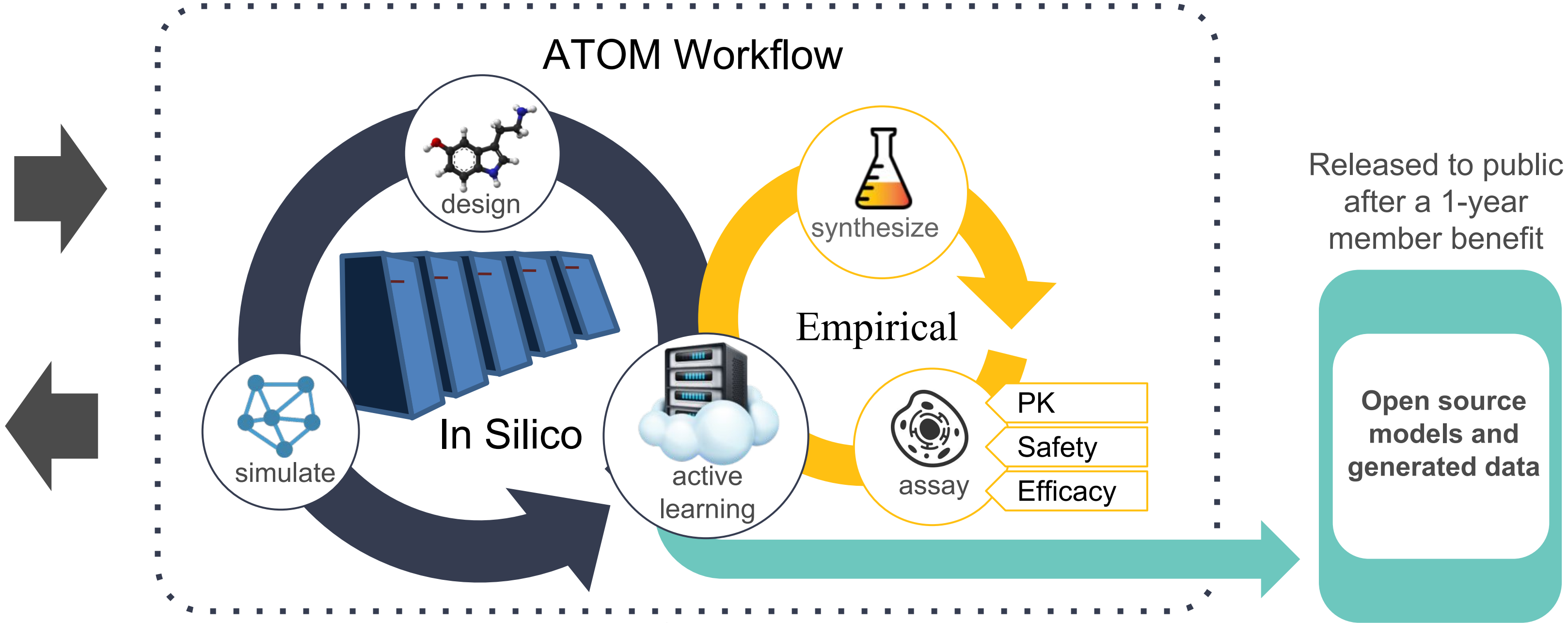
- **33% of total cost of medicine development**
- **Clinical success only ~12%, indicating poor translation in patients**

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

Accelerated drug discovery concept

Vision of ATOM workflow in practice

Patient-specific data and samples input to workflow to develop new therapeutics



ATOM Workflow

design

simulate

In Silico

active learning

synthesize

Empirical

assay

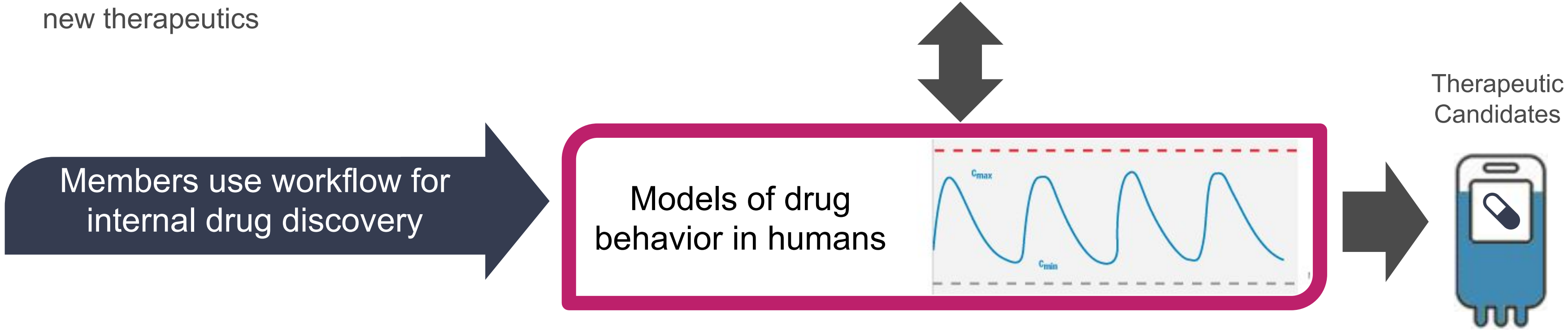
PK

Safety

Efficacy

Released to public after a 1-year member benefit

Open source models and generated data



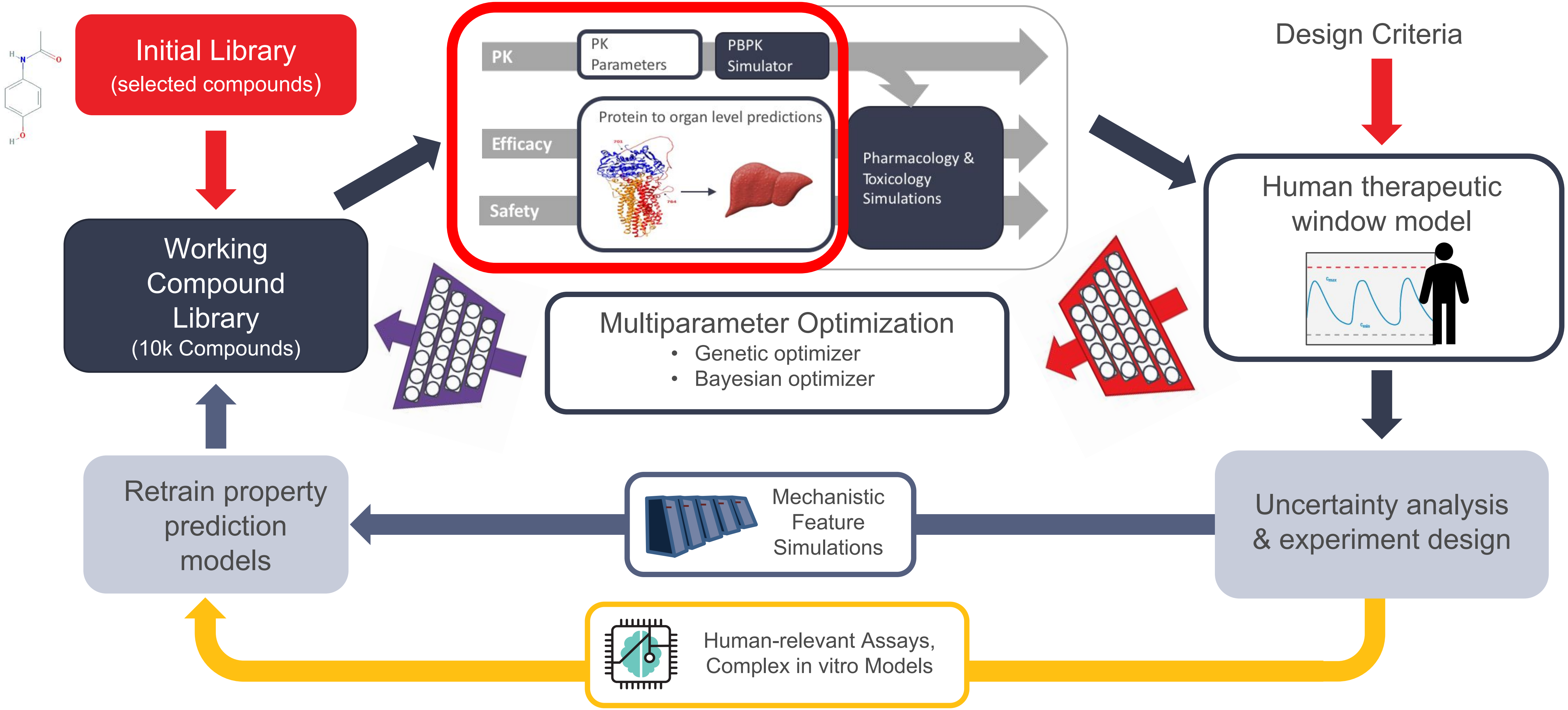
Members use workflow for internal drug discovery

Models of drug behavior in humans

Therapeutic Candidates

Commercialization by members for patient benefit

Top-level view of the ATOM molecular design platform



Initial Library
(selected compounds)

Working Compound Library
(10k Compounds)

PK

PK Parameters

PBPK Simulator

Efficacy

Safety

Protein to organ level predictions

Pharmacology & Toxicology Simulations

Design Criteria

Human therapeutic window model

Multiparameter Optimization
- Genetic optimizer
- Bayesian optimizer

Retrain property prediction models

Mechanistic Feature Simulations

Uncertainty analysis & experiment design

Human-relevant Assays, Complex in vitro Models

Software framework is being released as open source

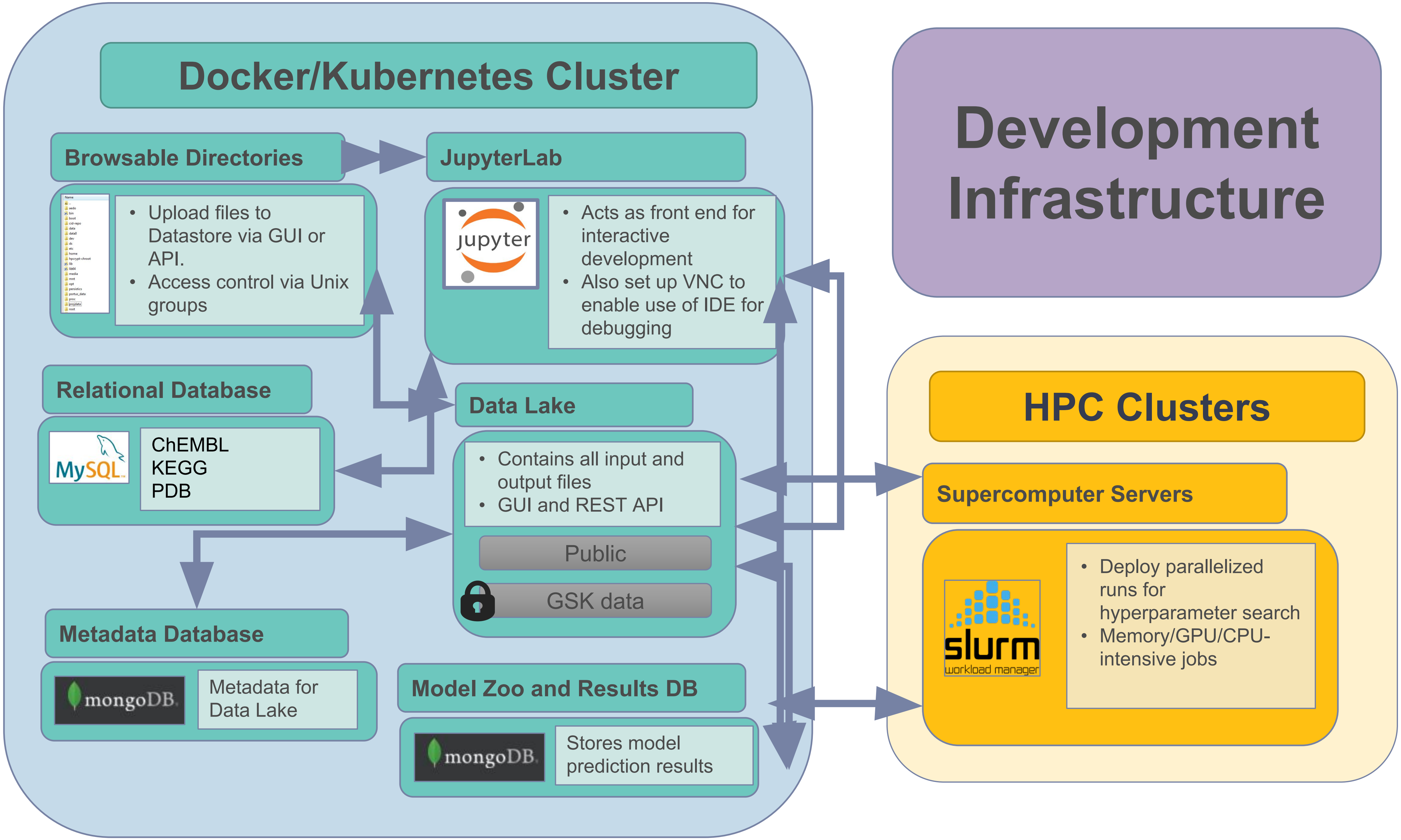
Roadmap

- Infrastructure and Architecture – what GPUs are we using?
- Data-Driven Modeling Pipeline – what have we built?
- Experiments – what have we been able to do?
- Future work – where are we going from here?

Roadmap

- Infrastructure and Architecture – what GPUs are we using?
- Data-Driven Modeling Pipeline – what have we built?
- Experiments – what have we been able to do?
- Future work – where are we going from here?

Development Infrastructure



Docker/Kubernetes Cluster

Browsable Directories

- Upload files to Datastore via GUI or API.
- Access control via Unix groups

JupyterLab

- Acts as front end for interactive development
- Also set up VNC to enable use of IDE for debugging

Relational Database

ChEMBL
KEGG
PDB

Data Lake

- Contains all input and output files
- GUI and REST API

Public

GSK data

Metadata Database

Metadata for Data Lake

Model Zoo and Results DB

Stores model prediction results

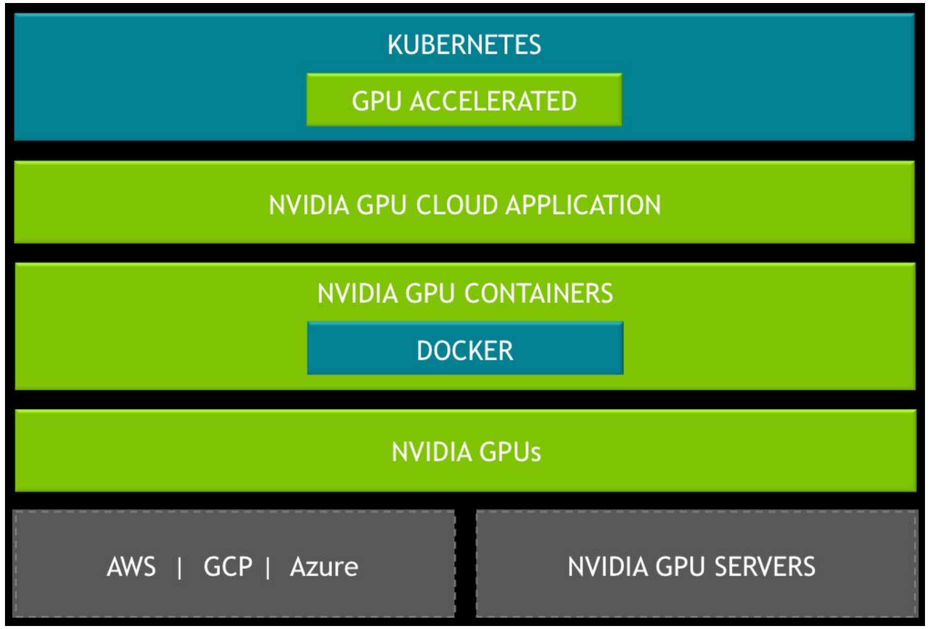
HPC Clusters

Supercomputer Servers

- Deploy parallelized runs for hyperparameter search
- Memory/GPU/CPU-intensive jobs

Kubernetes allocates GPU resources on our development server

kubernetes by Google



- Our development server has 4 GPU nodes with 4 Titan XPs in each node
- 1 data server (cephid), 1 login/head node
- Kubernetes is an open source container orchestrator
- Manages containerized workloads and services
- Use it to orchestrate allocation of GPUs, CPUs, and memory
- Handles Role-Based Access Control

LLNL HPC Software Specs and Computer Architecture

- Nodes: 164
- Cores/Node: 36
- Total Cores: 5,904
- Memory/Node: 256
- Total Memory: 41,984 GB
- GPU Architecture: NVIDIA Tesla P100 GPUs
- Total GPUs: 326

- GPUs per compute node: 2
- GPU peak performance (TFLOP/s double precision): 5.00

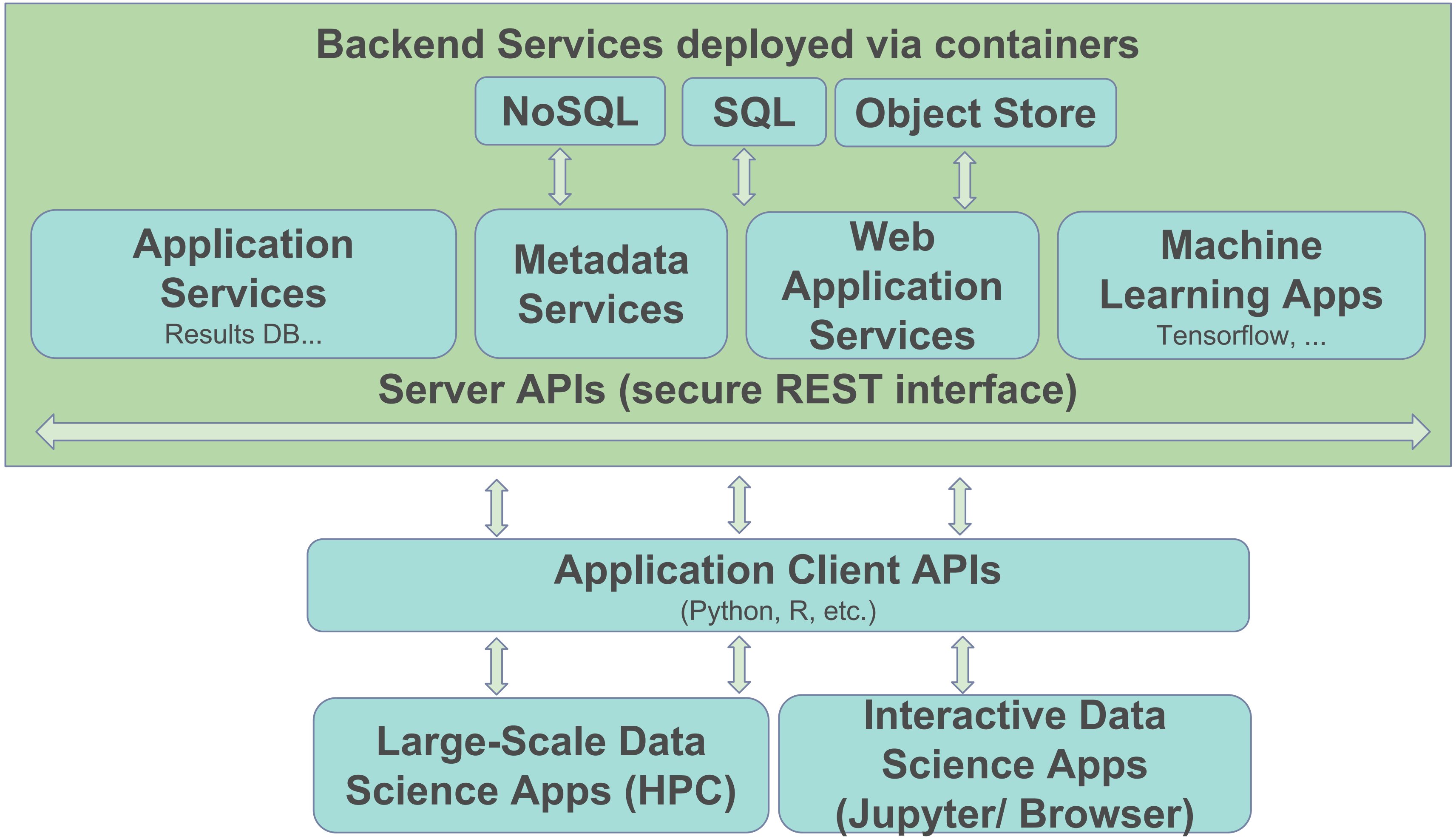
Data services are a necessity

- Data services are required to organize:
  - Raw data
  - Curated datasets
  - Model-ready datasets
  - Train/test/validation split of datasets
  - Serialized models
  - Performance results
  - Simulation output
- These data types vary in size, format, and level of organization/complexity

Have a variety of services to handle our needs

- Data Lake
  - In-house object store service
  - Allows for association of complex metadata with any type of file
  - Can access via GUI and REST API
- mongoDB
  - Used as backend for Data Lake metadata
  - Used as backend for Model Zoo metadata
  - Used for Results DB
- MySQL
  - Many public datasets are available in SQL format

Overall structure of data services



Backend Services deployed via containers

NoSQL

SQL

Object Store

Application Services
Results DB...

Metadata Services

Web Application Services

Machine Learning Apps
Tensorflow, ...

Server APIs (secure REST interface)

Application Client APIs
(Python, R, etc.)

Large-Scale Data Science Apps (HPC)

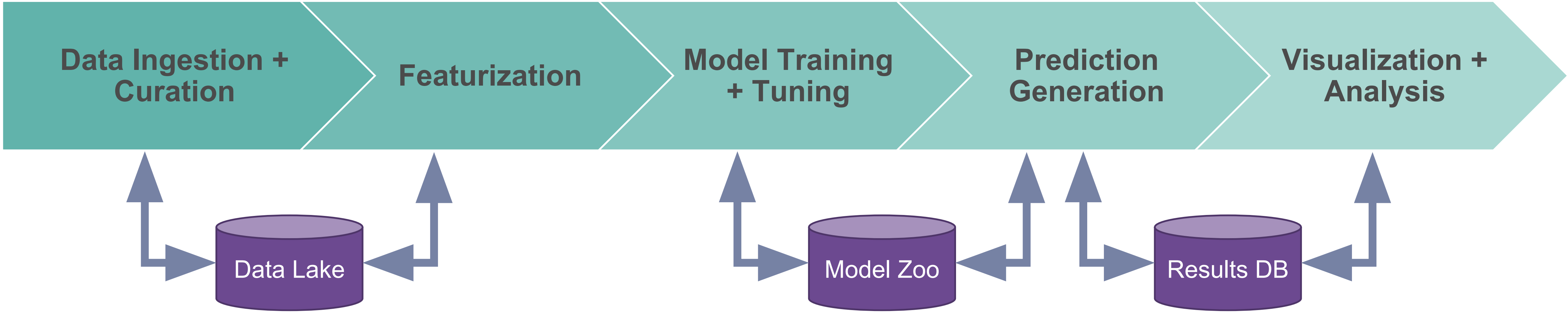
Interactive Data Science Apps (Jupyter/ Browser)

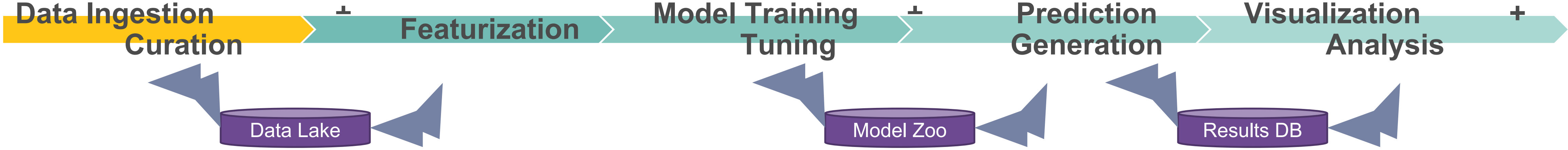
Roadmap

- Infrastructure and Architecture – what GPUs are we using?
- Data-Driven Modeling Pipeline – what have we built?
- Experiments – what have we been able to do?
- Future work – where are we going from here?

End-to-End Data-Driven Modeling Pipeline

Enables portability of models and reproducibility of results





Data Ingestion + Curation

Featurization

Model Training + Tuning

Prediction Generation

Visualization + Analysis

Data Lake

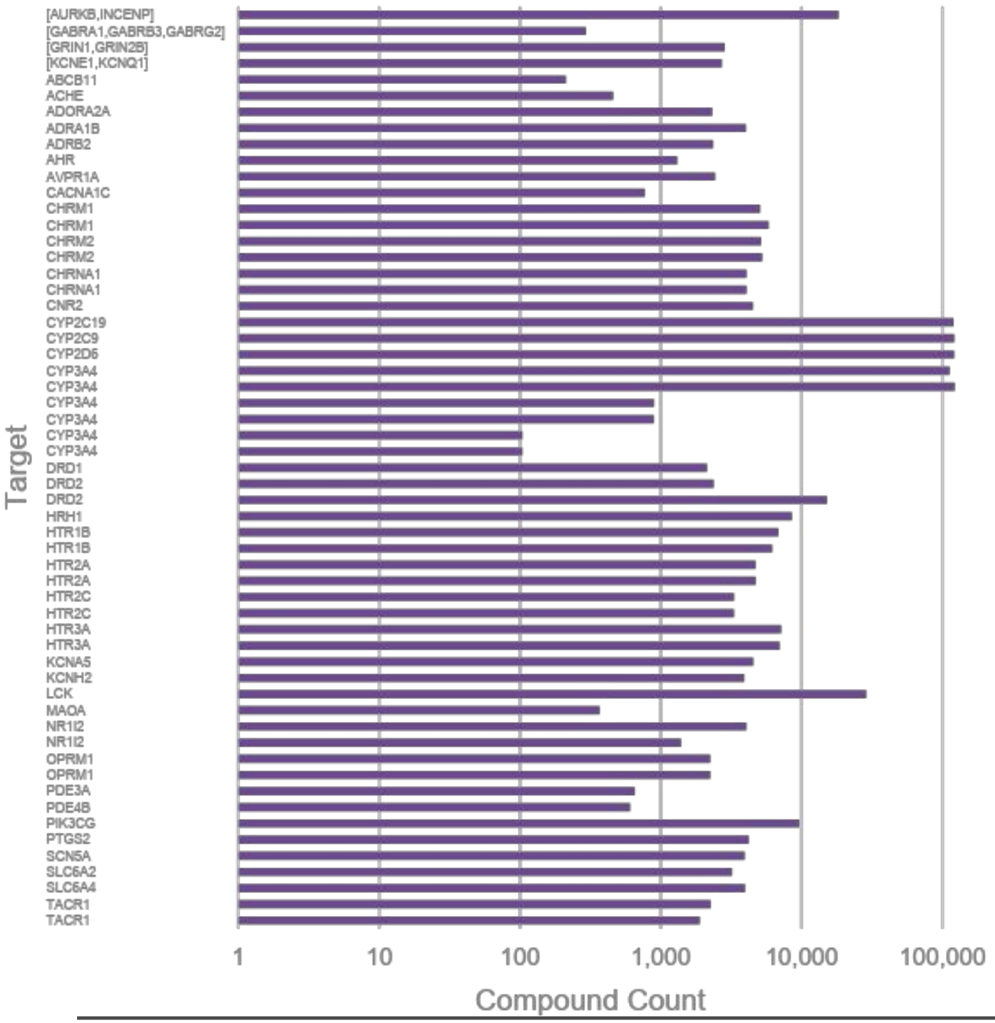
Model Zoo

Results DB

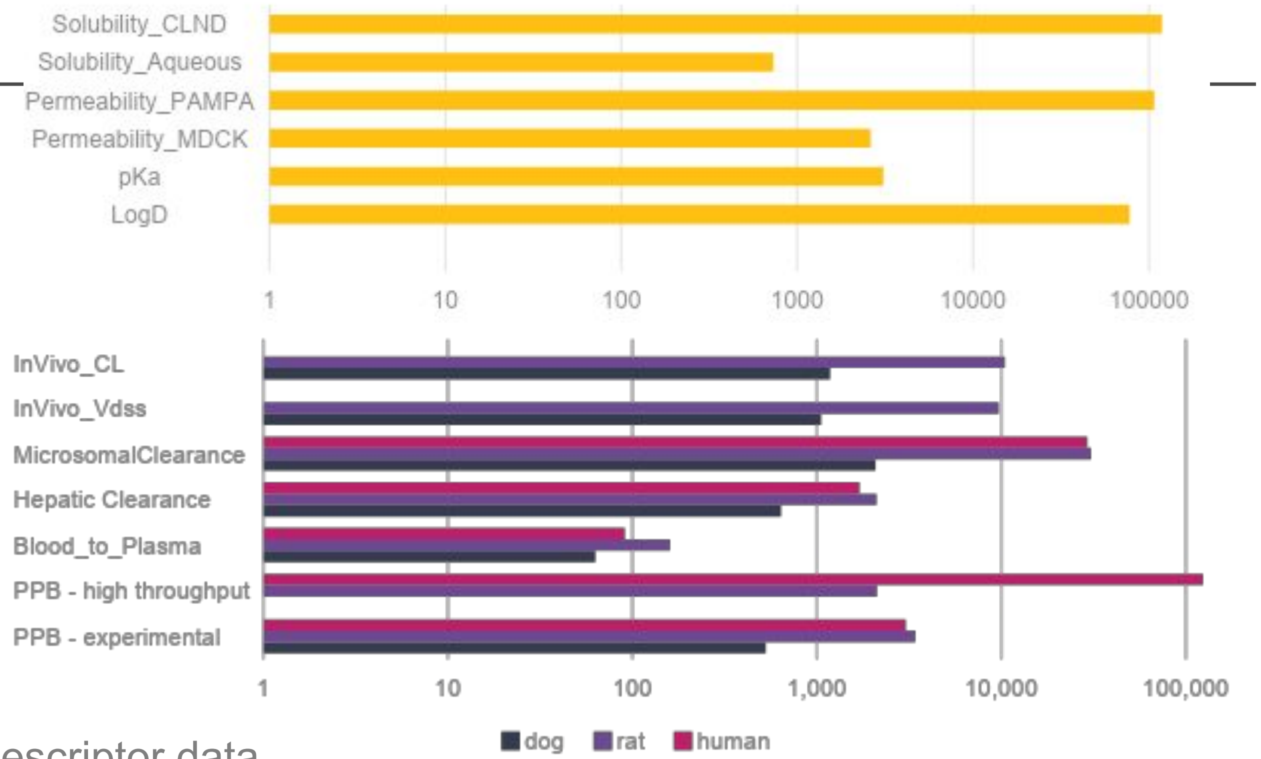
- Raw pharma data consists of 300 GB of a variety of bioassay and animal toxicology data on ~2 million compounds from GSK
- Proprietary or sensitive data must only be stored on approved servers
- Data may need to remain sequestered from other members

ATOM has curated ~150 model-ready data sets

GSK Safety Datasets

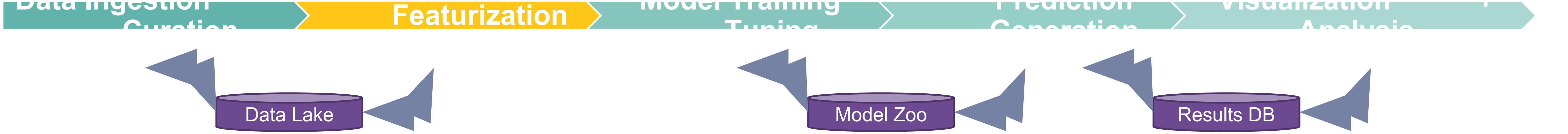


GSK Pharmacokinetic Datasets



Descriptor data

Data Set	MOE	3D Descriptors	Compounds
GSK			1.86M
ChEMBL			1.6M
Enamine			680M



Data Ingestion Curation

Featurization

Model Training Tuning

Prediction Generation

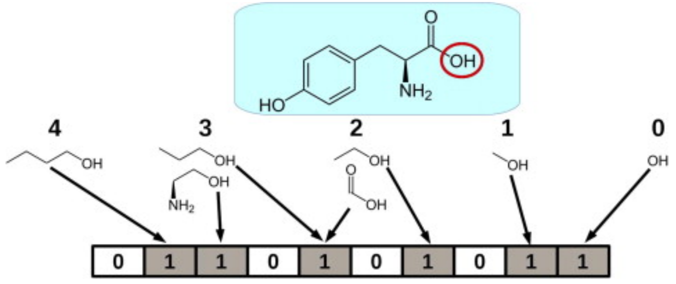
Visualization Analysis

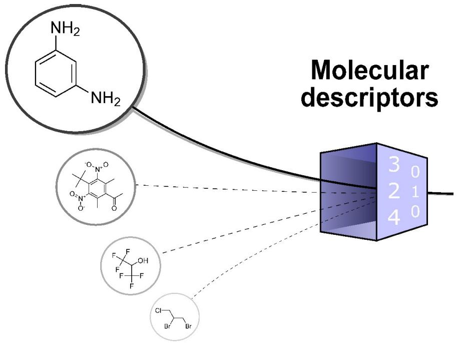
Data Lake

Model Zoo

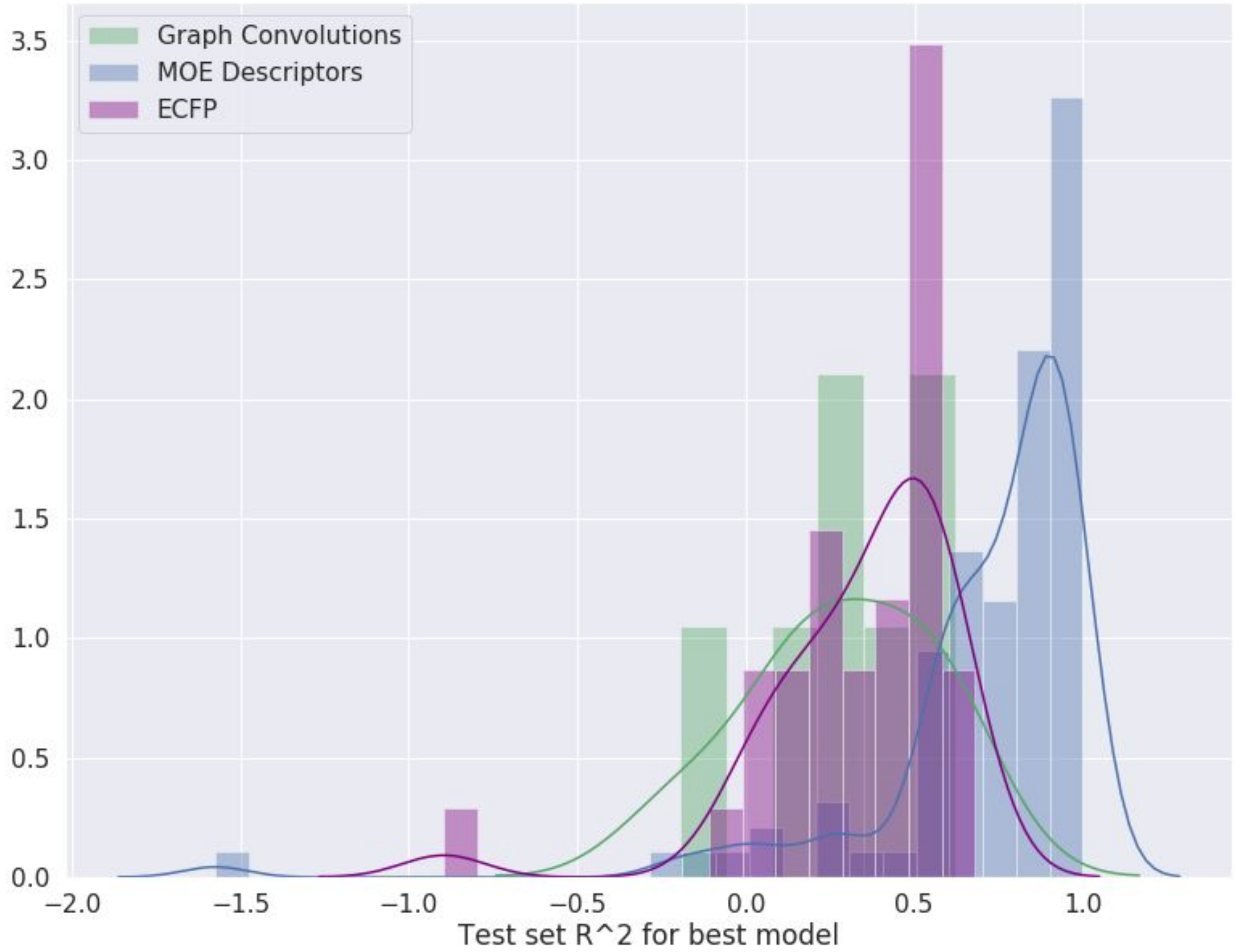
Results DB

- Support loading datasets from either Data Lake or filesystem
- Support a variety of feature types
  - Extended Connectivity Fingerprint
  - Graph-based features
  - Molecular descriptor-based features (MOE, DRAGON7, rdkit)
  - Autoencoder-based features (MolVAE)
  - Allow for custom featurizer classes
- Split dataset based on structure to avoid bias



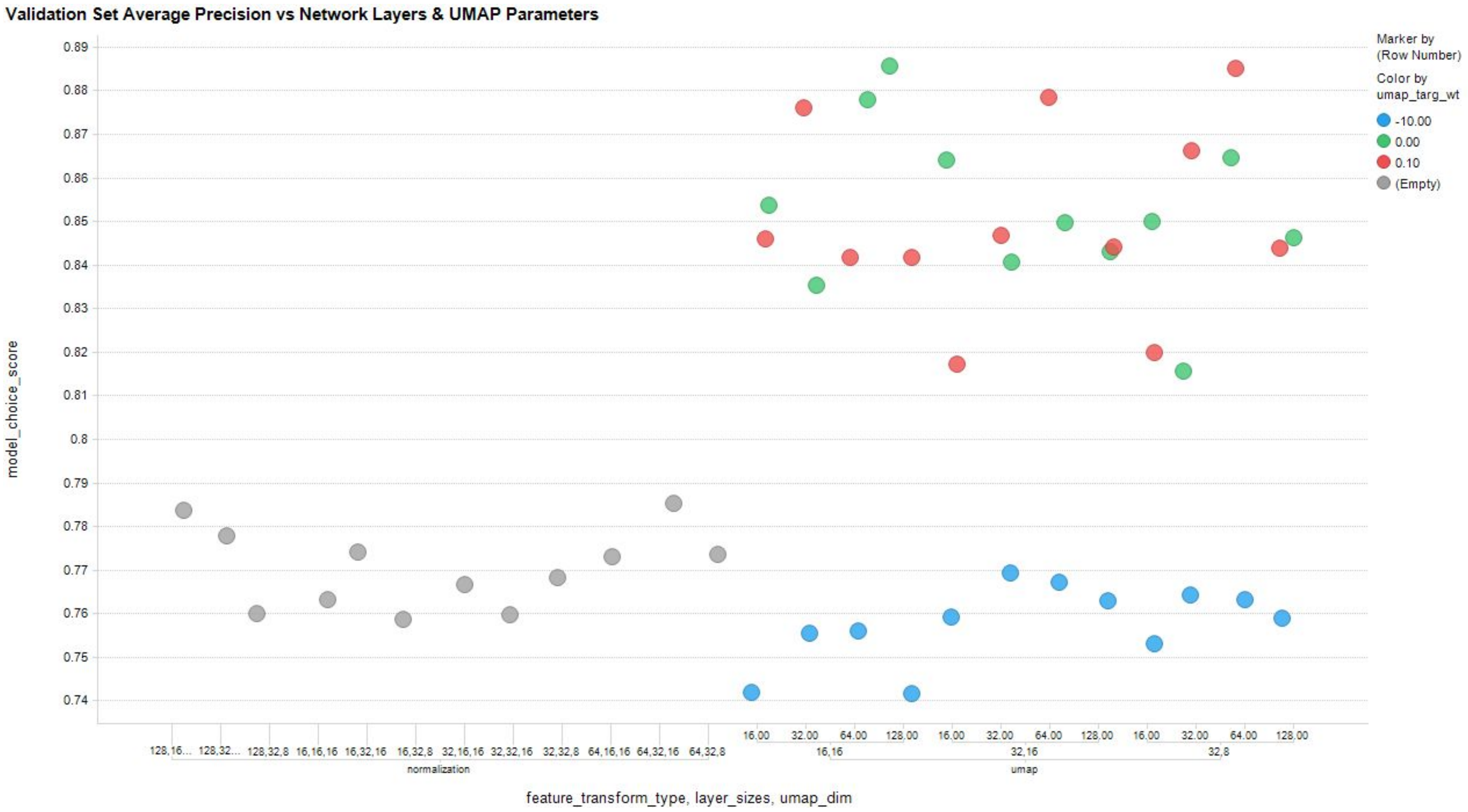


Featurization is key

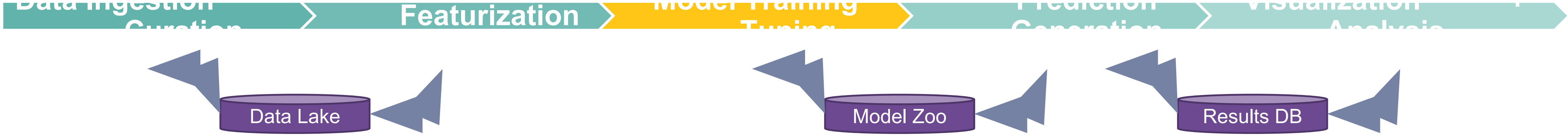


- We have found that the best-performing feature type varies by dataset
- In general chemical descriptors out-perform other feature types
- Graph Convolutions occasionally outperform others

Dimensionality reduction can improve performance



ATOM



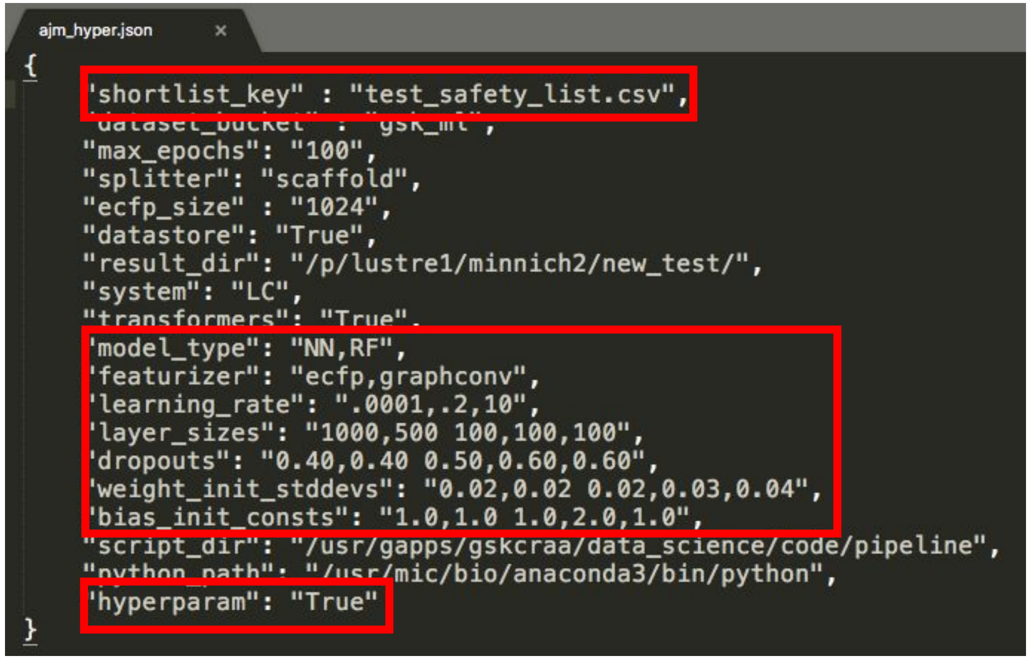
- Have built a train/tune/predict framework to create high-quality models
- Currently support:
  - sklearn models
  - deepchem models (wrapper for TensorFlow)
  - Allow for custom model classes
- Tune models using the validation set and perform k-fold cross validation

deepchem scikit learn TensorFlow PyTorch

Hyperparameter optimization

Support distributed hyperparameter search for dataset/feature/model combinations

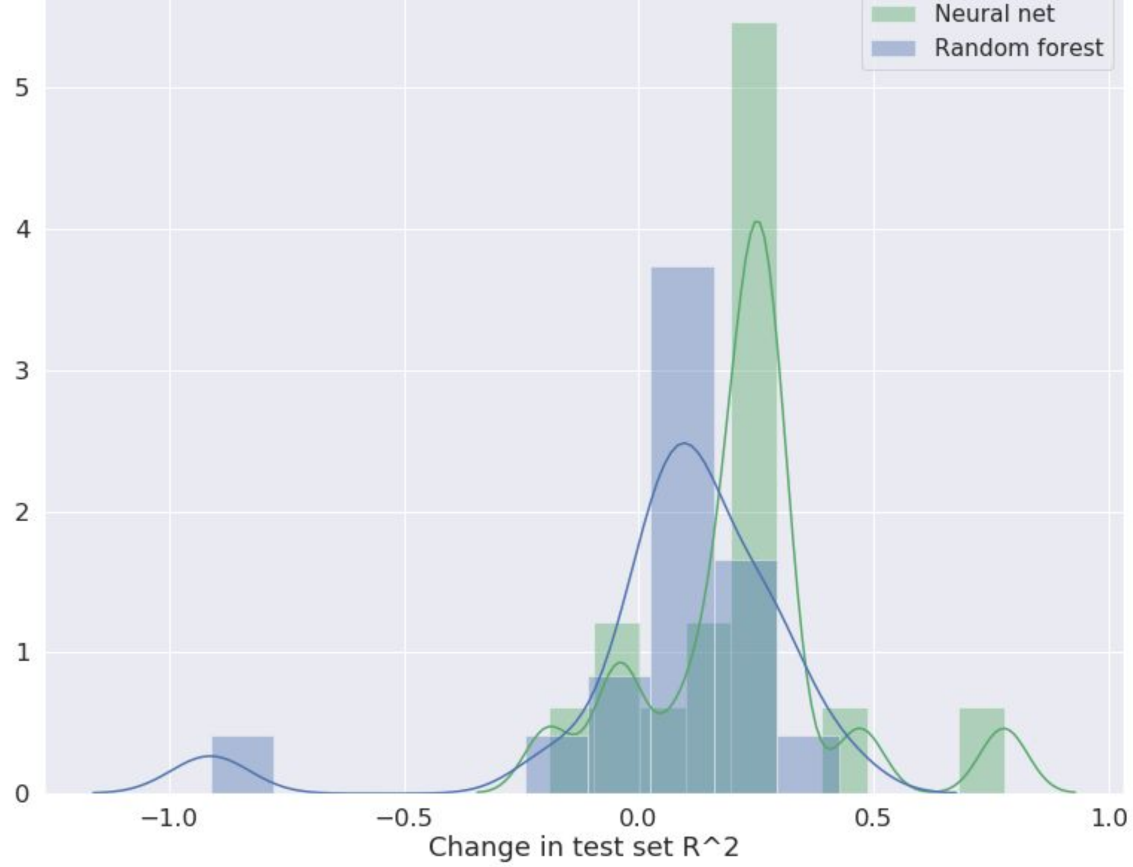
- Support linear grid, logistic grid, random, and user-specified steps
- Currently does not support optimization
- Specify input with JSON file or command line
- Generates all possible combinations of hyperparams, accounting for model type
- Groups neural net architecture combinations



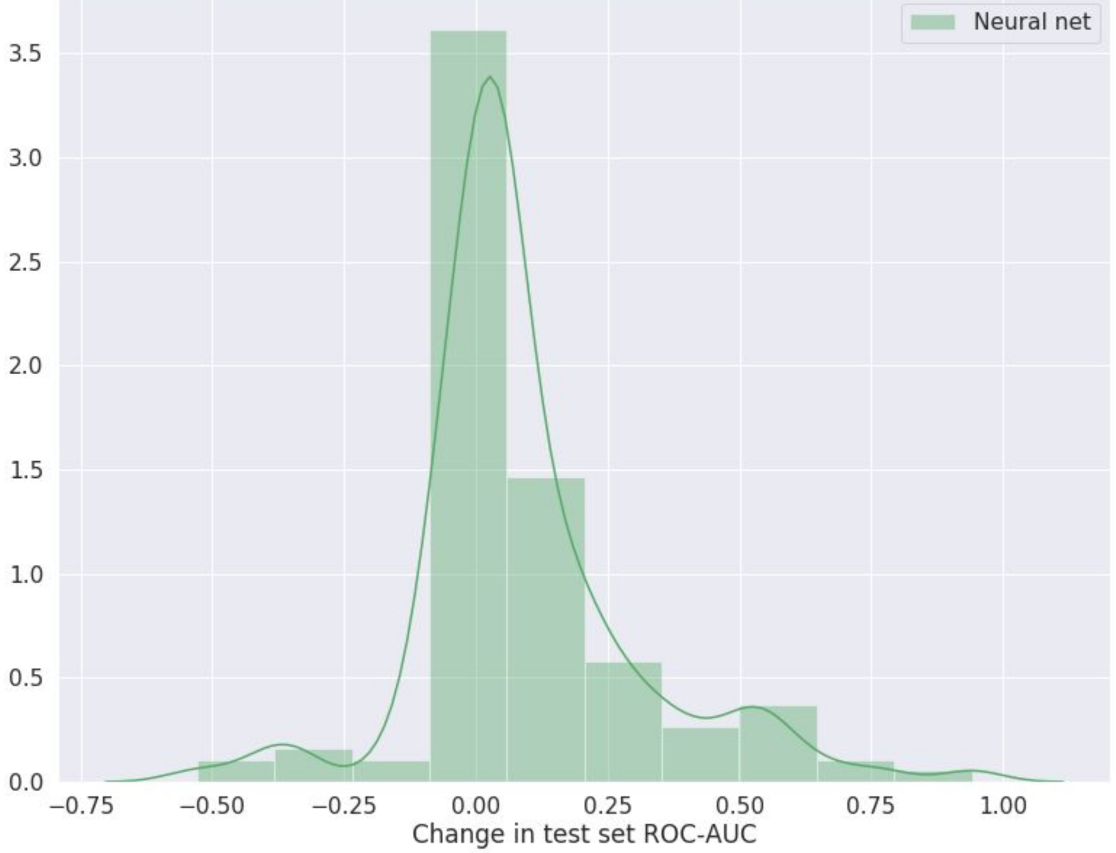
```
{
    "shortlist_key" : "test_safety_list.csv",
    "dataset_bucket" : "gsk_ml",
    "max_epochs": "100",
    "splitter": "scaffold",
    "ecfp_size" : "1024",
    "datastore": "True",
    "result_dir": "/p/lustre1/minnich2/new_test/",
    "system": "LC",
    "transformers": "True",
    "model_type": "NN,RF",
    "featurizer": "ecfp,graphconv",
    "learning_rate": ".0001,.2,10",
    "layer_sizes": "1000,500 100,100,100",
    "dropouts": "0.40,0.40 0.50,0.60,0.60",
    "weight_init_stddevs": "0.02,0.02 0.02,0.03,0.04",
    "bias_init_consts": "1.0,1.0 1.0,2.0,1.0",
    "script_dir": "/usr/gapps/gskcraa/data_science/code/pipeline",
    "python_path": "/usr/mic/bio/anaconda3/bin/python",
    "hyperparam": "True"
}
```

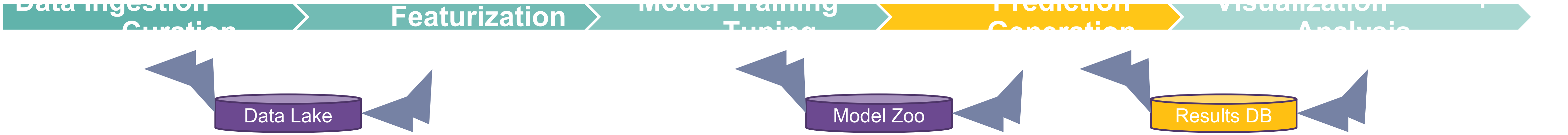
Hyperparameter search improves model accuracy for both regression and classification models

Regression Models



Classification Models





Data Ingestion Curation

Featurization

Model Training Tuning

Prediction Generation

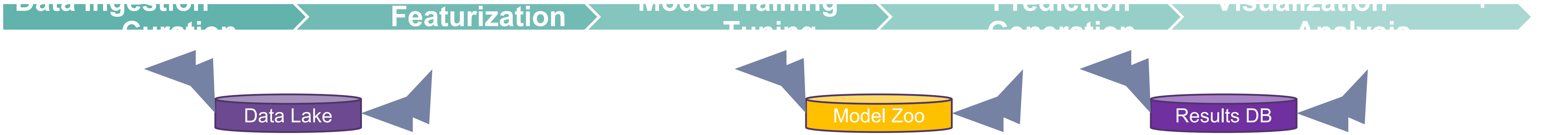
Visualization + Analysis

Data Lake

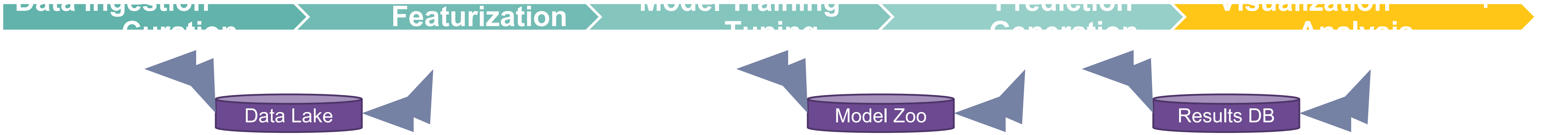
Model Zoo

Results DB

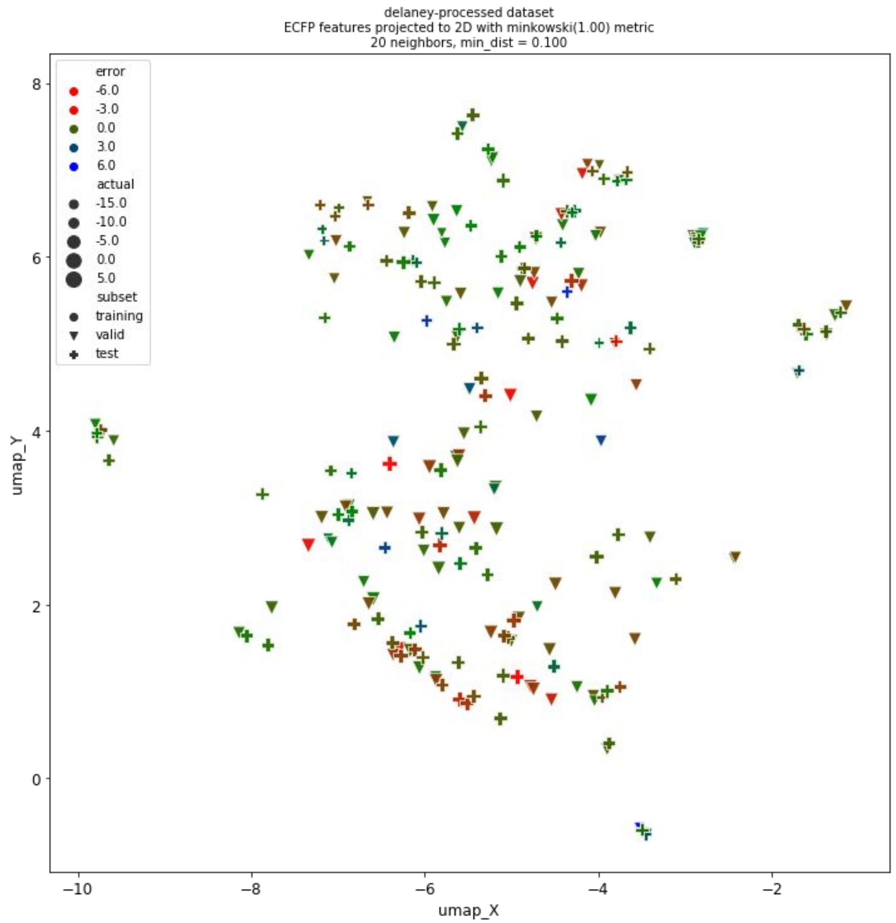
- Our models predict
  - Binding activation/inhibition values for safety-relevant proteins
  - Pharmacokinetic parameters for input into QSP models
  - Also working on hybrid ML/Molecular Dynamics models
- Calculate model-based uncertainty quantification metrics
- If ground truth provided, calculate a variety of prediction accuracy metrics
- All predictions and results saved to Results Database or file system based on user preference

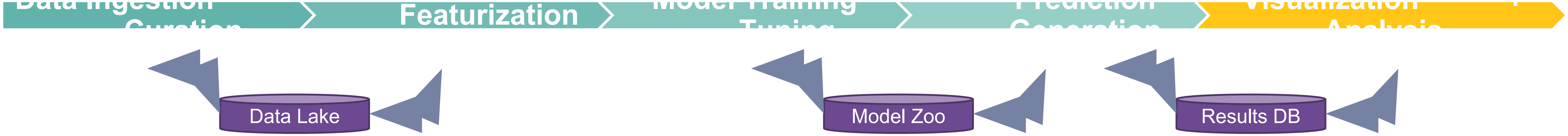


- Model Portability is key for:
  - Releasing to the public
  - Sending to partners for testing with internal data
  - Incorporating into Lead Optimization Pipeline for *de novo* compound generation
- Serialized models are saved to model zoo with detailed metadata
- Support complex queries for model selection
- One command generates queries from dictionary or JSON file, searches model zoo, and loads matching models

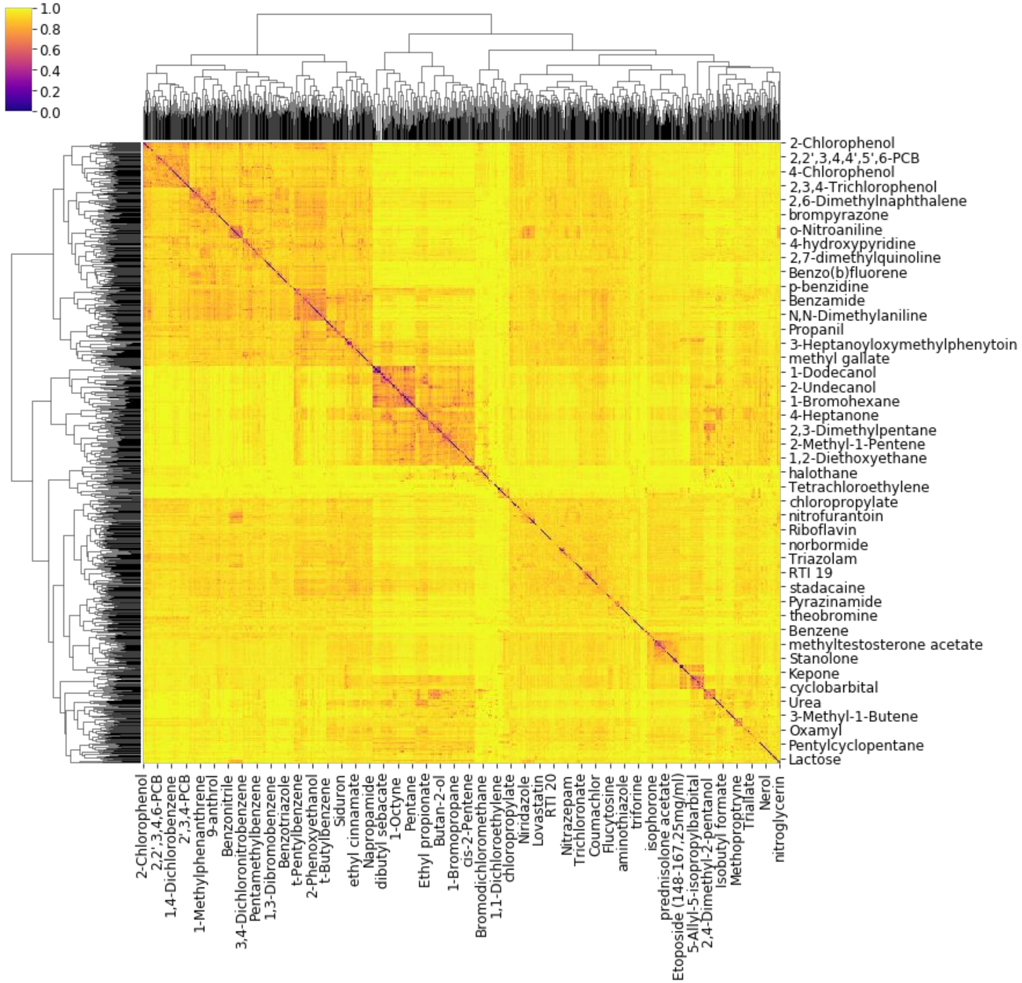


- Visualizations enable validation and evaluation of results
- Support variety of visualizations and also allow for custom functions
- Examples:
  - Predicted vs actual values
  - Learning curve
  - ROC curve/ precision vs. recall curve
  - 2-D projection of numeric features using UMAP





- Chemical diversity analysis is crucial for analyzing domain of applicability, bias in dataset splitting, and novelty of *de novo* compounds
- Support a number of input feature types, distance metrics, and a variety of clustering, analysis, and plotting methods



ATOM

Roadmap

- Infrastructure and Architecture – what GPUs are we using?
- Data-Driven Modeling Pipeline – what have we built?
- Experiments – what have we been able to do?
- Future work – where are we going from here?

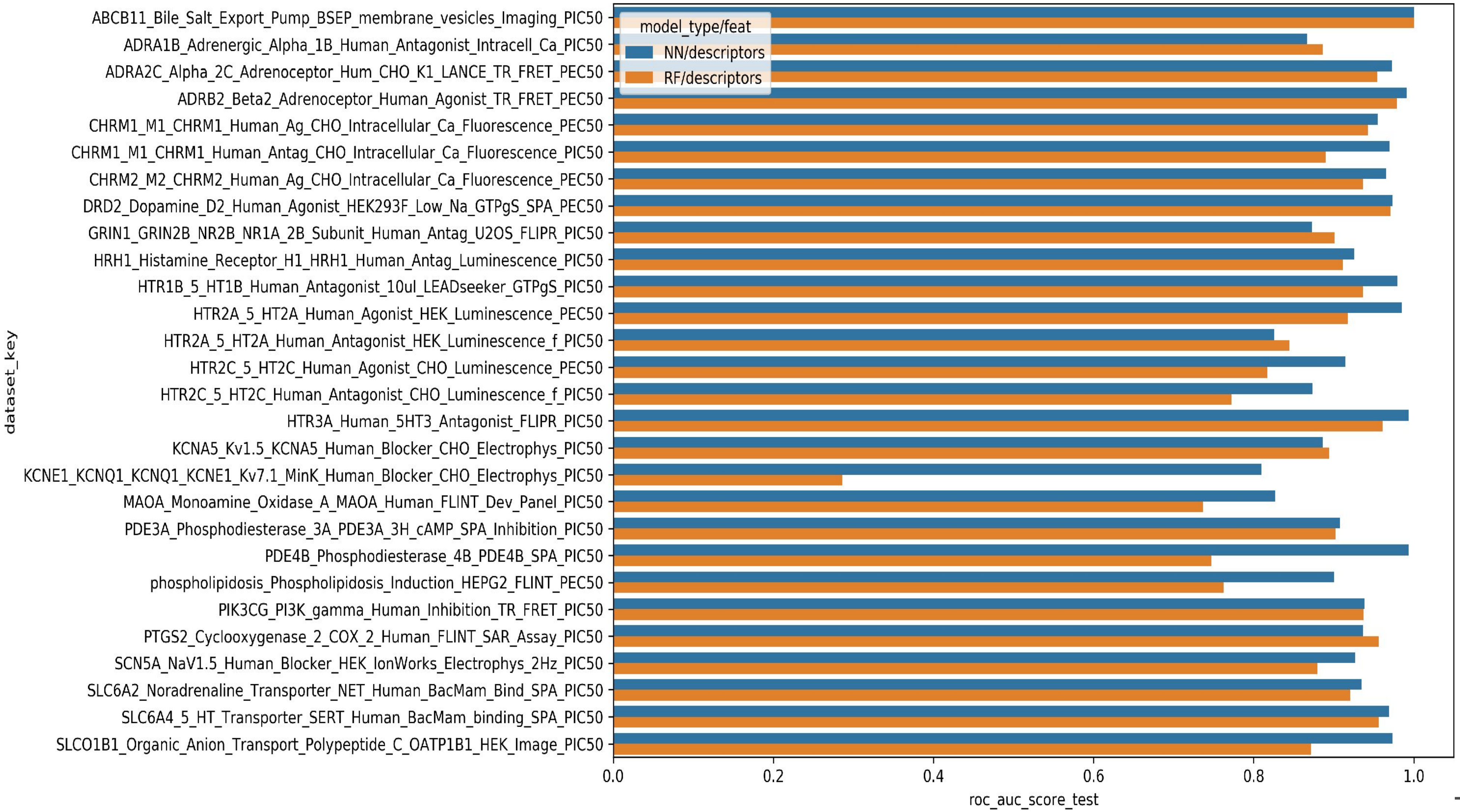
Experimental Design

- Neural Nets and Random Forest Models
- Extended Connectivity FingerPrints (ECFP), Molecular Operating Environment (MOE) descriptor vectors, and GraphConvolution-based features
- NN: Vary learning rates, number of layers, layer sizes, dropout rates
- RF: Vary max depth and number of estimators
- Train iteratively up to 500 epochs and pick best model based on validation set performance

Experimental Summary

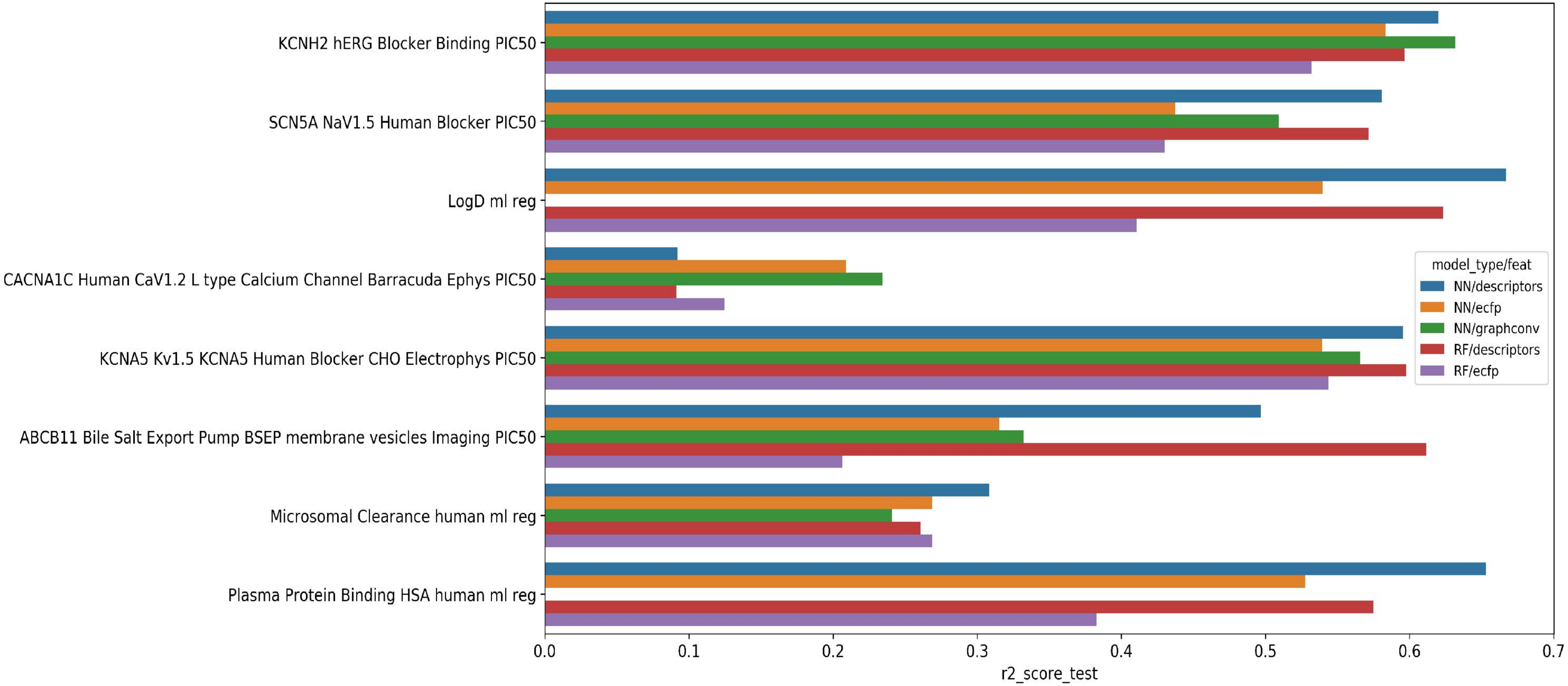
- 5,964 total models for 41 Safety and Pharmacokinetic datasets
- 4,696 Neural Net models
- 1,253 Random Forest models
- 3,819 Regression models
- 2,130 Classification models
- Models were trained on a wide range of proprietary GSK assay datasets, including ones that are larger than public datasets reported in the literature

Classification performance shows high accuracy for selected safety targets



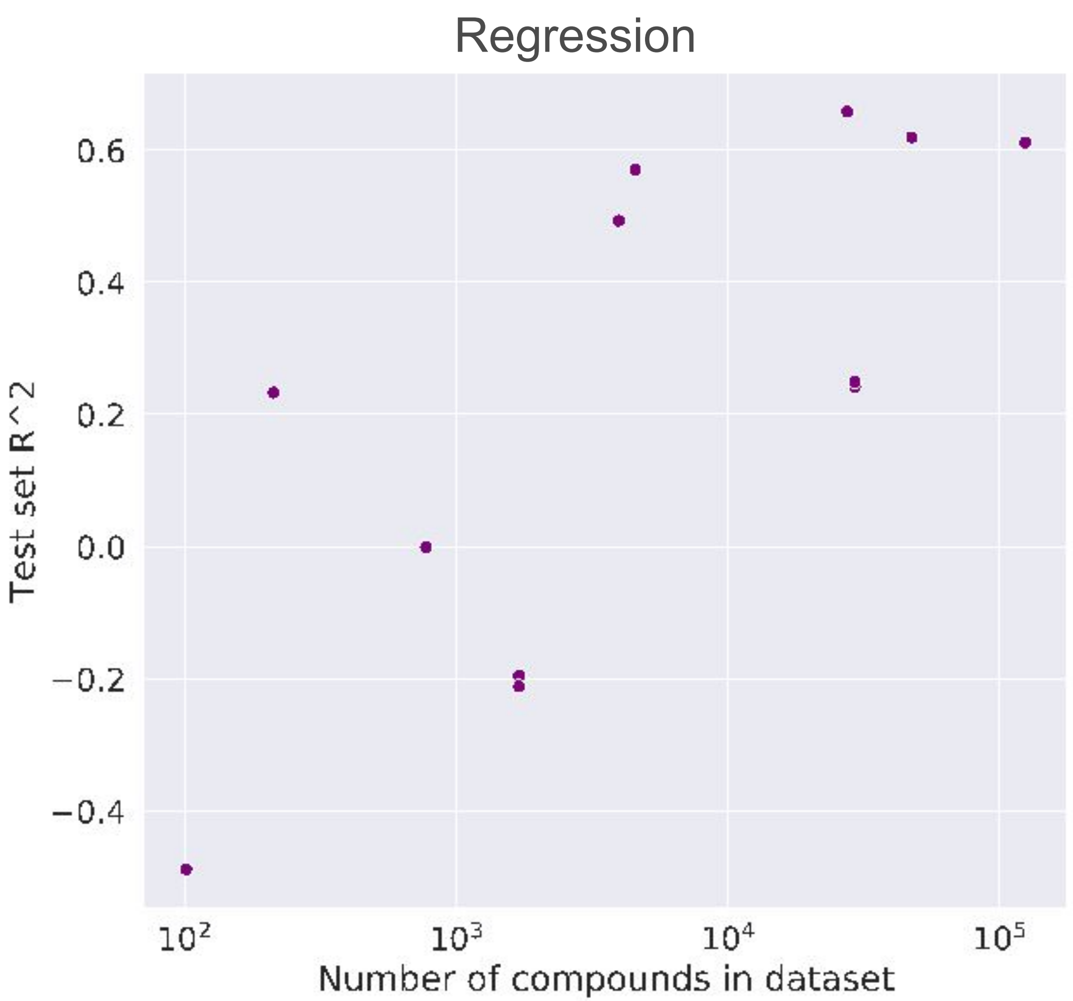
- Assays range in size from 187 to 9173 compounds
- 23 of 28 of the assays show improvement with NN
- KCNE1 shows largest improvement
- Classification accuracy appears to be relatively high (>0.8 ROC-AUC)

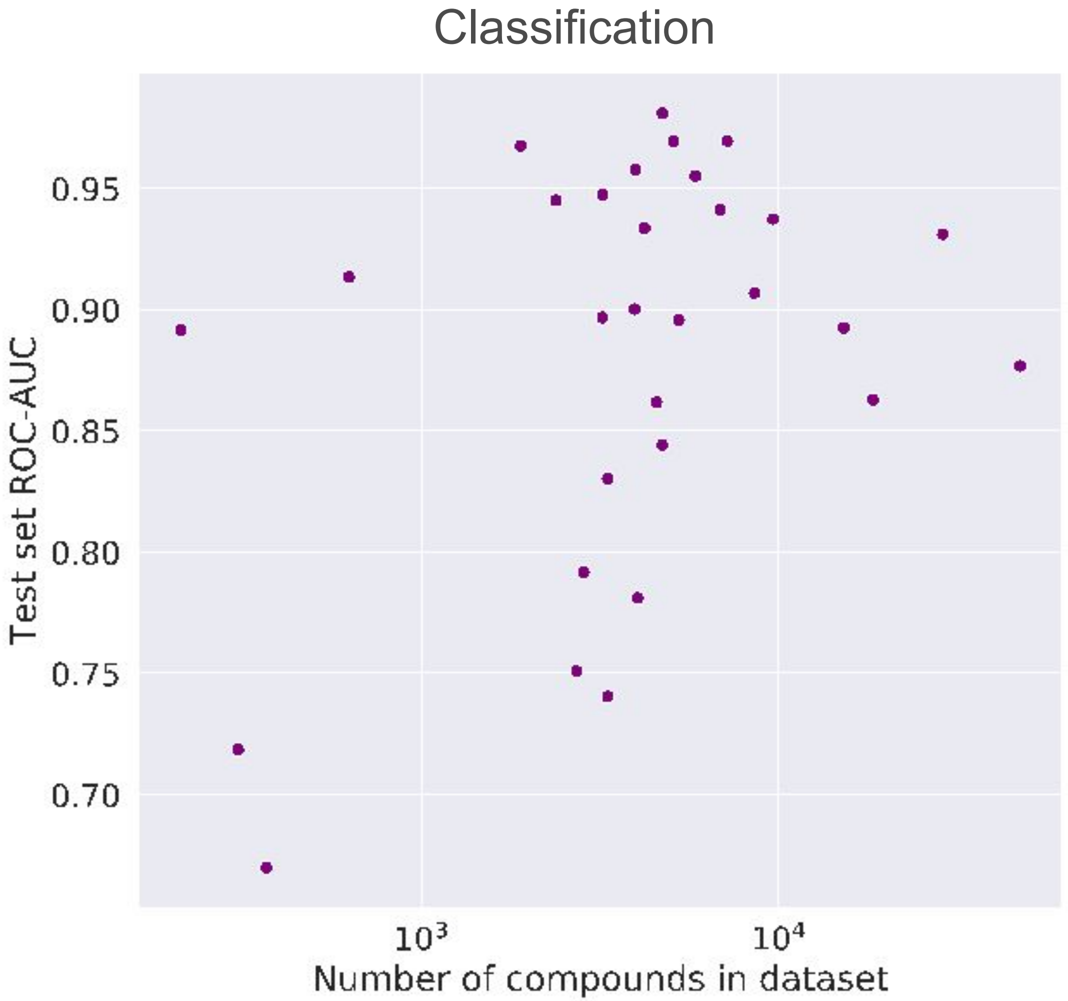
Regression models present a greater challenge



- Assays range in size from 101 to 123,759 compounds
- 4 of 8 of the assays show improvement with NN
- Descriptors and Graphconv outperform ECFP
- Test set R^2 ranges from ~0.1 to ~0.7

Test set accuracy varies with number of compounds in dataset





ATOM

Summary of Observations

- Classification results look good, but need to better handle class imbalance
- Regression models can be improved
- Adding data seems to help, so we are looking into:
    - Sourcing public datasets
    - Generated targeted experimental data
    - Transfer learning
    - Multi-task learning

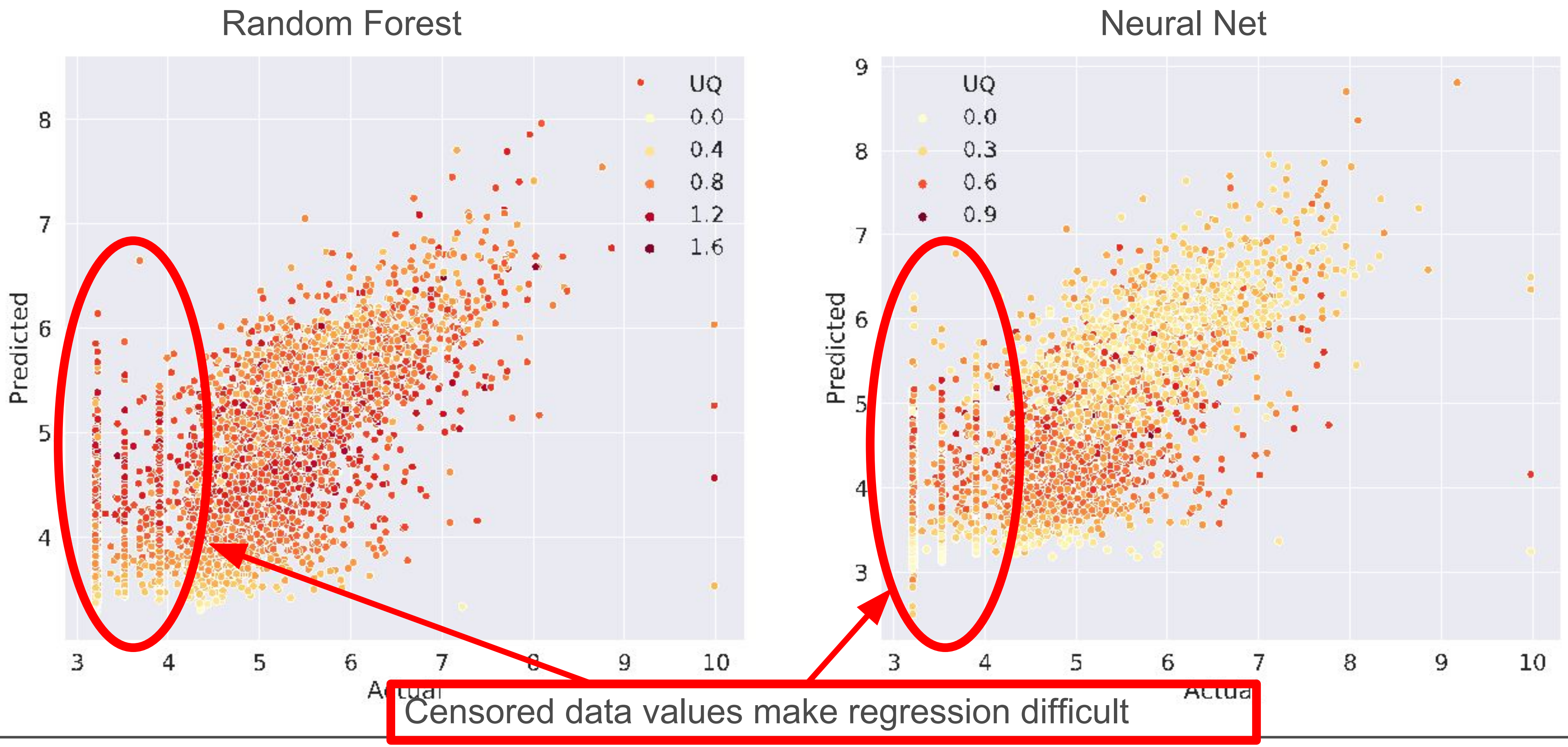
Uncertainty Quantification (UQ) Analysis

- UQ helps reveal what a model is not confident about
- Goals for data-driven model UQ:
  1. Accurately characterize confidence in model predictions as a function of UQ
  2. Use UQ to guide active learning
  3. Use UQ to weight model ensembles

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
  - 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 $\sigma_{total} = \sqrt{{\sigma_{aleatoric}}^2 + {\sigma_{epistemic}}^2}$

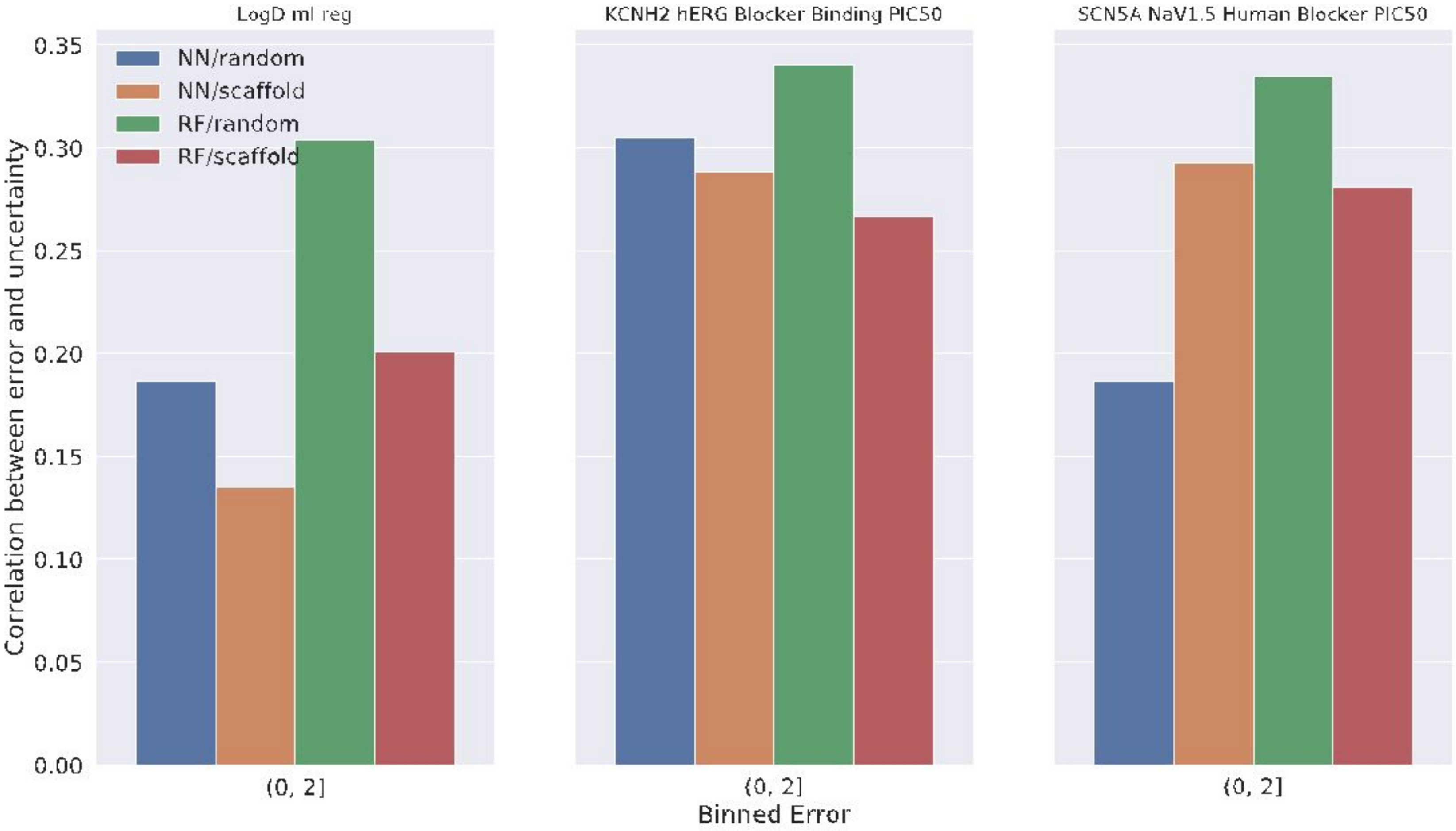
Goal is to quantify prediction uncertainty for assays such as hERG



Censored data values make regression difficult

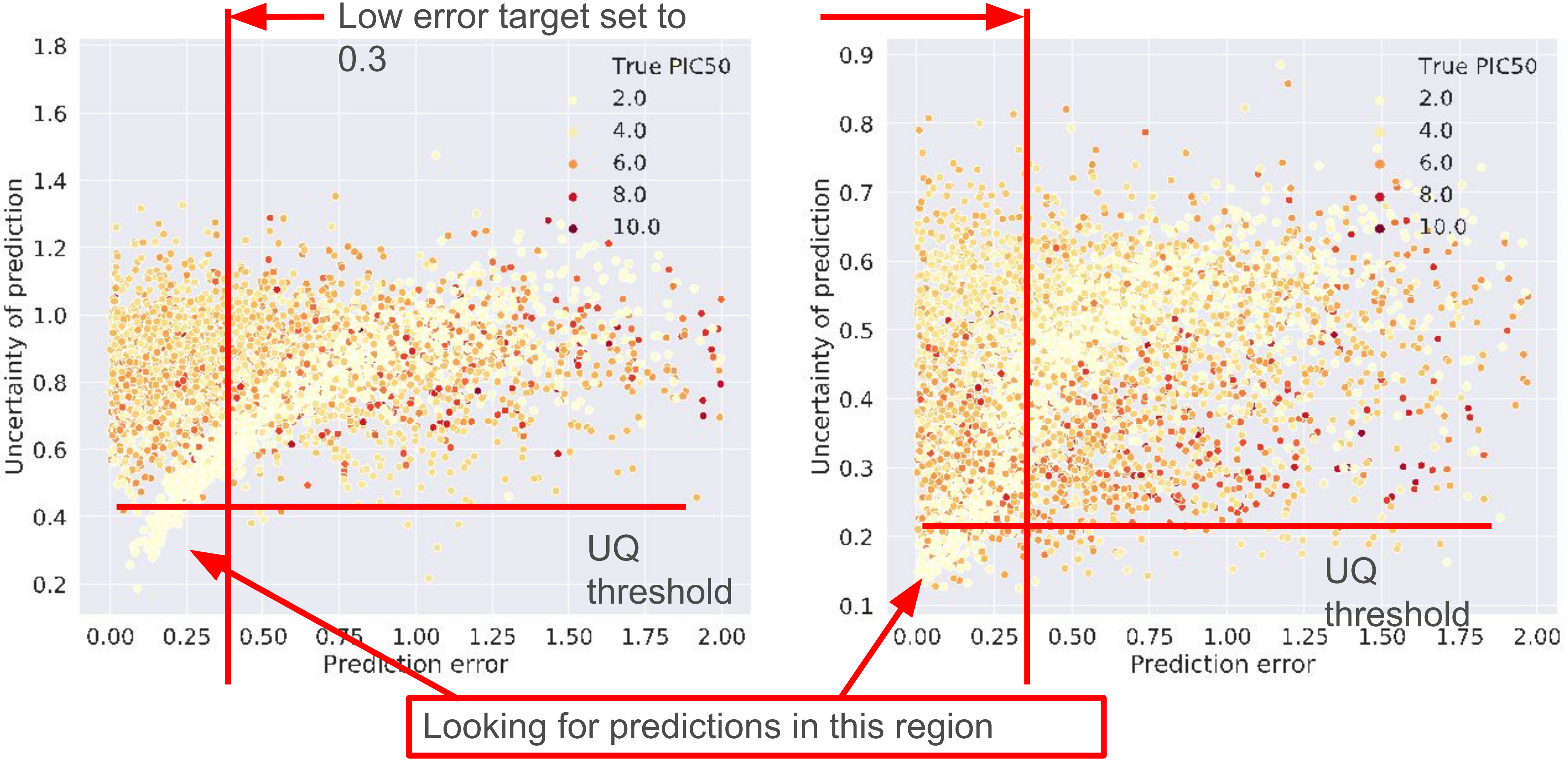
Correlation between error and UQ is fairly low

- Binned prediction error
- Kept bins with > 150 samples
- Calculated Pearson's Correlation between error and UQ
- Correlations range between ~0.14-0.35
- All p-values are <<< 0.01



ATOM

UQ threshold identifies a fraction of the "low error" predictions, which approximates experimental error



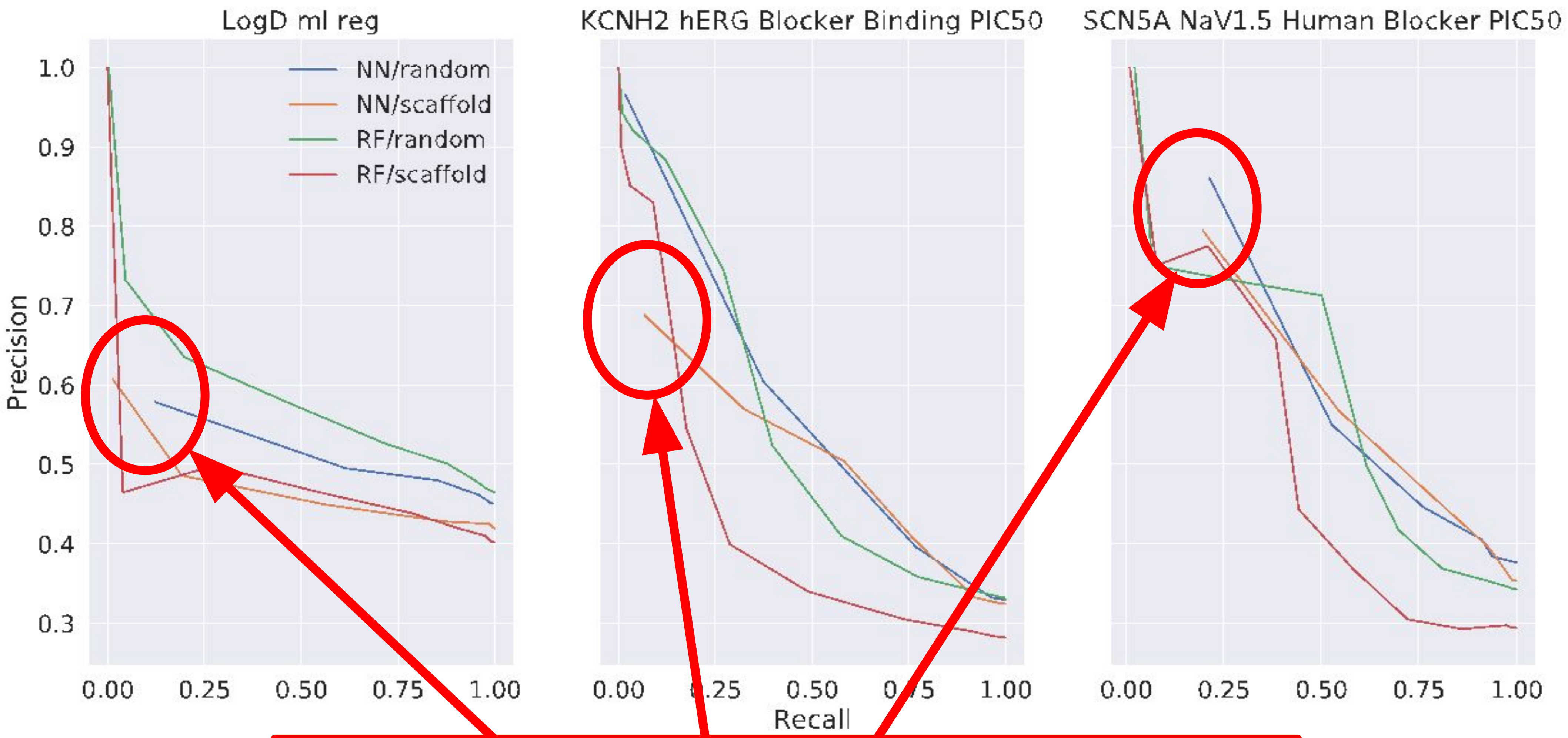
Low error target set to 0.3

UQ threshold

UQ threshold

Looking for predictions in this region

Precision-Recall curves with varying UQ threshold show greater challenges with scaffold splits and neural networks



NN does not reach a precision of 1 for any UQ threshold

ATOM

RF=Random Forest, NN=neural

Training time Analysis

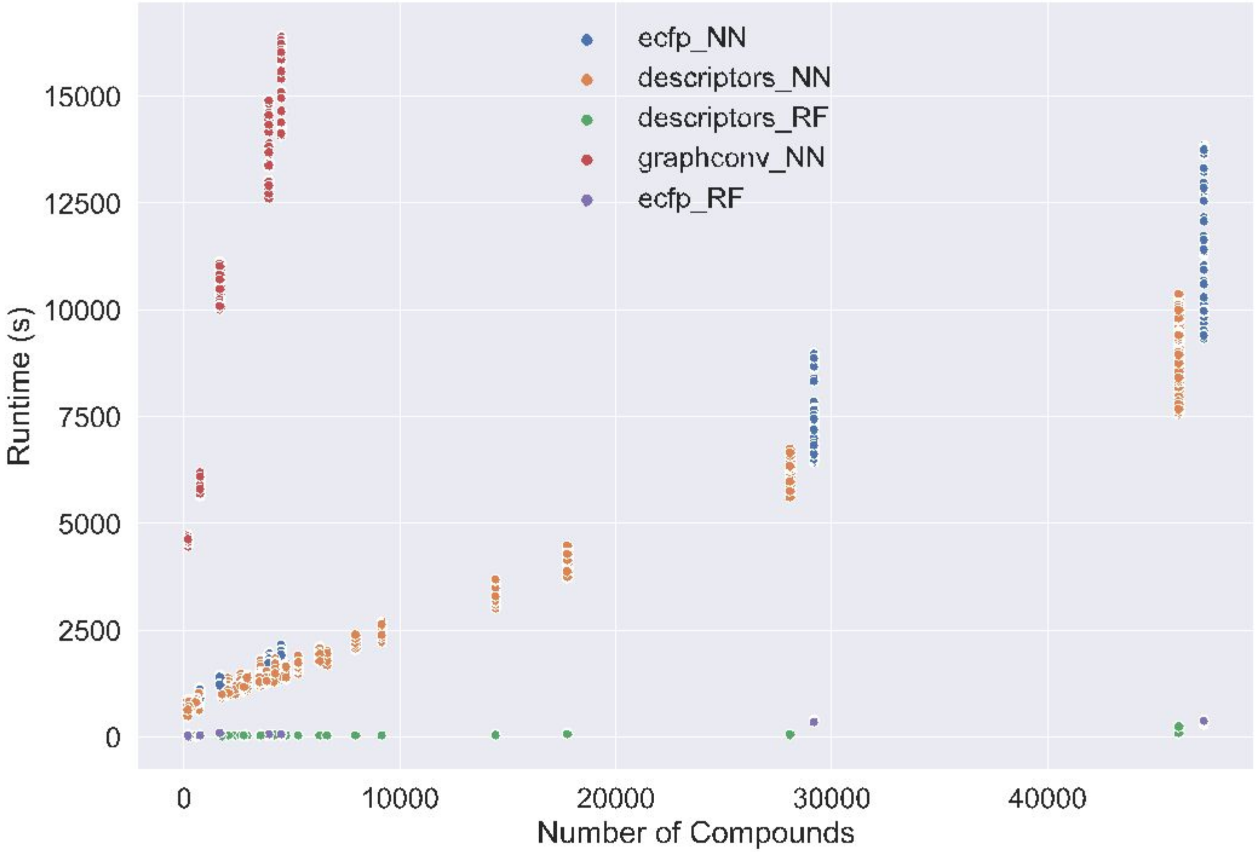
- In addition to understanding performance of models, need to understand efficiency
- Examined training runtimes for our models and a variety of variables
- All times were calculated for model building on supercomputers
- Can help to guide future experiments as we scale up

Training time is highly dependent on number of compounds

- Plotted runtime versus number of compounds

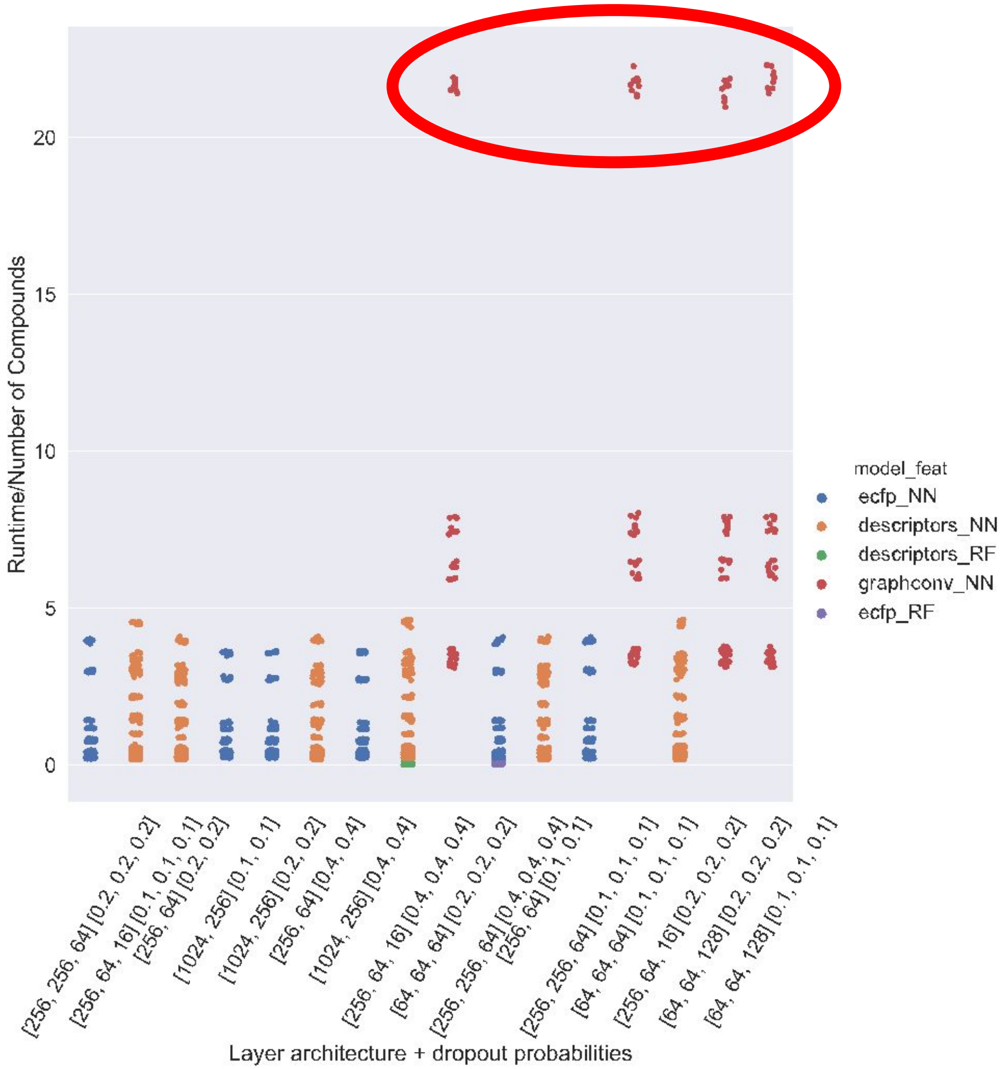
- Relationship looks linear for NN, with slope depending on feature type

- GraphConv NN models are very slow, while Random Forest is very fast



Layer architecture does not appear to have an effect on training time

- Plotted runtime normalized by dataset size versus Layer Architecture + Dropout Probability Combination
- Surprisingly, number of parameters in network does not affect training time
- Currently investigating why some Graph Convolution models are much slower



Roadmap

- Infrastructure and Architecture – what GPUs are we using?
- Data-Driven Modeling Pipeline – what have we built?
- Experiments – what have we been able to do?
- Future work – where are we going from here?

Current status

- Pipeline
  - Dev 1.0 release
  - Installable using pip as a whl file
  - Runs internally at GlaxoSmithKline for evaluation
- Models
  - Our models have been incorporated into our *de novo* compound generation active learning loop
  - We are able to export and share models with consortium members as well

Future Plans

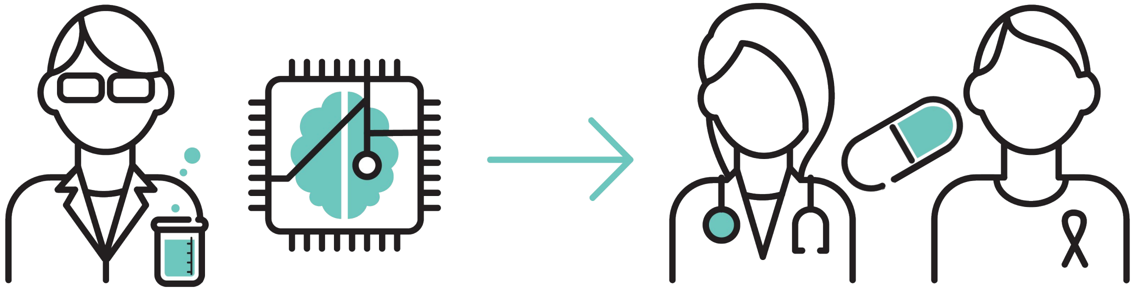
- Improving Portability
    - Release pipeline open source
    - Dockerize the entire pipeline
    - Release data services infrastructure as Kubernetes pods
- Improving performance
    - Add in optimized hyperparameter search function
    - Explore hyperparameters for uncertainty quantification
    - Transfer learning
    - Multi-task learning
    - Ensemble modeling

Join ATOM

Visit atomscience.org/membership

Contact info@atomscience.org

@ATOM_consortium #ATOMscience



Transform drug discovery, accelerate R&D, and integrate data, AI, and supercomputing to benefit patients

Consortium Members:

Frederick National Laboratory for Cancer Research gsk Lawrence Livermore National Laboratory UCSF

Feel free to email me with technical questions: minnich2@llnl.gov