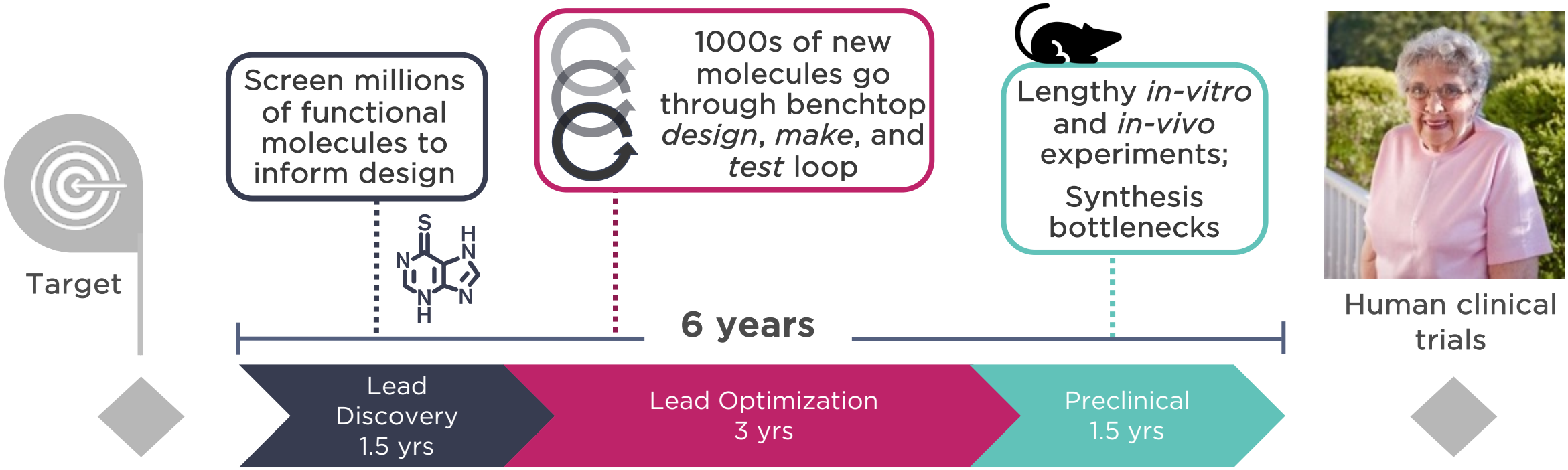
Accelerating Therapeutics for Opportunities in Medicine

Dr. Amanda J. Minnich

Current drug discovery

Is there a better way to get medicines to patients?



- 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

Accelerating Therapeutics for Opportunities in Medicine

Building a precompetitive platform to get better medicines to patients faster

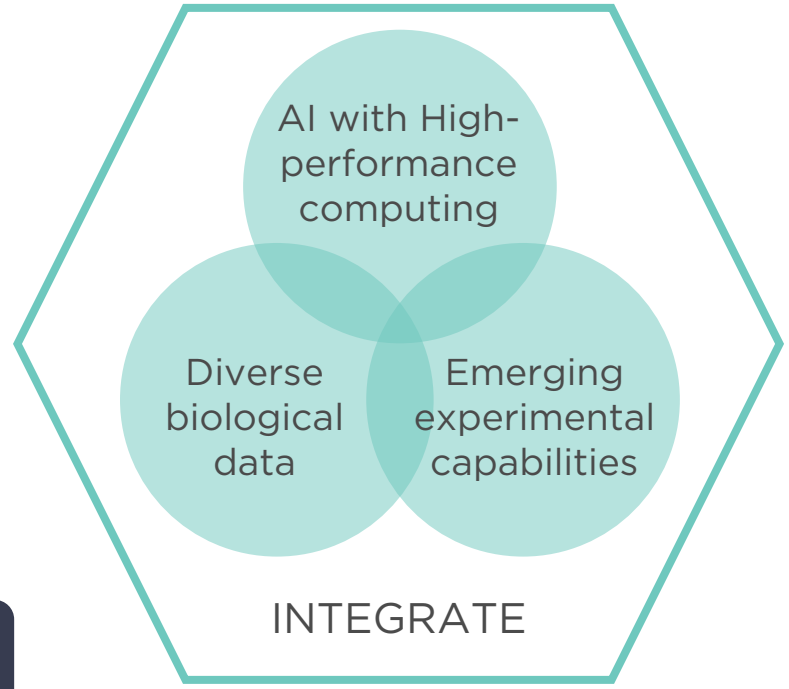
Mission

ACCELERATE the discovery & development of more effective therapies for patients

Vision

TRANSFORM drug discovery from experiment-driven, sequential, and high-failure to computation-driven, integrated, and patient-centric

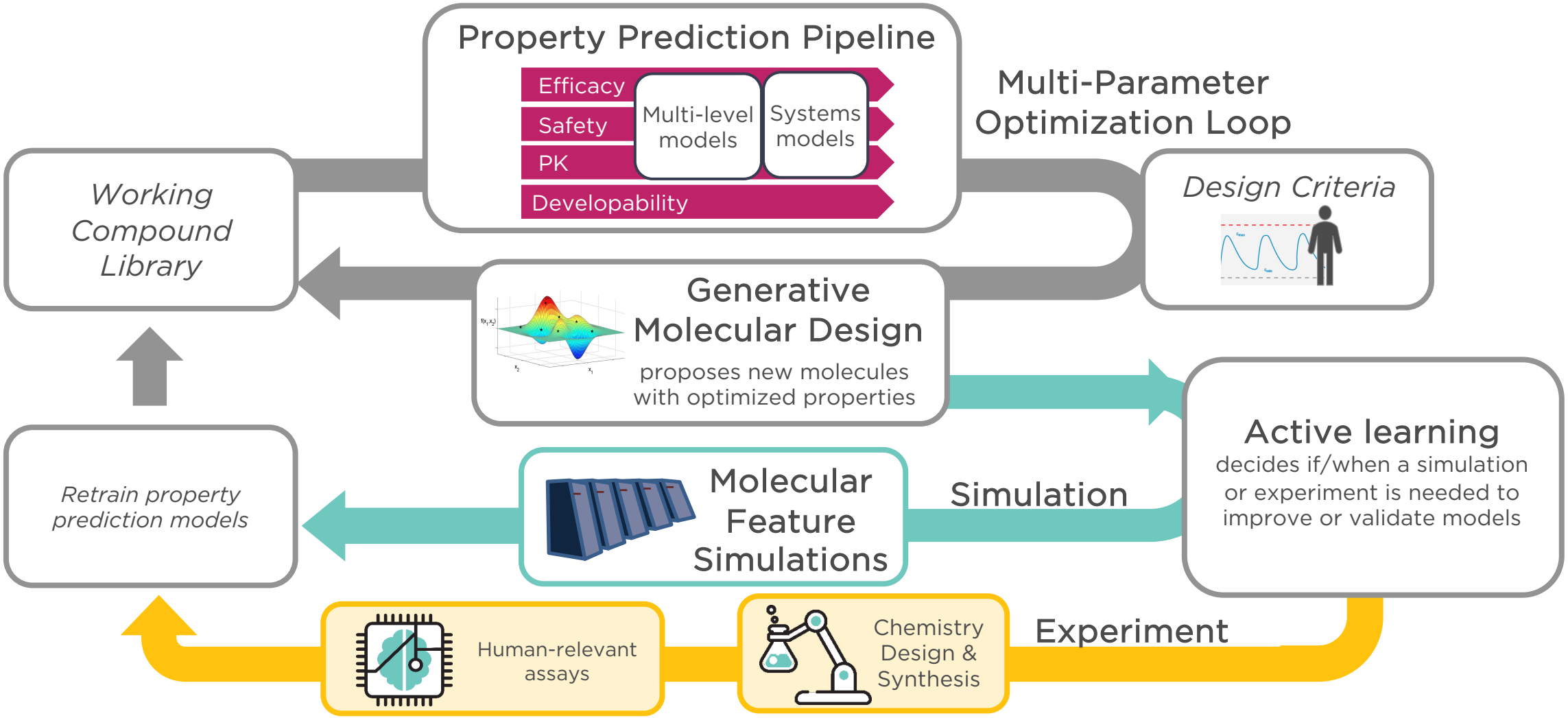
How?



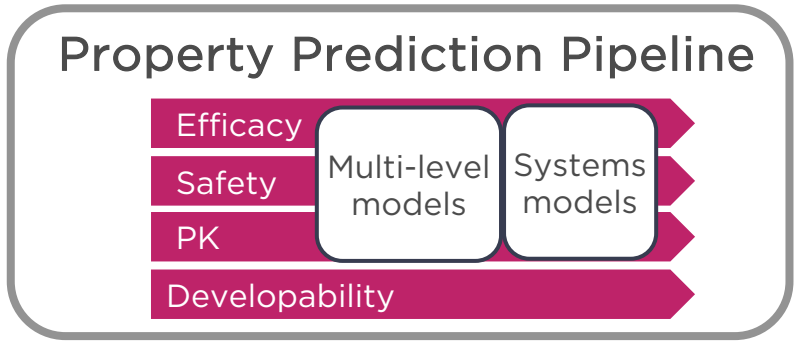
Scope

Platform 1.0: Target to candidate Platform 2.0: Target to patient

Our platform is an active learning drug discovery framework

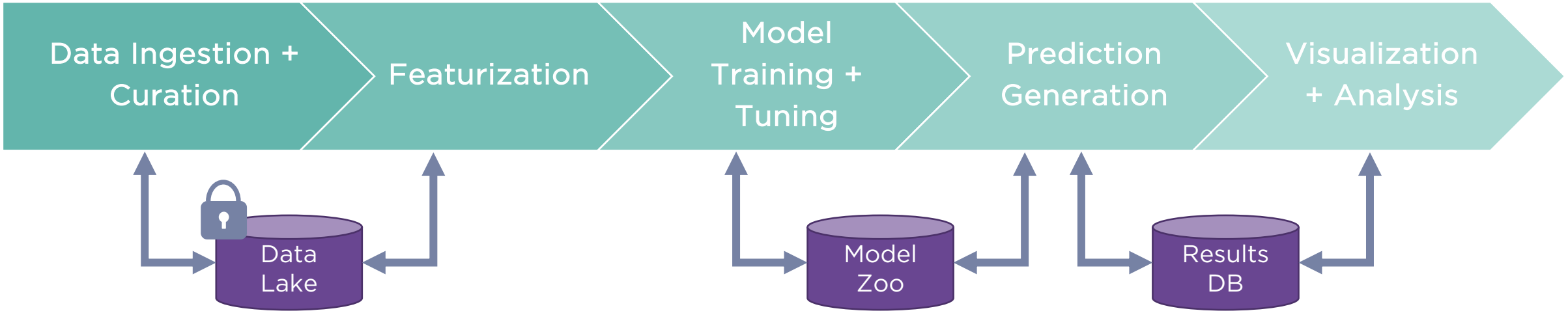


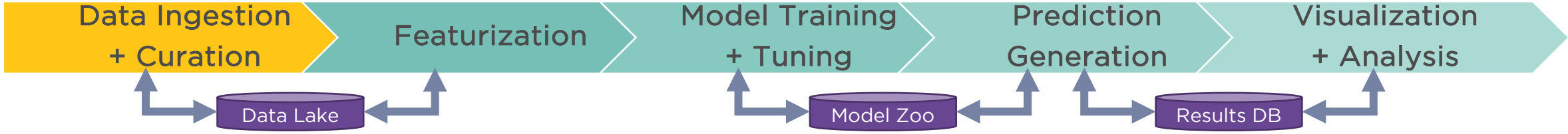
Our platform is an active learning drug discovery framework



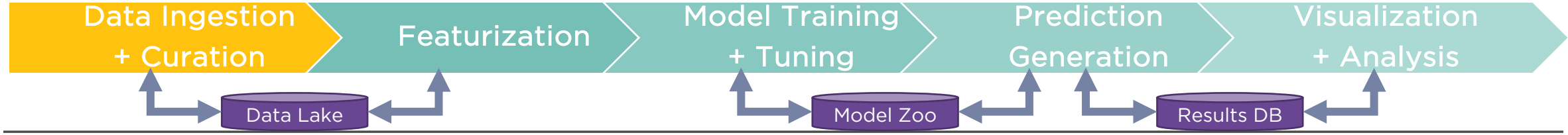
Property prediction pipeline

Will be released open source by November 2019

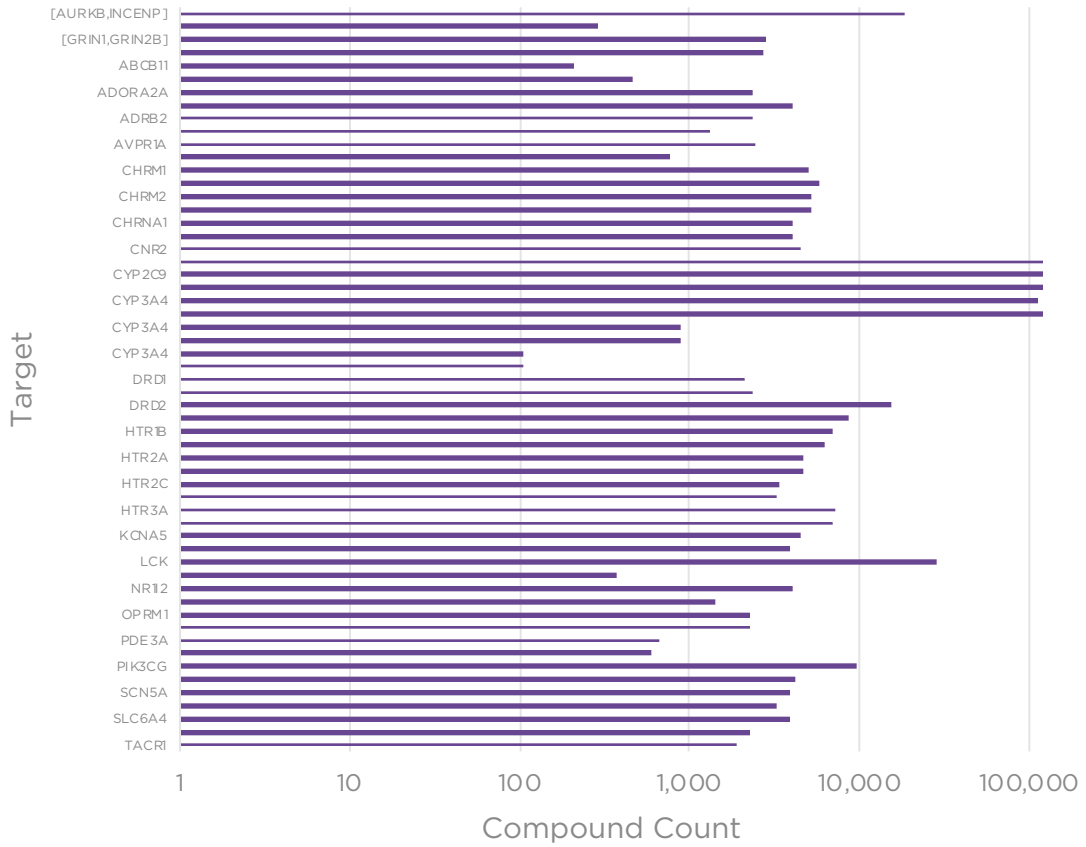




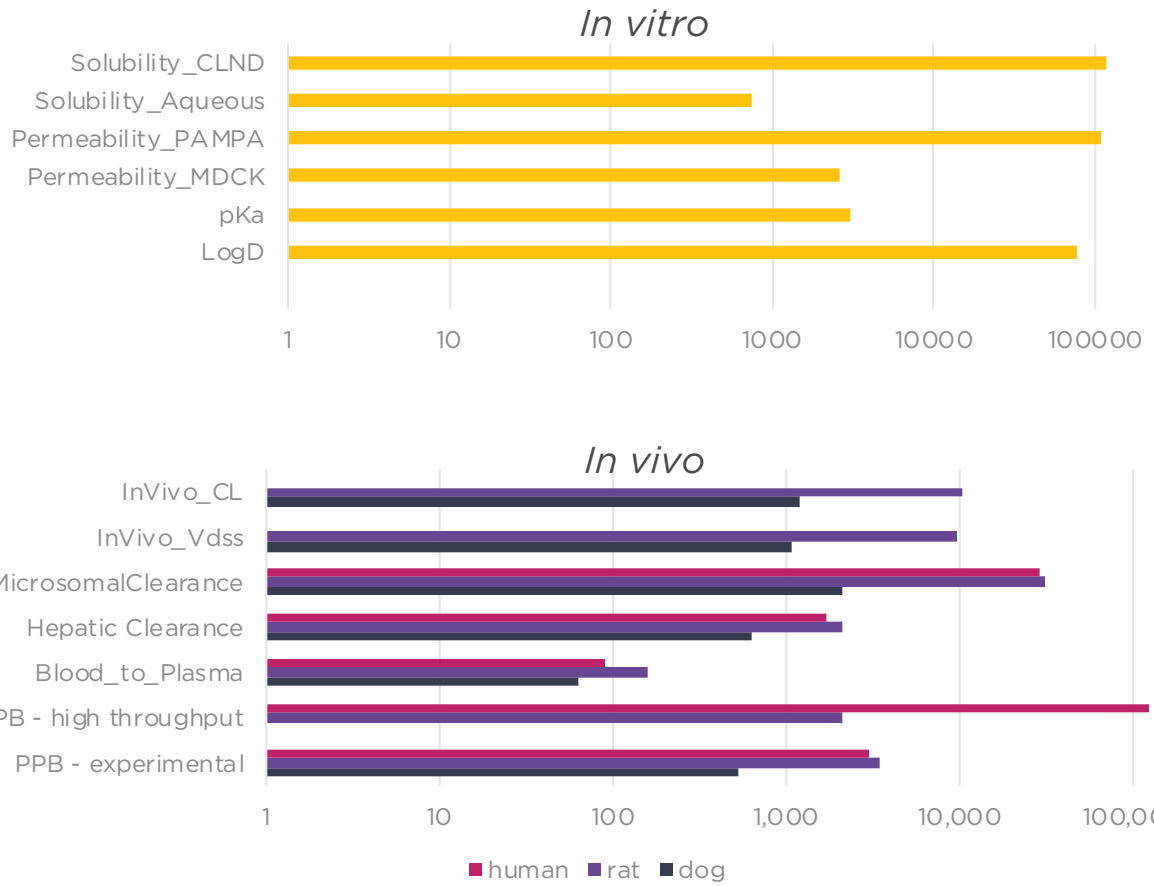
- Raw pharma data consists of 300 GB of a variety of bioassay and animal toxicology data on ~2 million compounds from GSK
- Domain experts created Jupyter notebooks to process data
- Serve as both code and record of modifications made to datasets



Example safety datasets (56 targets)

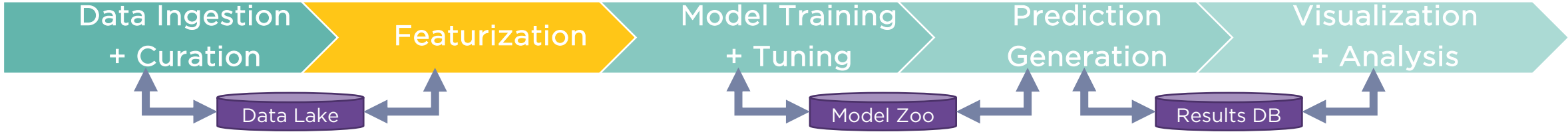


Example Pharmacokinetic datasets *In vitro* and *In vivo*

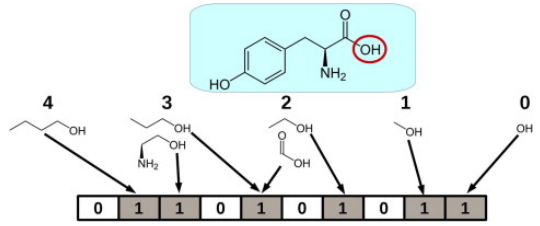


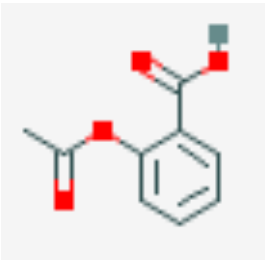
ATOM

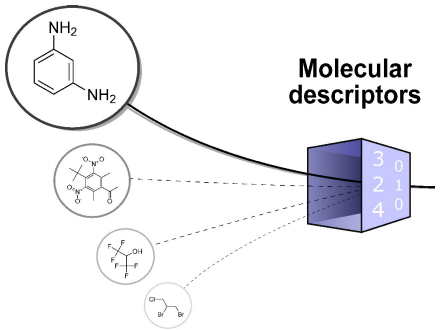
We have curated ~150 model-ready datasets

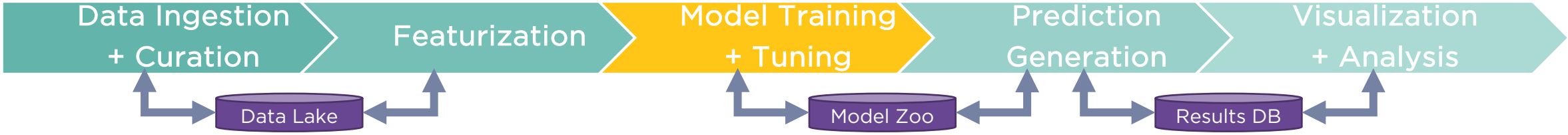


- Support loading datasets from either Data Lake or filesystem
- Support a variety of feature types
  - Extended Connectivity Fingerprint
  - Graph-based features
  - Molecular descriptor-based features (MOE, Mordred)
  - Autoencoder-based features (MolVAE)
  - Allow for custom featurizer classes
- Split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

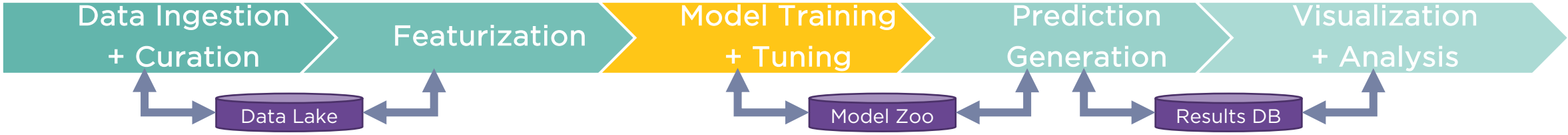








- Have built a train/tune/predict framework to create high-quality models
- Currently support:
  - scikit-learn models
  - Deepchem models (wrapper for TensorFlow)
  - XGBoost models
  - Allow for custom model classes
- Allow for iterative training of neural nets
- Tune models using the validation set and perform k-fold cross validation

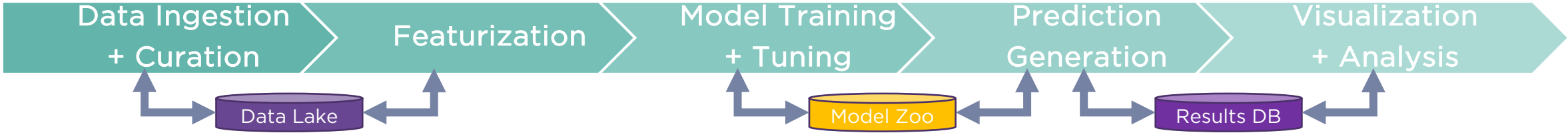


We have a module for distributed brute-force hyperparameter search

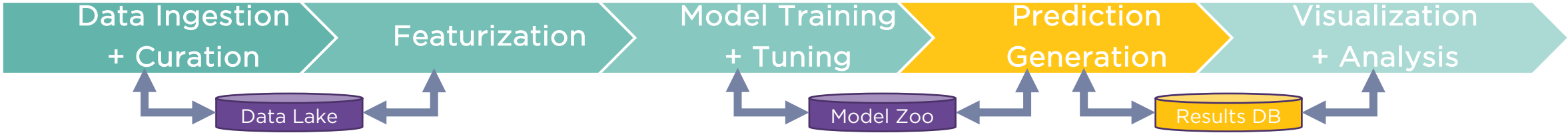
- Support linear grid, logistic grid, random, and user-specified steps
- Specify input with JSON file or command line
- Generates all possible combinations of hyperparams, accounting for model type
- Groups neural net architecture combinations
- Constrains number of parameters in NN based on dataset size
- Checks if model already exists in model zoo

ecfp_hyper_config_pk_round2.json

```
{
    "hyperparam": "True",
    "search_type": "geometric",
    "shortlist_key": "dskey_PK_MLready_LOG_transformed_dataset.csv",
    "result_dir": "/p/lustre1/minnich2/pk052819/",
    "collection_name": "pk052819",
    "uncertainty": "True",
    "transformers": "True",
    "splitter": "scaffold,random",
    "model_type": "NN,RF",
    "featurizer": "ecfp,graphconv,descriptors",
    "descriptor_type": "moe",
    "split_valid_frac": "0.1",
    "split_test_frac": "0.2",
    "learning_rate": ".0001,.01,5",
    "layer_nums": "1,2",
    "node_nums": "1024,256,128,64,32,16,8,4,1",
    "max_final_layer_size": "16",
    "dropout_list" : "0.1,0.2,0.4",
    "rf_max_depth": "20,100,5"
}
```



- Model Portability is key for:
  - Releasing to the public
  - Sending to partners for testing with internal data
  - Incorporating into Multi-Parameter Optimization Loop for generative molecular design
- Serialized models are saved to model zoo or disk 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



- 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-building summary

11,552 total models

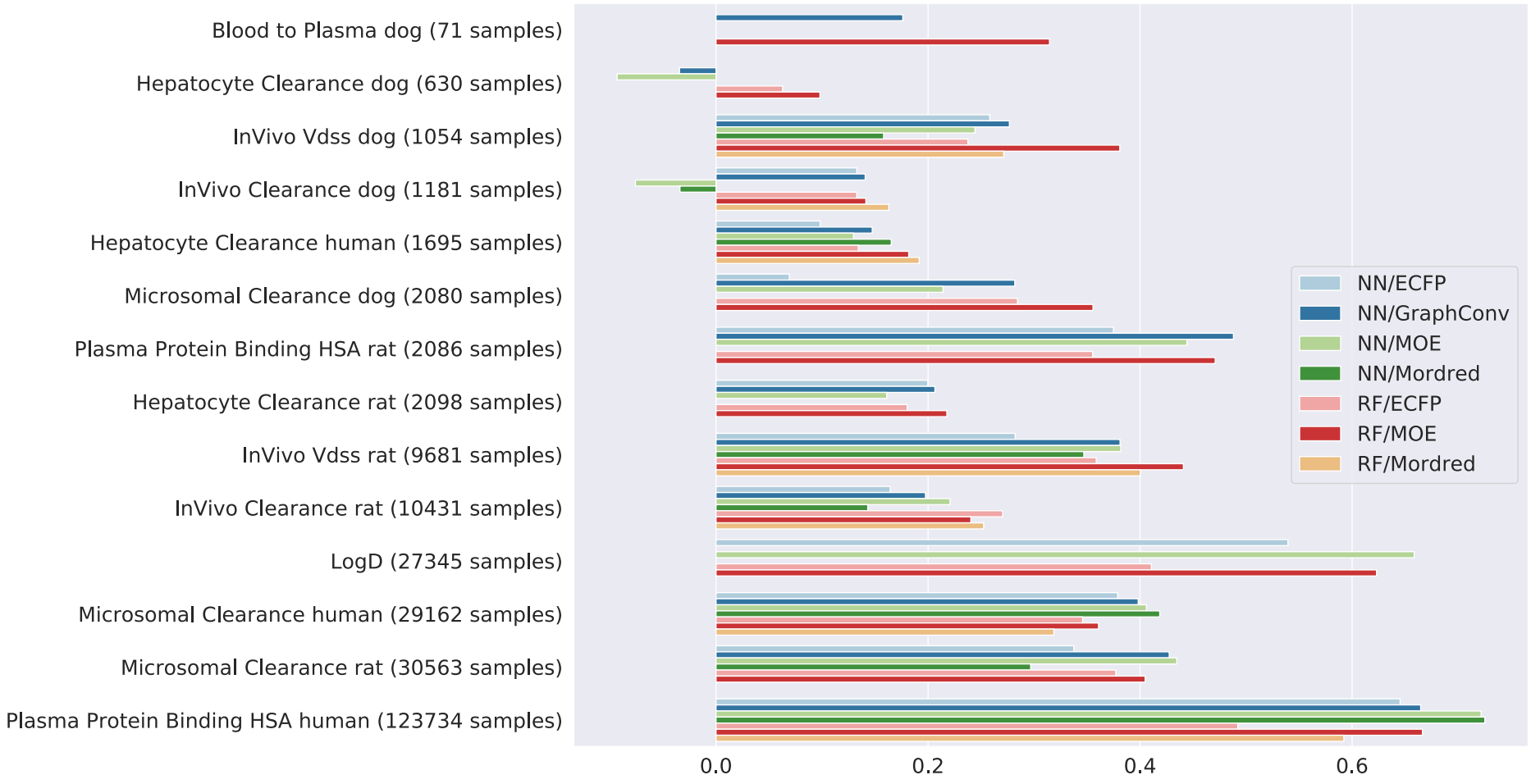
9,422 Regression models

2,130 Classification models

15 Pharmacokinetic datasets

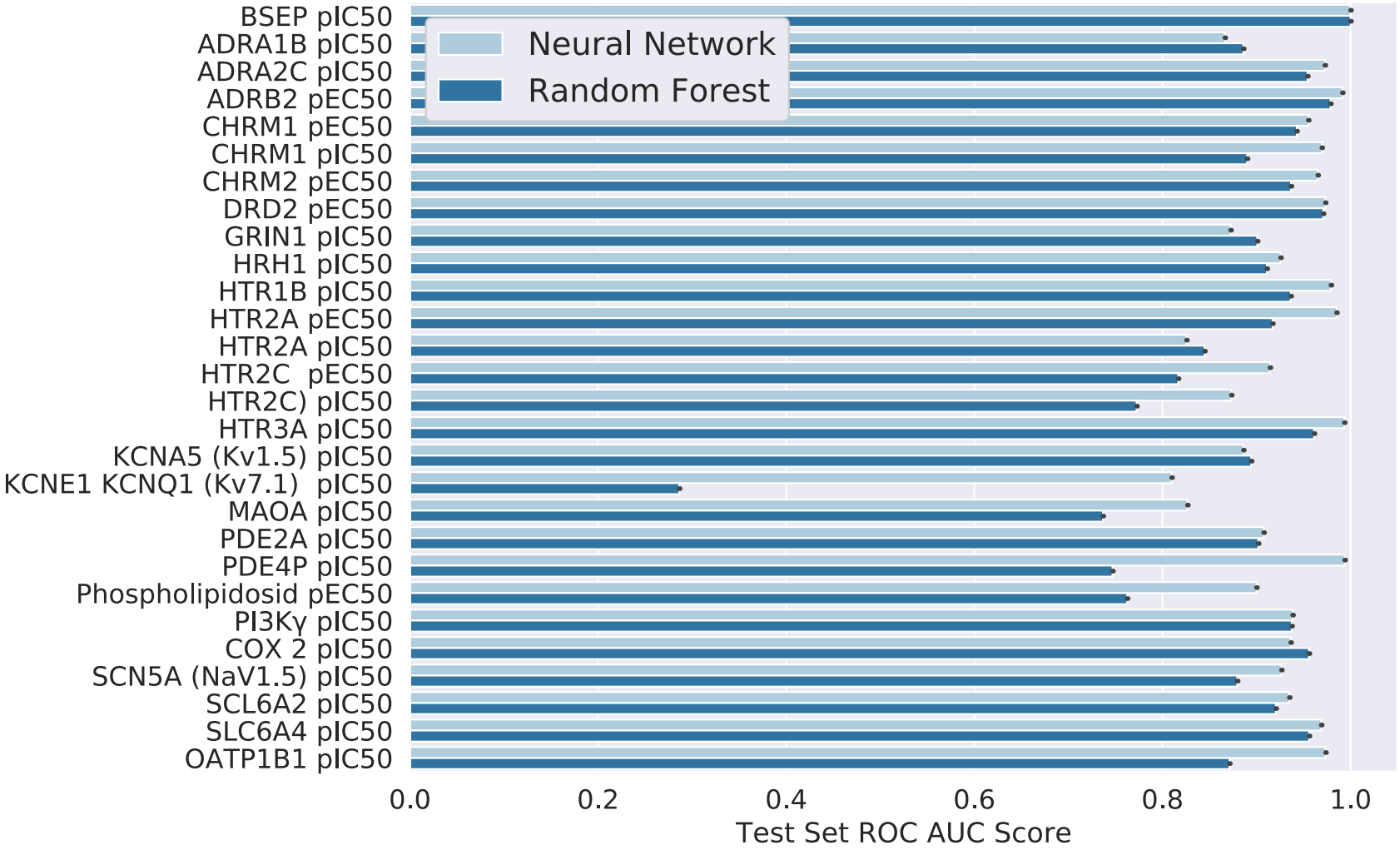
26 Safety Datasets

PK datasets vary in size and model accuracy



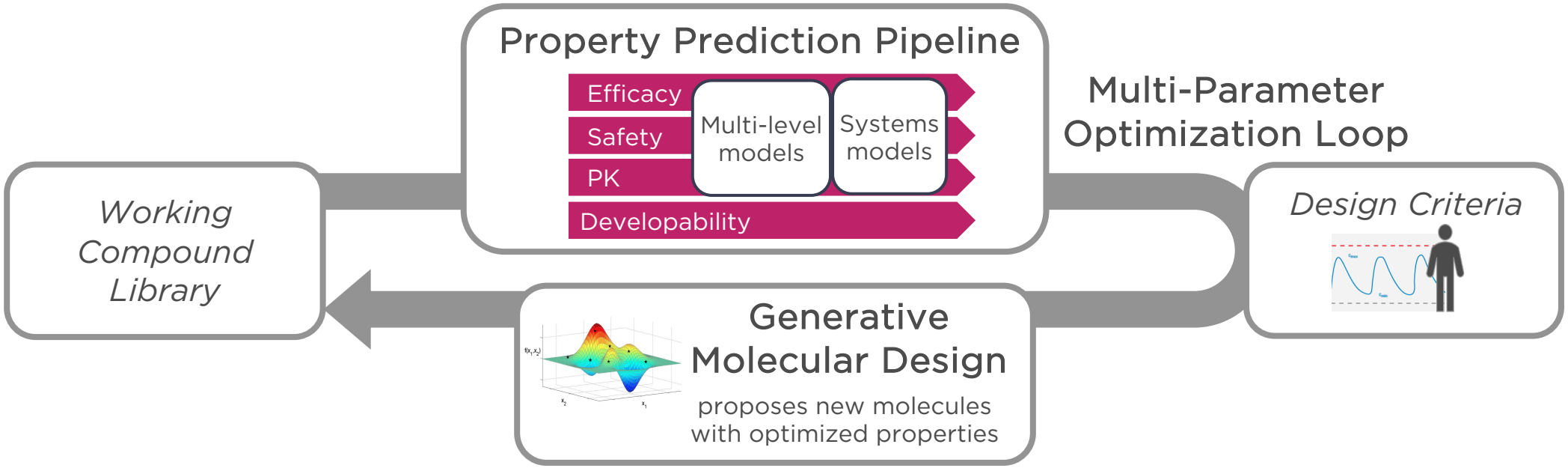
- Assays range in size from 71 to 123,759 compounds
- 5 of the assays show improvement with NN
- Descriptors and Graphconv outperform ECFP
- Test set R^2 ranges from <0 to 0.7

Classification performance shows high accuracy



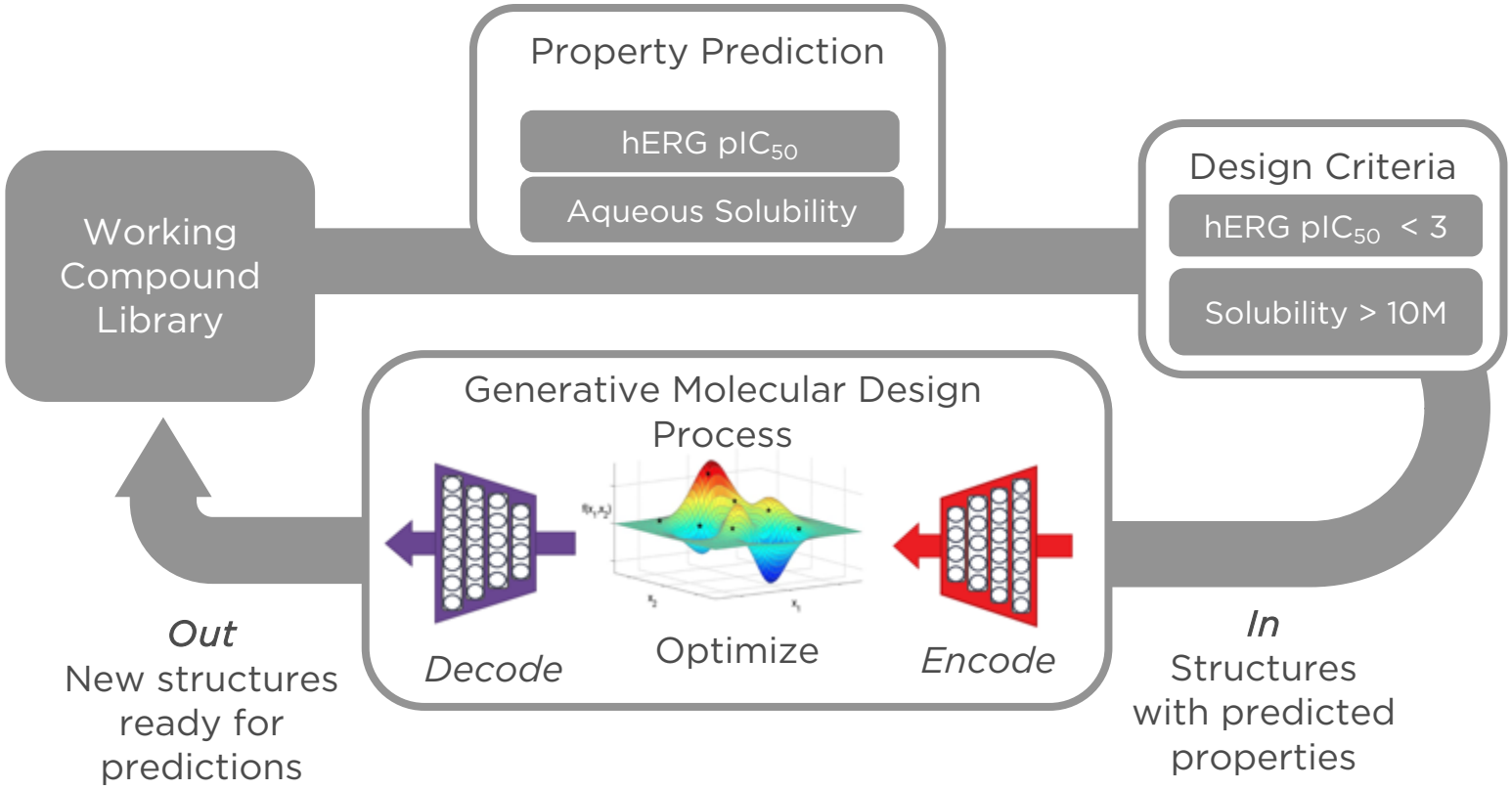
- Assays range in size from 187 to 9173 compounds
- 23 of 28 of the assays show improvement with NN
- KCNE1 shows largest improvement
- Classification accuracy appears to be relatively high (>0.8 ROC-AUC)

The second piece is the multi-parameter optimization loop for generative molecular design

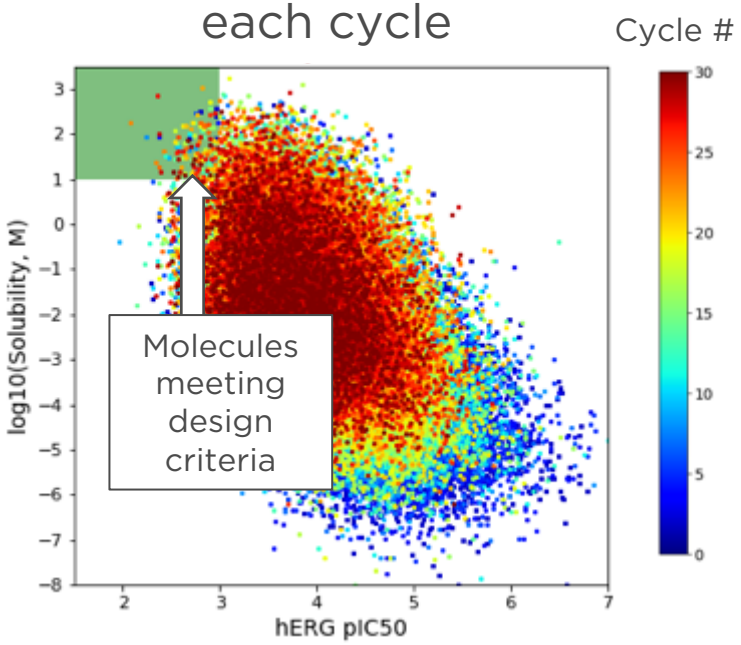


Generative Molecular Design (GMD)

Iteratively generate new compounds with better properties



Working compound library improving at each cycle



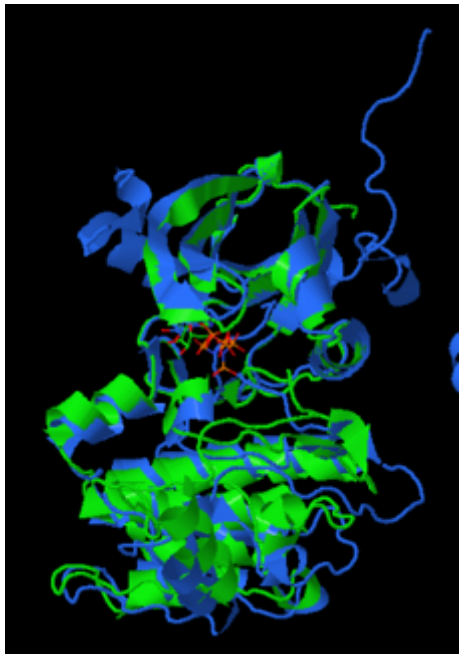
- Junction tree variational autoencoder transforms molecules into continuous vector
- Genetic algorithm perturbs these vectors to create new molecules

Prediction & Design loop validation

Generative molecular design of AURK B inhibitors

Proof-of-Concept

PILOT 1 | Starting point: Early program data | Lead Optimization | End point: Experimental validation



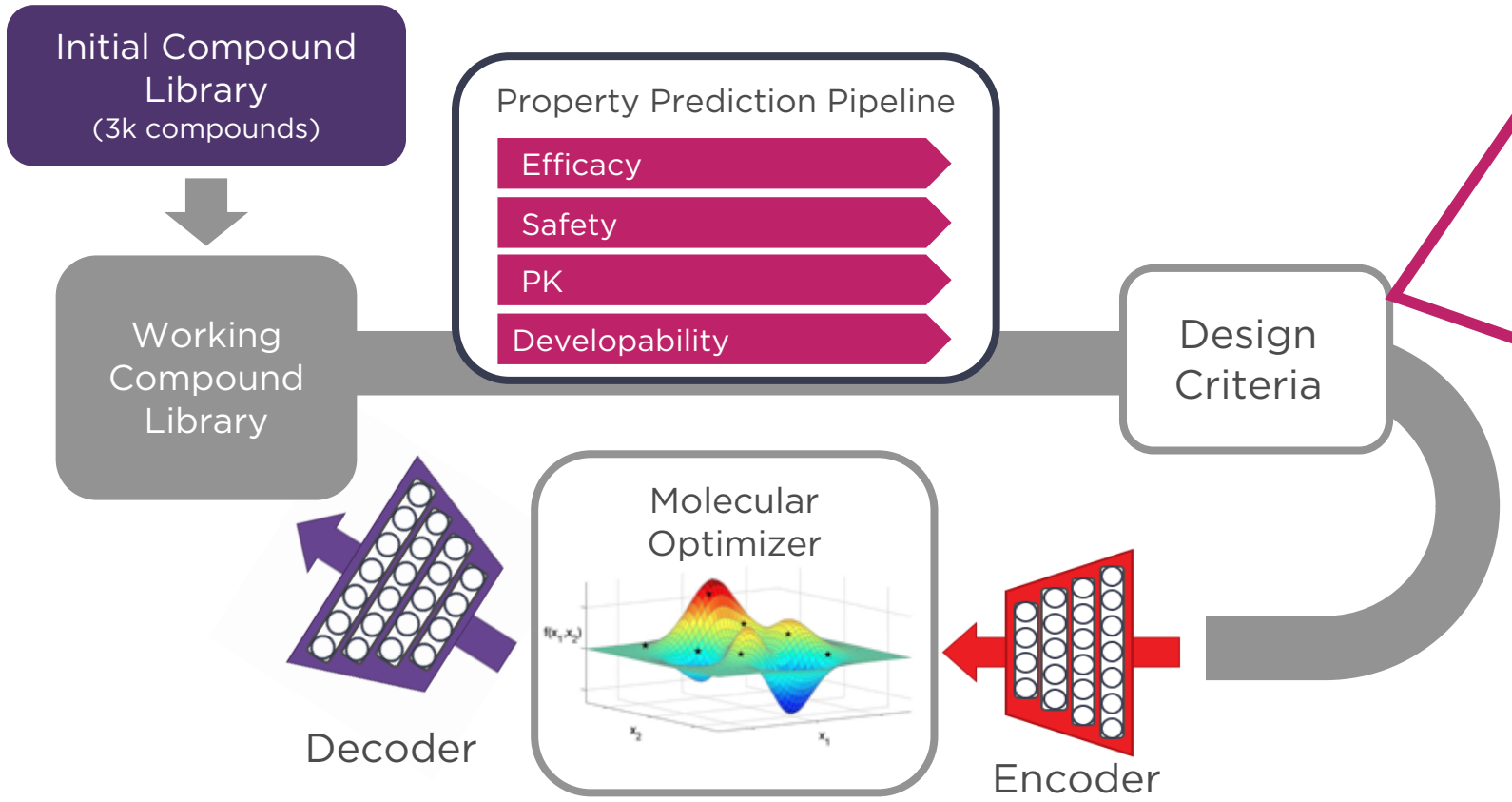
Structure overlay of AURK A and AURK B

Why Aurora Kinase?

- **Cancer relevant**: >30 clinical trials are ongoing or completed for AURKA selective, AURKB selective, and AURKA/B dual inhibitors
- **Internal data available:** Potency data on ~24k compounds available for AURK B and/or AURK A
- **Pharmaceutical discovery relevant problem:** Selectivity between kinases is an important and common pharmaceutical discovery problem

Design Criteria

Proof-of-Concept



Candidate Quality Panel

Efficacy
AURK B (pIC50 > 9)
AURK B/A Selectivity > 1000

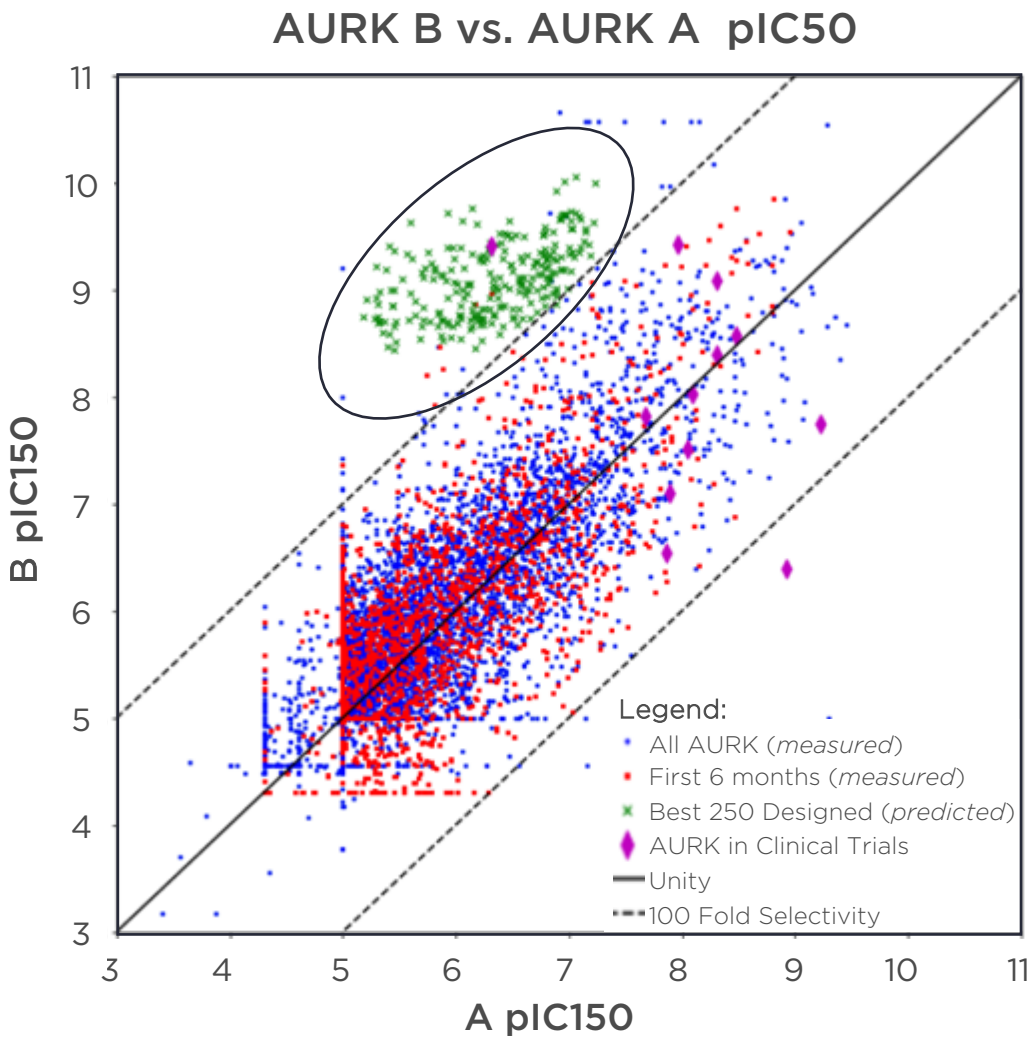
Safety
BSEP (pIC_{50} < 4)
hERG (pIC_{50} < 4)

PK
Solubility (>10uM)
CL_{int} (<3 mL/min/g)

Developability
Solubility (>10uM)
SAS: Synthetic Accessibility Score
QED: Quantitative Estimation of Druglikeness

Initial results: >200 new potent, selective AURK B compounds with favorable other properties

Proof-of-Concept



Other design criteria for top compounds:

B pIC50	A pIC50	B/A	hERG	BSEP	PK Sol	CL	Dev Sol	SAS
9.627	5.60	10772	3.260	4.010	6.022	1.819	412.492	2.640
9.724	5.92	6381	3.202	4.029	4.241	1.338	69.457	2.632
9.762	6.14	4174	3.197	4.027	4.535	1.322	93.249	2.410
9.298	5.98	2065	3.198	3.969	5.988	1.455	398.809	2.392
9.209	5.73	3024	3.200	4.027	7.000	4.371	1096.282	2.498
9.208	5.81	2477	3.195	4.027	5.413	1.868	224.400	2.397
9.626	6.18	2784	3.868	3.982	5.447	1.434	232.073	2.332
9.407	5.41	9984	3.259	4.018	3.704	1.252	40.620	2.784
9.353	5.75	4028	3.199	4.018	4.470	1.835	87.357	2.339
9.517	6.45	1160	3.223	3.976	4.353	2.024	77.733	2.222
9.252	5.79	2922	3.794	3.977	5.207	1.405	182.459	2.441
9.293	5.61	4851	3.197	3.994	4.006	1.479	54.916	2.627
9.334	5.56	5926	3.198	4.043	6.552	0.986	700.482	2.818
9.393	5.93	2911	3.198	4.026	5.343	1.595	209.163	2.624
9.397	6.05	2247	3.199	4.016	4.017	1.421	55.541	2.640
9.399	5.97	2682	3.211	3.993	3.554	1.632	34.955	2.255
9.193	5.96	1720	3.646	3.970	5.044	1.816	155.047	2.472
9.222	5.30	8342	3.215	4.048	5.936	0.888	378.391	2.628
9.327	6.25	1205	3.198	4.055	6.356	1.498	575.970	2.380
9.440	6.39	1116	3.380	3.968	4.635	1.775	103.039	2.361
9.129	5.88	1775	3.657	4.070	7.134	1.553	1254.501	2.278
9.338	6.14	1579	3.198	3.967	3.507	1.269	33.360	2.369
9.516	6.46	1136	3.202	4.067	6.818	0.858	913.920	2.464
9.278	6.21	1171	3.416	4.069	4.565	1.777	96.042	2.330
9.090	5.91	1509	3.210	4.052	6.827	0.880	922.242	2.433
9.365	6.43	869	3.210	4.020	6.059	0.936	427.917	2.686
9.107	5.53	3788	3.545	3.993	6.339	3.112	565.978	2.501
9.375	5.45	8340	3.199	4.022	2.663	1.340	14.338	2.754
9.650	6.05	3951	3.205	3.990	2.503	1.604	12.224	2.426
8.896	5.64	1821	3.209	4.027	6.561	1.374	707.083	2.181
9.648	6.48	1482	3.244	4.021	4.026	1.318	56.041	2.577
9.389	6.28	1284	3.198	4.055	5.250	1.037	190.604	2.754
9.075	5.98	1235	3.199	4.055	6.027	1.711	414.475	2.527
9.422	6.35	1179	3.202	4.014	3.214	1.569	24.878	2.503
9.298	6.27	1063	4.026	4.058	5.720	1.863	304.910	2.474
9.108	5.80	2028	3.198	4.115	5.448	0.973	232.219	2.608
8.969	5.60	2333	3.925	3.987	5.674	3.901	291.332	2.224
9.162	5.70	2884	3.198	4.069	4.387	0.800	80.426	2.520
9.154	5.87	1907	3.198	4.024	3.459	1.411	31.775	2.319
9.294	6.41	767	3.253	4.000	4.356	0.939	77.924	2.522

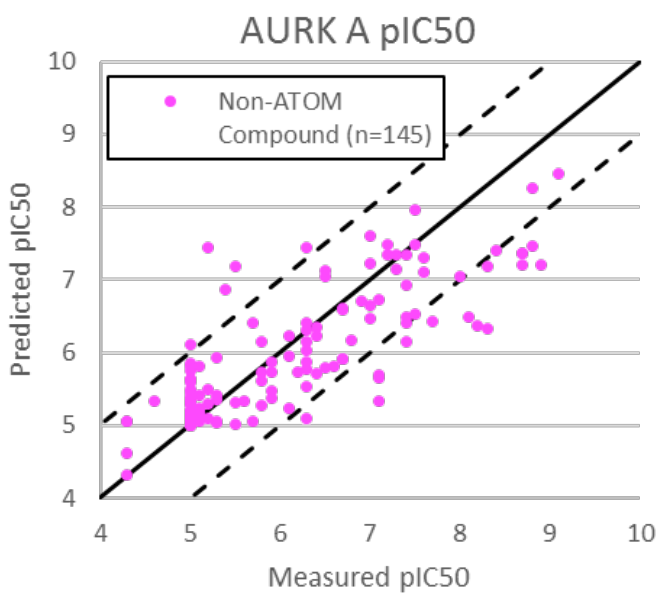
criteria met
close to criteria
criteria not met

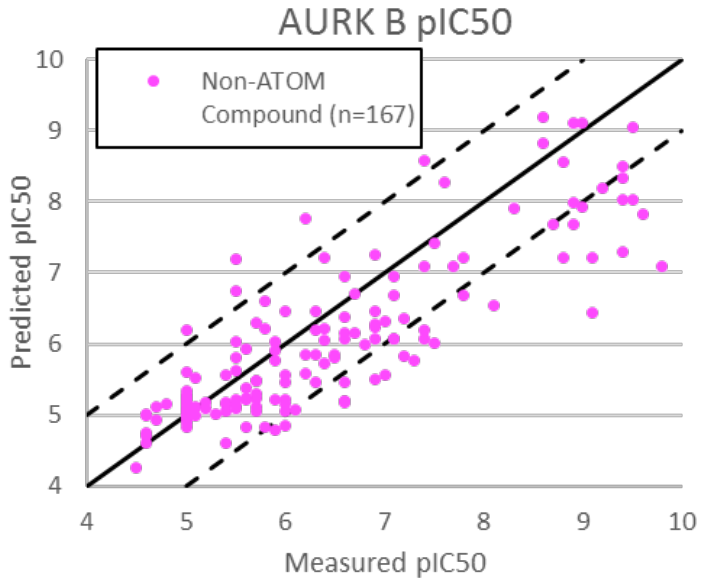
Generated compounds with existing data for comparison

High actual vs. predicted values for AURK

Generated compounds **not in** our internal dataset are well predicted

R^2:
AURK A : 0.68
AURK B: 0.75





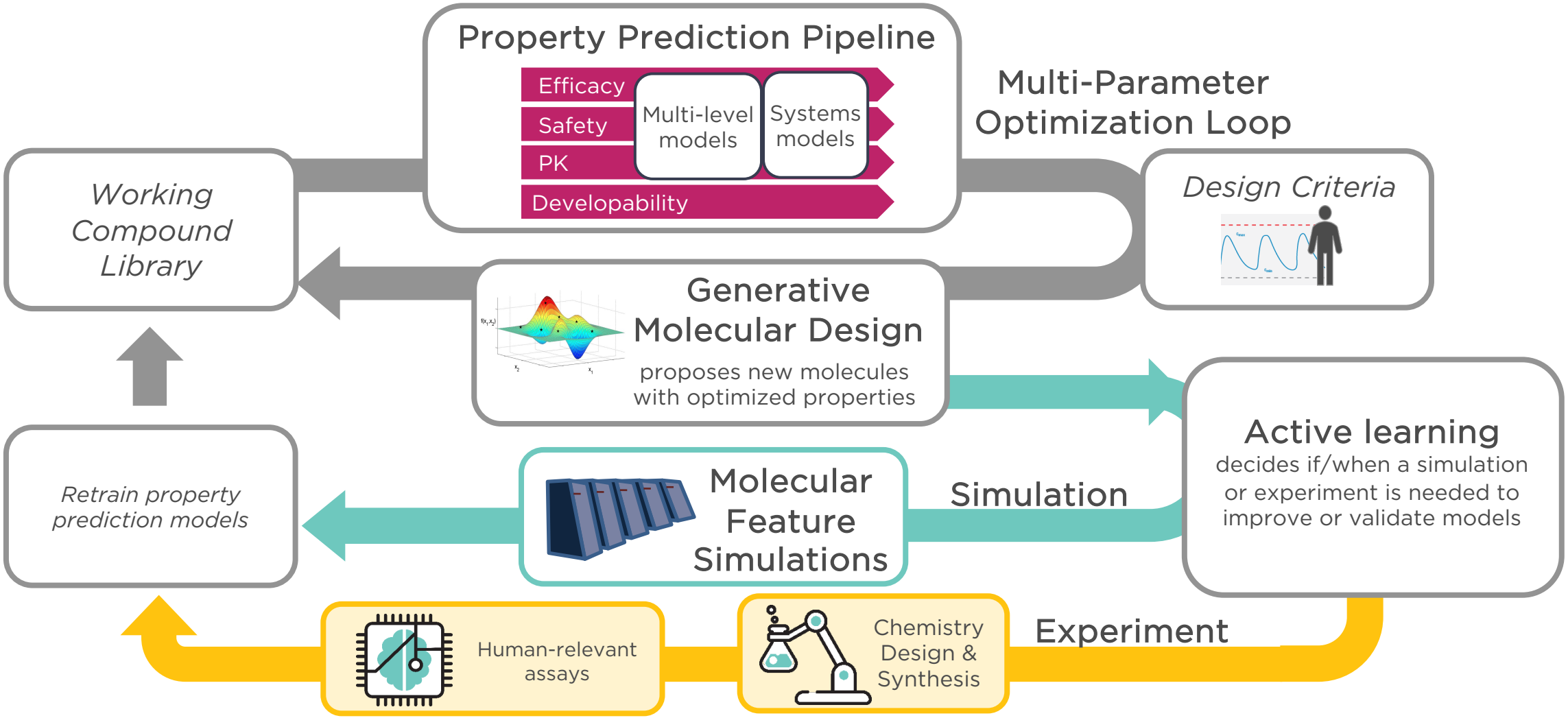
Initial Results for Make/Test Cycle Generally Favorable at Meeting Criteria

Criteria	Target	Total Returned	In Target Range (Predicted)	Within 1 log of target
AURK B	pIC_{50} > 9	37	15 (9)	33 (36)
Selectivity	>1000 fold	38	4-6 (2)	7 (11)
hERG*	pIC_{50} < 4	57	10 (33)	40 (55)
BSEP	pIC_{50} < 4	55	9 (1)	23 (36)
CL_{int}	< 3 mL/min/g	0	N/A	N/A
Solubility	>10 ug/mL	0	N/A	N/A

* hERG LLQ is <4.3, all compounds at LLQ considered in target

GMD Ranking	AURK B pIC50	Selectivity	hERG pIC50	BSEP pIC50	AURK A pIC50
36	>10.6	>316	<4.3	3.7	8.1
37	>10.6	>3162	5.5	5.7	7.1
38	>10.6	>1259	4.6	5.5	7.5
42	>10.6	>631	6.3	4.1	7.8
48	>10.6	>3162	5.8	4.4	7.1
68	>10.6	>1995	4.9	5.1	7.3
47	10.2	158			8.0

Next step: incorporating active learning



Future work

- Molecular design loop
  - Multi-target profile QSAR and model sharing framework
  - Scaled up generative models
  - Integrate human systems-level PK and safety models
- Active Learning Loops
  - Experiment - automated chemical synthesis and assay loop
  - Computational- Optimal experimental design and integration of mechanistic models
- Pilot design studies of increasing complexity
  - Genomic target efficacy models
  - Network-based design initialization
  - Broader chemical space design models

Frederick National Laboratory
for Cancer Research

Lawrence Livermore
National Laboratory

U.S. DEPARTMENT OF
ENERGY

UCSF

Thanks to our partners and funding organizations