ATOM

Jonathan Allen, Ph.D.
Informatics Thrust Leader
allen99@llnl.gov
https://atomscience.org/

July 15, 2020

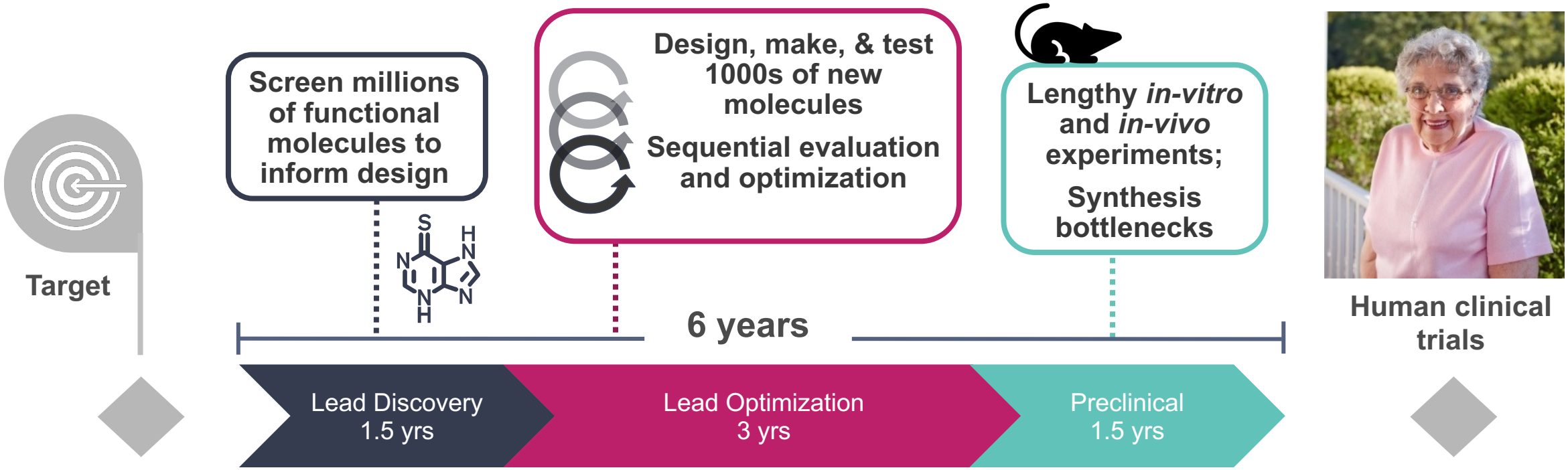
An Integrated Machine Learning Framework for Novel Small Molecule Drug Design

Roadmap of Talk

- Problem statement and introduction to ATOM Consortium
- Overview of ATOM computational platform
- Data and Data-Driven Modeling Pipeline
- Recent applications into systems modeling
- Applications of the small molecule generative design loop (GMD)

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

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

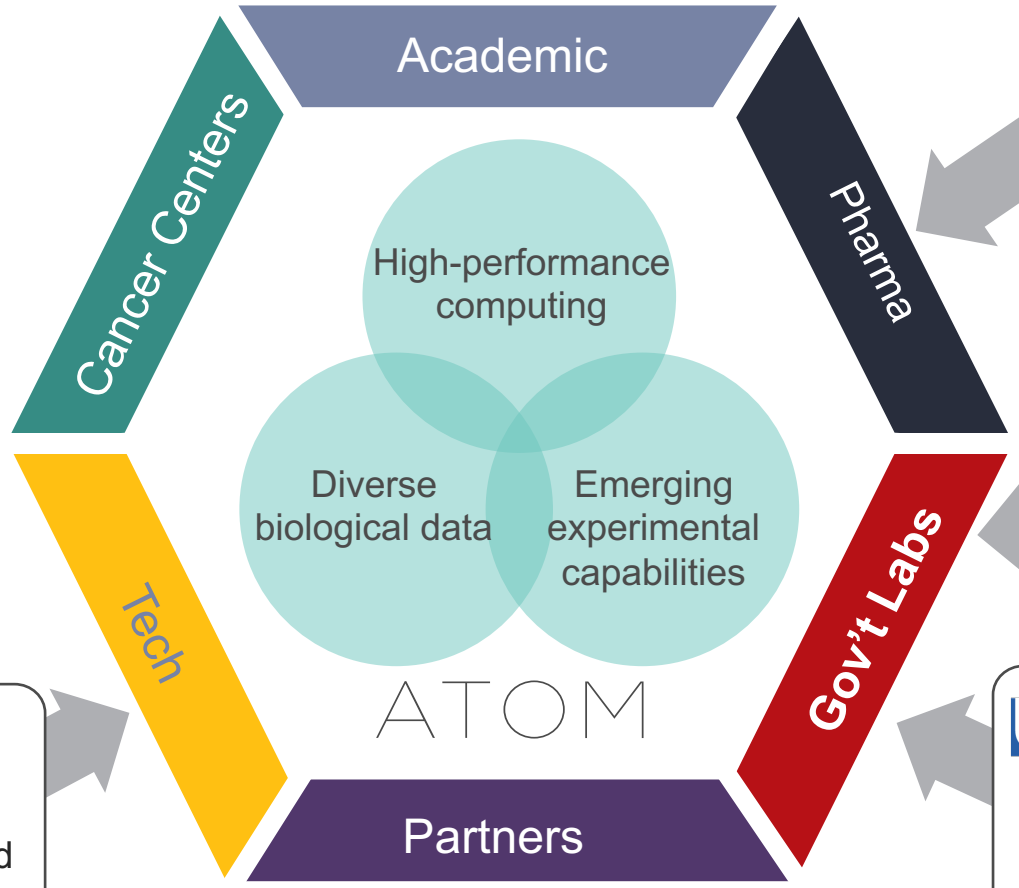
- Computation-driven drug design, supported and validated by targeted experiments
- Data-sharing to build models using everyone's data

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



Example member contributions

gsk

Drug discovery, chemistry, and dark data

Frederick National Laboratory for Cancer Research

Experimental biology and data analytics

Lawrence Livermore National Laboratory

High performance computing, data science, and honest broker

UCSF

Cancer biology and new assay development

ATOM provides an open platform, models, and data for a rapidly diversifying drug development ecosystem

University drug development teams

- Open tools to develop and optimize molecules
- Advancing further up the development value chain
- Education and training in new approaches

Pharma and biotech

- Precompetitive design optimization technology
- Open R&D on emerging technology, methods, and workflows

ATOM

- Open platform, models, and data
- Public-private R&D community

Computing technology community

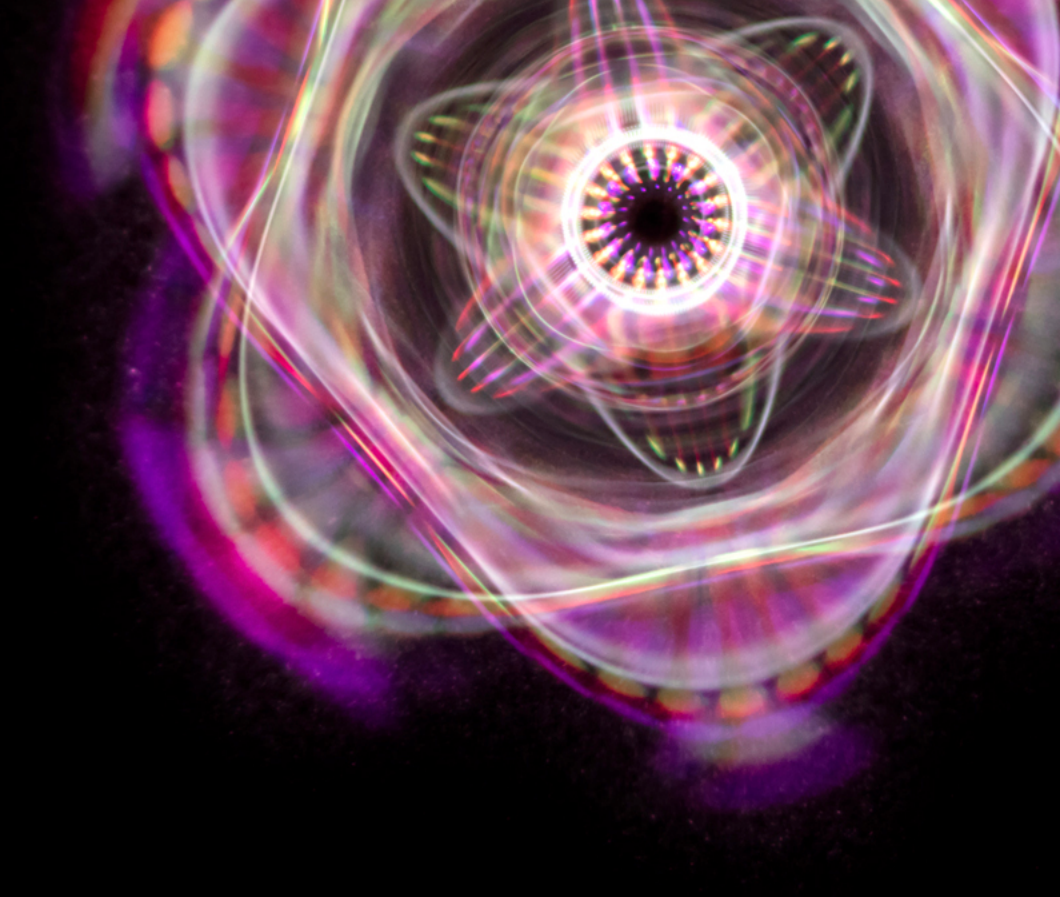
- Complex problems challenging and extending AI and HPC capabilities
- Scalable and supported approaches

Neglected disease communities

- Platform and research for drug design projects for public good
- New partnerships to expand the open research community
- Broad access to data

Government programs advancing public health and biosecurity

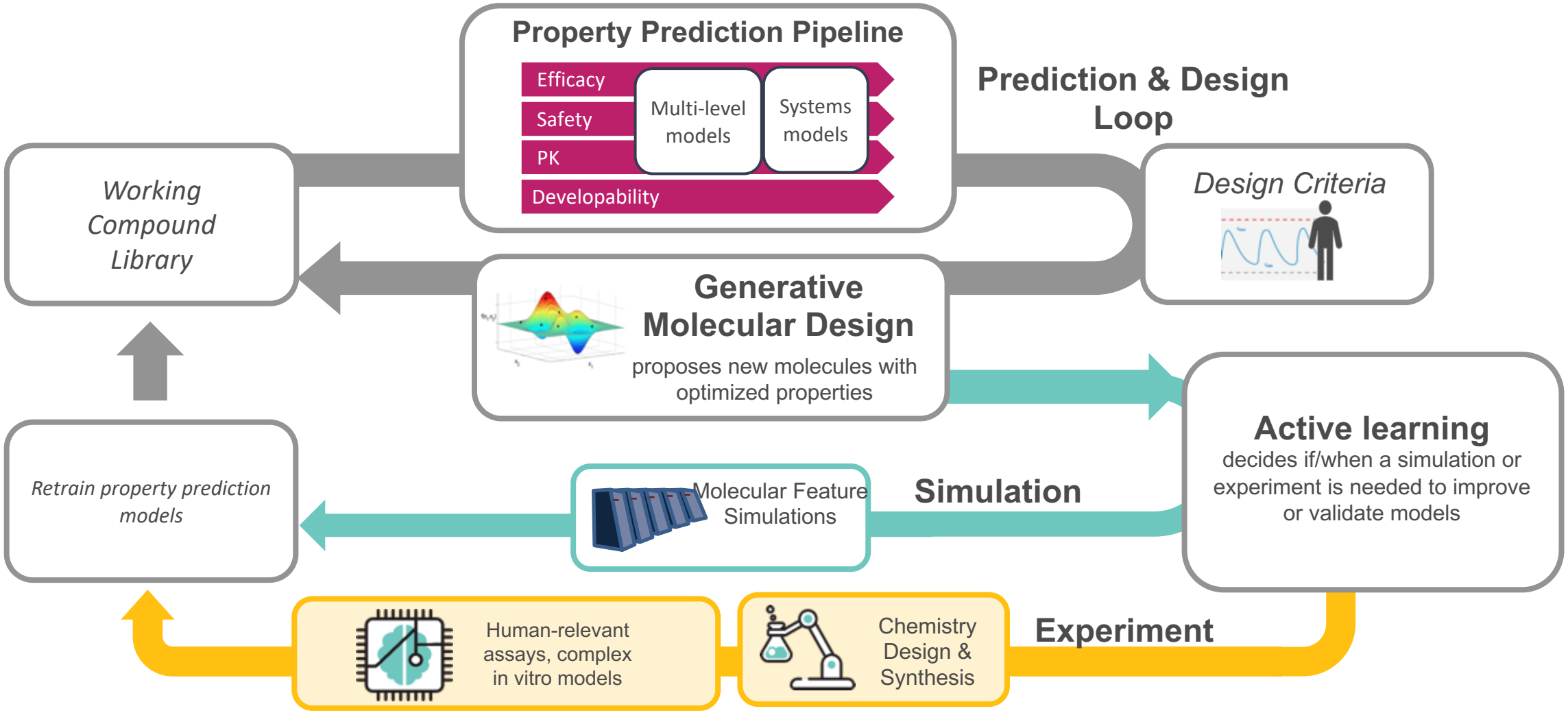
- Open platform supporting rapid response programs
- Public-private partnership programs
- Support for interagency collaboration



Building new predictive models

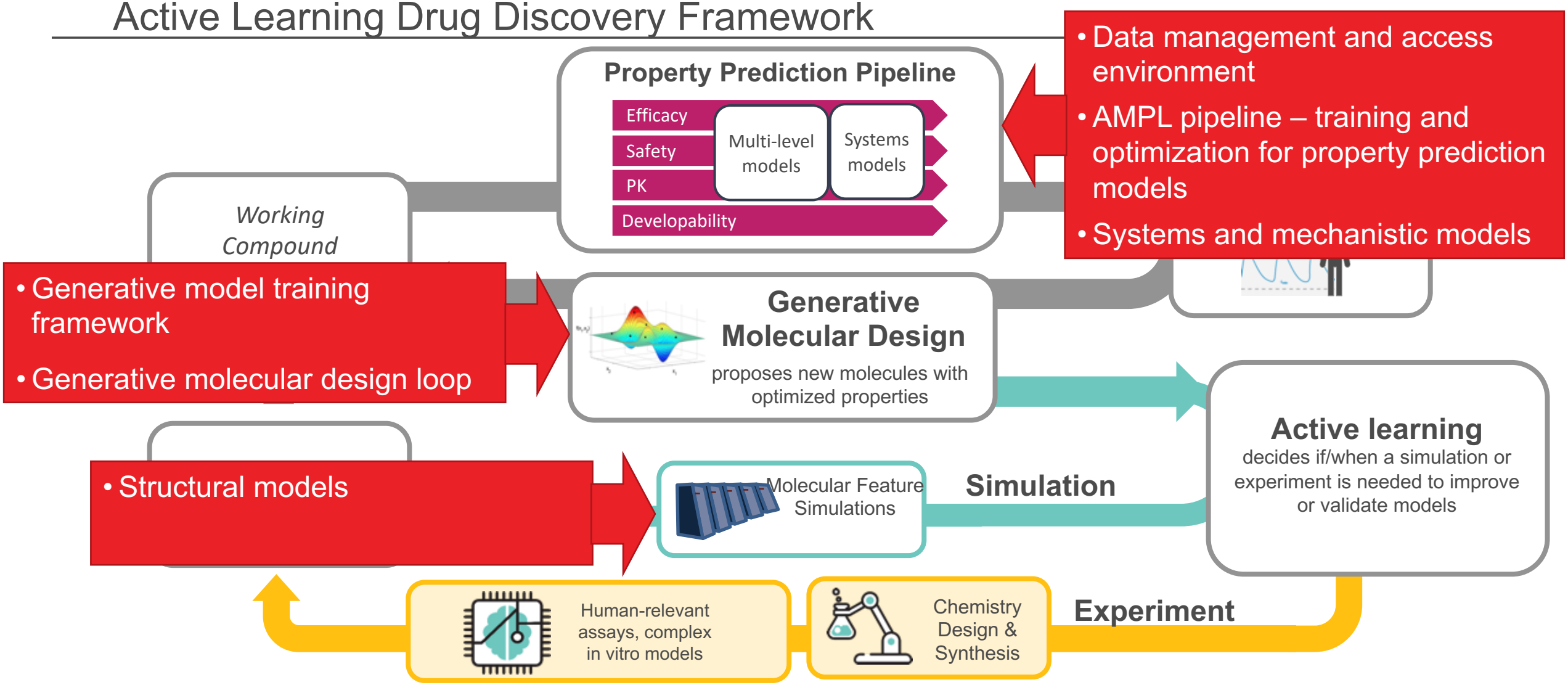
Machine learning frameworks

The ATOM Platform
Active Learning Drug Discovery Framework



The ATOM Platform

Active Learning Drug Discovery Framework



To build these workflows ATOM is focusing on several technical challenges

Challenges

Building a foundation of diverse open data sets

Predictive models for human relevant chemical properties

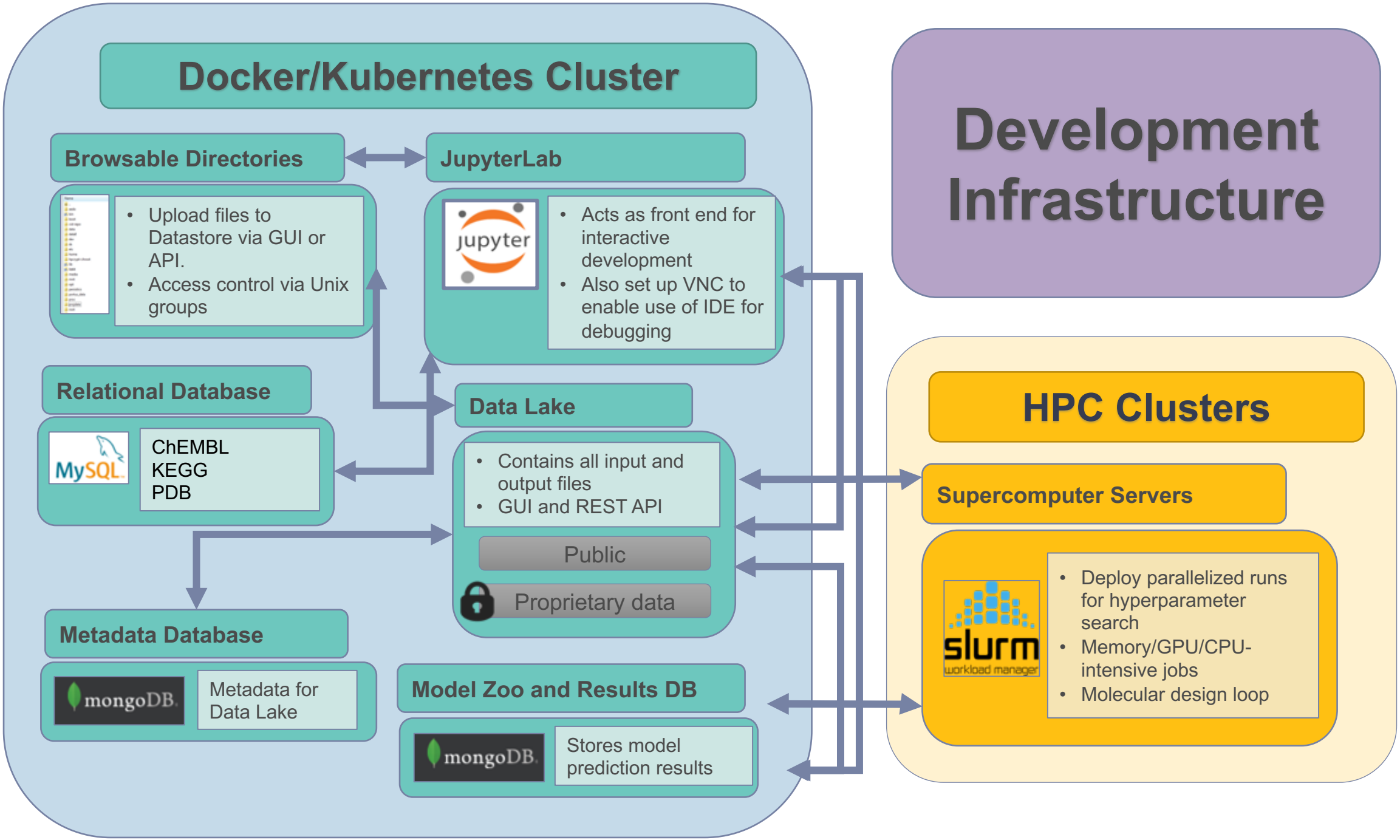
Active learning based molecular design

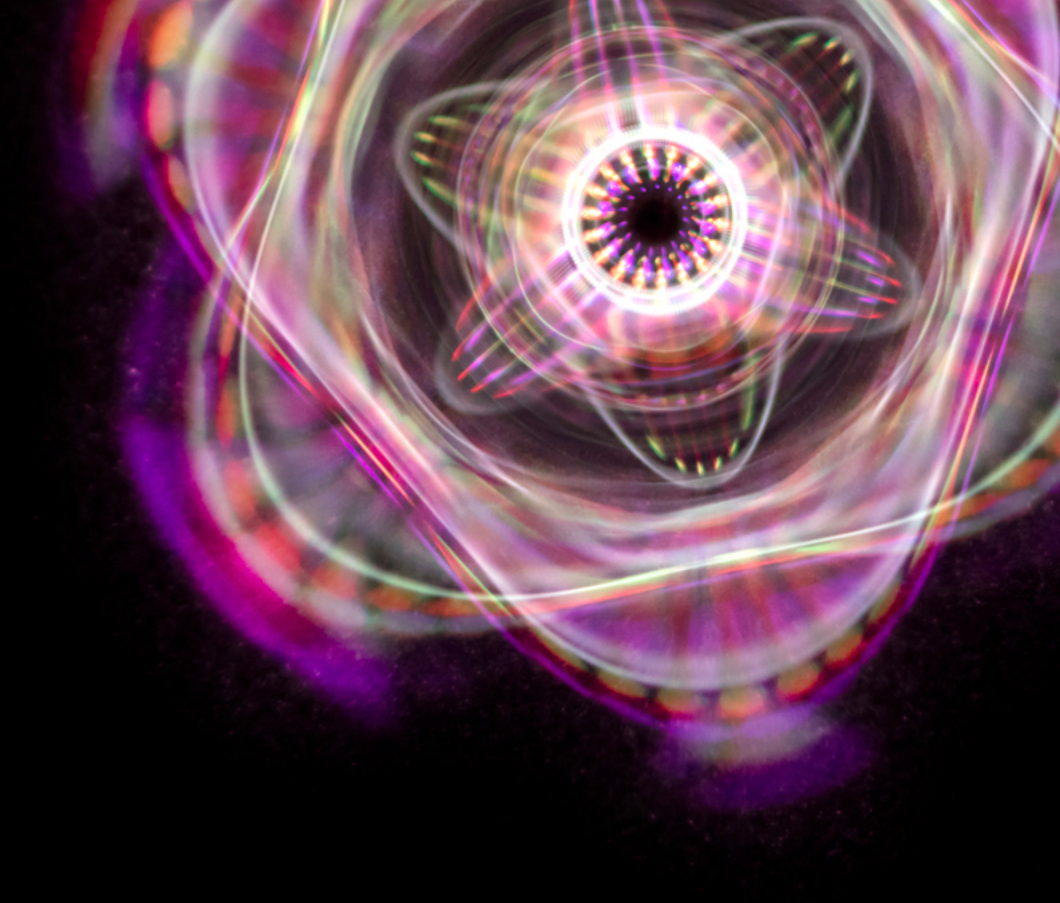
Approaches

- Creating partnerships to build and curate public data repositories
- Analysis and integration of data from multiple sources
- Targeted generation of new, human-relevant data

- Leveraging rapid progress in AI methods to improve performance and transfer to humans
- Broadening domain of applicability through secure integration of multisource models
- Using mechanistic models to effectively expand the datasets

- Design in a rigorous UQ framework
- Integrated multiparameter optimization of therapeutic window
- Optimal experimental design to drive active learning
- Automated chemical synthesis and assays in active learning loop

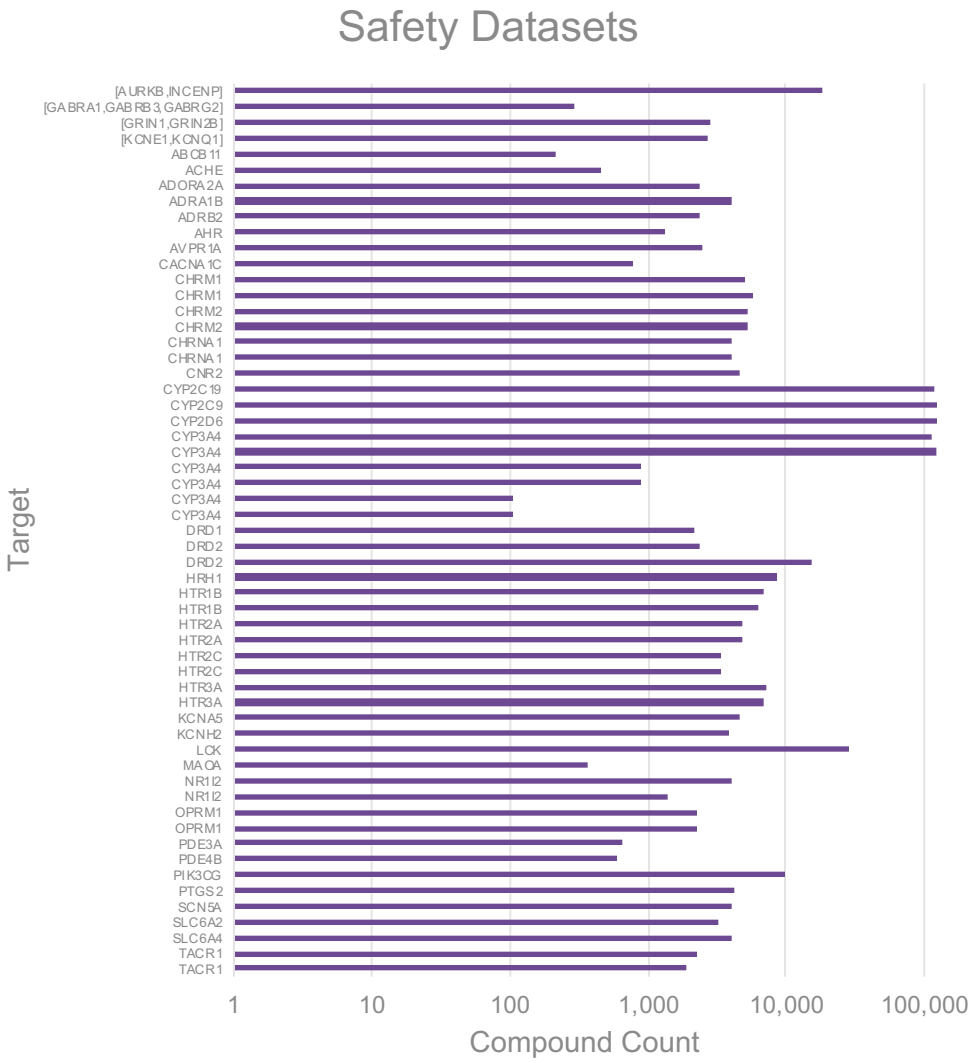

Docker/Kubernetes Cluster
Development Infrastructure
Browsable Directories
Upload files to Datastore via GUI or API.
Access control via Unix groups
JupyterLab
jupyter
Acts as front end for interactive development
Also set up VNC to enable use of IDE for debugging
Relational Database
MySQL
ChEMBL
KEGG
PDB
Data Lake
Contains all input and output files
GUI and REST API
Public
Proprietary data
Metadata Database
mongoDB
Metadata for Data Lake
Model Zoo and Results DB
mongoDB
Stores model prediction results
HPC Clusters
Supercomputer Servers
slurm
workload manager
Deploy parallelized runs for hyperparameter search
Memory/GPU/CPU-intensive jobs
Molecular design loop

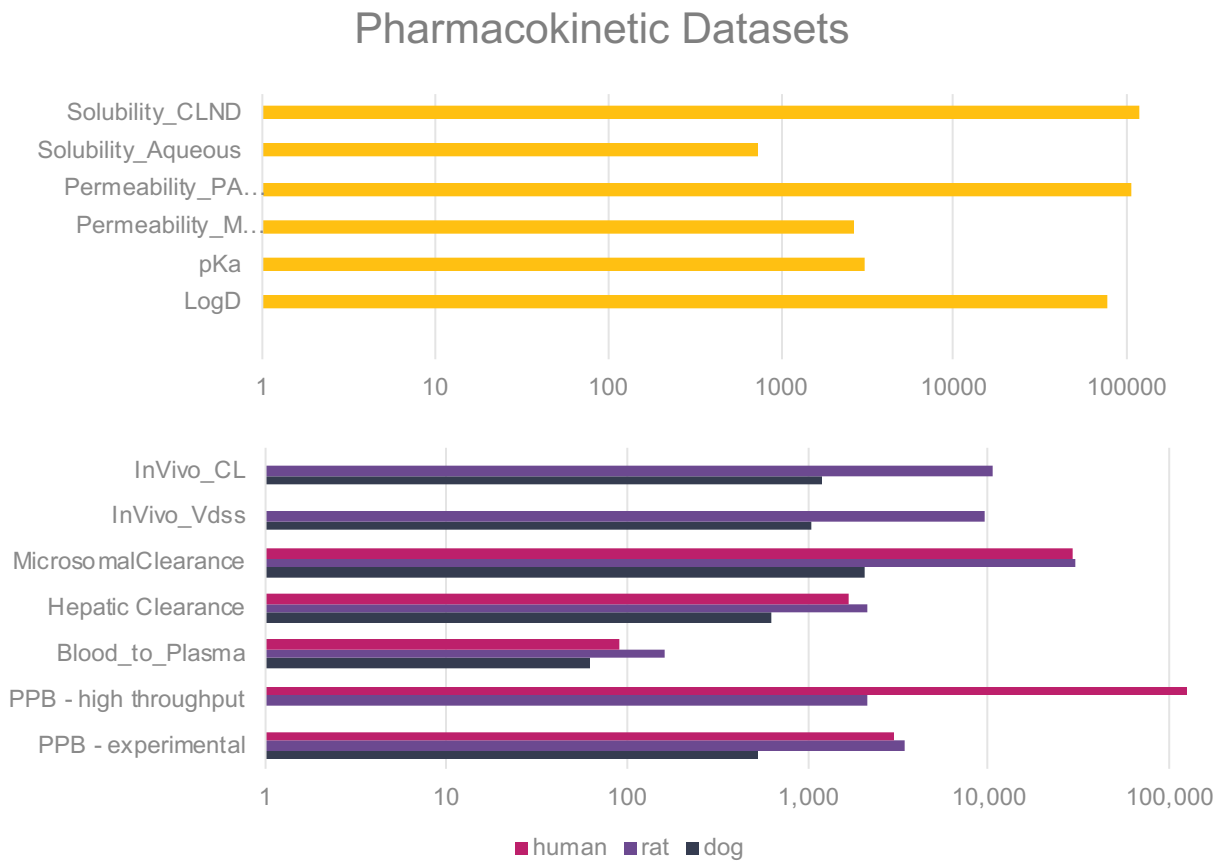


Expanding our data foundations

Curated model-ready datasets

ATOM has built models for hundreds of pharmaceutical data sets





ATOM

The ATOM data strategy

1. Work with available public data sources to build baseline safety and PK models
2. Expand databases using commercial data sources where required
3. Establish and enable open data partnerships to grow public data sources
4. Collect targeted data sets to fill gaps and emerging needs in open data

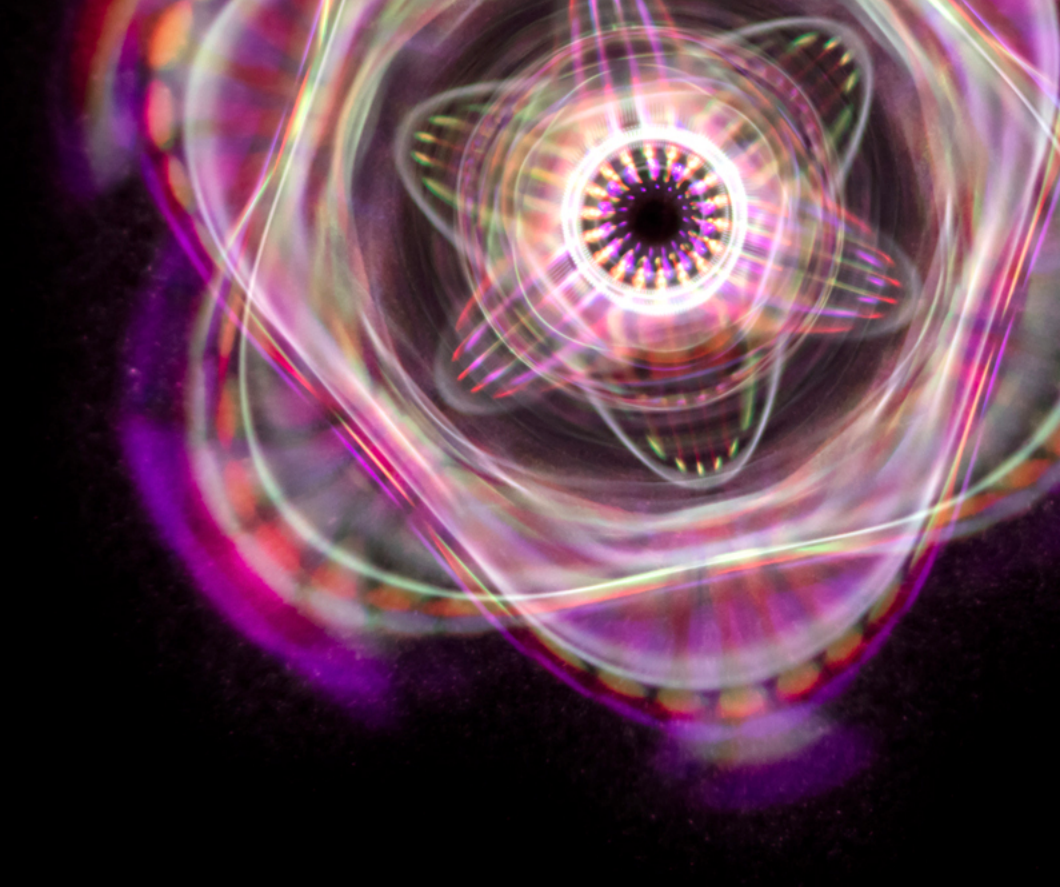
We're working with multiple data sources

CHEMBL – Manually curated repository of bioactive molecules
- 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

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

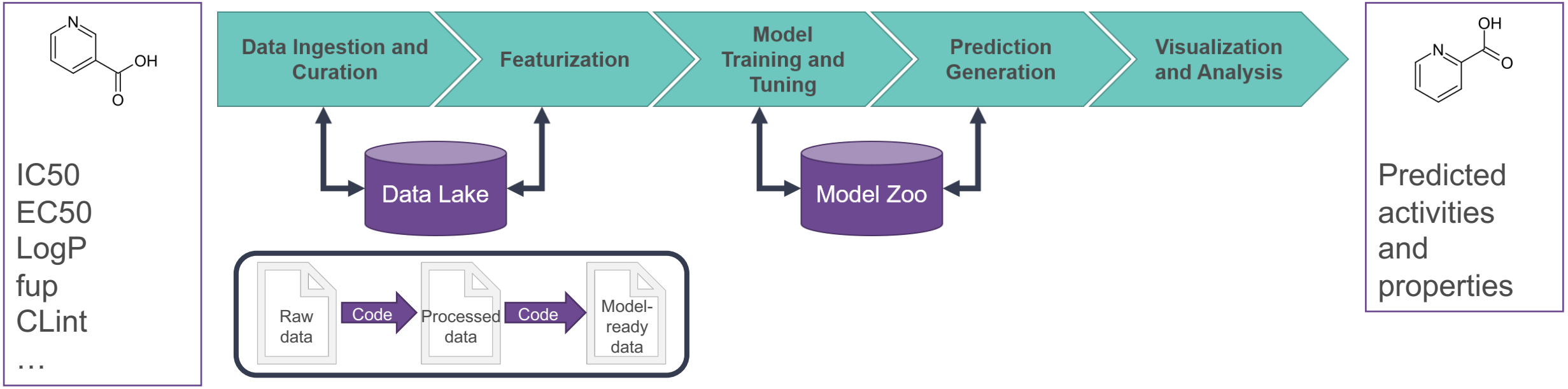
Drug Target Commons – An open multi-database platform for curation with common ontology
- *Sponsored by University of Helsinki*
- *Largest source is CHEMBL*
- *1.7M compounds, 13K targets*



ATOM Modeling PipeLine (AMPL)

The ATOM Modeling PipeLine (AMPL)

An open source software library for building and sharing machine learning models that predict bioassay activities or molecular properties from chemical structures



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

End-to-End Data-Driven Modeling Pipeline

Common infrastructure in place and ready to receive/transform new data

Data Ingestion & Curation
- Granular Access Control
- Python Query/Access APIs
- Data Taxonomy
- Web Interface
- Audit trail/ versioning

Splitting
- Random
- Scaffold
- Index
- Stratified
- Butina Clustering
- Asymmetric Validation Embedding (ATOMwise)
- *Dataset balancing*

Featurization
- ECFP
- MOE
- Mordred
- Graph Convolution
- 3D Features (E3FP)
- UMAP reduced features

Architecture
- Deep Neural Network
- Random Forest
- XGBoost
- *kNN-classifier*
- *SVM (SVC & SVR)*
- *Native Bayes*
- *SGD Classifier*

Model Training
- Grid Search
- Random Search
- *Bayesian Optimization*
- *Genetic Optimization*
- *GAN Optimization*

Visualization & Analysis
- Train/Test/Valid ROC AUC/R^2
- k-Fold Validation
- Confusion Matrix
- PCA, tSNE, UMAP
- Uncertainty Quantification

Prediction Pipeline
- Single/Batch Prediction
- Model Streaming
- Model Ensembling

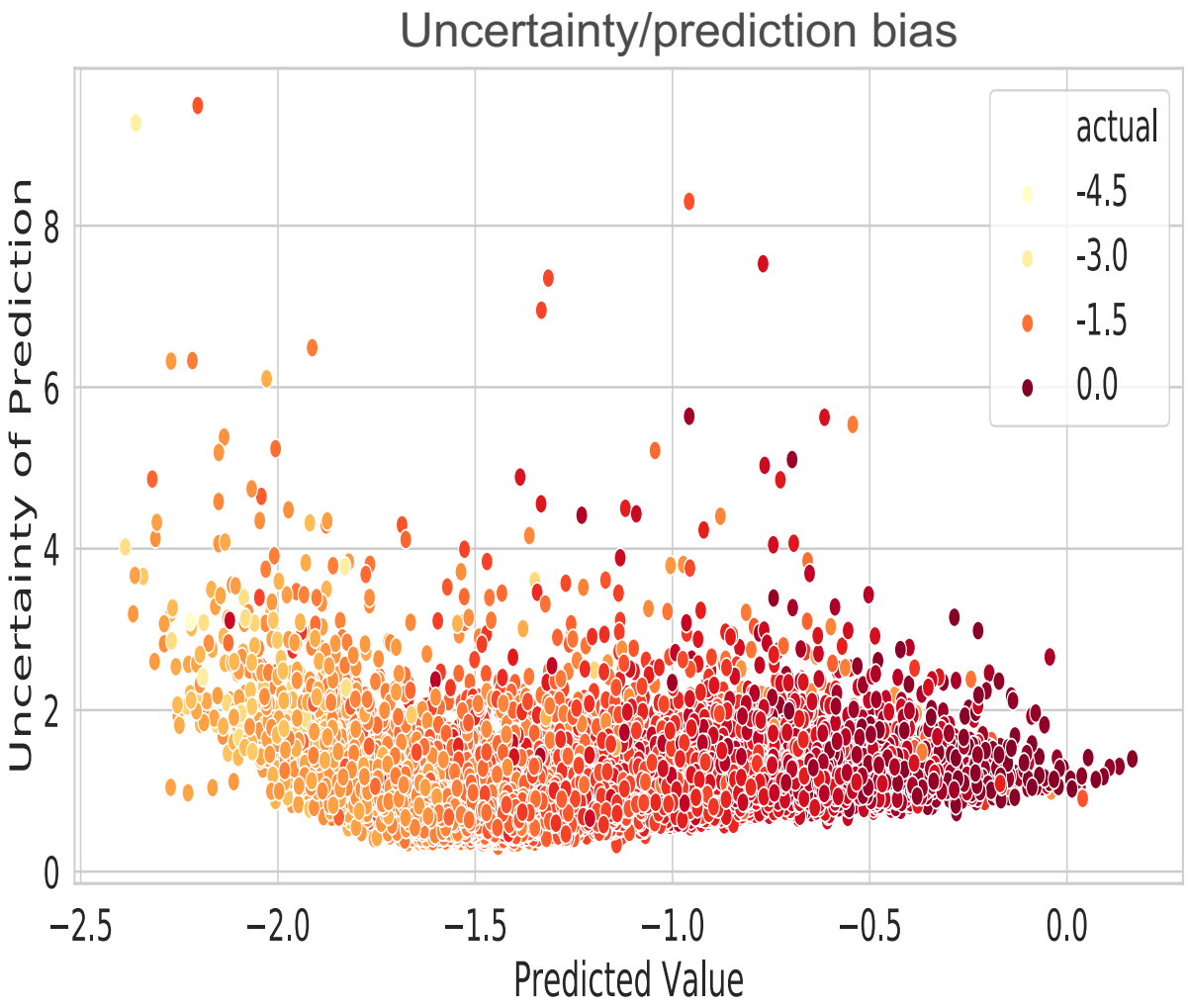
Benefits:

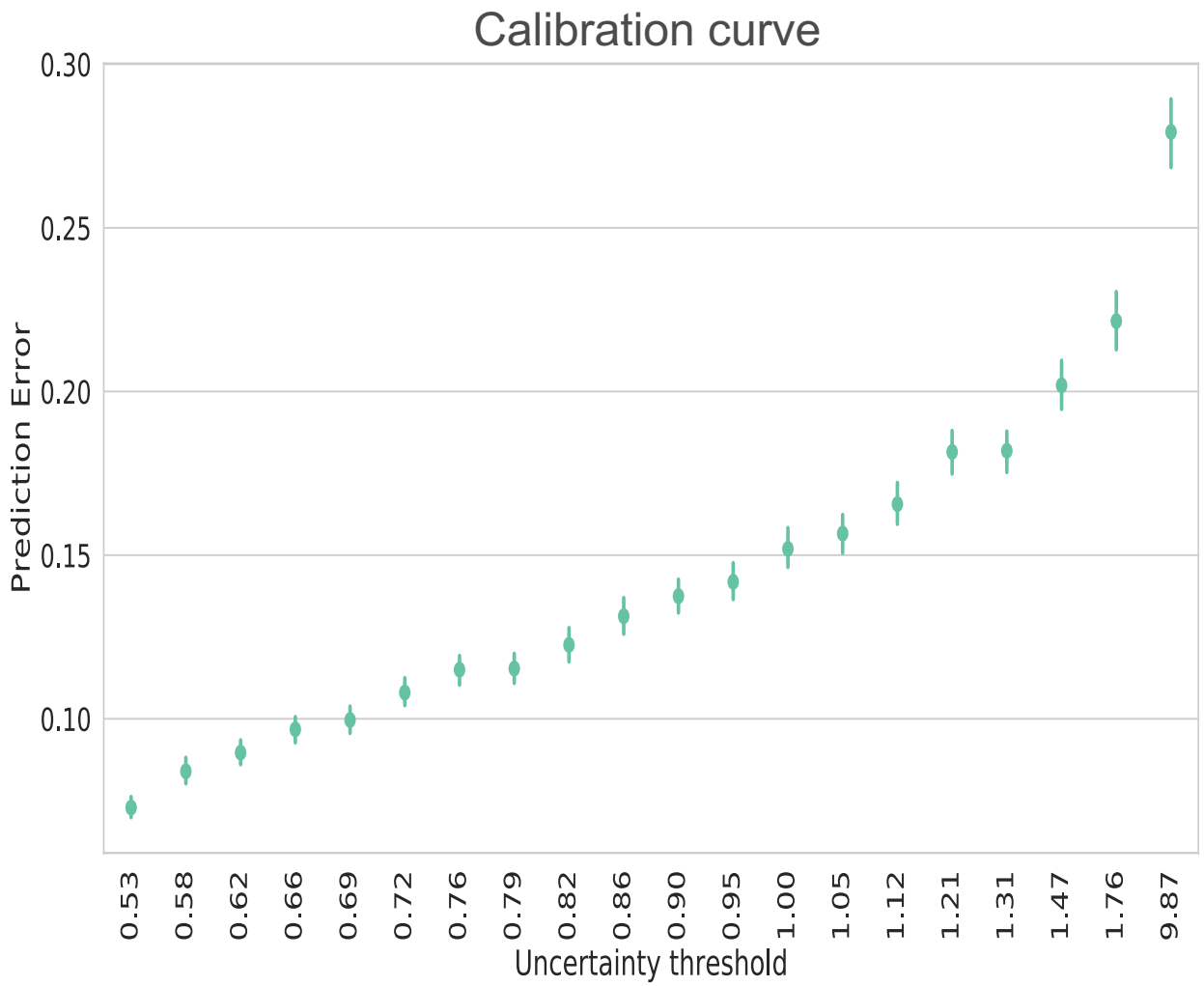
- Easy integration of diverse datasets
- Integration with scalable data and model services environment
- High-performance hyperparameter optimization
- Rapid evaluation of model architecture
- Seamless HPC integration using world-class compute systems
- Ensemble integration of models from multiple sources

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 (Kendal and Gal 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

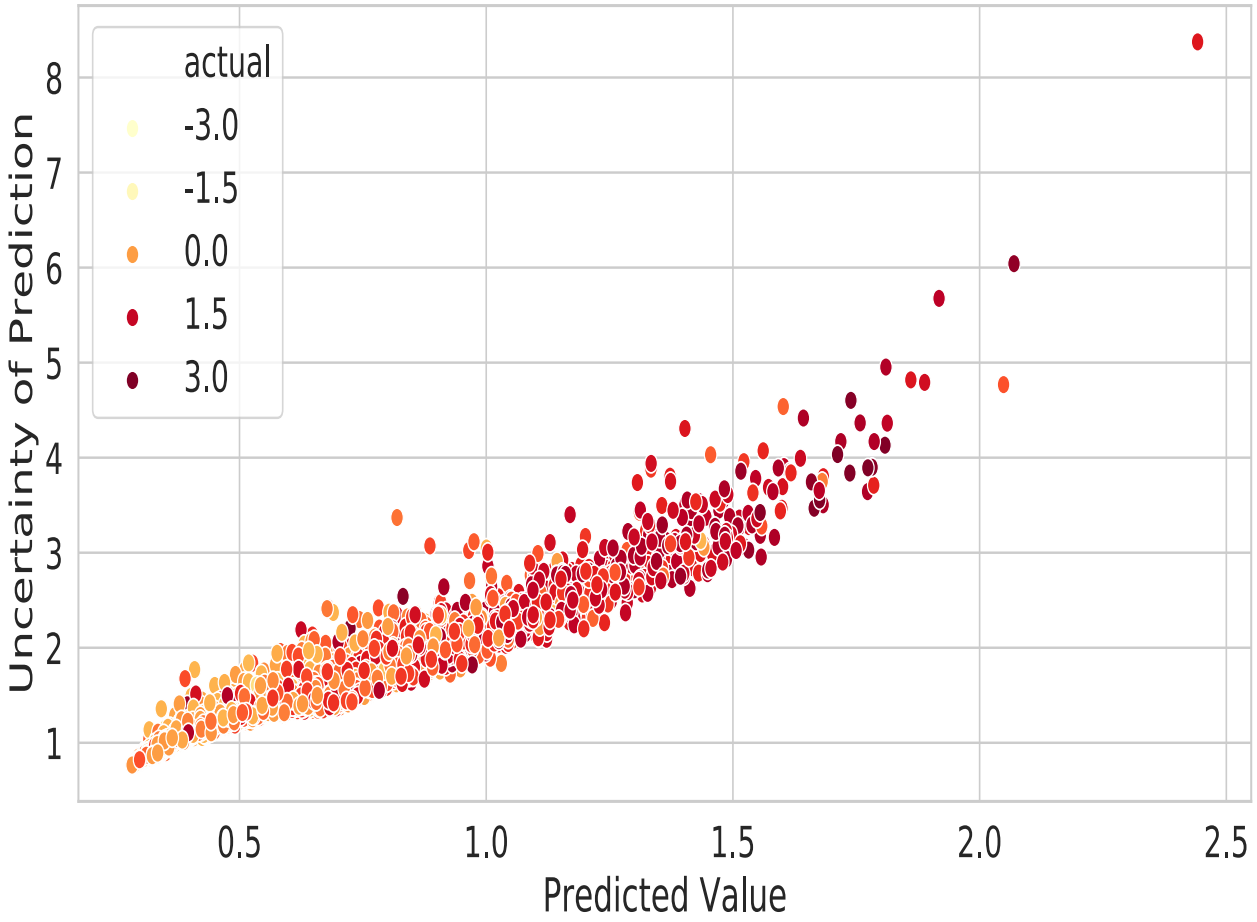




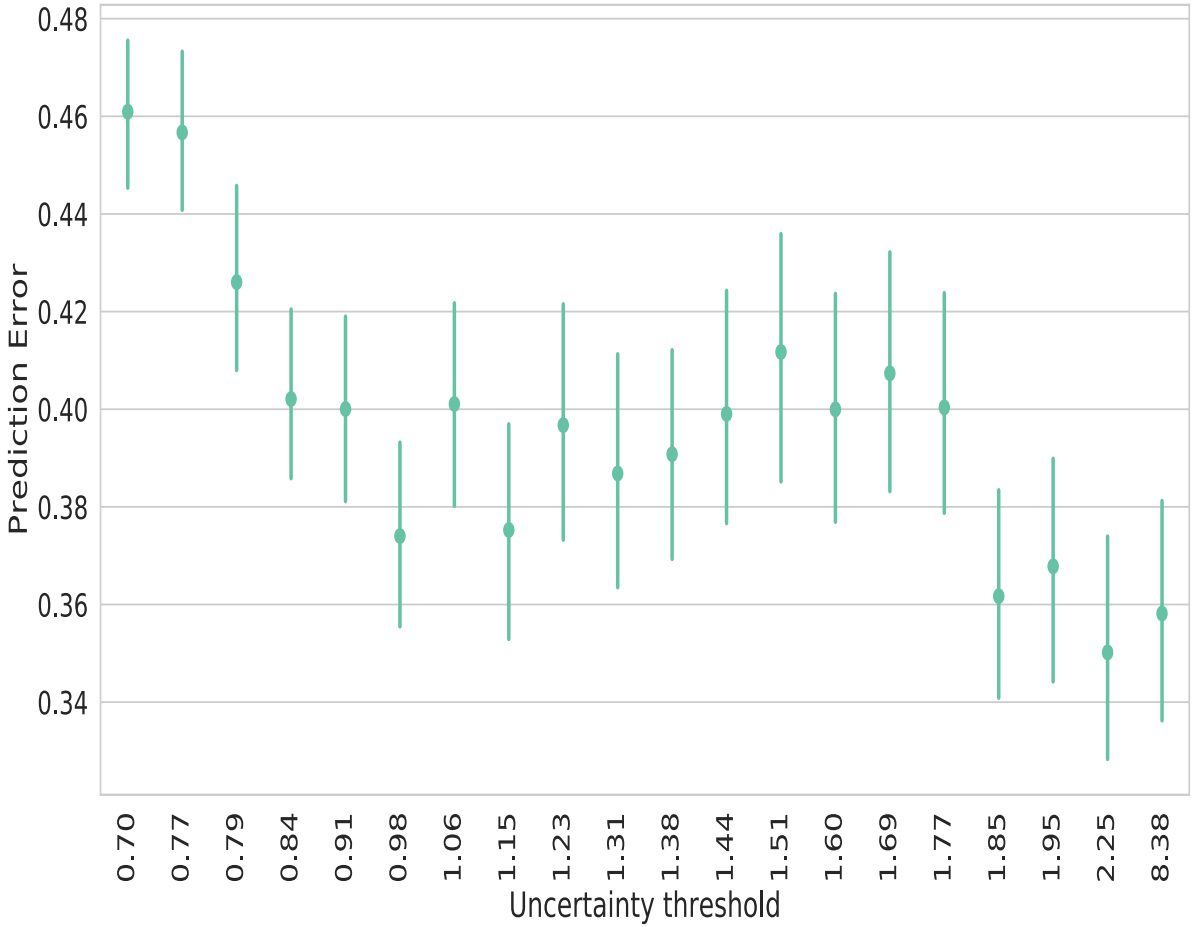
Blood plasma binding (HSA)

Model derived uncertainty varies depending on the model and dataset

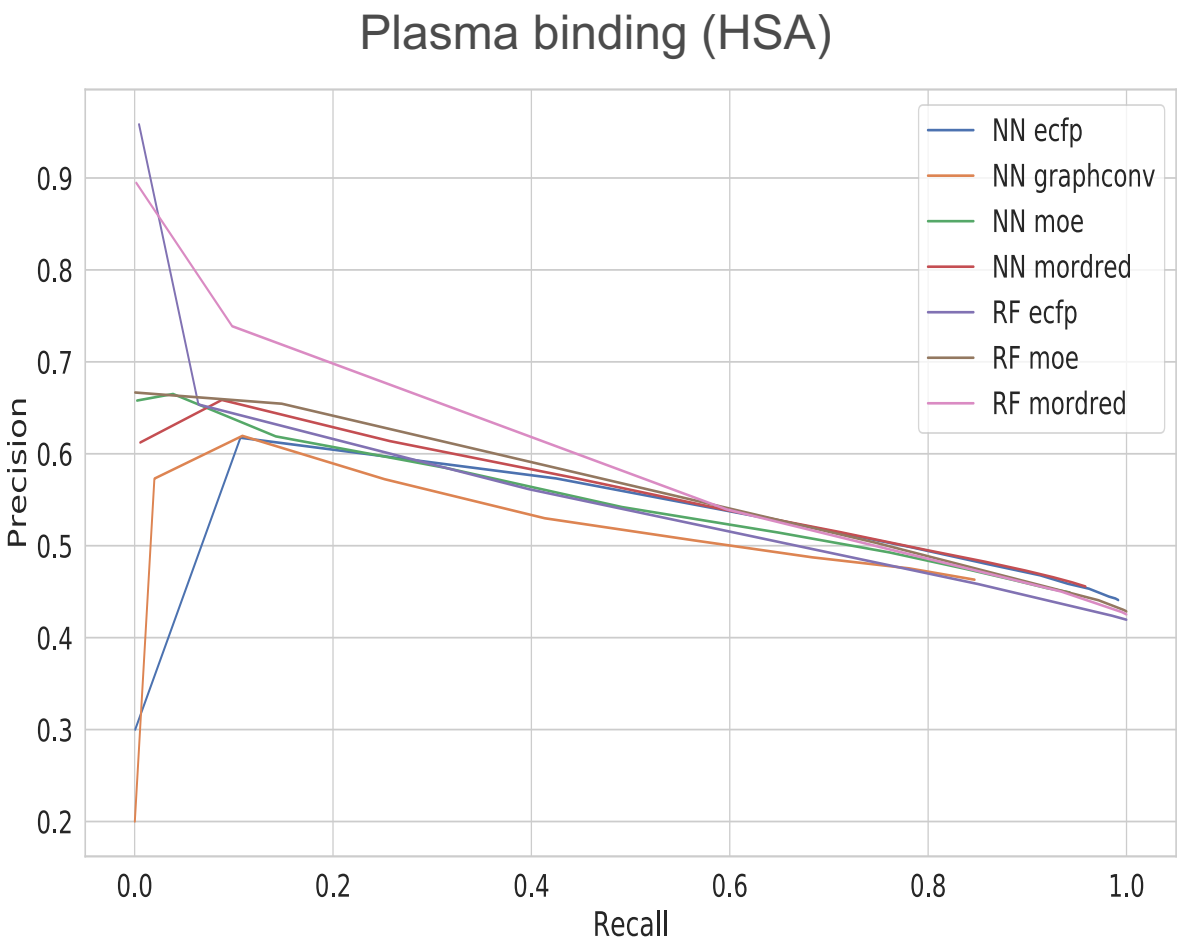
Uncertainty/prediction bias

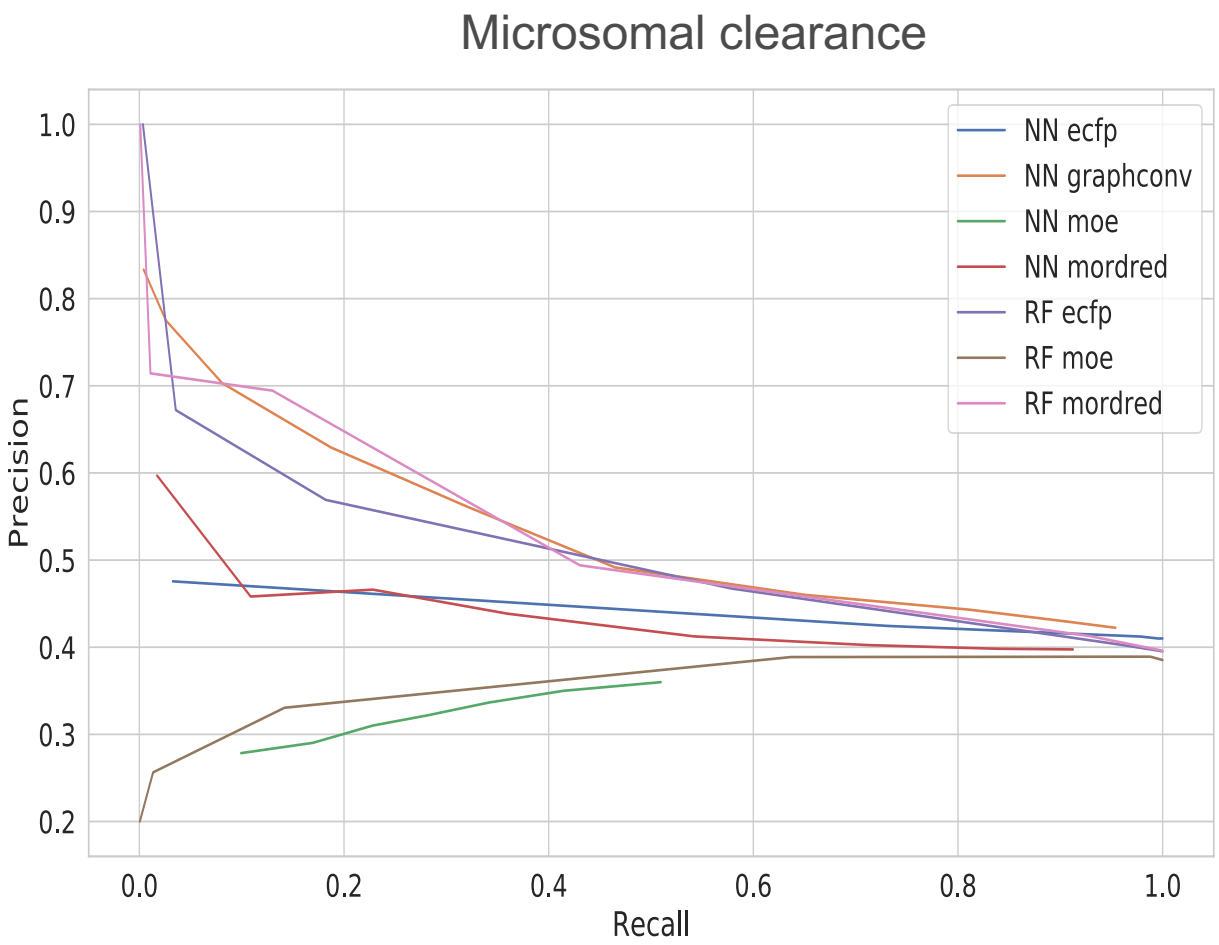


Calibration curve



Model derived uncertainty varies depending on the model and dataset





AMPL has been released open source

Fork me on GitHub

README.md

ATOM Modeling PipeLine (AMPL) for Drug Discovery

license mit

Created by the Accelerating Therapeutics for Opportunites in Medicine (ATOM) Consortium

ATOM

AMPL is an open-source, modular, extensible software pipeline for building and sharing models to advance in silico drug discovery.

> The ATOM Modeling PipeLine (AMPL) extends the functionality of DeepChem and supports an array of machine learning and molecular featurization tools. AMPL is an end-to-end data-driven modeling pipeline to generate machine learning models that can predict key safety and pharmacokinetic-relevant parameters. AMPL has been benchmarked on a large collection of pharmaceutical datasets covering a wide range of parameters.

A pre-print of a manuscript describing this project is available through ArXiv. readthedocs are available as well here.

AMPL: A Data-Driven Modeling Pipeline for Drug Discovery

Amanda J. Minnich, Kevin McLoughlin, Margaret Tse, Jason Deng, Andrew Weber, Neha Murad, Benjamin D. Madej, Bharath Ramsundar, Tom Rush, Stacie Calad-Thomson, Jim Brase, and Jonathan E. Allen*

Cite this: *J. Chem. Inf. Model.* 2020, 60, 4, 1955−1968
Publication Date: April 3, 2020
https://doi.org/10.1021/acs.jcim.9b01053

Article Views	Altmetric	Citations
1565	4	1

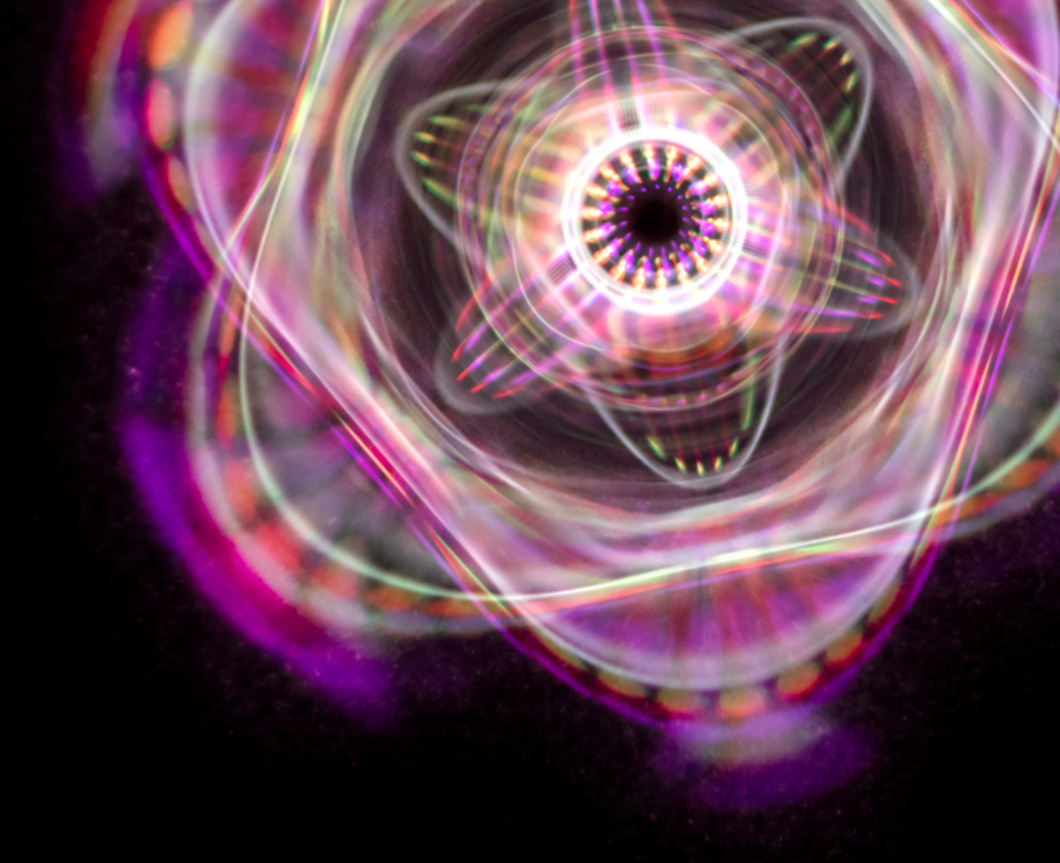
LEARN ABOUT THESE METRICS

Share Add to Export

RIS

https://pubs.acs.org/doi/full/10.1021/acs.jcim.9b01053

https://github.com/ATOMconsortium/AMPL

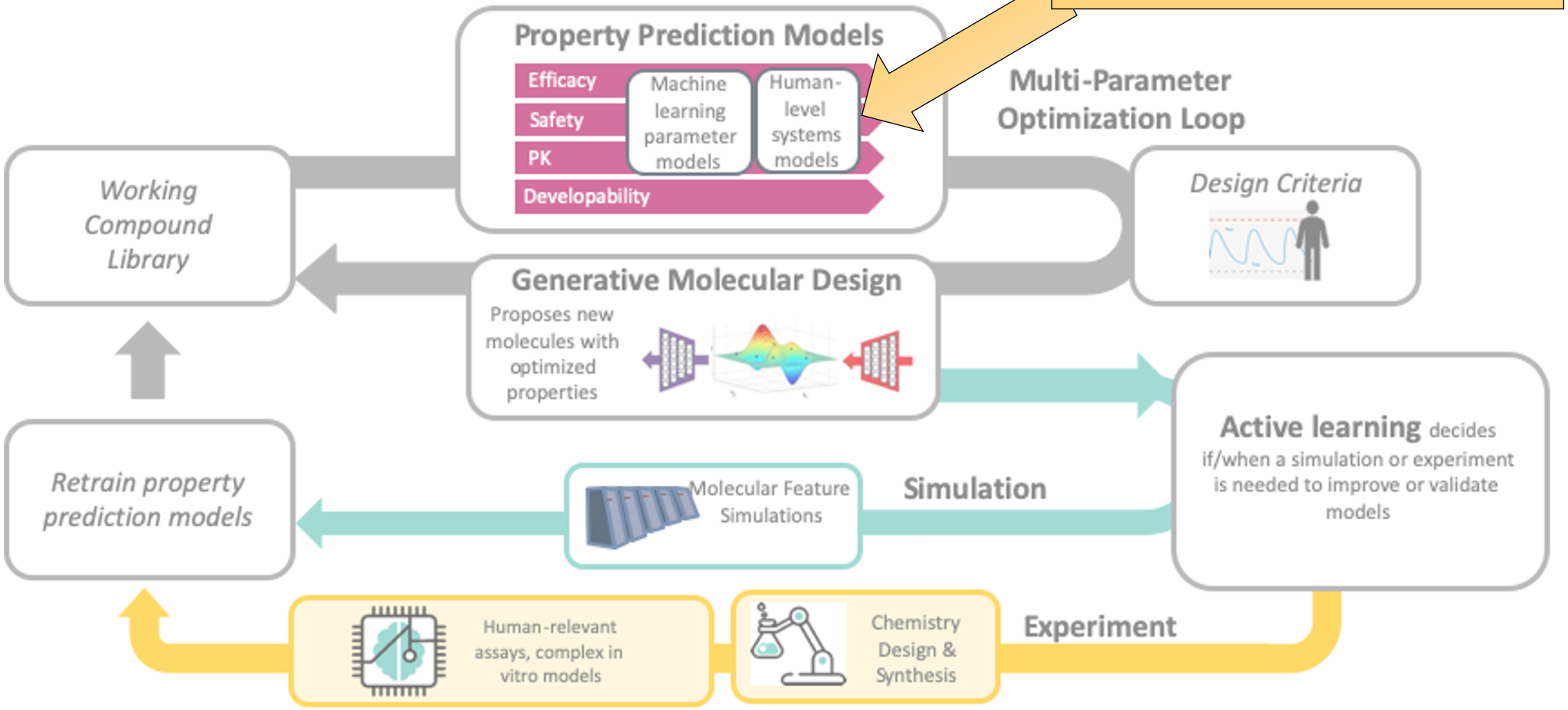


Building new predictive models

Human-level system models

Modeling frameworks support the ATOM workflow

- Human PBPK model
- Drug-induced liver injury model

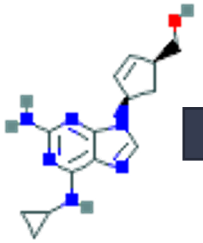


Current PK modeling activities

A limited first case – Single ML model

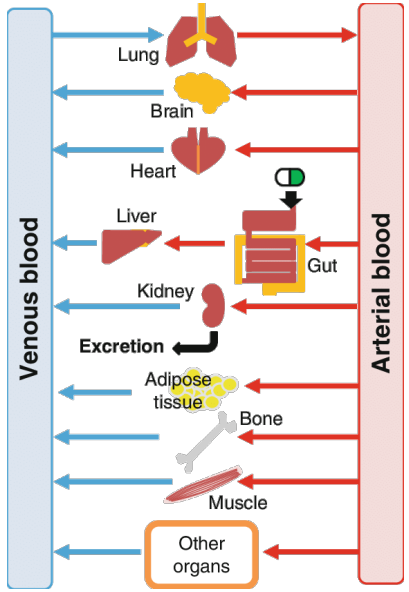
Structure to Input Parameters → **Human PK Simulator (PBPK)** → **Time-Concentration Profiles** → **Summary PK Parameters**

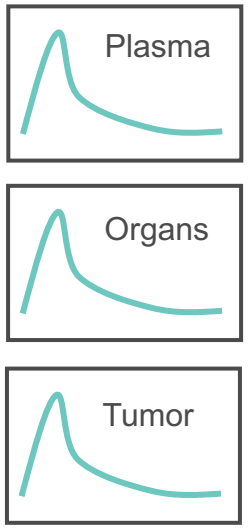
Parent Molecule(s)

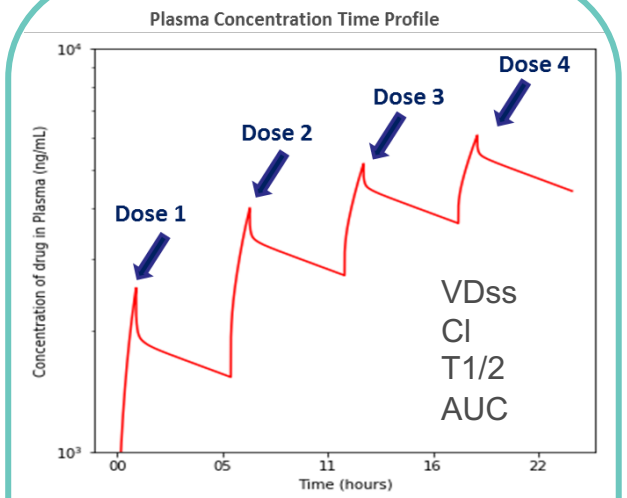


ATOM ML Models

- Parameters: solubility, permeability, pKa, B/P...
- Data curation of *in vitro* parameters from existing datasets
- *in vitro* experiments to fill data gaps







Curated human *in vivo* data (Obach, 2008; Lombardo, 2018)

Generating world class open data for PK modelling

Valuable data sets by combining curated and newly generated data

Started from *in vivo* Obach-Lombardo in-vivo data set and adding

ATOM Generated *In Vitro data*

300 compounds

- Hu Liver Microsomal Clearance
- Hu Liver Microsomal Protein Binding
- Plasma Protein Binding
- B/P Partitioning
- Log D (*in progress)*

Largest available set *in vitro* PK data with human *in vivo* data

ATOM Generated Novel Human Cell Line Data

200 Compounds

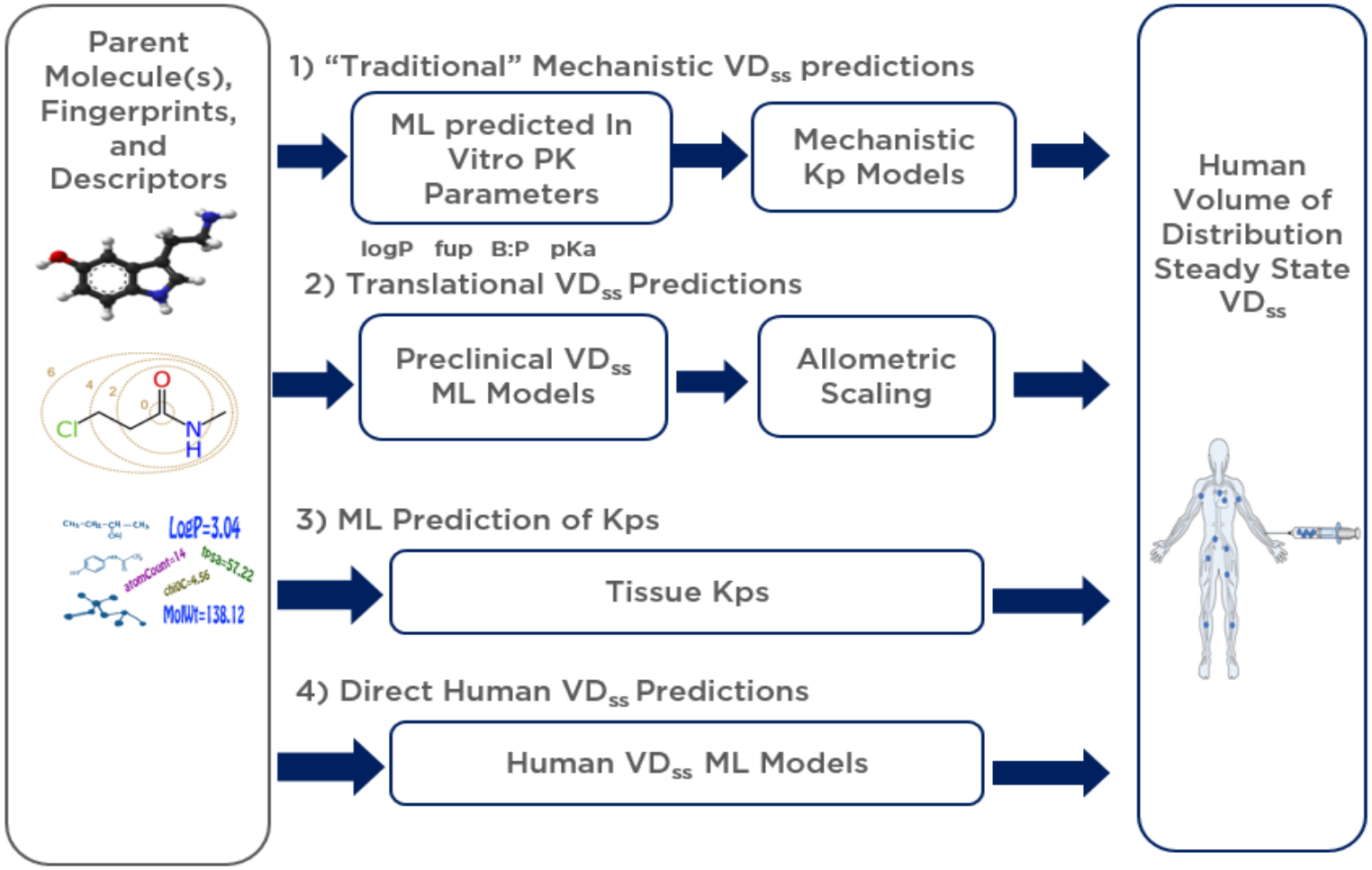
- Myocyte Partitioning
- Adipocyte Partitioning

Unprecedented human-relevant PK predictions

ATOM curated in-vitro and in-vivo data

- In vivo and in vitro data for Obach dataset (150-200 Compounds)
- ChEMBL datasets (ML ready)
  - Permeability
  - Log D
  - Log P
  - Fup
  - Microsomal Clearance (rat, dog, human)

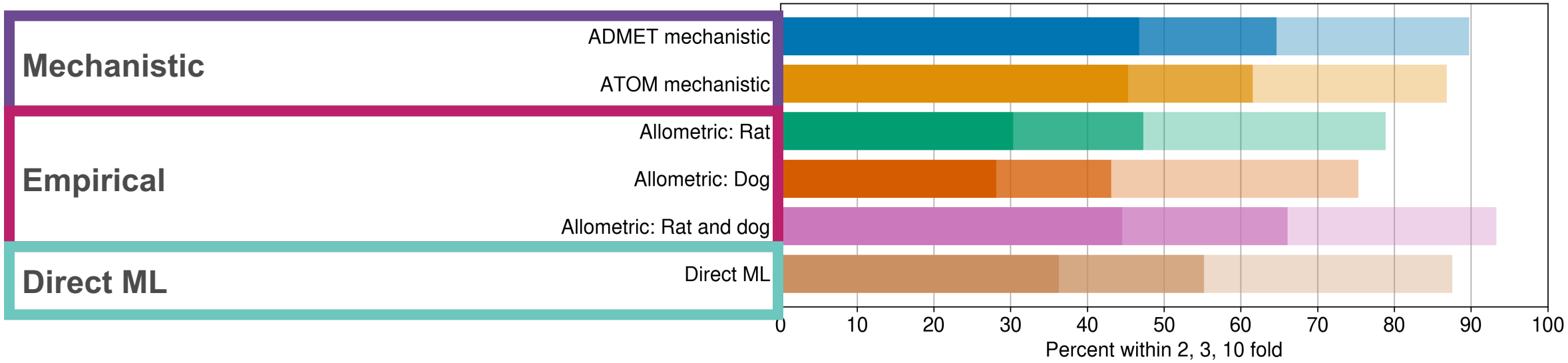
In silico methods to predict human steady state volume of distribution



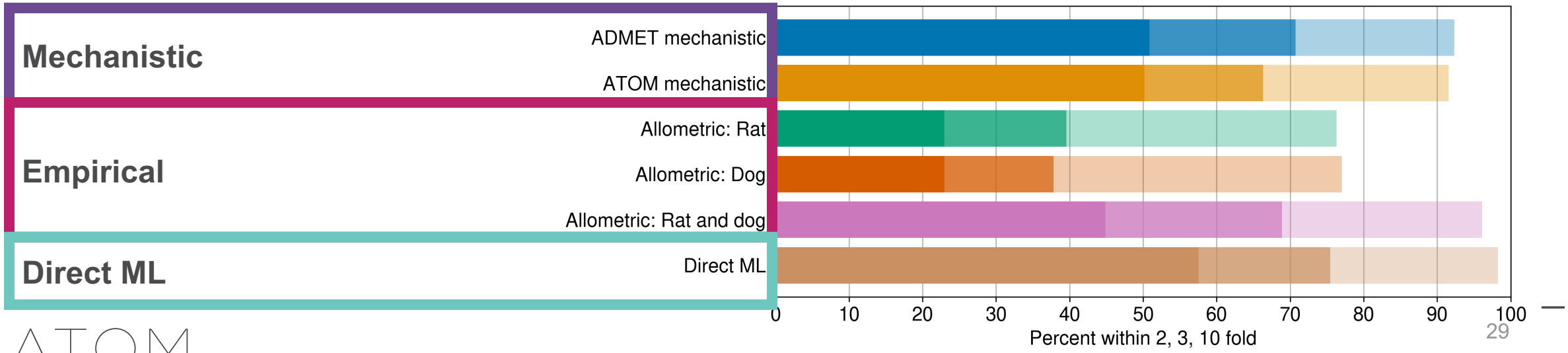
In silico models to predict human steady state volume of distribution on Lombardo Obach (2018) compounds

Type	Method	Description
Mechanistic	**ADMET mechanistic**	• Lukacova mechanistic model to predict tissue partitioning (Kp) • ADMET Predictor models
	ATOM mechanistic	• Lukacova mechanistic model to predict tissue partitioning (Kp) • ATOM models
Empirical	**Rat**	• Allometric scaling of predicted rat VDss
	Dog	• Allometric scaling of predicted dog VDss
	Rat and dog	• Allometric scaling of predicted rat and dog VDss
Direct ML	**Direct ML**	• Training and prediction on direct human VDss

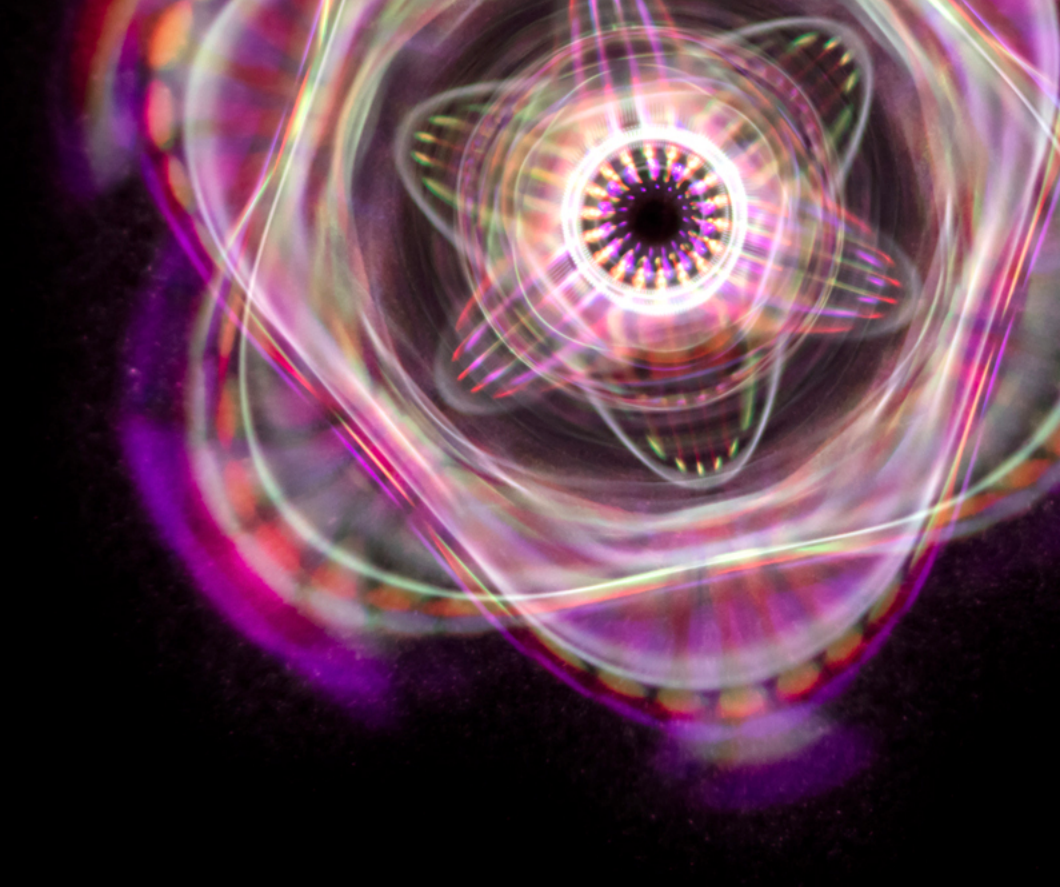
Predictions on ATOM *in silico* set (940)



Predictions on ATOM experimental set (250)
(Also compared to methods using experimental fup, RBP, adipocyte and myocyte Kp)



ATOM



Generative molecular design

ATOM Generative Molecular Design loop Proof-of-Concept

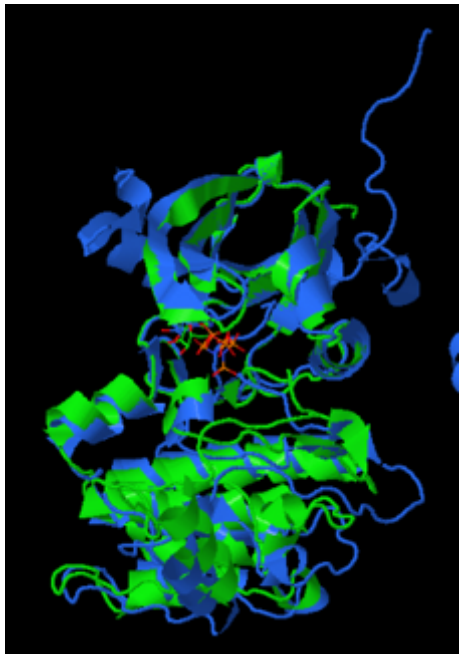
Generative molecular design of AURK B inhibitors

PILOT 1

Starting point:
Early program data

Lead Optimization

End point: Experimental validation



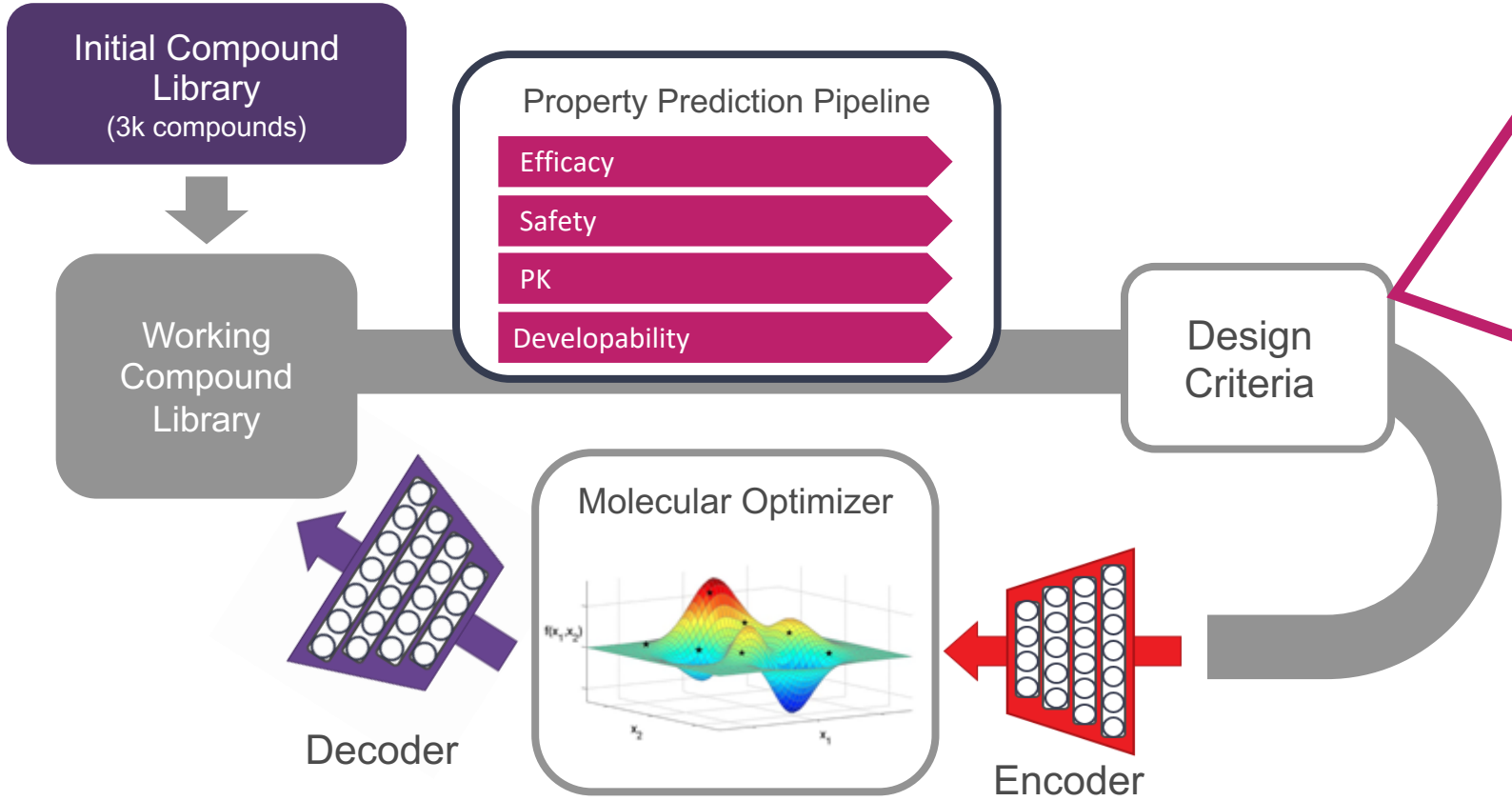
Structure overlay of AURK A and AURK B

Why Aurora Kinase?

- **Cancer relevant**: >30 clinical trials are ongoing or completed for AURKA selective, AURKB selective, and AURKA/B dual inhibitors
- **Data available at ATOM:** Potency data on ~24k compound 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

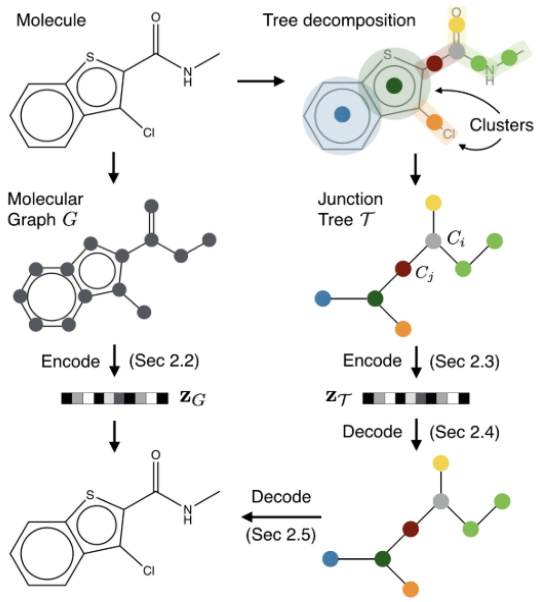
Safety
BSEP (pIC_{50} < 4)
hERG (pIC_{50} < 4)

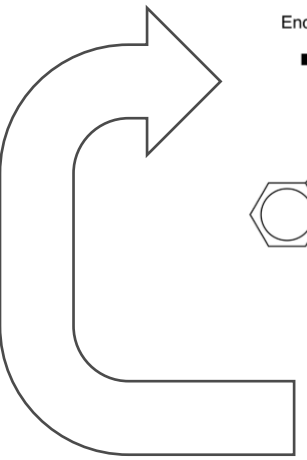
PK
Solubility (>10mμM)
CL_{int} (<3 mL/min/g)

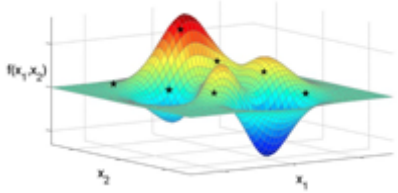
Developability
Solubility (>10μM)
SAS: Synthetic Accessibility Score
QED: Quantitative Estimation of Drug likeness

Our initial molecular design loop is based on multi-parameter optimization in a learned latent space

Junction Tree Variational Autoencoder (JT-VAE)
(Jaakkola, et al – published code)





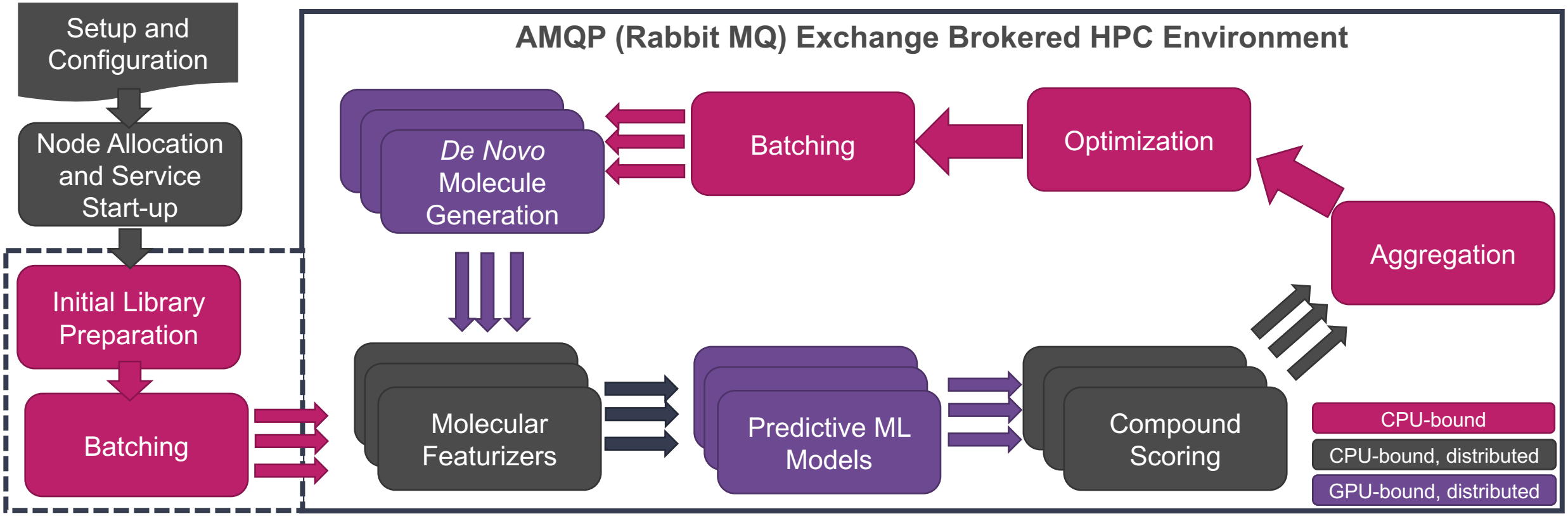


- Molecular modification in physical space lead to a very discontinuous cost surface
- Learned generative models provide a continuous latent space in which small design changes lead to small changes in properties
- Our initial implementations are based on junction tree autoencoders
- Trained on ~24K molecular structures associated with AURK A or B
- For expanding design into new chemical spaces we need to expand the training set to ~1B molecular structures
- Not possible with current learning frameworks and systems – a target for DOE ML learning systems: LBANN and CANDLE

High Performance Compute Facilitates Large Scale Search

Enables Scalable Management of Heterogeneous Compute Tasks

- Facilitated ideation and evaluation of **>3 million** compounds in **24 hour run time**
- Future **scaling by 10x or more** achievable on current, 100 node clusters
- Flexible, object-oriented worker framework allows for future addition of **systems and physics-based modelling**

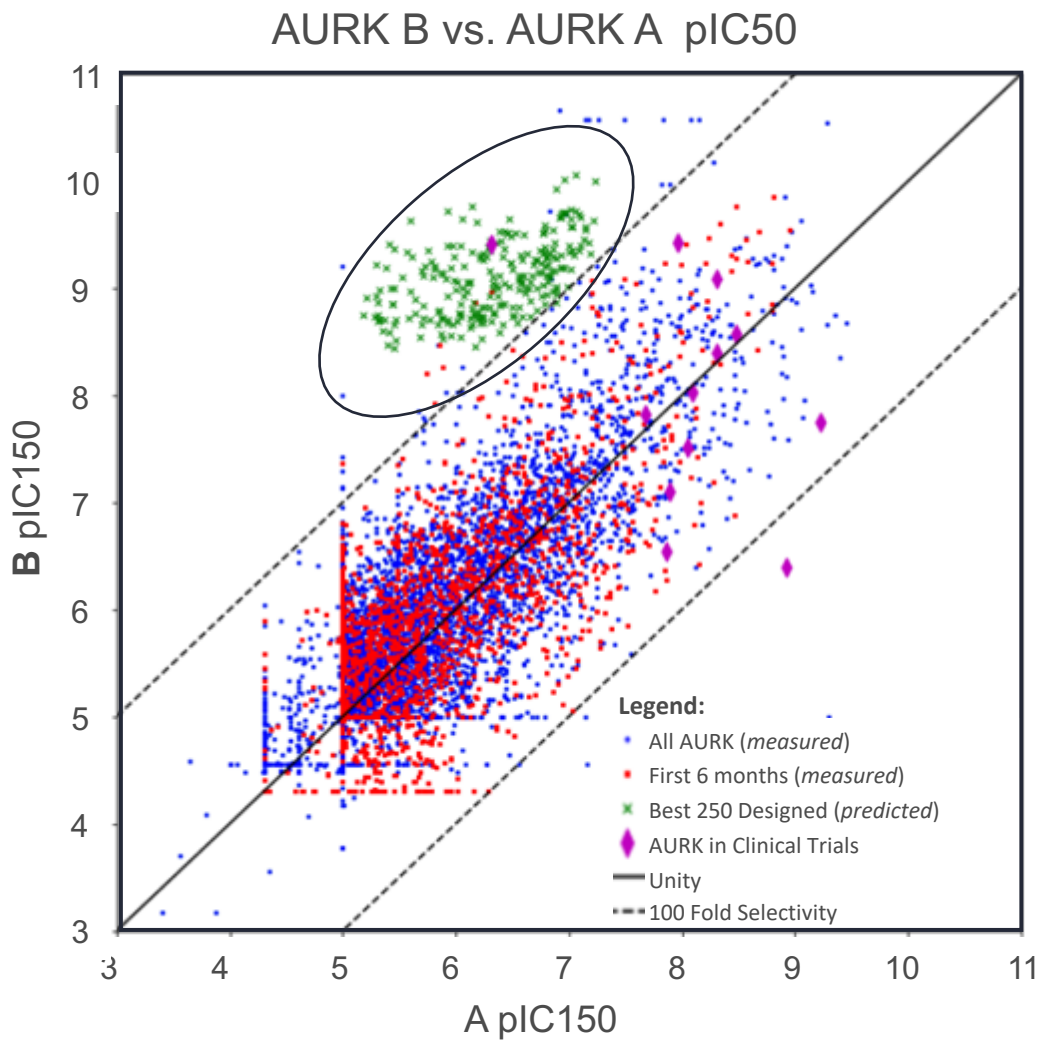


ATOM

Code going through open source process now, requires (for now) SLURM job schedular

~200 Compounds with high potency, selectivity, and other favorable properties

Proof-of-Concept



Other predicted properties for top compounds:

Generated Compound	AURK B pIC50	AURK B/A Selectivity (fold)	hERG pIC50	BSEP pIC50*	hLM Clearance mL/min/g	Solubility ug/mL	SAS
Cmpd 1	9.2	5287	3.3	4.3	3.6	1096	2.5
Cmpd 2	9.3	3233	3.2	4.2	2.5	399	2.4
Cmpd 3	9.6	11512	3.6	4.4	2.2	412	2.6
Cmpd 4	9.6	2449	3.2	4.3	2.5	60	2.3
Cmpd 5	9.7	3068	3.3	4.3	2.0	1155	2.5
Cmpd 6	9.6	5756	3.7	4.5	4.3	232	2.3
Cmpd 7	9.3	3296	3.3	4.4	2.6	33	2.4
Cmpd 8	9.1	1197	3.3	4.2	2.4	268	2.5
Cmpd 9	9.2	7724	3.3	4.3	2.3	733	2.7
Cmpd 10	10.1	2270	3.2	4.5	2.6	139	2.4
Cmpd 11	9.8	9948	3.2	4.8	5.0	93	2.4
Cmpd 12	9.7	3555	3.4	4.2	3.6	739	3.1
Cmpd 13	9.2	12116	4.1	4.1	2.1	1741	2.7
Cmpd 14	9.0	1951	3.2	4.2	5.0	343	2.5
Cmpd 15	9.3	3573	3.4	4.4	5.7	1248	2.7
Cmpd 16	9.2	5334	3.9	4.5	4.9	155	2.5
Cmpd 17	9.5	2277	3.2	4.8	5.3	78	2.2
Cmpd 18	9.2	5439	3.3	4.5	2.2	74	2.6
Cmpd 19	9.9	2372	3.2	4.7	4.4	689	2.5
Cmpd 20	9.3	8071	3.4	4.6	3.8	1332	2.8

criteria met Close to criteria criteria not met

Make Test Results – On Target Pharmacology

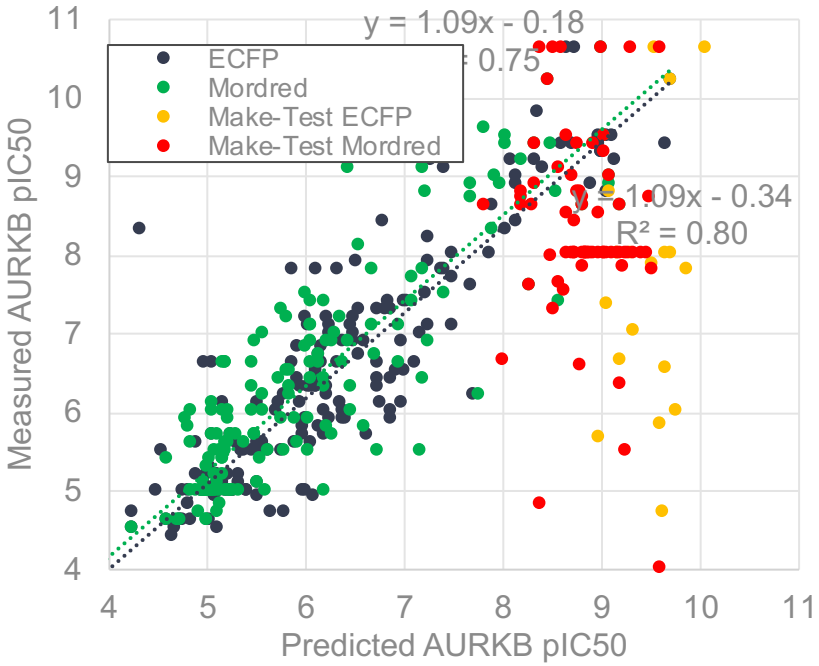
Significant Enrichment of High Quality Compounds!

De Novo Synthesis & Testing Confirms Enrichment of High Potency Compounds

	Count	AURK B Potency		AURK B/A Selectivity	
		Very High (pIC_{50} > 9)	High (pIC_{50} > 8)	Highly Selective (> 1000 fold)	Selective (> 100fold)
Initial Library	3114	18 (0.6%)	75 (2.4%)	0 (0%)	8 (0.3%)
Full ATOM AURK Library	18,582	69 (0.3%)	316 (1.7%)	7 (0.03%)	34 (0.2%)
Generated Compounds	**84**	**16-43 (19-51%)**	**58 (69%)**	**2-35 (2-42%)**	**9-42 (10-50%)**

42 *de novo* compounds successfully synthesized and tested
42 library available highly scored compounds sourced and tested

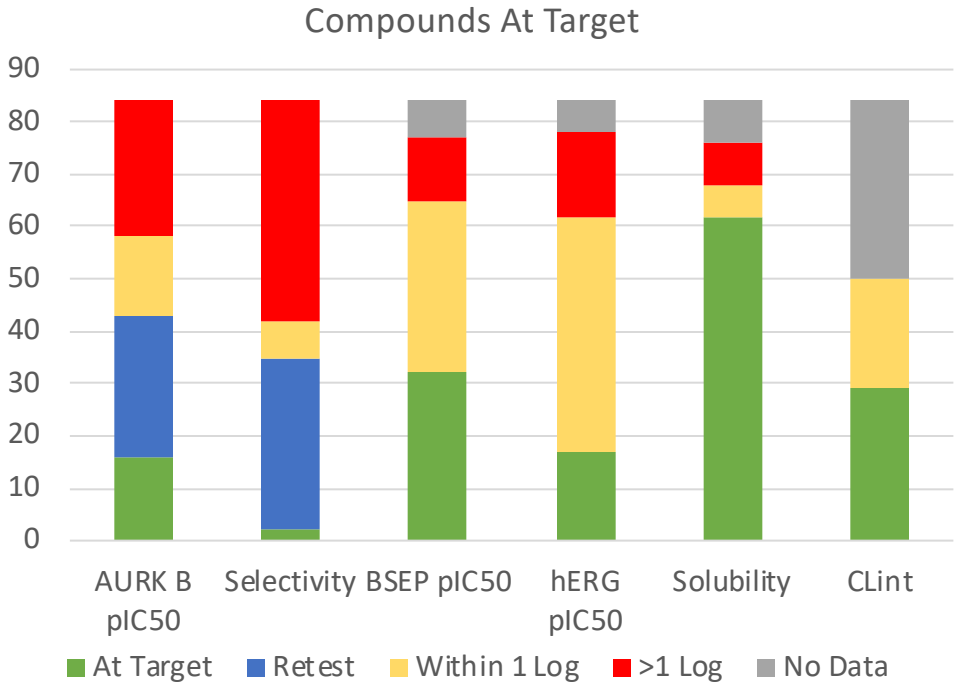
Rediscovery of Compounds in "Hold Out" Library Further Confirms Accuracy of Models for Generated Compounds



Achievement of Secondary Pharmacology In Alignment with Predictions

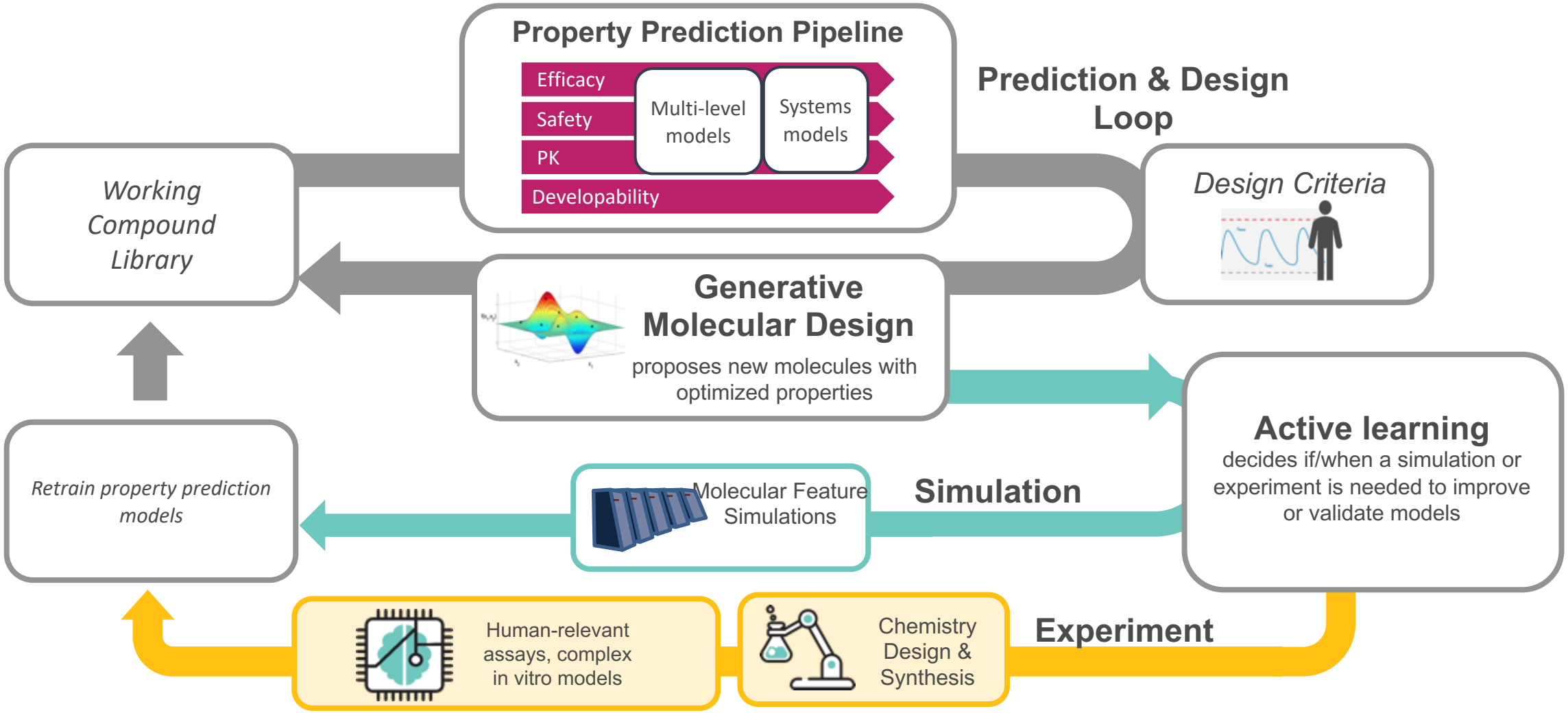
- Generally **>70%** of compounds tested are **as good or better** than the models predicted
- Selectivity and hERG are more difficult

Criteria	Target	Total Tested	In Target Range (Predicted)	Within 1 log of target
AURK B	$pIC_{50} > 9$	84	16-43 (37)	58 (81)
Selectivity	>1000 fold	84	2-35 (15)	9-42 (48)
hERG*	$pIC_{50} < 4.5$	78	17 (63)	62 (82)
BSEP	$pIC_{50} < 4.5$	77	32 (20)	65 (80)
CL_{int}	< 3 mL/min/g	50	29 (43)*	46 (82)*
Solubility	>10 uM	76	62 (79)	68 (82)



ATOM

The ATOM Platform
Active Learning Drug Discovery Framework



Acknowledgements

Computational Tech Team

- Benjamin Madej (SM)
- Kevin McLoughlin (GMD/DM)
- Amanda Paulson (SM)
- Jeff Mast (GMD)
- Derek Jones (GMD)
- Marisa Torres (DM)
- Sergio Wong (MM)
- Dan Kirshner (MM)
- Brian Bennion (MM)
- Stewart He (DM)
- Hiran Ranganathan (DM)
- Ya Ju Fan (DM)
- Soo Kim (DM)

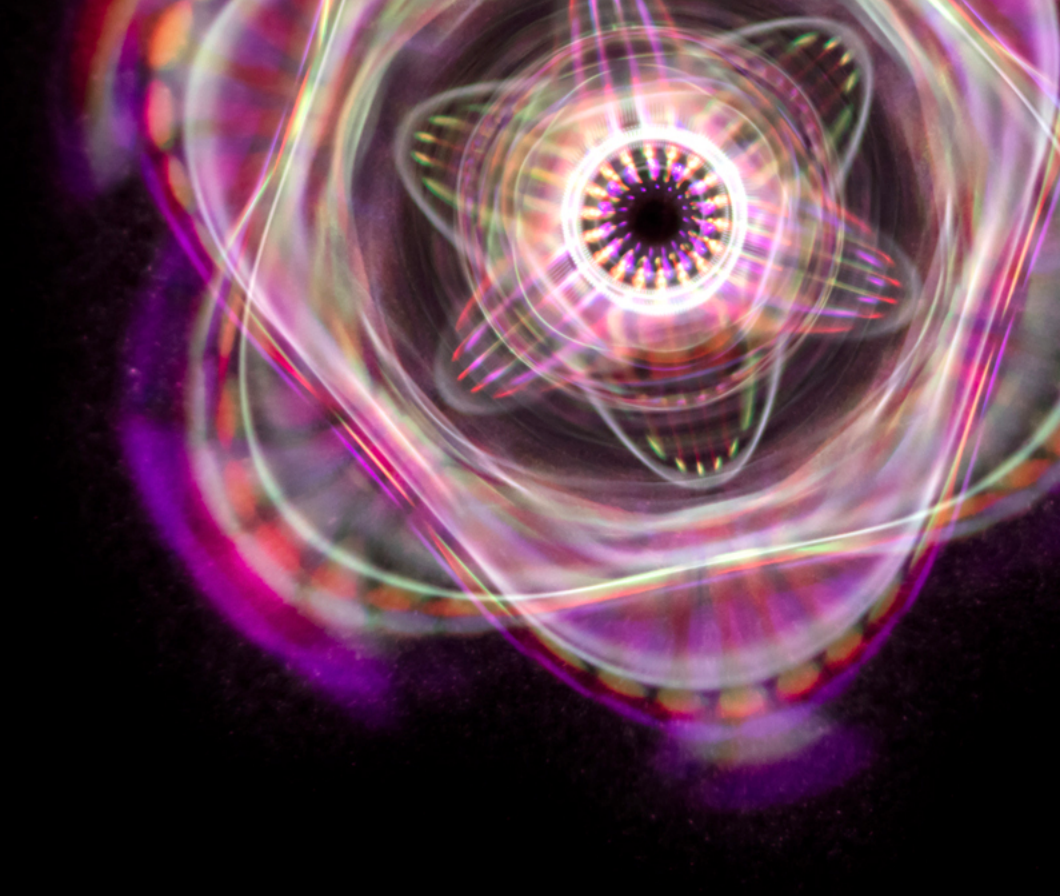
Past Team Members

- Kishore Pasikanti (SM)
- Jason Deng (GMD)
- Amanda Minnich (DM)
- Tom Sweitzer (GMD)
- Juliet McComas (GMD)
- Margaret Tse (SM/DM)
- Michael Gunshenan (DM)
- Andrew Weber (GMD/SM)
- Neha Murad (SM)
- Stacie Calad-Thomson (JRC)
- Tom Rush (JRC)
- Joe Polli (GMD/SM)
- Sabrina Crouch (SM)

ATOM Joint Research Committee (JRC)

- Eric Stahlberg
- Jim Brase
- Michelle Arkin
- Dwight Nissley

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



Questions?
More information at
https://atomscience.org/