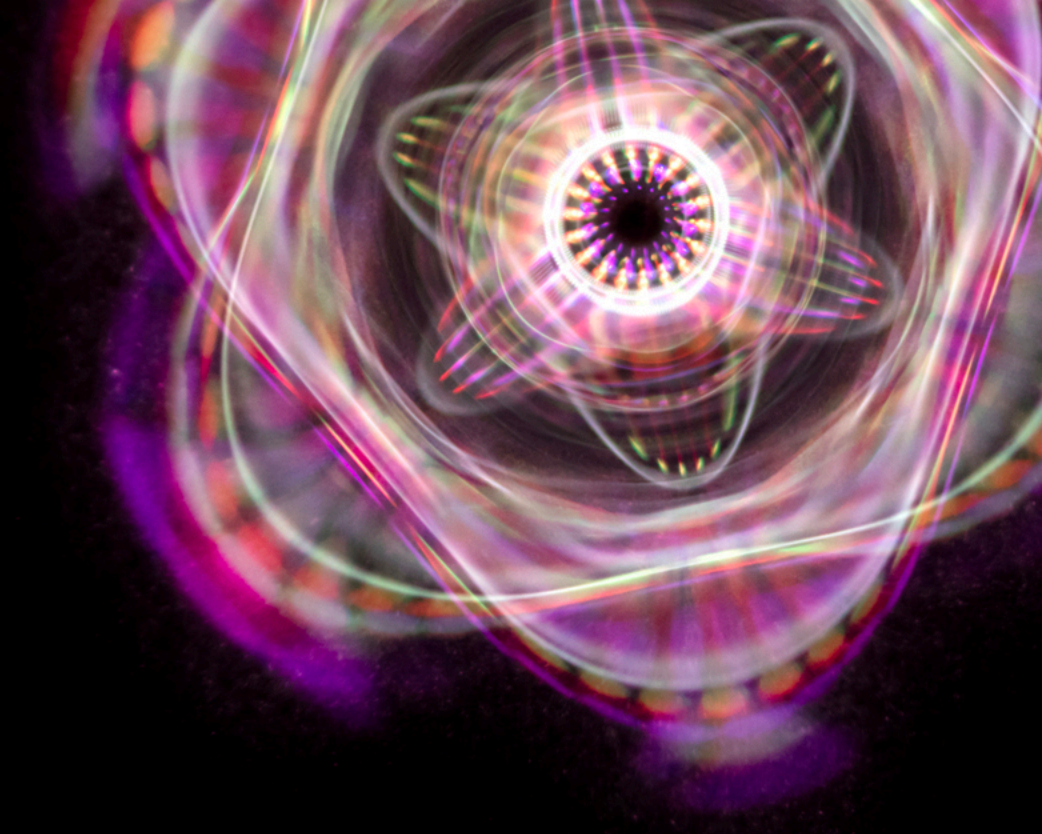


Deep Learning to Accelerate Cancer Drug Discovery

Kevin McLoughlin
Lawrence Livermore National Laboratory
August 8, 2018

Current drug discovery is slow, costly, & high failure

- It takes 16 years and over $1 billion to develop a new drug.
- Nine out of ten compounds fail in clinical trials.

The ATOM Consortium seeks to accelerate this process

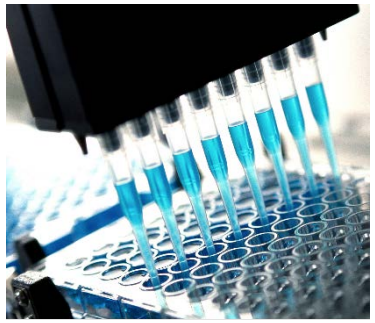
ATOM will accelerate cancer drug discovery by integrating HPC, machine learning and pharmaceutical science

LLNL provides HPC facilities, machine learning scientists, and data infrastructure.

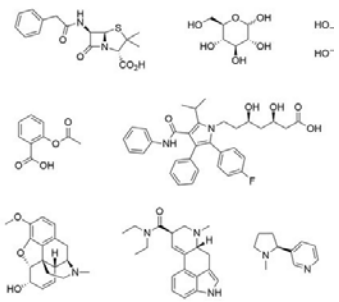
GSK contributes “dark” experimental data from past discovery projects, pharmaceutical expertise.

UCSF and FNLCR provide cancer specialists and laboratory facilities.

Many assays are performed during drug discovery

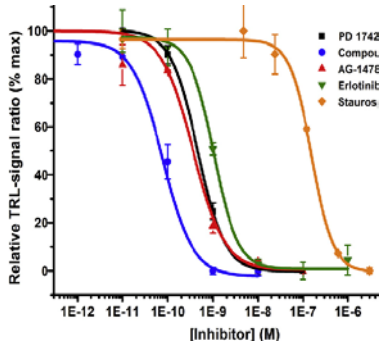


Screen 2M compounds for activity against target

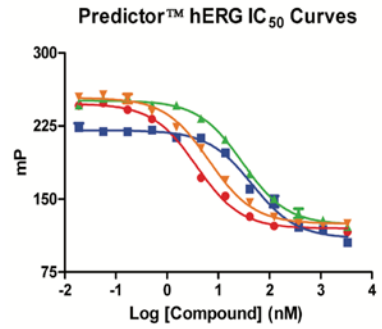


Select 5-10 leads

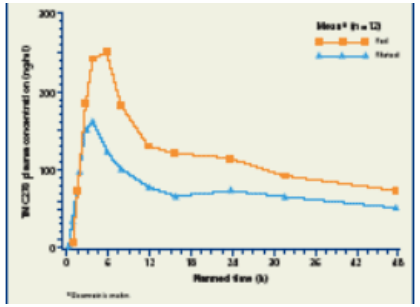
Design & synthesize variants



Test to find most active compounds against target



Assess off-target / toxic effects



Measure pharmacokinetic (ADME) properties

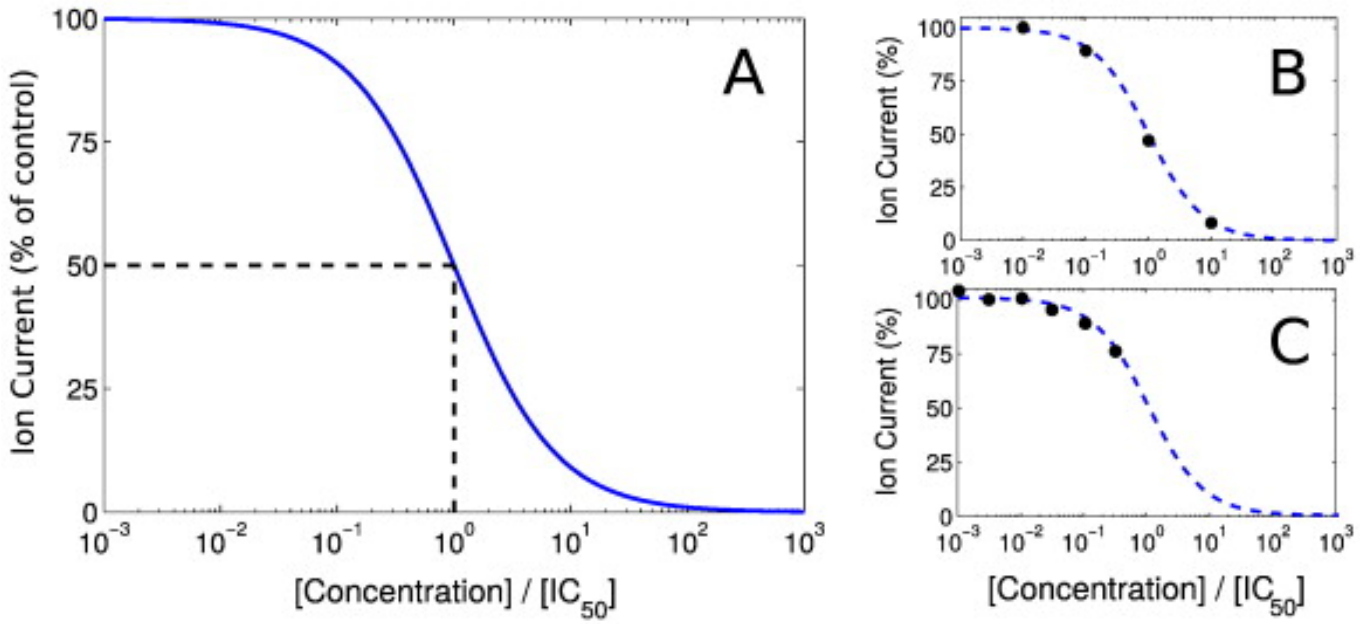
We are building machine learning models to replace many assays

Models use chemical structure features to predict compound properties:

- Efficacy:
  - How does the compound affect the function of a disease-related target?
  - What concentration is needed to achieve a therapeutic effect?
- Safety:
  - Does the compound interact with off-target proteins that cause adverse effects?
  - If so, at what concentration does it cause these effects?
  - What is the therapeutic window?
- Pharmacokinetics:
  - How well is the compound absorbed into the body?
  - What dose is needed to achieve the desired concentration?
  - How quickly is the compound metabolized and eliminated?

Many of our models predict bioactivity assays

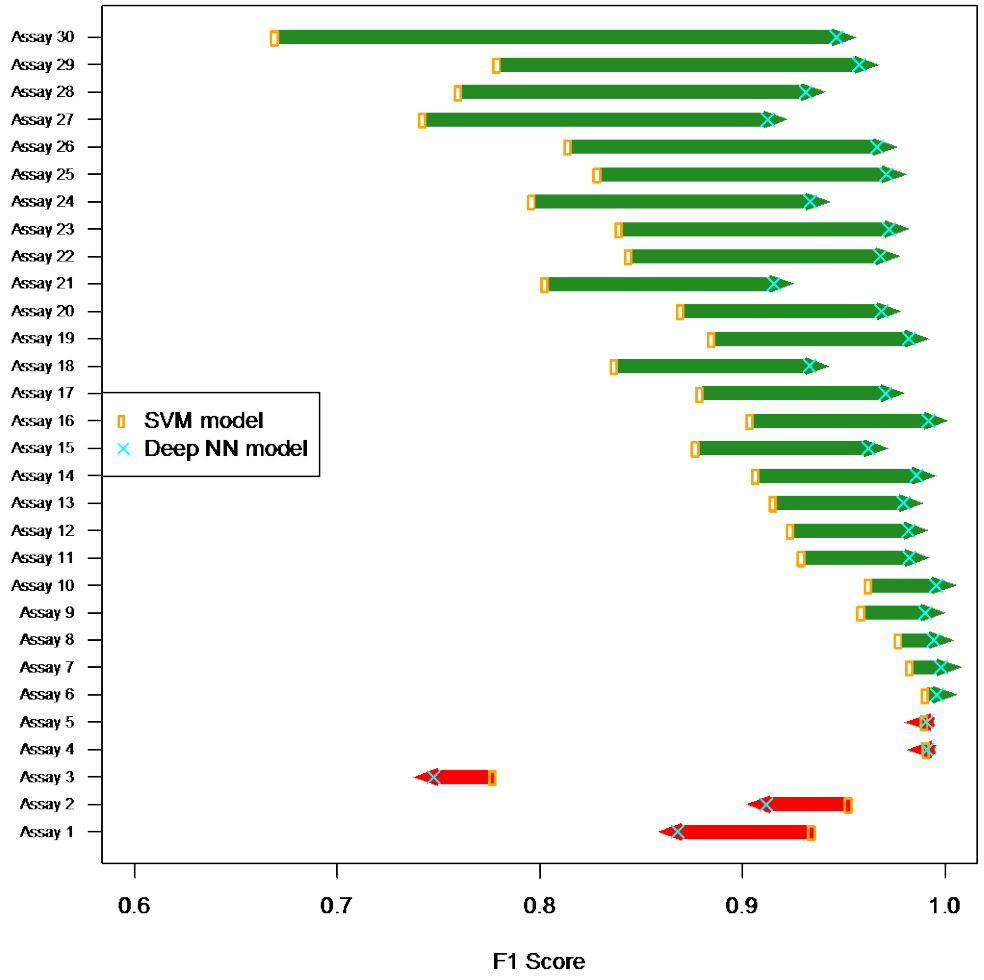
- Bioactivity assays measure %inhibition or %activation at one or more concentrations
- Compute IC_{50} or EC_{50} by fitting logistic curve to activity at 5-20 concentrations
- Reported values often censored, if 50% conc outside measured range.
- Thus most published models are classifiers: is bioactivity above some threshold?.



Geoff Williams & Gary Mirams, DOI: 10.1016/j.vascn.2015.05.002

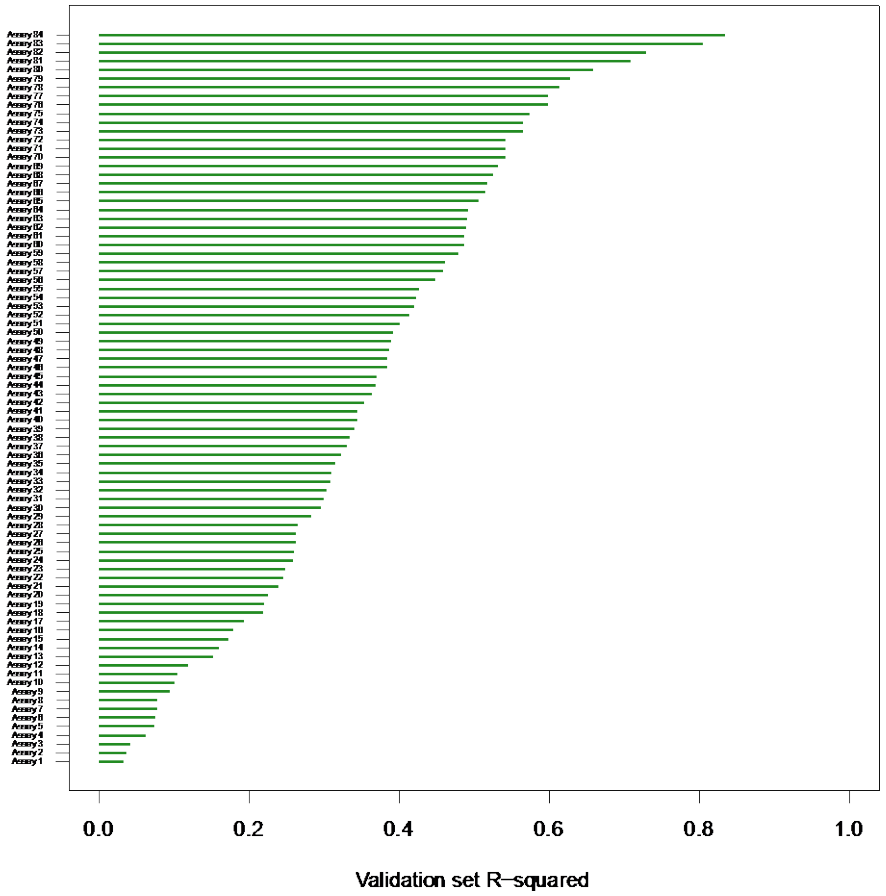
Classification model performance

- We trained neural network models to predict categorical results for 30 safety assays.
- Used DeepChem package from Stanford (https://deepchem.io) with graph convolution features.
- Classification performance of NN models was almost always better than SVM models used internally at GSK.



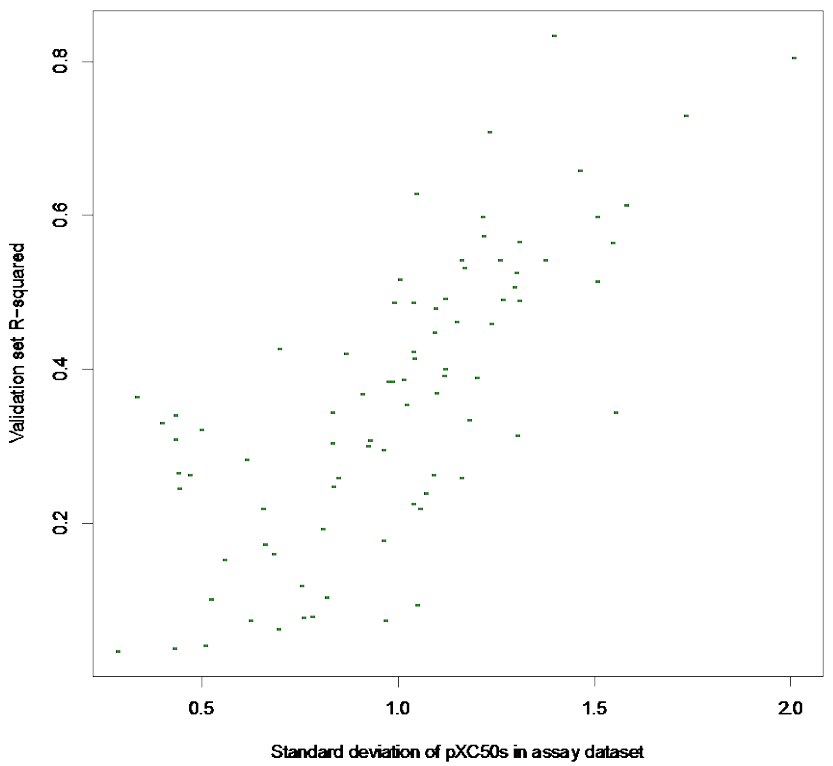
Regression model performance

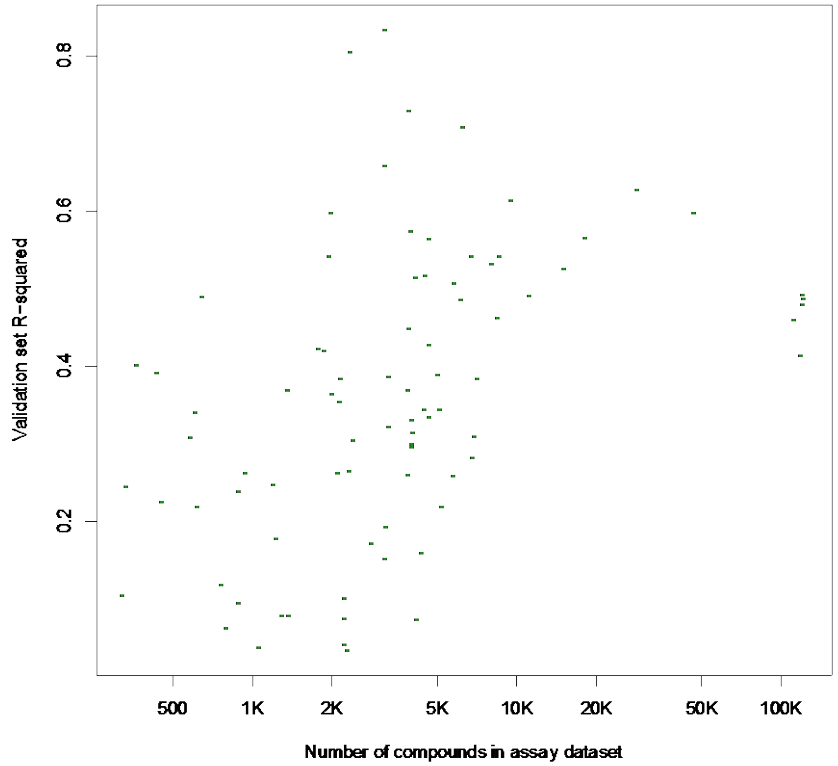
- We built DeepChem graph convolution regression models to predict results of 84 liability assays.
- Data was split by scaffolds into training, validation & test sets.
- Tuned learning rate and number of epochs for best performance on each validation set.



Factors affecting regression model performance

- Model performance is strongly correlated with the spread of activity values in the dataset.
- Dataset size was also important.
- To make better models, we need data from a more diverse set of compounds.





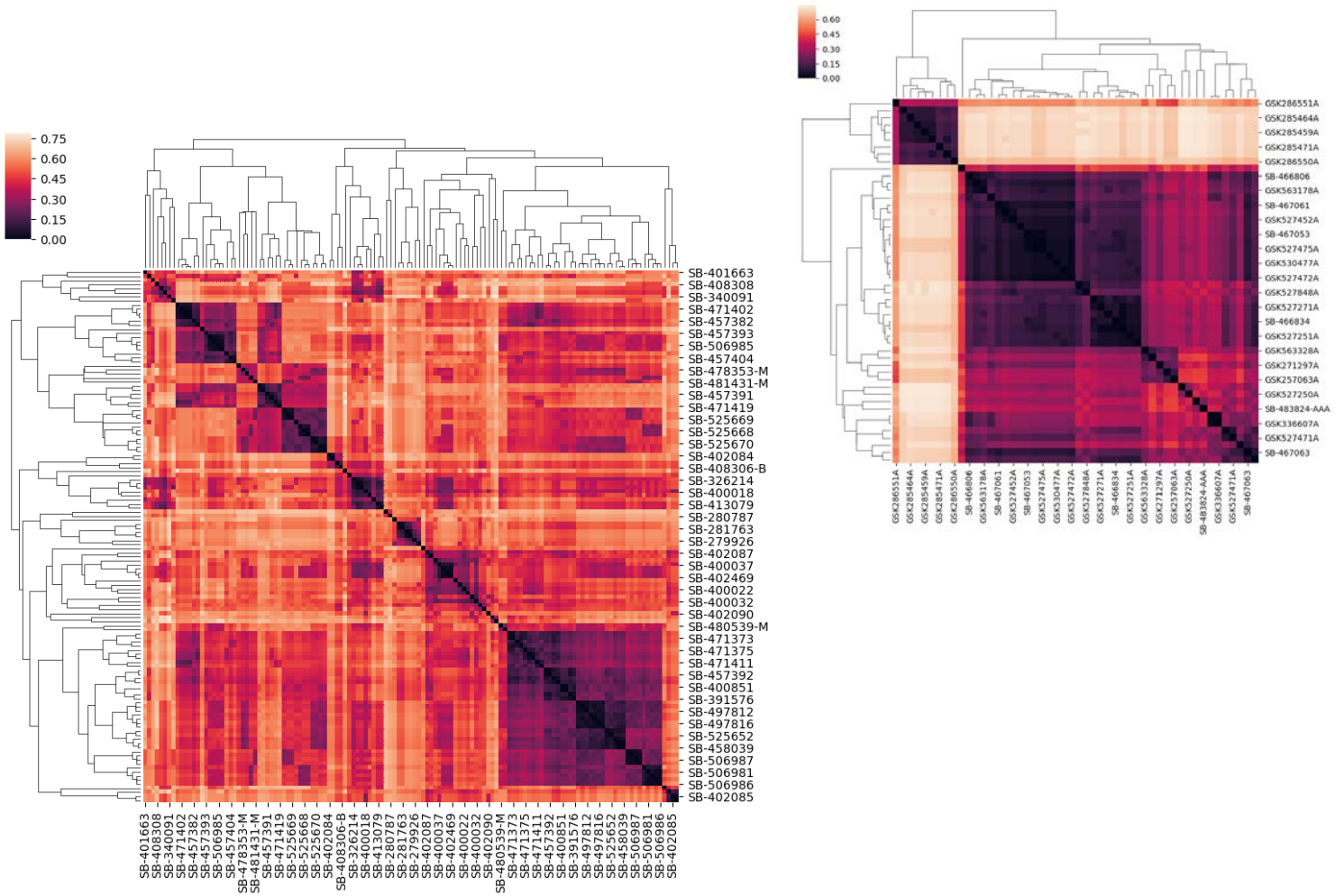
Structural diversity and functional diversity

Two kinds of diversity are required so that machine learning models can generalize to novel compounds:

- Structural diversity: The variety of chemical structures represented in a compound set.
- Functional diversity: The range of bioactivities measured for the set of compounds.

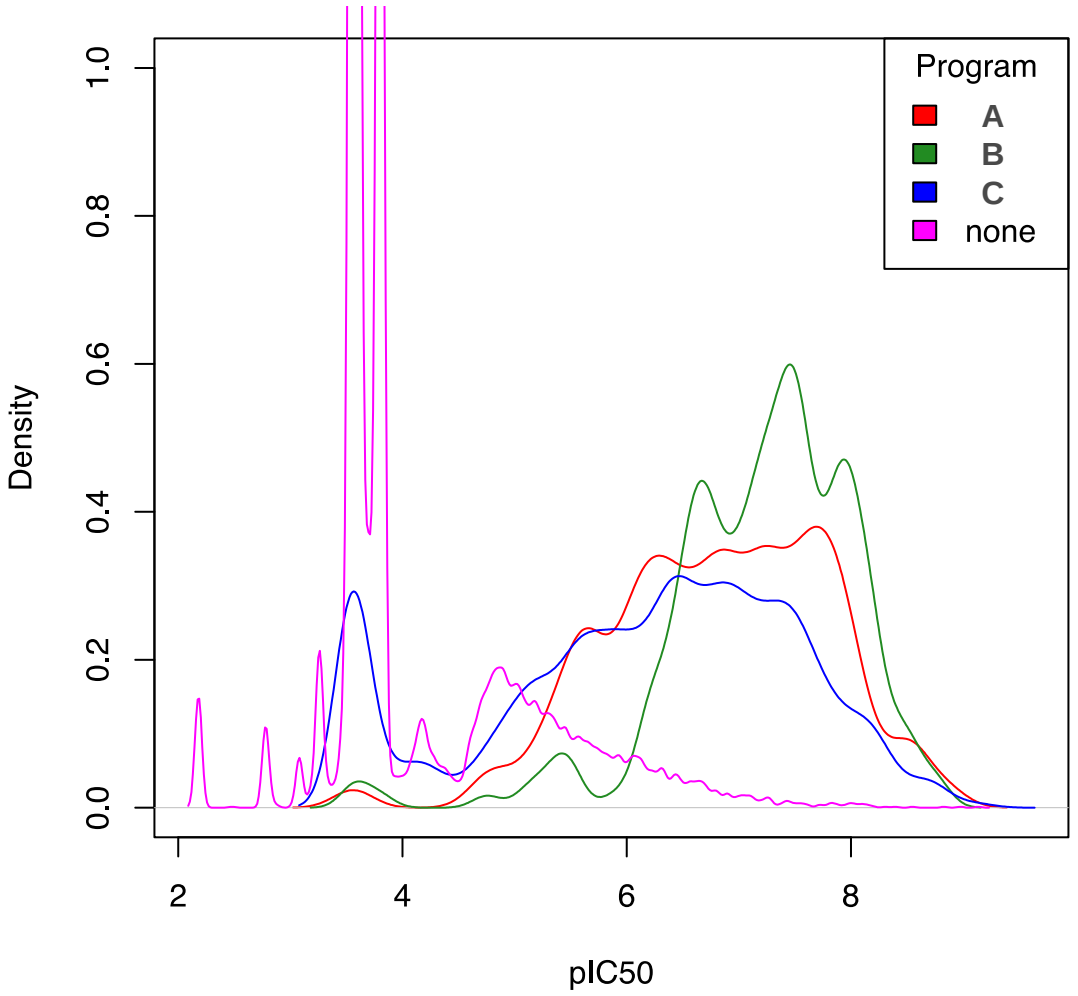
Many bioassay compound sets lack structural diversity

- Assay datasets include many GSK compounds synthesized for terminated drug discovery programs.
- Many of these compounds are derived from small sets of lead compounds => less structural diversity

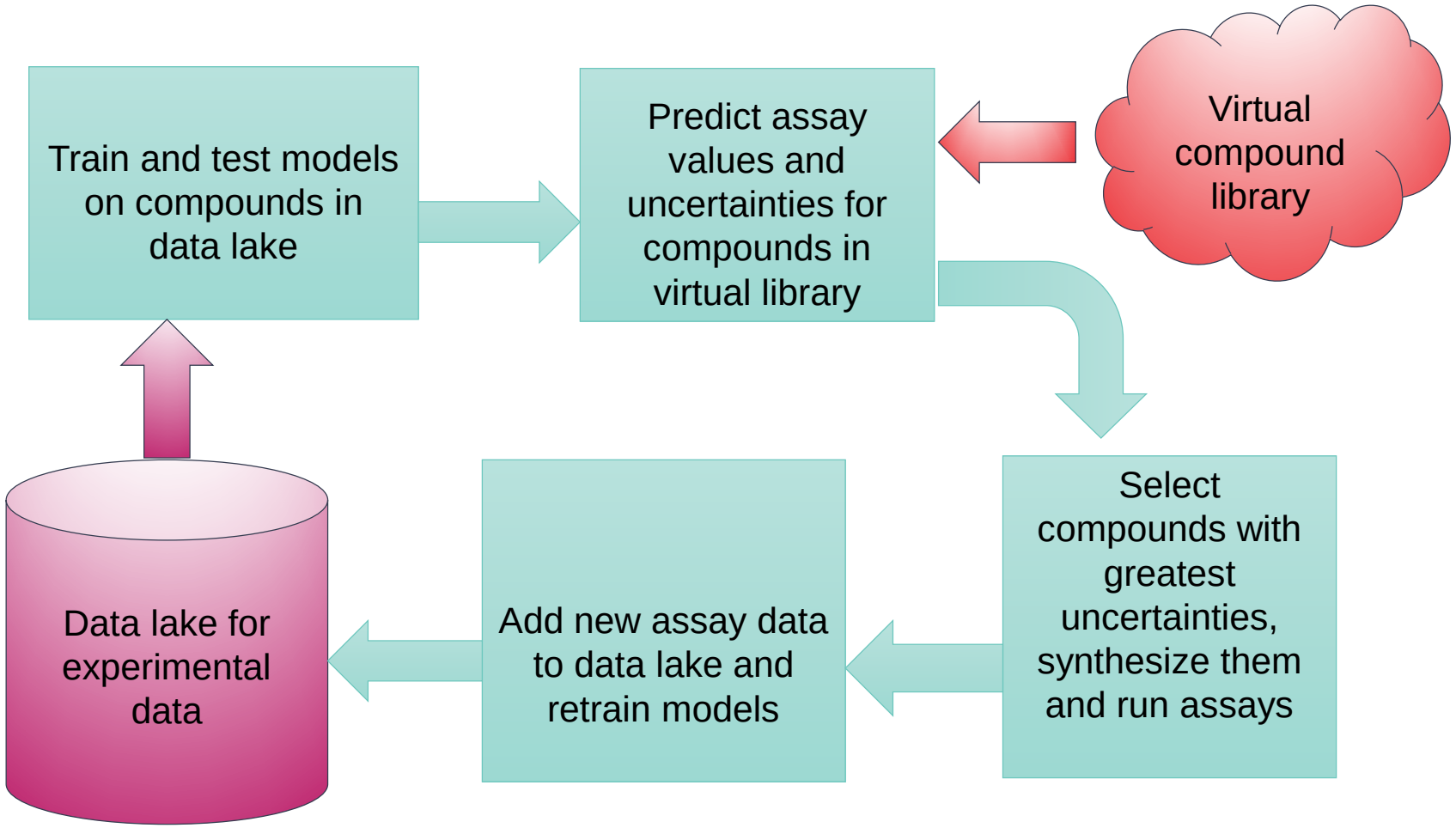


Structural diversity affects functional diversity and model performance

- A target inhibition assay was run against four sets of compounds. Three sets were lead optimization series for specific programs; the other was a diverse set from screening panels.
- Models trained on the diverse set predicted IC50's for the optimized sets reasonably well; but not vice versa.

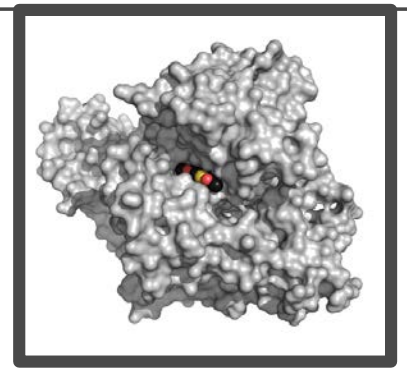


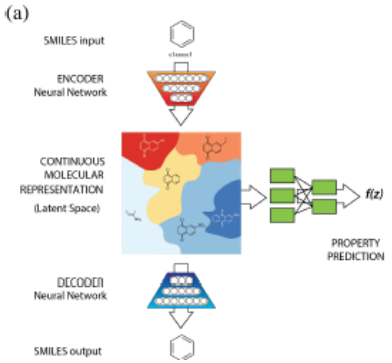
Expanding compound diversity through active learning



More work in progress

- Physics-based features for bioactivity modeling
- Uncertainty quantification for active learning
- Generative models to create optimized chemical structures
- Integration of ML predictions with physiologic models





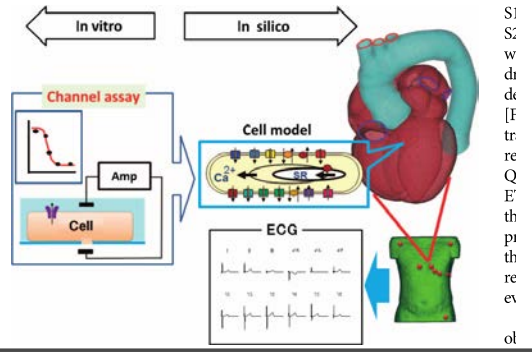


Fig. 1. Diagram of the assay system. Dose-inhibition curves of drugs

Conclusions

- ATOM's active machine learning approach will make drug discovery faster and more cost-effective.
- Diversity in compound datasets is the most important factor for building accurate models.
- Close integration of our machine learning team with GSK's drug development experts and UCSF's cancer research faculty will facilitate success.
- The models and tools we develop will have broad impact for all drug development and will ultimately save lives.

Acknowledgements

- LLNL:
  - Amanda Minnich
  - Jonathan Allen
  - Marisa Torres
  - Sergio Wong
  - Xiaohua Zhang
  - Brian Bennion
- DeepChem:
  - Bharath Ramsundar

GSK:
- Pragathi Kotha-Venkata
- Claire Jeong
- Sabrinia Crouch
- Tom Sweitzer
- Joe Polli
- Tom Rush
- Margaret Tse

FNLCR:
- Deb Hope