Good Morning/Evening Everyone,

**Intro** - My name is Titli Sarkar. I am working as a Data Scientist at Frederick National Laboratory for Cancer Research and today I am going to present the overview of the ATOM drug discovery framework.

**Impact** – Can you put a statement here? Talking points are:

* Drug discovery process is lengthy, expensive, and low success rates.
* With ATOM technology, we can accelerate the process faster and cheaper. That can benefit patients to get new treatments faster.
* Since ATOM technology is open-source, the data and models we build are publicly available and can share with the research community for their use.

**ATOM overview -** ATOM stands for Accelerating Therapeutics for Opportunities in Medicine (ATOM). This is public-private partnership with the mission of transforming drug and therapeutic discovery by accelerating the development of more effective therapies for patients~~, which can be useful for Cancer Drug Discovery.~~

It reduces drug-discovery ~~pipeline~~ timeline (by this I mean we hope to speed up the research process, rather than reduce a pipeline :-) through Artificial Intelligence and Machine Learning

This involves High Performance Computing using resources from Lawrence Livermore National Lab, Frederick National Laboratory (FRCE) and National Cancer Institute such as Biowulf, it curates and prepare diverse molecule data for ML, which can be used in AI and ML based emerging biotech capabilities. (Joe wants to point out that AMPL is the part of ATOM that curates/prepares diverse molecule data and it also does the ML/AI/Deep Learning steps too)

First, Molecular data from both publicly and privately available resource are collected and curated and featurized to create model ready datasets. Next, ML models are developed to predict useful properties of in silico Drug Discovery through AMPL pipeline. ~~Next, a Generative Molecular loop is created which can propose new molecules with optimized properties.~~ *~~This optimization loop involves domain knowledge as design criteria. Also, the GMD loop improves by Active Learning which analyzes the prediction uncertainty, and specify suitable experiments. The new predicted molecules then go through Chemistry synthesis and assays and only the best property prediction models are retained after execution the loops several times until satisfactory outcomes are generated.~~*

**AMPL -** The very first initiative of ATOM was AMPL or ATOM Modeling PipeLine (AMPL) which is an open-source, modular, extensible model pipeline for building and sharing ML models which can predict key safety and pharmacokinetic-relevant parameters. First step was data curation, then a set of featurizer were applied to create machine learning ready inputs. Next, a set of ML models were trained, and parameters were tuned and optimized models were read to generate prediction of properties (which ones?) on new test data. AMPL has been benchmarked on a large collection of pharmaceutical datasets covering a wide range of parameters.

**GMD** – The base AMPL model was used to create input latent vectors (?) for a Generative Molecular Design for Drug Development which can shorten the timeline for Cancer Drug Discovery. GMD is a High-performance platform for parallel optimization of important drug property like Efficacy, Safety and Pharmacokinetics. GMD uses an autoencoder for optimized molecule prediction. The initial molecular population is passed through an Encoder which represents inputs in an encoded latent space, then they are decoded to the physical space representation. This goes through AMPL property prediction loop and once the autoencoder is trained, the new molecules are predicted after n number of optimization loops.

AMPL is not used to create input latent vectors or train the autoencoder (as far as I understand it anyway). The autoencoder training is a completely different component.

GMD has 3 main components:

1. A trained Virtual Autoencoder (VAE)

a) An encoder for transforming molecules -> latent vectors (really just a vector (of numbers) for each molecule)

b) A decoder for translating latent vectors -> molecules (take a vector (of numbers) and create a molecule).

Clearly the best possible VAE is one that can take a very large set of molecules, encode them to latent vectors, and decode them back to the exact starting molecules. However, this is very difficult to achieve. There are several successful methods for this, and we use the Junction Tree Virtual Autoencoder (JTVAE) one.

2. Once you have the latent vectors for a set of molecules, a Model needs to be trained on them for the prediction of a property, or several properties. Then, we try to optimize molecules (in the latent space) to give better properties. Once the best latent vectors have been found, they must be decoded using the VAE to generate the representative molecules. This is also challenging as you’d like to produce ‘real’ molecules and that’s not as simple as it sounds. This is the crux of the GMD method, the molecule optimization step.

3. Once you have a set of new (optimized) molecules you’d like to predict, or compute, their properties if you can. For the problem of finding the best binding molecules, LLNL has a very high-throughput docking process to analyze a new set of ‘optimized’ molecules. The results of these docking calculations on a new ‘optimized’ set of molecules are used as their property values. These molecules and their properties are then included in another iteration of molecule optimization (in the latent space) to produce another round of new (optimized) molecules.

**GMD Example** -

The pilot project of ATOM to demonstrate the value GMD process with a realistic lead optimization problem was Neurocrine project. First generation H1-antihistamines have undesirable side effects mainly due to off-target activities against muscarinic receptors. Some of the second generation H1 inhibitors had to be taken off the market because of interference with the hERG channel, leading to cardiac side effects. For this case study, we have used the GMD active learning framework to generate and evaluate selective H1 antagonists, alternating computational design with experimental testing and model retraining. The GMD loop started with ~25K compounds and ran for 250 generations and generated ~ 7K novel components with optimized properties. The hybrid model used in GMD loop is trained to predict pKi first, then infer binding affinity for the given pKi . This example plot shows compounds with optimized efficacy with H1 histamine receptor pKi (pharmacokinetics) > 9 and M2 muscarinic receptor pKi < 6. The pilot project was developed with histamin receptor, but we are broadening it to cancer data.

**Capability Transfer** – ATOM offers capability transfer through publicly available data and models through github. It also developed a public-facing web interface for named MoDaC for model and data sharing. Data is managed via upload/download with configurable access level and data transfer supports AWS S3, Globus and Local.

We hope ATOM will create a platform which can act as a bridge between broader cancer research area.

========================================================================================================

Information which can be used:

* generation of a pool of candidates for computational characterization before experimental synthesis
* Novel method to boost pKi or pIC50 model performance using single-concentration data The hybrid model is implemented using PyTorch.  
    
     
  GMD Loop - Initial molecule set ~ 25K compounds

1. GMD loop with AD index constraints ran for 250 generations on initial 25K compound set
2. Produced and evaluated ~1.75M compounds over about 60 hours of elapsed time.
3. Toward end of run, ~ 7K/25K compounds evaluated each generation were novel.
4. Thus, GMD loop continues to explore new parts of chemical space even after many generations.

**Efficacy**

H1 histamine receptor pKi > 9

M2 muscarinic receptor pKi < 6

**Safety**

hERG (pIC50 < 5) 10 mmolr = 5 (which is good)

**PK**

ogP (-0.4 - 5.6) Within the range = good

**Developability**

SAS: Synthetic Accessibility Score < 3

Molecular features: # RotBond < 10, # HBondDonor < 5, # HBondAcceptor < 10, MW (180 ~ 500), TPSA < 140.

We selected the 2000 top scoring compounds, clustered them by structure and sent the best representative from each cluster to Neurocrine for evaluation.

Also found 95 close matches in Enamine REAL catalog that scored well.

Neurocrine chose 55 for synthesis and is testing them along with the Enamine compounds; expecting results soon.

We’ll use experimental results to retune models and do at least one more GMD round.  
  
  
Hybrid model is trained to predict pKi first, then infer % binding given pKi  
  
Training aims to minimize loss function:

* + pKi data: (pred\_pKi - real\_pKi)2 = normal L2 loss
  + % binding data: (1 - Fpred) - (1 - Freal) log(1-Fpred) = Poisson loss

Fpred and Freal are the predicted and measured fraction of receptors bound to drug

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.

**Mission:** provide integrated drug design and optimization technologies to shorten the drug discovery timeline and

reduce the cost of drugs.

**Objectives:** A pre-competitive drug discovery platform integrating high performance computing, chemical biology,

and related technologies to transform the current slow, high cost and high failure process into a rapid, integrated

design/make/test/refine cycle with access to all.

**Goals:**

• Drug Development: develop drugs for the public good using alternative business relationships

• Research: a research and development environment focused on developing computer-driven approaches

for pre-competitive, open-source science to modernize the technology and business of drug discovery

• Training: interdisciplinary training of scientists, providing a foundation in high-performance computing and

drug discovery technologies and knowledge

• Symposia and Workshops: host and participate in scientific symposia and workshops to demonstrate

advances

• Advocacy: fund research efforts related to its purposes

ARA Capabilities: High-performance computing, deep biological systems knowledge, drug discovery and

development experience, and releasing workable models in the public domain.