Hi everyone,

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

**Impact–**

* ATOM stands for Accelerating Therapeutics for Opportunities in Medicine.
* Usually, the Drug discovery process is **lengthy**, **expensive**, and it has **low success rates**.
* ATOM **shortens** the drug discovery **timeline** by using **AI** and **ML**.
* This can be useful for Cancer Drug Discovery as **it aims to deliver** **therapies** to **patients** **faster** and **cheaper** **than existing market**.

The top-left box demonstrates general ATOM overview. I will explain it in modules.

* In general,
* They very first step is **data curation** - Molecular data are collected and curated to create a machine learning model ready database.

**AMPL –**

* The very first ML initiative of ATOM was AMPL. Next box explains it. AMPL stands for ATOM Modeling PipeLine which is a **modular** and **extendable software pipeline** for **building** and **sharing** **ML models** formolecular **property prediction**.
* We get raw data, extract features from them and train ML models to learn the features using known data. The ML models can now predict properties/features of new unseen data. Isn’t that fabulous?
* Basically, AMPL can **identify** the molecules which are the **strongest candidate** for **next experimental investigation step.**

**GMD –** Next, we have created a Generative Molecular Design loop for **predicting new molecules** with **optimized property**.

The **backbone** of GMD is a ML model named **Autoencoder** (VAE) which have two parts: an encoder and a decoder.

a) The **encoder** **transforms** **molecules** to **latent vectors**, which is just a **bunch of numbers** for each molecule.

b) The **decoder** **translates** the **latent vectors** back to **molecules**.

Now, translating back to **exact** same molecules are **incredibly complex**. We trained the autoencoder which learns features from the known molecules. Then, we **generate new molecules** by **perturbing latent vectors and then decoding** them using the autoencoder.

2. Once we generate a set of **new** molecules, the **AMPL** module is used for **property prediction** and **molecules** with **optimized properties** are **added back to the molecule library** and it goes for second iteration for producing a new set of molecules with **better properties**. This multi-parameter optimization loop continues until there is no scope of optimization anymore.

**GMD Example –**

Neurocrine project was the pilot project of ATOM which demonstrated the capability of AI through GMD. We got antihistamine data from Neurocrine. **H1-antihistamines** have **side effects** like cardiac arrest, mainly due to **off-target activities against muscarinic receptors.**

Our goal was to generate new molecules with **maximized** on target **effects** and **minimized** off target effects.

**ML** makes co-**optimization of multiple parameters** **together** **cheaper**, and it **does not create huge number of molecules**.

The GMD loop started with a large number of compounds, ran for a few hundreds of generations and **retained** a small number of components with optimized properties. This example shows novel compounds with **on-target** or H1 **histamine receptor pharmacokinetics > 9** and **the same** **with <6** for **off-target** M2 **muscarinic receptors** which was **desired**.

The pilot project was developed with histamine receptor, but we are looking forward to broadening it to cancer data.

**Yellow Loop –** The yellow loop shows the experimental part of our workflow.

Once we have **a set of new (optimized) molecules, a very high throughput docking process** **analyzes** them. They go through lab experiments and the outputs are added in the molecule compound library. They again go through GMD and this loop continues **until** there **are no further scope** of optimization.

Finally, we get **high-quality** **optimized molecules** as **drug-binding candidates** with **a** **shortened timeframe**.

**Capability Transfer –** ATOM technology is **open-source**. We share our **data** and **models** through **github** and a **public-facing web interface** named **MoDaC**. Data **transfer** **supports** **AWS S3**, **Globus** and **Local**. We want to thank LL and Frederick Nattional Lab, NCI/NIH for providing High Performance Computing resources.

To conclude, We hope ATOM will **create** a **platform** which can **act as a bridge** between **broader** cancer research **community**.