 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**.
* This can be useful for Cancer Drug Discovery as **it aims to deliver** **therapies** to **patients** **faster** and **cheaper**.

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

* In general, ATOM **uses** a set of **ML models** to **design more-effective new drugs** for patients in **shorter-time**.
* The very first step is **data curation** - Molecular data are **collected** and **curated** to create a machine learning model ready database.

**AMPL –**

The first ML initiative of ATOM was AMPL. Next box explains it. AMPL stands for ATOM Modeling PipeLine. It **trains** a set of ML models for **molecular property prediction**.

**GMD –** Second module is GMD which stands for Generative Molecular Design loop, which can **predicting new molecules** with **optimized property**. The beauty of GMD is it can parallelly optimize multiple properties for new molecules.

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

a)  The **encoder** **transforms** **molecules** to **latent vectors** which is a bunch **of numbers**,

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

Now, translating back to **exact** same ‘Original’ molecule is **incredibly complex**. For this purpose, we trained the autoencoder to learns features from the known molecule compound libraries. 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**. This multi-parameter optimization loop continues until there is no scope of optimization anymore.  Human design criteria play a crucial role here.

**GMD Example –**

Neurocrine was the pilot project of ATOM. We got antihistamine data from Neurocrine. **H1-antihistamines** have undesirable **side effects** due to **off-target activities against muscarinic receptors.**

GMD uses ML to **co-optimize** multiple **properties** such as **maximize** **on-target** and **minimize off-target activities** and instead of creating huge number of molecules, **proposes** a **small** number of **new targeted molecules** with **optimized properties**.

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

**Yellow Loop –** Nexu module is the yellow loop which shows the experimental part of our workflow. The **new (optimized) molecules** from GMD can get **analyzed** and if needed, go through **simulations** and **lab experiments** and the outputs are added in the molecule compound library. They again go through GMD and this loop **continues to be optimized.**

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

**Capability Transfer –** ATOM technology is **open-source**. Our **data** and **models** are available through **github** and a **public-facing web interface** named **MoDaC.** We want to thank LL , Frederick National Lab and NCI/NIH for supporting us with High Performance Computing resources.

To conclude, we hope ATOM will **create** a **platform** which can **act as a bridge** between broader drug discovery **community** (by transforming a **lengthy**, **costly,** and **high-failure** process to a **rapid**, **fast** and **cost-effective** process with **high-success rates**) for **serving patients better**.