# **COLLABORATION PLAN FOR SUMMER 2022**

## **Research Scope**

Recent advances increasing the size of virtual libraries of molecules for drug discovery has enabled new approaches for virtual screening and virtual lead optimization. These methods have generally relied on machine learning (ML) models fit from previous screening assay data or other objectives such as structure-based methods like docking or relative binding free energy calculations. The success of such approaches may vary depending on the chemical space and diversity of the underlying training sets as well as the ML model type selected for fitting. Additional complicating factors to fitting ML models include the underlying uncertainties in the assay data and uncertainties in model predictions resulting in prediction errors.

Frederick National Laboratory for Cancer Research (FNLCR) as a member of the Accelerating Therapeutics for Opportunities in Medicine (ATOM) Consortium, drives research and technology development activities to support an AI-driver platform for drug discovery. Activities will involve tools including the ATOM Modeling Pipeline (AMPL) and generative molecular design (GMD).

ATOM’s AI-driven platform designs and optimizes drugs *in silico* considering their full pharmacological profile, and incorporates new experimental data, as needed, to advance discovery of new therapeutics.  ATOM is currently incorporating bioassay activity and property datasets from public and commercial sources for modeling. ATOM works with a variety of datasets from chemical bioassay activity to preclinical physicochemical, pharmacokinetic, and safety properties. In the following research projects, data sets will be analyzed or curated as preprocessing step using AMPL, next, AMPL machine learning models will be used to generate prediction using the data and prediction on new test set to further support ATOM’s active learning drug discovery platform such as GMD. Visualization will be incorporated as an important part of data analysis.

FNLCR Principal Investigator:

Titli Sarkar, PhD

Data Scientist

Frederick National Laboratory for Cancer Research

FNLCR Supporting Investigator:

Sean Black

Data Scientist

Frederick National Laboratory for Cancer Research

## **2022 Project scope:**

The students are expected to learn the basic pipeline of Machine Learning and how it can be useful in healthcare sector. They are expected to get some exposure on publicly available molecular databases like ChEMBL, ExCAPE-DB, DTC Database and how to use them. They are expected to go through some background research related to their project. Next, they should be able to use AMPL for data preprocessing, machine learning model training, validation and testing, using visualization tools to get an overview of the data and results, and be able to compare the results from different models.

Specific Aims are listed below:

*Aim1: ATOM data curation and model fitting for [Cytochrome P450s and hERG1 potassium channel related genes]*

FNLCR, as a co-founder of ATOM Consortium, is currently incorporating lead discovery which includes the property datasets from public and commercial sources for modeling into ATOM’s active learning drug discovery platform and using those datasets to design machine learning models for important drug property prediction. Given the importance of quality, curated datasets for fitting predictive models, we propose a summer student project to characterize and curate new datasets and fit machine learning models to those datasets. These datasets will include important targets of CYPSs or Cytochrome P450sand hERG1 potassium channel related genes. Cytochrome P450s have a central role in drug metabolism and understanding how they are inhibited or activated is critical for new drugs, especially in the context of drug-drug interactions when humans have to take multiple medicines. CYP enzymes are membrane-bound proteins that can control the speed at which drugs are metabolized in our body and the length of time that the drug will remain in our body this helps to control drug toxicity and safety. CYP enzyme inhibition is a principal mechanism for metabolism-based drug-drug interactions. Many chemotherapeutic drugs can cause drug interactions due to their ability to either inhibit or induce the CYP enzyme system. Genetic polymorphisms and epigenetic changes in CYP genes may be responsible for inter-individual and interethnic variations in disease susceptibility and the therapeutic efficacy of drugs. This project also aims to target hERG1 potassium channel which is an important ion channel in heart tissue that is blocked by lots of drugs causing dangerous toxicity. The human ether à-gogo related gene (hERG1) channel, responsible for the rapid component of the delayed rectifier potassium current (IKr), is one of the main determinants of action potential duration (APD). Functional hERG1 channels are formed by the assembly of four α-subunits encoded by the KCNH2 gene. The HERG channel is widely used for the assessment of proarrhythmic risk for new drugs. HERG channel blockers obstruct channel functions through various mechanisms, which usually show time dependence, voltage dependence, and state dependence. Students are expected to learn how to use The Predictive Oncology Model and Data Clearinghouse (MoDaC) which is a data-sharing repository developed to transition resources to the broader research community. In addition to working with existing data available on MoDaC, related to targets from three CYPs i.e. CYP2C9, CYP2D6, CYP3A4 and one hREG1 i.e. KCNH2, students are expected to explore more data related to these targets from literature study and working with some new data prepared within organization. The primary aims of this project will be to analyze the quality of underlying datasets, process the datasets to address heterogenous collection methods, and prepare the datasets for model fitting. These datasets are important components for prediction in virtual screening and lead optimization projects with the ATOM platform. Additionally, this project presents the opportunity to learn how to fit machine learning models with modern featurization approaches and model architectures.

## **Training Experience**

AI and data science underlie a revolution that is currently underway in the pharmaceutical and healthcare industry. Previously dormant data insights are transforming the long, costly, experimentally driven drug discovery process into a fast, AI-driven, patient-centric approach. To accelerate this process, ATOM is building the work force of the future – scientists and pharmacologists with integrated expertise in data science, AI, and drug discovery. FNLCR will provide a summer student training experience for select student(s). The experience will equip professional pharmacy student(s) with data science and machine learning expertise. During an approximately 10-12-week (40 hours/week) training period, students will be immersed in the AI-driven drug discovery process at ATOM, training with pharmaceutical industry, data, and machine learning scientists. Students will examine, analyze and build datasets that can support ATOM’s AI-driven drug discovery platform, as stated above. In addition to cross-training in data science and pharmaceutical domains, students will develop a repertoire of know-how in machine learning modeling, building prediction models and visualization for analysis of dynamic ODE modeling and drug discovery. The experience will be completely virtual. Students and mentors communicate through email and chat tools as well as ad hoc Zoom meetings.

*Planned experiences and goals for the training:*

* Python data science & ATOM software bootcamp to learn needed skills
* Weekly meetings with ATOM mentors covering various topics & assignments including:
  + Journal clubs to understand the latest science surrounding the project theme
  + Writing abstracts for individual student projects
  + Development of final presentations to the ATOM team & leadership team from member organizations
  + Submission of poster to a professional conference such as SC
* Weekly ATOM Technical Team Meeting to observe the real-world computational drug discovery efforts
* Daily interactions with other interns and mentors via Teams and video chats
* Weekly or biweekly job talks from ATOM team members & invited guests to give students ideas of data science careers in bioinformatics.
* Longer term: curated data, models and computer code contributions may be generated by students, incorporated into ATOM software, and will be appropriately attributed in any future publications.

During this experience, students may develop curated data sets, models, code, posters, and presentations. Some of these results may be captured in digital notebooks and hosting sites like GitHub. These results may be stored within the ATOM consortium databases and shared within the ATOM consortium for the future work or application. Students may present their results with ATOM consortium and other stakeholders in ATOM.

*Preferred experience by students:*

* Introductory programming with Python
* Familiarity with Jupyter notebook environment
* Linux/Unix
* Previous coursework in cheminformatics or computational chemistry, calculus, or advanced math
* Querying cheminformatics databases such as PubChem or ChEMBL