# **General Project Outline:**

Students will be assigned some machine learning related tasks

* + - Data curation
    - Model Training
    - Prediction on Test Data
    - Visualization

# **MoDaC Safety Screen Targets (?) Project:**

Students can request account on MoDaC and download Safety\_Screen\_Targets data from MoDaC. There are 9 kinds of genetic datasets on MoDaC ['CHRM1', 'CHRM2', 'CHRM3', 'CYP2C9', 'CYP2D6', 'CYP3A4', 'HRH1', 'KCNH2', 'PIK3CG']. Corresponding to each type, machine learning models were trained and the best models were developed om MoDaC. These models were developed using AMPL-1.1.0. Students can use the data from MoDaC, train different types of machine learning models eg. Random Forest, Graph Convolution , XGBoost etc. available in AMPL using newer version of AMPL (AMPL-1.3.0/1.4.0) and compare models for best accuracy.

**Which targets we want to address, will be determined soon. A more detailed plan will be outlined once we decide on the target(s).**

# **NEK Project:**

## **Overview:**

NEKs have been linked to a variety of illnesses when they are over or under expressed. The overall goal is to design very selective drugs that target a specific subset of NEKs while leaving the others alone. This is hard because the NEKs are somewhat similar and compounds that inhibit one can inhibit others. We also don’t have a lot of data, so we don’t have a lot of opportunities to learn the differences in how compounds bind to them. Some of the approaches listed in the plan assume that NEK inhibition is correlated, which, while true, does not help when finding selective compounds.

## **Related Papers:**

1. [https://doi.org/10.3390/molecules25081778](https://gcc02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.3390%2Fmolecules25081778&data=05%7C01%7Ctitli.sarkar%40nih.gov%7C883ea813aadb4ff1988308da377c6e9a%7C14b77578977342d58507251ca2dc2b06%7C0%7C0%7C637883306403693189%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=Ndh%2FBqqHV2uAmFVshQSviOK8fiQlSvmhdwP1Mbm8lZ8%3D&reserved=0)
2. <https://doi.org/10.1186/1747-1028-6-18>
3. <https://doi.org/10.1039/C7MD00510E>
4. <https://doi.org/10.1021/acs.jcim.7b00166>

## **Journal Club Tasks:**

* Read background on NEK family
* Conduct background survey
* Search ChEMBL for human kinases to use in data
* Collect references for existing GoStar targets (data will be supplied by Hiran, data have limited access- check with Naomi)

## **Goals:**

1. Design and test selective NEK inhibitors for use as probe molecules
2. Demonstrate model development for limited-data targets
3. Close the active learning loop with human-based synthesis

## **Background:**

* “Illuminating the Druggable Genome” is an NIH program for exploring the biology of understudied targets. Within that program NEK is a target family being studied by a group at UNC
* NEK
  + A family of kinases that participate in cell division, cilia formation, DNA damage repair
  + Implicated in multiple cancers, diabetes, CNS diseases, inflammation
  + Not well studied, data very limited
* *General problem:* Building property prediction models with very limited target data
* The idea of learning from related targets with more available data has been around for a long time. Kinases in general have a lot of data. Can we use it to create better NEK models?

## **Project Tracks:**

1. **Preprocessing – Data Curation**

* Learn Data Format from existing Livermore dataset (columns of the input .csv file)
* Learn Target (which property to predict)
* Explore other databases to collect more data
* Literature study for NEK data
* Search ChEMBL for human kinases to use in data
* Find GoStar references and double check that the references are Human.
* Finalize preliminary training data

1. **Baseline AMPL ML models**

* Collect and set up kinase training sets
* Train baseline models for kinase-specific representations
* Train and characterize NEK and selectivity models

1. **Visualize data and prediction results – NEK and general kinases**