Lawrence Livermore National Laboratory

LLNL - ATOM Consortium

Fall 2020 - Spring 2021

Table of Contents

Corporate Partner Mentors

Dr. Jonathan Allen
LLNL Corporate Partner Mentor
allen99@llnl.gov

Faculty Mentors

Dr. Mark Daniel Ward
Head of The Data Mine
mdw@purdue.edu

Justin Gould
Senior Data Scientist at The Data Mine
gould29@purdue.edu

Maggie Betz
Corporate Partners Senior Manager at The Data Mine
betz@purdue.edu

Student Biographies

Rosalie Wilfong
rwilfong@purdue.edu
Rose is a junior in Pharmaceutical Sciences at Purdue. She is involved in Purdue Club Tennis and Timmy Global Health, along with working in a lab on campus. She currently serves as the undergraduate teaching assistant for the LLNL team.

Shan Lu
lu310@purdue.edu
Shan is a third year PhD student in Industrial Physical and Pharmaceutical at Purdue College of Pharmacy. She works in Dr. Tonglei Li's lab for Drug Development. Her research focuses on oncology drug delivery. She currently serves as the graduate teaching assistant for the LLNL team.

Krystal Diaz
diaz161@purdue.edu
Krystal is a 2nd year Ph.D. student in Medicinal Chemistry and Molecular Pharmacology in the College of Pharmacy. She works in Dr. Rong Huang's lab in the Purdue Institute for Drug Discovery, where her research focus is the functional characterization of epigenetic pathways. Her contribution to this project includes the generation of models to characterize the interactions of various compounds with drug safety targets and the annotation of open-source Jupyter Notebooks for distribution to the public.

Camille Goenawan
cgoenawa@purdue.edu

Camille is a junior in Pharmaceutical Sciences in the College of Pharmacy. On campus, she is a member of Kappa Epsilon, a professional pharmacy fraternity. Her contribution to this project includes creating models to predict and characterize interactions between various compounds and the Histamine H1 Receptor.

Sylvia Liu
liu2715@purdue.edu
Sylvia is a sophomore in Computer Science in the College of Science. She is working with a subteam to create models that will predict how their team's compound interacts with others. In addition, she and her team have been annotating Jupyter notebooks provided by Dr. Allen.

Erika Meredith
meredie@purdue.edu
Erika is a freshman in Molecular Biology in the College of Science. She is currently performing data science research for Dr. Janice Evan's reproductive biology lab where she is developing a framework for identifying and analyzing the role of certain oocyte-enriched proteins in cardiovascular and other body systems. For this project, she is involved in the predictive modelling of the Aurora Kinase A Receptor's interactions with a variety of compounds.

Veer Pradhan
vvpradha@purdue.edu
Veer is a freshman in Data Science in the College of Science. He has assisted in annotating the Jupyter Notebooks given to the LLNL team by Dr. Allen meant for public consumption once completed, and is currently working in a group that has focused on the Aurora Kinase B receptor to generate predictive models between the various compounds.

Kiernan Schuerman
kschuerm@purdue.edu
Kiernan is a sophomore in Biomedical Engineering in the College of Engineering. Has assisted in annotating notebooks and is in the group that is focusing on studying the Aurora Kinase A receptor to generate models.

Sota Shishikura
sshishik@purdue.edu
Sota is a junior in Pharmaceutical Science in the College of Pharmacy. He is involved in Purdue Run club and pharmaceutical science club, and also works in a research lab on campus. He is working in a group to study the histamine H1 receptor.

Vidhi Singh
singh726@purdue.edu
Vidhi is a Computer Science major in the Class of 2024. Helped with annotating notebooks to make them more accessible to the public, as well as focused on generating and analyzing single task models for the Serotonin Receptor (HTR2A).

Sandi Shahini
sshahini@purdue.edu
Sandi is a junior majoring in Microbiology in the College of Science. He is working on creating models for the CHRM3 receptor.

Albert Zhang
zhan3112@purdue.edu
I am a junior in health science and for this project I am working in a group to study the protein target Aurora Kinase A. Interaction with this target is an indication that a molecule is unsafe for drug development. We have been working to create models that can accurately predict if a molecule will interact with our target. This will help our team screen out molecules that are not suitable for development.

Abigail Pati
apati@purdue.edu
Abigail is a junior in Chemical Engineering in the College of Engineering. She has assisted in annotating Jupyter notebooks provided by Dr. Allen. My subteam is currently working on generating models for the Aurora Kinase B receptor and its interactions with various compounds.

Project Statement

Currently our team is constructing machine learning models for various targets and comparing their accuracy results. We are also working with SLURM, a workload manager, to help digest our tasks and recently learned how to run it through RStudio. By the end of the semester, we hope to hopefully extract or identify potential drug candidates based on the biological targets we're analyzing. Between now and then, we hope to gain more experience working with SLURM and the various notebooks we've received from Dr. Allen. We also hope to get better at choosing best-fit models for our targets and optimizing them for our targets.

Executive Statement

The LLNL project is focused on work being performed within the ATOM consortium. Our team is working with various biological targets and developing various data-driven models to study potential drug candidate molecules, while working on developing best-fit models.

In the fall semester, we spent the first few weeks working on doing background research on various targets. This included locating the area of the body that the target effects, its protein name, UNIPROT ID, function, disease implications, and its 3D structure. The targets that we worked with in the fall include: PDE2A, PTGS2 (COX2), CYP3A4, CYP2D6, CYP2C9, OATP1B1, ADRA1B, ADRA2C, ADRB2, CHRM 1/2/3, DRD2, GRIN1, HRH1, HTR1B, HTR2A, HTR2C, HTR3A, PDE4D, MAOA, SLC6A2, SLC6A4, PI3Kγ.

Our next step was creating datasets for the targets. This was completed by extracting data relevant to our targets from the Drug Target Commons database and ExcapeDB database. Both databases are large public

databases, ExcapeDB which contains chemogenomics files and Drug Target Commons which is a crowd-sourcing platform that contains bioactivity data.

Once we had our datasets, we used data visualization to observe and analyze our targets. This was completed by utilizing several Jupyter notebooks that Dr. Allen shared with us. From these notebooks, we extracted various visuals, including pIC50 graphs, heat map distribution graphs based on Tanimoto similarity, Tanimoto distance distribution line graphs, and 2D cluster projections based on Tanimoto distance.

To wrap up fall semester, we focused on refining and comparing our datasets. We compared the ExcapeDB and the Drug Target Commons datasets for our targets and worked on removing any duplicates and extracting any outliers and analyzing them. We also worked on annotating, or adding helpful comments and explanations, to the notebooks that we worked with over the duration of the semester. We worked on this to help provide clarity about what the various functions in the notebook do, and to allow our results to be reproducible.

The spring semester is a continuation of the fall semester. In the spring semester, we are working with a refined set of targets. These targets include: AURKA, AURKB, CYP2C9, SLC6A2, CYP3A4, SLC6A4, HRH1, CHRM2, CHRM3, KCNH2, DRD2, CYP2C6, and HTR2A. We are mainly focused on working with AURKA, AURKB, CHRM2, CHRM3, and HRH1.

This semester, the team was divided into 4 sub-teams. The first group focused on comparing molecules active for AURKA and inactive for AURKB, while the second group will compare molecules active for AURKB and inactive for AURKA. The third group focused on CHRM2 and HRH1, while the fourth group focused on CHRM3 and HRH1.

The motivation behind AURKA and AURKB is that both are associated with various types of cancers. They are essential kinases for cell division via regulating mitosis, especially the process of chromosomal separation, they have also been implicated in regulating meiosis. The overexpression or gene amplification has been identified in various cancers and the inhibition of the kinases could potentially be used as treatment. The design challenge is that there are several drugs that are targeting both AURKA and AURKB and it is thought that it would be good to hit only one, preserving the function of the other.

The motivation behind CHRM2, CHRM3, and HRH1 is a desire in drug design to make antihistamines much more selective, because they often cause undesirable side effects, such as drowsiness. However, it can be very challenging to design a molecule that is receptive for one histamine receptor and ignore the others that cause undesirable side effects. Therefore, we are focusing on molecules active for HRH1 over CHRM2 and CHRM3.

To start the semester, we began with background research on our various targets and then started on splitting our datasets. We used two different types of splits: random and scaffold. Random splitting is just as it is named, it will randomly section off the molecules into groups. Scaffold splitting splits the molecules based on their two-dimensional framework. In comparison, random splitting isn't always the best mechanism for evaluating machine learning models because scaffold splitting offers a greater

challenge to the algorithms because the splits are separated based on structure. The datasets were split into three different sets: training, validation, and testing. After splitting the datasets, we receive a 2D UMAP of the similarity between our training and testing datasets based on their molecular framework. This creates a visualization of the differences between the two splits.

After splitting the datasets, we worked on building random forest models and graph convolutional neural networks to train on. To complete this, we worked with a Jupyter notebook Dr. Allen shared with us. Within the notebook, it displays how many models for each type (NN or RF) are associated with your target based on the type of split (random or scaffold) and its unique identifier. The models are trained on various different parameters and then assessed by their R^2 validation score. Within the notebook, we will receive different plots that visualize the R^2 scores. There are also plots that compare the models with the best R^2 score and the worst, and run predictions on them based on the training, test, and validation splits.

Currently, we are working on determining whether the models meet our multi-objective design criteria and optimizing our models to be best-fit with our targets based on their pIC50 values, R^2 values and ROC_AUC scores based on the validation dataset. We evaluated that the best random forest models produced test R^2 scores ranging from 0.5 to 0.6. Our goal is creating and reporting the best model for our targets.

Our work left to do is mostly choosing best-fit models and optimizing the best-fit models to work well on new molecules. We have goals to identify at least 1 or 2 promising drug candidate molecules that can go on for further evaluation. Finally, if time allows, we also have goals integrating the docking pipeline to score disease targets and build models that learn across multiple targets simultaneously.

Recommended Resources

I. General background knowledge on Neural Networks (written in R, but conceptually relevant):
   A. Building A Neural Net from Scratch Using R - Part 1 · R Views (rstudio.com)
   B. Building A Neural Net from Scratch Using R - Part 2 · R Views (rstudio.com)
II. Interesting progress on a related project:
   A. https://www.cnbc.com/2020/11/30/deepmind-solves-protein-folding-grand-challenge-with-alphafold-ai.html
III. Background on the Drug Target Commons dataset collection methodology:
   A. https://www.sciencedirect.com/science/article/pii/S2451945617304269
IV. Download Dr. Allen's introduction to the project webinar:
   A. https://www.rosaandco.com/webinars/2020/machine-learning-framework-small-molecule-drug-design
V. How to submit jobs through SLURM
   A. ITaP Research Computing - Knowledge (purdue.edu)
VI. Background Information on Spring 2021 Research Targets
   A. https://docs.google.com/spreadsheets/d/1U_59tYTe--xvmOhnQAv97RDlqZYS8OWs8hBXrYrLo4s/edit?usp=sharing
VII. Background Information on Fall 2020 Research Targets
   A. https://docs.google.com/spreadsheets/d/1_5928-WphNCB-jawEgI5NbcGTzhy84tZwLR3ZPcFOhI/edit?usp=sharing

VIII. Nature articles about drug discovery and AI
    A. https://www.nature.com/articles/s41573-019-0050-3
    B. https://www.nature.com/articles/nrd3845
IX. Deep Learning for the Life Sciences Textbook (use Purdue login)
    A. https://purdue-primo-prod.hosted.exlibrisgroup.com/permalink/f/vjfldl/PURDUE_ALMA51792556430001081
X. Meeting presentations
    A. intro_learning.pptx - Google Slides
    B. ml_review_20181212_3.pptx - Google Slides
    C. notes_02102021.pptx - Google Slides
    D. notes_02242021 (1).pptx - Google Slides
    E. semester_two_jan_2021.pptx - Google Slides
XI. Ravi's GitHub
    A. https://github.com/ravichas/AMPL-Tutorial

Data Used

I. Notebooks made by Dr. Allen (can be found in scholar @ /class/datamine/corporate/llnl/...):
    A. explore_data_excape_min_viable_one.ipynb
    B. explore_data_dtc_min_viable_one.ipynb
    C. split_dataset_example.ipynb
    D. explore_data_dtc_2_curate.ipynb
    E. build_rf_nn_example1.ipynb
    F. Split_dataset_example_with_binary_classes.ipynb
II. Screening libraries from Enamine
    A. Can be located at /class/datamine/corporate/llnl/allen99/S/…)
    B. Multiple folders in this directory that can be used, and they are grouped by size.

Status of Project

At the end of the 2021 Spring semester, we are wrapping up by screening the best model that each of us have created up until this point using the various notebooks that have been provided to us from Dr. Allen. This initial set of screened data is not only going to be added to the library, it is also a test of the process of screening so we can find out ways to make it easier, faster, and be able to screen more compounds at a larger scale. The team next semester should work on making their own models using the notebooks provided in this document and screen their best models as well to begin to build up the library.

Wish List

I. Use previous model training to create proper visualization and analysis tools
II. Use created models to develop an accurate prediction pipeline that can run multiple molecules at one time and score them
III. Run models through a virtual library to evaluate the created model against specified criteria
IV. Apply models we built to virtual screens and asses different strategies on how we apply them to virtual screens

V. Submit proven models to open source databases

Symposium Links (Posters and Video)

I. Link to Symposium poster and video for Fall 2020 - Spring 2021 project
    A. https://datamine.purdue.edu/symposium/atom/2021.html