# AI AND DRUG DISCOVERY

A Project

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science in Business Analytics

In

The College of Business Administration

By

Linxuan Zhang

Luke Liang

2021

**SIGNATURE PAGE**

**PROJECT:** AI AND DRUG DISCOVERY

**AUTHOR:** Linxuan Zhang

Luke Liang

**DATE SUBMITTED:** Summer 2021

College of Business Administration

Honggang Wang, Ph.D.

Project Committee Chair

Professor of Technology

and Operations Management _____

Jae Min Jung, Ph.D.

Project Committee Member

Professor of Marketing _____

## ACKNOWLEDGEMENTS

We would first like to thank our professor Jung, who helped us formulate the research questions and methodology aspects. His insightful feedback drove our progress. We would also like to thank our professor Wang for his invaluable guidance throughout our machine learning process.

## ABSTRACT

**Purpose**: Using ligand-based approach and Quantitative Structure Activity Relationship (QSAR), this project aims to apply machine learning (ML) to compounds screening stage, to speed up drug discovery processes and reduce failure rate.

**Study design**: In this project, QSAR model for exploring potential inhibitors of lung cancer target, C-Met, has been designed by using regression. 4771 molecules are collected from the ChEMBL database and calculated to chemical descriptors for ML model building. pIC50 is the focus as the higher the value, the better the potency of a drug.

**Findings**: Molecular weight, solubility, hydrogen bond donors and acceptors are four metrics that could measure drugs' absorption and toxicity. Random forest is chosen for predicting the potency of hit compounds for C-Met as its R2 (0.58) was decent. Also, it's a highly interpretable model.

**Originality/value**: Generating Shapley feature importance graph from random forest was the last step of the ML processes, but it's a starting point for lead-optimization because it could serve as a guideline for scientists. This project built on the foundation of QSAR, which is valuable for future research with the proteochemometric (PCM) model as it could generate interactions between multi protein targets instead of one target.

**Practical Implications**: Companies could expedite drug development using ML. They could filter out toxic drugs by following Lipinski's druglikeness rules. Instead of trial and error, drug developers would have a clear direction by using the Shapley graph when they alter compounds.



**Keywords**: AI, Drug Discovery, QSAR, Compound Screening, Random Forest

## SUMMARY OF INDIVIDUAL CONTRIBUTION

### Linxuan Zhang

I actively organized and participated in discussions and meetings with team members and professors. In the early stage, I researched and investigated the research background and drug discovery. I actively participated in testing the implementation and results of the computer code. After that, I verified the results and other research results again and analyzed the research data results. I wrote the first draft and participated in the production of the Business Analytics Project Proposal and final presentation slides. Finally, I prepared, created, and presented our final result with my team members.

### Luke Liang

I collected many datasets from ChEMBL, and I decided to go with C-Met. I mainly focused on python codes for data preprocessing, visualization, calculating descriptors, and model building. I researched much information such as drug discovery processes and relationship between IC50, target, and compound. Not only did I come up with recommendations from a business perspective, but I also derived solutions to potential issues they might be facing. As a team leader, I was also responsible for explaining ideas to my teammates.