

Chapman University

## Chapman University Digital Commons

---

Computational and Data Sciences (MS) Theses

Dissertations and Theses

---

Summer 8-2022

# Development of Machine Learning Models for Generation and Activity Prediction of the Protein Tyrosine Kinase Inhibitors

Ryan Kassab

*Chapman University*, rkassab@chapman.edu

Follow this and additional works at: [https://digitalcommons.chapman.edu/cads\\_theses](https://digitalcommons.chapman.edu/cads_theses)

---

### Recommended Citation

R. Kassab, "Development of machine learning models for generation and activity prediction of the protein tyrosine kinase inhibitors," M. S. thesis, Chapman University, Orange, CA, 2022. <https://doi.org/10.36837/chapman.000391>

This Thesis is brought to you for free and open access by the Dissertations and Theses at Chapman University Digital Commons. It has been accepted for inclusion in Computational and Data Sciences (MS) Theses by an authorized administrator of Chapman University Digital Commons. For more information, please contact [laughtin@chapman.edu](mailto:laughtin@chapman.edu).

Development of Machine Learning Models for Generation and Activity Prediction of  
the Protein Tyrosine Kinase Inhibitors

A Thesis by

Ryan H. Kassab

Chapman University

Orange, CA

Schmid College of Science and Technology

Submitted in partial fulfillment of the requirements for the degree of  
Master of Science in Computational and Data Sciences

August 2022

Committee in charge:

Gennady Verkhivker, Ph.D., Chair

Mohamed Allali, Ph.D.

Cyril Rakovski, Ph.D.



**CHAPMAN UNIVERSITY**  
SCHMID COLLEGE OF SCIENCE AND TECHNOLOGY  

---

Computational and Data Sciences

The thesis Ryan H. Kassab is approved.

*Gennady Verkhivker*

---

Gennady Verkhivker, Ph.D., Chair

*M. Allali*

---

Mohamed Allali, Ph.D.

*Cyril Rakovski*

---

Cyril Rakovski, Ph.D.

June 2022

Development of Machine Learning Models for Generation and Activity

Prediction of the Protein Tyrosine Kinase Inhibitors

Copyright © 2022

by Ryan H. Kassab

## ACKNOWLEDGEMENTS

I would like to thank Dr. Gennady Verkhivker for his support and encouragement during this process. It was a great opportunity to work with him and I appreciate everything he has done for me. If it wasn't for Dr. Verkhivker, I do not think would have fallen in love with computational biology research as much as I did, and I thank him for putting me down the path that I currently am on. I also want to thank Dr. Hesham El-Askary, Dr. Mohammed Allali, Dr. Cyril Rakovski, as well as all the other people in my life that supported me during this journey. It would have been impossible to succeed if not for the constant support that everyone provided me.

I would also like to extend an appreciation to both the Computer Science department and CADS program at Chapman University, as well as those in charge of them. I was unsure of what I wanted to pursue before beginning my undergraduate career, but the computer science department at Chapman University helped grow my talents and love for the field. In continuing my education with the CADS program, the diversity of classes helped me find the computational biology field that I fell in love with. If it wasn't for enrolling in the program, I don't think I would have ever found this field.

I also want to thank the Kay Foundation for its support of this research experiment and consequently my thesis project with its generous grant. I want to thank Dr. Keykavous Parang and his lab for his support with this project and all the collaboration done with the School of Pharmacy.

## ABSTRACT

Development of Machine Learning Models for Generation and Activity Prediction of the Protein

Tyrosine Kinase Inhibitors

by Ryan H. Kassab

The field of computational drug discovery and development continues to grow at a rapid pace, using generative machine learning approaches to present us with solutions to high dimensional and complex problems in drug discovery and design. In this work, we present a platform of Machine Learning based approaches for generation and scoring of novel kinase inhibitor molecules. We utilized a binary Random Forest classification model to develop a Machine Learning based scoring function to evaluate the generated molecules on Kinase Inhibition Likelihood. By training the model on several chemical features of each known kinase inhibitor, we were able to create a metric that captures the differences between a SRC Kinase Inhibitor and a non-SRC Kinase Inhibitor. We implemented the scoring function into a Biased and Unbiased Bayesian Optimization framework to generate molecules based on features of SRC Kinase Inhibitors. We then used similarity metrics such as Tanimoto Similarity to assess their closeness to that of known SRC Kinase Inhibitors. The molecules generated from this experiment demonstrated potential for belonging to the SRC Kinase Inhibitor family though chemical synthesis would be needed to confirm the results. The top molecules generated from the Unbiased and Biased Bayesian Optimization experiments were calculated to respectively have Tanimoto Similarity scores of 0.711 and 0.709 to known SRC Kinase Inhibitors. With calculated Kinase Inhibition Likelihood scores of 0.586 and 0.575, the top molecules generated from the Bayesian Optimization demonstrate a disconnect between the similarity scores to known SRC Kinase Inhibitors and the calculated Kinase Inhibition Likelihood score. It was found that

implementing a bias into the Bayesian Optimization process had little effect on the quality of generated molecules. In addition, several molecules generated from the Bayesian Optimization process were sent to the School of Pharmacy for chemical synthesis which gives the experiment more concrete results. The results of this study demonstrated that generating molecules through Bayesian Optimization techniques could aid in the generation of molecules for a specific kinase family, but further expansions of the techniques would be needed for substantial results.

# TABLE OF CONTENTS

	<u>Page</u>
<b>ACKNOWLEDGEMENTS</b> .....	<b>IV</b>
<b>ABSTRACT</b> .....	<b>V</b>
<b>LIST OF TABLES</b> .....	<b>IX</b>
<b>LIST OF FIGURES</b> .....	<b>X</b>
<b>LIST OF ABBREVIATIONS</b> .....	<b>XII</b>
<b>1 INTRODUCTION</b> .....	<b>1</b>
1.1 Protein Kinases for Oncological Applications.....	1
1.2 Machine Learning in Computational Biology .....	3
1.3 Machine Learning and Drug Discovery.....	5
1.3.1 Machine Learning Models Created for Drug Discovery .....	6
<b>2 ACCUMULATION AND REPRESENTATION OF MOLECULAR DATA</b> .....	<b>7</b>
2.1 Accumulating Molecular Data.....	7
2.1.1 Representations of Kinase Inhibitors.....	7
2.1.2 Chemical Databases .....	9
2.2 Latent Space Representation.....	11
<b>3 STRATEGIES FOR GENERATION OF NOVEL MOLECULES</b> .....	<b>15</b>
3.1 Utilizing the SMILE String Molecular Representation .....	15
3.2 Developing a Measurement for Kinase Inhibitor Success.....	16
3.3 Utilizing the Aggregation Techniques of Random Forest Models .....	17
3.4 Applying Bayesian Optimization for Generating Molecules.....	18
<b>4 MEASURING KINASE INHIBITION LIKELIHOOD</b> .....	<b>19</b>
4.1 Implementing Random Forest Learning Modeling.....	19
4.2 Binary and Multiclass Classification .....	23
4.2.1 Binary Classification Model .....	23
4.2.2 Multiclass Classification Model .....	24
4.3 Training on Latent Space Location and Chemical Features .....	25
4.3.1 Using Latent Space Coordinates.....	25
4.3.2 Chemical Feature-Based Training .....	26
4.4 Kinase Inhibition Likelihood .....	27
4.4.1 Measuring Accuracy of Kinase Inhibition Likelihood .....	27



4.5	Analysis of Random Forest Model Performance.....	30
4.5.1	Latent Space Location Based Results .....	30
4.5.2	Chemical Feature Based Results.....	32
<b>5</b>	<b>GENERATING MOLECULES USING BAYESIAN OPTIMIZATION .....</b>	<b>37</b>
5.1	Introduction.....	37
5.2	Objective Function.....	39
5.2.1	Training the Objective Function .....	40
5.3	Bias within the Optimization Function .....	41
5.3.1	Utilizing a Multiclassification Random Forest Model as the Scoring Function 41	
5.4	Similarity Testing.....	42
5.5	Analysis of Generated Molecules .....	43
5.5.1	Kinase Inhibition Likelihood Evaluation.....	44
5.5.2	Chemical Feature Evaluation .....	45
5.5.3	QED, logP, and SAS Evaluations .....	55
5.5.4	Tanimoto Similarity Testing Evaluation.....	57
5.5.5	Analysis of Top Generated Molecules and Bayesian Optimizers .....	59
<b>6</b>	<b>CONCLUSION .....</b>	<b>70</b>
6.1.1	Future Expansions.....	73
<b>7</b>	<b>REFERENCES.....</b>	<b>75</b>

## LIST OF TABLES

	<u>Page</u>
<b>Table 1.</b> Binary Latent Space-Based Random Forest Classification Report Results.....	30
<b>Table 2.</b> Multiclass Classification Latent Space-Based Random Forest Classification Report Results.....	31
<b>Table 3.</b> Binary Chemical Feature-Based Random Forest Classification Report Results	32
<b>Table 4.</b> Multiclass Classification Chemical Feature-Based Random Forest Classification Report Results.....	35

# LIST OF FIGURES

	<u>Page</u>
<b>Figure 1.</b> Diagram of Process of Encoding Molecules into Latent Space Representations	12
<b>Figure 2.</b> (A) 2-D latent space representation of the kinase inhibitor families and GDB small molecules, (B) 2-D density plot of latent space representation shown in (A), (C) 2-D latent space representation of only the kinase inhibitor families, and (D) a 2-D density plot of the latent space representation shown in (C)	13
<b>Figure 3.</b> The visual representation of one of the Trees in the Binary Chemical Feature-Based Random Forest Model	21
<b>Figure 4.</b> Diagram of Random Forest Model	22
<b>Figure 5.</b> (A) Receiver Operating Characteristic Curve, (B) Feature Importance Histogram, and (C) Precision-Recall Curve of the Binary Classification Random Forest model	33
<b>Figure 6.</b> Figurative Representation of the Exploration, (A) and (B), and Exploitation, (C), Processes of Bayesian Optimization	38
<b>Figure 7.</b> Bar Graphs of (A) the average Kinase Inhibition Likelihood scores of all generated molecules, (B) the max Kinase Inhibition Likelihood scores of all generated molecules from the Unbiased and Biased Bayesian Optimizers	44
<b>Figure 8.</b> Histograms of the distribution of the LabuteASA values from the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors	46
<b>Figure 9.</b> Histograms of the distribution of the molecular weights of the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors	48
<b>Figure 10.</b> Histograms of the distribution of the HallKier Alpha values of the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors	50
<b>Figure 11.</b> Histograms of the distribution of the number of aromatic rings of the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors	52
<b>Figure 12.</b> Histograms of the distribution of the number of aromatic carbocycles of the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors	54

<b>Figure 13.</b> Bar Graphs of (A) the average QED scores, (B) the average logP scores, and (C) the average SAS scores of the molecules generated from the Biased and Unbiased Bayesian Optimizer, in comparison to the known SRC Kinase Inhibitors.....	56
<b>Figure 14.</b> Bar Graphs of (A) the average similarity scores of all generated molecules, (B) the max similarity scores of all generated molecules, and (C) the average similarity scores of the generated molecules with a calculated Kinase Inhibition Likelihood score above 0.5 from the Biased and Unbiased Bayesian Optimizers.....	58
<b>Figure 15.</b> The Top Three Molecules Generated from the Unbiased Bayesian Optimizer with the Closest Known SRC Kinase Inhibitors.....	61
<b>Figure 16.</b> The Top Three Molecules Generated from the Biased Bayesian Optimizer with the Closest Known SRC Kinase Inhibitors.....	63
<b>Figure 17.</b> Top Molecules Generated Through Bayesian Optimization Sent to Pharmacy School for Synthesis (Molecules 1-3).....	66
<b>Figure 18.</b> Top Molecules Generated Through Bayesian Optimization Sent to Pharmacy School for Synthesis (Molecules 4-6).....	67
<b>Figure 19.</b> Top Molecules Generated Through Bayesian Optimization Sent to Pharmacy School for Synthesis (Molecules 7-9).....	68

## LIST OF ABBREVIATIONS

<b><u>Abbreviation</u></b>	<b><u>Meaning</u></b>
ASA	Accessible Surface Area
AUC	Area Under Curve
GAN	General Adversarial Network
PTK	Protein Tyrosine Kinase
ROC	Receiving Operating Characteristic
SMILES	Simplified Molecular-Input Line-Entry System
VAE	Variational Autoencoder



# 1 Introduction

Drug discovery applications seek to design small molecules that have specific targets. This typically requires extensive trial and error when creating and assessing molecules. Given the levels of success achieved in previous machine learning aided molecular design applications ( (Maziarka, et al., 2020) (De Cao & Kipf, 2018); (Kadurin, et al., 2017); (Yu, Zhang, Wang, & Yu, 2017)), we decided to see if the same idea could be applied to generate and alter molecules to display targeted properties and outcomes. Our goal was to create a framework in which we can generate molecules with the potential of being protein tyrosine kinase inhibitors specifically belonging to the SRC kinase inhibitor family.

## 1.1 Protein Kinases for Oncological Applications

Protein kinase inhibitors are small molecules that block the actions of enzymes called protein tyrosine kinases. Protein tyrosine kinases (PTK) are a class of proteins that contain tyrosine kinase activity and catalyze the transfer of phosphate groups on ATP to tyrosine residues on proteins, invoking phosphorylation and signal transferring in many cellular functions including metabolism, cell cycle regulation, survival, and differentiation (Kannaiyan & Mahadevan, 2018). Dysregulation of these protein kinases has been implicated in various carcinogenic processes (Kannaiyan & Mahadevan, 2018). The most common protein domains that are implicated in cancer are protein kinases and there are more than 500 kinase enzymes encoded in the human genome. Deregulation of these proteins leads to the development of cancer and other disorders. Many protein kinases have emerged as important therapeutic targets for combating diseases caused by abnormalities in signal transduction pathways.

Structurally, PTK's can be divided into two categories: Receptor PTK (RTK) and Non-receptor PTK (NRTK). The two categories further divide into multiple enzymes based on their structural homology. Analysis of human genome data shows that there are 518 kinase genes in the human body, of which 90 have been identified PTK, including RTK which contains 58 species and NRTK which contains 32 species. For example, RTKs include Epithelial Growth Factor Receptors (EGFR), a cell surface receptor that is pivotal in the regulation of survival and apoptosis of epithelial cells. Overexpression of EGFR can result in multiple epithelial cell-based tumors, such as lung cancer, breast cancer, and bladder cancer to name a few (Jiao, et al., 2018).

The SRC Family is an important member of NRTK. The SRC Kinases play a huge role in the regulation of many cells through the extracellular binding of the ligand to the cellular receptor, inducing multiple signaling pathways that affect cell adhesion, mobility, proliferation, and angiogenesis. These include the RAS/RAF/MEK/ERK pathways, the PI3K/AKT/mTOR pathway, and the STAT3 pathway that regulates the expression of c-Myc and Cyclin D1 (Jiao, et al., 2018). In normal circumstances, the activity site of the SRC Kinase is closed, inhibiting its expression (Jiao, et al., 2018). However, under the activity of exogenous and endogenous carcinogen factors, the kinase becomes hyperactivated, leading to uncontrolled cell proliferation and differentiation and in turn, tumorigenesis (Jiao, et al., 2018). The specific family of SRC Tyrosine Kinase Inhibitors are important due to the association of SRC Tyrosine Kinases activity with cancer. Tyrosine Kinase Inhibitors (TKI) can compete with the ATP binding site of the tyrosine kinase with ATP and reduce tyrosine kinase inhibition, therefore inhibiting cancer cell proliferation.



## 1.2 Machine Learning in Computational Biology

Advances in the machine learning field have driven the design of new computational systems that improve with experience and are able to model increasingly complex chemical and biological phenomena ( (Chen, Engkvist, Wang, Olivecrona, & Blaschke, 2018); (Dimitrov, Kreisbeck, Becker, Aspuru-Guzik, & Saikin, 2019); (Goh, Hodas, & Vishnu, 2017); (Korotcov, Tkachenko, Russo, & Ekins, 2017); (Mater & Coote, 2017); (Popova, Isayev, & Tropsha, 2018)). Machine learning techniques have been successfully applied to various computational chemistry challenges ( (Husic & Pande, 2018)), pharmaceutical data analysis, protein–ligand binding affinity prediction problems ( (Ballester & Mitchell, 2010), (Decherchi, Berteotti, Bottegoni, Rocchia, & Cavalli, 2015)), dissecting molecular determinants of protein mechanisms and biochemical reactions (Li, Kermode, & De Vita, 2015). Data-intensive machine learning modeling can be also applied for detection and classification of allosteric protein states. The integration of Markov modeling, simulations, and machine learning approaches into robust and reproducible computational pipelines with the experimental feedback can be explored for atomistic modeling and classification of allosteric states. Two key factors were necessary for them to see so much use. First, large amounts of rich data. We are generating more data today than ever before and the biochemistry field is no exception. Computational tools for molecular modeling (De Cao & Kipf, 2018), protein folding simulation, or mutation analysis have started to generate more data than can even be stored. Second, powerful computational tools like GPUs that augment our abilities to perform parallel processing allowing machine learning models to ingest these large datasets. This has allowed medicine to become more personalized, with current research catering solutions to specific genetic profiles rather than taking a one size fits all approach. Much of the benefit of these methods comes from their versatility: not only do they

both generate and analyze data, but often enhancements to machine learning techniques in one domain can be readily applied to techniques in any domain. For example, techniques designed for the image processing domain have been applied to molecular design (Maziarka, et al., 2020).

In addition, deep neural network methods were successfully applied to predict intrinsic molecular properties such as atomization energy based on simple molecular geometry and element types. Deep learning models were recently used for structure-functional prediction of cancer mutations and functional hotspots of ligand binding in cancer-associated genes (Agajanian, Odeyemi, Bischoff, Ratra, & Verkhivker, 2018). The developed models can capture about 90% of experimentally validated mutational hotspots and yield novel information about molecular signatures of driver mutations. Other studies have shown that deep learning models can learn high importance features from raw genomic information and produce reliable recognition and classification of functionally significant cancer mutation hotspots. Moreover, these deep learning models can often outperform computational predictors of cancer mutations that are based on protein sequence and structure features (Agajanian, Oluyemi, & Verkhivker, Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations, 2019). The success of deep learning tools in deciphering important functional phenotypes directly from primary sequence information is encouraging as these models can bypass the need for many empirically derived functional and structural features. However, machine learning methods often result in "black box" models with limited interpretability. There has been an explosion of interest in transparent and interpretable machine learning models to enable more efficient data mining and scientific knowledge discovery (Holzinger, Dehmer, & Jurisica, 2014). Our investigations have also provided a roadmap how to combine deep learning predictions of functional sites with

subsequent biophysical analysis to aid in the interpretability of machine learning models and facilitate their applications in biological problems (Agajanian, Odeyemi, Bischoff, Ratra, & Verkhivker, 2018).

### **1.3 Machine Learning and Drug Discovery**

By exploring the role of kinases and kinase inhibitors, it has not only helped advance the field of cancer biology but also led to the advent of ‘targeted therapy’ or ‘personalized medicine’ in cancer leading to a paradigm shift in cancer therapy (Kannaiyan & Mahadevan, 2018). This has led to excitement from the biochemical research community towards the creation of protein kinase inhibitors that can be used as anticancer therapeutic agents, and in recent times, utilizing computational tools such as Generative Machine Learning or Deep Learning models to achieve this goal.

Generative deep learning models have excelled as tools to aid in navigating the large space of known molecules and in the creation of new molecules. These models are fed various representations of molecules as inputs and learn to perform a variety of things, such as the optimization of these molecules towards a targeted property. This task requires a large amount of data to perform successfully, which in turn requires non-trivial computational resources. An additional functionality that generative models provide is making alterations to inputs to yield transformed outputs. An additional functionality that generative models provide is making alterations to inputs to yield transformed outputs.

### 1.3.1 Machine Learning Models Created for Drug Discovery

This idea has been applied in the chemistry domain with the Mol-CycleGAN (Maziarka, et al., 2020) and MolGAN (De Cao & Kipf, 2018). SeqGAN (Yu, Zhang, Wang, & Yu, 2017), created molecules one token at a time, which machine learning models had trouble doing before (Yu, Zhang, Wang, & Yu, 2017).

JT-VAE (Jin, Barzilay, & Jaakkola) and Chemical VAE (Gómez-Bombarelli, et al., 2018) are two successful VAE approaches that have high potential for assisting a molecular generation task. These variational autoencoders are trained using the Simplified Molecular Input Line Entry System (SMILES) format and outperform traditional string encoding techniques while providing a continuous representation. Notably, druGAN combined GAN and VAE by training an adversarial autoencoder to efficiently sample molecules from the latent space (Kadurin, et al., 2017).

MolGAN was designed to create new molecules that optimize a portfolio of different properties that include drug likelihood (QED), synthesizability (SAS) and water-octanol partition coefficient (logP). This requires the model to learn a probabilistic pattern about molecular structure. The authors of this paper achieved state of the art performance in the optimization of all these selected properties, while maintaining almost 100 percent validity of generated molecules (De Cao & Kipf, 2018). However, the MolGAN framework struggled with creating unique molecules, suffering from mode collapse since it could only sample nine atoms to create molecules (De Cao & Kipf, 2018). Regardless of the mode collapse issues, MolGAN exhibited impressive performance learning a complicated task and proving that machine learning models can execute molecular design. Maziarka et al., proved that machine learning models trained in a reinforcement learning

paradigm could successfully make structural alterations to small molecules and maintain the validity of these altered molecules with Mol-CycleGAN (Maziarka, et al., 2020).

These generative deep learning models allow us to build a bridge between the chemical and continuous space and understand the compromise between invoking small incremental changes to radical modifications to generate optimal molecules for therapeutic benefits. In our work, we create a platform of Machine Learning approaches to generate novel molecules with the goal of being Kinase Inhibitors. We will start by discussing the accumulation of data to use as a training set for our downstream Machine Learning processes and its representation in the continuous Latent Space created by ChemVAE. Once discussing our platform, we discuss our scoring metric of Kinase Inhibition Likelihood which will evaluate our generated molecules based on various structural and chemical-based characteristics. We use Bayesian Optimization approaches and the known SRC Kinase Inhibitors as a guide to generate novel molecules. We then use Similarity and drug-like metrics to analyze structural and chemical similarity of our novel compounds to those of SRC Kinase Inhibitors.

## **2 Accumulation and Representation of Molecular Data**

### **2.1 Accumulating Molecular Data**

#### **2.1.1 Representations of Kinase Inhibitors**

There is a wealth of active research dedicated towards determining the optimal representation for molecules with respect to machine learning models. Each comes with its own set of pros and cons, and there isn't a clear best choice. 3D molecules, in the form of voxels (Kuzminykh, et al., 2018), struggle with the invariance of their representations. Different rotations, translations, or

permutations of the atomic indexing can yield different 3D grids that all represent the same molecule. 2D representations cannot encode as much information as their 3D counterparts, but don't suffer from as many of the same invariance issues. They typically are represented by an 80x80 grayscale image, but RGB channel style representations have been attempted. 1D string representations, otherwise known as SMILES (Weininger, 1988), are the most widely used due to their simplicity and wide availability. SMILES strings can represent seven important characteristics of a molecule: atoms, bonds, rings, aromaticity, branching, stereochemistry, and isotopes. Another prominent representation is the Coulomb matrix (CM) introduced by (Rupp et al.), which is a square atom-by-atom matrix containing an approximate potential energy of the free atom along the diagonal and pair Coulombic potentials on the off-diagonal terms (Townshend, et al., 2020). An improvement over CM is typically observed using the Bag-of-Bonds (BoB) representation, where each atomic pair is placed in specific vectors (bags) based off the elemental pairs and sorted by value. Another representation of molecules is FCHL, a representation based on Gaussian distribution functions for the universal kernel ridge regression-based quantum machine-learning models. In addition, the smooth overlap of atomic positions (SOAP) representation calculates the local density of atoms around all atoms in each chemical environment but suffers from an increased computational cost over the pairwise CM and BoB representations (Townshend, et al., 2020).

Within numerous molecular modeling approaches, the consistent use of SMILES representations allows for an extensive dataset to be used during training.

### 2.1.2 Chemical Databases

Typically, models learn how to create or alter molecules by training with as much data as possible and only successfully accomplish the task by iterating hundreds of thousands of times over large datasets. Numerous large databases are available that contain molecules in a variety of representations including SMILES, 2D, and 3D. From these databases we will need a large dataset of known protein kinase inhibitors to analyze and emulate. In addition, we will need a baseline set of random similar small molecules to understand what differentiates our protein kinase inhibitor set from any other molecule and to act as a control group. There is no shortage of publicly available biochemical databases, enumerating up to 166 billion molecules (Ruddigkeit, van Deursen, Blum, & Reymond, 2012) in some cases.

These databases serve as catalogs to search through to choose molecules with desired properties. Examples of these databases include GDB (Ruddigkeit, van Deursen, Blum, & Reymond, 2012), PubChem (Kim, et al., 2019), ZINC (Irwin & Shoichet, 2005), CAS (Ruddigkeit, van Deursen, Blum, & Reymond, 2012), ChEMBL (Gaulton, et al., 2012), and DrugBank (Wishart, et al., 2008). GDB-17 (Ruddigkeit, van Deursen, Blum, & Reymond, 2012) was created by capitalizing on work previously performed on the GDB-13 (Ruddigkeit, van Deursen, Blum, & Reymond, 2012) database. The authors of the paper operated on the molecular graphs directly and enhanced the memory efficiency of their graph analysis to overcome the limitations of GDB-13 (Ruddigkeit, van Deursen, Blum, & Reymond, 2012). This allowed for the enumeration of much larger molecules, yielding the massive 166 billion molecules currently deposited in the database.

GDB-17's 166 billion molecule size is much larger than alternative options such as PubChem-17 or CAS-17 (Ruddigkeit, et al., 2012). The authors compared the size and composition of their molecular database with public archives from PubChem (Kim, et al., 2016), ChEMBL (Gaulton,

et al., 2012), and DrugBank (Wishart, et al., 2008). GDB-17 was the leading dataset for compliance of reference datasets while having the highest percentage of molecules with at least one small ring out of all datasets (Ruddigkeit, et al., 2012). Another attractive component of GDB-17 (Ruddigkeit, et al., 2012) is that it has a more uniform distribution of topologies than the other datasets. Notably, DrugBank-17 exhibited the most uniform distribution for the different categories (heteroaromatic, aromatic, heterocyclic, carboxylic, acyclic). This allows us to obtain a very large set of random small molecules to serve as the foundation for our molecular generation experiments that are sufficiently representative of the underlying distribution. We obtained a sample of this database corresponding to 163,953 random small molecules from a variety of domains.

To complement our random baseline molecules found in GDB-17, we took around 20,000 kinase inhibiting small molecules across 10 Protein Kinase Inhibitor families from the ZINC (Irwin & Shoichet, 2005) database for our training dataset: SRC, ABL1, CSF1R, EGFR, FLT3, KDR, LCK, MAPK10, MAPK14, and MET. ZINC maintains an active list of known protein kinase inhibitors, which allow us to create a training set for any downstream machine learning processes used in our experiments. Additionally, ZINC specializes in commercially available drug compounds enhanced with properties about these molecules such as molecular weight, calculated logP, and number of rotatable bonds (Irwin & Shoichet, 2005). Other chemical metrics were calculated using RDKit, a framework that is a collection of cheminformatics and machine learning software for chemical data analysis (Landrum, 2013). All molecules collected have a molecular weight less than 700, calculated logP between -4 and 6, less than 6 rotatable bonds, and only contain a set of 10 atoms (C, N, O, F, S, P, Cl, Br, I). The GDB baseline and kinase inhibiting small molecules data will be

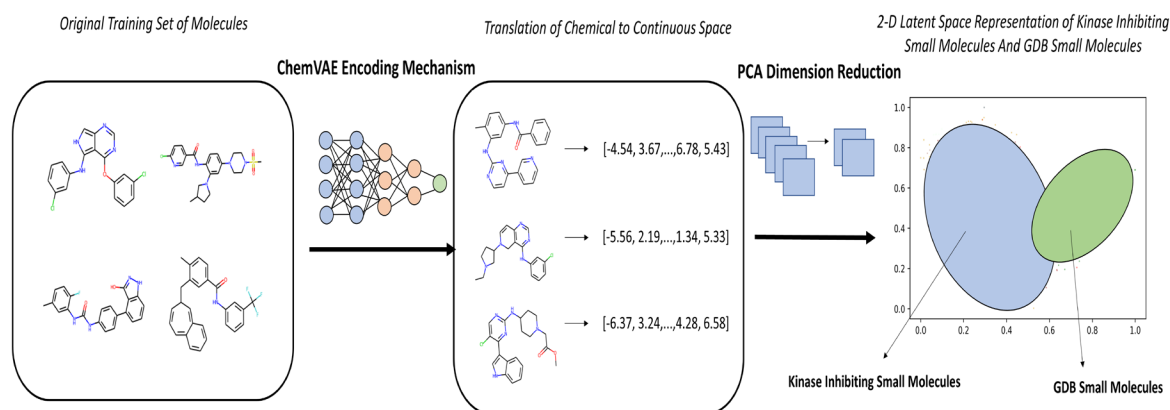


used by the Deep Learning framework ChemVAE (Gómez-Bombarelli, et al., 2018) for conversion into the kinase inhibitor latent space.

## 2.2 Latent Space Representation

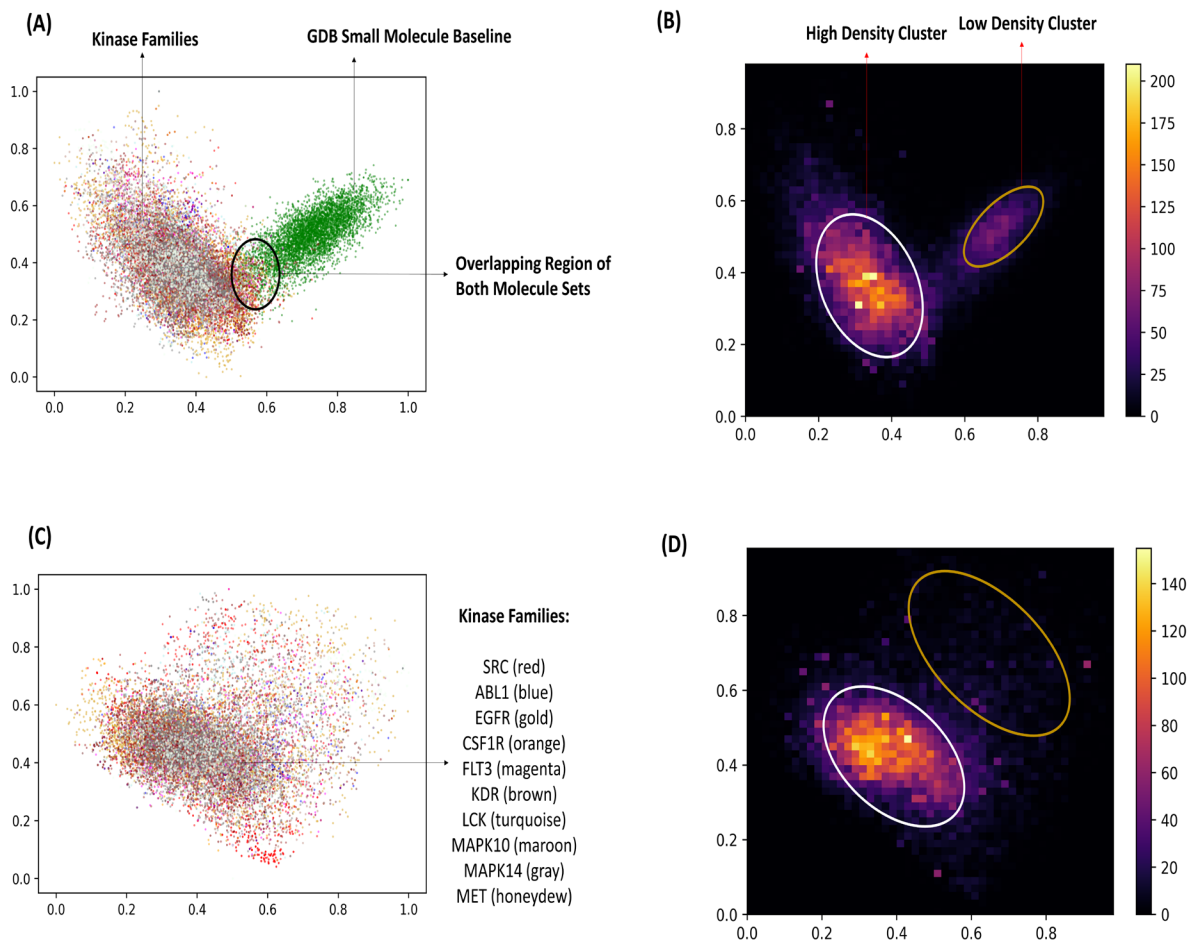
Chemical Variational Auto-Encoder or ChemVAE is a deep learning framework that works to convert discrete representations of molecules to and from a multidimensional continuous representation (Gómez-Bombarelli, et al., 2018). The ChemVAE framework contains 2 deep networks: an encoder that converts each SMILE string to a fixed dimensional vector and a decoder that converts the fixed dimensional vector back into SMILE strings (Gómez-Bombarelli, et al., 2018). The autoencoder works to minimize the error in reproduction of the original SMILE string, i.e. it learns the identity function (Gómez-Bombarelli, et al., 2018). The key of the autoencoder is attempting to map the SMILE string through an information bottleneck. The bottleneck forces the fixed length vector to learn the main statistically salient features in the data (Gómez-Bombarelli, et al., 2018). This vectorized representation is known as the latent space representation of the molecule.

In addition to providing benefits, such as compression, autoencoders provide another key benefit: the creation of a continuous space that can be searched (Hinton & Salakhutdinov, 2006). In the SMILES domain, there are discrete changes from one string to another and no clear graphical representation that can be observed or searched. This leads to difficulty in experimenting with or optimizing candidate molecules. Autoencoders solve this problem by mapping these string representations into a numeric domain where much of the structural chemical properties are maintained.



**Figure 1.** Diagram of Process of Encoding Molecules into Latent Space Representations

Chemical VAE tailored the variational autoencoder to the biochemical realm by simultaneously training the model to predict the QED, SAS, and logP values (Gómez-Bombarelli, et al., 2018). By optimizing the recreation objective function, the model is forced to learn information about the nature of the molecules causing the latent space to preserve important characteristics of their inputs (Gómez-Bombarelli, et al., 2018). Furthermore, this tool provides us with a complete and continuous representation of the known molecular landscape, where we can find any molecule with the right tools.



**Figure 2.** (A) 2-D latent space representation of the kinase inhibitor families and GDB small molecules, (B) 2-D density plot of latent space representation shown in (A), (C) 2-D latent space representation of only the kinase inhibitor families, and (D) a 2-D density plot of the latent space representation shown in (C)

To understand the multidimensional continuous representation of our molecules, we sampled a set of 100,000 random small molecules from GDB-17, 1883 ABL1 kinase inhibiting small molecules, 3477 SRC kinase inhibiting small molecules, 662 CSF1R kinase inhibiting small molecules, 3971 EGFR kinase inhibiting small molecules, 1058 FLT3 kinase inhibiting small molecules, 4252 KDR kinase inhibiting small molecules, 1554 LCK kinase inhibiting small molecules, 859 MAPK10 kinase inhibiting small molecules, 3659 MAPK14 kinase inhibiting small molecules, and 2081 MET kinase inhibiting small molecules all from ZINC and encoded them through the

ChemVAE framework. The 196-dimensional vectors representing these molecules were fed through principal component analysis (PCA) so that we could visualize the continuous space in two dimensions. The two-dimensional latent space representations and their respective density plots are shown in Figure 2. We assigned a different color for each data point so that they could be differentiated on the graph and saw that the kinase inhibiting molecules were heavily concentrated in one area. Our GDB small molecule dataset was in a different area, conveying to us there is regional difference between our GDB small molecule dataset and our kinase inhibiting families (Figure 2A and 2C). However, it is important to note that the cluster of kinase inhibitors isn't pure and there are molecules from GDB-17 present within them. This does not necessarily mean that the intersecting molecules are not kinase inhibitors. It is valid to assume that some of these might exhibit kinase inhibition potential because there are many molecules that might simply be undiscovered kinase inhibitors. The density plots display concentrations of our encoded molecules within various regions of the latent space. Areas with colors closer to or of yellow indicate higher density regions and areas with colors close to or of purple indicate the latter. The density plots emphasized the emergence of highly skewed clusters within the latent space by emphasizing highly concentrated regions and sparsely concentrated regions (Figure 2B and 2D). The organization of our molecules in the hyperplane and the emergence of these highly skewed clusters demonstrates potential for accurate classification and generation. To capitalize on this potential for molecular generation, we need to choose the best methods of generating molecules and assessing the generated molecules to determine the experiment's success.

## **3 Strategies for Generation of Novel Molecules**

### **3.1 Utilizing the SMILE String Molecular Representation**

The main objective of this experiment was to generate novel molecules using machine learning frameworks that would potentially be synthesized for chemical use. The process of discovering and creating new drugs before the adoption of machine learning techniques was a slow and expensive one. Machine learning approaches significantly increased the rate at which potential drugs could be synthesized due to its simulation abilities. By simulating thousands of molecules per second, only the molecules that showed significant potential for its designed purpose would be synthesized and tested which also greatly decreases the overall cost. In addition, the field of machine learning based drug design has expanded greatly in recent years due to the expansion of molecular data and the databases they are contained on.

Before any molecules can be generated using machine learning techniques, the large amount of molecular data needs to be converted into a computer readable format. Determining which representation of the molecular data to use is the first choice to be made when constructing the architecture of the project. Although there are several different molecular representations that a molecule could be converted into before used in training a machine learning algorithm, the SMILE string representation is the best choice when conducting a computational drug design experiment. The SMILE string representation's ability to encode several different chemical features while remaining a 1-dimensional object is its strongest aspect. While other molecular representation methods such as 2D or 3D modeling can encode a significantly larger amount of molecular data, the drawback is the decreased capability of a machine learning algorithm

designed to work with higher dimensional data. The number of computational resources required to run hundreds of thousands of 2D, or 3D objects is exponentially higher than the same amount of 1D strings. By utilizing the SMILE string representation, we can construct a machine learning framework that is able to utilize a large amount of molecular data and subsequently increase its performance. In addition, the expansion of the computational drug design field has allowed projects such as RDKit to be created. This collection of cheminformatics and machine learning software for chemical data analysis includes the ability to change the 1D SMILE string representation into a 2D representation of the molecule as well as a molecular fingerprint representation. By having the ability to change molecules into its higher dimensional representations, we can utilize the strongest aspects of the 1D SMILES representations, while foregoing the many of the downsides of the representation. The utilization of the SMILE string for molecular encoding allows for a greater amount of training for the machine learning frameworks. The increased amount of training consequently allows for a stronger machine learning framework which can produce more potential kinase inhibitors. To measure the success of each generated molecule, there needs to be a measurement method for determining a given molecule's kinase inhibition likelihood.

### **3.2 Developing a Measurement for Kinase Inhibitor Success**

There exists several metrics when determining different chemical or drug-like qualities of any given molecule. However, there is no one metric that encompasses all the qualities of a good kinase inhibitor. The lack of consistent measurement method is a significant problem when designing an experiment based around generating novel SRC Kinase Inhibitors. If there is no way of knowing if a generated molecule is a SRC Kinase Inhibitor without chemical synthesis, the advantage of utilizing machine learning techniques is minimized.

Kinase inhibitors are classified together due to their ability to inhibit either cytosolic or receptor tyrosine kinases; however, each class of kinase inhibitors have slight differences between their chemical features. Analyzing the chemical features of each family of kinase inhibitors for use as the measurement method for generated molecules would be a good idea; however, the differences in chemical features between each molecule is slight. One of the main reasons for this is due to the similar biochemical functions that each kinase family shares.

By not having a concordant metric for analyzing the generated molecules, a novel method of measurement was developed for this experiment. To develop a new metric in determining kinase inhibition likelihood, a reliable architecture had to be used that could combine several parameters of each kinase inhibitor to produce a single score.

### **3.3 Utilizing the Aggregation Techniques of Random Forest Models**

While there are several machine learning techniques that utilize an aggregation feature within its architecture, the decision tree aggregation and feature randomness aspects of the Random Forest model makes it the optimal choice in calculating a measurement metric. One of the main advantages the Random Forest model has over the single decision tree is its use of feature randomness. Feature randomness is the concept that generates a random subset of features which ensures a low correlation amongst the decision trees (Breiman, 2001).

When creating a new measurement metric for kinase inhibitors, the feature randomness aspect of the Random Forest is important. The metric cannot be based on any single chemical feature or group of chemical features because otherwise the metric would already be in widespread use within computational kinase inhibitor discovery experiments. By utilizing the feature randomness aspect of the model, we can ensure that the features used as parameters for the

model will be given equal weighting. The aggregation of decision tree predictions also plays an important role. Aggregating the results from each tree normalizes the feature randomness aspect so no single feature will take a strong priority within the scoring mechanism.

The Random Forest models aggregation of decision trees also allows for a greater understanding of the underlying decisions that constitute the predicted value. Many machine learning techniques have a “black-box” nature to them due to their complex architecture and mathematical foundations. Since the Random Forest model is an aggregation of several decision trees, each decision tree can be analyzed on their own. Every split of the decision tree as well as the feature and threshold determining the split can be assessed. Evaluating the features and thresholds of the decision can help foster a greater understanding of how the machine learning algorithm interprets the differences between each kinase family. This analysis can allow for a more reliable metric to be produced and inspire further study of how these chemical features impact the biochemical function of the kinase inhibitors.

### **3.4 Applying Bayesian Optimization for Generating Molecules**

The Random Forest model’s aggregation of decision trees and feature randomness are not the only aspects of the machine learning technique that encouraged its implementation in our experiment. The simplistic underlying function and reduced resource cost allowed for its application as the scoring function for the Bayesian Optimizer that would generate the molecules.

The Bayesian Optimization process is a hyperparameter tuning operation meant to further improve machine learning functions. Although the original goal of utilizing Bayesian Optimization was to improve the Random Forest model for better Kinase Inhibition Likelihood



predictions, its searching of the latent space inspired its use for the generation of novel molecules. By searching the molecular latent space for local optima, the Bayesian Optimizer discovers a significant amount of information about the encoding of kinase inhibitors. The capitalization of the new knowledge could help us improve similar machine learning based kinase inhibitor discovery experiments, while also demonstrating the potential for creating molecules belonging to a specific kinase family.

Choosing the Bayesian Optimization process as the machine learning technique for generating molecules allows us to generate potential SRC Kinase Inhibitors, while at the same time improving the underlying scoring function and consequently the Kinase Inhibition Likelihood metric. This combination effect makes the Bayesian Optimizer the best learning method for this experiment.

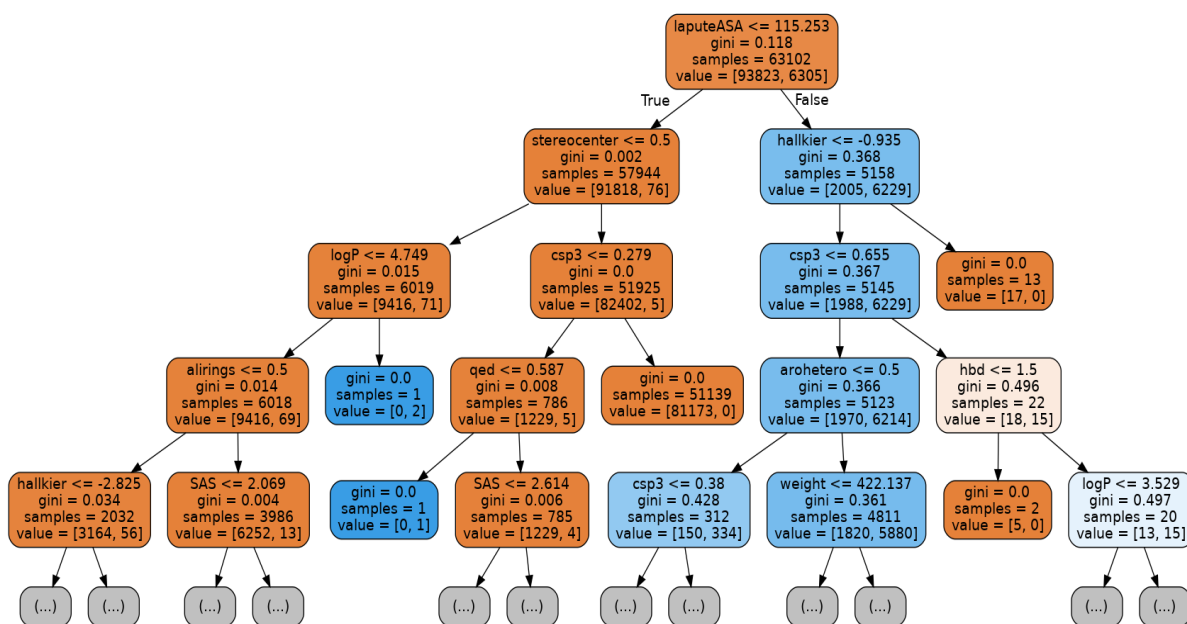
## **4 Measuring Kinase Inhibition Likelihood**

### **4.1 Implementing Random Forest Learning Modeling**

The lack of consistent measurement method for determining a molecule's kinase inhibition potential necessitates the creation of a new metric. In addition, the overlap of the chemical features of the kinase families further emphasized the need for a new measurement. By creating a metric, we could feasibly measure the potential the generated molecules have without having to rely on chemical synthesis or similarity testing methods. To develop this metric, we needed to pick a machine learning architecture that would maximize the accuracy, while minimizing the resource expenses.

The Random Forest model is a very popular classification methodology, especially in applications pertaining to biological and chemical systems. One of the reasons lies in the fact that it can cope with high dimensional data, which is beneficial in accurate predictions and classifications (Boulesteix, Janitza, Kruppa, & König, 2012). This also benefits in being applied to settings containing highly correlated predictors (Boulesteix, Janitza, Kruppa, & König, 2012). The Random Forest model is not based on a particular stochastic model, therefore allowing it to capture nonlinear association patterns between predictors and response (Boulesteix, Janitza, Kruppa, & König, 2012). Ultimately, it does not require the user to specify a certain distribution underlying the data (Boulesteix, Janitza, Kruppa, & König, 2012). All these characteristics provide a huge advantage as most biological and chemical data fall in one or more of these categories, containing high complexity and dimensionality.

The Random Forest model combines the performance and output of several decision trees to reach a single result. By reducing the correlation between decision trees, a Random Forest can increase its overall accuracy by ensuring that the decision trees contained within it will be determining how different features of the dataset interact with each other and their overall importance to the output.

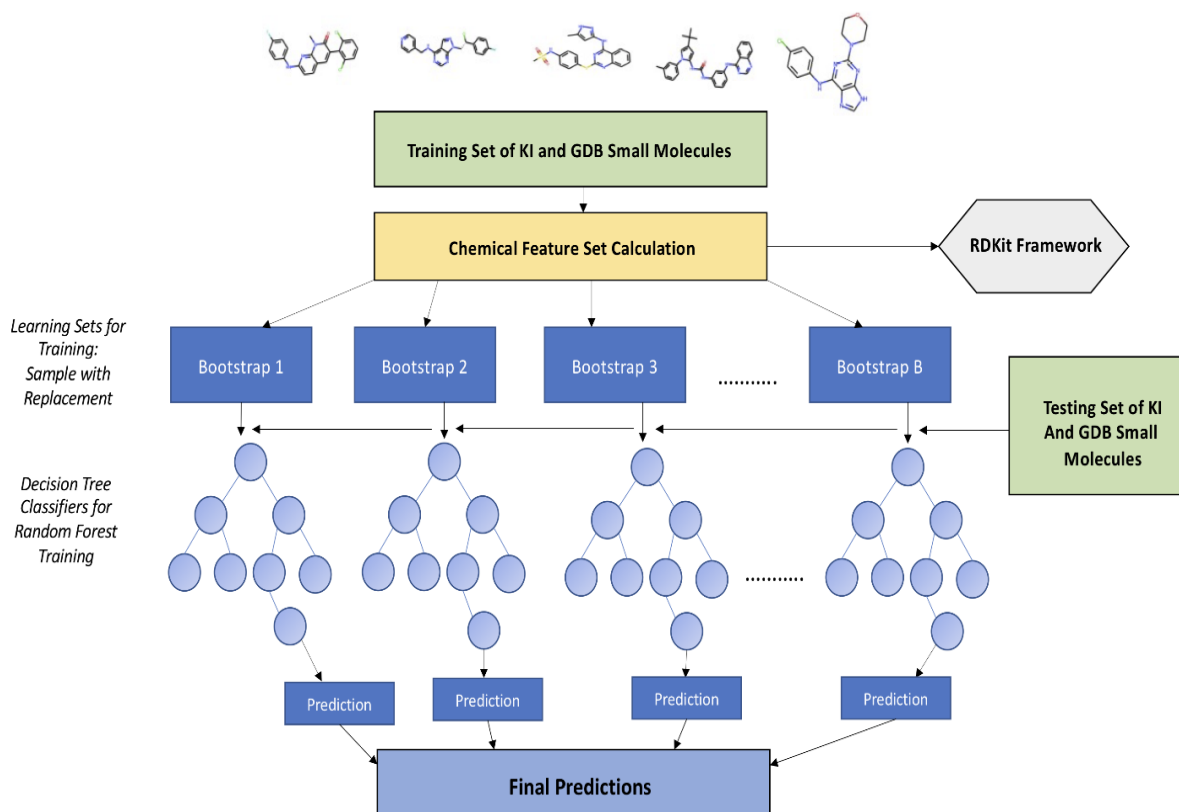


**Figure 3.** The visual representation of one of the Trees in the Binary Chemical Feature-Based Random Forest Model.

In Figure 3 we can visualize a single decision tree within the Random Forest model. Visualizing the tree helps us to understand the features that play a significant role in classifying our binary labels and shows us the role the dataset plays in training the model for accurate classification. Each split within the decision tree occurs based on a certain feature and the value of that feature. At every split, we also see the number of samples for each binary label. These splits are also determined by the GINI impurity, which acts as a measure of purity at each split for classification. For classification tasks, decision trees will attempt to minimize or maximize a particular metric at every node in the tree. This metric is often the GINI impurity coefficient defined as:

$$G = \sum_{i=1}^c p(i) * (1 - p(i))$$

Where  $C$  is the total number of classes and  $p(i)$  is the probability of picking a data point with class  $i$ . To optimize this metric, decision trees take a brute force approach to the problem and observe what the  $G$  would like with every possible split value for every possible column. Once they have calculated the target metric for every possible point, they choose the split value/column pair that optimizes most. This process occurs in every tree of the Random Forest; however, due to the feature randomness of the Random Forest model, the feature that determines each split is different. The model combines all the decision trees into a more complete architecture that allows for a greater accuracy in its predictions.



**Figure 4.** Diagram of Random Forest Model

Figure 4 depicts the architecture of the Random Forest models used during the experiment. The model is initiated with the training set of molecules from all kinase families as well as GDB

molecules. Each molecule within the training set was processed through RDKit to calculate several chemical features. Binary Decision Trees are created, and the chemical attributes were used as parameters to determine the most important features in determining the target variable. Each decision tree makes a prediction on the value of the target variable and the predictions are then aggregated and averaged to get a value between 0 and 1. If there are more than two classes, the predictions are normalized and then averaged to maintain a predicted value between 0 and 1. This would ensure that a target value would still be between 0 and 1, while allowing for multiple classification variables.

## **4.2 Binary and Multiclass Classification**

Due to our experiment containing eleven molecular classifications, ten kinase families and one GDB molecular set, there were two different Random Forest models we could have pursued: the Binary Classification model, and the Multiclass Classification model.

### **4.2.1 Binary Classification Model**

The Binary Classification model would take all kinase families except our target family, the SRC Kinase Family, as well as the GDB molecules and assign them all a target value of 0. The SRC Kinase Family would have a target value of 1. By sacrificing some of its ability to distinguish each of the kinase families from each other within the Random Forest classification, the model emphasizes the overall accuracy of its predictions.

One downside of the Binary Classification model is it is uneven in the weighting of the target variables. The Random Forest model does not utilize the entirety of the training data, but instead takes a proportional sample from each set to get a more even distribution of training data for more accurate decision tree training. The overwhelming amount of GDB molecules used in the training

set causes the model to predict a target value of 0 extremely well but becomes less accurate when predicting a target value of 1. This does not mean that the Binary Classification model cannot predict the SRC Kinase Inhibitors well, it only implies that the model will always be significantly better at identifying the non-SRC Kinase Inhibitors. The downsides of this uneven training data are outweighed by the overall accuracy in the model's predictions.

#### **4.2.2 Multiclass Classification Model**

The Multiclass Classification model would take all kinase families used within training and assign each of them a different target variable value. The goal of this model would be to be able to identify the differences between the different Kinase Families for better isolation techniques during molecular generation. By increasing the amount of classification variables, the Random Forest model consequently sacrifices accuracy during the identification process.

One of the downsides of using the multiclass classification model is the lack of distinction between many of the kinase families. In Figure 2C it is shown that within the latent space all kinase families have significant overlap, making it harder to distinguish between each of the families. Even by implementing a Random Forest model designed to find and predict the differences, the task is still extremely difficult. The number of classes ultimately lowers the model's accuracy in its prediction of the target value, while maintaining a minimal amount of accuracy in distinguishing the different kinase families.

Due to our experiment being focused on attempting to generate a novel molecule with the potential of being an SRC Kinase Inhibitor, the discernment between each of the Kinase Families was less important than the accuracy of the model's ability to identify an SRC Kinase Inhibitor. For this

reason, we decided to pursue the Binary Classification Model to increase the confidence of the values the Random Forest model assigned to each generated molecule.

### **4.3 Training on Latent Space Location and Chemical Features**

#### **4.3.1 Using Latent Space Coordinates**

After deciding on the Binary Classification model over the Multiclass Classification model, the next decision was made on what data the model should be trained on. The first Random Forest model was trained on based on the coordinates of each molecule within the latent space. The model was trained with 25,032 molecules each with 197 feature variables: 196 latent space dimension variables and one classification variable. The initial reasoning behind using the latent space coordinates to train the Random Forest model was due to the differences between the latent space locations of the molecules of all the kinase families and the GDB molecules. By training a model to identify the differences in locations, the goal was to have an accurate model that can discern the differences between a Kinase Inhibitor and a non-Kinase Inhibitor molecule. While there were distinct clusters of molecules within the latent space, the GDB and Kinase Inhibitor clusters, the small overlap that did occur decreased the overall accuracy of the latent space location-based model. In addition, the large number of features that each molecule had when converted to its latent space coordinates also decreased the overall accuracy. With 196 different parameters and 1 target variable, a greater amount of decision trees was necessary for an accurate prediction. Although a greater amount of decision trees theoretically would be more useful, the larger number of trees decreases the importance of each individual tree due to the Random Forest's aggregation of results. The large number of decision trees also significantly increases the resources needed to accurately run the model making it less feasible on large datasets.

### 4.3.2 Chemical Feature-Based Training

Due to the disadvantages of implementing the latent space Random Forest model, we decided to calculate several chemical features of each molecule and train a Random Forest model on those. Since the differences between each kinase family are small, we needed to use many features to increase the accuracy of the Random Forest model. We calculated twenty chemical features for each molecule used during training and testing. The features are: number of rings, the exact molecular weight, the fraction of carbon Sp<sup>3</sup>, the HallKier alpha value, the Labute ASA value, the number of aliphatic carbocycles, the number of aliphatic heterocycles, the number of aliphatic rings, the number of amide bonds, the number of aromatic carbocycles, the number of aromatic heterocycles, the number of aromatic rings, the number of stereocenters, the number of bridgehead atoms, the number of H-bond acceptors, the number of H-bond donors, the QED value, the SAS value, and the logP value. The drug property measures QED, SAS, and logP, pertain to the molecules' chances of not only being successfully synthesized, but their ability to be absorbed and used by the body. Notably, we ensure that the logP adheres to Lipinski's rule of 5 (Lipinski, 2004) for maximum chance of success in synthesis. Apart from the QED value, the SAS value, and the logP value, all the features that were selected were selected at random. One of the main reasons for this random selection is due to the lack of a consistent metric that separates the kinase families. Even though each kinase family is separated based on how each molecule behaves when in the presence of either cytosolic or tyrosine kinase receptors, many molecules share several chemical features across different families. The random selection of features helps encourage a greater accuracy of the model due to the larger range of possible selection metrics. The randomization of features was also utilized for the possibility of discovering a chemical feature that was consistently shared amongst members of the same kinase family. That would allow for not only greater



accuracy in the Random Forest model, but any scoring function created for developing Kinase Inhibitors.

#### **4.4 Kinase Inhibition Likelihood**

The resulting score the Random Forest models output represents the probability or “likelihood” that a molecule can be deemed an SRC Kinase Inhibitor. Values closer to 0 indicate that the molecule has low Kinase Inhibition Likelihood whereas values closer to 1 indicate that the molecules have a high Kinase Inhibition Likelihood. We wanted to create an all-encompassing metric with both structural and drug-like properties to represent kinase inhibition likelihood. Adding on, we did not just want to represent Kinase Inhibition Likelihood in general, but more so in relation to SRC Kinase Inhibitor molecules. Our main objective was to develop novel SRC Kinase Inhibitors, so it was critical for us to create this metric as one of the measures to show how chemically similar our generated molecules from other families are to the SRC Kinase Inhibitor space.

##### **4.4.1 Measuring Accuracy of Kinase Inhibition Likelihood**

For the metric of Kinase Inhibition Likelihood to be a reliable one, it is important to establish the criteria needed to measure its usefulness. The Kinase Inhibition Likelihood score is dependent on the performance of the Random Forest models used during the implementation of the experiment. Thus, utilizing the metrics in determining the accuracy of each Random Forest model would be the most efficient way of measuring the accuracy of the calculated Kinase Inhibition Likelihood score. The three most common metrics used to measure the performance of Random Forest models are precision, recall and the F1-score.

The precision value of a model indicates the rate that the model accurately assigns a positive value to a given result.

$$(1) \textit{ Precision} = \frac{\textit{ True Positive}}{\textit{ True Postive} + \textit{ False Positive}}$$

As seen in equation (1), the precision value takes the true positive value and divides it by all positive values, both true and false. The precision value shows that out of those predicted positive, how many of them are truly positive which makes it a good measure of overall accuracy of the model.

The recall value of a model indicates the rate the model accurately captures the true positive results.

$$(2) \textit{ Recall} = \frac{\textit{ True Positive}}{\textit{ True Postive} + \textit{ False Negative}}$$

Shown in equation (2), the recall value takes the true positive value and divides it by the total positives, or the true positives plus the false negatives. This equation gives the rate at which the model is accurately registering a positive value as being positive.

The F1-score seeks to combine the results of the precision and recall values into a single metric.

$$(3) \textit{ F1} = 2 * \frac{\textit{ Precision} * \textit{ Recall}}{\textit{ Precision} + \textit{ Recall}}$$

One of the benefits of utilizing the F1-score is that it was designed to work on imbalanced data. By averaging the mistakes the model made when predicting a positive value, the score provides a more objective view at the performance of the model.

The most used metric is the overall accuracy metric. It measures the percentage of correct predictions the model has made and divides it by the total amount of predictions made. However, this metric can be misleading when dealing with an imbalanced data set. With all the kinase families being very different in their quantities, our imbalanced data makes the accuracy metric functionally obsolete for our experiment.

The F1-score seems to be the most promising measure to use when calculating the reliability of our model because it combines both the precision and recall values into a single metric. One of the problems is that the precision and recall values each have cases where they are more important. Recall works best when the cost of producing a false negative is high, while precision works best when the cost of producing a false positive is high. Due to our experiment being centered around generating new molecules, the cost of labeling a molecule as having a high SRC Kinase Inhibitor Likelihood when it in fact does not is much higher than labeling a good molecule as having no SRC Kinase Inhibition potential. When conducting drug design experiments, the end goal is always to synthesize a given molecule and test its viability in biological applications; however, the cost of synthesizing a molecule is high. By assigning a molecule that should have no Kinase Inhibition Likelihood with a high score, the chances of that molecule making into the synthesis stage is higher.

When looking only at the metrics that pertain to the calculation of the Kinase Inhibition Likelihood score, the most important value would be the precision value. Though the recall value and the F1-score are still useful and should not be discounted. Due to the importance of the Kinase Inhibition Likelihood score in our experiment, it is better to analyze multiple metrics to maximize the effectiveness of the developed metric.

## 4.5 Analysis of Random Forest Model Performance

### 4.5.1 Latent Space Location Based Results

Even though the latent space based Random Forest models were not utilized during the main molecular generation phase of our experiment, it is useful to include the results of these models to better understand why these models were not implemented.

**Table 1.** Binary Latent Space-Based Random Forest Classification Report Results

	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0</b>	0.98	0.99	0.99	27032
<b>1</b>	0.51	0.33	0.40	679
<b>Accuracy</b>			0.98	27711
<b>Macro Avg</b>	0.75	0.66	0.69	27711
<b>Weighted Avg</b>	0.97	0.98	0.97	27711

The classification report for the binary latent space-based model is shown in Table 1. In the macro averages, the precision score was 0.75, the recall score was 0.66, and the F1-Score was 0.69. In the weighted average, the precision was 0.97, the recall score was 0.98, and the F1-Score was 0.97 (Table 1). One of the commonalities seen in both the binary latent space based Random Forest model and binary chemical-feature based Random Forest model is the large discrepancy in accuracy metrics in predicting a 0-target value and predicting a 1-target value. Although this model has a precision of 0.98, and a recall value of 0.99 when predicting a 0, the precision when predicting a target value of 1 being 0.51 demonstrates that the model is not good at determining if a molecule is an SRC Kinase Inhibitor (Table 1). In addition, the macro average F1-score of 0.69 reinforces this conclusion as it shows its poor performance when given equal samples. The main purpose of utilizing a Random Forest model was to predict the potential any given generated

molecule would have at being an SRC Kinase Inhibitor making the poor performance of this model detrimental to our overall experiment.

**Table 2.** Multiclass Classification Latent Space-Based Random Forest Classification Report Results

	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>ABL1</b>	0.50	0.53	0.51	390
<b>SRC</b>	0.56	0.51	0.53	669
<b>CSF1R</b>	0.46	0.24	0.32	132
<b>EGFR</b>	0.60	0.69	0.64	763
<b>FLT3</b>	0.36	0.16	0.22	234
<b>KDR</b>	0.42	0.59	0.49	888
<b>LCK</b>	0.44	0.30	0.36	289
<b>MAPK10</b>	0.73	0.30	0.43	178
<b>MAPK14</b>	0.66	0.73	0.69	763
<b>MET</b>	0.76	0.55	0.63	429
<b>Accuracy</b>			0.55	4735
<b>Macro Avg</b>	0.55	0.46	0.48	4735
<b>Weighted Avg</b>	0.56	0.55	0.54	4735

The classification report for the multiclass latent space-based model is shown in Table 2. In the macro averages, the precision score was 0.55, the recall score was 0.46, and the F1-Score was 0.48. In the weighted average, the precision was 0.56, the recall score was 0.55, and the F1-Score was 0.54 (Table 2). The poor performance of this model was expected due to the significant overlap of kinase families in the latent space shown in Figure 2C. The only families that have an F1-score above 0.6 are EGFR, MAPK14, and MET which reinforces the idea that the model is not good at distinguishing the families from each other. With a macro average precision value of 0.55, the model has an overall poor performance of identifying the true positive values when given equal samples of each kinase family.

## 4.5.2 Chemical Feature Based Results

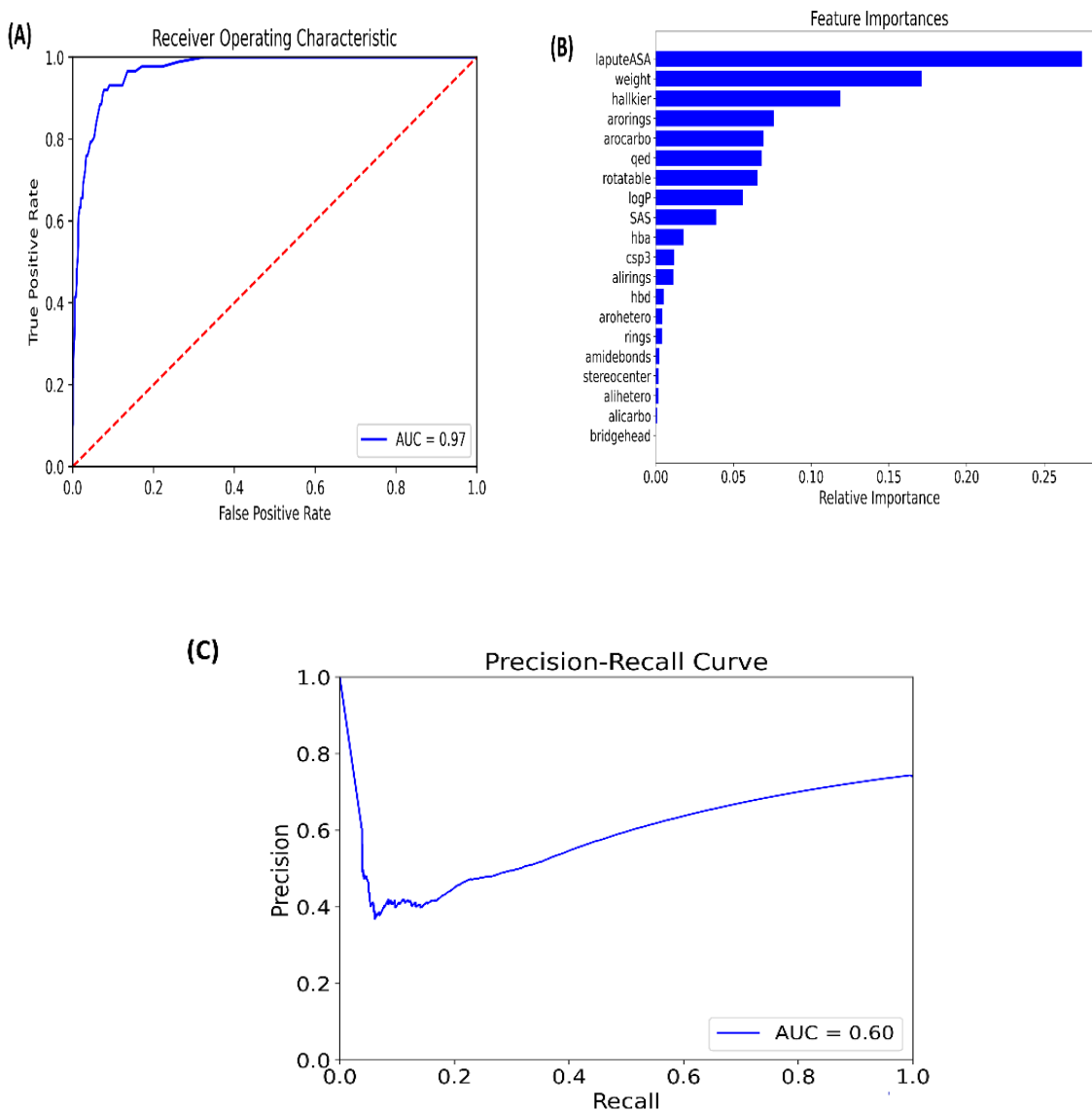
By analyzing the results of the chemical feature-based Random Forest models, we can get a better understanding of the overall accuracy of the Kinase Inhibition Likelihood metric. In addition, the analysis of each chemical feature-based model can give more insight into how to improve the metric for future experiments. Starting the analysis with the classification reports can give more insight into the strengths and weaknesses of each Random Forest model.

**Table 3.** Binary Chemical Feature-Based Random Forest Classification Report Results

	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0</b>	0.99	0.98	0.98	23530
<b>1</b>	0.71	0.86	0.78	1502
<b>Accuracy</b>			0.97	25032
<b>Macro Avg</b>	0.85	0.92	0.88	25032
<b>Weighted Avg</b>	0.97	0.97	0.97	25032

The classification report for the binary chemical feature-based model is shown in Table 3. In the macro averages, the precision score was 0.85, the recall score was 0.92, and the F1-Score was 0.88. In the weighted average, the precision was 0.97, the recall score was 0.97, and the F1-Score was 0.97 (Table 3). The precision, recall, and F1-scores are significantly better when predicting a 0 over predicting a 1; however, this discrepancy is caused by the large difference in support, or number of molecules used for each classification (Table 3). The macro average precision score of 0.85 reinforces the overall good performance of the model because it means that the model was accurate in predicting if a given molecule was an SRC Kinase Inhibitor 85% of the time. For classification models, an accuracy score of 0.85 is extremely strong. In addition, the macro average

recall score of 0.92 validates the excellent performance of the model that the precision value helped establish. All these metrics indicate good classification performance of the model.



**Figure 5.** (A) Receiver Operating Characteristic Curve, (B) Feature Importance Histogram, and (C) Precision-Recall Curve of the Binary Classification Random Forest model

Our Random Forest Model to determine kinase inhibition likelihood classified well, with an overall accuracy of being able to distinguish kinase inhibiting molecules around 97% (Table 3).

The AUC of the model was 0.97, indicating that the coverage of variation within the dataset is around 97% and that the model can distinguish both classes with 97% certainty (Figure 5A). The ROC-AUC or Receiving Operating Characteristic graph in Figure 5A represents the performance measurement of the Kinase Inhibition Likelihood Random Forest classifier that differentiates kinase inhibitor molecules from our small molecule baseline based on a variety of chemical, structural, and drug-like features. The AUC is a metric that measures how well our classifier model can distinguish the classes. Values close to 1 indicate good differentiation of classes by the model, and values close to 0 indicates poor differentiation of classes by the model. The Feature Importance graph in Figure 4B represents the Chemical Feature Importance of the Random Forest model. It will assign scores to input features to a predictive model that indicates the relative importance of each feature when making a prediction. This graph portrays which features fed into the random forest play a huge role in classifying and differentiating the kinase inhibitors from our small molecule baseline. The importance of features is listed in descending order, with the feature having the most importance on top and the feature with the least importance at the end. The top 10 features that contribute the most relative importance to the model's prediction are the labute accessible surface area (labuteASA), weight, HallKier Alpha, the number of aromatic rings, aromaticity, the QED score, number of rotatable bonds, the logP score, the SAS score, and the number of hydrogen bond acceptors (Figure 5B). These are some of the chemical features and attributes that differentiate kinase inhibitors from the GDB small molecule dataset, and in turn, will play a huge role in determining kinase inhibition likelihood of our generated novel compounds. The Precision-Recall graph shown in Figure 5C gives a good idea of how accurate the model really is. Ideally, you would want to have both a high precision value and a high recall value. Although the model begins with a high precision value, it



eventually begins to trade precision for recall to reach an equilibrium between the two metrics. This helps support the idea that our model does perform well and can be trusted to have accurate predictions.

**Table 4.** Multiclass Classification Chemical Feature-Based Random Forest Classification Report Results

	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>ABL1</b>	0.51	0.58	0.55	409
<b>SRC</b>	0.57	0.56	0.56	660
<b>CSF1R</b>	0.69	0.54	0.61	142
<b>EGFR</b>	0.69	0.74	0.71	795
<b>FLT3</b>	0.55	0.46	0.50	194
<b>KDR</b>	0.58	0.59	0.58	916
<b>LCK</b>	0.47	0.41	0.44	313
<b>MAPK10</b>	0.77	0.55	0.64	163
<b>MAPK14</b>	0.75	0.80	0.78	722
<b>MET</b>	0.74	0.72	0.73	421
<b>Accuracy</b>			0.63	4735
<b>Macro Avg</b>	0.63	0.59	0.61	4735
<b>Weighted Avg</b>	0.63	0.63	0.63	4735

The classification report for the multiclass chemical feature-based model is shown in Table 4. In the macro averages, the precision score was 0.63, the recall score was 0.59, and the F1-Score was 0.61. In the weighted average, the precision was 0.63, the recall score was 0.63, and the F1-Score was 0.63 (Table 4). When comparing the classification report of the binary classification model in Table 3 to the one of multiclass classification model in Table 4, the reasoning behind utilizing the binary classification model throughout the experiment becomes clear. Due to the greater number of classification values in the multiclass classification model, the model does not excel for any single class or group of classes. The model had the greatest metric values when predicting kinase

inhibitors from the MAPK14 kinase family, with good metric values when predicting the MET kinase family as well. However, the other kinase families performed poorly in precision values, recall values or the F1-scores. In addition, the macro average F1-score of the multiclass random forest model is 0.61 compared to the 0.88 F1-score of the binary random forest model shown in Table 3. The metrics within Table 4 indicate that the multiclass random forest model performs poorly when distinguishing SRC Kinase Inhibitors which is important for use in the scoring function of our experiment.

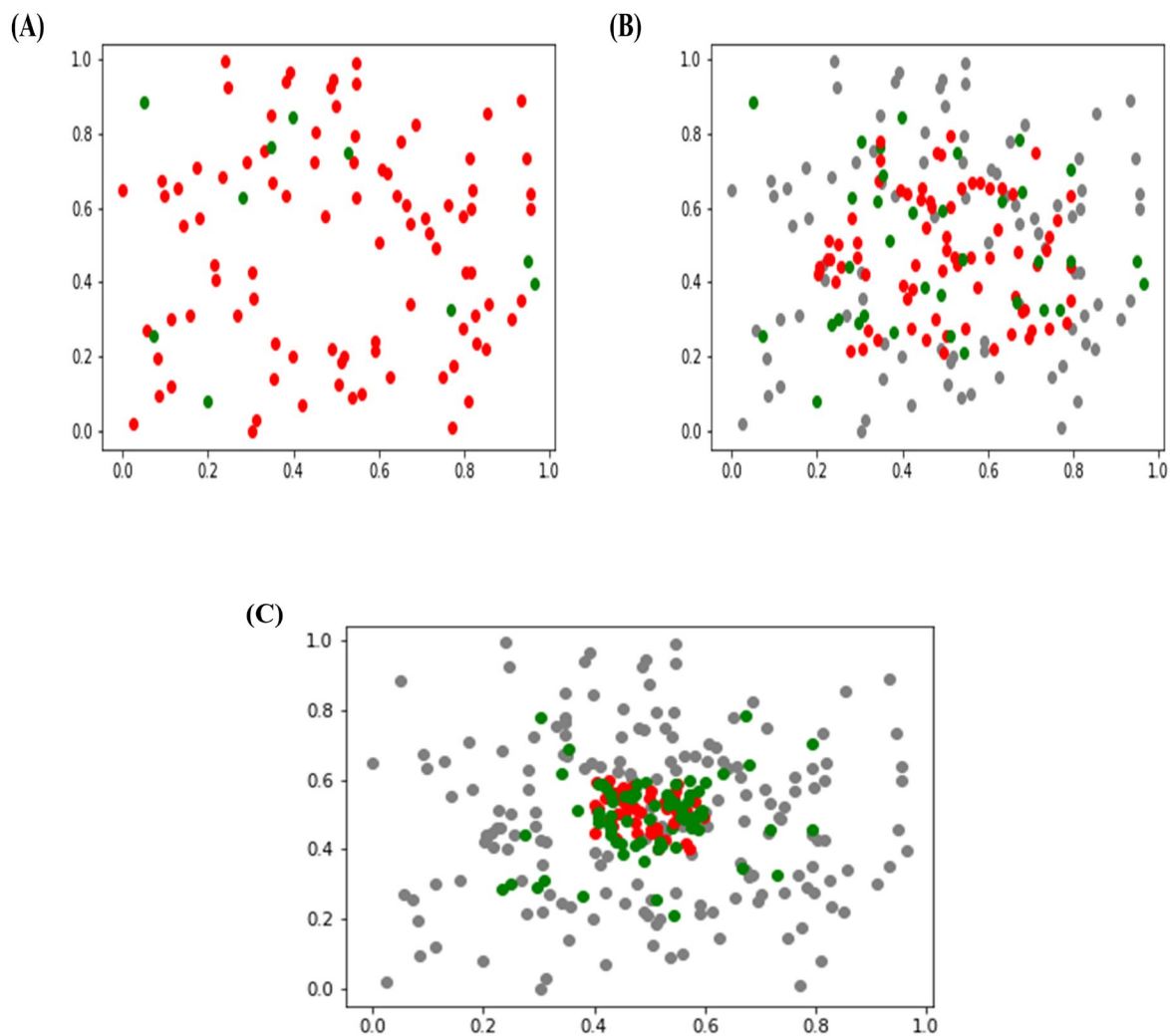
While the initial results shown in Table 4 indicate that the multiclass chemical feature-based Random Forest model is not feasible for our experiment, the ability to distinguish different kinase families is not without its uses. By improving on this model, we can utilize its differentiating ability to possibly classify generated molecules as different kinase inhibitors. Our experiment being focused on the SRC family makes this model unusable but extending the experiment to multiple classes could give greater feasibility to the multiclass models.

Due to the binary chemical-feature based Random Forest model performing well in its predictions of an SRC Kinase Inhibitor, the experiment progresses with the implementation of the binary chemical feature-based Random Forest model into a generative learning method to generate novel molecules.

## 5 Generating Molecules using Bayesian Optimization

### 5.1 Introduction

Bayesian Optimization is a technique for efficient sampling and testing of parameters with respect to a scoring function (Frazier, 2018). Heavily used in the hyperparameter tuning step of machine learning experiments (Snoek, Larochelle, & Adams, 2012), Bayesian Optimization is designed to learn information about how different parameters affect an observed function with minimal sampling. Bayesian Optimization incorporates prior knowledge about the objective function and updates it with random samples drawn from the objective function to better approximate it. Bayesian Optimization approaches use training a gaussian process to approximate a function that represents expected scores with respect to hyperparameters. In addition, Bayesian Optimization uses acquisition functions that helps direct the random sampling to areas with a higher chance of improving the objective function. The acquisition functions trade off exploitation and exploration. Exploitation means sampling where the surrogate model predicts a high objective and exploration means sampling at locations where the prediction uncertainty is high. Both correspond to high acquisition function values and the goal is to maximize the acquisition function to determine the next sampling point. By repeating this process several thousand times, the Bayesian Optimization model can minimize the number of steps necessary for an objective function to reach an optimal accuracy level. Bayesian Optimizers have been shown to perform well in molecular generation applications as well (Jin, Barzilay, & Jaakkola).



**Figure 6.** Figurative Representation of the Exploration, (A) and (B), and Exploitation, (C), Processes of Bayesian Optimization

A figurative representation of the exploration and exploitation processes of Bayesian Optimizers can be shown in Figure 6. Figure 6A represents the initialization of several points within the optimizer's searching process. Towards the beginning, the points are initialized completely randomly to find any local optima that would increase the chances of molecular generation. The green points represent a successful molecule, whereas the red ones indicate a failure. Since the points are created randomly, the success rate of creating a valid molecule is very small and

improves slightly over time during an exploration phase. Figure 6B shows the continuation of Bayesian Optimizer's exploration phase. The optimizer has learned a little more about the space its searching, so it begins to narrow the searching parameters around a potential local optimum. By searching around a potential optimum, the success rate of molecule generation increases. Figure 6C shows an exploitation phase of the optimizer. It discovered a local optimum during an exploration phase and now is creating points mainly around the optimum it found. The specificity of the searching parameters greatly increases the success rate of molecular generation compared to the exploration phases. To find the local optima during an exploration phase, the objective function of the optimizer needs to be as accurate as possible when predicting the scoring value.

## 5.2 Objective Function

The objective function, or "scoring function", of the Bayesian Optimization process is normally predefined, with the goal being to improve the function for use in a separate learning model. For our experiment, our scoring function was based on the Random Forest models that we built to calculate the Kinase Inhibition Likelihood value of the molecules generated through perturbation. The optimization process began with the original Random Forest Binary Classifier that was trained on 196-dimension latent space data of each molecule. However, due to the high computational complexity of the Bayesian Optimization's hyperparameter tuning process, it is usually run on datasets with less than twenty features. The scoring function based on the latent space trained Random Forest Binary Classifier was too expensive in computational resources as well as having too large of a feature set to have a significant enough improvement to the quality of the generated molecules. To better improve our molecules, we needed to change the Random Forest model to be trained on a smaller feature set to get more meaningful optimizations.

### 5.2.1 Training the Objective Function

To accomplish this, the Random Forest model was trained on the twenty chemical features of each inputted molecule. By compressing the feature set to twenty, we allowed the Bayesian Optimization process to make more impactful improvements in the hyperparameter tuning process.

To train the Bayesian Optimization model, we needed to train the model on similar molecules we used during the Random Forest Binary Classifier model. We used the ABL1 and GDB molecules that were used for training the Random Forest model and converted all the molecules into their corresponding 196-dimension latent space locations. The SRC molecules were split into training and testing sets with a split of 0.67/0.33. The testing set was used to validate the results from the Bayesian Optimizer, and the training set was saved for future use. After encoding all the molecules into their latent space representations, the scoring function was established as the same function used to train the Random Forest Binary Classifier based on the chemical features of each molecule. Once the scoring function was in place, the Bayesian Optimizer function was created with 7,000 initial points, and 1,500 iterations, or sampling points. The initial points are used by the Bayesian Optimizer to sample random points with no regard for exploitation, record their reward functions, and warm up your prior distribution over the latent space. This is necessary so that the breadth of the parameter space can be searched without getting stuck in local optima due to scoring bias. Many pockets of high potential scoring values are found during this step to force the model to explore more areas of the search space rather than becoming too focused on one area. There is typically a relationship between the number of parameters to be optimized and the amount of exploration steps that need to be performed for sufficient training.

After all the exploration steps have been completed, the optimizer begins to exploit this information.

### **5.3 Bias within the Optimization Function**

In addition to training the objective function, the initialization of the optimizer could play a significant role in the quality of the results. Creating a Bayesian Optimization function can be done in two different ways: as an Unbiased Optimization Function, or as a Biased Optimization Function. The main difference between an Unbiased Optimization Function and a Biased Optimization Function is the probing a Biased Optimization Function performs before initiating the iteration training. Probing is the process by which the optimizer takes known data points within the exploration phase to get more targeted information to exploit during the exploitation phase. By implementing a bias in the Bayesian Optimization process, the molecules generated from the model become a lot more in line with the target SRC kinase family; however, the model does lose the pure gaussian process of hyperparameter tuning because it is starting off with known molecules within the same family. For our biased optimization function, we did create it with 7,000 initial points as well as 1,500 iterations; however, we used the training set of the SRC Kinase Inhibitors to probe the model. When probing the model, the Bayesian Optimizer explores the latent space areas of the molecules in the training data before initiating any of the 7,000 points which allows the model to be a lot more accurate in its search for the optimal parameters.

#### **5.3.1 Utilizing a Multiclassification Random Forest Model as the Scoring Function**

In addition to the Binary Bayesian Optimizer listed above, another method that was implemented was utilizing the multiclassification Random Forest model as the scoring function for a potential

Multiclass Bayesian Optimizer. There were two different Multiclass Bayesian Optimizers created: an unbiased one and a biased one. The training for the multiple class Bayesian Optimizers were slightly different to the training of the binary Bayesian Optimizers. Unlike the binary optimizers, the multiple class optimizers were each created with 2,500 initial points and 500 iterations. The biased multiple class Bayesian Optimizer was probed using the same SRC Kinase Inhibitors that were used to probe the binary Bayesian Optimizer. Unfortunately, after each of the models finished running, the multiple class Bayesian Optimizers did not produce any valid molecules. I believe the main reason for this is due to the probabilistic nature of the Bayesian Optimizer. By introducing more than two classes for the Bayesian Optimizer to optimize for, the model could not accurately tune the hyperparameters to match the predictions of the underlying multiclassification random forest model. Since each of the multiclassification Bayesian Optimizers failed to produce any molecules, they will be omitted from the results and discussion portion of this chapter.

## **5.4 Similarity Testing**

After the Bayesian Optimizers generate novel molecules, there needs to be a way to analyze them to filter the molecules with higher potential of being an SRC Kinase Inhibitor. While the calculated Kinase Inhibition Likelihood is the metric developed for this purpose, the high cost of producing a false positive necessitates an additional metric to be used in conjunction with the calculated Kinase Inhibition Likelihood.

We utilized Tanimoto Similarity Testing as the additional metric to further evaluate these generated molecules to understand which molecules would have the highest potential of being SRC Kinase Inhibitors. The Tanimoto similarity coefficient is a metric that compares the molecular similarity of two compounds using Morgan fingerprint analysis. Molecules with



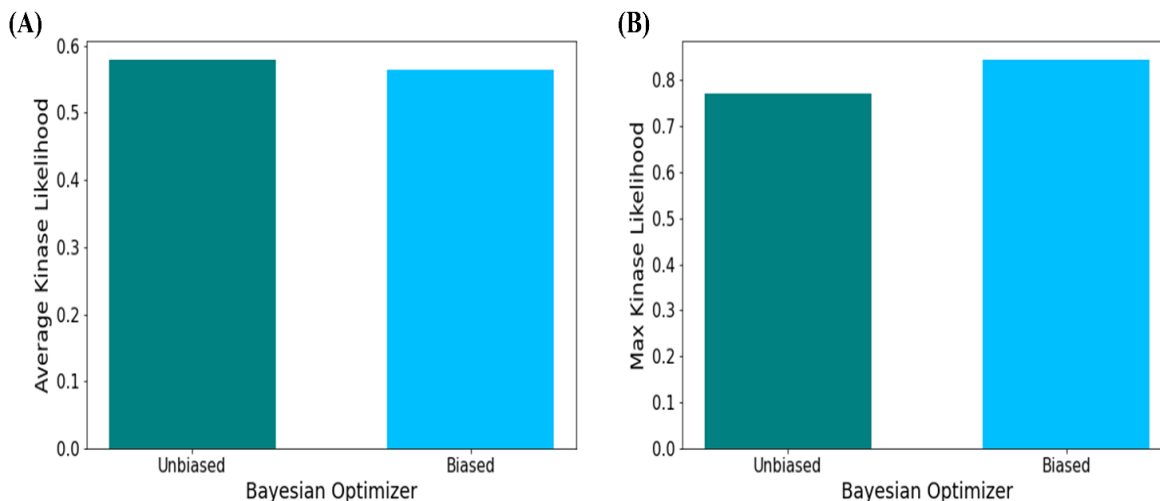
Tanimoto coefficient values that are above 0.75 are considered to have high similarity with the comparison molecule whereas values lower than that value are considered to have lower similarity. To calculate the similarity, we obtain the Morgan fingerprints of each molecule using RDKit and compute the Tanimoto coefficient between it and all the molecules within the SRC Kinase Inhibitor testing set. We then return the highest similarity value that is obtained and the SRC Kinase Inhibitor that the molecule is closest in similarity to.

By using the similarity testing aspect of our experiment, the value of our calculated Kinase Inhibition Likelihood metric could be either validated or rejected. The extra layer of testing gives a stronger evidential background for the results of the experiment. While the calculated Tanimoto coefficient was not used as the sole determining factor for a molecule's kinase inhibition, the coefficient was determined to have a greater significance than the calculated Kinase Inhibition Likelihood metric.

## **5.5 Analysis of Generated Molecules**

To evaluate the performance of the different Bayesian Optimizers, we investigated the relationship between each optimizer and the molecules they produced in relation to the Kinase Inhibition Likelihood score, the calculated chemical features, and the Tanimoto similarity scores to known SRC kinase inhibitors. When looking at the quantity of molecules produced by each Optimizer, the biased optimizer produced 589 molecular smile strings and the unbiased optimizer produced 440 molecular smile strings. Out of these molecules, several were discarded due to the inability to convert the produced smile strings into valid molecules. After conversion from the molecular smile strings to molecular image, the biased optimizer maintained 492 valid molecules, and the unbiased optimizer maintained 390 valid molecules.

## 5.5.1 Kinase Inhibition Likelihood Evaluation



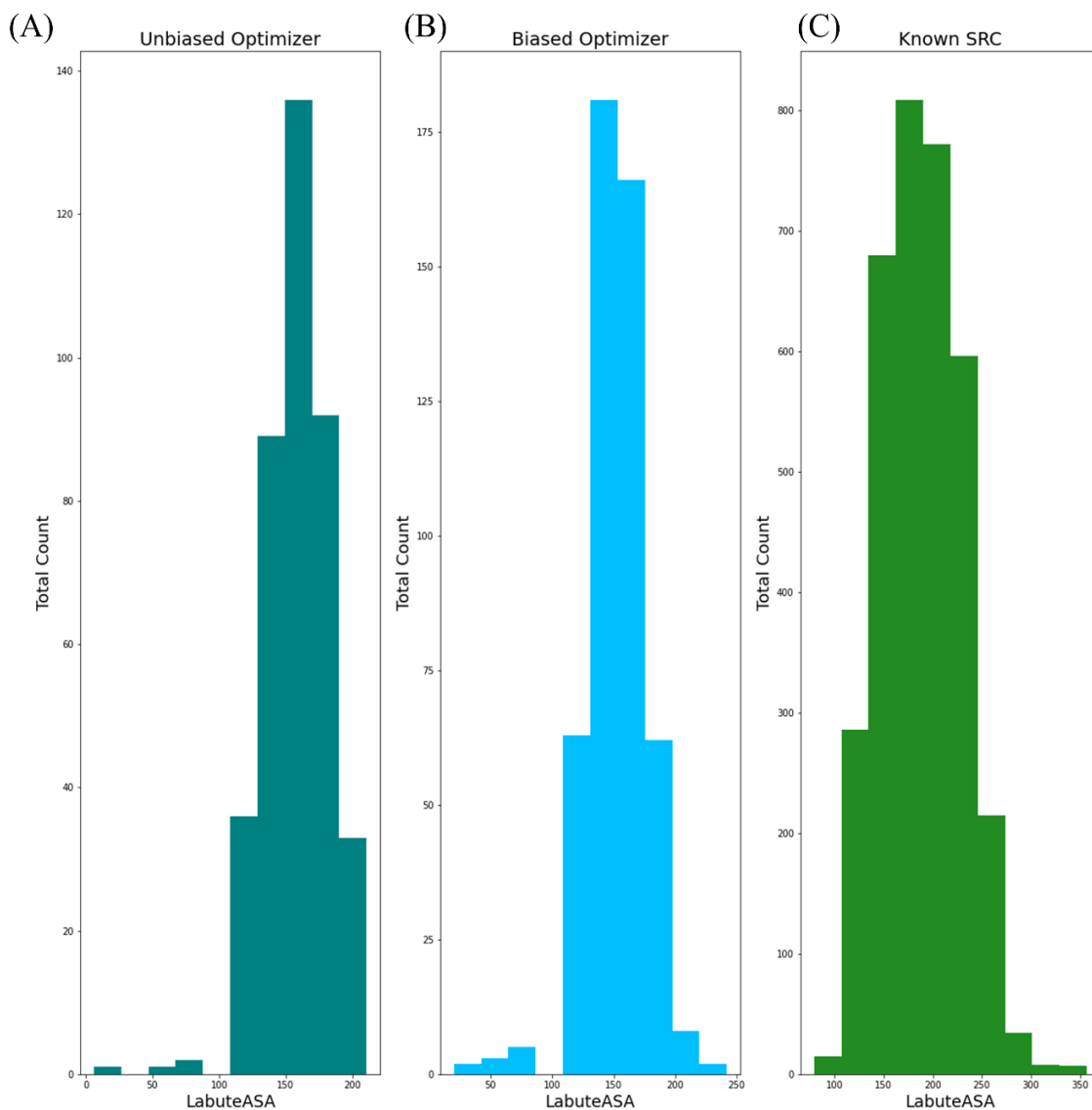
**Figure 7.** Bar Graphs of (A) the average Kinase Inhibition Likelihood scores of all generated molecules, (B) the max Kinase Inhibition Likelihood scores of all generated molecules from the Unbiased and Biased Bayesian Optimizers

Due to the random nature of the Bayesian Optimizer a threshold calculated Kinase Inhibition Likelihood score of 0.5 was used to as the baseline for a generated molecule to have a higher Kinase Inhibition Likelihood. Out of the valid molecules produced from each Optimizer, 153 molecules out of the original 492 molecules produced, or 31.10%, from the Biased Optimizer had a calculated Kinase Inhibition Likelihood value greater than 0.5. The Unbiased Optimizer maintained 145 of its original 390 valid molecules produced, or 37.18%, with a calculated Kinase Inhibition Likelihood value greater than 0.5. When analyzing the molecules with a calculated Kinase Inhibition Likelihood score greater than the 0.5 threshold, the Unbiased Optimizer had a higher average calculated Kinase Inhibition Likelihood of 0.5783 compared to an average of 0.5639 for the molecules generated by the Biased Bayesian Optimizer (Figure 7A). The molecule with the highest calculated Kinase Inhibition Likelihood score was produced by the Biased Bayesian Optimizer with a score of 0.8425. The molecule with the highest calculated Kinase

Inhibition Likelihood score produced by the Unbiased Optimizer had a score of 0.7693 (Figure 7B).

### **5.5.2 Chemical Feature Evaluation**

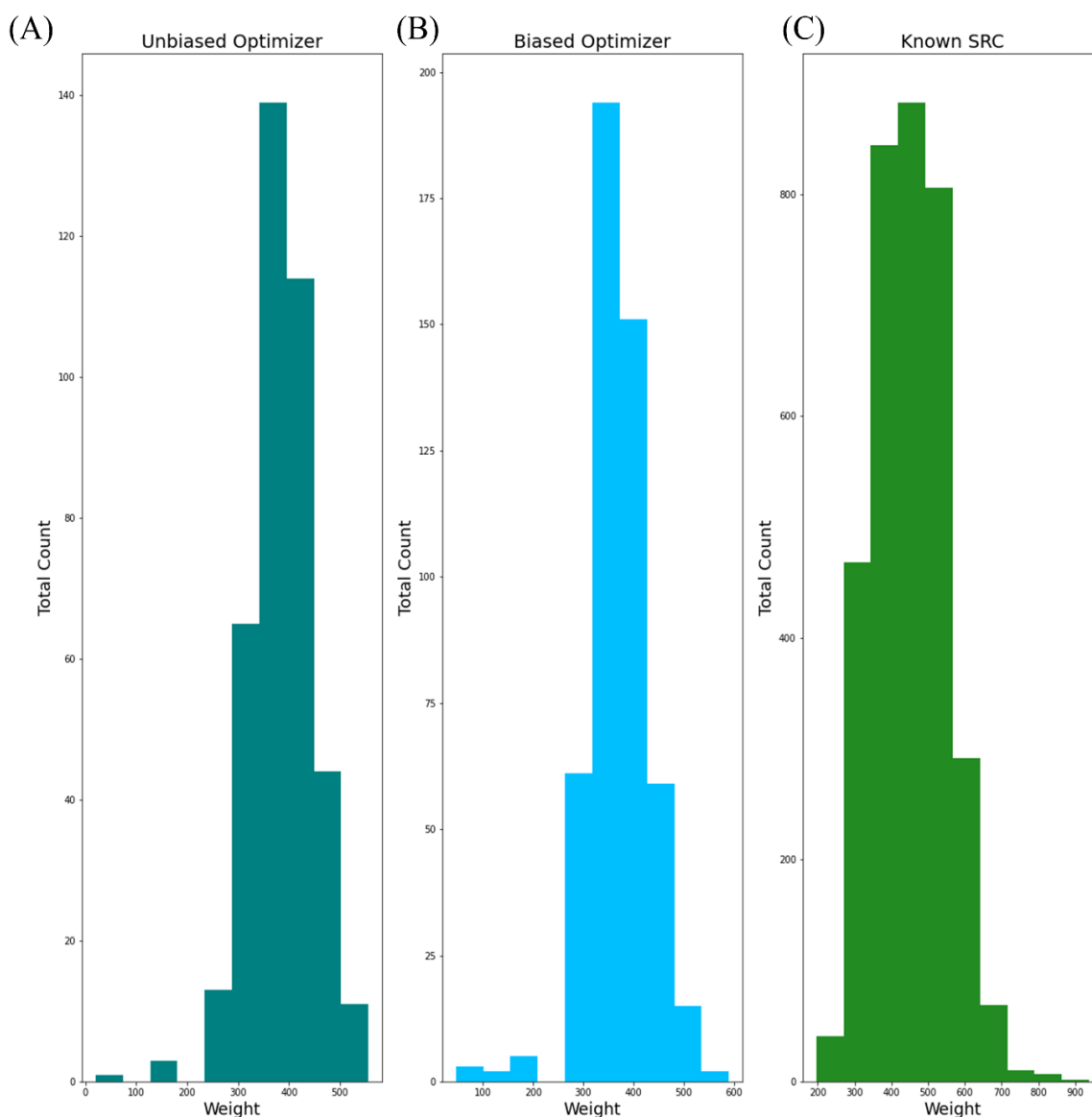
After analyzing the calculated Kinase Inhibition Likelihood values of the molecules produced from each Bayesian Optimizer, the chemical features of the generated molecules were compared to the chemical features of the known SRC Kinase Inhibitors. Since the scoring function of the Bayesian Optimizers were based on the binary Random Forest model, we used the top five features determined by the feature importance graph from (Figure 5B) and compared the results from two Bayesian Optimizers to the same chemical features of the known SRC Kinase Inhibitors. The strategy was to use a histogram to analyze the distribution of each of the chemical features to determine chemical similarity to the known SRC Kinase Inhibitors.



**Figure 8.** Histograms of the distribution of the LabuteASA values from the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors

According to the Binary Random Forest model, the feature with the highest importance in determining whether a molecule is an SRC Kinase Inhibitor was the LabuteASA value. The LabuteASA value, or the accessible surface area (ASA), is the area of a molecule that is accessible to the solvent (e.g., water). The ASA is calculated by summing the surface area of each atom in a

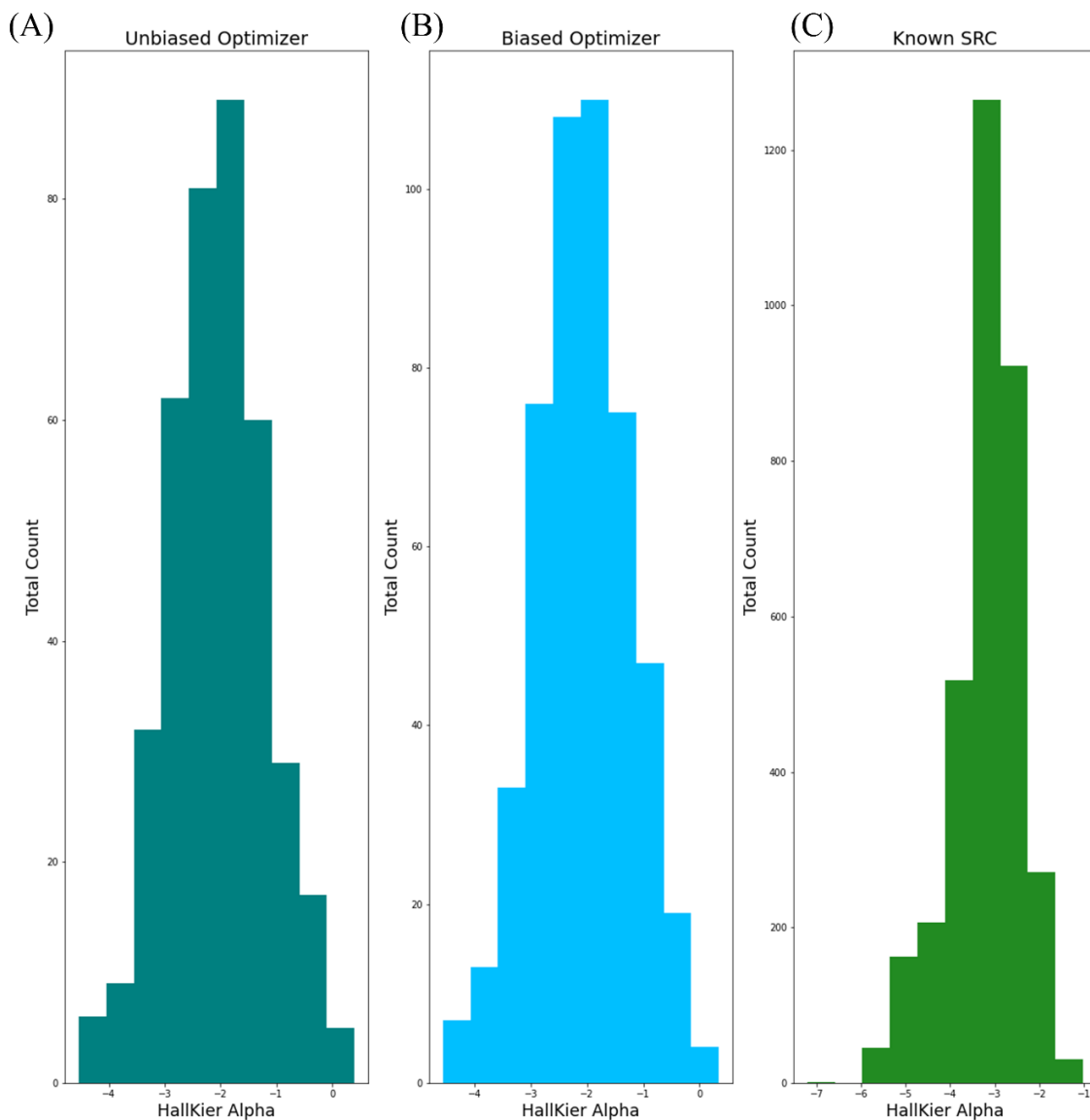
molecule (Hemmerich, Troger, Fuzi, & Ecker, 2020). The Biased and Unbiased Bayesian Optimizers had a similar distribution of the LabuteASA values of their generated molecules, with the highest concentrations between 150 and 200. The optimizers did have a similar overall distribution of the LabuteASA values compared to the known SRC Kinase Inhibitors; however, the highest concentration of molecules for the known SRC Kinase Inhibitors was at a LabuteASA value around 200 (Figure 8). Even though the average LabuteASA values were different between the Bayesian Optimizers and the known SRC Kinase Inhibitors, their similar distributions of molecules reinforce the importance that the feature had when calculated through the hyperparameter tuning of the Bayesian Optimization function.



**Figure 9.** Histograms of the distribution of the molecular weights of the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors

The molecular weight of each molecule was determined to have next highest importance in determining the kinase likelihood of a generated molecule. The Bayesian Optimizers had similar distributions of the molecular weights of the generated molecules as well as strong similarity in the areas with the highest concentration of molecules being around the molecular weight of 400.

The Bayesian Optimizers once again had a similar distribution pattern compared to the known SRC Kinase Inhibitors; however, the known SRC Kinase Inhibitors had the strongest concentration of molecules around a molecular weight of 500 (Figure 9). The Bayesian Optimizers also produced more molecules outside the range of the known SRC Kinase Inhibitors. The known SRC Kinase Inhibitors had a general range from 200 to 700, with some molecules having a weight greater than 700, the Bayesian Optimizers produced most molecules within the weight range of 250 to 550, with some molecules generated having weights around 100. The narrower margin for the weights could suggest that the scoring function behind the Bayesian Optimizers is reducing the calculated Kinase Inhibition Likelihood score of the generated molecules after some percentage of the standard deviation of the mean weight of the known SRC Kinase Inhibitors.

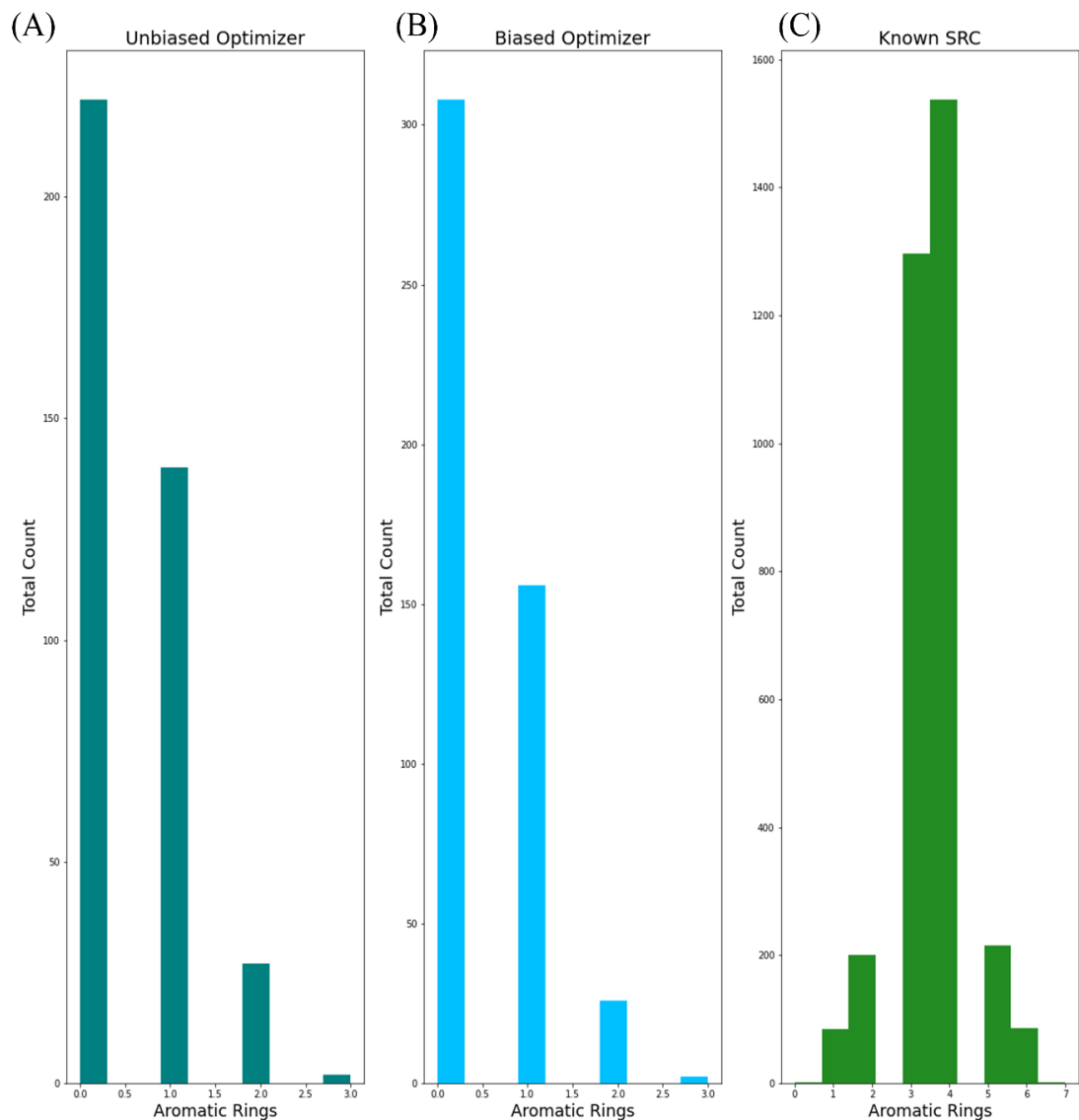


**Figure 10.** Histograms of the distribution of the HallKier Alpha values of the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors

The HallKier Alpha value had the next highest importance value according to the Binary Random Forest. The HallKier Alpha value represents the connectivity index of a given molecule (Kier & Hall, 2002). It utilizes the count of neighboring atoms bonded to an atom in the hydrogen-suppressed graph to determine molecular connectivity (Kier & Hall, 2002). The Bayesian



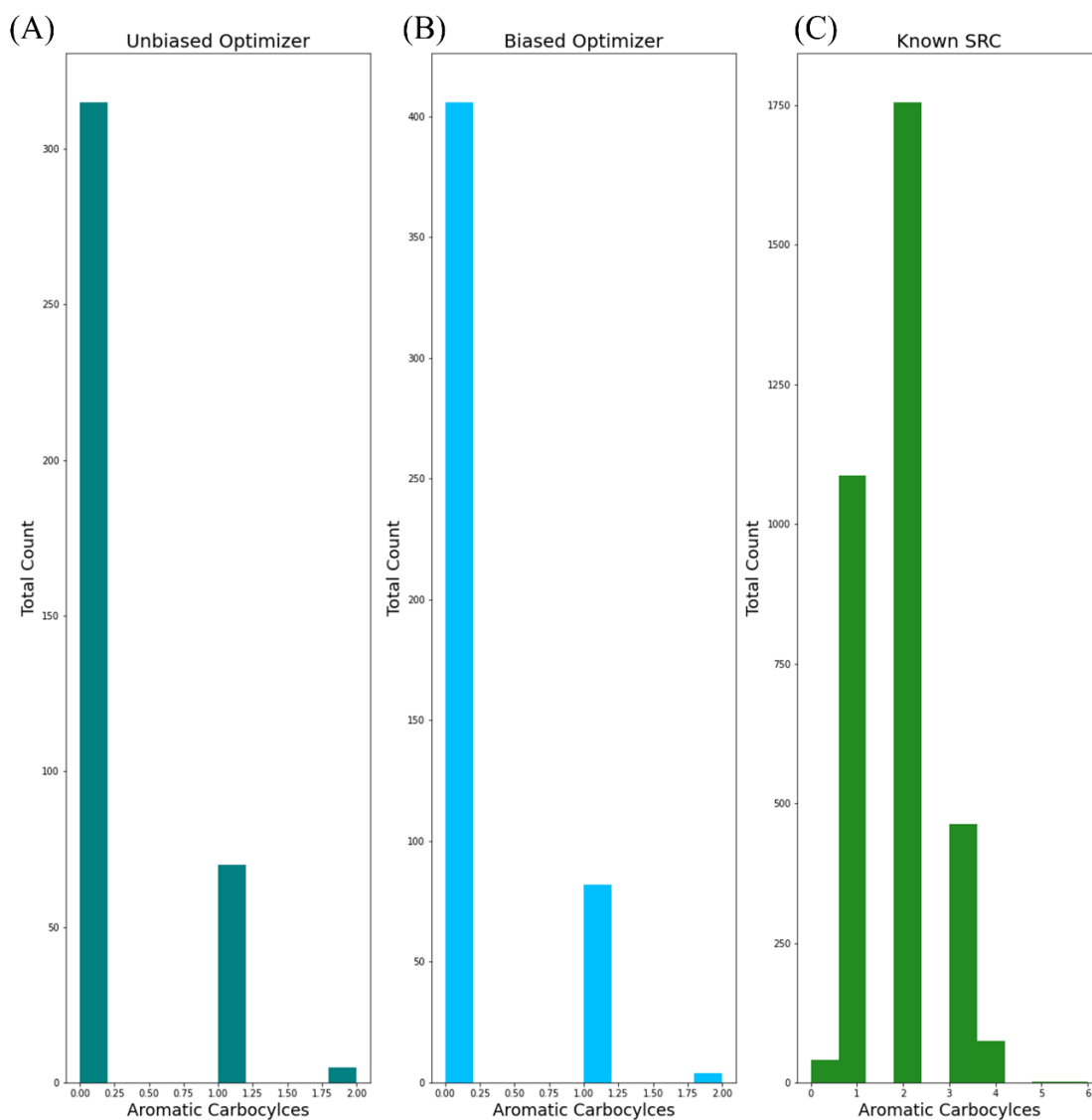
Optimizers performed similarly to each other regarding the HallKier Alpha values of the generated molecules, both in distribution of values and area of the highest concentration of around -2; however, the optimizers did have different distributions and calculated areas of highest concentrations of the known SRC Kinase Inhibitors, which for the known inhibitors was around -3 (Figure 10). The distributions of the values of the molecules generated through the Bayesian Optimizers had a wider range and was more symmetrical in its overall distribution of values, whereas, the known SRC Kinase Inhibitors had significant concentrations of molecules between -3.5 and -2.5 but was not evenly distributed for the rest of the values within the calculated ranges (Figure 10). The similarities between shape of distribution and neighboring areas of highest concentrations imply that the HallKier Alpha parameter was a priority feature for the Bayesian Optimizer to tune. The differences of in these values could suggest that the Bayesian Optimizer weighted the tuning of the HallKier Alpha parameter incorrectly, or that the optimizer needed more iterations to find the global optima of the feature.



**Figure 11.** Histograms of the distribution of the number of aromatic rings of the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors

After the HallKier Alpha scores, the next feature with the highest importance is the number of aromatic rings. Within the molecular structure, aromatic rings are hydrocarbons which contain benzene, or some other related ring structure (Polêto, et al., 2018). Aromatic rings are extensively used in drugs due to their well-known synthetic and modification paths (Polêto, et al., 2018). The

Bayesian Optimizers performed nearly identical in their distribution of values, as well as the highest concentrations. Most of the molecules generated from the Bayesian Optimizers had either 0 or 1 aromatic rings with more molecules having 0 (Figure 11). The Bayesian Optimizers greatly differ from the known SRC Kinase Inhibitors regarding the number of aromatic rings. With the known SRC Kinase Inhibitors, the general distribution of molecules is even between values of 1 and 6 for the number of aromatic rings, and the highest concentration of molecules has either 3 or 4 aromatic rings (Figure 11). Neither of the Bayesian Optimizers produced any valid molecules with the number of aromatic rings being greater than 3, which suggests that the hyperparameter tuning calculations of both optimizers underperformed on the number of aromatic rings feature. In addition to the underperformance of tuning of the aromatic ring feature, the lack of valid molecules could imply the need for greater iterations of each Bayesian Optimizer.



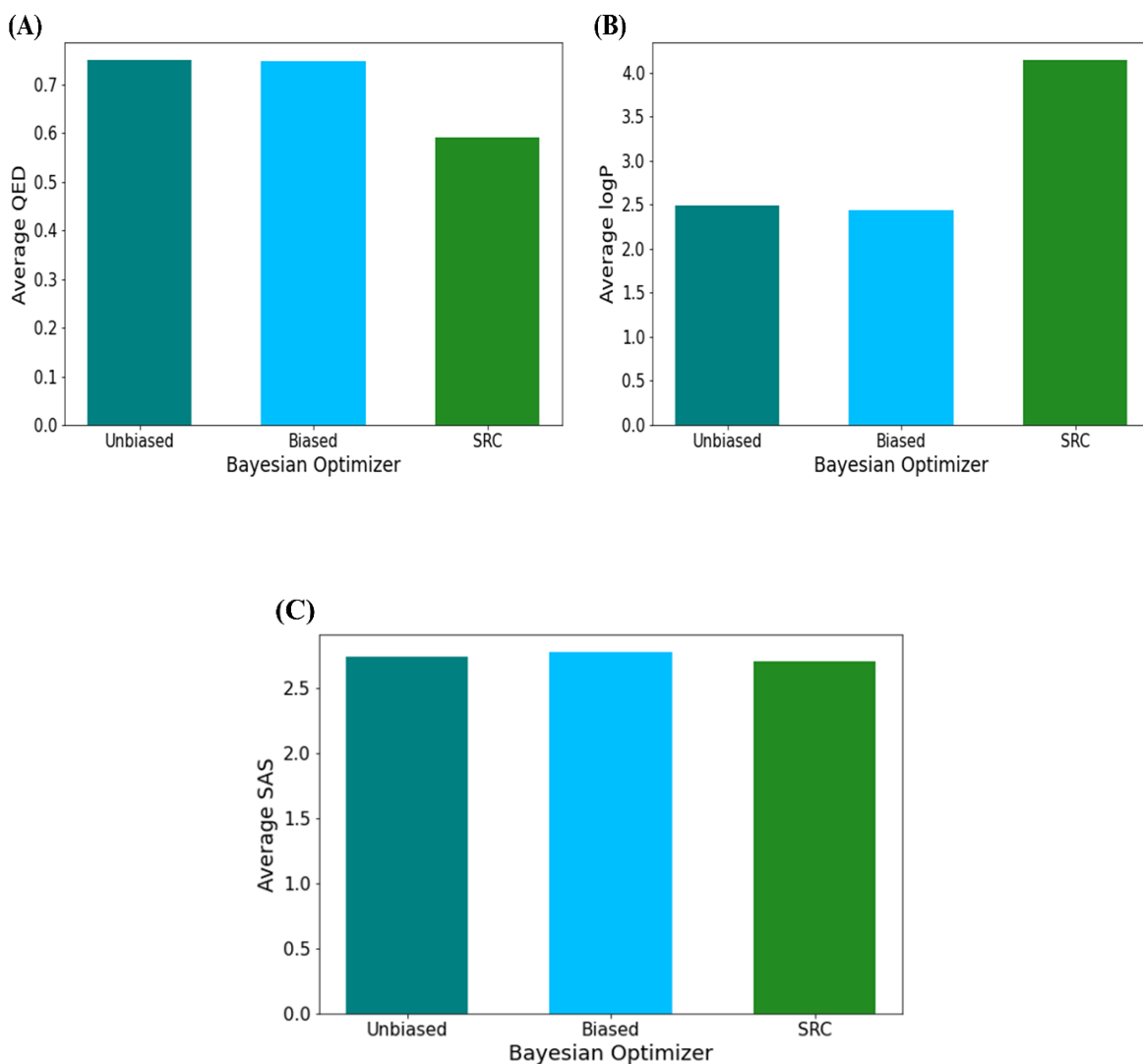
**Figure 12.** Histograms of the distribution of the number of aromatic carbocycles of the generated molecules of (A) the Unbiased Bayesian Optimizer, (B) the Biased Bayesian Optimizer, and (C) the set of Known SRC Kinase Inhibitors

The last feature that was compared between the Bayesian Optimizers and the known SRC Kinase Inhibitors was the number of aromatic carbocycles. Carbocycles are organic molecules that contain one or more rings. Aromatic carbocycles are carbocycles where the rings of the molecule are benzene rings (Hu, Li, Han, Min, & Li, 2020). The Bayesian Optimizers performed almost

identical in its overall distribution, as well as its area of highest concentration. Most of the molecules generated from the Bayesian Optimizer were calculated to have 0 aromatic carbocycles, with 1 being the next highest concentration (Figure 12). The known SRC Kinase Inhibitors had an even distribution of molecules between the values of 0 and 4 aromatic carbocycles; however, the number of aromatic carbocycles with the highest concentration of molecules was 2 (Figure 12). Although the Bayesian Optimizers underperformed in generating molecules with a similar number of aromatic carbocycles to the known SRC Kinase Inhibitors, the optimizers were more accurate in its calculations in comparison to the number of aromatic rings. The lack of similarity between the number of aromatic rings and number of aromatic carbocycles in the molecules generated from the Bayesian Optimizers and the known SRC Kinase Inhibitors could suggest that Bayesian Optimizers could not determine the best way to modify the latent space coordinates to affect these features in significant way. The extreme differences between the generated molecules and the known SRC Kinase Inhibitors also imply that simply modifying the iteration or initial point values of the Bayesian Optimizers would not be enough to improve the aromatic carbocycles of the generated molecules.

### **5.5.3 QED, logP, and SAS Evaluations**

After analyzing the chemical features determined to be the most important by the chemical feature based Random Forest model, the calculations of the QED, logP, and SAS values were evaluated in the comparison to the known SRC Kinase Inhibitors. One of the main reasons for this is because of their importance in chemical synthesis and cellular absorption.

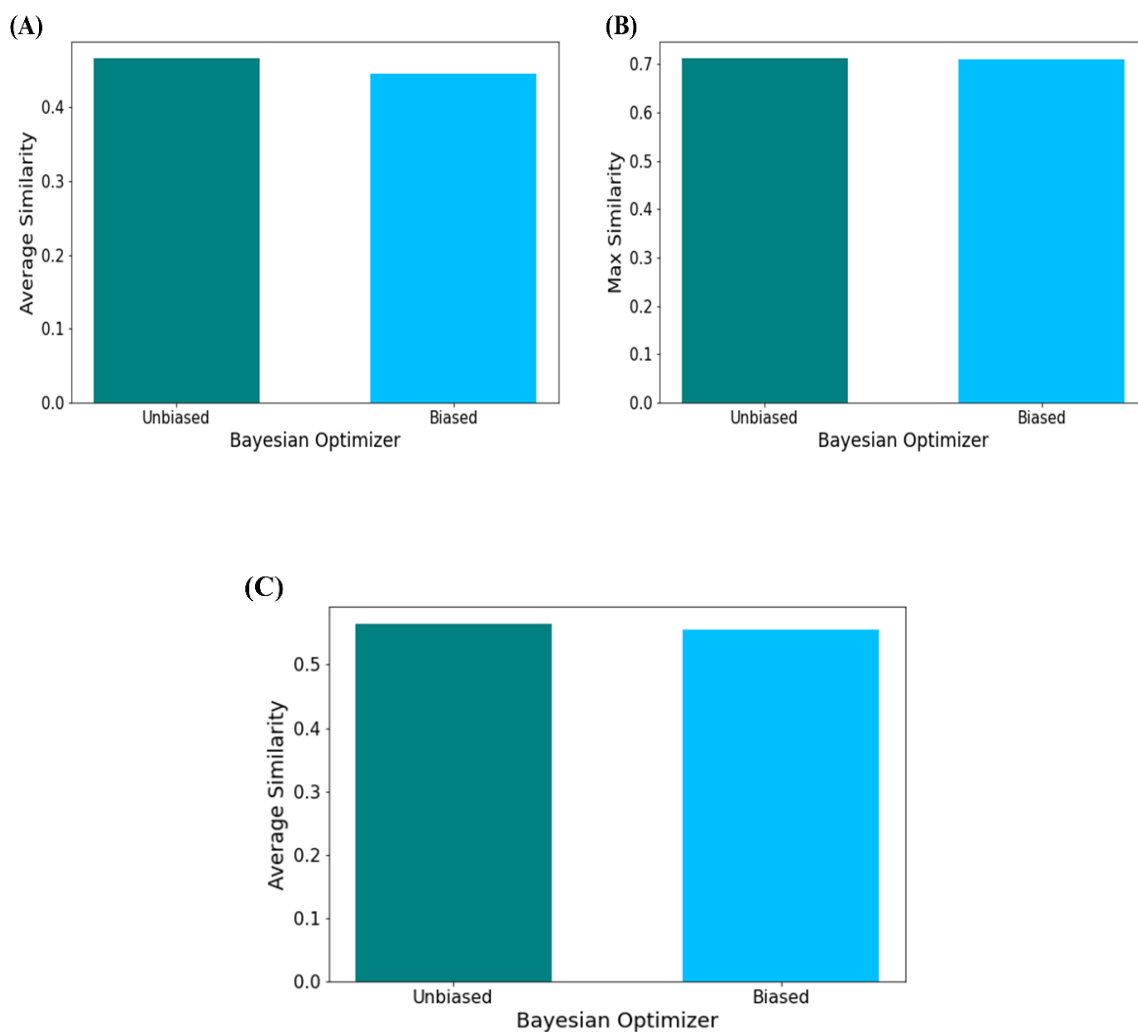


**Figure 13.** Bar Graphs of (A) the average QED scores, (B) the average logP scores, and (C) the average SAS scores of the molecules generated from the Biased and Unbiased Bayesian Optimizer, in comparison to the known SRC Kinase Inhibitors

When determining the performance of the Bayesian Optimizers in relation to the chemical feature values of QED, logP, and SAS, the generated molecules from the optimizers had similar average SAS scores compares to the known SRC Kinase Inhibitors but had significant differences in the average QED and logP scores. The average QED scores for the Unbiased and Biased Bayesian Optimizers' generated molecules were 0.7499 and 0.7486 respectively, in comparison to the

known SRC Kinase Inhibitors' average QED score of 0.5908 (Figure 13A). The average logP scores for the Unbiased and Biased Bayesian Optimizers' generated molecules were 2.488 and 2.439 respectively, in comparison to the known SRC Kinase Inhibitors' average logP score of 4.137 (Figure 13B). The average SAS scores for the Unbiased and Biased Bayesian Optimizers' generated molecules were 2.742 and 2.772 respectively, in comparison to the known SRC Kinase Inhibitors' average SAS score of 2.706 (Figure 13C). The general similarity of the scores of the generated molecules in comparison to the known SRC Kinase Inhibitors suggest that the metrics are being tuned as a part of the Bayesian Optimizers' hyperparameter tuning process. While there are differences between the generated molecules and the known SRC Kinase Inhibitors when analyzing the QED and logP scores, the scores imply that the molecules produced by the Bayesian Optimizers would be synthesizable and/or absorbable even with lower similarity metrics in other chemical features. This suggests that improving the Bayesian Optimizers to achieve higher quality molecules would be overtly beneficial as the higher quality molecules would retain the practicality of their synthesizability or bodily absorbability. To ensure that higher quality molecules would be produced, the molecules' similarity metrics would also need to be tested against known SRC Kinase Inhibitors to ensure the feasibility of their relation to kinase families.

#### **5.5.4 Tanimoto Similarity Testing Evaluation**



**Figure 14.** Bar Graphs of (A) the average similarity scores of all generated molecules, (B) the max similarity scores of all generated molecules, and (C) the average similarity scores of the generated molecules with a calculated Kinase Inhibition Likelihood score above 0.5 from the Biased and Unbiased Bayesian Optimizers

To evaluate the similarity testing metrics, we investigated the performance of each of the Bayesian Optimizers based on average similarity scores of the generated molecules, as well as the maximum similarity score that each model produced. When analyzing all the molecules generated from each Bayesian Optimizer, the average Tanimoto similarity scores for the Unbiased and Biased Bayesian



Optimizers were 0.4656 and 0.4446 respectively (Figure 14). The maximum Tanimoto similarity scores for the Unbiased and Biased Bayesian Optimizers were 0.7115 and 0.7091 respectively (Figure 14). After removing all the generated molecules that had a calculated Kinase Inhibition Likelihood score below the baseline value of 0.5, the average similarity scores of the Unbiased and Biased Bayesian Optimizers were raised to 0.5635 and 0.5551 respectively (Figure 14). The maximum Tanimoto similarity scores did not change after removing all the generated molecules with a Kinase Inhibition Likelihood score below the baseline.

The similarity scores produced by each Bayesian Optimizer indicate that the Optimizers require significant improvement before the desired quality of molecules are produced. In similar experiments, the threshold for determining the feasibility of generated molecules was a Tanimoto similarity score above 0.75. Neither Bayesian Optimizers produced a single molecule above the high similarity threshold. Although the cause of the generation of lower similarity scoring molecules could be the scoring function, it would be impossible to come to that conclusion at this time. This does not mean the experiment was not a success as there were molecules generated with a similarity score close to the threshold indicating some level of SRC Kinase Inhibition potential.

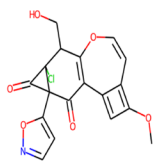
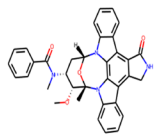
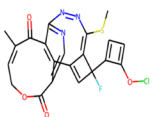
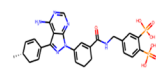
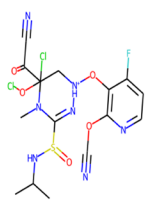
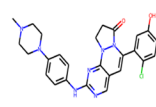
### **5.5.5 Analysis of Top Generated Molecules and Bayesian Optimizers**

Through our targeted based experimentation of the computational space of kinase inhibitors, we were successful in generating novel molecules with the potential of being SRC Kinase Inhibitors. Using kinase inhibition likelihood, the drug-like property comparison, and similarity metrics, both the Unbiased and Biased Bayesian Optimizers performed similarly in the creation of molecules most resembling SRC Kinase Inhibitors. The Biased Bayesian Optimizer did generate a larger quantity of molecules compared to the Unbiased Optimizer; however, a larger percentage of

molecules the Biased Optimizer generated were removed when accounting for a baseline value of 0.5 for the calculated Kinase Inhibition Likelihood. When comparing the Bayesian Optimizers on the chemical features and Tanimoto similarity score calculations, the two Bayesian Optimizers performed similarly with the Unbiased Optimizer having a slight advantage with respect to the Tanimoto similarity scores to known SRC Kinase Inhibitors. The Biased Bayesian Optimizer produced a molecule with a much higher calculated Kinase Inhibition Likelihood score than any molecule produced by the Unbiased Optimizer, but the Unbiased Optimizer's generated molecules had a higher average calculated Kinase Inhibition Likelihood score. In addition, the Kinase Inhibition Likelihood score was used to determine a molecule's potential feasibility of being a SRC Kinase Inhibitor; however, the Tanimoto Similarity scores to known SRC Kinase Inhibitors was given more significance in interpreting the results. The results from this experiment demonstrated that a bias within Bayesian Optimization functions may not be as impactful as originally thought. Even though the Biased Bayesian Optimization function was trained on 2258 SRC Kinase Inhibitors, the biased and unbiased models performed very similar.

Both the calculated Kinase Inhibition Likelihood and the Tanimoto similarity scores do not individually have a strong enough impact in determining the quality of the generated molecule. When used in conjunction, the two scores help give stronger evidence of a given molecule's success as a novel kinase inhibitor; however, it is not until the chemical features are incorporated that a more concrete conclusion of a given molecule can be drawn. Since the specific chemical features of QED, logP, and SAS have been shown to play a role in a molecule's drug like behavior, it is necessary to include those metrics along with the calculated Kinase Inhibition Likelihood, and the Tanimoto similarity scores with the generated molecules. Otherwise, a generated molecule could have high Kinase Inhibition Likelihood and Tanimoto similarity scores, but the chemical

features indicate the molecule would fail in either synthesis or fail to be absorbed by the body, making the molecule unsuccessful.

Generated Molecule	Known SRC Kinase Inhibitor	Tanimoto Similarity Score	Kinase Inhibition Likelihood	QED Score	logP Score	SAS Score
		0.71153	0.58567	0.63763	3.6400	2.3322
		0.69567	0.59816	0.71623	4.24062	1.80823
		0.66666	0.60690	0.76593	-0.15078	3.17604

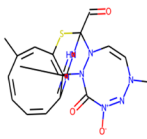
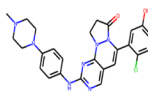
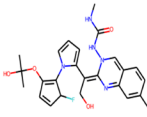
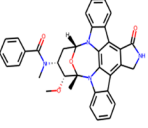
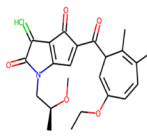
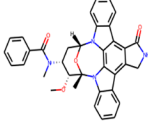
**Figure 15.** The Top Three Molecules Generated from the Unbiased Bayesian Optimizer with the Closest Known SRC Kinase Inhibitors

After initially analyzing the top molecules generated from the Unbiased Bayesian Optimizer shown in Figure 15, the molecules that were generated seem to be different from the known SRC Kinase Inhibitors they are most like. The Tanimoto similarity scores of the top molecules were not significant with the highest similarity score being only 0.71153 (Figure 15). In addition, the calculated Kinase Inhibition Likelihood scores were above the baseline threshold of 0.5; however, the scores were not as high as we would have liked them to be. Apart from the logP score of the third molecule being -0.15078, the other QED, logP, and SAS metrics were within the acceptable

range for and SRC Kinase Inhibitor, indicating that the molecules generated could be feasible if synthesized.

The final molecule within the top three from the Unbiased Optimizer having a negative logP value is slightly concerning with the overall accuracy of the Optimizer. While the molecule did perform well with the Tanimoto similarity score and had an average calculated Kinase Inhibition likelihood score, the poor chemical metrics indicate that synthesis of this molecule would be infeasible. For this specific experiment, there are no plans for synthesizing any molecules; however, the high scoring molecule with poor metrics imply that the Unbiased Optimizer may need more iterations during training or an overall improvement to the scoring function used within it.

The Unbiased Optimizer's top generated molecules were diverse in their molecular structure which suggests that the Unbiased Optimizer performed well in its searching of the molecular latent space. By expanding its searching parameters, the Optimizer could better find the local optima and produce more diverse molecules each with high potential of being an SRC Kinase Inhibitor. With a greater number of iterations or initial points, the Optimizer could improve more and produce a set of molecules with Tanimoto similarity scores and calculated Kinase Inhibition Likelihood scores closer to 1.

Generated Molecule	Known SRC Kinase Inhibitor	Tanimoto Similarity Score	Kinase Inhibition Likelihood	QED Score	logP Score	SAS Score
		0.70910	0.57519	0.63193	1.87408	3.15589
		0.64016	0.75332	0.62254	4.24315	2.2926
		0.63440	0.58261	0.78757	0.51124	2.80652

**Figure 16.** The Top Three Molecules Generated from the Biased Bayesian Optimizer with the Closest Known SRC Kinase Inhibitors


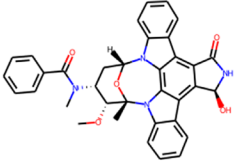
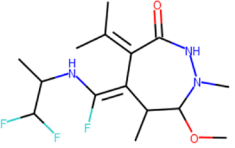
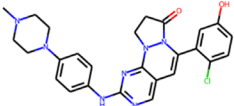
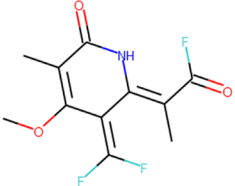
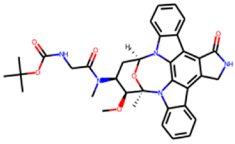
In the results of the Biased Bayesian Optimizer there was a pattern amongst the relationship between the generated molecules and the corresponding known SRC Kinase Inhibitor that they were most like (Figure 16). Of the top generated molecules, based on their Tanimoto similarity score to known SRC Kinase Inhibitors, there were two known SRC Kinase Inhibitors that were the most like the generated molecules. The same two known inhibitors score the highest in similarity until the nineteenth generated molecule in the list of all generated molecules. It could be that the bias within the Bayesian Optimizer had little effect on the overall outcomes due to the strong similarity of all the metrics used for testing the success; however, the repeated known SRC Kinase Inhibitors demonstrates that the bias ultimately skews the results. Since this pattern only occurred with the results from the Biased Optimizer, the repeated similar known molecules could

be the result of the probing before the optimizer was initiated. By probing in the areas of the known SRC Kinase Inhibitors, the Bayesian Optimizer potentially learned which areas of the latent space had the highest potential of valid molecular generation. It is possible that in the process of probing on the training set of known SRC kinase inhibitors that certain molecular motifs had a higher success rate of generation than others. This would cause the Bayesian Optimizer to focus on those areas of the latent space thereby generating more molecules with similar motifs. The generated molecules would all be similar to the same known SRC kinase inhibitors in the testing set because of their shared molecular motifs explaining the repeated known molecules. One of the downsides of this skewed searching of the latent space is regarding the local optima of the latent space. By staying within one of the local optima, the Optimizer fails to find a potentially higher scoring region. The similarity in scores to the Unbiased Optimizer suggests that the Biased Optimizer did find a local optimum but is hindered by the initial probing and does not properly search for more concentrated areas. By introducing less molecules for the Biased Optimizer to be trained on before initializing points, the Optimizer could start out with the knowledge of highly concentrated areas of the latent space, while also not being overtrained to stay within any given optimum. This would allow the Biased Optimizer to properly utilize the information given before training without the bias hindering the overall results.

The results from the Biased Optimizer indicate that a bias within a Bayesian Optimization function does not necessarily increase the overall success of the optimization function. Bias within the optimization function seems to affect which parameters are most impacted during the hyperparameter tuning which causes only minor differences between the molecules generated. While the Biased Bayesian Optimizer did produce about 26.15% more valid molecules than the

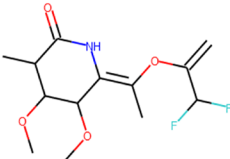
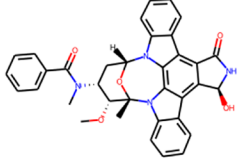
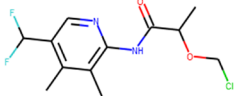
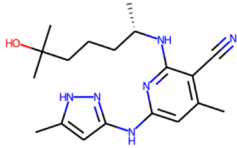
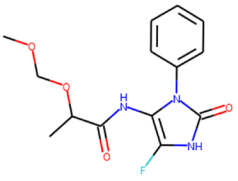
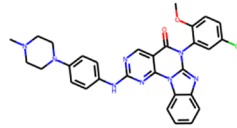
Unbiased Bayesian Optimizer, the overall quality of the molecules was similar across both optimizers.

Due to a grant by the Kay Foundation, we were able to send several molecules to Professor Keykavous Parang in the Pharmacy school during our experimentation process. I would like to extend my gratitude to him and the other Pharmacy collaborators for participating in this experiment. By providing monetary support, the Kay Family Foundation allowed us to not only conduct the initial experiments to generate the novel molecules using machine learning techniques, but also encouraged the collaboration with the School of Pharmacy which furthered our experiment with concrete results. The molecules that were sent to Dr. Parang and other Pharmacy School collaborators had simpler chemical structures allowing for a greater success rate with chemical synthesis. Along with that, the other chemical metrics used to calculate the Kinase Inhibition Likelihood score indicated that the molecules would be significant enough to propose for chemical synthesis.

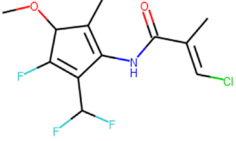
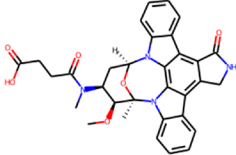
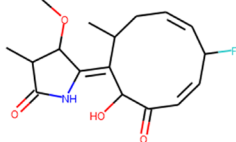
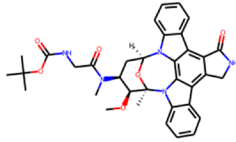
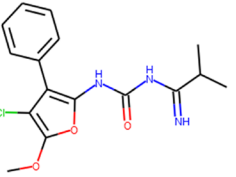
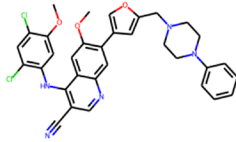
Generated Molecule	Tanimoto Similarity Score	Known SRC Kinase Inhibitor	Kinase Inhibition Likelihood
	0.5		0.649317
	0.4778		0.4066866
	0.44560		0.719936

**Figure 17.** Top Molecules Generated Through Bayesian Optimization Sent to Pharmacy School for Synthesis (Molecules 1-3)



Generated Molecule	Tanimoto Similarity Score	Known SRC Kinase Inhibitor	Kinase Inhibition Likelihood
	0.4351		0.5299372
	0.4335		0.4585961
	0.4296		0.470089

**Figure 18.** Top Molecules Generated Through Bayesian Optimization Sent to Pharmacy School for Synthesis (Molecules 4-6)

Generated Molecule	Tanimoto Similarity Score	Known SRC Kinase Inhibitor	Kinase Inhibition Likelihood
	0.4233		0.5802295
	0.4212		0.8649606
	0.4161		0.160643

**Figure 19.** Top Molecules Generated Through Bayesian Optimization Sent to Pharmacy School for Synthesis (Molecules 7-9)

The molecules that were sent to Dr. Parang's lab are shown in Figure 17, along with the known SRC Kinase Inhibitor each molecule is most similar, the Tanimoto Similarity score, and the calculated Kinase Inhibition Likelihood value. There was a total of 206 molecules generated through the Bayesian Optimization process that were sent to the School of Pharmacy for chemical synthesis. Due to the significance of synthesizing molecules generated from this experiment, more sample molecules with their most similar known SRC Kinase Inhibitor are shown through Figure 17. One of the noticeable aspects of this set of top molecules compared to

the top molecules from the Biased and Unbiased Bayesian Optimizers are the simpler chemical structures of the generated molecules. By having simpler chemical structures, the process of synthesization is easier and more cost effective. The top scoring molecule in terms of Tanimoto Similarity score to known SRC Kinase Inhibitors had a value of exactly 0.5 and a Kinase Inhibition Likelihood score of 0.649 (Figure 17). Although the Tanimoto similarity scores to the known molecules are not as high as some of the other generated molecules, these molecules being sent for chemical synthesis gives a greater amount of substance to this experiment.

The distinct spread of the calculated Kinase Inhibition Likelihood score in comparison to the Tanimoto Similarity score is shown with the top molecule of this set having a similarity score of 0.5 and Kinase Inhibition Likelihood score of 0.649, while the next top molecule having a similarity score of 0.4778, but a Kinase Inhibition Likelihood score of 0.406. The differences in scores of the generated molecules further emphasizes the room for improvement that the calculated metric has. The possible success of chemical synthesis of the generated molecules sent to Dr. Parang's lab gives another potential feature to include in the scoring function of the calculated Kinase Inhibition Likelihood. The additional feature of chemical synthesis success could further improve the measurement metric leading to higher quality molecules in the future. By improving the developed metric, the significance of this set of molecules is expanded to not only giving a more concrete outcome of this experiment, but also allowing future experiments to benefit and produce molecules with greater potential of being SRC Kinase Inhibitors.

## 6 Conclusion

The field of computational drug discovery and development continues to grow at a rapid pace, with generative deep learning being at the forefront. Due to a generous grant from the Kay Foundation, we were able to motivate the progress of generating new compounds for potential synthesis and therapeutic benefits using data-driven methods and various Machine Learning tools and mathematical applications. Our dataset consisted of 10 different kinase inhibitor families and numerous small molecules, all taken from ZINC and the GDB-17 Small Molecule Database. After encoding and mapping our molecules in a continuous latent space representation, we implemented a chemical feature based Random Forest Binary Classifier and predictor model to create an all-encompassing metric of kinase inhibition likelihood. The Random Forest Binary Classifier was then used as the scoring function for a Bayesian Optimization process that generated novel molecules through latent space search methods.

Through our analysis, we found that the Bayesian Optimization process was successful at generating novel molecules and showing it has the potential to produce molecules that are SRC Kinase Inhibitors. When analyzing the top molecules from each Bayesian Optimizer implemented, the highest scoring molecules had a Tanimoto Similarity score of 0.711 and 0.709 to a testing set of known SRC Kinase Inhibitors. Those molecules had calculated Kinase Inhibition Likelihood scores of 0.586 and 0.575 respectively. These scores of the top molecules show that the Optimizers can create molecules with the potential of being SRC Kinase Inhibitors indicated by the Tanimoto Similarity scores; however, the disconnect between the similarity scores and the calculated Kinase Inhibition Likelihood scores demonstrates the necessity to improve the underlying scoring function. The highest scoring molecule in terms of the calculated Kinase Inhibition Likelihood

came from the Biased Bayesian Optimizer with a score of 0.842, but the Tanimoto Similarity score of the molecule was only 0.548 further emphasizing the disconnect between the two metrics.

In determining the kinase inhibition potential of a generated molecule, the Tanimoto Similarity score and the calculated Kinase Inhibition Likelihood value are used in conjunction with respect to the baseline threshold of each metric. For Tanimoto similarity score, the threshold for determining high similarity to a known SRC Kinase Inhibitor is 0.75. For Kinase Inhibition Likelihood, the threshold for this experiment was 0.5, but in similar experiments the threshold was determined to be 0.75. While no generated molecule from either optimizer reached the 0.75 threshold for Tanimoto Similarity score, there were a total of 7 generated molecules with a Kinase Inhibition Likelihood value above the unadjusted threshold of 0.75. The Unbiased Optimizer produced 4 molecules above the unadjusted threshold, and the Biased Optimizer produced 3 molecules. These results show that the Bayesian Optimizer, in its current state, is not the best model for generating Kinase Inhibitors for a specific target family. If the scoring function were to be improved, the Bayesian Optimizer model could produce significantly better results; however, the Bayesian Optimizer's learning process may be too inaccurate for molecular generation. By searching through the latent space of molecules to find local optima, the optimizers risk finding a shallow optimum and missing more concentrated areas of the molecular latent space. In this experiment, the Biased Optimizer experienced this problem during its training. After using known SRC Kinase Inhibitors to probe the model before initiating random points, the optimizer learned the locations of some local optima where molecules from the training set were located. The overall low similarity scores and Kinase Inhibition Likelihood scores indicate that the optimum the optimizer found during training was not the best location within the latent space. Due to the exploitation processes of the Bayesian Optimizers, the model stayed within the local optimum

rather than expanding its searching process for a potentially more concentrated area demonstrating the shortcomings of this model within molecular generation or drug discovery.

This experiment also demonstrated that the Bayesian Optimization process would need to be improved prior to any continuation experiments for the generation of higher quality molecules. It was determined that training the Bayesian Optimization function before initiating the optimization process had little effect on the quality of the molecules generated. When conducting drug discovery experiments, creating diverse molecules is more important due to the chemical synthesis process. If all molecules generated are one atom or one bond off, if one fails to synthesize properly it could mean that the rest of the generated molecules would also fail during synthesis. By creating a more diverse set of molecules, the rate of success in the synthesization process could increase thereby causing the experiment to be a success as well. Though the Biased Bayesian Optimizer did produce similar molecules in terms of the similarity scores and Kinase Inhibition Likelihood scores, the repeated motifs of the known SRC Kinase Inhibitor most like the generated molecules indicate the optimizer's lack of ability to generate diverse molecules. The experiment showed that the Unbiased Bayesian Optimizer would be the better optimizer to use in future experiments due to the diversity of molecular motifs found in the generated molecules.

The additional set of generated molecules that were sent to Dr. Parang's lab for chemical synthesis gives a better baseline set of models for the Bayesian Optimizers. Although the molecules sent to Dr. Parang's lab were created earlier in the experiment's lifespan, the success rate of the molecules' chemical synthesis creates an opportunity to further improve the Kinase Inhibition Likelihood metric developed for this experiment. The top molecules from this set had Tanimoto Similarity scores of 0.5, 0.4778, 0.445 while their corresponding Kinase Inhibition Likelihood scores were 0.649, 0.406, and 0.719 respectively. Since the Tanimoto Similarity score metric is a

heavily utilized measurement method within the field of computational drug discovery, the similarity scores carry a stronger weight in determining the potential for a molecule to be an SRC Kinase Inhibitor. However, the structural or chemical makeup of a molecule is not a reliable sole factor in determining the kinase inhibition ability of a generated molecule. The chemical synthesis success of a generated molecule can be used in training of the Random Forest model to vastly improve the accuracy of the calculated score. By improving the accuracy of the score, the calculated Kinase Inhibition Likelihood value can become more valuable than the Tanimoto Similarity score in determining the potential of a given molecule to belong to a specific target kinase family. Due to the differences in structural motifs within a given kinase family, it would not be unreasonable to assume that molecules that demonstrate kinase inhibition properties, but that do not share structural similarities could still belong to a target family. This shows one of the drawbacks of using the Tanimoto similarity score with greater emphasis than the calculated Kinase Inhibition Likelihood score which is based on a more diverse set of chemical features.

### **6.1.1 Future Expansions**

Improving the scoring function to be more accurate at scoring molecules would be the next step in expanding our research. Currently the calculated Kinase Inhibition Likelihood score is a good metric at determining the general potential a generated molecule has, but it is not as all-encompassing of a metric as we would have liked. One possible idea for improving the Random Forest function that is used as the basis for the scoring function is incorporating the Tanimoto similarity score into its calculations. By using the similarity score of any generated molecule as a possible feature for the Random Forest, the calculated Kinase Inhibition Likelihood score could become more accurate as it incorporates more knowledge of the SRC Kinase Inhibitors before producing a score.

The next step would be implementing a General Adversarial Network (GAN) into our experiment to improve the quality of the generated molecules. The GAN would take the scoring function used in the Bayesian Optimization process to make kinase inhibiting alterations to small molecules. By utilizing the scoring function, the GAN would make more targeted alterations than the Bayesian Optimizer which would allow for the greater production of high-quality molecules. To produce higher quality molecules, the scoring function would need to be improved to have more accurate predictions of the Kinase Inhibition Likelihood score.



## 7 References

- Agajanian, S., Odeyemi, O., Bischoff, N., Ratra, S., & Verkhivker, G. (2018). Machine learning classification and structure-functional analysis of cancer mutations reveal unique dynamic and network signatures of driver sites in oncogenes and tumor suppressor genes. *J Chem. Inf. Model*, *58*, 2131-2150.
- Agajanian, S., Oluyemi, O., & Verkhivker, G. (2019). Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations. *Front Mol Biosci*, *6*, 44.
- Ballester, P., & Mitchell, J. (2010). A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, *26*(9), 1169-1175.
- Boulesteix, A., Janitza, S., Kruppa, J., & König, I. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(6), 493-507.
- Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.
- Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov Today*, *23*(6), 1241-1250.
- De Cao, N., & Kipf, T. (2018). *MolGAN: An implicit generative model for small molecular graphs*. arXiv preprint arXiv: 1805.11973.
- Decherchi, S., Berteotti, A., Bottegoni, G., Rocchia, W., & Cavalli, A. (2015). The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics machine learning. *Nat Commun*, *6*, 6155.
- Dimitrov, T., Kreisbeck, C., Becker, J., Aspuru-Guzik, A., & Saikin, S. (2019). Autonomous Molecular Design: Then and Now. *ACS Appl Mater Interfaces*, *11*(28), 24825-24836.
- Frazier, P. (2018). A Tutorial on Bayesian Optimization. *arXiv[stat.ML]*, arXiv: 1807.02811.
- Gaulton, A., Bellis, L., Bento, A., Chambers, J., Davies, M., Hersey, A., . . . Overington, J. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, *40*(D1), D1100 - D1107.
- Goh, G., Hodas, N., & Vishnu, A. (2017). Deep learning for computational chemistry. *J. Comput. Chem.*, *38*, 1291-1307.

- Gómez-Bombarelli, R., Wei, J., Duvenaud, D., Hernández-Lobato, J., Sánchez-Lengeling, B., Sherberla, D., . . . Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2), 268-276.
- Hemmerich, J., Troger, F., Fuzi, B., & Ecker, G. (2020). Using Machine Learning Methods and Structural Alerts for Prediction of Mitochondrial Toxicity. *Molecular Informatics*, 39(5).
- Hinton, G., & Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786), 504-507.
- Holzinger, A., Dehmer, M., & Jurisica, I. (2014). Knowledge Discovery and Interactive Data Mining in Bioinformatics -- State-of-the-Art, future challenges and research directions. *BMC Bioinformatics*, 15, Suppl6.
- Hu, Y.-J., Li, L.-X., Han, J.-C., Min, L., & Li, C.-C. (2020). Recent Advances in the Total Synthesis of Natural Products Containing Eight-Membered Carbocycles (2009–2019). *Chem. Rev.*, 120(13), 5910-5953.
- Husic, B., & Pande, V. (2018). Markov State Models: From an Art to a Science. *J Am Chem Soc*, 140(7), 2386-2396.
- Irwin, J., & Shoichet, B. (2005). ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model*, 45(1), 177-182.
- Jiao, Q., Bi, L., Ren, Y., Song, S., Wang, Q., & Wang, Y. (2018). Advances in studies of tyrosine kinase inhibitors and their acquired resistance. *Molecular cancer*, 17(1), 1-12.
- Jin, W., Barzilay, R., & Jaakkola, T. (n.d.). Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv [cs.LG]*, arXiv:1802.04364v4.
- Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., & Zhavoronkov, A. (2017). The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, 8(7), 10883.
- Kannaiyan, R., & Mahadevan, D. (2018). A comprehensive review of protein kinase inhibitors for cancer therapy. *Expert review of anticancer therapy*, 18(12), 1249-1270.
- Kier, L., & Hall, L. (2002). The Meaning of Molecular Connectivity: A Bimolecular Accessibility Model. *Croatica Chemica ACTA*, 72(2), 371-382.
- Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., & Bae, H. (2019). Deep Learning in Medical Imaging. *Neurospine*, 16(4), 657-668.
- Korotcov, A., Tkachenko, V., Russo, D., & Ekins, S. (2017). Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol Pharm*, 14(12), 4462-4475.

- Kuzminykh, D., Polykovskiy, D., Artur Kadurin, A., Zhebak, A., Baskov, I., Nikolenko, S., & Zhavoronkov, A. (2018). 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks. *Mol. Pharmaceutics*, *15*(10), 4378-4385.
- Landrum, G. (2013). RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling.
- Li, Z., Kermode, J., & De Vita, A. (2015). Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys Rev Lett*, *114*(9), 096405.
- Lipinski, C. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, *1*(4), 337-341.
- Mater, A., & Coote, M. (2017). Deep Learning in Chemistry. *J Chem Inf Model*, *59*(6), 2545-2559.
- Maziarka, L., Pocha, A., Kaczmarczyk, J., Rataj, K., Danel, T., & Warchol, M. (2020). Mol-CycleGAN: a generative model for molecular optimization. *Journal of Cheminformatics*, *12*(1), 1-18.
- Polêto, M., Rusu, V., Grisci, B., Dorn, M., Lins, R., & Verli, H. (2018). Aromatic Rings Commonly Used in Medicinal Chemistry: Force Fields Comparison and Interactions With Water Toward the Design of New Chemical Entities. *Frontiers in pharmacology*, *9*(395).
- Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Sci Adv*, *4*(7).
- Ruddigkeit, L., van Deursen, R., Blum, L., & Reymond, J. (2012). Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, *52*(11), 2864-2875.
- Snoek, J., Larochelle, H., & Adams, R. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv [stat.ML]*, *arXiv:1206.2944*.
- Weininger, D. (1988). SMILES, a chemical language and information system. *Journal of Chemical Information and Computer Sciences*, *31*(6).
- Wishart, D., Knox, C., Guo, A., Cheng, D., Shrivastava, S., Tzur, D., . . . Hasanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, *36*(Supp\_1), D901-D906.
- Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). Seqgan: Sequence generative adversarial nets with policy gradient. *Proceedings of the AAAI conference on artificial intelligence*, *31*.