

**AI In Chemistry: Accelerating Early-Stage Drug Discovery Through the Iterative
Use of Machine Learning And Improved Data Quality**

By

Hunter Sturm

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba

in partial fulfillment of the requirements of the degree

Doctor of Philosophy

Department of Chemistry

University of Manitoba

Winnipeg

Copyright © 2025 by Hunter Sturm

Abstract

Currently, it takes around 2 billion dollars and 15 years to bring a drug to market. The excessive cost and time can be partly attributed to issues within the drug discovery pipeline. This research focuses on speeding up the time it takes to bring a drug to market as well as the associated cost. This is done by applying machine learning methods to early-stage drug discovery to improve hit rates, improve the quality of data, and investigate the size of the dataset needed for training ML models. More specifically, this study investigates the predictive power of ML models in the classification of growth inhibitory activity, colloidal aggregation, and enzyme inhibition. Furthermore, in the case of colloidal aggregation and enzyme inhibition, a subset of ML known as xAI (explainable AI) is utilized to gain insights into how the model makes predictions by having said model explain itself. While it is not yet clear whether this work has directly accelerated the drug discovery pipeline, it has contributed meaningfully to the process. Notable outcomes include: the development of a ML model used in identifying a lead compound, a predictive xAI model for small molecules for identifying and modifying aggregation, and a foundation for an xAI model that predicts enzyme inhibition of small molecules. Additionally, this work has improved data quality and provided insights into the optimal size and balance of training datasets for ML models.

Acknowledgments

First, I would like to thank my committee members Dr. Rebecca Davis, Dr. Frank Schweizer, Dr. John Sorensen, and Dr. Silvia Cardona. Without your guidance and support, I would not have finished my thesis.

Secondly, I would like to thank every one of my collaborators without whom I would not have been able to finish this thesis.

Third, I would like to thank the members of the Davis group past and present for always being willing to talk about science and assisting with research-related problems.

Dedication

This thesis is dedicated to anyone and everyone who helped me get to where I am today.

Table of Contents

1.1 Introduction to The Drug Discovery Pipeline	1
1.2: Early Drug Discovery.....	1
1.2.1 High Throughput Screening	3
1.2.3 Machine Learning Methods	14
1.2.3.1 Machine Learning Overview	14
1.2.3.2 Machine Learning Architectures	15
1.2.3.3 Explainable Artificial Intelligence	17
1.2.3.4 Data Requirements	17
1.2.3.5 Evaluation of ML Models.....	20
1.3 Molecular Features	23
1.4 ML for Antibiotic Discovery	25
1.5 Objectives of Thesis Work.....	26
1.6 References:.....	27
2.1 Introduction	36
2.2 Methods.....	41
2.2.1 Library Selection and Curation.....	41
2.2.2 Library Analysis.....	41
2.2.3 ML Model.....	42
2.3 Results and Discussion	42
2.4 Conclusions.....	52
2.5 References.....	54
3.1 Introduction	59
3.2 Methods.....	62
3.2.1 Dataset preparation	62
3.2.2 Feature Determination	63
3.2.3 Machine Learning Classification Modeling	63
3.2.4 Random Forest Training.....	63
3.2.5 MEGAN Training	64
3.3 Results.....	64
3.3.1 Training Set Diversity	64

3.3.2 Model Performance.....	66
3.3.3 Explanation Examination	68
3.3.4 Cluster Analysis	72
3.4 Conclusion	75
3.5 References.....	76
4.1 Introduction	78
4.2 Results and Discussion	82
4.2.1 The MEGAN model	82
4.2.2 Prediction accuracy and benchmarking	84
4.2.3 Sensitivity of the model to small structural modifications	85
4.2.4 DFT assessment of physical relevance for MEGAN model structural sensitivity ..	87
4.2.5 Global explanations	90
4.3.6 Counterfactuals.....	93
4.2.7 Experimental validation of predictions and counterfactuals.....	96
4.2.8 External validation	98
4.2.9 Conclusion	100
4.3 Methods.....	101
4.3.1 Dataset curation and preparation	101
4.3.2 Explainable graph neural networks, MEGAN.....	103
4.3.3 Counterfactuals.....	104
4.3.4 Fingerprint and descriptor-based ensemble methods	104
4.3.5 DFT modeling of interaction energies	105
4.3.6 Dynamic light scattering experiments	106
4.4 References.....	106
5.1 Conclusions.....	113
5.1.1 Chapter 2	113
5.1.3 Chapter 4.....	114
5.2 Future work.....	115
5.2.2 Chapter 3.....	116
5.2.3 Chapter 4.....	116
5.3 Final Thoughts	117

List of tables

Table 2.1. The nine libraries that were curated for this study, with target and MIC cutoff reported as well as size and balance.....	50
---	----

List of Figures

Figure 1.1. The drug discovery pipeline is broken down into the five processes required to bring a drug to market.	1
Figure 1.2 An example of the 12 scaffolds of antibiotics	6
Figure 1.3 overview of CADD methods illustrating how CADD branches into different approaches but ultimately can lead to the same screening in the end.....	8
Figure 1.4 (left) The 2D representation of identified pharmacophores for the molecule. (right) the 3D representation of the figure on the left, and the input for calculating color tanimoto.....	9
Figure 1.5 A) the underlying architecture of RF and XGBoost. ³⁰ B) the underlying architecture of DMPNN, GAT, and SVM.	15
Figure 1.6 A) Labeled confusion matrix B) Example ROC plot ⁴³ C) example PRC plot	20
Figure 2.1. Tanimoto combo scores for the pairwise comparison of the 4160 active molecules in the dataset taken from Johnson et. al. the x axis shows the range of Tanimoto combo scores in increments of 0.1 and the y axis shows the number of molecular com	44
Figure 2.2. Common “drug-like” features A) the number of H-bond acceptors B) the number of H-bond donors, C) The number of rotatable bonds, D) the calculated Log P, E) the molecular weight range of the molecules, and F) the globularity range. Error! Bookmark not defined.	
Figure 2.3. The compounds A and B that showed growth inhibitory activity against the ESKAPE pathogens.	47
Figure 2.4. The simplest tetrahydro- β -carboline scaffold with transformable carboxylic acid (top) with generated bioisosteric side chains shown in red to replace the acid(A-C).	48
Figure 3.1 A simplified MEGAN architecture and what explanations look like from a trained MEGAN model.....	62
Figure 3.2 Bar graphs of the number of molecules (y-axis) compared to instances of a feature (x-axis) for active (blue) and inactive (orange) compounds. Features examined are molecular weight (top left) hydrogen bond donors (top right), LogP (bottom left), and hydrogen bond acceptors (bottom right).	65
Figure 3.3 The confusion matrix, accuracy, and F1 score for the MEGAN (left) and RF (right) models.	67
Figure 3.4 UMAP projection for predicted inactive masks(top) and active masks(bottom)	69

Figure 3.5 Explanations representative to each cluster with cluster number (inactive clusters are masks that contribute to inactivity and active clusters show masks that contribute to activity) note only clusters with one common explanation shown.....	71
Figure 3.6 Representative compounds from the cluster of β -lactam containing compounds with common explanation shown, note all 11 cluster members showed the above explanation for activity.	73
Figure 3.7 shows the oxacillin derivative with explanations generated by MEGAN compared to the interactions present in the active site crystal structure of oxacillin bound to AmpC with bond lengths for hydrogen bonds(right)	74
Figure 4.1. Summary of (a) experimental and (b) in silico approaches to aggregation determination (c) compared with our work. Accuracies are those reported in the respective publications and are thus not directly comparable due to differences in training and testing datasets.....	80
Figure 4.2. MEGAN model overview. (a) Molecular graph structures are used as input to multiple attention-based message-passing layers. Node representations are aggregated and passed to a fully connected network to output the predicted class. Explanation masks are derived from the internal attention values. (b) Attention-based explanations are explicitly split into separate channels for each possible output class. (c) Overarching structural explanations can be found by identifying clusters in the latent space of subgraph embeddings. Analyzing all members of a concept cluster yields general trends associated with certain structural motifs. Explicit explanation masks and values are constructed for illustration purposes.	83
Figure 4.3. MEGAN predictions and explanation masks for two pairs of structurally similar molecules that exhibit experimentally contrasting aggregation behavior. ...	86
Figure 4.4. MEGAN predictions compared with DFT interaction energies of dimers for compounds E - P. Predictions are reported on a scale from 0 to 1, with 0 indicating a prediction of non-aggregator and 1 indicating a prediction of aggregator. Also shown are the 3D structures of the optimized dimer geometries and MEGAN explanation masks for each molecule. The light-colored explanation masks for I indicate that the MEGAN model detects no pattern systematically associated with either class.	90
Figure 4.5. Counterfactuals of experimentally validated aggregator A and non-aggregator B with MEGAN model's prediction confidence and explanation masks.	94

Figure 4.6. (a) MEGAN model's prediction confidence and explanation masks for clioquinol and methylclioquinol. (b) Overview of DLS method and experimental DLS results for clioquinol (orange) and methylclioquinol (green). 98

List of abbreviations

ADMET	Absorption-Distribution-Metabolism-excretion-toxicity
AMR	Antimicrobial Resistance
AUC	Area Under the Curve
CADD	Computer Aided Drug Discovery
DLS	Dynamic Light Scattering
DMPNN	Directed Message Passing Neural Network
ECFP	Extended Connectivity Fingerprint
ESKAPE	<i>Enterococcus faecium</i> , <i>Staphylococcus aureus</i> , <i>Klebsiella pneumoniae</i> , <i>Acinetobacter baumannii</i> , <i>Pseudomonas aeruginosa</i> , and <i>Enterobacter species</i>
FN	False Negative
FP	False Positive
GAT	Graph Attention Network
GN	Gram-Negative
GP	Gram-Positive
HTS	High throughput Screening
HTVS	High throughput Virtual Screening
LBDD	Ligand Based Drug Discovery
MEGAN	Multi Explanation Graph Attention Network
MIC	Minimum Inhibitory Concentration
ML	Machine Learning
MOA	Mechanism of Action
PAINS	Pan Assay Interferences Compounds
PRC	Precision-Recall Curve
QSAR	Quantitative Structure-Activity Relationship
RF	Random Forrest
ROC	Receiver Operator Characteristic Curve
SBDD	Structure Based Drug Discovery
SCAMS	Small Colloidally Aggregating Molecules
SMILES	Simplified Molecular Input Line Entry System
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

Chapter 1: ML in Early Drug Discovery

1.1 Introduction to The Drug Discovery Pipeline

The process of bringing a new drug to market typically takes 12-15 years and requires an investment of around 2.6 billion dollars.¹ It is also estimated that out of every 20,000 drug candidates tested in animals, only ten will make it to human drug trials, and of those 10, only 1 compound will make it to market. The often-overlooked aspect of the above numbers is that to get 20,000 compounds that are able to move on to testing in animals, millions, of compounds must first be screened.¹ This process of taking millions of molecules and finding a novel drug is known as the drug discovery pipeline (Figure 1.1). Drug discovery consists of five steps, the first two steps will be referred to as early drug discovery and will be examined further in the coming sections of this chapter.

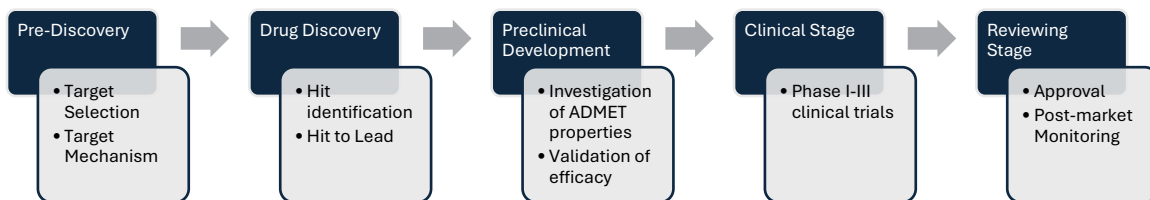


Figure 1.1. The drug discovery pipeline is broken down into the five processes required to bring a drug to market.^{2,3}

1.2: Early Drug Discovery

Early drug discovery encompasses the first three stages of the drug discovery pipeline (Figure 1.1). It begins with the selection of a suitable therapeutic target, such as a protein, gene, or pathway that plays a crucial role in disease pathology.² This initial stage involves validating the target's relevance to the disease and assessing its

Chapter 1: ML in Early Drug Discovery

druggability, which is an assessment of the target to determine if it can be modulated with a small molecule.⁴ Once a target is identified, stage 2 of the pipeline is entered, and researchers engage in hit discovery, using a variety of screening methods to identify molecules that can interact with the target.^{2, 3} At the end of stage 2, hit compounds, or compounds that have the desired effect, are evaluated for drug-likeness and favorable compounds move on to stage 3 as lead compounds. Drug-likeness is a qualitative measure of a compound's similarity to known drug compounds. Drug-likeness can look at many molecular features but often looks at log P, molecular weight, and number of hydrogen bond donors and acceptors, to name a few.⁵ In stage 3, lead compounds are further refined and optimized for potency and selectivity. These compounds are then subjected to biological testing through in vitro and in vivo assays, which evaluate their efficacy, toxicity, and pharmacokinetics, providing critical data that will inform further development in the drug discovery pipeline.³ After stage 3, lead compounds with desired effects become drug candidates, and clinical trials begin.

The initial stages of drug development are critical to the identification of new drug targets and scaffolds, and as a result, both academic and industrial researchers are focused on ways in which to improve the early stages of drug development. Researchers, however, face significant challenges, including identifying novel and biologically relevant therapeutic targets, as many targets are not easily druggable. Hit identification and lead optimization are also complex, as only a small fraction of

Chapter 1: ML in Early Drug Discovery

screened compounds demonstrate the desired efficacy, specificity, and safety profiles. Additionally, the biological testing of lead compounds often reveals unforeseen toxicity or poor pharmacokinetic properties, leading to high attrition rates, above 90%, even before clinical trials.⁶

While there are issues listed above for stage 3 of the pipeline which leads to high attrition rates, it should be noted that issues in stage 2 also contribute to high attrition rates, therefore, it is important to understand how lead compounds are identified. There are a couple different methods for finding a lead compound one of the most widely utilized methods in the wet lab is HTS, and in-silico it is HTVS.

1.2.1 High Throughput Screening

High-throughput screening is a key technology in early drug discovery, enabling the rapid evaluation of thousands to millions of compounds for biological activity against a specific target. This approach has revolutionized early drug discovery by significantly accelerating the lead identification process, reducing the time required to find promising compounds.⁷

Although HTS has accelerated the testing of large compound libraries, challenges remain in producing high-quality data that can be utilized effectively in the drug discovery process. The first issue with HTS is that it can generate a lot of false positives, causing researchers to investigate compounds that will be of no use.⁸ False hit compounds are commonly referred to as PAINS (**P**an **A**ssay **I**Nterference compound**S**).⁹ PAINS can be broken down into different classes based on their

Chapter 1: ML in Early Drug Discovery

mechanism of action. Classes of PAINS encompass many structural features like epoxides and nitro alkenes. However, the most prevalent PAINS are small colloiddally aggregating molecules (SCAMS). SCAMS account for around 80% of false hits in HTS data, and it has been found that ~20% of small molecules in publicly available compound libraries aggregate at a concentration of 30 μM , a common assay concentration.¹⁰

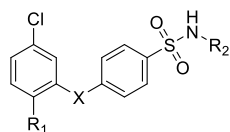
The second issue with HTS is the lack of diversity in the compound libraries that are being tested. Most libraries of small molecules are constructed of “drug-like compounds” that have similar qualities and characteristics when compared to drugs that are on the market.¹¹ Common drug-like characteristics are generally structural features, like size and number of Hydrogen-bond donors/acceptors, but can also be nonstructural features like logP.¹² These criteria for inclusion in the screening process limit the structural and chemical diversity of the libraries and mean that there are potential compounds and even classes of compounds being overlooked as new drugs because they do not fit the classical definition of a drug scaffold. While an obvious fix to this would be to synthesize many different classes of compounds, this becomes an issue as synthesizing millions of compounds requires a large investment in time and money. Also, due to the nature of parallel synthesis, if millions of compounds are to be synthesized in a short period of time, the diversity of the chemical space would still not ensure the structural diversity was appropriate for the screen.¹¹

Chapter 1: ML in Early Drug Discovery

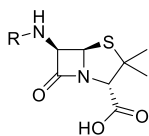
The diversity issue of HTS is further exacerbated when one is looking specifically for antibiotic compounds. With only a few exceptions, antibiotics fall into one of twelve classes (Figure 1.2).

Chapter 1: ML in Early Drug Discovery

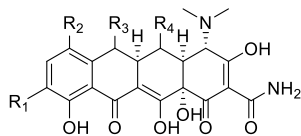
Sulfa Drugs



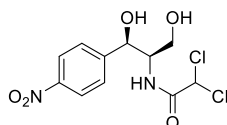
B-lactams



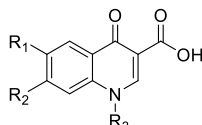
Tetracyclines



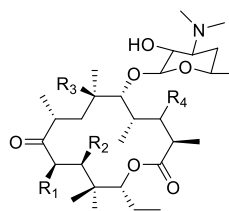
Amphenicol



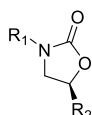
Quinolones



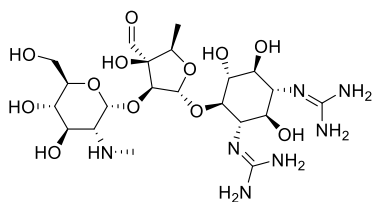
Macrolides



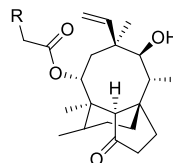
Oxazolidinones



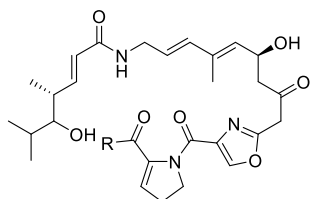
Aminoglycosides



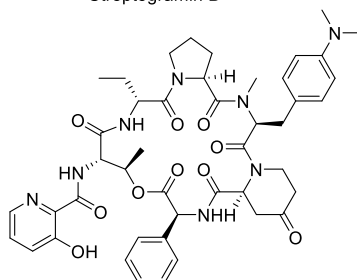
Mutilins



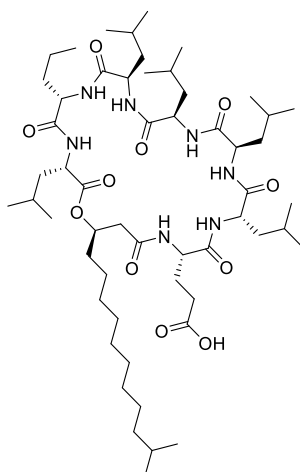
Streptogramin A



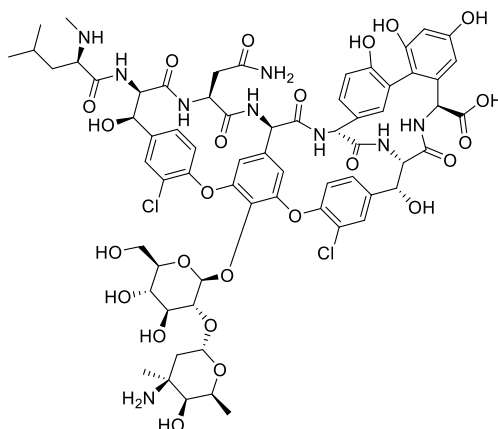
Streptogramin B



Lipopeptides



Glycopeptides



Chapter 1: ML in Early Drug Discovery

Figure 1.2 An example of the 12 common classes of antibiotics, where the structures of the common antibiotic scaffold appears under the name of said class¹³

1.2.2 *In-Silico* Methods

In order to combat some of the issues that arise from a lack of diversity in the data, researchers rely on *in-silico* methods as more diverse compounds can be selected for a screen, and parallel synthesis of all compounds is not required.¹⁴ One common method used is High Throughput Virtual Screening (HTVS). HTVS follows the same idea as HTS, however the activity of a compound toward a target is predicted *in-silico* instead of being evaluated *in-vivo/vitro*. HTVS is classified as one of many Computer Aided Drug Design (CADD) methods.

There are many different approaches to CADD with the two main approaches being: ligand-based and structure-based design (Figure 1.3).¹⁵ Figure 1.3 shows some of the many options researchers have in utilizing CADD, specifically researchers can take the overall methods of CADD and apply them in different ways to achieve their goals. Whether scientists take a structure-based approach or ligand-based approach, in most cases CADD leads to HTVS. Note: only the methods that are applied in later chapters of the thesis will be introduced in this section.

Chapter 1: ML in Early Drug Discovery

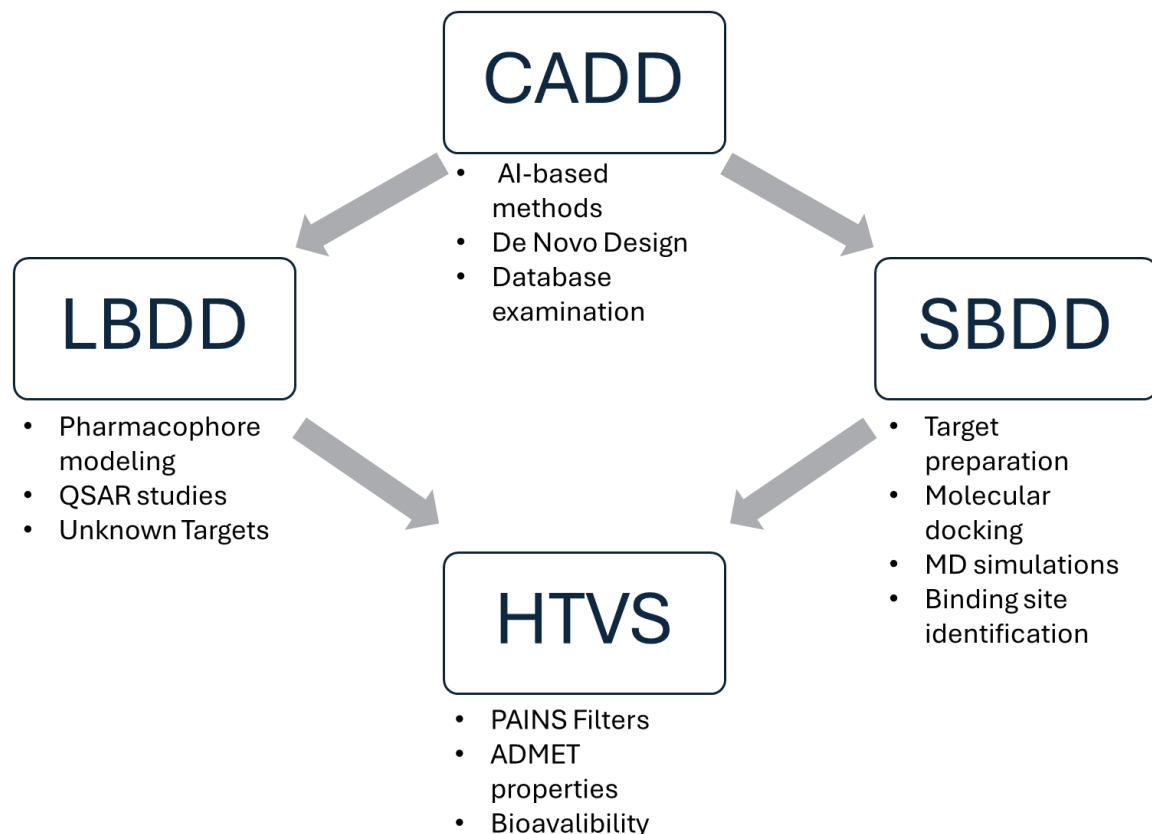


Figure 1.3 generalized workflow of CADD methods illustrating how CADD can branch into different approaches but ultimately can lead to the same screening in the end.

Ligand based drug design (LBDD) is employed in CADD when the 3D structure of the target protein is unknown.¹⁶ This approach focuses on taking a ligand with known activity and modifying its structural and electronic properties to develop a more potent drug. LBDD typically involves comparing the lead compound to known active molecules to identify shared features. Once these features have been identified, structural modification to other functional groups on the ligand can be made to enhance the efficacy of the drug candidate.

Chapter 1: ML in Early Drug Discovery

Another approach to LBDD involves the use of 2D or 3D quantitative structure-activity relationships (QSAR), pharmacophore modeling, and statistical models to validate the data.^{16, 17} For example, pharmacophore modeling software, such as that from OpenEye Scientific, assigns each functional group in the molecule a color based on its hydrogen bonding, hydrophobic, aromatic, or ionic characteristics (Figure 1.4). These properties are represented as spheres, and when two molecules are compared, their assigned colors are overlaid. The degree of overlap is then calculated, and a similarity score is assigned to each molecule based on their similarity.

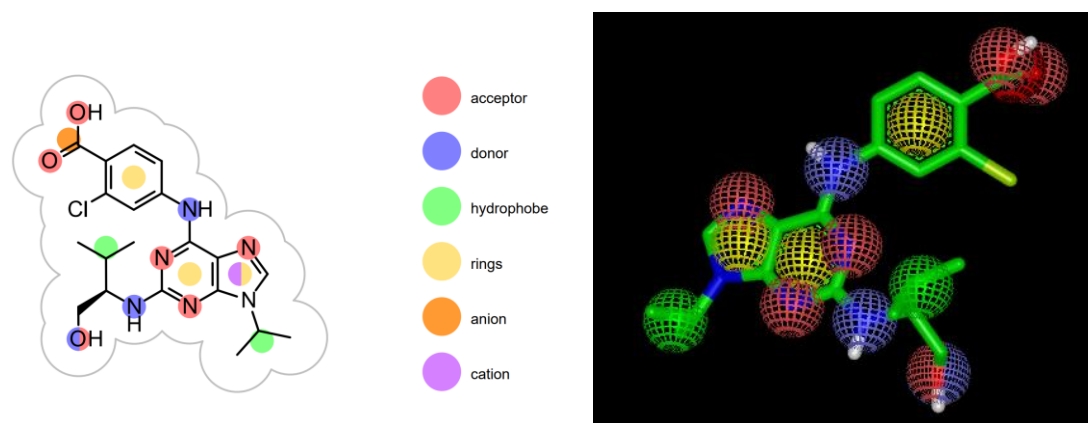


Figure 1.4 (left) The 2D representation of identified pharmacophores for the molecule. (right) the 3D representation of the figure on the left, and the input for calculating color tanimoto.¹⁸

Apart from calculating similarity, pharmacophore modeling can also be utilized in the generation of bioisosteres. Bioisosteres can be split into two categories, being classical and non-classical. Classical bioisosteres are compounds that

Chapter 1: ML in Early Drug Discovery

contain substitutions from one another but have physical similarities and related molecular shapes while maintaining similar activity. An example of a classical bioisosteric replacement would be replacing a methyl group with an alcohol. Alternatively non-classical bioisosteres do not have the same number of atoms and also differ in their steric and electronic properties while maintaining similar activity. An example of a non-classical bioisosteric replacement would be replacing a benzene ring with a thiophene.¹⁹

Structure-based approaches to CADD look at and predict how small molecules will sit in a protein's active site and try to find molecules that similarly fill that space.²⁰ These approaches are primarily based on atomic level structures of the protein-ligand complex, typically determined via X-ray crystallography or NMR spectroscopy. Structure based drug design (SBDD) can be broken down into two major approaches. The first approach in SBDD is HTVS, which uses software to predict the properties of a small molecule and how said small molecule will interact with a protein. A subset of HTVS that is often used is molecular docking. Molecular docking is an *in-silico* method where the software calculates the preferred pose of a small molecule in the active site of the target protein as well as the binding energy of that pose. Benefits of SBDD are that HTVS can be used to increase the hit rate of HTS, and that the number of compounds being screened can increase as you only need 3D coordinates of a ligand and do not need to synthesize it.²¹

Chapter 1: ML in Early Drug Discovery

Property prediction in CADD is very important as different trends have been observed when it comes to “drug-likeness” so being able to accurately describe a molecule will allow for one to see if their compound is “drug-like”. Various properties known as molecular descriptors can be calculated, such as logP (the logarithm of the octanol-water partition coefficient), globularity (a measure of how closely the small molecule represents a sphere), or presence of certain functional groups. These functional groups can be examined, and other researchers have found trends allowing for the formation of “rules” to evaluate a small molecules “drug-likeness”. Perhaps the most famous of these “rules” are Lipinski’s rules, which state a small molecule with oral bioavailability will have no more than 5 Hydrogen-bond donors, no more than 10 Hydrogen-bond acceptors, a molecular weight under 500 Daltons, and a logP lower than 5.²² Compounds are considered to satisfy Lipinski’s rules if they break no more than 2 of the above rules. Lipinski’s rule of five was developed to predict whether or not a small molecule will have oral bioavailability. Lipinski’s rules were developed when looking at a database of known approved drug compounds. There are many exceptions to these “rules”, and as such it is widely accepted that Lipinski’s rules are flawed and many variations (Hiremath, Ghose, Veber) of the rules have been proposed.²³⁻²⁵ When considering all “rules” mentioned above, a small molecule is likely to be a drug if it is rigid, has minimal rotatable bonds, a log P between -0.4 and 5, has a total polar surface area less than 140 Å², has between 20 and 70 atoms with less than 12 Hydrogen bonding atoms, and finally is somewhere in the weight range of 180-500 Daltons.²²⁻²⁵

Chapter 1: ML in Early Drug Discovery

Property prediction becomes especially important with dealing with antibiotics, which often exhibit distinct features compared to other drug classes. For example, macrolides tend not to satisfy Lipinski's rules as they are larger than 500 Daltons and have more than 5 hydrogen bond donors and more than 10 hydrogen bond acceptors yet are still viable drugs. Indeed, other antibiotics tend not to follow these variants of Lipinski's rules. The reason antibiotics have different features than drugs targeting other enzymes, as identified by Lipinski's rules, is because of the target. Different features of the molecule are important for antibiotics to have because of the nature of targeting a prokaryote cell vs a eukaryote cell. The difference in features becomes further apparent when looking at Gram-positive (GP) vs Gram-negative (GN) targeting antibiotics. It has been found that when comparing GP and GN agents to each other GP agents generally have higher molecular weights, higher polar surface areas, lower lipophilicity, and a higher number of Hydrogen-bond donor and acceptors.²⁶ Alternatively, GN agents are more polar, often charged at physiological pH, and are size-limited (MW <600 Da) due to porin-imposed diameter limits.²⁶

While antibiotics tend not to follow Lipinski's rules, GN agents have had the **eNTRy** rules created to predict the accumulation of a small molecule within the GN bacteria. The **eNTRy** rules state that compounds need a non-sterically encumbered ionizable **N**itrogen, with primary amines being preferred, low **T**hree-dimensionality with a globularity ≤ 0.25 and are relatively **R**igid with rotatable bonds ≤ 5 .²⁷

Chapter 1: ML in Early Drug Discovery

Another important feature of CADD is the ability to rapidly calculate structural similarity of molecules using Tanimoto scores. A shape Tanimoto score is calculated by first calculating the van der Waals radii for each molecule, aligning the centers of mass, and then calculating the amount of overlap in radii between the molecules. This is then reported on a scale between 0-1 wherein 0 is no overlap and 1 is full overlap. This can be added to a color Tanimoto score which assigns functional groups a color for each molecule and calculates overlap in a similar fashion again reporting on a scale of 0-1. Upon addition to these two scores, one gets a Tanimoto combo score which is on a scale of 0-2 where once again 0 is no overlap and 2 is full color and shape overlap (identical molecules).¹⁸

While SBDD and LBDD are established *in silico* approaches for identifying lead compounds other *in silico* methods that utilize HTVS are currently being developed as potentially powerful predictive models in the field of drug discovery. The models that are currently being developed tend to fall under the umbrella of Machine Learning (ML).²⁸ The results of HTVS (i.e., docked score or RMSD's of docking poses) can be used to train ML algorithms. This is important as a large amount of data is needed to train ML models, models which have the ability to further speed up the drug discovery pipeline. Whether it be through preprocessing methods to filter out PAINS, or identify negative ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties, or it be as a predictive tool to identify novel antibiotics, ML is a fast method that can speed up the drug discovery pipeline.

Chapter 1: ML in Early Drug Discovery

1.2.3 Machine Learning Methods

1.2.3.1 Machine Learning Overview

Machine learning models used in drug discovery typically learn from data through one of two primary approaches: supervised and unsupervised learning.^{29, 30} These methods differ in how the model is exposed to data and how it draws conclusions. In supervised learning, the model is provided with a dataset that consists of known input-output pairs. The input might represent features such as chemical structures or biological activity, while the output corresponds to the desired result, such as whether a compound is effective in a particular context. By training on these input-output pairs, the model learns to predict the outcome for new, unseen inputs by leveraging the patterns it has recognized during training. Unsupervised learning involves giving the model only input data without any corresponding output labels. In this case, the model must analyze the data on its own and identify any underlying patterns, clusters, or relationships. This can be particularly valuable for exploratory tasks where there is no predefined outcome, such as grouping compounds with similar chemical properties or discovering new relationships between molecular features and biological effects.

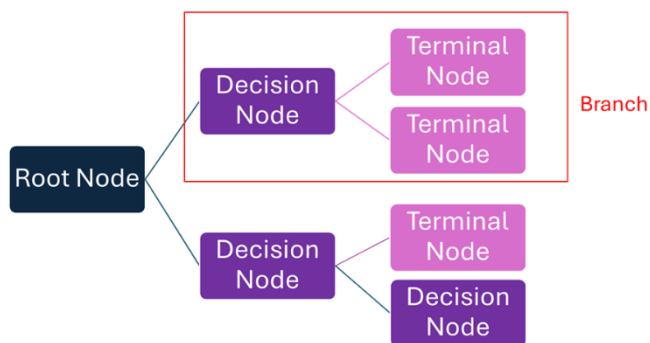
Several machine learning algorithms have become predominant in drug discovery. These include methods such as Random Forest (RF), which is a decision tree-based ensemble model commonly used for classification and regression tasks; Extreme Gradient Boosting (XGBoost), which builds strong predictive models by sequentially improving upon errors made by previous models; Directed Message

Chapter 1: ML in Early Drug Discovery

Passing Neural Networks (DMPNN), which are designed to capture the relationships within molecular graphs, making them particularly suitable for understanding chemical structures; Graph Attention Networks (GATs), which focus on the most important nodes within molecular graphs, thereby improving accuracy in tasks like predicting drug-target interactions; and Support Vector Machines (SVM), a popular method for binary classification tasks, often used to predict whether a compound is active or inactive in biological assays.^{30, 31}

1.2.3.2 Machine Learning Architectures

A.



B.

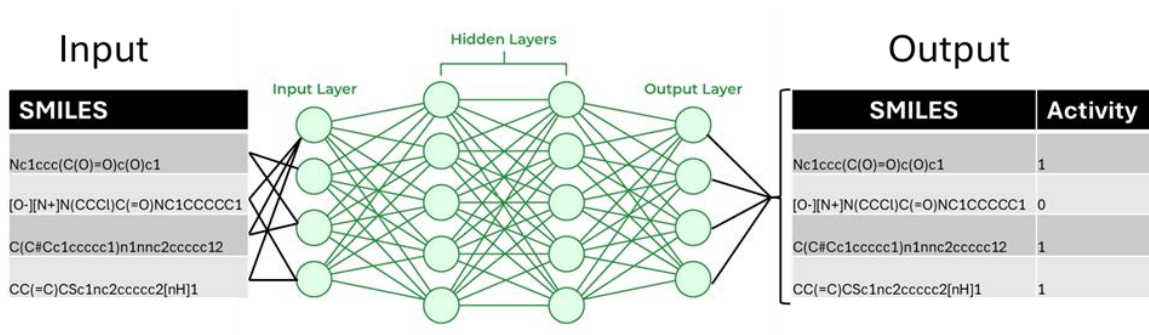


Figure 1.5 A) Example of the underlying architecture of RF and XGBoost.³⁰ (Decision Tree), wherein the root node is the data being input into the model, each decision

Chapter 1: ML in Early Drug Discovery

node represents a decision task meaning the input is examined by the model and based on features categorized by decision nodes, and each terminal node is the output prediction by the model. B) Example of the underlying architecture of DMPNN, GAT, and SVM.³¹ (Neural Network) In this example the input is a list of SMILE strings, the hidden layer is where the model does its learning and is a black box, and the output is the input list plus the predictions made by the model.

A decision tree is a supervised learning method that serves as the foundation for algorithms like Random Forest and XGBoost (Figure 1.5 A). In a decision tree, an input is provided at the root node, and the algorithm selects which branch to follow based on how well the criteria for that branch is met, ultimately reaching a leaf or terminal node, which represents the output. Both RF and XGBoost both use multiple decision trees, but they differ in their approach. RF builds all trees in parallel and combines their results to improve accuracy, whereas XGBoost builds trees sequentially, with each tree aiming to correct the errors of the previous ones.³²

A Neural Network can be used as a supervised or unsupervised learning method, it is inspired by the human brain and is the foundation of algorithms like DMPNN or GAT models. SVM is a supervised learning model, but its architecture is more closely related to that of a neural network (Figure 1.5 B) Neural networks are also the underlying architecture for DMPNN and GAT models. Both models are unsupervised learning models meaning the user inputs data that is then run through N number of hidden layers before an output is then given. Due to the unknown nature

Chapter 1: ML in Early Drug Discovery

of the hidden layers and what is being connected in them, Explainable Artificial Intelligence (XAI) has been developed to shine a light on some of the connections being made in these layers.

1.2.3.3 Explainable Artificial Intelligence

Most ML models function as black boxes, making it difficult to understand how they arrived at their predictions, XAI provides interpretability to the decisions made by these models. Unlike the other algorithms, XAI can be applied across different architectures, whether they are tree-based models, such as RF and XGBoost, or neural network-based models like DMPNN and GAT. XAI provides transparency and can help researchers and regulators understand the reasoning behind the model's predictions, which is particularly important in the highly regulated field of drug development.³³

While ML in drug discovery shows great promise, the data that is fed into the algorithm to train it has some shortcomings. A major issue with using machine learning to predict active compounds is that there is rarely enough available data to train and test the algorithm. To appropriately train and test a data set, one generally needs tens of thousands of data points. This can be a challenge as finding that many compounds that have activity data against one target is extremely difficult.³⁴

1.2.3.4 Data Requirements

Data is paramount when it comes to ML as one needs to both train and test the algorithm, which is done by first inputting data, in this case, features that describe

Chapter 1: ML in Early Drug Discovery

small molecules as well as the molecule's classification as active or inactive. The algorithm then “learns” and predicts what compounds are active and what compounds are inactive using unseen connections. Once trained, different known active and inactive compounds are given to the algorithm so that the user can determine if the algorithm is correctly predicting the activity of compounds. A nice example of this is reported by Johnson et al. where they discovered new classes of inhibitors for *M. tuberculosis*.³⁵ Using *M. tuberculosis* hypomorphs, which are mutant strains of the bacteria lacking essential targets, Johnson and coworkers discovered the mechanism of action for 45 new small molecules against *M. tuberculosis*, many of which included new scaffolds; this was completed using ML algorithms and experimentally validated results.³⁵ While the Johnson paper shows that their method is successful, for this project the greater interest is the inputs provided to the algorithm in the learning stage. For example, in the Johnson paper the researchers report that their inputs had 400 dimensions, meaning they had 400 descriptors for each molecule input into their algorithm.³⁵ In general, researchers that are using both supervised and unsupervised methods are using molecular features that are represented as both multidimensional and 1D vectors.^{35–39} Molecular features, which are discussed more in depth below, are different ways in which a molecule can be described.

One reason large data sets are important is that with smaller data sets one runs the risk of over or underfitting the algorithm. Over fitting occurs when the

Chapter 1: ML in Early Drug Discovery

algorithm learns the nonpredictive features of the data well enough that it negatively effects the output generated by the model.⁴⁰ In contrast underfitting occurs when the algorithm is unable to find a relationship between the input and output data even in the training set.⁴⁰

A second shortcoming of the training data used in drug discovery ML algorithms is the presence of PAINS. While there are many PAINS filters available for *in silico* methods, not all PAINS are currently classified. This makes properly determining if a molecule is a PAINS or just a false positive a daunting task. As such, PAINS can often be published as a good lead compound before other researchers discover that it is a PAINS.^{41, 42} While PAINS filters can be applied, the available PAINS filters have some issues. The issues from PAINS filters stem from the fact that they are based on molecular fragments not on the molecule as a whole. Thus, a compound can be misclassified relatively easily if it has a nonactive fragment that is considered a PAINS compound. This misclassification is then an issue for training ML models as accurate data is needed in order to get an accurate model.

Finally, researchers need to account for the balance of the training dataset as well. A datasets balance refers to the amount of one class to another in a classification task. An example of this would be a dataset of drug compounds and how many are active compounds verse how many are inactive compounds, if there is near 50% active to 50% inactive split the dataset is balanced if there is 75% active to 25% inactive split that would be an imbalanced dataset.

Chapter 1: ML in Early Drug Discovery

1.2.3.5 Evaluation of ML Models

ML algorithms are often evaluated using the Area Under the Curve (AUC) of a Receiver Operating Characteristics (ROC) curve. A ROC is the plot of the true positive rate vs the false positive rate (Figure 1.6 B). The true and false positives are determined using the test set data where the activity of the compounds is already known and is often reported in a confusion matrix. A confusion matrix (Figure 1.6 A) is a way to represent data in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

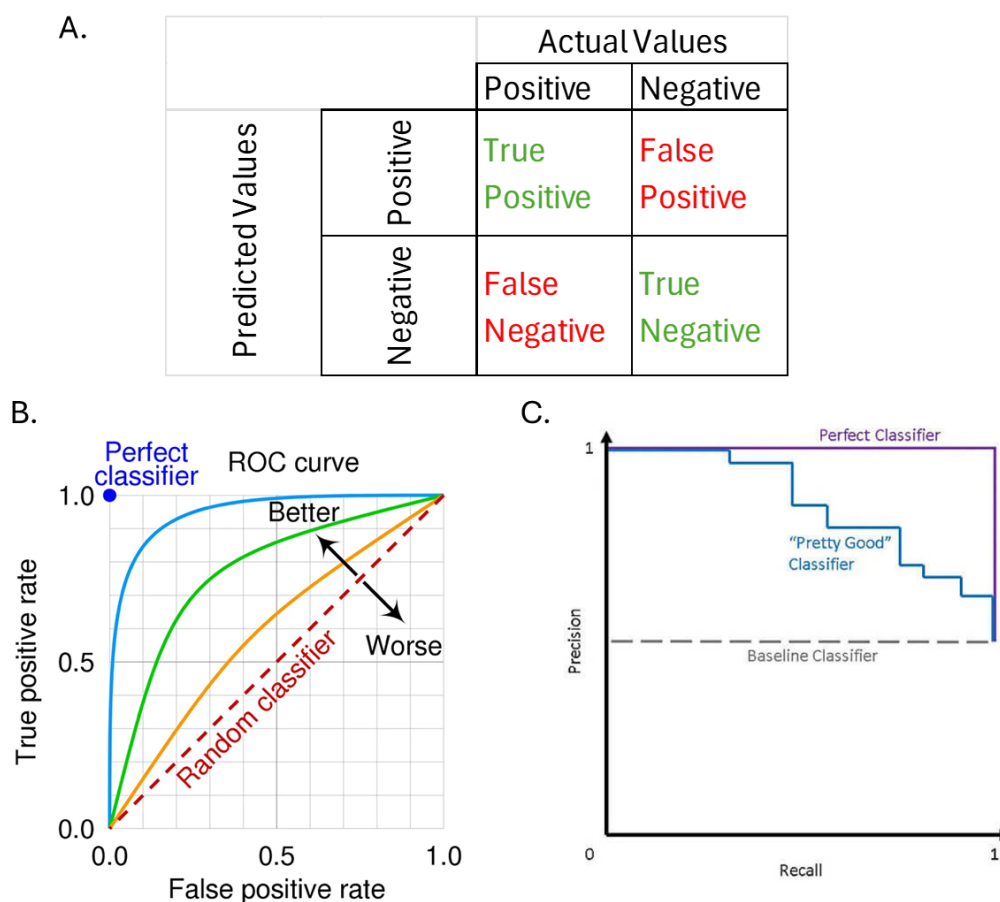


Figure 1.6 A) Labeled confusion matrix B) Example ROC plot⁴³ C) example PRC plot⁴⁴

Chapter 1: ML in Early Drug Discovery

In lay terms, Figure 1.6 A shows a confusion matrix which is just a comparison to the actual label of your data and the model's prediction of your data. From a confusion matrix, one can calculate many different metrics for evaluating a model's performance. Figure 1.6 B shows an example of a ROC plot. The ROC plot is generated by plotting the true positive rate against the false positive rate both of which can be calculated using a confusion matrix. The area under the ROC curve allows for insight into a ML model's performance. If a model's plot looks like the red dashed line, the model performs poorly and is no better than flipping a coin. As a ROC plot moves away from that red dotted line towards 1 the model is performing better, meaning in our example the model that produced the yellow line is the worst, the one that produced the green line is better but the best model in this example is the blue line. Alternatively, one can plot the precision of their model against the recall as explained below and take the area under their precision-recall curve (PRC) Figure 1.6 C. The area under the PRC is once again a measure of model performance, with the closer a value is to one the better the model is.

As stated above from a confusion matrix, one can calculate the true positive rate and the true negative rate. This aids in the generation of the ROC plot. However, a confusion matrix can tell much more about the model's performance like precision and recall. It is important that a confusion matrix can provide all this information, as while the ROC plot is most often used as a measure of a model's performance, it is only able to accurately describe the performance of models if they are trained on a

Chapter 1: ML in Early Drug Discovery

balanced data set, or one where there is a near 50-50 split of positives and negatives. This is due in part to the fact that with highly imbalanced datasets, the false positive rate tends not to fall when the total number of negatives is large. In other words, due to the algorithm learning on mostly negative compounds, it does not incorrectly classify negatives, which gives a higher AUC-ROC value whether or not it has the predictive power to classify positive classes.^{43, 44} In drug discovery research, datasets are often inherently imbalanced due to the characteristics of chemical compounds, as most compounds tested in high-throughput screening do not exhibit the properties necessary to become lead compounds. This imbalance poses challenges for model evaluation and requires the use of specific performance metrics tailored to imbalanced datasets. To evaluate performance on unbalanced datasets you needed to use precision, recall, f1 scores, and the AUC of the PRC. Precision describes the proportion of compounds that a model correctly identifies as true positives (active) out of all the compounds it predicted as active (Eq 1). In other words, it measures how often the model is right when it says a compound is a good lead. Recall describes refers to how well a model can correctly identify all the compounds that are actually true positives (Eq 2). It measures the model's ability to find as many of the relevant compounds as possible, without missing potential leads. The F1 score measures a model's accuracy by considering both how often it correctly identifies relevant compounds (precision) and how many relevant compounds it successfully finds (recall) (Eq 3). It can be used as a measure of the model's overall performance, with a higher F1 score indicating better performance.

Chapter 1: ML in Early Drug Discovery

$$\text{Eq1. Precision} = \frac{\Sigma TP}{\Sigma TP + FP}$$

$$\text{Eq2. Recall} = \frac{\Sigma TP}{\Sigma TP + FN}$$

$$\text{Eq3. F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Additionally, the AUC-PRC, or area under the precision-recall curve (Figure 1.6 C), is another measure used to evaluate how well a classification model performs when the dataset is unbalanced. The precision-recall curve plots precision (the proportion of true positives among the predicted positives) against recall (the proportion of true positives among all actual positives). The AUC-PRC provides a summary of the model's performance at various thresholds, with a higher value reflecting a more favorable balance between precision and recall. The AUC-PRC is frequently employed in evaluating a model, as it provides a more informative evaluation of performance under class imbalance compared to traditional metrics such as accuracy or the AUC-ROC.⁴⁵

1.3 Molecular Features

Molecular features play a critical role in training machine learning models for early drug discovery. Features are able to capture the structural, physical, and chemical properties of molecules in a way computers understand which is essential for the model in predicting biological activity and/or drug-like behavior. Structural features include the characteristics of the molecule such as the number of a specific functional group present in the molecule, and how the atoms are arranged and

Chapter 1: ML in Early Drug Discovery

bonded. Physical properties may include solubility, melting point, and physical state. Chemical properties, such as acidity and polarity can also provide insight into how a molecule might interact in biological systems. These features, as well as the molecule's activity, serve as inputs for machine learning models, allowing for the model to predict a molecule's potential as a lead compound.

In order for a model to understand the features that are used in the training, researchers must represent the molecules in certain ways, known as molecular representations. Molecular representations can take many forms but the most common in drug discovery currently are SMILES (Simplified Molecular Input Line Entry System), molecular fingerprints, or having the molecules described by graphs via graph encoders. While each of these representations has a different utility, they all accomplish the same thing, being alphanumeric/vector-based ways to describe complex molecules.

SMILES notation is widely used by chemists due to its ease of interpretation, but it presents certain limitations. For example, SMILES struggles to accurately represent stereochemistry, and a single molecule can be described by multiple valid SMILES strings. To overcome these issues, SMILES are often canonicalized, meaning the conversion from 3D to 1D is kept the same for all molecules and as such there is only one “true” canonical smile for each molecule. SMILES notation is generally only useful for humans and when using SMILES in training ML models the SMILES are often first converted to fingerprints or graphs so the computer can interpret them.

Chapter 1: ML in Early Drug Discovery

Molecular Fingerprints capture important aspects of a molecule's structure, with different types giving different data. Two commonly used fingerprints are Morgan fingerprints (also known as Extended Connectivity Fingerprint (ECFP)) and RDKit fingerprints (RDKfingerprint).^{46, 47} The major difference between these two types of fingerprints is that RDKfingerprints are generated based on similarity to predefined molecular fragments and are path-based fingerprints, whereas ECFP's are generated based on distance between atoms in a molecule. While these methods are both made up of some complex math, the results end up being the "same", as both methods describe the bonding of atoms in molecules.

In contrast to molecular fingerprints, molecular graphs represent molecules as complex matrices, which encode a molecule's structural identity in a format that machine learning models can interpret and analyze. This graph-based approach enables more comprehensive representation of a molecule's overall structure and properties for computational analysis. The flexibility of molecular graphs makes them suitable for handling a wide variety of molecular features, including stereochemistry and 3D conformation, which are critical in predicting biological activity.

1.4 ML for Antibiotic Discovery

There are a few recent success stories when utilizing ML for the discovery of antibiotics, namely Halicin and Abaucin.^{48, 49} Despite the success of the algorithms that identified these compounds, the fact remains that these are the only highly cited compounds identified by ML in the past decade when looking for antibiotics. The lack

Chapter 1: ML in Early Drug Discovery

of novel antibiotics identified by ML methods while concerning is unsurprising, and is attributed to both the quantity and quality of the data accessible.^{50, 51} This work while described in greater in section 1.5 aims to accelerate early drug discovery using *in-silico* methods, and part of that goal is an examination of quantity and quality of data used in training ML models for the prediction of antibiotic activity.

1.5 Objectives of Thesis Work

Described in greater detail in the coming chapters, my work over the course of this PhD was focused on applying different *in-silico* methods (as described above) to early drug discovery with the goal of speeding up the process. While there are many databases, software's and ML methods available this work will look at applying neural networks to complex issues, like drug discovery and molecular aggregation (PAINS), and will give insights into how these models “think” (XAI) as well as making screening methods that will speed up the drug discovery process.

Specifically, Chapter 2 will focus on the work I have done using deep learning methods in the prediction of growth inhibitory activity. Furthermore, in Chapter 2 work will be discussed that was completed to determine what size, balance, and diversity is needed in order to train an accurate ML model. Finally, Chapter 2 will build a foundation for which the research completed in chapters 3 and 4 were built upon.

Chapter 3 will focus on the work completed using explainable AI in the prediction of enzyme inhibition specifically AmpC which is responsible for antibiotic resistance. This chapter will focus on training a XAI model that can explain some of

Chapter 1: ML in Early Drug Discovery

the chemical insights it has gained and will look to validate the predictions/explanations that are being produced by the model.

Chapter 4 will focus on the expansion of my explainable AI work in looking at molecular aggregation. Small molecule aggregation is responsible for a large number of false hits in HTS and as such a predictive model was developed to determine the likelihood that a small molecule will aggregate. This model also provides insights into why the model believes the aggregators aggregate.

Finally, Chapter 5 is a concluding chapter that will include potential future work on all projects.

1.6 References:

1. AI for Drug Discovery Q1 2023 | Deep Pharma Intelligence. Deep Pharma.
<https://www.deep-pharma.tech/ai-in-dd-q1-2023-subscribe> (accessed 2024-07-02).
2. Pandey, A. *Drug Discovery and Development Process*. NorthEast BioLab.
<https://www.nebiolab.com/drug-discovery-and-development-process/> (accessed 2024-07-02).
3. Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K.L. Principles of early drug discovery. *Br. J. Pharmacol.* **2010**, 162(6), 1239-1249,
<https://doi.org/10.1111/j.1476-5381.2010.01127.x>.
4. Owens, J. Determining druggability. *Nat Rev Drug Discov* **2007**, 6, 187,
<https://doi-org.uml.idm.oclc.org/10.1038/nrd2275>

Chapter 1: ML in Early Drug Discovery

5. Daina, A.; Michielin, O.; Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017**, 7, 42717, <https://doi-org.uml.idm.oclc.org/10.1038/srep42717>
6. Sun, D.; Gao, W.; Hu, H.; Zhou, S. Why 90% of clinical drug development fails and how to improve it?. *Acta Pharm. Sin. B.* **2022**, 12(7), 3049-3062, <https://doi.org/10.1016/j.apsb.2022.02.002>
7. Szymanski, P.; Markowicz, M.; Mikiciuk-Olasik, E. Adaptation of High-Throughput Screening in Drug Discovery—Toxicological Screening Tests. *Int. J. Mol. Sci.* **2011**, 13(1), 427-452, doi: 10.3390/ijms13010427
8. Blucher, A. S.; Mcweeney, S. K. Challenges in secondary analysis of high throughput screening data. *Pac Symp. Biocomput.* **2014**, 114-24. PMID: 24297539
9. Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. *ACS Chem. Biol.* **2018**, 13 (1), 36–44. <https://doi.org/10.1021/acscchembio.7b00903>.
10. Reker, D.; Bernardes, G. J. L.; Rodrigues, T. Computational Advances in Combating Colloidal Aggregation in Drug Discovery. *Nat. Chem.* **2019**, 11 (5), 402–418. <https://doi.org/10.1038/s41557-019-0234-9>.
11. Henrich, C. J.; Beutler, J. A. Matching the power of high throughput screening to the chemical diversity of natural products. *Nat. Prod. Rep.*, **2013**, 30, 1284-1298, <https://doi-org.uml.idm.oclc.org/10.1039/C3NP70052F>

Chapter 1: ML in Early Drug Discovery

12. Bickerton, G., Paolini, G., Besnard, J. *et al.* Quantifying the chemical beauty of drugs. *Nature Chem* **2012**, 4, 90–98, <https://doi-org.uml.idm.oclc.org/10.1038/nchem.1243>
13. Fischbach, M. A.; Walsh, C. T. Antibiotics For Emerging Pathogens. *Science* **2009**, 325 (5944), 1089–1093. <https://doi.org/10.1126/science.1176667>.
14. Pyzer-Knapp, E. O.; Suh, C.; Gomez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What is High-Throughput Virtual Screening? A perspective from organic materials discovery. *Annu. Rev. Mater. Res.* **2015**, 45, 195-216, <https://doi.org/10.1146/annurev-matsci-070214-020823>
15. Ece, A. Computer-aided drug design. *BMC Chemistry* **2023**, 17, 26, <https://doi.org/10.1186/s13065-023-00939-w>
16. Aparoy, P.; Kumar Reddy, K.; Reddanna, P. Structure and Ligand Based Drug Design Strategies in the Development of Novel 5-LOX Inhibitors. *Curr. Med. Chem.* **2012**, 19 (22), 3763–3778. <https://doi.org/10.2174/092986712801661112>.
17. Acharya, C.; Coop, A.; Polli, J. E.; MacKerell, A. D. Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. *Curr. Comput. Aided Drug Des.* **2011**, 7 (1), 10–22.
18. ROCSReport – Applications. <https://docs.eyesopen.com/applications/rocs/rocsreport.html> (accessed 2025-01-10)

Chapter 1: ML in Early Drug Discovery

19. Silverman, R. B.; Holladay, M. W. *The Organic Chemistry of Drug Design and Drug Action*, 3rd ed.; Academic Press, 2014.
20. Moro, S.; Bacilieri, M.; Deflorian, F. Combining Ligand-Based and Structure-Based Drug Design in the Virtual Screening Arena. *Expert Opin. Drug Discov.* **2007**, 2 (1), 37–49. <https://doi.org/10.1517/17460441.2.1.37>.
21. Polgar, T.; Keseru, G. M. Integration of Virtual and High Throughput Screening in Lead Discovery Settings. *Comb. Chem. High Throughput Screen.* 2011, 14(10) 889-897, DOI: 10.2174/138620711797537148
22. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings¹. *Adv. Drug Deliv. Rev.* **2001**, 46 (1), 3–26. [https://doi.org/10.1016/S0169-409X\(00\)00129-0](https://doi.org/10.1016/S0169-409X(00)00129-0).
23. Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, 45 (12), 2615–2623. <https://doi.org/10.1021/jm020017n>.
24. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, 1 (1), 55–68. <https://doi.org/10.1021/cc9800071>.

Chapter 1: ML in Early Drug Discovery

25. Hiremath, C. N. Abbreviated Profile of Drugs (APOD): Modeling Drug Safety Profiles to Prioritize Investigational COVID-19 Treatments. *Heliyon* **2021**, 7 (8), e07666. <https://doi.org/10.1016/j.heliyon.2021.e07666>.
26. Silver, L. L. Challenges of Antibacterial Discovery. *Clin. Microbiol. Rev.* **2011**, 24 (1), 71–109. <https://doi.org/10.1128/CMR.00030-10>.
27. Richter, M. F.; Drown, B. S.; Riley, A. P.; Garcia, A.; Shirai, T.; Svec, R. L.; Hergenrother, P. J. Predictive Compound Accumulation Rules Yield a Broad-Spectrum Antibiotic. *Nature* **2017**, 545 (7654), 299–304. <https://doi.org/10.1038/nature22308>.
28. Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. *Expert Opin. Drug Discov.* **2021**, 16 (9), 949–959. <https://doi.org/10.1080/17460441.2021.1909567>.
29. Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A Review on Machine Learning Approaches and Trends in Drug Discovery. *Comput. Struct. Biotechnol. J.* **2021**, 19, 4538–4558. <https://doi.org/10.1016/j.csbj.2021.08.011>.
30. K, A. *Decision Tree Machine Learning Algorithm*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2022/01/decision-tree-machine-learning-algorithm/> (accessed 2024-07-02).

Chapter 1: ML in Early Drug Discovery

31. *Artificial Neural Networks and its Applications - GeeksforGeeks.*
<https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/> (accessed 2024-07-02).
32. Lev, A. *XGBoost versus Random Forest – qwak*
<https://www.qwak.com/post/xgboost-versus-random-forest> (accessed 2025-02-11)
33. Alizadehsani R.; Oyelere, S. S.; Hussain, S.; Jagatheesaperumal, S. K.; Calixto, R. R.; Rahouti, M. Explainable Artificial Intelligence for Drug Discovery and Development: A Comprehensive Survey, *IEEE Access*, 2024, 12, 35796-35812, doi: 10.1109/ACCESS.2024.3373195.
34. Goodarzi, M.; Dejaegher, B.; Heyden, Y. V. Feature Selection Methods in QSAR Studies. *J. AOAC Int.* **2012**, 95 (3), 636–651.
https://doi.org/10.5740/jaoacint.SGE_Goodarzi.
35. Johnson, E. O.; LaVerriere, E.; Office, E.; Stanley, M.; Meyer, E.; Kawate, T.; Gomez, J. E.; Audette, R. E.; Bandyopadhyay, N.; Betancourt, N.; Delano, K.; Da Silva, I.; Davis, J.; Gallo, C.; Gardner, M.; Golas, A. J.; Guinn, K. M.; Kennedy, S.; Korn, R.; McConnell, J. A.; Moss, C. E.; Murphy, K. C.; Nietupski, R. M.; Papavinasasundaram, K. G.; Pinkham, J. T.; Pino, P. A.; Proulx, M. K.; Ruecker, N.; Song, N.; Thompson, M.; Trujillo, C.; Wakabayashi, S.; Wallach, J. B.; Watson, C.; Ioerger, T. R.; Lander, E. S.; Hubbard, B. K.; Serrano-Wu, M. H.; Ehrt, S.; Fitzgerald, M.; Rubin, E. J.; Sassetti, C. M.; Schnappinger, D.; Hung, D. T. Large-Scale Chemical–Genetics Yields New M. Tuberculosis

Chapter 1: ML in Early Drug Discovery

Inhibitor Classes. *Nature* **2019**, 571 (7763), 72–78.

<https://doi.org/10.1038/s41586-019-1315-z>.

36. Patel, L.; Shukla, T.; Huang, X.; Ussery, D. W.; Wang, S. Machine Learning Methods in Drug Discovery. *Molecules* **2020**, 25 (22), 5277.
<https://doi.org/10.3390/molecules25225277>.
37. Fang, J.; Yang, R.; Gao, L.; Zhou, D.; Yang, S.; Liu, A.; Du, G. Predictions of BuChE Inhibitors Using Support Vector Machine and Naive Bayesian Classification Techniques in Drug Discovery. *J. Chem. Inf. Model.* **2013**, 53 (11), 3009–3020. <https://doi.org/10.1021/ci400331p>.
38. Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead Hopping Using SVM and 3D Pharmacophore Fingerprints. *J. Chem. Inf. Model.* **2005**, 45 (4), 1122–1133. <https://doi.org/10.1021/ci049732r>.
39. Wang, Y.; Zeng, J. Predicting Drug-Target Interactions Using Restricted Boltzmann Machines. *Bioinformatics* **2013**, 29 (13), i126–i134.
<https://doi.org/10.1093/bioinformatics/btt234>.
40. Brownlee, J. *Overfitting and Underfitting With Machine Learning Algorithms*. MachineLearningMastery.com.
<https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/> (accessed 2024-07-02).
41. Dahlin, J. L.; Walters, M. A. How to Triage PAINS-Full Research. *Assay Drug Dev. Technol.* **2016**, 14 (3), 168–174. <https://doi.org/10.1089/adt.2015.674>.

Chapter 1: ML in Early Drug Discovery

42. Jasial, S.; Hu, Y.; Bajorath, J. How Frequently Are Pan-Assay Interference Compounds Active? Large-Scale Analysis of Screening Data Reveals Diverse Activity Profiles, Low Global Hit Frequency, and Many Consistently Inactive Compounds. *J. Med. Chem.* **2017**, 60 (9), 3879–3886.
43. Kılıç, İ. *ROC Curve and AUC: Evaluating Model Performance*. Medium.
<https://medium.com/@ilyurek/roc-curve-and-auc-evaluating-model-performance-c2178008b02> (accessed 2024-10-17).
44. Steen, D. *Precision-Recall Curves*. Medium.
<https://medium.com/@douglassteen/precision-recall-curves-d32e5b290248> (accessed 2025-01-10)
45. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*. **2015**, 10(3), e0118432, doi: 10.1371/journal.pone.0118432
46. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, 50 (5), 742–754. <https://doi.org/10.1021/ci100050t>.
47. RDKit: Open-Source Cheminformatics. <https://www.rdkit.org>.
48. Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M., Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell*, **2020**, 180(4), 688-702, <https://doi.org/10.1016/j.cell.2020.01.021>

Chapter 1: ML in Early Drug Discovery

49. Liu, G. ; Catacutan, D. B. ; Rathod, K. ; Swanson, K. ; Jin, W. ; Mohammed, J. C. ; Chiappino-Pepe, A.; Syed, S. A.; Fragis, M.; Rachwalski, K.; Magolan, J.; Surette, M. G.; Coombes, B. K.; Jaakkola, T.; Barzilay, R.; Collins, J. J.; Stokes, M. M. Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nat Chem Biol* **2023**, 1342–1350, <https://doi-org.uml.idm.oclc.org/10.1038/s41589-023-01349-8>
50. Branda, F.; Scarpa, F. Resistance: Innovations, Global Challenges, and Healthcare’s Future. *Antibiotics* **2024**, 13(6), 502; <https://doi.org/10.3390/antibiotics13060502>
51. Cesaro, A.; Hoffman, S.C.; Das, P.; Fuente-Nunez, C. D. L. Challenges and applications of artificial intelligence in infectious diseases and antimicrobial resistance. *npj Antimicrob Resist* **2025**, 3, 2, <https://doi-org.uml.idm.oclc.org/10.1038/s44259-024-00068-x>

Chapter 2: Database Curation for ML in Antibiotic Discovery

2.1 Introduction

Antimicrobial resistance (AMR) is recognized by the World Health Organization as one of the most critical threats to global health and development.¹ In 2019, AMR was linked to more than 5 million deaths worldwide.¹ Projections by the World Bank suggest that if left unaddressed, it could result in an annual global GDP loss of up to 3.4 trillion USD by 2030.¹ This growing global health and economic crisis underscores the urgent need for innovative solutions to combat resistant pathogens.

Currently, antibiotic discovery is limited primarily due to a lack of return on investment for large pharmaceutical companies, but a few intrinsic issues also plague the field. In a review by Silver, the lack of progress in antibiotic discovery is attributed, in part, to two key bottlenecks: target selection and limited chemical diversity.² Target selection is a critical yet time-consuming step, as it requires the identification of targets that are not prone to the rapid development of resistance. The second limitation involves the chemical diversity of compounds, particularly the challenge of designing many diverse molecules capable of overcoming cellular entry barriers or ones that are unaffected by efflux pumps. While these steps present significant challenges in laboratory settings, they have the potential to be accelerated through the use of *in silico* methods, which could enhance the rate of antibiotic discovery.

Chapter 2: Database Curation for ML in Antibiotic Discovery

Machine Learning has become an increasingly valuable tool for addressing complex challenges in chemistry, particularly in the discovery of novel antibiotics. Perhaps the most notable examples of the successful application of ML in antibiotic discovery, as discussed in chapter 1, are the discoveries of Halicin and Abaucin as antibiotics.^{3, 4} ML has also been successfully applied to the discovery of antimicrobial peptides, as demonstrated by the identification of the novel "P9" peptide, which exhibited strong activity against multidrug-resistant *Acinetobacter baumannii*.⁵

Despite these early successes in the field, many challenges remain in applying ML models to the discovery of novel antibiotics. One major obstacle is the limited availability of data. ML models rely on large, high-quality datasets for effective training, and while there are vast numbers of non-antibiotic compounds, the number of known antibiotic compounds is relatively small. This scarcity of true antibiotics hinders the ability to train models effectively, reducing their predictive accuracy. Given the difficulty of building reliable models with such limited datasets, this chapter explores the requirements for a small molecule library that can be used to train an ML model to predict the growth inhibitory activity of knockdown mutants. Understanding these requirements could help address the data scarcity issue and improve the application of ML in antibiotic discovery.

The first consideration in curating a training library is its size. In ML, it is better to have large datasets with many (millions) instances of activity and inactivity for a model's training. However, in antibiotic discovery, there is a lack of "active"

Chapter 2: Database Curation for ML in Antibiotic Discovery

antibiotics available to train models, which leads to datasets with a balance of less than 1%.⁶

The balance of a library refers to the amount of positive vs negative instances.⁷ In the case of small molecules in this study, the balance refers to the ratio of the binary active/inactive label of each molecule. A perfectly balanced library would have 50% of the molecules in it as active and 50% of the molecules as inactive. However, the limitation in the number of active antibiotic compounds means that any library used in the training of an ML model will likely be unbalanced. While this is not an issue in itself, it is important to know the balance of one's library so the proper metrics can be used to determine if the model performs well. To aid in the proper training of a ML model with a small, highly imbalanced library, a diverse set of molecules can be used to help stop overfitting.

The most complex and challenging aspect of library curation is ensuring molecular diversity within the dataset, which is critical for effective machine learning model training. The limited availability of antibiotic data is further complicated by the lack of diversity within available datasets. Currently, there are only 12 commonly used antibiotic scaffolds, and antibiotics generally follow one of five specific mechanisms of action (MOA).⁸ Each MOA targets different biological pathways, resulting in significant structural variation between antibiotic classes. While essential for biological activity, this structural diversity presents challenges

Chapter 2: Database Curation for ML in Antibiotic Discovery

when building training libraries that capture sufficient variation within and across the different molecular scaffolds.

While it might seem straightforward, molecular diversity is multifaceted and depends on the molecular features used to describe the compounds. There are thousands of features that can be used to describe molecules and their properties, both structural and electronic, as well as 1D, 2D, and 3D features. Given this complexity, two molecules can be considered very similar in one respect but very diverse if looking at a different feature. For example, molecules may share common functional groups or molecular fragments but differ in spatial arrangement or electronic distribution. Therefore, when curating a library of compounds for training a ML model, diversity of structures should be calculated in different ways to determine how the molecules are different and how they are similar concerning one another.

To ensure a diverse representation of chemical space in a dataset, structural diversity must be quantified using appropriate metrics. These metrics can include molecular fingerprints, graph-based descriptors, or more complex 3D shape and electronic feature representations. Additionally, the diversity of a library of small molecules can be compared to the features present in different “rules” of drug-likeness, such as Lipinski’s rules or eNTRY rules.⁹⁻¹¹ These provide guidelines on the physicochemical properties that are typically associated with orally bioavailable and cell-permeable drugs. Comparing a library’s diversity to these drug-likeness

Chapter 2: Database Curation for ML in Antibiotic Discovery

criteria helps ensure that the compounds in the dataset not only exhibit structural variation but also fall within the desired range of properties that are relevant for antibiotic development.

The initial goal of the work outlined in this chapter was to develop a diverse library of compounds with unknown activity that could be run through a pre-trained model to identify lead compounds that showed growth inhibitory activity against *B. cenocepacia*. This goal was quickly “met”; however, the results of our predictions raised concerns that the data we were utilizing in our training was fraught with promiscuous molecules. To that end, a new library needed to be curated for training. Here, it was decided that a more systemic approach needed to be taken, and instead of finding a library we could train on, we decided to look at what we needed in a library for training. Reported in this chapter are the initial results of our predictions as background to show the need for the investigation into size and balance of ML libraries. These identification of promiscuous molecules in the training data is an important realization, as training ML models for antibiotic discovery already deals with small imbalanced libraries, and removing promiscuous molecules makes these libraries smaller. Therefore, this chapter investigates the size of training libraries and the balance of the libraries. Chapter 4 focuses on how to remove promiscuous molecules from small molecule libraries.

Chapter 2: Database Curation for ML in Antibiotic Discovery

2.2 Methods

2.2.1 Library Selection and Curation

For design of a training library for a ML model for growth inhibitory activity a number of small molecule libraries were examined. All libraries in this study are open-source with molecules being represented in SMILES notations. Prior to use, the molecules in the libraries undergo cleaning, and cleaning consists of canonicalizing smiles using RDKit, keeping the largest fragment, and finally filtering out problematic compounds (duplicates, known undesired activity). The following libraries were utilized as follows: ZINC Docking (all purchasable natural products ~330K molecules), Co-add (see table 2.1), Pubchem (diverse 10K selected from all available structures), and ChEMBL (~130K AmpC inhibitors).¹²⁻¹⁵

2.2.2 Library Analysis

Molecular shape similarity for the library was evaluated using Openeye Scientific's ROCS software which compares 3D structural overlap and calculates a shape and color Tanimoto score.¹⁶ Tanimoto similarity has also been calculated using Morgan fingerprints generated using RDKit. RDKit's Picker was also utilized to choose a diverse subset of a database.^{17, 18} Picker takes Morgan fingerprints as an input and calculates dice similarity coefficients the algorithm then chooses the user input number of diverse compounds as an output. RDKit was also used to calculate different features of molecules (i.e., H-bond donors/acceptors, number of sp³ carbons, num aromatic rings, etc. these features of molecules can then be compared to other molecules, and similarity can be calculated.

Chapter 2: Database Curation for ML in Antibiotic Discovery

2.2.3 ML Model

ML model trained and used in Chapter 2 comes from CHEMPROP which is an open-source D-MPNN model. This CHEMPROP model was chosen as there was precedence of the model working well in the prediction of antibiotics (halicin).¹⁹

2.3 Results and Discussion

Prior to my joining of the Lab our collaborators had trained a ML model that was used to predict growth inhibitory activity. My first step was to curate a diverse library of compounds that we could be used as input in a pre-trained CHEMPROP model to predict growth inhibitor activity against *Burkholderia cenocepacia*.²⁰ Before a library could be established, a similarity metric needed to be found. To do this, a reference dataset was first needed to determine the relevant features in the antibiotic chemical space. I completed this task using the database curated by Johnson et.al. in their screening of compounds active against *Mycobacterium tuberculosis* hypomorphs.²¹ This dataset contains 4160 compounds that were experimentally determined to have some growth inhibitory activity against *M. tuberculosis* mutants which is a larger amount of “actives” than other publicly available antimicrobial libraries at the time. The large number of actives were important as it allowed us to assess what features of the molecule correlate with their activity. This dataset was chosen as each compound had been experimentally validated.

Chapter 2: Database Curation for ML in Antibiotic Discovery

Using ROCS, Tanimoto combo scores were generated for each molecule compared to every other molecule giving approximately seventeen million values to examine. Tanimoto combo scores were used as they evaluate the shape and electronics of the molecule, and based on the amount of overlap, one can determine the diversity of the dataset. Based on the data presented in Figure 2.1, it was determined that the compounds were diverse in regard to structural features. This was determined as the Tanimoto combo score is the sum of the shape Tanimoto and color Tanimoto scores; each score is represented by a number between 0 and 1, wherein 0 would consist of no overlap, and 1 would be complete overlap, in the Tanimoto combo score this scale is then between 0-2. Most of the molecules had Tanimoto combo scores between 0.4 and 0.9 out of 2 (Figure 2.1). This means that the molecules are less than 50% similar according to this metric.

Chapter 2: Database Curation for ML in Antibiotic Discovery

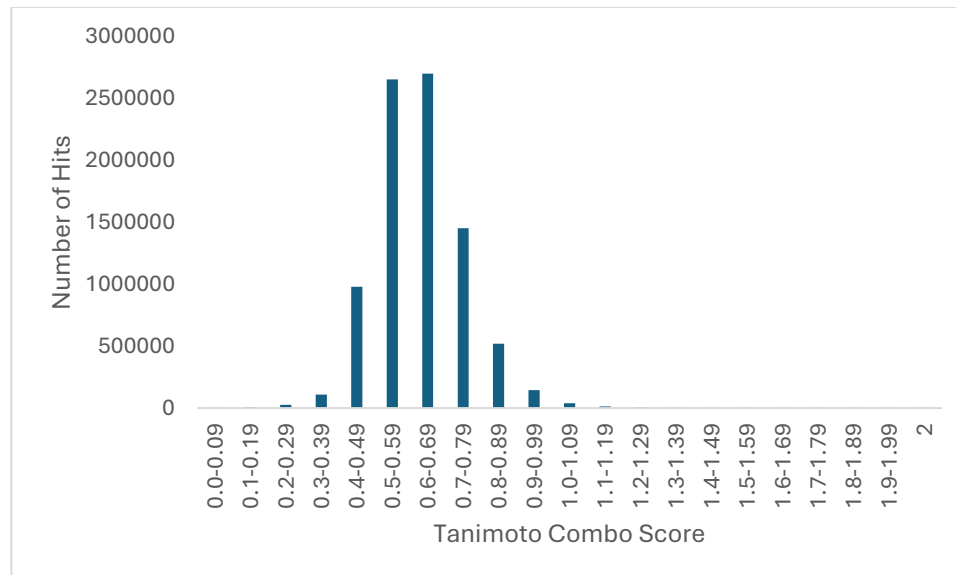


Figure 2.1. Tanimoto combo scores for the pairwise comparison of the 4160 active molecules in the dataset taken from Johnson et. al. the x axis shows the range of Tanimoto combo scores in increments of 0.1 and the y axis shows the number of molecular com

A closer look was taken at the 4160 compounds mentioned above, and the molecules were examined for “drug-likeness”, specifically the features that are present in the Lipinski rules as well as the entry rules, meaning the distribution of the number of H-bond donors and acceptors, Log p, molecular weight, rotatable bond count, and globularity (Figure 2.2).

Chapter 2: Database Curation for ML in Antibiotic Discovery

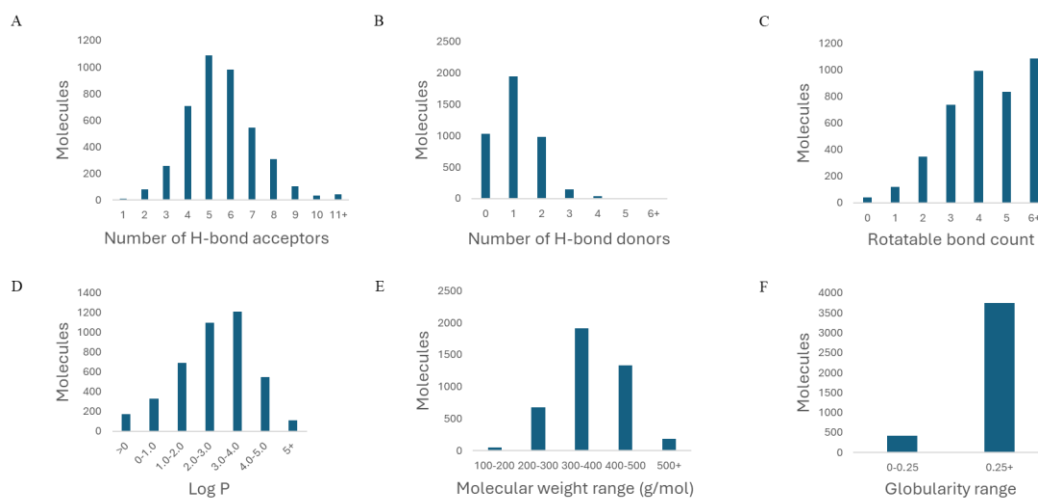


Figure 2.2. Common “drug-like” features A) the number of H-bond acceptors B) the number of H-bond donors, C) The number of rotatable bonds, D) the calculated Log P, E) the molecular weight range of the molecules, and F) the globularity range.

The features of molecules examined in Figure 2.2 are features that make up both Lipinski’s rules and the entry rules. Lipinski’s rules state that for a small molecule to be likely to have oral bioavailability, it must not break more than two of the following rules:^{9, 10}

1. A molecular weight of no more than 500 Da
2. No more than 5 H-bond donors
3. No more than 10 H-bond acceptors
4. Log P no greater than 5

These rules do have many exceptions, especially when it comes to antibiotics which is why the features of molecules listed in the entry rules were also examined. The

Chapter 2: Database Curation for ML in Antibiotic Discovery

entry rules were designed specifically to determine the likelihood a small molecule will accumulate inside of a gram-negative bacterium; more specifically, small molecules are likely to accumulate in a gram-negative bacterium if they have a non-sterically hindered ionizable nitrogen, have globularity ≤ 0.25 , and are rigid.¹¹

Upon evaluation of the Lipinski and entry rules for the 4160 active molecules examined above, it was found that while 4107 molecules satisfied Lipinski's rules, only 25 satisfied the entry rules and were likely to accumulate in gram-negative bacteria. This was unsurprising as the initial screening of the compounds was completed on gram-positive bacteria so the dataset not accumulating in Gram-negative bacteria makes sense. This also led to the determination that future examination of the features of this library should be halted as most of the compounds are predicted to not accumulate inside the cell of Gram-negative bacterium. While there are many ways to define diversity and drug-likeness within the chemical space, it was decided in this study to keep using structural diversity as it is computationally easy to calculate, and an examination of the above dataset proved to adequately explain diversity in that dataset.

Utilizing structural similarity as a diversity metric and using the natural products present on ZINCdocking, a library of compounds was generated and then filtered to exclude compounds that had reported antibiotic activity. The remaining data ended up being 224,205 diverse natural products, that were used as inputs into our trained model and as an output, each compound had its likelihood of having

Chapter 2: Database Curation for ML in Antibiotic Discovery

growth inhibitory activity predicted. The natural products were chosen for this study as most of the compounds are for sale, which would aid in the subsequent testing of the predicted compounds. The top one hundred unique compounds from this natural product prediction were examined, and forty-six were purchased for further testing.

As reported in our publication titled “A machine learning model trained on a high throughput antibacterial screen increases the hit rate of drug discovery” the testing of the top 46 compounds revealed that two of the forty-six compounds purchased for further testing showed growth-inhibitory properties against the ESKAPE pathogens (residual growth ≤ 0.8) (Figure 2.3).²²

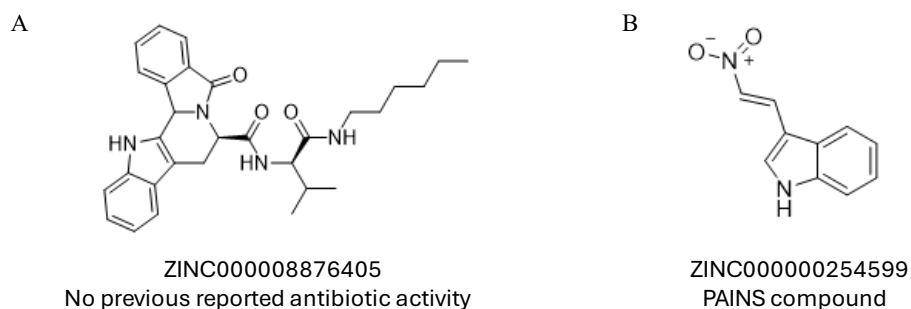


Figure 2.3. The compounds A and B that showed growth inhibitory activity against the ESKAPE pathogens.

The ESKAPE pathogens are six bacteria that have shown multidrug resistance, and thus, inhibiting their growth is an important step. The first compound turned out to be in a class of compounds known as tetrahydro- β -

Chapter 2: Database Curation for ML in Antibiotic Discovery

carbolines (ZINC8876405). Tetrahydro- β -carbolines are readily synthesizable in literature, and due to a carboxylic acid moiety in the simplest structure and the ability for acids to be readily transformed, bioisosteres were generated for compound one. This was done using Openeye Scientific's BROOD software (Figure 2.4).²³

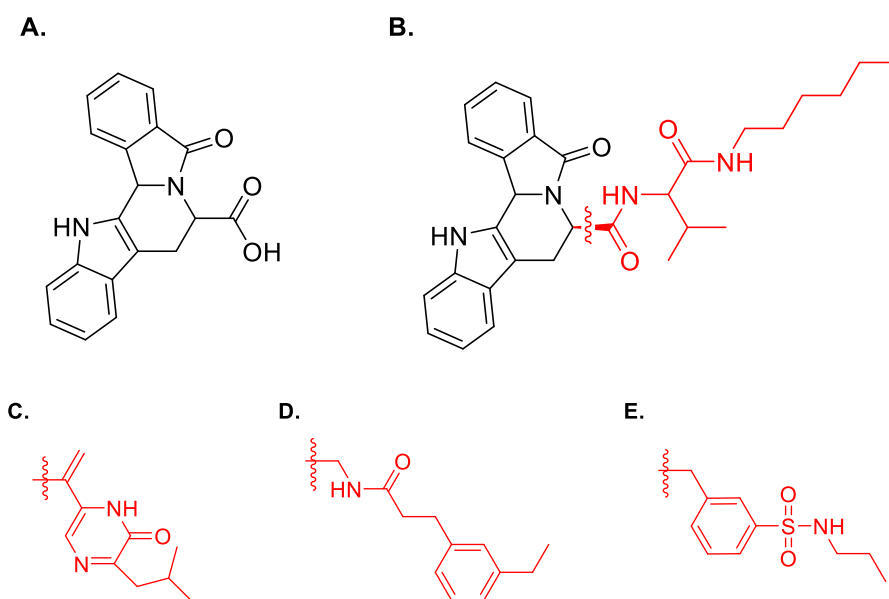


Figure 2.4. (A) The simplest tetrahydro- β -carboline scaffold (B) Zinc8876405 with side chain to be replaced to make bioisosteres highlighted red. (C-E) generated bioisosteric side chains shown in red to replace

The second compound (ZINC000000254599) also showed growth inhibitory activity against the ESKAPE pathogens. However, we determined this to be a PAINS compound. Compound two is a Michael acceptor and alkylating agent, which is why it shows promiscuous activity, meaning that it is not inhibiting the growth of the

Chapter 2: Database Curation for ML in Antibiotic Discovery

ESKAPE pathogens selectively but instead will kill the growth of any cell it comes into contact with. Furthermore, the presence of Michael acceptors in our top one hundred compounds from our natural product screen indicates that there are PAINS compounds in the initial training set we are using for our CHEMPROP model.

Having PAINS compounds present in our model's training set is concerning as it means that we will predict PAINS compounds, and much time and money can be wasted on testing compounds that have no druggable value. To overcome this issue, PAINS were further examined, and Chapter 4 will discuss the work done to eliminate them from training datasets.

With the determination of PAINS being present in our initial dataset, our small training set was about to get smaller. As mentioned above, large training sets are not feasible with antibiotics as the number of active compounds is not high enough to have millions of compounds in a dataset. These two issues coupled together raise the question: “What is the smallest data set you can use to accurately train a ML algorithm?” To answer this question my collaborator from the Cardona lab compiled multiple libraries from the co-add library as summarized in Table 2.1. These nine libraries were curated not only for the bacteria or cell line they target but to be diverse and have different sizes and balances to determine what size and balance one needs to train an accurate ML model.

Chapter 2: Database Curation for ML in Antibiotic Discovery

Table 2.1 The nine libraries that were curated for this study, with target and MIC cutoff reported as well as size and balance

Library	Activity type	Strain/Cell-line	Cut-off point	Total compounds	# Active compounds	Active (%)	AUROC	PRC-AUC	F1
1	Antimicrobial	<i>P. aeruginosa</i> ATCC 27853	MIC $\leq$ 4 μ g/ml	5529	40	0.72	0.98	0.77	0
2	Antimicrobial	<i>P. aeruginosa</i> ATCC 27853	MIC $\leq$ 8 μ g/ml	4923	52	1.06	0.93	0.46	0.23
3	Antimicrobial	<i>S. aureus</i> ATCC 43300	MIC $\leq$ 4 μ g/ml	4794	228	4.76	0.86	0.2	0
4	cytotoxicity	HEK293	CC50 $\geq$ 32 μ g/ml	3972	1333	33.56	0.74	0.6	0.58
5	Hemolytic Activity	hRBC	HC50 $\geq$ 32 μ g/ml	2146	100	4.66	0.9	0.78	0.68
6	Antimicrobial	<i>P. aeruginosa</i> ATCC 27853	MIC $\leq$ 16 μ g/ml	4112	86	2.09	0.97	0.36	0.38
7	Antimicrobial	<i>S. aureus</i> ATCC 43300	MIC $\leq$ 16 μ g/ml	3930	796	20.25	0.78	0.55	0.45
8	Antimicrobial	<i>S. aureus</i> ATCC 43300	MIC $\leq$ 8 μ g/ml	4164	502	12.06	0.43	0.2	0.15
9	Antimicrobial	<i>P. aeruginosa</i> ATCC 27853	MIC $\leq$ 16 μ g/ml	27002	9390	34.78	0.49	0.39	0.31

Table 2.1 shows that libraries 1, 2, 6, and 9 were all tested against the same bacterial species, *Pseudomonas aeruginosa*, but are reported at different cutoff values for what is considered active, thus allowing for different sizes of libraries. The same can be said for libraries 3, 7, and 8. Libraries 4 and 5 look at cytotoxicity and hemolytic activity respectively. While there may be different targets each library had its “activity” binarized where 0 was considered no activity and 1 was considered active. From there each library underwent a random split where 80% of the molecules were used for training, 10% were used for testing and 10% were used for validation. To ensure active compounds in all three splits first the active compounds were split into the training test and validation sets, followed by the inactive compound. A Chemprop model was then trained, and the results can be seen in Table 2.1.

Chapter 2: Database Curation for ML in Antibiotic Discovery

The reason that PRC-AUC and F1 are used in conjunction is exemplified by the model trained by library 1. The library 1 trained model was able to produce a PRC-AUC of 0.77, this is nearly double that of the next best model in the field.²³ This model should not be used, however, as the F1 score is 0. What that means is that while there was great precision in our model it was unable to correctly classify any of the true positive cases in the test set, meaning the recall is 0. This is not good in an ML model and is likely because less than 1% of the compounds were active in this library. This means that likely the model predicted everything as an inactive compound and because 99% of the compounds were inactive it had great accuracy or in other words was overfit.

It should be noted that there were a few models (trained from libraries 4,5,7) that performed well. While these libraries may not have the best scores from a ML standpoint, they currently outperform the models people in the field are using in the prediction of antibiotics. Specifically, there is a paper by Wong et.al wherein a D-MPNN model is trained looking for antibiotic scaffolds, and they report successes they experimentally validated.²⁴ The experimental validation shows their model works, even though they do not report an F1 score, and their best PRC-AUC score is only 0.33. By comparison, 7 of our 9 models outperform their model, yet only 3 of our models have F1 scores to back it up, once again exemplifying the need for both metrics. While this alludes to being able to train ML models with small, diverse datasets, it also shows that a better balance is potentially more important than the

Chapter 2: Database Curation for ML in Antibiotic Discovery

size; this is seen by the fact that the models trained with libraries with the best balance between the number of actives and inactive performed the best, at least in the case of antibiotics. In the case of library 5, which looked at hemolytic activity, a small, highly unbalanced library performed better than any of the other models. This can lead one to conclude that while yes small imbalanced libraries can be utilized in ML training, the activity one hopes to predict is perhaps more important in determining whether or not a small library will be useful. This also led to the conclusion that perhaps we were asking too much of the model in terms of antibiotics, and instead of training on data from whole cell studies with unknown MOA, we should focus on compounds with one MOA. This was further examined by the work in Chapter 3.

2.4 Conclusions

This work examined *in silico* chemical libraries, specifically what needs to be included in the generation of libraries for training ML algorithms. Initially, a library of natural products was curated and run through our ML algorithm, giving both positive results in the case of tetrahydro- β -carboline and indicating the initial dataset was plagued with PAINS in the case of Michael acceptors. This allowed for my work to move in two directions simultaneously. The first direction was the investigation into the compound ZINC8876405 and the activity it showed, and the second direction was the investigation into both filtering out PAINS compounds and examining the size and weight of a compound library,

Chapter 2: Database Curation for ML in Antibiotic Discovery

Although my work with the tetrahydro- β -carboline is ongoing, the investigation into PAINS led to the work discussed in Chapter 4. As for the investigation into the size and balance of ML training datasets, many interesting results were observed. First, it was observed that libraries 4 and 5 outperformed the training of our ML algorithm when compared to the training done all other libraries. It is important that libraries 4 and 5 outperform the other seven libraries as they look at hemolysis and cytotoxicity not antibiotics. Interestingly both the size of the database and balance of the library 5 dataset were smaller than the trainings of the other libraries, yet library 5 performed the best in my trainings. This indicates that each target requires a different amount of data to have an algorithm trained to make predictions. While that is not what I initially set out to see I find this to be important when it comes to library curation as it shows there is not one size fits all model.

As for the trainings of the algorithm specifically for growth inhibition prediction, it was determined that balance was more important than library size. This is seen when comparing Library 7 to Libraries 3 and 8. All three of the libraries had the same target bacterium of *s. aureus*. However, library 7 was the smallest yet most balanced library of the three. Library 7 was able to achieve “okay” results despite the lack of size. It should be noted that I describe the results as “okay” because, from a ML standpoint, they are not good. That being said the training of the algorithm produced better AUC-PRC results than what is currently published in the ML in the antibiotic field.

Chapter 2: Database Curation for ML in Antibiotic Discovery

2.5 References

1. Antimicrobial resistance | World Health Organization.
<https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance>
(accessed 03-11-2025)
2. Silver, L. L. Challenges of Antibacterial Discovery. *Clin. Microbiol. Rev.* **2011**, 24 (1), 71–109. <https://doi.org/10.1128/CMR.00030-10>.
3. Booq, R. Y.; Tawfik, E. A.; Alfassam, H. A.; Alfahad, A. J.; Alyamani, E. J.
Assessment of the Antibacterial Efficacy of Halicin against Pathogenic Bacteria. *Antibiotics* **2021**, 10 (12), 1480.
<https://doi.org/10.3390/antibiotics10121480>.
4. Liu, G. ; Catacutan, D. B. ; Rathod, K. ; Swanson, K. ; Jin, W. ; Mohammed, J. C. ; Chiappino-Pepe, A.; Syed, S. A.; Fragis, M.; Rachwalski, K.; Magolan, J.; Surette, M. G.; Coombes, B. K.; Jaakkola, T.; Barzilay, R.; Collins, J. J.; Stokes, M. M. Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nat Chem Biol* **2023**, 1342–1350, <https://doi-org.uml.idm.oclc.org/10.1038/s41589-023-01349-8>
5. Veltri, D.; Kamath, U.; Shehu, A. Deep Learning Improves Antimicrobial Peptide Recognition. *Bioinformatics* **2018**, 34 (16), 2740.
<https://doi.org/10.1093/bioinformatics/bty179>.
6. Fischbach, M. A.; Walsh, C. T. Antibiotics For Emerging Pathogens. *Science* 2009, 325 (5944), 1089–1093. <https://doi.org/10.1126/science.1176667>.

Chapter 2: Database Curation for ML in Antibiotic Discovery

7. Tripathi, H. What Is Balance And Imbalance Dataset?. Analytics Vidhya.
<https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5> (accessed 2024-05-20).
8. Neu, H. C.; Gootz, T. D. Antimicrobial Chemotherapy. In Medical Microbiology; Baron, S., Ed.; University of Texas Medical Branch at Galveston: Galveston (TX), 1996.
9. Lipinski, C. A. Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* **2000**, 44 (1), 235–249.
[https://doi.org/10.1016/S1056-8719\(00\)00107-6](https://doi.org/10.1016/S1056-8719(00)00107-6).
10. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **2001**, 46 (1), 3–26.
[https://doi.org/10.1016/S0169-409X\(00\)00129-0](https://doi.org/10.1016/S0169-409X(00)00129-0).
11. Richter, M. F.; Hergenrother, P. J. The Challenge of Converting Gram-Positive-Only Compounds into Broad-Spectrum Antibiotics. *Ann. N. Y. Acad. Sci.* **2019**, 1435 (1), 18–38. <https://doi.org/10.1111/nyas.13598>.
12. Desselle, M. R.; Neale, R.; Hansford, K. A.; Zuegg, J.; Elliott, A. G.; Cooper, M. A.; Blaskovich, M. A. Institutional Profile: Community for Open Antimicrobial Drug Discovery – Crowdsourcing New Antibiotics and Antifungals. *Future Sci. OA* **2017**, 3 (2), FSO171. <https://doi.org/10.4155/fsoa-2016-0093>.

Chapter 2: Database Curation for ML in Antibiotic Discovery

13. Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, 60 (12), 6065–6073. <https://doi.org/10.1021/acs.jcim.0c00675>.
14. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, 51 (D1), D1373–D1380. <https://doi.org/10.1093/nar/gkac956>.
15. Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; Magarinos, M. P.; Bosc, N.; Arcila, R.; Kizilören, T.; Gaulton, A.; Bento, A. P.; Adasme, M. F.; Monecke, P.; Landrum, G. A.; Leach, A. R. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2024**, 52 (D1), D1180–D1192. <https://doi.org/10.1093/nar/gkad1004>.
16. Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, 50 (1), 74–82. <https://doi.org/10.1021/jm0603365>.
17. Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of Diverse Database Subsets Using Property-Based and Fragment-Based Molecular Descriptions. *Quant.*

Chapter 2: Database Curation for ML in Antibiotic Discovery

Struct.-Act. Relatsh. **2002**, 21 (6), 598–604.

<https://doi.org/10.1002/qsar.200290002>.

18. RDKit: Open-Source Cheminformatics. <https://www.rdkit.org>.

19. Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.;

Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, 64 (1), 9–17.

<https://doi.org/10.1021/acs.jcim.3c01250>.

20. Liu, C.; Hogan, A. M.; Sturm, H.; Khan, M. W.; Islam, Md. M.; Rahman, A. S. M.

Z.; Davis, R.; Cardona, S. T.; Hu, P. Deep Learning-Driven Prediction of Drug Mechanism of Action from Large-Scale Chemical-Genetic Interaction

Profiles. *J. Cheminformatics* **2022**, 14 (1), 12. <https://doi.org/10.1186/s13321-022-00596-6>.

21. Johnson, E. O.; LaVerriere, E.; Office, E.; Stanley, M.; Meyer, E.; Kawate, T.;

Gomez, J. E.; Audette, R. E.; Bandyopadhyay, N.; Betancourt, N.; Delano, K.;

Da Silva, I.; Davis, J.; Gallo, C.; Gardner, M.; Golas, A. J.; Guinn, K. M.;

Kennedy, S.; Korn, R.; McConnell, J. A.; Moss, C. E.; Murphy, K. C.; Nietupski,

R. M.; Papavinasasundaram, K. G.; Pinkham, J. T.; Pino, P. A.; Proulx, M. K.;

Ruecker, N.; Song, N.; Thompson, M.; Trujillo, C.; Wakabayashi, S.; Wallach,

J. B.; Watson, C.; Ioerger, T. R.; Lander, E. S.; Hubbard, B. K.; Serrano-Wu, M.

H.; Ehrt, S.; Fitzgerald, M.; Rubin, E. J.; Sassetti, C. M.; Schnappinger, D.;

Hung, D. T. Large-Scale Chemical–Genetics Yields New M. Tuberculosis

Chapter 2: Database Curation for ML in Antibiotic Discovery

Inhibitor Classes. *Nature* **2019**, 571 (7763), 72–78.

<https://doi.org/10.1038/s41586-019-1315-z>.

22. Rahman A. S. M. Z., Liu C., Sturm H., Hogan A. M., Davis R., Hu P., Cardona S.

T., “A machine learning model trained on a high-throughput antibacterial

screen increases the hit rate of drug discovery” *PLoS Comput. Biol.* **2022**, 18,

e1010613

23. BROOD 3.2.2.3. OpenEye, Cadence Molecular Sciences, Santa Fe, NM.

[Http://Www.Eyesopen.Com](http://www.eyesopen.com).

24. Wong, F.; Zheng, E. J.; Valeri, J. A.; Donghia, N. M.; Anahtar, M. N.; Omori, S.;

Li, A.; Cubillos-Ruiz, A.; Krishnan, A.; Jin, W.; Manson, A. L.; Friedrichs, J.;

Helbig, R.; Hajian, B.; Fiejtek, D. K.; Wagner, F. F.; Soutter, H. H.; Earl, A. M.;

Stokes, J. M.; Renner, L. D.; Collins, J. J. Discovery of a Structural Class of

Antibiotics with Explainable Deep Learning. *Nature* **2024**, 626 (7997), 177–

185. <https://doi.org/10.1038/s41586-023-06887-8>.

Chapter 3: Training Interpretation Using XAI

3.1 Introduction

Machine learning models in antibiotic discovery have been largely focused on whole-cell studies for the discovery of antibiotics. This approach is advantageous as it often leads to the discovery of compounds that are more likely to be druggable, given their ability to penetrate the cell. However, the issue with this method is that it neglects MOA - with all active compounds having the same features, despite having vastly different mechanisms of action. To date, very little work has been done in using ML methods to identify compounds for a specific MOA. This is likely due to the lack of activity data for diverse sets of molecules for many of the known MOAs. The rationale of training a model on compounds with one MOA is that a stronger trend in the data should be observed as the compounds are binding to one active site as opposed meaning theoretically there should be one structural or electronic feature shared by all active molecules compared to other studies where compounds can bind to one or more of many enzyme targets that are located within the cell of study.

With the above rational in mind a target needed to be picked wherein a lot of activity data is available to that end AmpC was chosen as the target. AmpC beta-lactamases are a class of cephalosporinases that confer resistance to a broad spectrum of beta-lactam antibiotics, including penicillin's, cephalosporins, and beta-lactamase inhibitors.^{1,2} The mechanism of action of AmpC beta-lactamases involves the hydrolysis of the beta-lactam ring, a key structural feature responsible for the antibiotic's bactericidal activity.¹ A serine residue in the active site of AmpC

Chapter 3: Training Interpretation Using XAI

nucleophilicity attacks the beta-lactam ring, leading to its cleavage and rendering the antibiotic inactive. If inhibition of the AmpC enzyme can be achieved resistance to β -lactam antibiotics should also decrease.

The discovery of new inhibitors targeting AmpC beta-lactamases has been challenging. Expression of AmpC is generally low but is induced by the presence of β -lactam and unlike some other beta-lactamases, AmpC is not typically inhibited as well by standard beta-lactamase inhibitors like clavulanic acid.¹⁻³ Given the importance of this enzyme and the challenges associated with finding effective, selective inhibitors of AmpC, large screens of small molecules have been performed to identify potential inhibitors. Recent efforts have focused on developing both covalent and non-covalent inhibitors to counteract this mechanism of resistance.²⁻⁴ However, the clinically significant inhibitors developed in recent years remain those that act nonselectively against multiple beta-lactamases (e.g. avibactam, which is effective against a broad range of serine beta-lactamases, including AmpC).

As discussed in previous chapters, ML models have utility in chemistry and can aid in drug discovery. Not mentioned in those chapters is that there is the fact that many ML models used are black boxes when it comes to why they make the decisions they make. An example of this is the data presented in Chapter 2. While the compounds predicted were able to be tested in the lab by collaborators, we have no definitive answer as to why the molecules that were tested were chosen by the model as potentially having activity. This is where XAI shows its utility, as now

Chapter 3: Training Interpretation Using XAI

we can get the model to not only predict activity but tell us why it is predicting that activity. The explanations generated can then be examined and trends and commonalities can be compiled with the hope of finding a better drug.

The XAI model used in this chapter is called the MEGAN (Multi-Explanation Graph Attention Network) model. The model works by having the user input data, in the case of this study SMILES strings that are then encoded into graphs that the model can read. These graphs are just complex matrices where in atoms are represented as nodes and bonds are represented as edges. Once graphs are produced the model is trained and able to learn what it deems as important to a compound's classification. This is where MEGAN and other XAI models differ, once training is completed with most XAI a secondary algorithm is then run on the output data for explanations to be generated. MEGAN on the other hand takes it a step further by first creating masks for the edges and nodes before producing an output for the user. These masks represent explanations as to the decision-making process of the GAT and are represented visually in the form of highlighted bonds and atoms. These explanations describe visually what the model has learned as important features of the molecule and an example can be seen in Figure 3.1.

Chapter 3: Training Interpretation Using XAI

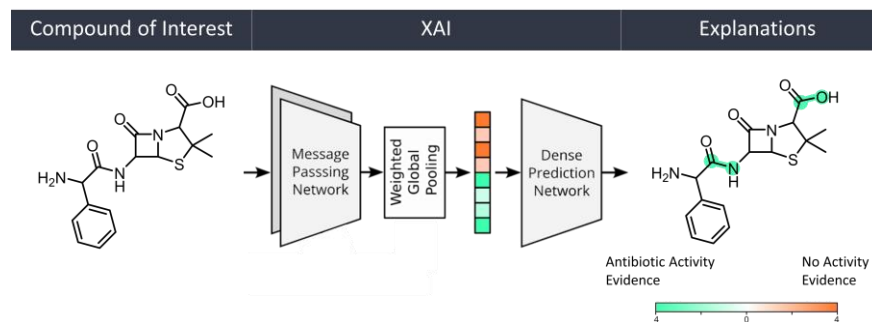


Figure 3.1 A simplified MEGAN architecture and what explanations look like from a trained MEGAN model

Apart from the XAI MEGAN model, a RF model was trained in this study to be used for a baseline that the GAT could be compared to.

3.2 Methods

3.2.1 Dataset preparation

The library that was ultimately chosen for this was the AID-485341 (on PubChem) which contained 331,108 compounds that were experimentally tested as inhibitors of AmpC.⁵ Of the 331,108 compounds only 1730 were considered active as they had a AID_ACTIVITY_SCORE larger than 40. PUBCHEM_ACTIVITY_SCORE is a metric PubChem requires where the results from a HTS are normalized and given a score between 0-100; Anything that has a score greater than 40 is considered to be active.⁶ With only 1730 molecules being active, the library balance was 0.5% which would make it difficult to train a model that is not overfit. To overcome that issue, the compounds were split into active and inactive groups, and all active compounds were kept. The inactive compounds underwent 4 rounds of clustering using the CHEMFp software.⁷ In each round of clustering, clusters were generated using a

Chapter 3: Training Interpretation Using XAI

distance-based method on fingerprints. The center of each cluster was kept and utilized in the next round of clustering until only 35,886 diverse representative inactive compounds were left. This gave a total training set of 37,616 compounds with a new balance of 4.5% which is still highly imbalanced but is better than having less than 1% active compounds. Note: all compounds are reported in SMILES string notation and the dataset is cleaned following the procedure in Chapter 2. After cleaning, the input SMILES were transformed into graphs for the MEGAN model and Morgan fingerprints for the RF model.

3.2.2 Feature Determination

As mentioned in Chapter 2, the diversity of a training set is important for training a robust model. To analyze the diversity of the training set in section 3.2.1, “drug-like” metrics were once again calculated. Features looked at for this dataset included the number of hydrogen bond donors/acceptors, molecular weight, and logP (Figure 3.2).

3.2.3 Machine Learning Classification Modeling

The dataset identified above was used in training the MEGAN model (Described above) and a random forest model was also trained as a control.

3.2.4 Random Forest Training

The RF model was first tuned using the hyperparameters of max_depth, max_features, max_leaf_nodes, and n_estimators. When the model was sufficiently tuned, training occurred using the following hyperparameters: max_depth= None, max_features= None, max_leaf_nodes= None, n_estimators = 50. The data was

Chapter 3: Training Interpretation Using XAI

randomly split with 201 compounds being taken into the test set (While this is a small number of molecules this was kept in order to be consistent with the size of the test set for the MEGAN model). Using the above metrics training occurred and the results are reported below (Figure 3.3).

3.2.5 MEGAN Training

For training a MEGAN model the data was first split into three groups. The first group was 200 randomly chosen molecules wherein the criteria that 100 molecules were active, and 100 molecules were inactive had to be met, and this was to be called the validation set. The second group had the same split and criteria but was labeled as the test set. Finally, the rest of the molecules were kept as a training set. For training, oversampling of the minority class occurred wherein every molecule in the minority class (active) was duplicated 26 times in order to bring balance to the dataset. 30 epochs were used (meaning the model was trained 30 times) and the model evaluated the results of each epoch using the validation set. At the end of the training, the epoch with the best performance was taken as the trained model.

3.3 Results

3.3.1 Training Set Diversity

To ensure a diverse representative subset of the compounds from the original dataset was being used, the features outlined in Figure 3.2 were examined for both the active and inactive compounds. These features consist of molecular weight, log p, and hydrogen bond donors/acceptors as described in Chapter 2.

Chapter 3: Training Interpretation Using XAI

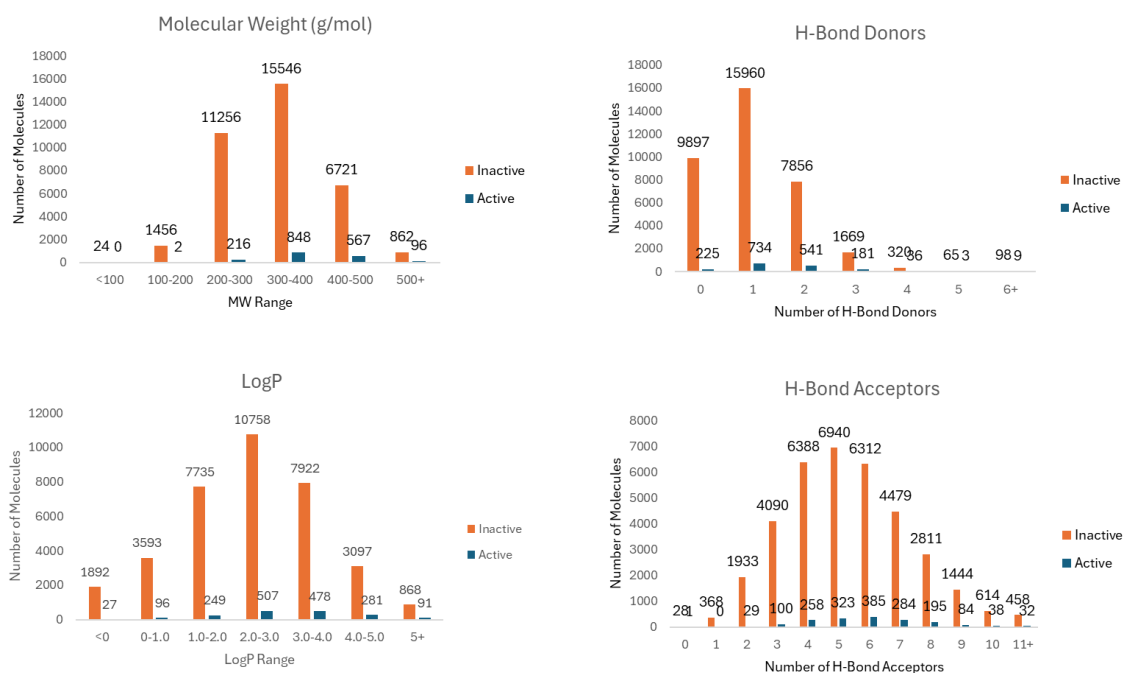


Figure 3.2 Bar graphs of the number of molecules (y-axis) compared to instances of a feature (x-axis) for active (blue) and inactive (orange) compounds. Features examined are molecular weight (top left) hydrogen bond donors (top right), LogP (bottom left), and hydrogen bond acceptors (bottom right).

The plots in Figure 3.2 show that the active and inactive compounds both had Gaussian distributions for the selected features meaning they were diverse in regard to those features. The gaussian distribution implies diversity as it shows that a range of compounds with the selected feature is present in the training data for both active and inactive compounds. Interestingly, the graphs showed the same trends between active and inactive compounds for all features examined except hydrogen bond acceptors, wherein the active β -lactamase inhibitors had on average more hydrogen bond acceptors than the inactive compounds. This shows that while our

Chapter 3: Training Interpretation Using XAI

active and inactive chemicals are in the same chemical space (based on size and lipophilicity) there are differences within the molecules that can allow for differentiation between the two classes.

3.3.2 Model Performance

With the diversity examination complete, training commenced on both a RF and MEGAN model. A Random Forest model was trained as a baseline as it has been shown that RF models tend to outperform neural networks for many tasks, therefore to prove the utility of a model one should show it outperforms the RF.⁸ The MEGAN model was chosen for this task as previous work (as described in Chapter 4) has shown the utility of this model in the classification of small molecules concerning their ability to form colloidal aggregates, as well as other molecular features such as mutagenicity. Figure 3.3 shows the confusion matrices that were generated from training the two models and shows that MEGAN outperforms the RF model.

Chapter 3: Training Interpretation Using XAI

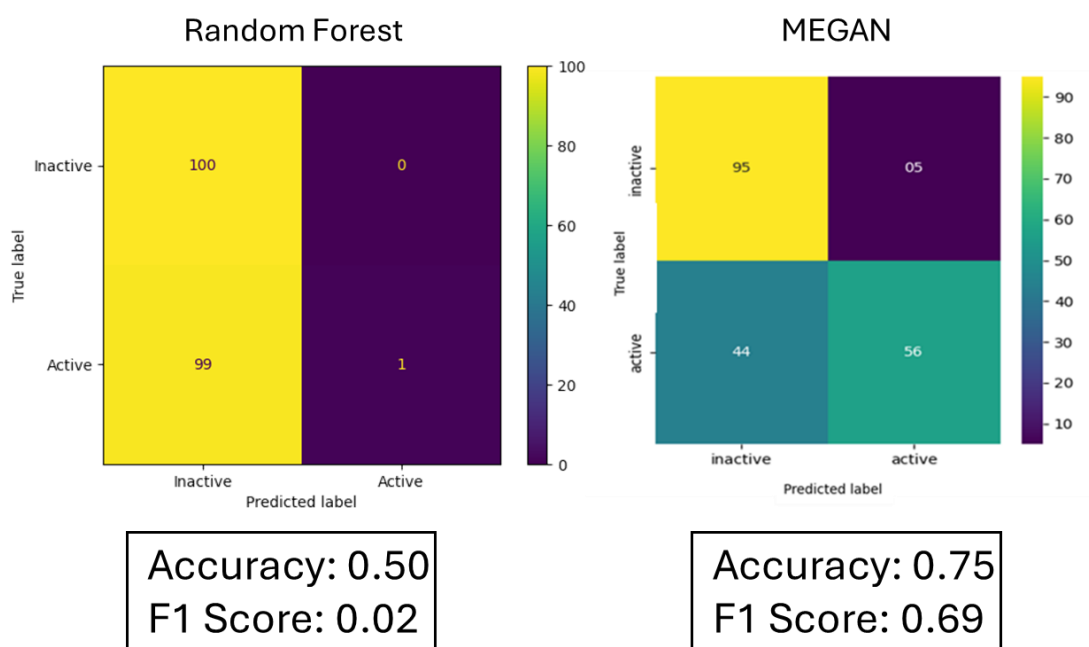


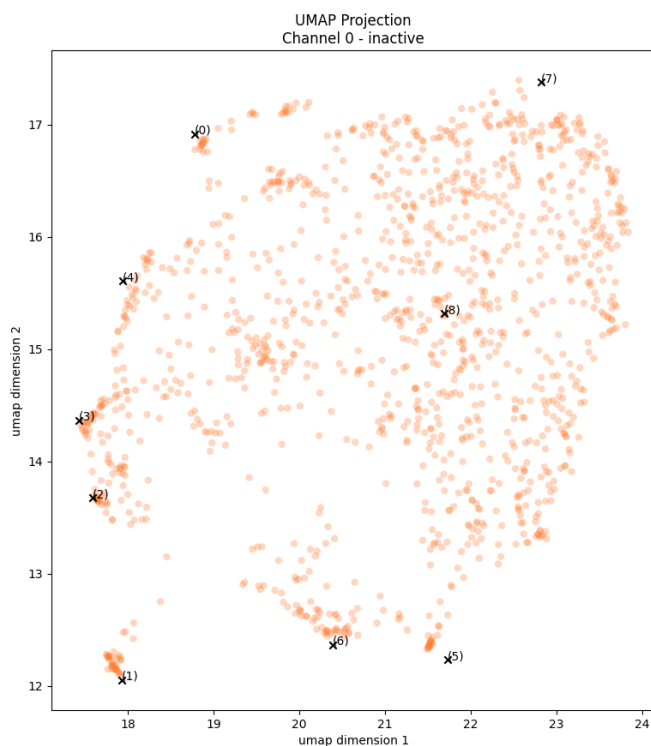
Figure 3.3 The confusion matrix, accuracy, and F1 score for the MEGAN (left) and RF (right) models.

While the RF model had a higher accuracy the MEGAN model outperforms the RF model as our training datasets are highly imbalanced. An F1 score (as calculated by eq 3 in Chapter 1) should be used as a metric to judge model performance. The F1 score for the MEGAN model was calculated to be 0.69 while the F1 score of the RF model was 0.02. While it is clear MEGAN outperforms the RF model MEGAN is the first of its kind in the sense that it is targeting one specific MOA of antibiotic resistance, so it is unable to be compared to other ML models in publications. Future work outlined in Chapter 5 will look at the predictive power of this model.

Chapter 3: Training Interpretation Using XAI

3.3.3 Explanation Examination

With training complete, the explanations provided by MEGAN were examined to determine if the model produced any chemically relevant insights. To examine the results of the training UMAP (Uniform Manifold Approximation & Projection) clustering was performed on the explanations that were generated, and 34 clusters were identified for validation as seen in Figure 3.4. It should be noted that the MEGAN model clusters each molecule in the entire dataset twice, once with the explanations or masks that contribute to the inactivity of the molecule, and once for the masks that contribute to the activity of the molecule. In this study, active refers to compounds that inhibit AmpC and inactive is compounds that do not inhibit AmpC.



Chapter 3: Training Interpretation Using XAI

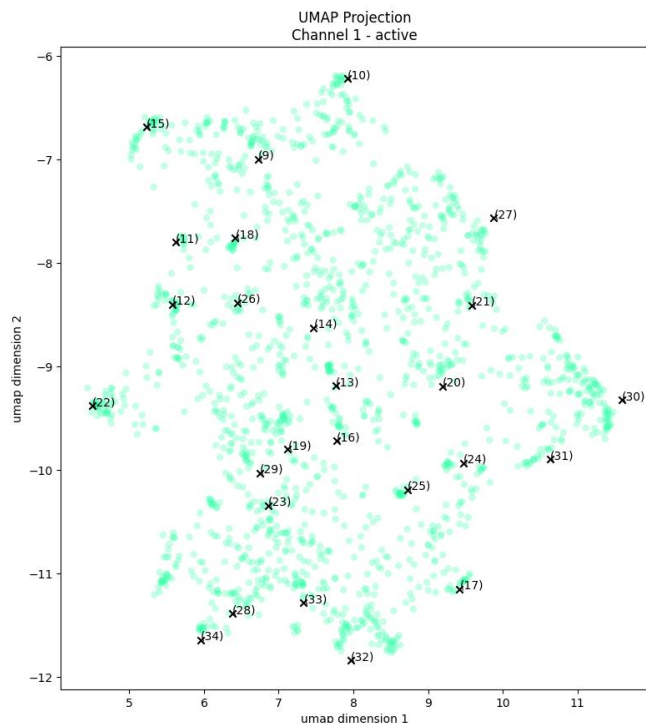


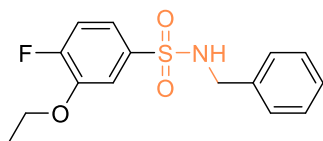
Figure 3.4 UMAP projection for predicted inactive masks(top) and active masks(bottom) the numbered x's represent the clusters present in the data as given by MEGAN

Upon cluster analysis, some trends were noted as seen in Figure 3.5 which shows representative explanations for compound clusters.

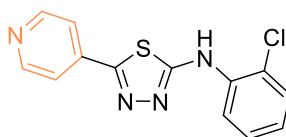
Chapter 3: Training Interpretation Using XAI

Inactive Clusters

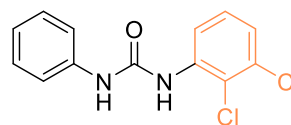
0



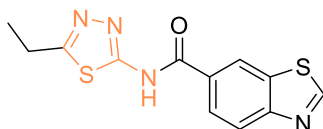
1



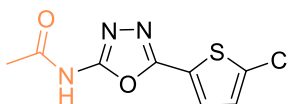
2



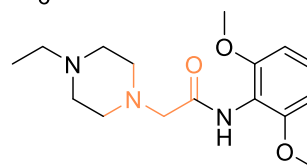
4



5

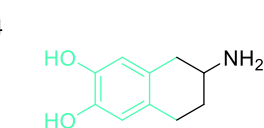
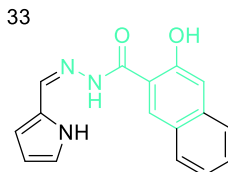
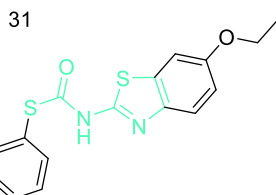
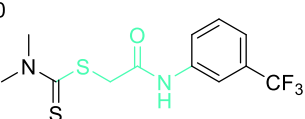
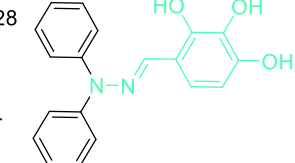
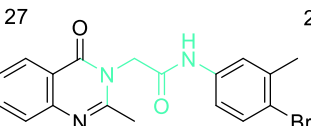
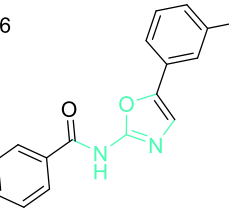
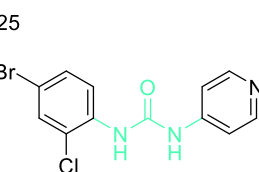
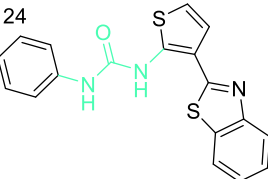
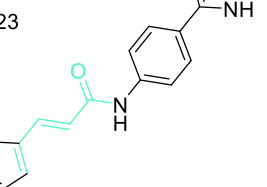
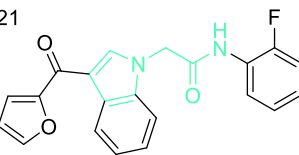
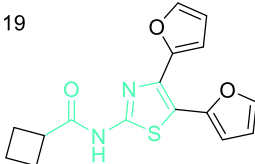
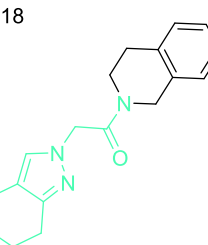
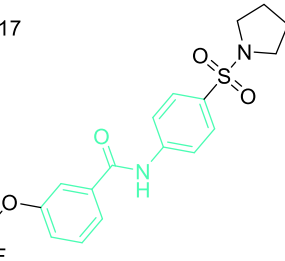
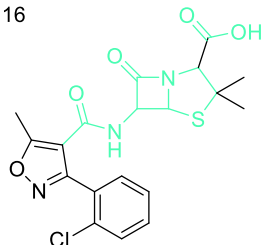
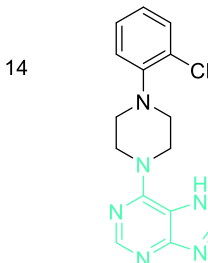
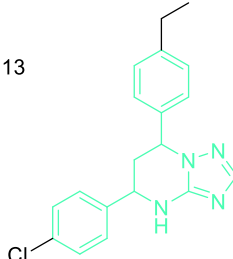
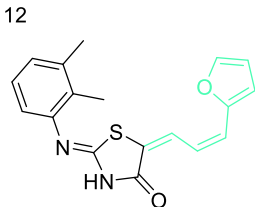
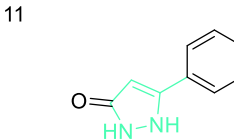
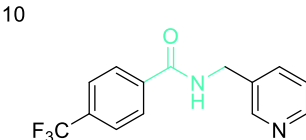
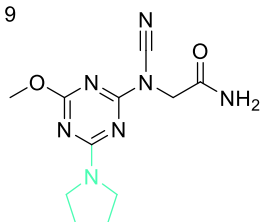


6



Chapter 3: Training Interpretation Using XAI

Active Clusters



Chapter 3: Training Interpretation Using XAI

Figure 3.5 Representative examples of explanations (colored atoms/bonds) for each cluster with cluster number (inactive clusters are masks that contribute to inactivity and active clusters show masks that contribute to activity) note only clusters with one common explanation shown.

The first trend that was noted was for the inactive clusters. Where it was noted that MEGAN identified halogens, and sulfonyl groups as contributing to a compound not having inhibitory activity against Ampc. Alternatively, when looking at the clusters of active compounds MEGAN identified tertiary amines, secondary amides, urea, and conjugated systems as all contributing to a compound's activity. It should be noted the functional groups located in the above trends are present in both active and inactive molecules and due to the imbalance of the dataset all functional groups appear more often in the inactive compounds, which is a testament to the model being able to identify which features are present in the active compounds.

The scaffold shown in cluster 16 figure 3.5 is a scaffold for an active cluster containing compounds that have a β -lactam. This cluster was examined further as it is widely known that the β -lactam is responsible for the activity of the compounds.

3.3.4 Cluster Analysis

With the model performing well based on the reference RF model, the generated explanations needed to be validated as they will provide insight into the performance of the explainability aspect of the model. The explanations were

Chapter 3: Training Interpretation Using XAI

examined by choosing to look at cluster 16 which was the cluster that contained β -lactam compounds. MEGAN was able to correctly identify the β -lactam as being important in the model's decision to classify it as active (Figure 3.6) Further analysis of the scaffolds in this cluster revealed that the model also identified the amide adjacent to the β -lactam ring and the carboxylic acid group to also be important to their classification as an inhibitor of AmpC. This was of special interest, as only 58 compounds in the original ~37K used for training contained a β -lactam ring, and of those only 13 were active. This shows that the model is not overfit with this data, as most of the β -lactam containing compounds are inactive and even then, only 0.15% of the dataset contains this substructure, meaning MEGAN was able to identify chemically relevant information about these molecules despite never seeing an active site or literature on these compounds.

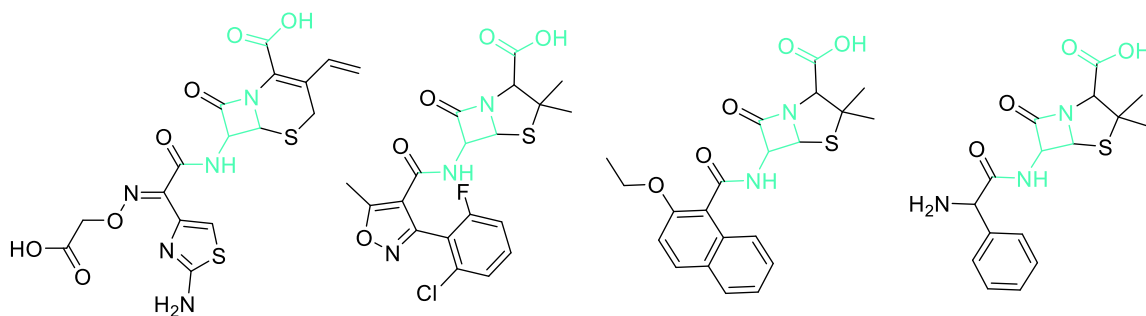


Figure 3.6 Representative compounds from the cluster of β -lactam containing compounds with common explanation shown, note all 11 cluster members showed the above explanation for activity.

Chapter 3: Training Interpretation Using XAI

With the initial analysis completed showing the model's ability to learn, a closer look was taken at cluster 16 as the amide and acid highlighted in Figure 3.6 were being used as an explanation of activity for these molecules that was just as important as the β -lactam ring itself. Furthermore, this was an interesting development as there is not much literature where these features of the molecules are discussed. To determine if these features highlighted by our model were important in the binding of small molecules to the enzyme active site, a crystal structure of oxacillin bound in the active site of AmpC (pdbid=4JXG) was examined (Figure 3.7).⁹ This ligand was chosen as it was contained in the cluster and was structurally similar to many of the other compounds present in cluster 16.

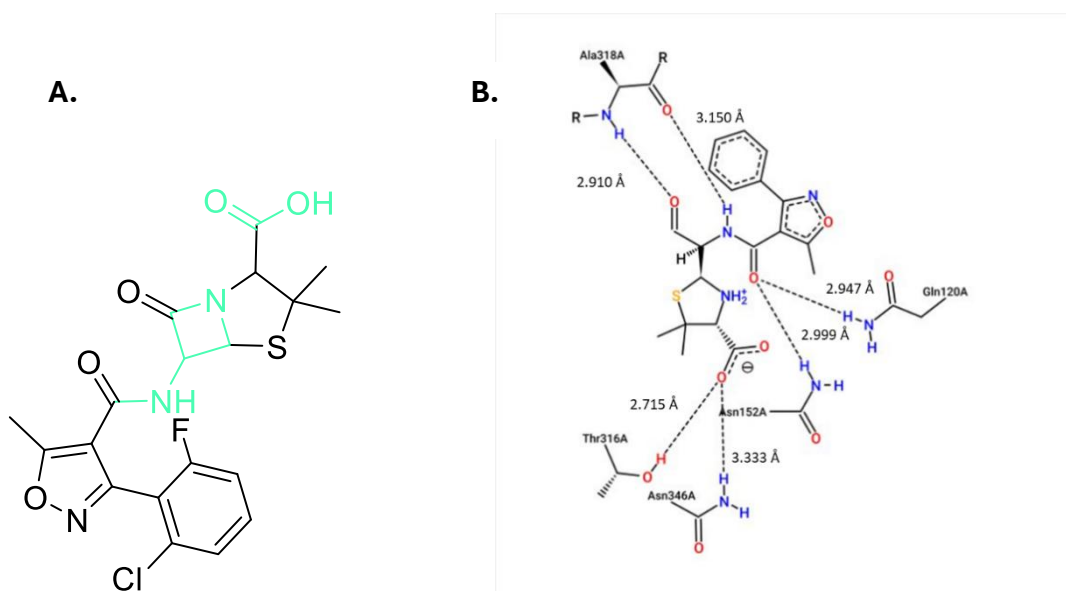


Figure 3.7 A) The oxacillin derivative with explanations generated by MEGAN compared to B) the interactions present in the active site crystal structure of oxacillin bound to AmpC with bond lengths for hydrogen bonds(right)⁹

Chapter 3: Training Interpretation Using XAI

Figure 3.7 shows that the portions of the molecule identified as being important to the model's decision to classify the oxacillin derivative as active are important to the interactions of the small molecule in the active site of Ampc. Firstly, the model identified the β -lactam ring as important. This is widely known to be true as it is responsible for binding in the active site of Ampc. Another portion of the molecule identified is the carboxylic acid that is responsible for hydrogen bonding, specifically to THR 316 and ASN 346 in the active site. Finally, the model identifies the amide adjacent to the β -lactam ring as important, this was found to be a hydrogen bond to ALA 318, GLN 120, and ASN 152. This is an exciting discovery as the model was trained from SMILE strings with a binary active or inactive and had never seen a protein active site yet was able to identify which portions of the molecule are responsible for the hydrogen bonding as well as the β -lactam itself as being important to the model's decision.

3.4 Conclusion

The work presented in this chapter shows that XAI models can give chemical insight into molecules based on the limited data given to the model in training. This chapter shows that MEGAN can be trained using a library of β -lactamase inhibitors to achieve performance that is better than a RF model as well as better than what is currently in literature sources. It should be noted that while the explanations generated for the β -lactam containing compounds were correct, future work will study some of the other clusters to gain further insights into inhibitors of Ampc.

Chapter 3: Training Interpretation Using XAI

3.5 References

1. Tamma, P. D.; Doi, Y.; Bonomo, R. A.; Johnson, J. K.; Simner, P. J. A Primer on AmpC β -Lactamases: Necessary Knowledge for an Increasingly Multidrug-Resistant World. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **2019**, 69 (8), 1446–1455. <https://doi.org/10.1093/cid/ciz173>.
2. Tooke, C. L.; Hinchliffe, P.; Bragginton, E. C.; Colenso, C. K.; Hirvonen, V. H. A.; Takebayashi, Y.; Spencer, J. β -Lactamases and β -Lactamase Inhibitors in the 21st Century. *J. Mol. Biol.* **2019**, 431 (18), 3472–3500. <https://doi.org/10.1016/j.jmb.2019.04.002>.
3. Romero, E.; Oueslati, S.; Benchekroun, M.; D'Hollander, A. C. A.; Ventre, S.; Vijayakumar, K.; Minard, C.; Exilie, C.; Tlili, L.; Retailleau, P.; Zavala, A.; Elisée, E.; Selwa, E.; Nguyen, L. A.; Pruvost, A.; Naas, T.; Iorga, B. I.; Dodd, R. H.; Cariou, K. Azetidinimines as a Novel Series of Non-Covalent Broad-Spectrum Inhibitors of β -Lactamases with Submicromolar Activities against Carbapenemases KPC-2 (Class A), NDM-1 (Class B) and OXA-48 (Class D). *Eur. J. Med. Chem.* **2021**, 219, 113418. <https://doi.org/10.1016/j.ejmech.2021.113418>.
4. Powers, R. A.; Morandi, F.; Shoichet, B. K. Structure-Based Discovery of a Novel, Noncovalent Inhibitor of AmpC β -Lactamase. *Structure* **2002**, 10 (7), 1013–1023. [https://doi.org/10.1016/S0969-2126\(02\)00799-2](https://doi.org/10.1016/S0969-2126(02)00799-2).

Chapter 3: Training Interpretation Using XAI

5. AID 485341.

<https://pubchem.ncbi.nlm.nih.gov/bioassay/485341#section=BioAssay-Annotations> (accessed 2024-09-17).

6. PubChem. BioAssay Tag Names.

<https://pubchem.ncbi.nlm.nih.gov/docs/bioassay-tag-names> (accessed 2024-09-22).

7. Dalke, A. The Chemfp Project. *J. Cheminformatics* **2019**, 11 (1), 76.

<https://doi.org/10.1186/s13321-019-0398-8>.

8. Fernandez-Delgado, M.; Cernadas, E.; Barro, S.; Amorim, D. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *JMLR*

2014, 15, 3133-3181

9. Waller B. J.; Powers, R. A.; Docter, B. E. Complexed structures of AmpC beta-lactamase. *RCSB PDB*. **2014**.

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

4.1 Introduction

Interest in the application of machine learning in lead discovery has grown substantially in recent years, driven by academic and industrial initiatives to apply ML methods during early-stage drug discovery.^{1,2} This trend is largely attributed to the availability of extensive datasets containing activity data generated through high-throughput screening campaigns. The activity data produced by HTS has long been essential for hit identification in early-stage drug discovery and is becoming even more critical with the growing interest in ML approaches for predicting lead compounds.

A persistent challenge in HTS-based hit identification is the prevalence of false hits. While large-scale HTS campaigns typically generate numerous initial hits, only a small proportion represents the desired interactions between compounds and their target biomolecules. Many screening libraries contain a significant number of false positive and negative data points, with up to 80–95% of the hits from initial screening representing artifacts.^{3,4} This long-standing challenge in medicinal chemistry now extends to ML-based approaches for drug discovery. Models trained on datasets containing large numbers of false hits are prone to predicting compounds that are not viable as leads. Addressing these issues in data quality is essential to enhance the hit discovery efforts of medicinal chemists as well as the predictive prowess of ML models.

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

Colloidal aggregation represents a significant source of false positives in HTS.⁵ Aggregation occurs when molecules form supramolecular complexes, or colloids, at or above a critical aggregation concentration.⁶ Small colloiddally aggregating molecules can interact nonspecifically with proteins, leading to local unfolding and functional disruption, or they can interfere through mechanisms such as aggregation-induced emission, where self-assembled molecules fluoresce upon reaching their critical aggregation concentration.^{6, 7} Estimates suggest that 15–20% of small molecules in public chemogenomic databases aggregate under standard screening conditions, underscoring the need for accurate prediction of aggregation to mitigate its impact on drug discovery.⁸

The experimental detection of SCAMS is both expensive and time-consuming which has led to numerous *in silico* methods being developed to screen aggregating compounds from HTS datasets (Figure 4.1).⁸⁻¹⁵ One of the earliest tools, Aggregator Advisor, evaluates molecules represented as SMILES strings and determines their similarity to known aggregators based on LogP and Tanimoto similarity (Figure 4.1a).⁹ In addition to providing a rule-based *in silico* method for aggregation screening, Aggregator Advisor has also provided the field with a valuable database of experimentally validated aggregators.⁹ This dataset has been pivotal in the development and training of most of the ML models for aggregation prediction,

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

enabling these models to classify small molecules as aggregators with greater accuracy and reliability (Figure 4.1b).^{8, 10-13}

a Experimental Aggregation Detection Methods

- Dynamic light scattering (DLS)
- Detergent-based biological assays
- UV-Vis spectroscopy

Pros

- ✓ Accurate detection
- ✓ Conditions variable

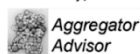
Cons

- ✗ Slow
- ✗ Expensive
- ✗ Non-predictive
- ✗ Individual molecule testing

b In silico Aggregation Prediction

Early Computational Approach

Aggregator Advisor



- Based on logP and similarity to experimentally known aggregators
- Can't extrapolate chemical patterns
- Source of data used to train ML models

Machine Learning Approaches



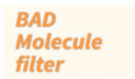
Deep SCAM
Neural Network
82% Accuracy



SCAM Detective
Gradient Boosting
77% Accuracy



ChemAgg
Gradient Boosting
94% Accuracy



BAD Molecule Filter
Gradient Boosting
84% Accuracy



Isometric Stratified Ensemble
Consensus Modeling
84% Accuracy

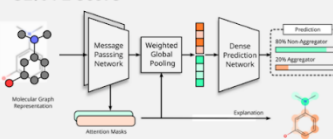
- Not easily scaled to filter large libraries of compounds
- Offer no insight into classification or alternatives

c This Work



MEGAN Aggregator

Explainable Graph Neural Network
82% Accuracy
81% F1 Score



- ✓ Database or single molecule screening
- ✓ Local and global explanations
- ✓ Counterfactual explanations
- ✓ Experimentally validated
- ✓ Open-source code
- ✓ Web interface

Figure 4.1. Summary of (a) experimental and (b) in silico approaches to aggregation determination (c) compared with our work. Accuracies are those reported in the respective publications (ref. 8-13) and are thus not directly comparable due to differences in training and testing datasets.

ML-based methods for the classification of aggregation have demonstrated accuracies exceeding 80% in predicting aggregators. However, these methods often do not provide scalable implementations or open-source code for filtering large

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

molecular libraries¹¹. Scalable approaches, such as SCAM Detective and DeepSCAM, typically achieve accuracies in the range of 70-80%.^{8, 10} Consequently, there remains a critical need for an accurate, scalable, and interpretable method capable of efficiently screening large libraries of compounds.

Aggregation is a complex phenomenon that is influenced by many variables, including concentration, pH, temperature, buffer, and solvent.^{16, 17} While numerous studies have explored the molecular features that contribute to small molecule aggregation, no clear consensus has been reached on the key factors driving this phenomenon.¹⁸ Features such as logP, the number of hydroxyl groups, the number of sulfur atoms, and the number of aromatic rings are often proposed to contribute to aggregate formation; however, the complexity of molecular aggregation has led to difficulty in identifying reliable, universal trends distinguishing aggregating and non-aggregating compounds.^{8, 9, 11}

This study addresses the need for an accurate and scalable model capable of detecting SCAMs while providing interpretable explanations and the ability to create non-aggregating counterfactuals. The explainable AI (xAI) model employed in this study, a multi-channel graph attention network (MEGAN), achieves an accuracy of 82% in predicting SCAMs and is suitable for screening both large molecular libraries and individual compounds.¹⁹ Furthermore, the model generates explanations for its classifications, offering insights into why a compound is

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

predicted as a SCAM or non-SCAM. The accompanying web server allows users to screen individual molecules and provides a user-specified number of counterfactual explanations. These counterfactuals are structurally similar to the query molecule but possess flipped classification labels (e.g., counterfactuals for a molecule predicted to aggregate are structurally similar molecules predicted to be non-aggregating, and vice versa). To validate our model and its application of counterfactuals, we synthesized and experimentally tested a non-aggregating derivative of clioquinol—an established aggregator—proposed by the model. The experimental results confirm the model's prediction and demonstrate the experimental relevance of our model in the informed design of molecules with tailored aggregation properties.

4.2 Results and Discussion

4.2.1 The MEGAN model

To develop an accurate and scalable model for detecting SCAMs and elucidating the structure-property relationships underlying molecular aggregation, we leverage xAI techniques applied to predictions generated by graph neural network models. To apply a graph neural network to the task of chemical property prediction, each molecule is first converted into a molecular graph where atoms are represented as nodes and bonds are represented as edges. Based on this graph-structured input information, the graph neural network is trained on the dataset to

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

predict the binary classification label of a given molecule as either an aggregator or non-aggregator (Figure 4.2).

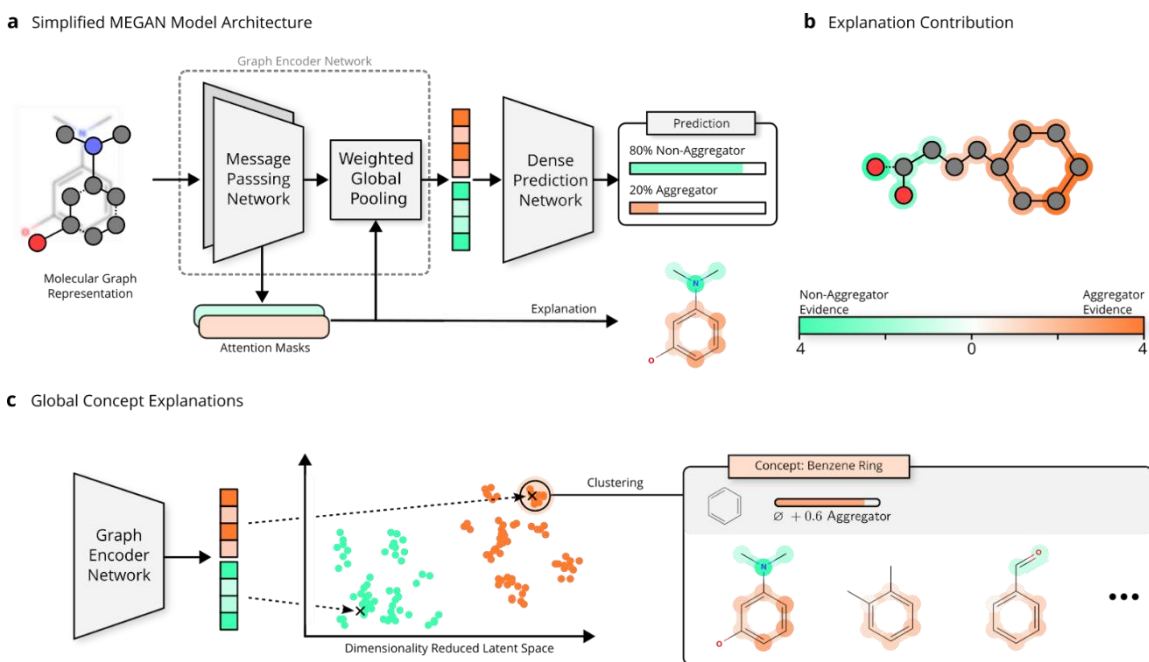


Figure 4.2. MEGAN model overview. (a) Molecular graph structures are used as input to multiple attention-based message-passing layers. Node representations are aggregated and passed to a fully connected network to output the predicted class. Explanation masks are derived from the internal attention values. (b) Attention-based explanations are explicitly split into separate channels for each possible output class. (c) Overarching structural explanations can be found by identifying clusters in the latent space of subgraph embeddings. Analyzing all members of a concept cluster yields general trends associated with certain

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

structural motifs. Explicit explanation masks and values are constructed for illustration purposes.

In this work, we apply the multi-explanation graph attention network (MEGAN) to the task of aggregation prediction.¹⁹ MEGAN is a self-explaining graph neural network model architecture for which node and edge attributional explanations are directly derived from the model's internal attention and masking mechanism. These attributional explanations assign an importance value between 0 and 1 to each node and edge of a given graph to indicate which substructures of a given graph are especially influential for the predicted outcome. The MEGAN model specifically generates one such attributional explanation for each of the possible classification outcomes—one explanation highlighting the structural evidence in favor of an aggregator classification (orange) and the other explanation highlighting substructures associated with the non-aggregator class (green) (Figure 4.2b).

4.2.2 Prediction accuracy and benchmarking

We trained a MEGAN model on a dataset comprised of 12,338 aggregating and 177,048 non-aggregating molecules (see Methods). For a quantitative evaluation of our trained model, we used a separate test set of 1500 aggregators and 1500 non-aggregators. Similar to most previously reported ML models, we achieved an accuracy of 82%. However, as we are using a training set with a class imbalance, accuracy can be misleading as it may achieve high accuracy simply by predicting

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

the majority class. To further evaluate the MEGAN model's performance, we employed the F1 score, which gives a more balanced performance estimate for both classes. The MEGAN model achieved an F1 score of 81%, indicating that the model has a good balance between precision and recall. The similarity in the accuracy (82%) and F1 score (81%) also suggests that the model's performance is well-balanced across both classes (aggregators and non-aggregators). The performance of the MEGAN model was compared against the XGboost model of Yang et al. (ChemAgg), using our balanced test set. It was found that the MEGAN model provided a higher accuracy and F1 score than the ChemAgg model (acc = 73%, F1 = 74%), demonstrating the superior performance of graph neural networks in capturing molecular structures.

4.2.3 Sensitivity of the model to small structural modifications

When evaluating the performance of the MEGAN model in predicting molecular aggregation, its predictions were observed to exhibit a high degree of sensitivity to subtle modifications in molecular structure. To illustrate this sensitivity, two groups of compounds with experimentally validated aggregation behavior are presented: one comprising molecules from the training dataset (**A**, **B**) and the other consisting of molecules external to the training dataset (**C**, **D**).

Azacarbazoles, **A** and **B**, were in the library used to train the model and have previously had their aggregation behavior experimentally validated (Figure 4.3).

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

Despite the azacarbazoles differing by only one methyl group, the MEGAN model correctly predicted, with high confidence, **A** as an aggregating and **B** as a non-aggregating compound. The model also generates distinct explanation masks for each compound. Examination of the explanation mask for **A** indicates that the methyl group is not important in the model's predicted outcome. Instead, the presence of the methyl group appears to cause the model to discriminate between different parts of the shared heterocyclic core of **A** and **B** in its classifications of each compound. This indicates that for molecules sharing the same scaffold, the model utilizes different substructures as evidence for the aggregation prediction, which demonstrates model learning that extends beyond scaffold memorization.

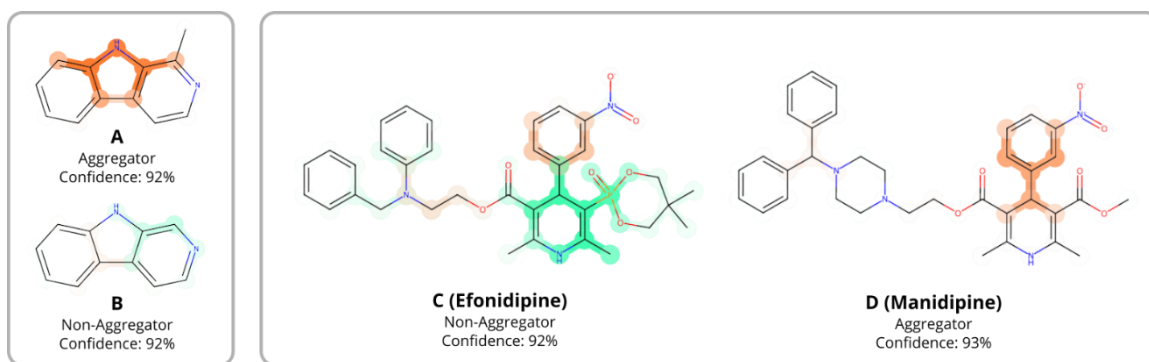


Figure 4.3. MEGAN predictions and explanation masks for two pairs of structurally similar molecules that exhibit experimentally contrasting aggregation behavior.

In comparing efonidipine (**C**) and manidipine (**D**), we again observe the MEGAN model's ability to correctly classify compounds based on changes in substructure. In both **C** and **D**, the model highlights the aromatic ring of the

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

nitrobenzyl group as evidence for aggregation. In **C**, the bulky phosphonate group is identified as providing evidence against aggregation. From a chemical perspective, the explanation masks for **C** seem reasonable, as the phosphonate group likely reduces the molecule's capacity to form closely packed arrangements or stable intermolecular interactions that promote aggregate formation. In contrast to **C**, the methyl ester group of **D** decreases the attributional explanations for nonaggregation resulting from the dihydropyridine. This would suggest that the methyl ester is less disruptive of favorable intermolecular interactions that induce aggregation in **D**. These findings suggest that the model effectively identifies key structural features that either promote or inhibit aggregation, highlighting the potential for targeted modifications to reduce aggregation behavior.

4.2.4 DFT assessment of physical relevance for MEGAN model structural sensitivity

To examine the physical relevance of the MEGAN model's sensitivity to small changes in molecular structure, quantum chemical modeling of a group of structurally similar compounds having different prediction labels was performed. As a significant portion of the molecules in the training dataset contain aromatic heterocycles, pyridine derivatives were selected as the primary focus of this study. This scaffold serves both as a simple example of an aromatic heterocycle and is easily modified to examine the effects of different intermolecular interactions. Although the selected pyridine derivatives have not been experimentally evaluated

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

for aggregation, their analysis provides insights into the relationship between aggregation prediction confidence and interaction energies. Interaction energies for the pyridine derivatives were calculated using density functional theory (DFT) and compared with the MEGAN predictions and explanation masks (Figure 4.4).

The intermolecular interactions of pyridine derivatives predicted with high confidence as aggregators (**E–H**) were calculated to be at least 1.5 kcal/mol more energetically favorable than those of compounds predicted with high confidence as non-aggregators (**I–P**). All 2-hydroxypyridine derivatives were classified as aggregators with high confidence. Conformational searches of the dimers revealed that the strongest interactions between these molecules involved the formation of two intermolecular hydrogen bonds. The only other compound to form two intermolecular hydrogen bonds was 2-aminopyridine (**K**), which was classified as a non-aggregator. However, its interaction energy was only -6 kcal/mol, likely due to the relatively weaker nature of N-H hydrogen bonds compared to O-H hydrogen bonds. Hydroxypyridines **M**, **N**, **O**, and **P** were classified as non-aggregators and were capable of forming only one hydrogen bond, leading to weak interaction energies of greater than -9 kcal/mol. The remaining compounds, **I**, **J**, and **L**, classified as non-aggregators by the MEGAN model, all preferred conformations that promoted π - π interactions and had interaction energies of greater than -8

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

kcal/mol. Overall, this demonstrates the ability of the model to discriminate based on intermolecular interaction strength.

In examining the MEGAN model's explanation masks for the hydroxypyridines, the model is found to distinguish between the various OH substitution patterns. For the predicted aggregators (**O-R**), the MEGAN model finds the HO-C-N substructure, which is responsible for the hydrogen bonding interactions in the dimers, as the key contributor to the aggregator classification (Figure 4.4). In contrast, for the isomers with weaker interaction energies (**J, M, N, O, P**) the model identifies the hydroxyl group as contributing to the classification of these compounds as non-aggregators, demonstrating that the model's prediction is influenced by the substitution pattern. The MEGAN model is also observed to be sensitive to the strength of the hydrogen bond donating abilities of the substituents. While the strongly hydrogen-bonding hydroxyl group of 2-hydroxypyridine (**G**) is highlighted as contributing to the aggregation prediction, the slightly weaker hydrogen-bonding amino group of 2-aminopyridine (**K**) is highlighted as contributing to a non-aggregator prediction. Therefore, not only do the interaction energies correlate with the MEGAN predictions, but the MEGAN model is able to identify the structural features of these molecules, which may be contributing to intermolecular interactions and aggregation.

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

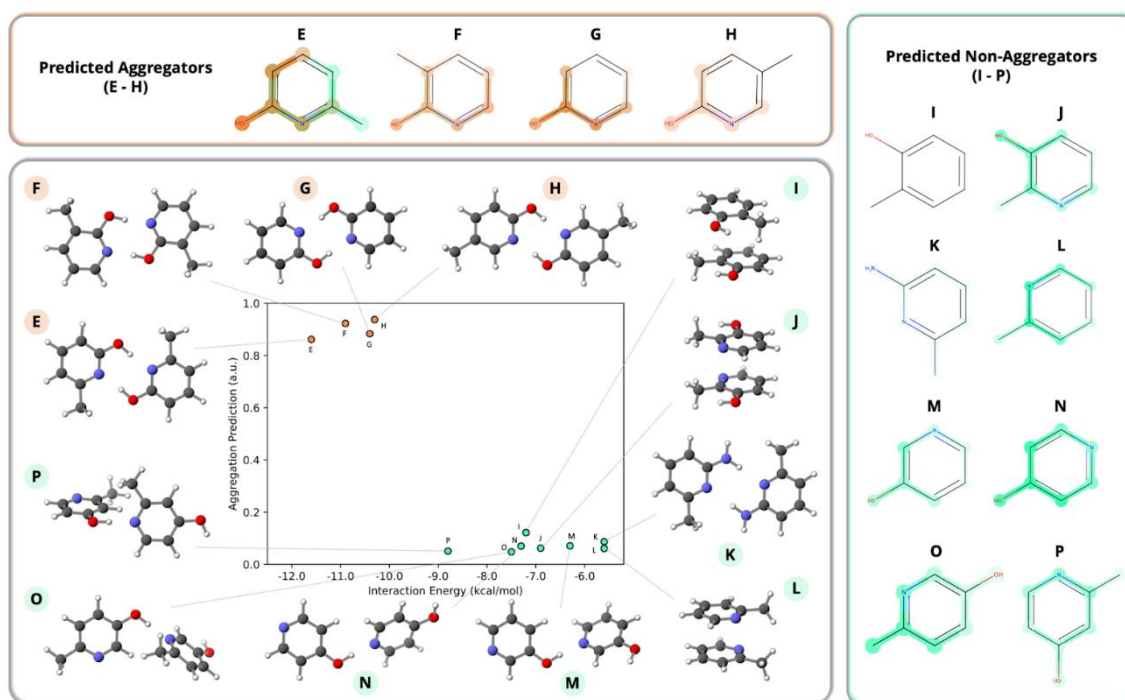


Figure 4.4. MEGAN predictions compared with DFT interaction energies of dimers for compounds E - P. Predictions are reported on a scale from 0 to 1, with 0 indicating a prediction of non-aggregator and 1 indicating a prediction of aggregator. Also shown are the 3D structures of the optimized dimer geometries and MEGAN explanation masks for each molecule. The light-colored explanation masks for I indicate that the MEGAN model detects no pattern systematically associated with either class.

4.2.5 Global explanations

To investigate the molecular substructures commonly contributing to the MEGAN model's prediction of molecules as aggregators or non-aggregators, a global

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

concept extraction was performed.²⁰ In this process, the pooled graph explanations for each channel undergo dimensionality reduction and clustering to provide clusters of molecules that share structural explanations (Figure 4.2c). Each cluster represents one specific molecular substructure which occurs as an important explanation in many individual samples in the training data. The concept extraction produced 159 clusters, with 25 associated with the non-aggregator channel and the remaining 134 clusters associated with the aggregator channel.

Visual analysis of the clustered concepts suggests that flexible molecules and molecules containing groups that have the potential to disrupt π -stacking through spherics often contribute to the non-aggregation prediction. Alternatively, flat, and rigid molecules, as well as molecules with functional groups that can act as both hydrogen bond donors and acceptors, contribute to the prediction of a molecule as an aggregator. Many of the trends that were identified in our clustering analysis are consistent with those that have been identified previously in literature for classifying molecular aggregation potential in small molecules.⁸⁻¹³

One notable trend identified in our cluster analysis, not previously reported in the literature, is the influence of thioureas and ureas on the classification of molecular aggregation. The model frequently associates ureas with aggregation, whereas the thiourea moiety is strongly linked to non-aggregator labels. However, when thioureas are adjacent to an electron-withdrawing group, such as a carbonyl,

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

they instead contribute to the aggregator label. Interestingly, while the model has identified the urea and thiourea substructures as important to the classification of molecules as aggregating or non-aggregating, there is no notable difference in the relative frequency of the urea and thiourea substructures between the aggregating and non-aggregating datasets used for model training. This indicates that the (local) environment of those groups plays a decisive role, and not only the presence or absence of those groups individually. Typical feature importance approaches based, e.g., on molecular fingerprints, would only indicate the presence or absence of those groups as relevant but do not provide further insights into the local environment effects, which are contained in MEGAN explanations.

To analyze whether the explanations generated based on the MEGAN model can be interpreted in terms of structural and physicochemical properties in an automated way, we experimented with large language models, specifically GPT-4o. The goal was to see whether the graph explanations can be automatically related to more broad chemical concepts and human-understandable explanations in order to stimulate further ideas for more detailed analysis by experts and to potentially derive design rules. In most cases, the connection to physicochemical properties did not reveal informative insights. Where non-aggregating motifs were queried, the response of the GPT-4o model explaining the non-aggregation behavior referred to functional groups that enhanced hydrogen bonding with water molecules and steric

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

hindrance of π -stacking, which are potentially correct but also rather obvious explanations. For aggregating motifs, the responses frequently included functional groups that enhanced hydrogen bonding to form molecular networks, as well as planar structures that promote π -stacking. A few explanations provided greater insight that might be validated and quantified in further experiments, e.g., electron-donating effects, which can lead to increased electron density in the π system of the benzene ring, potentially facilitating better solvent interactions over intermolecular π - π stacking interactions. However, additional insights that inspire more detailed analysis or even lead to immediate understanding are still lacking.²¹

4.3.6 Counterfactuals

In addition to MEGAN's attributional explanations, we also employ counterfactual explanations to gain further insights into the model's behavior and decision-making process. For a given original input molecule, we define a counterfactual as a molecule with a minimal structural change from the original molecule which causes the greatest deviation in the model's prediction. Counterfactuals explain which kinds of local perturbations to chemical structure the model, and by extension the underlying aggregation property, is most sensitive to. To probe some of the structural features and modifications influencing the model's classification predictions, counterfactuals were generated for azacarbazoles **A** and **B** (Figure 4.5). Recalling from the previous discussion, the

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

model correctly classifies **A** as aggregating and **B** as non-aggregating despite their subtle structural differences (Figure 4.3).

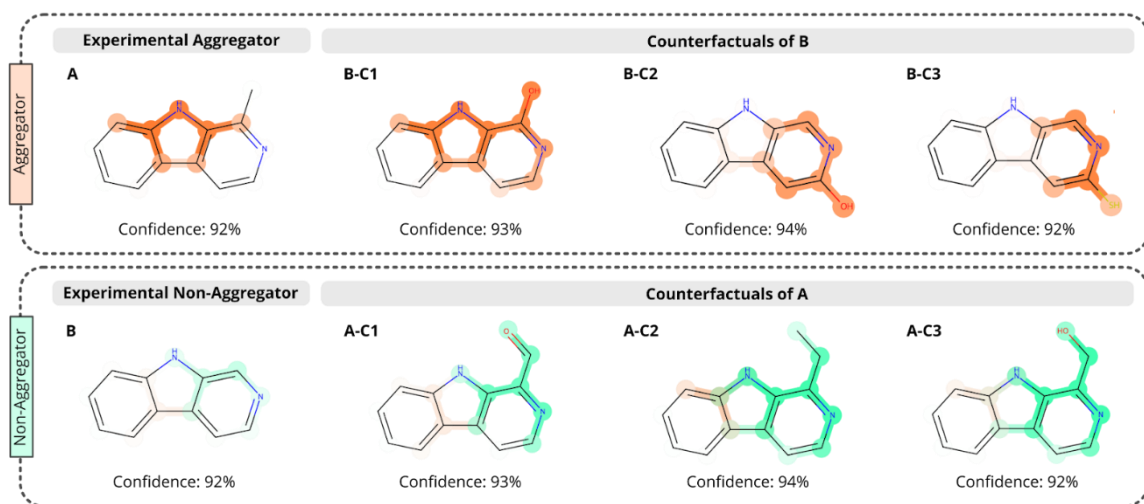


Figure 4.5. Counterfactuals of experimentally validated aggregator **A** and non-aggregator **B** with MEGAN model's prediction confidence and explanation masks.

In general, the counterfactuals generated for aggregator **A** include molecules with electron-withdrawing groups added to the pyridine ring of the azacarbazole as well as molecules with additional steric bulk added to the carbon of the methyl group at the C1 position of the azacarbazole. The addition of an electron-withdrawing group, such as an aldehyde (**A-C1**), to the pyridine ring was found to switch the classification label from aggregator to non-aggregator. Based on this observation, we suggest that electron-withdrawing groups could be capable of disrupting or altering the π -stacking arrangements that may produce aggregation in

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

A by modifying the electronics of the azacarbazole ring system. Additionally, steric bulk added to the methyl group of the azacarbazole (**A-C1**, **A-C2**, **A-C3**) results in larger, more flexible substituents, which sit out of the plane of the ring system, potentially disrupting favorable π -stacking arrangements and preventing aggregation. The non-aggregator channel's explanation masks for the counterfactuals of **A** are nearly identical to those of the non-aggregator **B**, with the only substantial difference being that the additional substituent is also highlighted in the counterfactuals.

The counterfactuals generated for non-aggregator **B** included molecules containing either a hydroxyl (**B-C1**, **B-C2**) or thiol group (**B-C3**) adjacent to the nitrogen atom of the pyridine ring. These analogous substructures, which are ideally suited for forming hydrogen bonding interactions that may contribute to aggregation, are frequently observed in molecules classified by the model as aggregators. These findings are consistent with trends observed in the study of pyridine derivatives (**E-H**) where the HO-C-N substructure was found to be critical for aggregation.

The explanation masks for each of the counterfactuals of **B** (**B-C1**, **B-C2**, **B-C3**) highlight the hydroxy/thiol pyridine substructures in red, indicating that these substructures provide the model with evidence for the aggregation class. The identification of these functional groups is also consistent with the results of other studies, in which thiols and hydroxyl groups have been identified as functional

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

groups commonly inducing aggregation.⁸⁻¹³ Interestingly, compound **A-C3**, which contains a hydroxyl group not directly bound to the pyridine, is predicted as non-aggregating, suggesting that the hydroxyl group alone is not a predictor of aggregation for these compounds and further emphasizing the importance of the hydroxy/thio pyridine substructure.

4.2.7 Experimental validation of predictions and counterfactuals

In addition to serving as a means for identifying the structural features of molecules contributing to their classification as aggregators or non-aggregators by the MEGAN model, counterfactuals also allow for the prediction of new molecules highly similar to an input molecule with alternate aggregation properties. Applying the model in such a manner provides a method for the design of new compounds that maintain the desired structural aspects of a molecule, which may be crucial to forming necessary interactions within a biological target while modifying its aggregation tendencies. The following example provides experimental validation of the MEGAN prediction classification for one of the counterfactuals and demonstrates the utility of counterfactuals for altering the aggregation properties of a compound through minor structural modifications.

We investigated clioquinol, which has been previously reported to aggregate and is correctly predicted by the MEGAN model, and its flipped label counterfactual, methylclioquinol, which has never previously been experimentally

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

examined for its aggregation properties (Figure 4.6). To compare the aggregation behavior of these two molecules, dynamic light scattering (DLS) experiments were performed for each compound for a concentration range that is relevant for HTS assays (see Methods). Consistent with previous reports, clioquinol was found to aggregate, as demonstrated by a significant increase in its hydrodynamic radius at higher concentrations. In contrast, methylclioquinol exhibited minimal fluctuation in its hydrodynamic radius, indicating that aggregation does not occur in the concentration range studied. These findings highlight the potential power of MEGAN model's ability to aid in the rational design of molecules with tailored aggregation properties, offering a versatile approach for optimizing compound behavior in biological systems while preserving critical structural features.

To understand the model's rationale for suggesting methylclioquinol as a counterfactual, the explanation masks for both molecules were examined. The explanation masks for clioquinol highlight the majority of the quinoline scaffold, with a focus on the area containing the nitrogen and alcohol, as providing evidence for the aggregator class. In contrast, for methylclioquinol the methoxy group as well as the quinoline scaffold are both highlighted, albeit to a lesser extent, as providing evidence for the non-aggregator class. The methylation of clioquinol at the oxygen atom eliminates its ability to engage in hydrogen bonding interactions. These

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

findings suggest that the model recognizes hydrogen bonding motifs and the steric bulk introduced by substituents as key factors influencing aggregation.

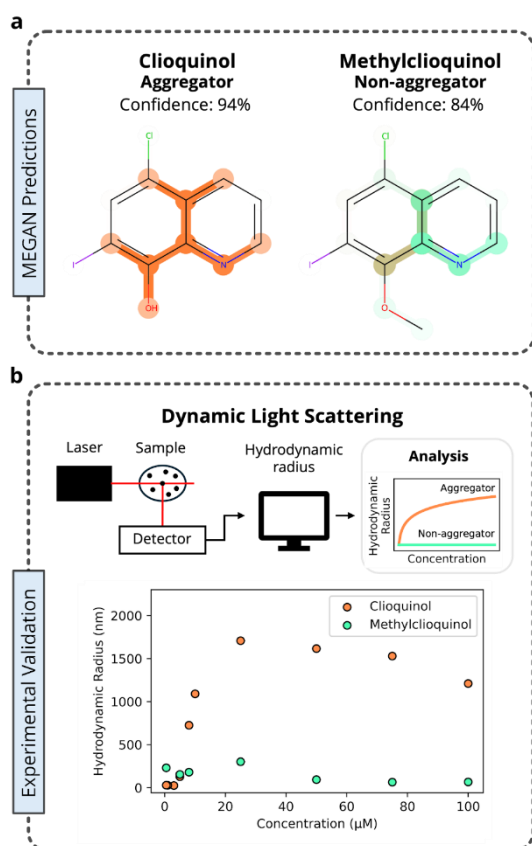


Figure 4.6. (a) MEGAN model's prediction confidence and explanation masks for clioquinol and methylclioquinol. (b) Overview of DLS method and experimental DLS results for clioquinol (orange) and methylclioquinol (green).

4.2.8 External validation

To further validate the predictive accuracy of the MEGAN model, a dataset of 58 structurally diverse drugs that have been experimentally characterized as aggregators or non-aggregators was compiled and assessed using the model. This external validation dataset was sourced from two literature references, where the

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

compounds had been experimentally screened to evaluate the impact of colloidal aggregation on SARS-CoV-2 drug repurposing efforts.^{22, 23} The dataset includes 30 aggregating and 28 non-aggregating compounds, none of which were present in the training or test sets for the MEGAN model, ensuring that they were previously unseen by the model.

The MEGAN model was capable of correctly classifying 15 of the 30 aggregating compounds and 24 of the 28 non-aggregating compounds, resulting in an accuracy of 68%. For comparison, we evaluated the performance of CHEMAgg, a widely recognized model for aggregation prediction, using the same validation dataset. CHEMAgg achieved an accuracy of only 50%, significantly lower than the MEGAN model. While CHEMAgg accurately identified 27 of the 28 non-aggregating compounds, it only correctly classified 2 of the 30 aggregating compounds. CHEMAgg's classification of 55 of the 58 molecules as non-aggregators suggests a strong bias in this model towards non-aggregator classification. In contrast, the MEGAN model provides a more balanced classification and demonstrates superior performance in accurately predicting molecular aggregation. This result highlights the ability of our model to outperform existing state-of-the-art models, particularly in identifying aggregating compounds, which is a

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

critical capability for accurately screening datasets for promiscuous molecules.

4.2.9 Conclusion

The MEGAN model presented in this study demonstrates a high level of accuracy for predicting SCAMs, achieving 82% on a balanced in-distribution test dataset and similar performance on a smaller external dataset of drug molecules. While commonly accepted features such as logP values and the number of aromatic rings are relevant for detecting SCAMs, the model's high accuracy and explanations indicate that it learns more complex relationships between molecular structure and aggregating behavior.

Through local and global model explanations and counterfactual analysis, we identified and systematically examined specific molecular motifs associated with aggregation. By combining expert knowledge, quantum mechanical modeling, and automated interpretations using large language models, we provide deeper insights into molecular aggregation. The MEGAN model not only highlights molecular motifs strongly linked to aggregation but also identifies small structural modifications that can significantly alter aggregating behavior, providing valuable tools for molecular design.

The MEGAN model's accuracy, combined with the accessibility and extended functionality provided through a publicly available web interface, enables improved

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

detection of SCAMs. Furthermore, the use of molecular counterfactuals offers a practical approach for designing alternatives to problematic compounds. This capability facilitates the filtering of HTS libraries, helping to reduce false positives in drug discovery databases and improving overall screening reliability.

4.3 Methods

4.3.1 Dataset curation and preparation

For model training, a dataset of molecules that have been experimentally validated as either aggregating or non-aggregating was curated. In order to do so, data from the Shoichet laboratory utilized for the development of their Aggregator Advisor tool was used.⁹ For the development of Aggregator Advisor, Stoichet et al. identified over 12,600 likely aggregators from their previously published experimental screens for colloidal aggregation. These compounds were used as our dataset of experimentally validated aggregators. To curate our dataset of non-aggregating molecules, one of the libraries screened for aggregation by Stoichet consisting of approximately 198,000 compounds was filtered to exclude any molecules in our dataset of aggregating molecules.²⁴ As the 198,000 compound dataset was utilized to populate the 12,600 aggregators in Aggregator Advisor, any compound in the dataset that was not identified by Stoichet as an aggregator has been experimentally validated as a non-aggregator.

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

Upon data curation, the resulting datasets consisted of 12,607 aggregating and 197,846 non-aggregating molecules represented as SMILES strings. In order to clean the data for use in training our machine learning model, the SMILES strings were canonicalized using Rdkit.²⁵ If multiple fragments were present in the SMILES strings, the longest one was kept, and the others (generally counterions or solvents) were removed. If there was more than one long (>12 characters) fragment, the data point was discarded. All duplicate molecules and molecules producing Rdkit errors were also discarded from the datasets. Additionally, for the non-aggregating dataset, molecules containing metal atoms and inorganic molecules (i.e. containing zero carbon atoms) were discarded. After filtering the molecules, the remaining datasets consisted of 12,338 aggregating and 177,048 non-aggregating molecules. Molecular descriptors were analyzed to examine the differences in structural features and properties of the compounds in the aggregating and non-aggregating datasets.

For the training of the graph neural network, the dataset was first pre-processed using the publicly available protonation software library DimorphiteDL to extract all possible protonation states at physiological pH for each molecule.²⁶ A value of 7.4 was used for both the min_ph and max_ph parameters of the DimorphiteDL software. The protonated molecule variants were then converted into graph representations using Rdkit to extract numeric node and edge attribute

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

vectors.²⁵ Node features include a one-hot representation of the atom type, as well as additional information such as the number of connected hydrogen and local charge distribution. Similarly, the edge features contain a one-hot encoding of the bond type and additional information such as whether the bond is part of a ring structure.

4.3.2 Explainable graph neural networks, MEGAN

The Multi-Explanation Graph Attention Network (MEGAN) is a graph neural network designed to make predictions while also explaining how it arrived at those predictions. Unlike many other models, MEGAN can provide multiple explanations for each prediction, such as highlighting features that contribute positively or negatively. It uses attention mechanisms to focus on important parts of the graph and a self-consistent, masking-based training process to ensure that its explanations align with the expected meanings. This makes MEGAN especially useful for tasks like predicting molecular properties, where understanding the role of specific substructures is critical. It can also be trained to match known explanations, improving both its accuracy and interpretability. An interactive interface to query the MEGAN model for aggregation prediction is available at https://megan.aimat.science/predict/megan_aggregator. Code and data are available on GitHub (See Data and Code Availability).

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

4.3.3 Counterfactuals

Counterfactuals are generated by exploring the immediate graph neighborhood of each original molecule and then selecting those that result in the highest prediction difference relative to the original prediction. We generate the local neighborhood of the molecular graphs with a procedure inspired by Riley et al. in which all chemically feasible atom and bond insertions and deletions are applied to the molecule recursively.²⁷ Depending on the recursion depth, this procedure typically produces 100-1000 perturbed graphs out of which we present the 10 graphs with the highest prediction difference as counterfactual explanations.

4.3.4 Fingerprint and descriptor-based ensemble methods

To compare the accuracy of our model to the aggregation prediction model reported by Yang et al., we re-implemented their computational approach and trained it on the dataset presented in this work. We used three types of descriptors, namely fingerprint features (circular fingerprints as well as ECFP4/Morgan fingerprints as implemented in Rdkit), MACCS features as implemented in Rdkit, and CATS features as implemented in <https://github.com/alexarnimueller/cats-descriptor>.²⁸⁻³⁰ Within each set of descriptors, we used a feature importance analysis based on a Random Forest classifier to determine how many features are required to achieve the highest accuracy on the validation set. This resulted in a total selection of 33 features. After concatenation and hyperparameter optimization, we found the optimal hyperparameters of the Random Forest

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

classifier to be an ensemble size of 500, with no limit on depth, and entropy as an impurity measure; of the Gradient Boosting classifier to be an ensemble size of 500, a maximum depth of 9, and a learning rate of 0.5; and of the XGBoost model to be an ensemble size of 500, a maximum depth of 8, and a learning rate of 0.1.

4.3.5 DFT modeling of interaction energies

DFT modeling of pyridine derivatives **E - P** was performed to assess the energetic favorability of the intermolecular interactions formed between compounds as well as the favored geometries of the dimers. To locate the lowest energy geometries of the monomers and dimers, initial geometries were constructed, and conformer ensembles were generated using CREST based on GFN2-xTB calculations.^{31,32} The lowest energy monomer and dimer geometries were then optimized using wB97xD/def2-TZVP with the SMD solvent model (H₂O) in Gaussian 16.³³⁻³⁶ Vibrational analysis performed at 273.15 K confirmed that all optimized geometries corresponded to minima on the potential energy surfaces, as indicated by the absence of imaginary frequencies. The interaction energies were calculated following Eq. 4.1 using the sum of the electronic energy (E) and zero-point correction to the electronic energy (E_{ZPE}). In all cases, the most energetically favorable geometry of the dimer that could be located was used for the calculation of the interaction energy.

$$\text{Interaction Energy} = (E_{\text{dimer}} + E_{\text{ZPE, dimer}}) - 2(E_{\text{monomer}} + E_{\text{ZPE, monomer}}) \quad (\text{Eq. 4.1})$$

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

4.3.6 Dynamic light scattering experiments

DLS was used as an aggregation detection method for the examination of clioquinol and methylclioquinol. The instrument used in this study was a Nanotemper Prometheus Panta and standard 10 μ L capillaries were used. Clioquinol was purchased from TCI chemicals via Fisher Scientific, and methylclioquinol was prepared and purified following standard methylation conditions using iodomethane.³⁷ Stock solutions of 1 mM were prepared for each compound in DMSO. The stock solutions underwent serial dilution using a 40 mM sodium phosphate buffer at pH 7.4 to prepare solutions of 100, 75, 50, 25, 10, 8, 5, 5, 3, 1, and 0.5 μ M. The concentrations tested were chosen based on the concentrations that standard HTS screens are run at. Each solution was run in triplicate on the instrument, and aggregation was determined based on hydrodynamic radius.

4.4 References

1. A. S. M. Z. Rahman, C. Liu, H. Sturm, A. M. Hogan, R. Davis, P. Hu, S. T. Cardona, "A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery" *PLoS Comput. Biol.* **2022**, 18, e1010613
2. F. Wong, E. J. Zheng, J. A. Valeri, N. M. Donghia, M. N. Anahtar, S. Omori, A. Li, A. Cubillos-Ruiz, A. Krishnan, W. Jin, A. L. Manson, J. Friedrichs, R. Helbig, B. Hajian, D. K. Fiejtek, F. F. Wagner, H. H. Soutter, A. M. Earl, J. M. Stokes, L. D.

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

- Renner, J. J. Collins, “Discovery of a structural class of antibiotics with explainable deep learning” *Nature*. **2024**, 626, 177-185
3. M. Klumpp, “Non-stoichiometric inhibition in integrated lead finding – a literature review” *Expert Opin. Drug Discov.* **2016**, 11, 149-162
 4. N. Thorne, D. S. Auld, J. Inglese, “Apparent activity in high-throughput screening: origins of compound-dependent assay interference” *curr. Opin. Chem. Biol.* **2010**, 14, 315-324
 5. Z. Y. Yang, J. H. He, A. P. Lu, T. J. Hou, D. S. Cao, “Frequent hitters: nuisance artifacts in high-throughput screening” *Drug Discov. Today*. **2020**, 25, 657-667
 6. K. E. D. Coan, B. K. Shoichet, “Stoichiometry and physical chemistry of promiscuous aggregate-based inhibitors” *J. Am. Chem. Soc.* **2008**, 130 9606-9612
 7. J. Mei, N. L. C. Leung, R. T. K. Kwok, J. W. Y. Lam, B. Z. Tang, “Aggregation-induced emission: together we shine, united we soar!” *Chem. Rev.* **2015**, 115, 11718-11940
 8. K. Lee, A. Yang, Y. C. Lin, D. Reker, G. J. L. Bernardes, T. Rodrigues, “Combating small-molecule aggregation with machine learning” *Cell Rep. Phys. Sci.* **2021**, 1005573

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

9. J. J. Irwin, D. Duan, H. Torosyan, A. K. Doak, K. T. Ziebart, T. Sterling, G. Tumanian, B. K. Shoichet, "An Aggregation Advisor for Ligand Discovery" *J. Med. Chem.* **2015**, 58, 7076-7087
10. V. M. Alves, S. J. Capuzzi, R. C. Braga, D. Korn, J. E. Hochuli, K. H. Bowler, A. Yasgar, G. Rai, A. Simeonov, E. N. Muratov, A. V. Zakharov, A. Tropsha, "SCAM Detective: Accurate Prediction of Small, Colloidally Aggregating Molecules" *J. Chem. Inf. Model.* **2020**, 60, 4056-4063
11. Z. Y. Yang, Z. J. Yang, J. Dong, L. L. Wang, L. X. Zhang, J. J. Ding, X. Q. Ding, A. P. Lu, T. J. Hou, D. S. Cao, "Structural Analysis and Identification of Colloidal Aggregators in Drug Discovery" *J. Chem. Inf. Model.* **2019**, 59, 3714-3726
12. C. Molina, L. Ait-Ouarab, H. Minoux, "Isometric Stratified Ensembles: A Partial and Incremental Adaptive Applicability Domain and Consensus-Based Classification Strategy for Highly Imbalanced Data Sets with Application to Colloidal Aggregation" *J. Chem. Inf. Model.* **2022**, 62, 1849-1856
13. A. A. Hajal, R. A. Bryce, B. B. Amor, N. Atatreh, M. A. Ghattas, "Boosting the Accuracy and Chemical Space Coverage of the Detection of Small Colloidal Aggregating Molecules Using the BAD Molecule Filter" *J. Chem. Inf. Model.* **2024**, 64, 4991-5005

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

14. A. J. Ryan, N. M. Gray, P. N. Lowe, C. W. Chung, “Effect of detergent on “promiscuous” inhibitors” *J. Med. Chem.* **2003**, 46, 3448-3451
15. B. Y. Feng, B. K. Shoichet, “A detergent-based assay for the detection of promiscuous inhibitors” *Nat. Protoc.* **2006**, 1, 550-553
16. A. N. Ganesh, E. N. Domders, B. K. Shoichet, M. S. Shoichet, “Colloidal aggregation: from screening nuisance to formulation nuance” *Nano Today* **2018**, 19, 188-200
17. A. K. Doak, H. Wille, S. B. Prusiner, B. K. Shoichet, “Colloid formation by drugs in simulated intestinal fluid” *J. Med. Chem.* **2010**, 53, 4259-4265
18. D. Reker, G. J. L. Bernardes, T. Rodrigues, “Computational advances in combating colloidal aggregation in drug discovery” *Nat. Chem.* **2019**, 11, 402-418
19. J. Teufel, L. Torresi, P. Reiser, P. Friederich, arXiv preprint 2023, DOI: arXiv:2211.13236
20. J. Teufel, P. Friederich, arXiv preprint 2024, DOI: arXiv:2404.16532
21. M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. D. P. Gomes, F. Hase, A. Jinich, A. Nigam, Z. Yao, A. Aspuru-Guzik, On scientific understanding with artificial intelligence. *Nature Reviews Physics* **2022**, 4, 761-796

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

22. H. R. O'Donnell, T. A. Tummino, C. Bardine, C. S. Craik, B. K. Shoichet, "Colloidal aggregators in biochemical SARS-CoV-2 repurposing Screens" *J. Med. Chem.* **2021**, 64, 17530-17539
23. I. S. Glenn, L. N. Hall, M. M. Khalid, M. Ott, B. K. Shoichet, "Colloidal Aggregation Confounds Cell-Based Covid-19 Antiviral Screens" *J. Med. Chem.* **2024**, 67, 10263-10274
24. A. Jadhav, R. S. Ferrerira, C. Klumpp, B. T. Mott, C. P. Austin, J. Inglese, C. J. Thomas, D. J. Maloney, B. K. Shoichet, A. Simeonov, "Quantitative Analyses of Aggregation, Autofluorescence and Reactivity Artifacts in a Screen for Inhibitors of Thiol Protease" *J. Med.Chem.* **2010**, 53, 37-51
25. RDKit: Open-Source Cheminformatics. <https://www.Rdkit.Org>.
26. P. J. Ropp, J. C. Kaminsky, S. Yablonski, J. D. Durrant, "Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules" *J. Cheminformatics* **2019**, 11, 14
27. Z. Zhou, S. Kearnes, L. Li, R. N. Zare, P. Riley, "Optimization of molecules via deep reinforcement learning" *Sci. Rep.* **2019**, 9, 10752
28. D. Rogers, M. Hahn, "Extended-Connectivity Fingerprints" *J. Chem. Inf. Model.* **2010**, 50
29. J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, "Reoptimization of MDL Keys for use in drug discovery" *J. Chem. Inf. Model.* **2002**, 42, 1273-1280

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

30. M. Reutlinger, C. P. Koch, D. Reker, N. Todoroff, P. Schneider, T. Rodrigues, G. Schneider, “Chemically Advanced Template Search (CATS) for scaffold-hopping and prospective target prediction for ‘orphan’ molecules” *Molecular informatics* **2013**, 32, 133-138
31. P. Pracht, F. Bohle, S. Grimme, “Automated exploration of the low-energy chemical space with fast quantum chemical methods” *Phys. Chem. Chem. Phys.* **2020**, 22, 7169-7192
32. C. Bannwarth, S. Ehlert, S. Grimme, “GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions” *J. Chem. Theory Comput.* **2019**, 15, 1652-1671
33. J. D. Chai, M. Head-Gordon, “Long-Range Corrected Hybrid Density Functionals with Damped Atom-Atom Dispersion Corrections” *Phys. Chem. Chem. Phys.* **2008**, 10, 6615-6620
34. J. Laun, D. V. Oliveira, T. Bredow, “Consistent gaussian basis sets of double- and triple-zeta valence with polarization quality of the fifth period for solid-state calculations” *J. Comput. Chem.* **2018**, 39, 1285-1290
35. A. V. Marenich, C. J. Cramer, D. G. Truhlar, “Universal solvation model based on solute electron density and on a continuum model of the solvent defined

Chapter 4: Mitigating Molecular Aggregation in Drug Discovery with Predictive Insights from Explainable AI

by the bulk dielectric constant and atomic surface tensions” *J. Phys. Chem.*

B. **2009**, 113, 6378-6396

36. Gaussian 16, Revision C.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Malatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gompers, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zeng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D. J. Fox, Gaussian, Inc., Wallingford CT, **2016**.
37. A. R. Joaquim, P. Reginatto, M. S. Lopes, L. C. G. Bazana, M. P. Gionbelli, M. A. D. Cesare, T. F. A. Kaminski, M. L. Teixeira, M. A. Abegg, A. M. Fuentefria, S. F. D. Andrade, “New 8-hydroxyquinoline derivatives highlight the potential of this class for treatment of fungal infections” *New J. Chem.* **2021**, 38

Chapter 5: Conclusions and Future Work

5.1 Conclusions

5.1.1 Chapter 2

This work in chapter 2 explored in silico chemical libraries and their role in training ML algorithms. A natural product library was analyzed, revealing promising results for tetrahydro- β -carboline but highlighting issues with PAINS compounds, particularly Michael acceptors. This led to two research directions: investigating the activity of ZINC8876405, which is an ongoing endeavor, and refining library curation by filtering PAINS compounds and evaluating library size and weight.

Also presented in chapter 2 was the training of nine small molecule libraries. Libraries 4 and 5 showed the best results, despite Library 5 being smaller it was more balanced, suggesting that training data requirements vary by target. Additionally, for growth inhibition prediction, dataset balance proved more critical than size, as seen in the comparative analysis of Libraries 3, 7, and 8. Despite its small size, Library 7 yielded reasonable results, surpassing existing ML benchmarks in antibiotic research.

5.1.2 Chapter 3

The work presented in this chapter shows that XAI models can give chemical insight into molecules based on the limited data given to the model in training. This chapter shows that MEGAN can be trained using a library of β -lactamase inhibitors to achieve performance that is better than a RF model as well as better than what is currently published in literature sources. MEGAN was also able to correctly explain

Chapter 5: Conclusions and Future Work

the activity of β -lactam containing β -lactamase inhibitors by highlighting the lactam ring as well as other atoms that seem to participate in the bonding of the compound in the protein active site. This was very promising as MEGAN had only ever ‘seen’ SMILES strings, so identifying interactions it has no previous knowledge of is important.

5.1.3 Chapter 4

As per our prepared manuscript, “The MEGAN model presented in this work was able to accurately predict SCAMS with an accuracy of 82.5% on a balanced in-distribution test data set and similar accuracies on smaller external data sets. While traditional features, e.g. logP values or number of aromatic rings are relevant to detect SCAMS, the high accuracy of the MEGAN model can only be explained if we assume that the model learns more complex relations between molecular structure and aggregating behavior. Using intrinsic model explanations as well as counterfactuals, we identified and further analyzed selected motifs that were found through systematic analysis of the model. Using expert knowledge-driven interpretations, further simulations, as well as automatic interpretations with large language models, we show that additional insight into molecular aggregating can be obtained.

We find specific molecular motifs to be strongly related to aggregation while at the same time, the MEGAN model helps to reveal small modifications that strongly change aggregating behavior which can be used in molecular design approaches.

Chapter 5: Conclusions and Future Work

We furthermore find that the predictions by the MEGAN model are strongly correlated with molecular binding energies, indicating that dimerization can play a relevant role in the formation of larger molecular aggregates. First tests of combining large language models with model predictions and counterfactuals show that interesting additional insight can be obtained which potentially accelerates or inspires further manual analysis by experts.

The accuracy of the MEGAN model coupled with the accessibility and extra functionality provided through a publicly available web interface allows for reliable detection of SCAMs. Furthermore, molecular counterfactuals, where minimal structural changes reverse the aggregation prediction, can be utilized in the design of alternatives for compounds that are likely to aggregate and will still maintain the main structure of the scaffold. This overall allows for the filtering of HTS libraries to combat false positives in the actual screens.”

5.2 Future work

Chapter 2 was work completed in collaboration with the Cardona lab at the U of M and the Hu lab at Western. This ongoing collaboration is still looking to train, test, and validate ML methods to aid in identifying compounds with growth-inhibitory activity. While the main goal remains the same the methods utilized are evolving with the field and new types of ML are to be investigated in the future for this task.

Chapter 5: Conclusions and Future Work

5.2.2 Chapter 3

The work outlined in Chapter 3 is far from complete. Our trained model targeting only one MOA is the first of its kind and initially looks as if it will outperform models trained on whole cell assay data. This is yet to be seen but can be validated in the lab by using the model to make predictions and experimentally testing the top compounds in the lab. A second direction this project can go is further examinations of clusters to identify key features of molecules present to curate a new scaffold of molecules that have the potential to have inhibitory activity. The work outlined in Chapter 3 should also be expanded to include different ML approaches as there is potential for improvement being that the f1 score was only 0.69. Finally retraining the algorithm with a different subset of molecules from the initial data set could be interesting to see if we can improve model performance that way.

5.2.3 Chapter 4

The work outlined in Chapter 4 has some interesting avenues that can be taken. One way our work could be expanded is to curate a library of AIE molecules that can be included in the training data so that our model can encompass medicinal and material chemical space. A second area that can be improved upon is the counterfactuals. While there is currently no other aggregation model that predicts counterfactuals, including a model that predicts a synthetic route to the counterfactuals could be interesting. In the early stage of this project, we got a lot of counterfactuals that were synthetically challenging or impossible to make so

Chapter 5: Conclusions and Future Work

allowing users to get a counterfactual and a synthetic route could have utility.^{5.3}

Publication record

Currently, I have 2 published works:

1. Liu, C., Hogan, A.M., **Sturm, H.** *et al.* Deep learning-driven prediction of drug mechanism of action from large-scale chemical-genetic interaction profiles. *J Cheminform* **14**, 12 (2022). <https://doi.org/10.1186/s13321-022-00596-6>
2. A. S. M. Z. Rahman, C. Liu, **H. Sturm**, A. M. Hogan, R. Davis, P. Hu, S. T. Cardona, “A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery” *PLoS Comput. Biol.* **2022**, 18, e1010613

Also, chapter four has been prepared as a manuscript and submitted to *Angewandte chemie*. Work completed in chapter 2 is currently in the process of being written up with collaborators in the Cardona Lab. Finally chapter 3 is being prepared into an invited manuscript.

5.3 Final Thoughts

This thesis has conveyed the work I have completed over the past four years. While there is no new approved drug from my work, I believe through my PhD work I have contributed to the drug discovery process, maybe not by finding a new drug, but by the development of predictive models that others can use to aid in their

Chapter 5: Conclusions and Future Work

scientific endeavors. Because after all science is collaborative and we must look to the past to make progress for the future.