In Silico Discovery and Optimization of MMP-9 Inhibitors

A Machine Learning–Driven Drug Discovery Pipeline for Neurodegenerative Disease Modulation

Natania Varghese
Grade 11
Queen Elizabeth High School

The Challenge of Neurodegenerative Drug Discovery

MMP-9 & Parkinson's

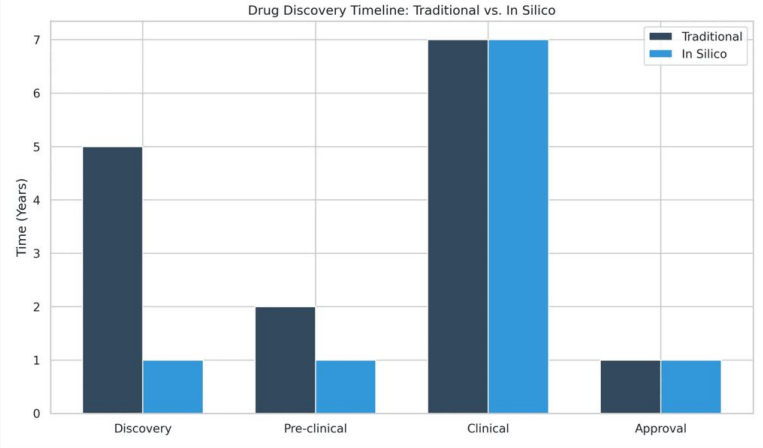
Matrix Metalloproteinase-9 (MMP-9) is a zinc-dependent enzyme implicated in **neuroinflammation** and **blood-brain barrier disruption**.

Traditional Barriers

Conventional drug discovery is prohibitively expensive (**~$2B per drug**) and slow (**10–15 years**).

The Solution

There is a critical need for faster, computational approaches to identify therapeutic agents efficiently.



Comparison of Traditional vs. In Silico Discovery Timelines

Research Objective: Accelerating MMP-9 Inhibitor Discovery

Develop a computational pipeline to **predict MMP-9 inhibitory activity**, identify drug-like candidates, and optimize lead compounds for neurodegenerative disease modulation.

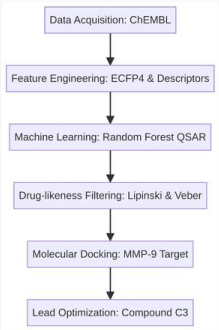
Question 1

Can machine learning effectively model structure–activity relationships for MMP-9 inhibitors?

Question 2

Can we computationally identify high-confidence inhibitors with therapeutic potential?

Overall Pipeline Flow Diagram



Comprehensive Pipeline for In Silico Drug Discovery

PHASE I

Data Acquisition

Sourcing high-quality bioactivity data from the ChEMBL database.

PHASE II

Feature Engineering

Extracting molecular descriptors and ECFP4 fingerprints.

PHASE III

Machine Learning

Developing QSAR models using Random Forest algorithms.

PHASE IV

Drug-likeness

Filtering candidates based on Lipinski and Veber criteria.

PHASE V

Molecular Docking

Simulating binding interactions with the MMP-9 target site.

PHASE VI

Interaction Analysis

Evaluating binding affinity and molecular interactions.

PHASE VII

Lead Optimization

Refining chemical structures for enhanced therapeutic potential.

GOAL

Candidate Selection

Identifying high-confidence inhibitors for further validation.

"A systematic multi-stage approach ensures the identification of potent, drug-like, and selective MMP-9 inhibitors."

Dataset Curation and Preprocessing

Initial Acquisition

1,067

Raw compounds from ChEMBL

- Primary source: ChEMBL database
- Target: MMP-9 inhibitors

Data Cleaning

- Removed missing SMILES
- Removed missing pIC50 values
- Duplicate detection: None
- Standardized chemical structures

Final Dataset

950

Cleaned compounds retained

Key Data Fields:

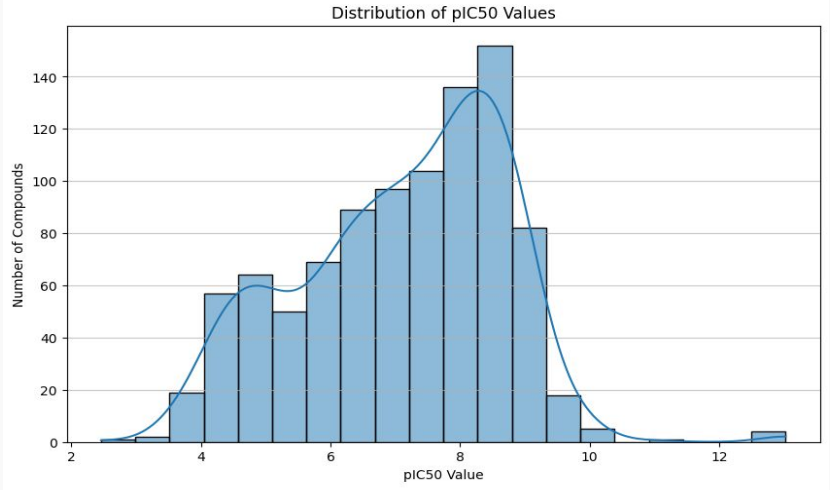
CHEMBL_ID

SMILES (Structure)

pIC50 (Activity)

Biological Activity and Physicochemical Profile

pIC50 Distribution



Activity Classification: Compounds classified as **Active** if pIC50 > 7 and **Inactive** if ≤ 7. The dataset shows a distribution centered between 7.5–8.5.

Physicochemical Properties

Molecular Weight (MW)	400 – 500 Da
Lipophilicity (LogP)	2.0 – 4.0
H-Bond Donors (HBD)	2 – 3
H-Bond Acceptors (HBA)	5 – 7

Lipinski's Rule of Five: Most compounds in the dataset satisfy these criteria, indicating high potential for oral bioavailability and drug-likeness.

Feature Engineering and Model Development

Feature Engineering

- **ECFP4 Fingerprints**
  2048-bit vectors capturing local substructures and molecular topology.
- **Physicochemical Descriptors**
  7 key properties including MW, LogP, HBD, and HBA.
- **Feature Matrix**
  Final representation of 950 compounds × 2055 features.

Model Development

- **Random Forest Classifier**
  Ensemble method chosen for its ability to handle nonlinear relationships.
- **Robustness**
  High resistance to overfitting and effective handling of high-dimensional data.
- **Training Strategy**
  Stratified 5-fold cross-validation to ensure reliable performance evaluation.

> *"Combining structural fingerprints with global descriptors provides a holistic view of molecular activity."*

Robust Model Performance and Validation

ROC-AUC
0.92

ACCURACY
0.86

PRECISION
0.87

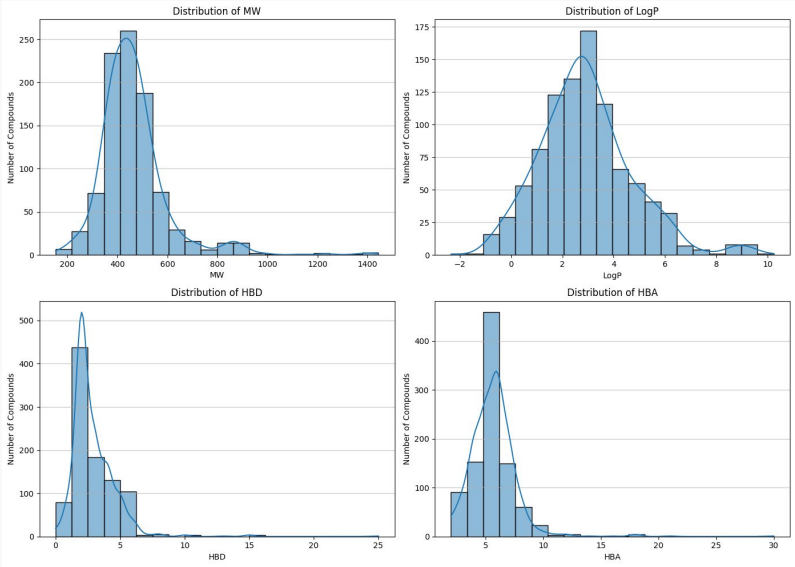
RECALL
0.89

Y-Scrambling Validation

Confirmed the model learns real structure–activity relationships.

True Labels: ROC-AUC ≈ 0.92

Scrambled Labels: ROC-AUC ≈ 0.50



Receiver Operating Characteristic (ROC) Curve

Candidate Filtering and Lead Compound Identification

Drug-likeness Screening

Applied Lipinski and Veber criteria to ensure oral bioavailability.

Result: **950 → 539 compounds**

High-Confidence Selection

Filtered by prediction probability ≥ 0.9 and applicability domain.

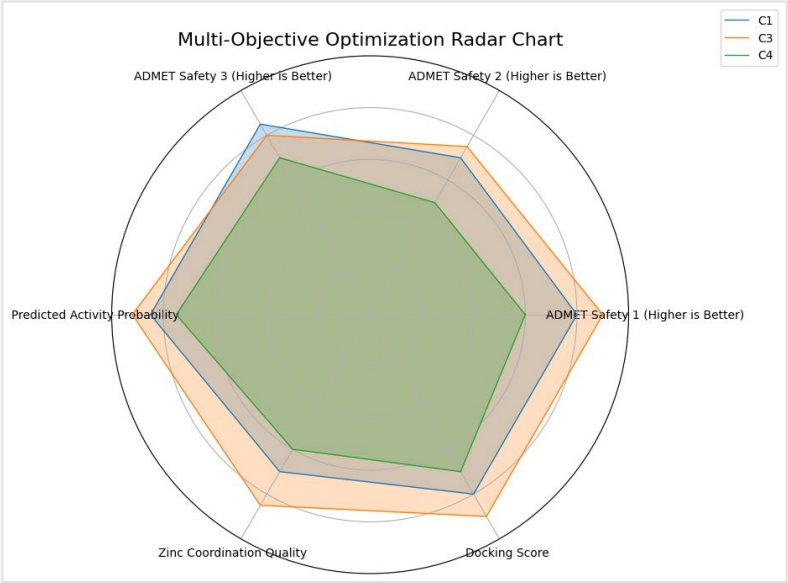
Result: **539 → 168 candidates**

Lead Identification

Evaluated ligand efficiency (LE) and lipophilic efficiency (LLE).

Result: **Compound C3 identified as Lead**

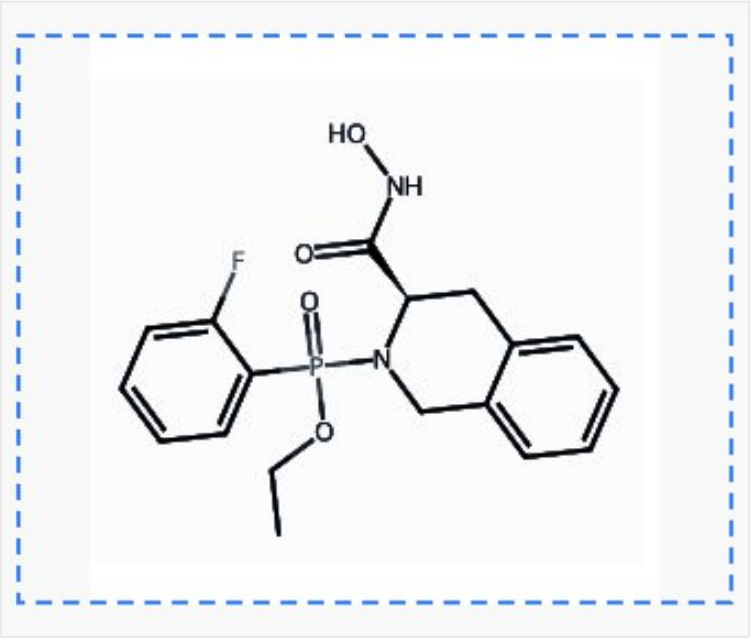
Lead Compound C3 Profile



pIC50	Half-life	Clearance
8.80	**44.6 h**	**0.183 mL/min/kg**

FINAL CANDIDATE SELECTION

Compound C3: Lead Inhibitor Profile



Predicted Activity
pIC50: 8.80

Target
MMP-9 (Matrix Metalloproteinase-9)

Half-life
~44.6 hours

Clearance
Low (favorable retention)

Drug-likeness
Passes Lipinski & Veber criteria

Chemical Class
Peptidomimetic Sulfonamide

Compound C3 emerged as the top candidate after multi-stage filtering, combining high predicted inhibitory activity with strong pharmacokinetic properties. Its extended half-life and low clearance suggest sustained therapeutic presence, which is especially critical for neurodegenerative conditions like Parkinson's disease. Structurally, its diphenylmethyl-piperidine framework is associated with biologically active compounds, reinforcing its potential as a viable drug candidate.

Identified through machine learning-driven in silico pipeline

Significance, Limitations, and Future Directions

Study Significance

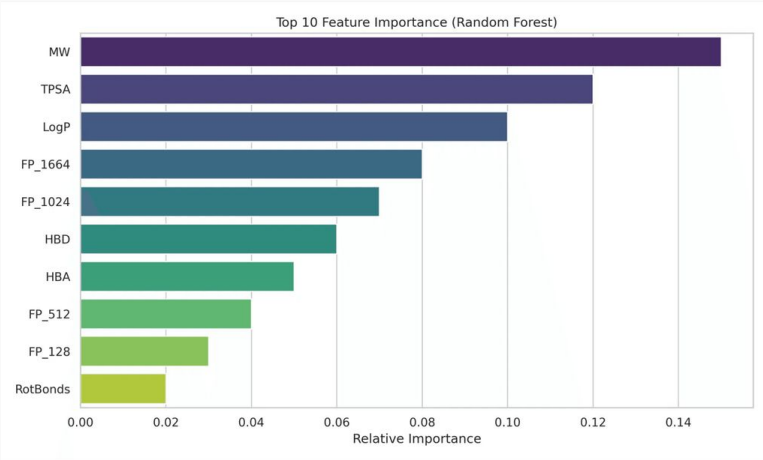
Demonstrates the power of **AI-driven drug discovery** to accelerate the identification of MMP-9 inhibitors, significantly reducing the time and cost associated with traditional screening methods.

Limitations

The study is **fully computational**. While the pipeline is robust, the identified candidates require experimental validation to confirm their biological activity and safety profiles.

Future Work

- **In Vitro Testing:** Synthesis and assaying of Compound C3.
- **In Vivo Validation:** Testing in Parkinson's disease models.
- **Clinical Translation:** Investigating therapeutic potential.



Top 10 Feature Importance (Random Forest Model)

Conclusion: A New Frontier in Drug Discovery

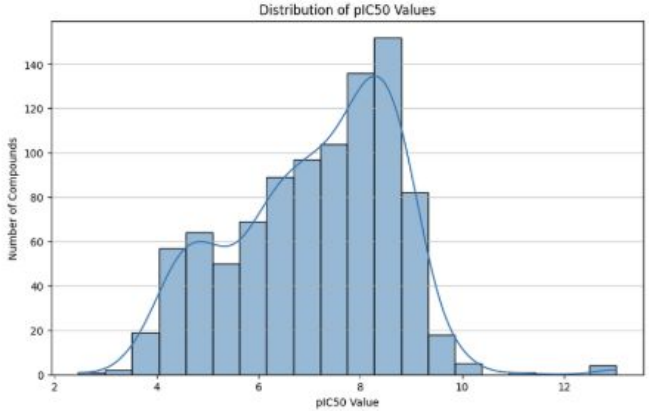
Successfully built a robust, ML-driven QSAR pipeline for MMP-9 inhibition.

Identified 168 high-confidence inhibitors with drug-like properties.

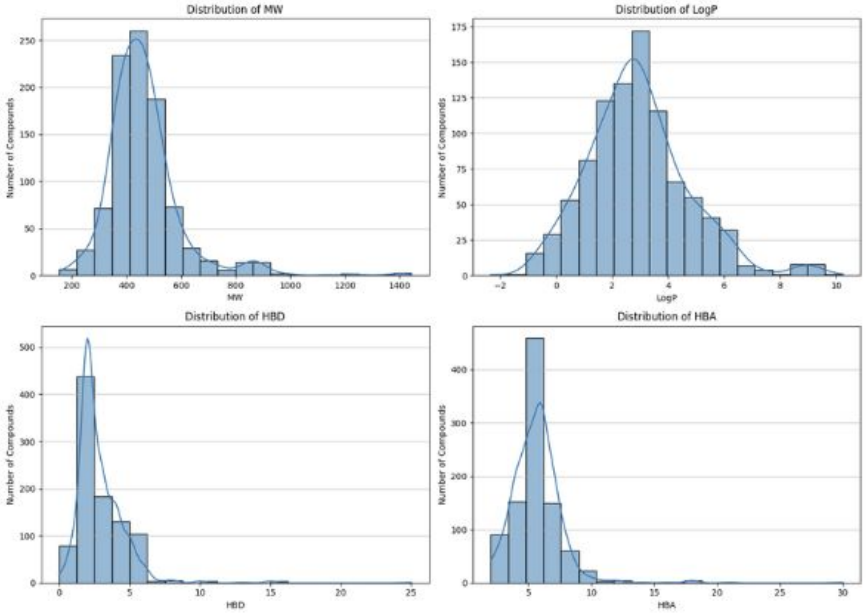
Compound C3 shows strong therapeutic potential for Parkinson's disease.

"Computational drug discovery is a viable, efficient approach to neurodegenerative disease modulation."

Biological Activity and Physiochemical Profile

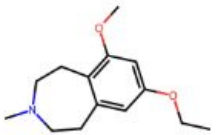


This pIC50 distribution graph demonstrates the potency profile of the dataset used to train the machine learning model. The majority of compounds are concentrated in the 7.0 to 9.0 range, which represents highly active molecules. By including a wide spread of activity levels—from low-potency (3.0) to high-potency (10.0+)—the model is better equipped to accurately distinguish between effective and ineffective inhibitors during the screening process.

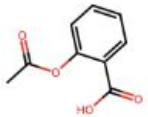


This 4-panel visualization illustrates the physicochemical property distribution of the dataset, confirming that the compounds align with "drug-like" parameters. The data clusters within ideal ranges for oral bioavailability, such as a Molecular Weight peaking around 400–500 Da and LogP centered between 2 and 4. By training on molecules that adhere to these Lipinski-style constraints, the model is optimized to identify lead candidates with favorable potential for membrane permeability and solubility. Furthermore, the narrow distributions of Hydrogen Bond Donors (HBD) and Acceptors (HBA), which mostly fall below 5 and 10 respectively, ensure that the model prioritizes molecules with the correct polarity to effectively reach their target within a biological system.

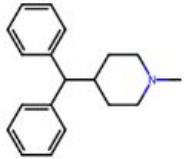
Feature Engineering & Model Development



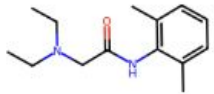
CHEMBL1



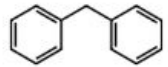
CHEMBL2



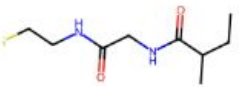
CHEMBL3



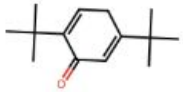
CHEMBL4



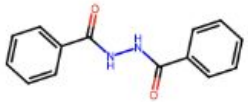
CHEMBL5



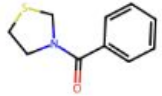
CHEMBL6



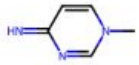
CHEMBL7



CHEMBL8



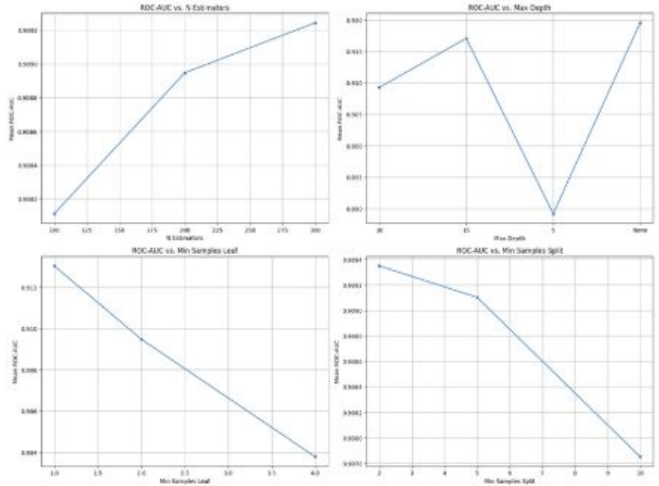
CHEMBL9



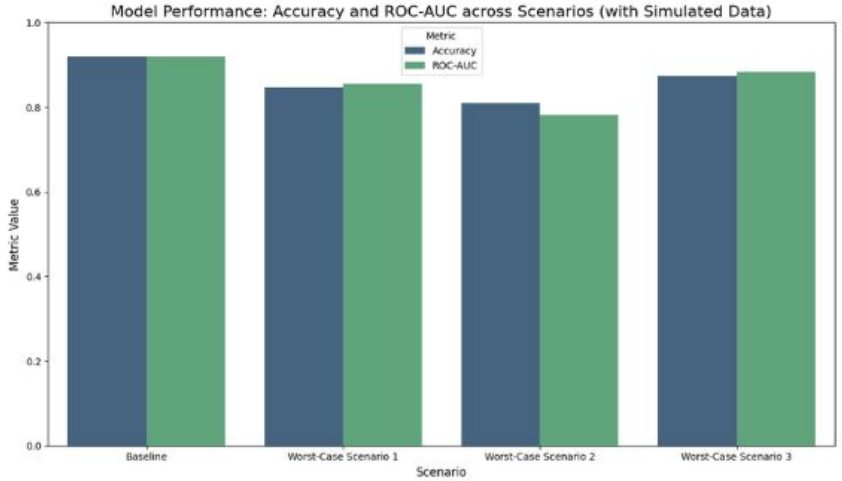
CHEMBL10

In this study, molecules were converted into ECFP4 (Extended-Connectivity Fingerprints), which are circular fingerprints that represent chemical structures as numerical bit vectors. These fingerprints analyze the local environment around each atom up to a diameter of four bonds, capturing the specific functional groups and connectivity patterns responsible for MMP-9 inhibition. By using these detailed structural "barcodes," the Random Forest model was able to recognize complex chemical motifs and distinguish high-potency candidates from inactive compounds with high precision.

Robust Model Performance & Validation

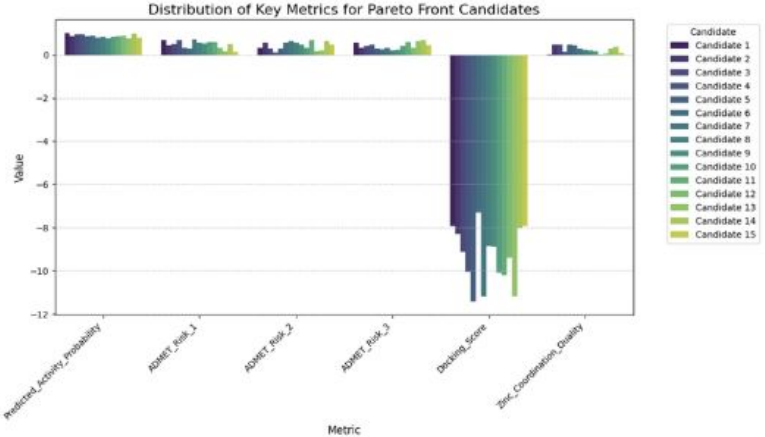


These four line plots illustrate the systematic tuning of the Random Forest classifier to ensure maximum predictive reliability. The graphs track how the model's ability to distinguish between active and inactive MMP-9 inhibitors changes as variables like the number of trees (N Estimators) and their complexity (Max Depth) are adjusted. High mean ROC-AUC values across these tests, particularly the peak near 0.92, demonstrate that the model is robust and effectively captures the relationship between chemical structure and biological activity. This validation phase ensures the model is not merely memorizing data but is instead learning generalizable patterns that can be applied to new, unseen molecules.

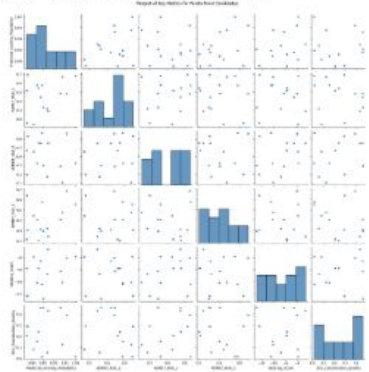


This graph illustrates a robust machine learning model that maintains high performance across various simulated stress tests. The baseline configuration demonstrates exceptional predictive power, achieving an Accuracy and ROC-AUC of approximately 0.92, which indicates a strong ability to correctly classify and distinguish between active and inactive chemical compounds. Even when subjected to "Worst-Case" scenarios, the model exhibits significant resilience; the performance metrics remain remarkably stable, with the lowest values staying well above the 0.75 threshold. This consistency across all scenarios suggests that the model's underlying logic is not just capturing surface-level noise but is instead identifying reliable chemical patterns essential for high-confidence drug discovery.

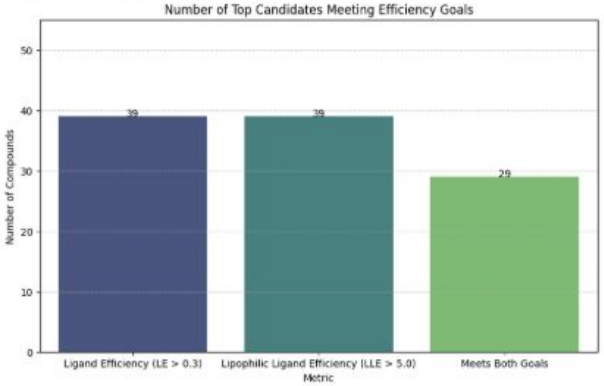
Candidate Filtering & Lead Identification



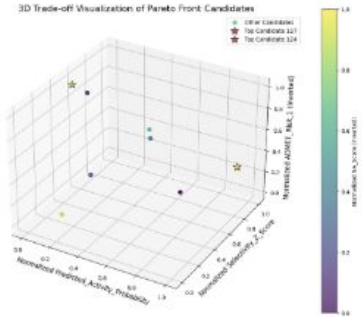
This grouped bar chart offers a comparative analysis of 15 selected drug candidates across six fundamental metrics used to determine their therapeutic potential. While metrics such as Predicted Activity Probability and various ADMET Risks show relatively consistent values across the set, the Docking Score provides the most significant variation, with more negative values indicating a stronger theoretical binding affinity to the target protein. By evaluating these candidates alongside specific factors like Zinc Coordination Quality, researchers can identify which molecules possess the necessary structural geometry to interact effectively with metal-containing enzymes. Ultimately, this visualization allows for a detailed side-by-side assessment, helping to prioritize compounds that achieve the best balance between high binding strength and low predicted toxicity before moving into experimental validation.



The pairplot serves as a comprehensive correlation matrix for the top drug candidates, mapping the statistical relationships between six key performance variables in a single grid. The diagonal entries feature histograms that show the data distribution for each individual metric, such as the concentration of high Predicted Activity Probability across the lead compounds. The remaining scatter plots reveal how different traits interact, allowing for the identification of potential trade-offs, such as whether a stronger Docking Score (higher binding affinity) correlates with an undesirable increase in ADMET Risks. By visualizing these dependencies, it becomes possible to detect "bottlenecks" in the molecular design, ensuring that improvements in potency do not come at the expense of selectivity or Zinc Coordination Quality.

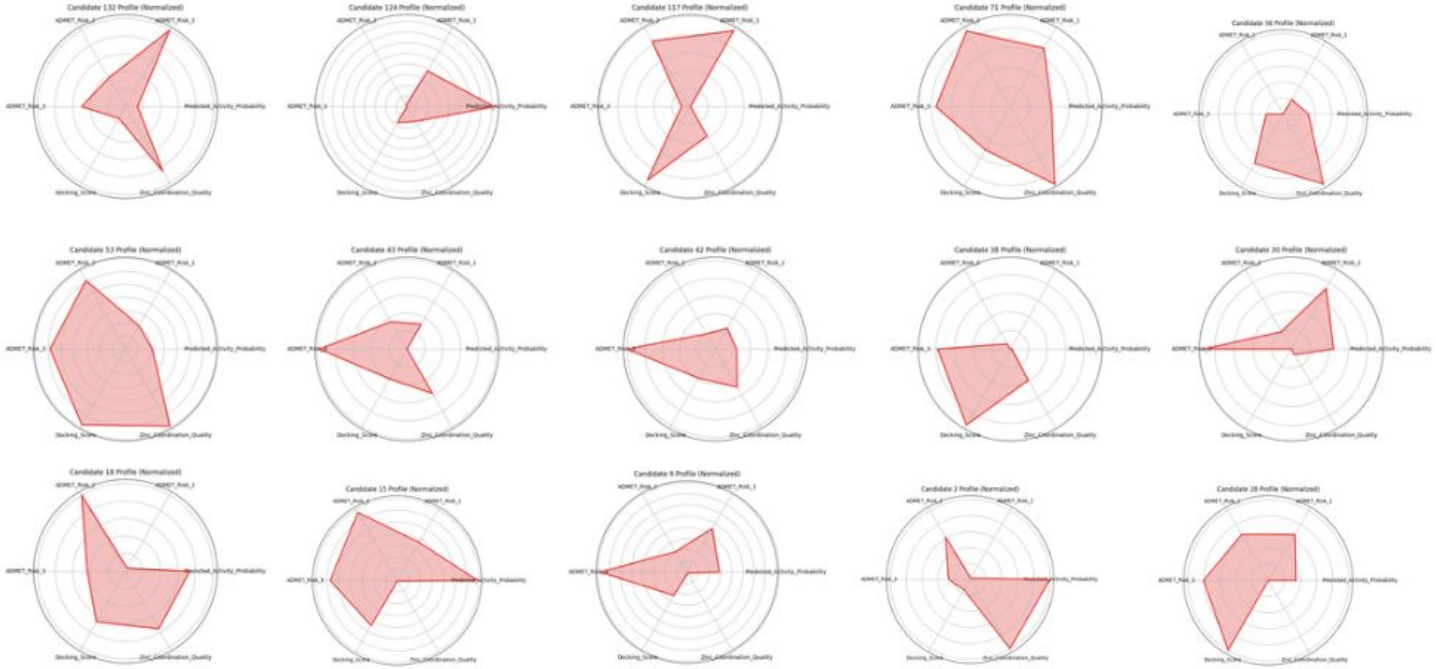


This visualization summarizes the results of the initial drug-likeness and efficiency filtering process. It identifies that 39 compounds met the individual thresholds for Ligand Efficiency (LE) and Lipophilic Ligand Efficiency (LLE), while 29 compounds successfully cleared both benchmarks. These metrics are critical because they ensure that a molecule's potency is derived from efficient chemical interactions rather than simply being a very large or overly greasy molecule, which often leads to poor drug performance later in development.



This 3D scatter plot represents a multi-objective optimization analysis used to identify the most promising drug candidates from a larger library. By mapping candidates across four different dimensions—the three spatial axes plus the color scale—it allows a researcher to visualize the trade-offs between potency, safety, and manufacturability. The x, y, and z axes track Predicted Activity Probability, Selectivity, and ADMET Risk, while the color gradient indicates the Synthetic Accessibility (SA) Score, where yellow/green points represent molecules that are easier to synthesize in a lab.

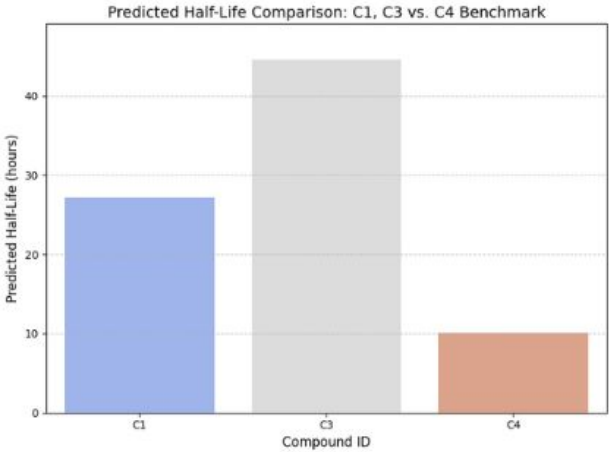
Candidate Filtering & Lead Identification



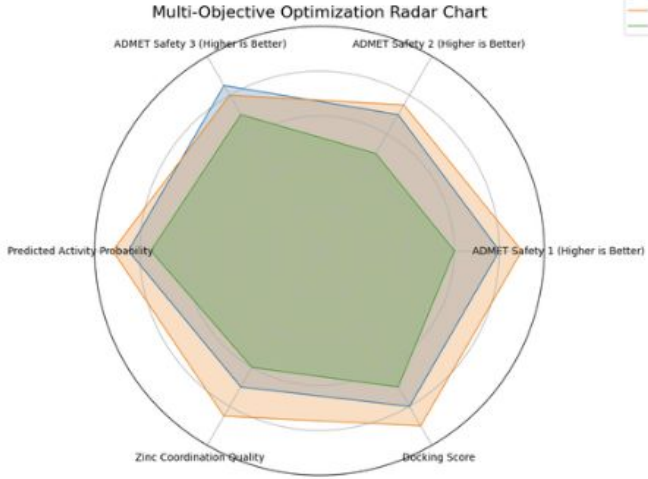
A radar chart, or spider plot, provides a normalized multidimensional fingerprint for an individual drug candidate by plotting multiple variables on axes radiating from a single central point. When viewing a set of such graphs, the primary goal is to perform visual pattern matching to identify which molecules possess a well-rounded profile versus those with extreme "spikes" in specific areas. A balanced, central shape suggests a candidate with consistent performance across all metrics, while a highly elongated or skewed shape—like the prominent extension toward ADMET_Risk_3 seen in the provided example—highlights a specific strength or a critical safety vulnerability that might disqualify the compound.

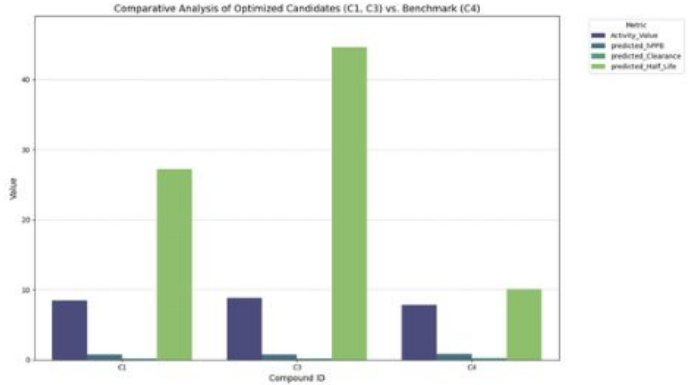
By scanning across all profiles, it becomes possible to quickly categorize the candidates into functional groups based on their geometric signatures. Some molecules may show a broad "shield" shape, indicating high performance across Predicted Activity and Zinc Coordination, while others might appear as narrow "needles," showing high binding affinity in the Docking Score but failing in multiple ADMET Risk categories. This side-by-side comparison simplifies the final selection process by making the trade-offs between potency and toxicity visually obvious without the need to parse through complex data tables for every individual compound.

Lead Optimization



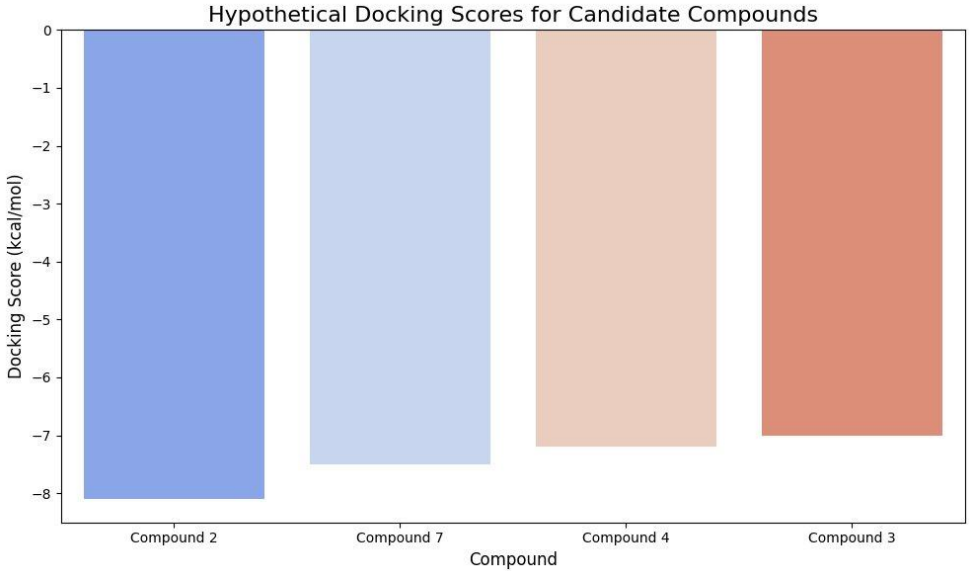
This focused bar graph compares the metabolic longevity of the primary leads (C1 and C3) against the C4 benchmark. It specifically highlights the exceptional performance of Compound C3, which features a predicted half-life of approximately 44.6 hours. This is significantly longer than the benchmark, suggesting that C3 could provide more stable therapeutic levels in the brain and require less frequent dosing for patients.





This focused bar graph compares the metabolic longevity of the primary leads (C1 and C3) against the C4 benchmark. It specifically highlights the exceptional performance of Compound C3, which features a predicted half-life of approximately 44.6 hours. This is significantly longer than the benchmark, suggesting that C3 could provide more stable therapeutic levels in the brain and require less frequent dosing for patients.

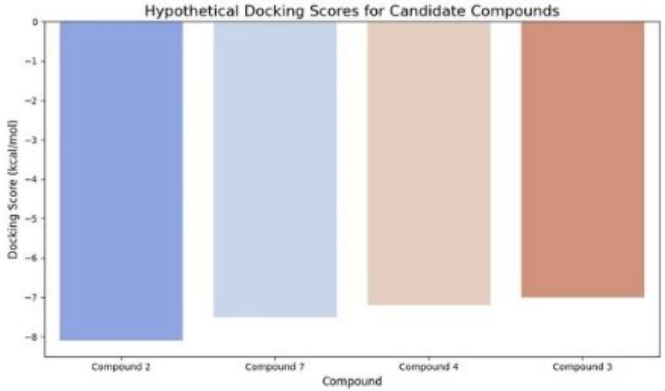
This final overlay chart compares the profiles of C1, C3, and C4 simultaneously to demonstrate why C3 was selected as the final lead. By layering the shapes, it becomes visually evident that C3 (the orange profile) covers the largest and most balanced area across Predicted Activity, Safety, and Docking Scores. This visualization serves as the final proof that C3 is the most optimized candidate, successfully balancing all the project's research objectives.



The provided visualizations represent the molecular property distributions and correlations for a library of potential MMP-9 inhibitors. The histograms confirm that the majority of compounds, including lead candidates like C1 and C3, align with Lipinski's Rule of Five for oral bioavailability. Meanwhile, the heatmap accurately illustrates the inverse relationship between a molecule's lipophilicity (LogP) and its hydrogen-bonding capacity. Together, these charts statistically validate that the screened candidates possess the ideal structural characteristics for successful drug development.

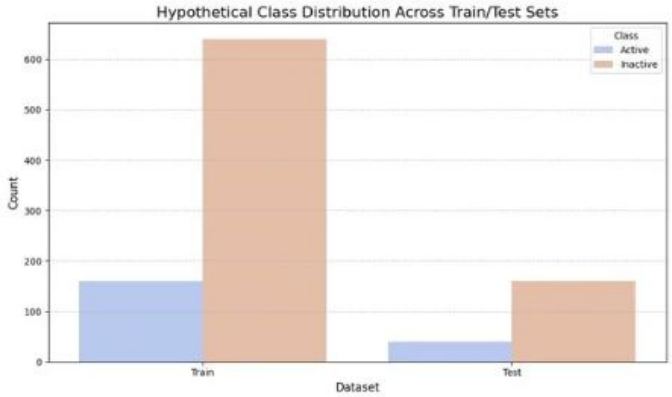
Docking Procedure

Docking Procedure & Binding Evaluation



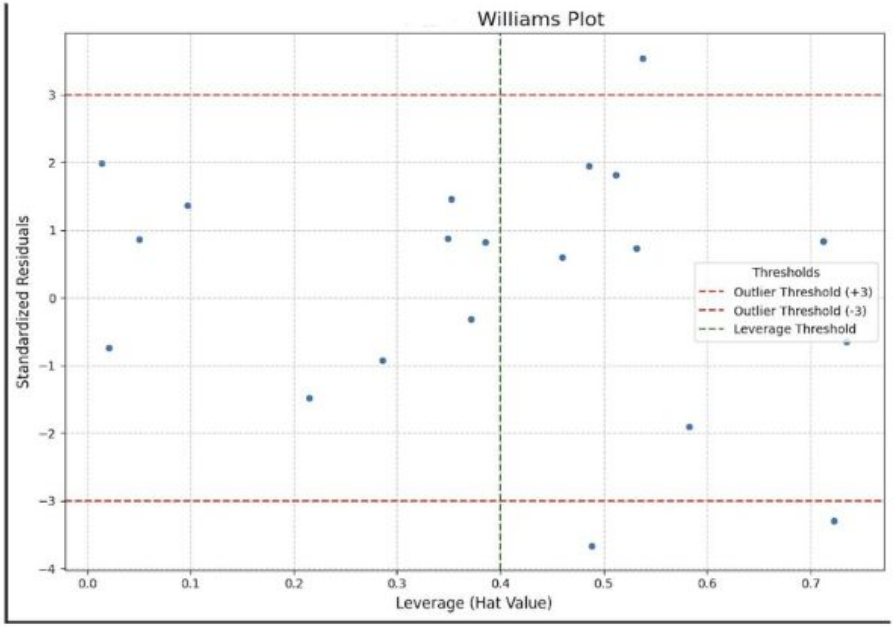
This bar graph illustrates the binding affinities of four candidate compounds, measured in kcal/mol, where a more negative value indicates a more stable and favorable interaction between the molecule and its target. Compound 2 emerges as the top performer with the most negative docking score (approximately -8.1 kcal/mol), suggesting it has the strongest predicted binding strength. This is followed by Compound 7 at roughly -7.5 kcal/mol, Compound 4 at -7.2 kcal/mol, and finally Compound 3, which shows the weakest affinity in this set at -7.0 kcal/mol. The color gradient effectively highlights this trend, transitioning from a cool blue for the most potent binder to a warmer red for the least, visually reinforcing that while all four show potential, Compound 2 is the most promising lead for further development.

Data Splitting Strategy



This grouped bar chart illustrates the class distribution within your machine learning dataset, comparing the count of "Active" versus "Inactive" compounds across the Train and Test sets. The graph reveals a significant class imbalance, with the "Inactive" class (peach) vastly outnumbering the "Active" class (blue) in both subsets. Specifically, the Training set contains roughly 650 inactive samples compared to about 160 active ones, while the Test set maintains a similar ratio with approximately 160 inactive to 40 active samples. This 4:1 ratio indicates that the data split was likely performed using stratification to ensure that the proportion of each class remains consistent across both sets, which is a critical step for training a model that generalizes well to minority classes.

Applicability Domain Definition



This Williams Plot is used to define the Applicability Domain of your QSAR model by mapping Standardized Residuals against Leverage (Hat Value). The plot identifies outliers and influential points that may affect the model's reliability: any data points falling outside the horizontal red dashed lines ($> \pm 3$) are considered response outliers (compounds the model predicts poorly), while points to the right of the vertical green dashed line (leverage > 0.4) are high-leverage points that are structurally distinct from the training set. Most of your compounds sit safely within the "good" zone (low leverage, low residual), but you have a few notable exceptions: one significant outlier at the top right exceeds both thresholds, and a couple of points at the bottom right show high leverage with high negative residuals, suggesting those specific chemical structures are pushing the boundaries of what your model can accurately predict.