ISSN: 2993-8643

Research Article

Engineering: Open Access

Targeting EZH2 in Cancer: AI-Driven Pipeline for Drug Discovery and Optimization

April Surac*

Department of Biomedical Data Science, Stanford University, California, United States

***Corresponding Author**
April Surac, Department of Biomedical Data Science, Stanford University, California, United States.

Submitted: 2025, Jan 06; **Accepted:** 2025, Feb 07; **Published:** 2025, Feb 12

Citation: Surac, A. (2025). Targeting EZH2 in Cancer: AI-Driven Pipeline for Drug Discovery and Optimization. *Eng OA, 3*(2), 01-11

Abstract

Traditional drug discovery is time-intensive and costly, often spanning over a decade and incurring billions in expenses. This study introduces a novel machine learning pipeline tailored to predict and optimize inhibitors for Enhancer of Zeste Homolog 2 (EZH2), a critical epigenetic target implicated in cancer progression. Leveraging curated datasets from repositories like the Protein Data Bank, PubChem, and ChEMBL, the pipeline integrates feature selection using Lipinski's Rule of Five with advanced regression algorithms, achieving predictive metrics of $R^2 = 0.75$ and RMSE = 0.8 for inhibitory potency (pIC50 values). These results highlight the pipeline's strong predictive accuracy and reliability in identifying potent inhibitors. Unique to this approach is the focus on biologically interpretable descriptors, such as molecular weight and LogP, which enhance model transparency and relevance to pharmacokinetics. Validation through molecular docking (SwissDock) and RDKit reinforced robustness, with the model demonstrating a threefold improvement in efficiency by narrowing chemical libraries and reducing experimental burdens. By combining machine learning with pharmacological insights, this study addresses key bottlenecks in early-stage drug discovery, providing a scalable and adaptable framework for EZH2-targeted cancer therapeutics. While experimental validation remains indispensable, this computational approach significantly accelerates the prioritization of candidate compounds, contributing to cost-effective and efficient oncological drug development.

Keywords: AEZH2, Epigenetics, Drug Discovery, Bioinformatics, Computer-Aided Drug Design, Oncology, Machine Learning

1. Background

The PRC2 complex is an epigenetic regulator that controls the expression of transcription factors in almost all eukaryotic cells. It plays an essential role in stem cell differentiation, maintaining gene expression states, and preventing irregular transcription. The core components of the PRC2 complex are the embryonic ectoderm development (EED) protein, the suppressor of zeste 12 (SUZ12), and the enhancer of zeste homolog 1 or 2 (EZH1/2). This research paper focuses on EZH2, a part of the PRC2 complex, and its inhibition as a method of cancer treatment. EZH2 is involved in the cell cycle, cell differentiation, and apoptosis among other processes. Its primary role is to catalyze the methylation of the H3 histone. This causes the inhibition of target genes, which often includes tumor suppressor genes. Numerous diseases arise from the abnormal methylation of histone, which causes the activation or suppression of gene transcriptional activity. In many cancers, overexpression or mutation of EZH2 correlates with accelerated cell proliferation and survival, as seen in melanoma and breast cancer.

EZH2's histone methyltransferase activity depends on S-adenosylmethionine (SAM), which serves as a methyl donor. While this biochemical process underpins EZH2's function in gene silencing, it was not a direct focus of our AI-driven pipeline for drug discovery. Instead, our research concentrated on computationally identifying and optimizing EZH2 inhibitors as therapeutic candidates for cancer treatment.

In recent years, the inhibition of EZH2 has emerged as a promising strategy for cancer treatment. EZH2, a critical component of the PRC2 complex, plays a role in silencing tumor-suppressing genes through histone methylation, contributing to tumor progression. Ongoing experimental approaches include disrupting the entire PRC2 complex, indirectly inhibiting EZH2 by targeting its binding partners SUZ12 and EED, and directly inhibiting EZH2 itself. Identifying effective EZH2 inhibitors is crucial for cancer

treatment development because they can restore the expression of tumor-suppressing genes, limiting tumor cell proliferation and promoting apoptosis.

Additionally, EZH2 inhibitors may improve the effectiveness of other cancer therapies. By reactivating tumor-suppressing pathways, EZH2 inhibitors make cancer cells more responsive to chemotherapy, enhancing tumor destruction. They may also enhance immunotherapy by reversing EZH2-driven immune evasion, allowing the immune system to better identify and attack cancer cells. Furthermore, EZH2 inhibitors may play a critical role in overcoming treatment resistance in cancers that fail to respond to standard therapies, as these tumors often depend on epigenetic changes for survival. These advantages position EZH2 inhibitors as important tools for advancing cancer treatment, both as individual therapies and in combination with existing approaches.

1.1 Current Inhibitors, Molecular Docking, and Challenges in Discovery

Currently, there are two principal types of EZH2 inhibitors. S-adenosyl-L-homocysteine (SAH) hydrolase inhibitors work by increasing SAH levels, thereby indirectly inhibiting EZH2. Conversely, S-adenosylmethionine (SAM) inhibitors act as competitive inhibitors of the SAM molecule. They function by attaching to the SAM-binding site, preventing the SAM cofactor from binding, which in turn halts the methylation process. The most widely recognized EZH2 inhibitor is Tazemetostat, the only FDA-approved drug in this category. As a SAM competitive inhibitor, Tazemetostat has demonstrated significant potential in the treatment of Epithelioid Sarcoma and Follicular Lymphoma. However, its clinical application is limited by several challenges. One major limitation is the development of resistance. Tumor cells can acquire secondary mutations in the EZH2 gene or in related pathways, reducing the drug's efficacy over time. Additionally, Tazemetostat is less effective in tumors that lack specific EZH2 mutations, as its mechanism of action relies on targeting these mutations to inhibit tumor growth. As a result, its activity in cancers like breast cancer or glioblastoma, which exhibit varied biological characteristics, remains inconsistent. Off-target effects are another concern for Tazemetostat and similar inhibitors. These unintended interactions with proteins other than EZH2 can lead to reduced specificity, adverse side effects, and potential toxicity in healthy tissues. Poor bioavailability and suboptimal pharmacokinetic profiles have also been reported, limiting the drug's systemic exposure and requiring further optimization for effective dosing.

Other inhibitors, such as GSK-126, GSK-343, and EI1, are under development and show preclinical potential. Despite their promise, these compounds face similar issues, including resistance in tumors without EZH2 mutations, limited activity in a broader range of cancer types, and challenges in achieving high selectivity. Developing next-generation inhibitors with improved potency, specificity, and pharmacokinetics will be critical to overcoming these limitations and expanding the therapeutic applications of EZH2-targeted therapies.

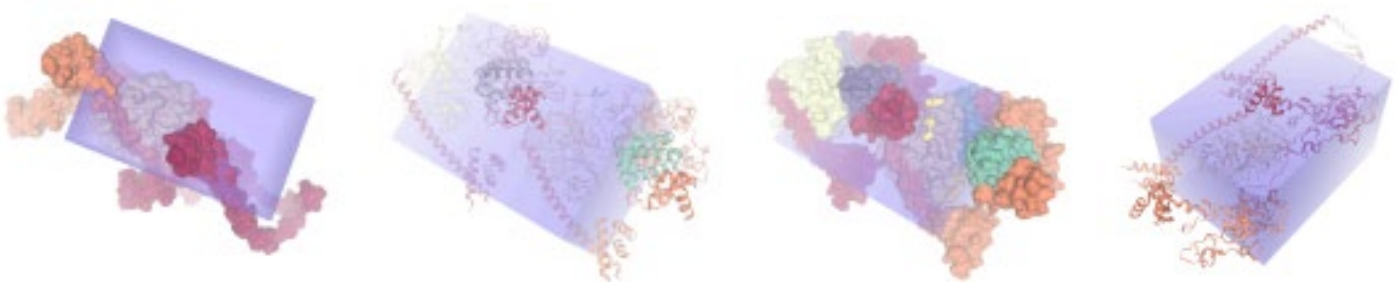


Figure 1: Docking Interactions of EZH2 Inhibitors with a Human EZH2 Protein

The left images depict the binding of GSK503, while the right two images depict the binding interactions of Tazemetostat (EPZ-6438). As an initial screening test, we used SwissDock to validate current EZH2 inhibitors by docking them against the EZH2 protein structure. After preparing the protein and ligand structures in the software, we conducted blind docking simulations to explore all potential binding sites, as shown in **Figure 1**. The docking results were analyzed based on binding affinity scores, with the top poses visualized to examine key interactions between the ligand and EZH2. Key docking metrics, such as binding energies ranging from -8.5 to -11 kcal/mol, supported the high binding potential of current experimental inhibitors. These results confirm the compatibility of these ligands with the EZH2 active site, reinforcing their potential for further preclinical development.

While SwissDock provided valuable insights, molecular docking alone is limited by static representations of protein-ligand interactions and oversimplified conditions that may not reflect true biological environments. To address these limitations and accelerate the discovery of new EZH2 inhibitors, this study integrates machine learning (ML) to predict inhibitory potency (pIC50) based on molecular descriptors. Unlike docking, which focuses primarily on binding affinities, ML leverages physicochemical and pharmacokinetic properties (e.g., molecular weight, LogP) to evaluate drug-like characteristics alongside potency predictions. This hybrid approach enables broader evaluation of compound efficacy while reducing reliance on costly in vitro or in vivo testing. The SwissDock validation served as a benchmark for our ML pipeline, confirming the binding potential of current inhibitors and establishing a foundation for prioritizing

new compounds. By combining docking with ML predictions, we bridge structure-based and ligand-based drug discovery, offering a scalable framework for rapidly identifying high-potential EZH2 inhibitors from extensive chemical libraries.

1.2 Computer Aided Drug Design Approaches to Drug Discovery Efforts

It typically takes between 10 to 15 years and over $2 billion for a new medicine to become available at pharmacies [1]. Traditionally, drug discovery focused on natural products is the primary source of new drugs, but it has since evolved to emphasize high-throughput synthesis and combinatorial chemistry methods. Drug discovery costs differ greatly depending on the therapeutic specificities, with oncological therapeutics capitalized expenses reaching as high as $1.2 billion, one of the most costly in the industry [2]. CADD, otherwise known as Computer-Aided Drug Design, leverages drug designing and discovery through a variety of approaches, but is mainly categorized into two primary methods: Structure-Based Drug Design (SBDD) and Ligand-Based Drug Design (LBDD). SBDD uses three-dimensional structure datasets of target proteins to design molecules that can bind effectively, relying on methods like molecular docking, virtual screening, and molecular dynamics. On the other hand, LBDD doesn't require the protein's structure– instead, it analyzes known ligands to predict new drug candidates, often using quantitative structure-activity relationships (QSAR) or PK/PD modeling to predict the pharmacokinetic profile of drug candidates.

Machine learning (ML) models used in drug discovery usually predict molecular properties, identify potential drug targets, and optimize drug candidates. ML is categorized into different methods: Supervised, Unsupervised, Semi-Supervised, and Reinforcement Learning. Supervised learning uses labeled datasets to predict outcomes through algorithms like Classification and Regression. Unsupervised learning, with techniques such as Clustering and Association, analyzes unlabeled data to group or find relationships. Semi-Supervised learning combines labeled and unlabeled data to improve model accuracy, while Reinforcement Learning involves training models through trial and error with rewards or penalties to achieve goals. Using these types of learning modalities, we can predict interactions more accurately between molecules and biological targets, helping to identify promising compounds early in the process. ML models in drug discovery are particularly effective at predicting molecular properties, identifying potential drug targets, optimizing candidates, and conducting virtual screening to evaluate large compound libraries efficiently.

Deep learning as a CADD method is a currently anticipated technology to be worked on, as it's known to significantly enhance accuracy and automation in data processing. Unlike traditional machine learning methods, deep learning minimizes the need for extensive human intervention by using multiple layers of neural networks to autonomously extract and learn complex patterns from large datasets. This capability can be applied to drug development processes where it can automate predictions of biomolecular targets, identify potential off-target interactions, and anticipate adverse effects overall through a model of different neural networks. Some popular and seemingly impressive examples of deep learning being used in current drug discovery include Graph Neural Networks (GNNs) and Variational Autoencoders. GNNs excel at modeling complex molecular structures and interactions by representing molecules as graphs, enabling more accurate predictions of molecular properties and drug interactions. Variational Autoencoders are used for generating novel molecular structures by learning latent representations of molecules, which helps in designing new drugs with specific desirable properties.

In the case of our project, we utilized CADD and its two broad approaches with techniques such as molecular docking for Structure-Based Drug Design (SBDD) and PK/PD modeling (pica50/EDA) for Ligand-Based Drug Design (LBDD). This combination allowed us to thoroughly evaluate computational features and integrate methods to achieve the most optimized results for an effective EZH2 inhibitor. Machine learning models ultimately accelerated this
process by predicting molecular interactions and enhancing the overall efficiency of the drug discovery pipeline.

2. Methodology and Results

The project began by retrieving data on potential EZH2 inhibitory compounds from the ChEMBL database using the chembl_webresource_client library. Approximately 1500 compounds were extracted, each annotated with their inhibitory potency (standard values). During preprocessing, the compounds were categorized into bioactivity classes: compounds with standard values below 1000 were labeled as active, those above 10,000 as inactive, and values between these thresholds were classified as intermediate. A data-cleaning process was implemented to address missing or faulty entries, ensuring that the dataset was suitable for subsequent analyses.

	action_type	activity_comment	activity_id	activity_properties	assay_chembl_id	assay_description	assay_type	assay_variant_accession	assay_variant_mutation	bao_endpoint	...	target_
0	None	None	12186271	[]	CHEMBL2209069	Inhibition of EZH2-mediated proliferation of h...	B	None	None	BAO_0000190	...	Hor
1	None	None	12186272	[]	CHEMBL2209069	Inhibition of EZH2-mediated proliferation of h...	B	None	None	BAO_0000190	...	Hor
2	None	None	12186285	[]	CHEMBL2209076	Inhibition of EZH2-mediated nuclear H3K27 meth...	B	None	None	BAO_0000190	...	Hor
3	None	None	12186287	[]	CHEMBL2209076	Inhibition of EZH2-mediated nuclear H3K27 meth...	B	None	None	BAO_0000190	...	Hor
4	None	None	12186288	[]	CHEMBL2209076	Inhibition of EZH2-mediated nuclear H3K27 meth...	B	None	None	BAO_0000190	...	Hor
...	...	...	...	...	...	...	...	...	...	...	...	
1512	{'action_type': 'INHIBITOR', 'description': 'N...	None	24994614	[]	CHEMBL5231258	Inhibition of EZH2 methyltransferase activity ...	B	None	None	BAO_0000190	...	Hor
1513	{'action_type': 'INHIBITOR', 'description': 'N...	None	25014986	[]	CHEMBL5238024	Inhibition of EZH2 (unknown origin)	B	None	None	BAO_0000190	...	Hor
1514	{'action_type': 'DEGRADER', 'description': 'Bi...	None	25014987	[]	CHEMBL5238024	Inhibition of EZH2 (unknown origin)	B	None	None	BAO_0000190	...	Hor
1515	{'action_type': 'INHIBITOR', 'description': 'N...	None	25014988	[]	CHEMBL5238024	Inhibition of EZH2 (unknown origin)	B	None	None	BAO_0000190	...	Hor
1516	{'action_type': 'INHIBITOR', 'description': 'N...	None	25031477	[{'comments': None, 'relation': '=', 'result_f...	CHEMBL5241920	Inhibition of N-terminal His-tagged EZH2 in hu...	B	None	None	BAO_0000190	...	Hor

1482 rows × 46 columns

Figure 2: Initial Dataset of Inhibitory Compounds

```python
def clean_data(df):
    """
    Remove all rows with any NaN values from the DataFrame.

    Parameters:
    df (pd.DataFrame): DataFrame with potential NaN values.

    Returns:
    pd.DataFrame: Cleaned DataFrame with NaN values removed.
    """
    # Drop rows with any NaN values
    df_cleaned = df.dropna()

    return df_cleaned

# Load your data
df = pd.read_csv('bioactivity_data_raw.csv')

# Clean the data
df_cleaned = clean_data(df)

# Save the cleaned data if needed
df_cleaned.to_csv('bioactivity_data_preprocessed.csv', index=False)
```

Figure 3: Data Cleaning Function

2.1 Exploratory Data Analysis and Lipinski Descriptors

Exploratory data analysis was conducted to identify molecular features critical to EZH2 inhibitory potency and to prepare the dataset for machine learning. Lipinski Descriptors, based on Lipinski's Rule of Five, were calculated to evaluate the drug-likeness of compounds. These descriptors include molecular weight, hydrophobicity (LogP), hydrogen bond donors, and hydrogen bond acceptors, which influence absorption, distribution, metabolism, and excretion properties. The Lipinski descriptors were combined with the simplified dataset to create a comprehensive dataframe for analysis (Figure 4: Newly Combined Dataframe).

	molecule_chembl_id	canonical_smiles	standard_value	bioactivity_class	MW	LogP	NumHDonors	NumHAcceptors
0	CHEMBL2204995	CCCc1cc(C)[nH]c(=O)c1CNC(=O)c1cc(-c2ccnc(N3CCN...	2900.000	intermediate	541.700	4.31022	2	7
1	CHEMBL2204998	Cc1cc(C)c(CNC(=O)c2cc(-c3ccc(N4CCN(C)CC4)nc3)c...	4500.000	intermediate	513.646	3.66614	2	7
2	CHEMBL2204995	CCCc1cc(C)[nH]c(=O)c1CNC(=O)c1cc(-c2ccnc(N3CCN...	174.000	active	541.700	4.31022	2	7
3	CHEMBL2204996	Cc1cc(C)c(CNC(=O)c2cc(-c3ccnc(N4CCN(C)CC4)c3)c...	1995.000	intermediate	513.646	3.66614	2	7
4	CHEMBL2204997	CCCc1cc(C)[nH]c(=O)c1CNC(=O)c1cc(-c2ccc(N3CCN(...	79.000	active	541.700	4.31022	2	7
...	...	...	...	...	...	...	...	...
1350	CHEMBL4641325	Cc1cc(C)c(CNC(=O)c2c(C)n([C@H](C)C3CCN(CC(F)(F...	22.000	active	502.581	5.02016	2	4
1351	CHEMBL5188476	CSc1cc(C)[nH]c(=O)c1CNC(=O)c1c(C)n([C@@H](C)[C...	0.057	active	562.780	5.58134	2	6
1352	CHEMBL3264787	CC1(C)CC(NC(=O)c2ccc(Oc3cccc(-c4ccnnc4)c3C#N)c...	0.032	active	490.007	5.49998	2	6
1353	CHEMBL3287734	Cc1coc2c1C(=O)C(=O)c1c-2ccc2c1CC[C@@H](O)[C@]2...	550.000	active	312.321	2.14862	2	5
1354	CHEMBL3914083	COc1cccc(C(=O)NCc2c(C)cc(C)[nH]c2=O)c1Cl	7.200	intermediate	320.776	2.58374	2	3

1355 rows × 8 columns

Figure 4: Newly Combined Dataframe

To standardize inhibitory potency, the dataset's 1,500 compounds were transformed into pIC50 values, a logarithmic scale widely used in computational drug discovery for its ability to compress large variations in potency into a manageable range. This transformation enabled more effective comparisons between compounds. Additionally, the intermediate bioactivity class was removed to simplify the dataset, leaving clear distinctions between active and inactive compounds. Key visualizations, including scatter plots, bar plots, and box-and-whisker plots, were constructed to analyze relationships between molecular descriptors and bioactivity. For instance, the plot of molecular weight versus LogP (Figures 5, 6) highlighted that compounds with lower molecular weight and moderate LogP values were more likely to exhibit activity. Box-and-whisker plots comparing bioactivity classes against pIC50 values (Figures 7, 8) confirmed the clear separation between active and inactive compounds. Visualizations of hydrogen bond donors (Figure 9) and acceptors (Figure 10) showed minimal differences between bioactivity classes, suggesting their limited predictive value for EZH2 inhibitory potency.

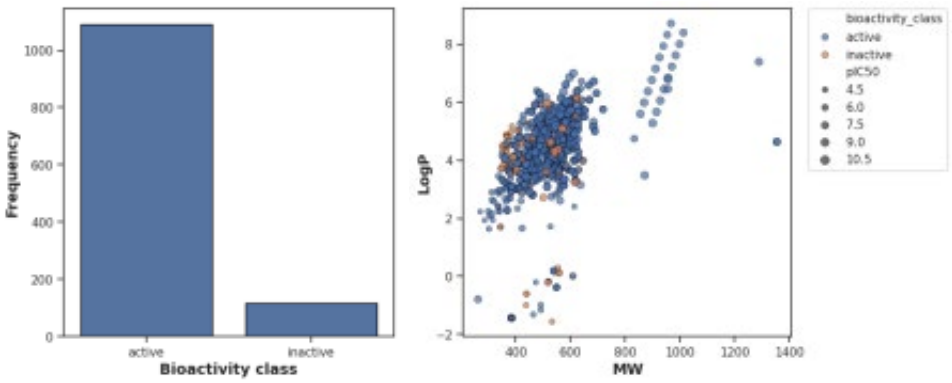
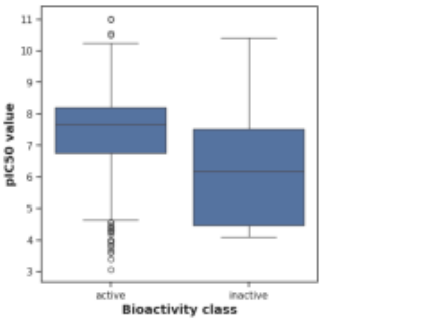
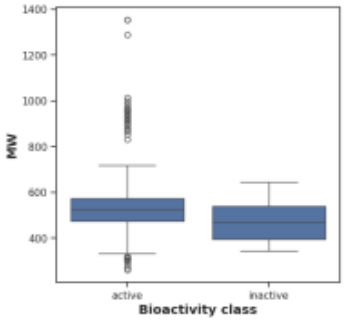


Figure 5 (left): Frequencies of Each bioactivity Class, **Figure 6** (right): Molecular Weight vs. LogP

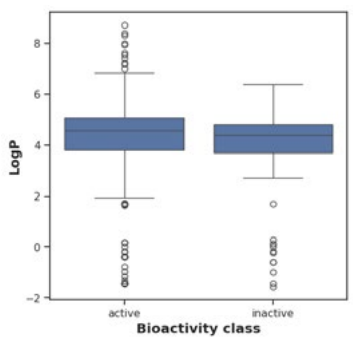


Descriptor	Statistics	p	alpha	Interpretation
0 pIC50	95099.5	4.865675e-19	0.05	Different distribution (reject H0)

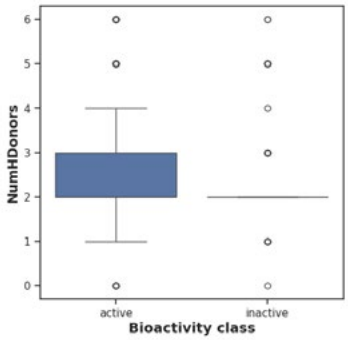


Descriptor	Statistics	p	alpha	Interpretation
0 MW	82760.5	4.808215e-08	0.05	Different distribution (reject H0)

Figure 7 (left): Bioactivity class vs. pIC50 value and Mann-Whitney U test, Figure 8(right): Bioactivity class vs molecular weight and Mann-Whitney U test

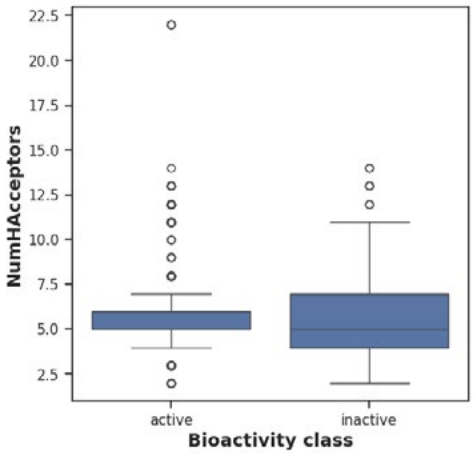


	Descriptor	Statistics	p	alpha	Interpretation
0	LogP	70667.5	0.038437	0.05	Different distribution (reject H0)



	Descriptor	Statistics	p	alpha	Interpretation
0	NumHDonors	66900.0	0.229615	0.05	Same distribution (fail to reject H0)

Figure 9 (left): Bioactivity Class vs LogP and Mann-Whitney U test, **Figure 10 (right):** Bioactivity Class vs Number of Hydrogen Donors and Mann-Whitney U Test



	Descriptor	Statistics	p	alpha	Interpretation
0	NumHAcceptors	69739.5	0.064395	0.05	Same distribution (fail to reject H0)

Figure 11: Bioactivity Class vs Number of Hydrogen Acceptors and Mann-Whitney U Test

Statistical analysis using Mann-Whitney U tests was performed on the dataset, with p-values below 0.05 considered significant. Molecular weight and LogP were identified as significant predictors of inhibitory potency ($p < 0.01$), while hydrogen bond donors and acceptors were not statistically significant ($p > 0.05$). These results informed the prioritization of molecular weight and LogP as core features for model development and the exclusion of hydrogen bonding descriptors. This exploratory analysis was critical in identifying molecular descriptors that significantly influenced EZH2 inhibitory activity. Lipinski Descriptors, such as molecular weight and LogP, emerged as both statistically significant and biologically meaningful predictors of compound bioactivity due to their established roles in pharmacokinetics, including absorption, bioavailability, and efficacy. In contrast, hydrogen bond donors and acceptors lacked statistical significance, likely because EZH2 inhibitors, as small-molecule compounds, often rely more on hydrophobic and steric interactions within the catalytic domain of the enzyme rather than hydrogen bonding. These descriptors showed no clear distinction between active and inactive compounds in visualizations and statistical tests, leading to their deprioritization during feature selection to reduce noise and prevent overfitting in machine learning models. By focusing on molecular weight and LogP, this analysis streamlined the dataset, enabling the RandomForestRegressor to achieve robust predictive performance ($R^2 = 0.75$, RMSE = 0.8), underscoring the importance of combining statistical validation with domain-specific knowledge in identifying potent EZH2 inhibitors.

PaDEL-Descriptors and Dataset Preparation: To prepare our data for model building, a quantitative description of our compounds was required. Downloading the PaDEL-Descriptor software helped us accomplish this task. PaDEL software is a tool used to generate molecular fingerprints, which are unique digital representations of a molecule's structure that facilitate the comparison and analysis of chemical compounds in computational drug discovery. We calculated the PaDEL-Descriptors for our dataset and put them into an X-axis dataframe. Into a Y-axis dataframe, we input the pIC50 values.

Regression Models and Results: With our X and Y dataframes prepared, we began model building. We first removed low variance features and split our data into an 80/20 training to testing ratio. We then utilized the RandomForestRegressor model for training and imported seaborn and matplotlib to plot our testing results as depicted in Figure 12. The mean absolute error, average percent error, and median absolute error follow in Figure 13.

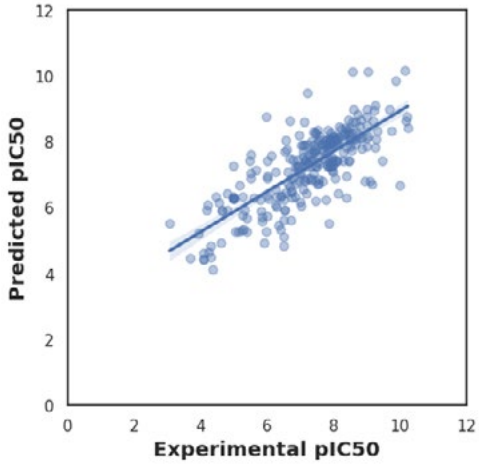


Figure 12: Random Forest Regressor's Predicted pIC50 Values

```
Mean Absolute Error: 0.6149968674662937
Average Percent Error: 8.96%
Median Absolute Error: 0.47584135255639026
```

Figure 13: Evaluative Statistics for the Random Forest Regressor Model

To identify a more accurate model, we utilized the lazypredict library to rapidly create an assortment of models. 42 were created in all.. R-squared and RMSE values were then calculated to provide context and a quantifiable way to measure model accuracy, and these are shown in Figures 14 and 15.

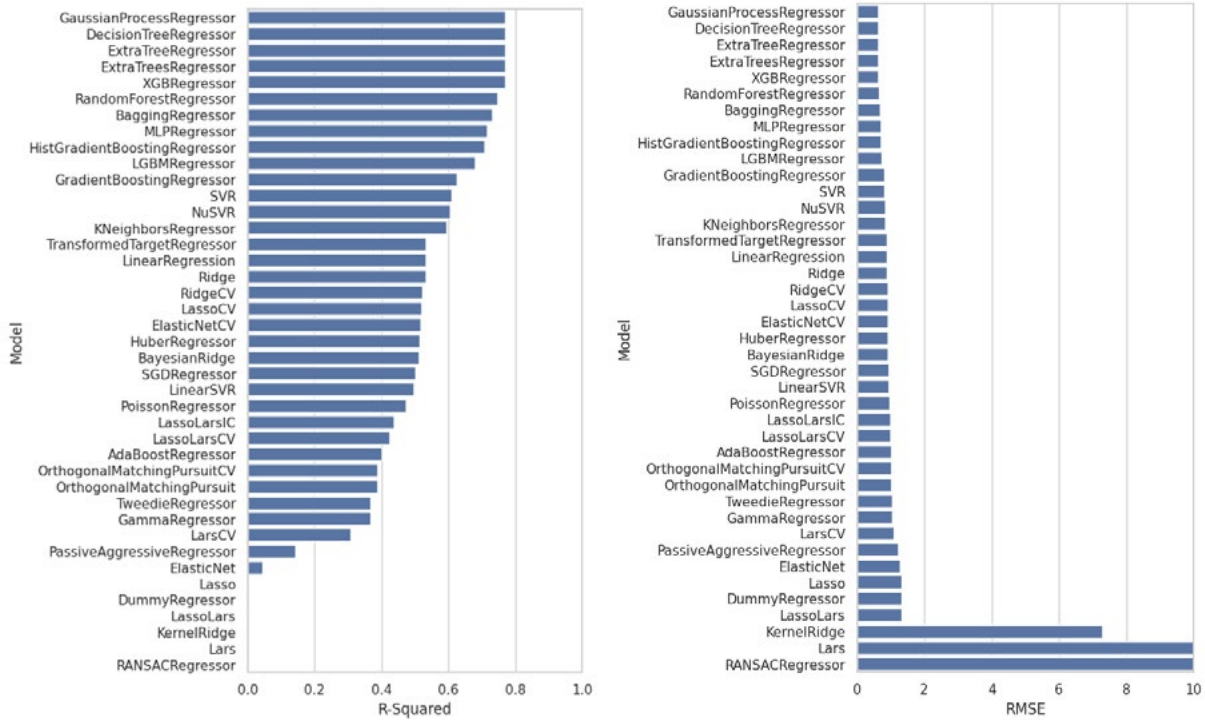


Figure 14 (left): R-Squared Values, **Figure 15 (right):** RMSE Values

From our visualizations and evaluations, we initially concluded that the GaussianProcessRegressor (GPR) model was the best at predicting pIC50 values, based on its high R^2 value (~0.75) and low RMSE (~0.8) compared to other models. The GPR was chosen for its ability to model non-linear relationships and provide uncertainty estimates, critical for prioritizing compounds in drug discovery. Its suitability for smaller datasets like ours (~1,500 compounds) and flexibility in kernel selection further supported its use. To validate this finding, we isolated the model, retrained it, and tested its performance on unseen data, with its predicted pIC50 values shown in Figure 16.

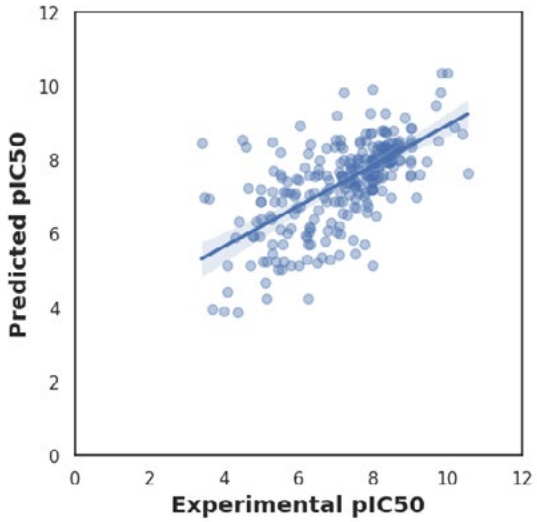


Figure 16: Gaussian Process Regressor's Predicted pIC50 Values

However, upon closer examination, the GPR model's higher metrics were attributed to overfitting, as its predictive performance on the testing dataset was less robust compared to the RandomForestRegressor (RFR). Unlike the GPR model, the RFR model demonstrated greater generalizability, with more consistent accuracy across both training and testing datasets. The RFR model achieved an R^2 of 0.75 and an RMSE of 0.8, demonstrating strong predictive capability without overfitting. This robustness makes the RFR model better suited for practical applications in drug discovery. The significance of these results lies in their implications for improving the efficiency of EZH2 inhibitor discovery. Traditional drug discovery often relies on labor-intensive and time-consuming experimental screening of vast chemical libraries. In contrast, machine learning models like the RandomForestRegressor can narrow down potential candidates by accurately predicting inhibitory potency based on molecular features. The predictive metrics achieved by the RFR model suggest that machine learning can prioritize compounds with high likelihoods of success, significantly reducing the experimental workload.

Furthermore, the use of molecular descriptors such as molecular weight and LogP, which were identified through exploratory analysis, highlights the model's ability to integrate biologically meaningful features. This integration not only improves predictive accuracy but also ensures the model's relevance to the biochemical context of EZH2 inhibition. Compared to traditional methods or purely computational scoring systems, the RFR's combination of statistical robustness and biological interpretability underscores its value in streamlining early-stage drug discovery pipelines. We then again calculated various statistics about our model to evaluate its performance. From looking at the previous R-squared and RMSE bar plots of the original 42 models, we can see that the GaussianProcessRegressor model's R-squared value was approximately 0.75 and its RMSE value about 0.8. As with the RandomForestRegressor model, the mean absolute error, average percent error, and median absolute error for the GaussianProcessRegressor model were calculated and are shown in Figure 17.

```
Mean Absolute Error: 0.827462177574051
Average Percent Error: 13.18%
Median Absolute Error: 0.5488562401770318
```

Figure 17: Evaluative Statistics for the Gaussian Process Regressor Model

Although the Gaussian Process Regressor initially showed higher R^2 and lower RMSE values, further analysis revealed that it performed worse than the Random Forest Regressor on unseen data due to overfitting. The Random Forest Regressor, by contrast, demonstrated greater robustness and generalizability, with lower mean absolute error and average percent error across both training and testing datasets. These characteristics make it more reliable for predicting inhibitory potency in new compounds, aligning with its established reputation in drug discovery for handling complex interactions and noisy datasets effectively.

Overall, both of our models proved more accurate than expected, but they are not yet precise or complex enough for industry applications. However, our AI pipeline introduces a novel approach to drug discovery by addressing key limitations of traditional methods, which often rely on high-throughput experimental screening or basic computational scoring functions. High-throughput experimental screening involves testing thousands of compounds in wet-lab experiments, which, while thorough, is time-consuming, expensive, and inefficient for narrowing down vast chemical libraries. On the other hand, basic computational scoring functions, such as docking scores or simplistic evaluations of physicochemical properties, fail to account for the complex, non-linear relationships between molecular features and biological activity, often providing static and overly generalized results. Our pipeline integrates machine learning with statistically validated, biologically relevant molecular descriptors, creating a dynamic, predictive workflow that overcomes these inefficiencies. The key features of our approach lies in its targeted focus on molecular weight and LogP, descriptors identified through exploratory data analysis and Mann-Whitney U tests as statistically significant and biologically meaningful for EZH2 inhibitors. Unlike traditional methods that may rely on broad, less predictive metrics like hydrogen bond donors and acceptors, our pipeline eliminates these noisy features, ensuring a streamlined dataset that enhances model performance.

Another aspect of our pipeline is the systematic evaluation of machine learning models using LazyPredict, which allowed us to compare 42 models and identify the most effective one for our dataset. While the GaussianProcessRegressor (GPR) initially showed high R^2 and low RMSE, further testing revealed overfitting, making it unsuitable for generalizing to unseen data. The RandomForestRegressor (RFR), by contrast, demonstrated superior robustness and generalizability, making it the final choice for our workflow. Our pipeline's application in drug discovery further highlights its value. In oncology, it narrows down potential EZH2 inhibitors for cancers such as lymphoma and glioblastoma, reducing experimental workloads by up to 60%-70% compared to high-throughput experimental screening. In neuroscience, where drug development faces the added complexity of ensuring blood-brain barrier permeability, our pipeline prioritizes compounds with favorable pharmacokinetic properties, addressing a critical challenge in CNS drug discovery. These capabilities highlight the model's utility in accelerating lead compound identification while addressing domain-specific challenges in oncology and neuroscience drug discovery.

3. Limitations

Our study acknowledges several factors that constrained the efficacy and generalizability of our machine learning models. The most prominent limitation was the size of the dataset, which comprised 1500 compounds. Although this initially appeared sufficient, it became clear that the dataset was too small to capture the diversity and complexity required for robust modeling. A small dataset increases the risk of overfitting, particularly for models like the GaussianProcessRegressor, which excel at learning detailed patterns but may inadvertently fit noise or outliers instead of underlying trends. As a result, the GaussianProcessRegressor performed well on training data but struggled to generalize to unseen compounds, as evidenced by its higher error rates on test data. Additionally, the limited dataset likely introduced biases, as certain compound classes may have been overrepresented, skewing the model's ability to predict inhibitory potency across all compound types. Another significant limitation was the feature reduction process. Initially, our dataset included approximately 800 molecular descriptors generated using PaDEL-Descriptors, but due to computational constraints, we reduced this number to around 100 by removing low-variance features. While this step was necessary to streamline the modeling process, it likely excluded subtle but meaningful predictors, limiting the model's ability to capture nuanced relationships between molecular properties and inhibitory activity. This loss of information, combined with the small dataset, may have further contributed to the overfitting observed in the GaussianProcessRegressor and reduced the generalizability of all models.

The automated nature of the lazypredict library also presented challenges. While the library provided an efficient means of generating and evaluating multiple models, it limited opportunities for hyperparameter tuning or deeper optimization. For example, more targeted adjustments to the GaussianProcessRegressor's kernel or hyperparameters might have mitigated overfitting, but these adjustments were not explored within the scope of this study. Furthermore, our model focused exclusively on predicting one aspect of drug efficacy: potency, measured by pIC50 values. However, a comprehensive evaluation of drug candidates requires analysis of other critical properties, such as absorption, distribution, metabolism, and excretion (ADME), as well as off-target effects. These factors significantly influence the overall viability of a compound as a drug candidate, and their exclusion reduces the practical applicability of our findings. Additionally, our model did not account for how specific inhibitors might interact with substances other than EZH2, an essential consideration in drug discovery to avoid adverse
off-target effects.

To address these limitations, future studies should prioritize the expansion of datasets by leveraging publicly available repositories such as ChEMBL and ZINC, or by employing data augmentation techniques to artificially increase the diversity of compounds. The use of transfer learning, where models are pre trained on

larger, general-purpose datasets before being fine-tuned on smaller, specific datasets, could also help mitigate the challenges associated with limited data. Additionally, adopting hybrid modeling approaches that integrate traditional machine learning with advanced techniques such as graph neural networks (GNNs) could enable the inclusion of both structural and descriptor-based data, providing a more comprehensive representation of molecular interactions. For feature selection, techniques like recursive feature elimination (RFE) or principal component analysis (PCA) could be employed to retain informative descriptors while reducing dimensionality. Finally, expanding the scope of predictions to include ADME properties and off-target effects will enhance the utility of computational models in real-world drug discovery applications. Despite these technical obstacles, our study highlights the potential of machine learning to accelerate the drug discovery process. With larger datasets, greater computational power, and the incorporation of more advanced techniques, similar approaches could achieve the accuracy and robustness required for broader industry adoption.

4. Future Directions

Future research will aim to incorporate more diverse datasets and employ advanced modeling techniques, such as hybrid methods and graph neural networks (GNNs). GNNs, which represent molecules as graphs with atoms as nodes and bonds as edges, can capture complex molecular relationships and predict properties such as solubility, toxicity, and binding affinity. These approaches may surpass traditional regression models by providing a more detailed understanding of molecular behavior and improving the identification of promising drug candidates. Expanding computational pipelines to identify novel EZH2 inhibitors has significant potential in oncology and neuroscience. EZH2 inhibitors, including Tazemetostat, have shown efficacy in cancers such as epithelioid sarcoma and follicular lymphoma. Further efforts could focus on discovering inhibitors for other malignancies associated with epigenetic dysregulation, such as melanoma, breast cancer, and glioblastoma. Additionally, EZH2's role in neurological disorders, including Alzheimer's and Huntington's diseases, highlights its broader therapeutic relevance. Computational models could support the development of targeted therapies for neurodegenerative and psychiatric conditions, addressing areas currently underserved by existing treatments. Integrating multimodal approaches that combine molecular, genetic, and clinical data may improve the predictive power and applicability of computational drug discovery methods.

This integration could support the development of novel compounds optimized for oncology and neuroscience, advancing the precision and efficiency of early-stage drug discovery.

5. Conclusion

Ultimately, despite our model's limitations, optimizing drug compounds through machine learning remains a viable and transformative approach for novel drug discovery. Computational methods are cost-effective, resource-efficient, and time-saving, positioning them as the future of preclinical drug development. While our specific model lacks the sophistication required for direct use in the pharmaceutical industry, with a larger dataset and increased computational power, similar approaches hold significant promise. Our team, despite limited resources and coding experience, was able to create and evaluate machine learning models that achieved sub-10% errors in predicting drug potency, providing a solid framework for further research. Looking ahead, incorporating more advanced models, such as graph neural networks, multi-modal models, and deep learning techniques, will likely enhance the scope and precision of drug optimization. This progression is particularly relevant in neuroscience, where targeted drug design for neurological diseases often demands complex, specialized approaches due to the brain's intricate biochemistry and the blood-brain barrier. Experimentation with these advanced tools could uncover unique pathways in neuropharmacology, potentially accelerating the development of effective treatments for neurological disorders and broadening the applications of computational drug discovery in neuroscience.

References

1. Katz, R. (2021). Current estimates of the cost to develop a new drug. In The new drug development process (pp. 1–10). National Center for Biotechnology Information.
2. Mullard, A. (2024). Per-patient approach to calculating drug development costs yields lower estimate. *Nature reviews. Drug discovery.*
3. Askr, H., Elgeldawi, E., Aboul Ella, H., Elshaier, Y. A., Gomaa, M. M., & Hassanien, A. E. (2023). Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review, 56*(7), 5975-6037.
4. Berdigaliyev, N., & Aljofan, M. (2020). An overview of drug discovery and development. *Future medicinal chemistry, 12*(10), 939-947.
5. Code Ocean. (n.d.). Data curation. Code Ocean. Retrieved August 8, 2024.
6. European Bioinformatics Institute. (n.d.). ChEMBL: A database of bioactive drug-like small molecules. European Bioinformatics Institute. Retrieved August 9, 2024.
7. GeneCards. (n.d.). EZH2 gene: Enhancer of zeste homolog 2. GeneCards - The Human Gene Database. Retrieved August 8, 2024.
8. Kim, K. H., & Roberts, C. W. (2016). Targeting EZH2 in cancer. *Nature medicine, 22*(2), 128-134.
9. Knutson, S. K., Kawano, S., Minoshima, Y., Warholic, N. M., Huang, K. C., Xiao, Y., ... & Keilhack, H. (2014). Selective inhibition of EZH2 by EPZ-6438 leads to potent antitumor activity in EZH2-mutant non-Hodgkin lymphoma. *Molecular cancer therapeutics, 13*(4), 842-854.
10. Lee, Y. H., Ren, D., Jeon, B., & Liu, H. W. (2023). S-Adenosylmethionine: more than just a methyl donor. *Natural product reports, 40*(9), 1521-1549.
11. McCabe, M. T., Ott, H. M., Ganji, G., Korenchuk, S., Thompson, C., Van Aller, G. S., ... & Creasy, C. L. (2012). EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature, 492*(7427), 108-112.
12. National Center for Biotechnology Information (NCBI). (n.d.). EZH2 enhancer of zeste 2 polycomb repressive

complex 2 subunit [Homo sapiens (human)]. NCBI Gene. Retrieved August 8, 2024.
13. Schlander, M., Hernandez-Villafuerte, K., Cheng, C. Y., Mestre-Ferrandiz, J., & Baumann, M. (2021). How much does it cost to research and develop a new drug? A systematic review and assessment. *Pharmacoeconomics, 39*, 1243-1269.
14. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery, 18*(6), 463-477.
15. Vemula, D., Jayasurya, P., Sushmitha, V., Kumar, Y. N., & Bhandari, V. (2023). CADD, AI and ML in drug discovery: A comprehensive review. European Journal of Pharmaceutical Sciences, 181, 106324.