ChemPrint: An AI-Driven Framework for Enhanced Drug Discovery

Tyler Umansky[1]*, Navya Ramesh[1], Virgil Woods[1], Sean M. Russell[1], Davey Smith[2], Daniel Haders[1]**

[1]Model Medicines, [2]University of California San Diego

ABSTRACT

In the evolving landscape of drug discovery, traditional methods face significant challenges due to low efficiency and high resource demands. This study introduces our GALILEO AI drug discovery platform and its foundational ChemPrint model, designed to improve the efficiency of drug discovery. Addressing the challenges of low hit rates and the difficulty in exploring novel chemical spaces, this platform adopts adaptive molecular embeddings and stringent model training environments to enhance predictive capabilities and navigate uncharted molecular territories. We validate this approach through a case study targeting the AXL and BRD4 oncology targets, where ChemPrint achieved a 45.5% *in vitro* hit rate and identified 20 novel acting compounds. These compounds exhibited substantial chemical novelty, with an average Tanimoto similarity score of 0.38 to their training set, significantly diverging from known compounds and demonstrating ChemPrint's capability to extrapolate, surpassing traditional benchmarks and industry standards. This underscores the platform's ability to bridge the gap between the potential of AI and its practical application in therapeutic discovery.

1. INTRODUCTION

Conventional drug discovery and development are characterized by extensive resource demands and low efficiency, which pose significant challenges. In a typical early drug discovery program, the process of screening over 10,000 compounds with High-Throughput Screening (HTS) methods yields a mere 2.5% hit rate [1]. This is compounded by subsequent lead optimization screenings, which aim to bridge potency with other necessary features of a complete Target Product Profile (TPP). While critical to translating hit discovery to clinical readiness, this process traditionally yields even lower success rates (0.1-0.2%) [1]. In total, 10-15 years and up to ~$6 billion in capital is typically required to discover a compound and bring it to FDA approval [1,2]. The inefficiency of traditional drug discovery practices has necessitated the exploration of advanced methods, especially those that apply artificial intelligence (AI) to the task of molecular property prediction.

To date, no AI-discovered therapeutic has achieved FDA approval, and few have progressed beyond non-clinical animal model proof of concept. Despite advancements, current AI models struggle to improve hit rates sufficiently, as evidenced by Schrödinger's pipeline-wide average *in vitro* hit rate of 26% [3], alongside other groups reporting rates of 12%, 14.2%, and 26% [4,5]. Of the compounds discovered to be hits few, if any, demonstrate significant chemical diversity. A clear critical obstacle

persists with the inability of these models to predict the properties of compounds beyond well-tread chemical space [6,7]. These challenges arise because AI models are trained, tuned, and evaluated on datasets that lack chemical diversity using conventional performance indicators, a method that demonstrates the models' technical ability but fails to discover drugs that work in the real world. These standard approaches to model validation inadvertently bias models and limit competency to the same chemical space as their training data. New evaluation strategies must be deployed that prioritize model extrapolative ability in order to infer beyond known chemical territories and discover genuinely novel compounds that optimize desired traits.

In this study, our AI drug discovery platform, GALILEO, and its zero-shot capable Molecular-Geometric Deep Learning (Mol-GDL)[8] model, ChemPrint, are presented as solutions to these challenges. The success of our methodology is demonstrated in an oncology program targeting AXL and BRD4, two targets that demonstrate a particularly egregious attrition rate in the post HTS steps of drug discovery, rendering them the classification of 'undruggable' [9,10]. Our study addresses the limitations of conventional approaches and surpasses them through the employment of adaptive molecular embeddings and stringent model training environments, enabling our models to navigate unseen chemistry accurately.

Adaptive Molecular Embeddings

AI drug discovery models are limited by molecular embeddings that do not evolve or adapt as the model gains knowledge. The reliance on static molecular embeddings such as Extended Connectivity FingerPrints (ECFP) [11] limit the potential for groundbreaking discoveries. ECFP encodings use bits that treat similar molecular characteristics as if they were unrelated, and require machine learning algorithms to relearn relationships. As a result, stagnant fingerprints, such as ECFP, have limited predictive capabilities in prospective drug discovery studies. Conventional molecular representations like ECFP, while useful, fail to capture the complete multidimensional complexity of molecules in a useful way, rendering similar molecular features as disparate entities [12]. To address this, our study applies an adaptive molecular embedding technique within ChemPrint that evolves with the model's learning process, enabling the in silico identification of multiple novel therapeutic candidates.

Stringent Model Training Environments

In contrast to other machine learning (ML) applications, traditional data splitting techniques such as stratified random splitting fail to adequately quantify the prospective performance of ML models for molecular property prediction. Test sets generated by this method of splitting often limit models by presenting them with an oversimplified prediction task. As noted in Feinberg et al., "common cross-validation strategies, like random splitting, significantly overestimate the generalization and extrapolation capabilities of ML methods in drug discovery."[12] Alternative dataset splitting techniques, like the temporal split method utilized by Feinberg et al., aim to mitigate this by organizing data based on the publication order [12,13]. This method operates on the principle that the sequential evolution of substructural chemistries over time can provide a more realistic challenge when splitting data for model validation. However, assuming publication dates adhere to medicinal chemistry pedigree can introduce discrepancies, as these dates may not align with the actual order of discovery. While the intention behind the temporal split employed by Feinberg et al. is creative, *in vitro* results that would validate its impact on

real-world drug discovery remain unpublished. In response, Model Medicines herein introduces a novel train-test splitting methodology for ChemPrint that aims to directly maximize dataset dissimilarity and model training challenge, paired with *in vitro* results that suggest a tangible advance in the development of drugs for challenging targets AXL and BRD4.

2. RESULTS

2.1 Data Preparation: AXL and BRD4

AXL and BRD4 Data Curation and Binarization

To discover potent, novel compounds that have high drug discovery potential, it is critical that training datasets are built-to-purpose. Using our GALILEO data acquisition and curation pipeline, we procured large, high-quality datasets (see methods) from primary source literature for our two oncology targets of interest, AXL and BRD4, which affect multiple cancer types, such as breast cancer, pancreatic cancer, and glioblastoma [14,15]. We sought to identify the most highly potent compounds, so we defined compounds with a bioactivity measurement below 1 μM as active and above as inactive. **Table 1** summarizes our cleaned data set and the binarization into active and inactive classes for each protein target.

Table 1: Active and Inactive counts for the AXL and BRD4 datasets.

Target	Active Compounds	Inactive Compounds	Total
AXL	1823	982	2805
BRD4	3668	1987	5655

AXL and BRD4 Train-Test Splits

We split our AXL and BRD4 data into training and testing sets with both a 80%-20% stratified random split and with our novel splitting methodology that utilizes t-distributed Stochastic Neighbor Embedding (t-SNE) (see methods) [16]. This t-SNE splitting protocol aims to increase the structural and pharmacophoric dissimilarity between training and testing sets to enable the model to generalize beyond seen chemistry. We then evaluated the respective datasets splits by calculating their average-maximum test-train Tanimoto coefficients [17, 18] to demonstrate the chemical dissimilarities between the two datasets. The results for each method are reported below:

Table 2: Comparison between stratified random and our t-SNE split for AXL and BRD4 data.

Target	Split Method	Train Actives	Train Inactives	Test Actives	Test Inactives	Test-Train Similarity
AXL	Stratified Random	1456	785	367	197	0.75
	t-SNE	931	454	892	528	**0.59**
BRD4	Stratified Random	2932	1589	736	398	0.80
	t-SNE	1639	1164	2029	823	**0.53**

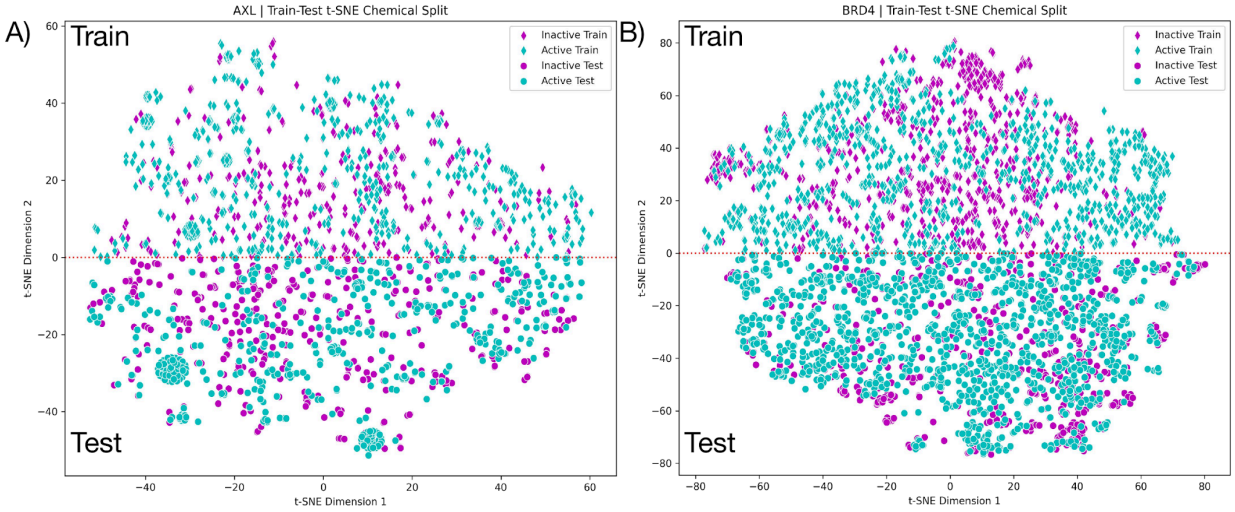

Figure 1: Visualization of our t-SNE train-test split for (**A**) AXL and (**B**) BRD4.

From the above results in **Table 2**, we see that the average-maximum Tanimoto similarity scores for AXL and BRD4 stratified random splits are 0.75 and 0.80, respectively. In comparison, the t-SNE splits report Tanimoto similarity scores of 0.59 and 0.53, respectively for AXL and BRD4. A comparison of the stratified random and our t-SNE method demonstrates a 27% increase in chemical similarity for AXL and 51% increase in chemical similarity for BRD4, for the stratified random method relative to the t-SNE method reported here. As the average-maximum Tanimoto similarity score is reduced, the dissimilarity between the test-train split increases. From this quantitative analysis, we observe that t-SNE data splitting provides better chemical separation between train and test data for both AXL and BRD4. Additionally, **Figure 1** gives us a qualitative insight into the functionality of our t-SNE splitting method. Clusters of compounds can be seen in the figure, suggesting that these molecules have similar structural and pharmacophoric properties. When the cutoff boundary is drawn, these similarly typed compounds are shown to be retained in either the train or test set, resulting in two chemically distinct data sets. In

principle, this t-SNE split provides the added challenge needed to produce a model capable of accurate extrapolation outside of train-similar compounds. We applied a more stringent 50-50% proportion of data within the t-SNE split train-test sets than the standard 80-20% train-test composition of stratified random splitting to imitate a data-limited scenario.

2.2 Performance Evaluation: ChemPrint vs. Baseline Models

We performed a comparative case study analyzing the performance of our zero-shot, Mol-GDL model, ChemPrint and five baseline models: logistic regression (LR), random forest (RF), k-nearest neighbors (KNN), support vector machine (SVM), and extreme gradient boosting (XGB) [19,20]. To further validate our t-SNE data splitting method on AXL and BRD4 datasets, we assessed the performance of the six different models with both our t-SNE split and the stratified random split. Each model architecture and splitting technique pair was trained and tested ten times, with the area under the receiver operating characteristic (AUROC)[21] value recorded for each replicate. The averages of these AUROC values are reported in **Table 3** and the single best model run is reported as an ROC curve in **Figure 2**.

Table 3: Average AUROC ± 95% confidence interval scores for all six evaluated models using both the stratified random split and our t-SNE split on AXL and BRD4 data. Statistics are averaged over ten distinct model trainings.

	AXL		**BRD4**	
	t-SNE	Stratified Random	t-SNE	Stratified Random
ChemPrint	**0.897 ± 0.002**	0.934 ± 0.001	**0.876 ± 0.003**	0.948 ± 0.001
SVM	0.885 ± 0.000	0.949 ± 0.000	0.852 ± 0.000	0.945 ± 0.000
RF	0.879 ± 0.003	0.961 ± 0.001	0.842 ± 0.004	0.951 ± 0.001
LR	0.871 ± 0.000	0.957 ± 0.000	0.814 ± 0.000	0.941 ± 0.000
XGB	0.872 ± 0.000	0.938 ± 0.000	0.806 ± 0.000	0.929 ± 0.000
KNN	0.826 ± 0.000	0.930 ± 0.000	0.801 ± 0.000	0.936 ± 0.000
****Average of Each Split****	0.872	0.945	0.832	0.942

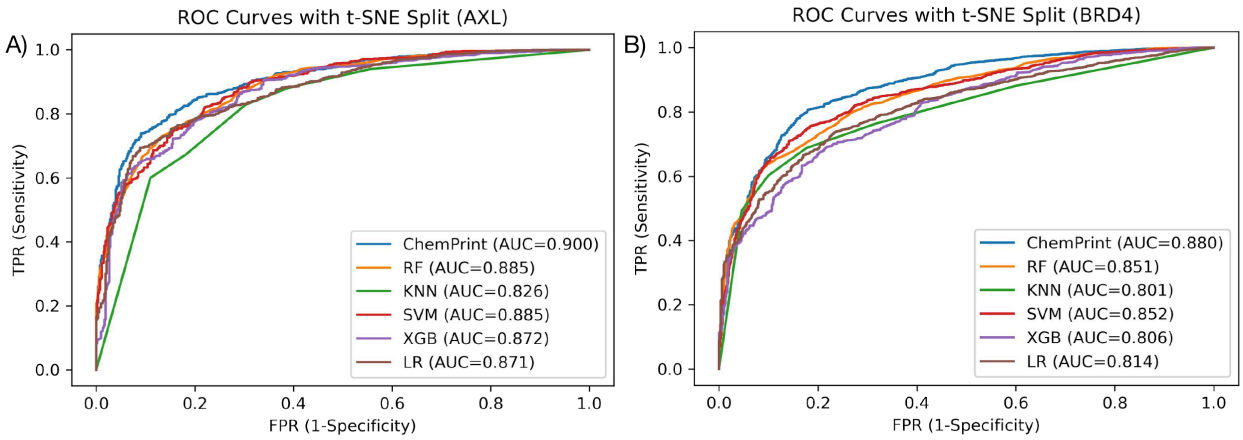

Figure 2: Reported ROC curves and associated AUROC values for the best performing model across all replicates for (**A**) AXL and (**B**) BRD4. Metrics are generated from our t-SNE test split.

The AXL and BRD4 ChemPrint models with t-SNE splitting demonstrate the highest performance based on both average (**Table 3**) and best AUROC scores (**Figure 2**). With the t-SNE split, ChemPrint has an average AUROC score of 0.897 for AXL, whereas baseline model scores range from 0.826 to 0.885. ChemPrint also outperforms all baseline models for BRD4 using the t-SNE split with an average AUROC of 0.876 compared to the baseline models that range from 0.801 to 0.852. The best ChemPrint iterations report higher AUROC scores than each baseline model across both proteins for the t-SNE split, with score improvements of up to ~10%, as seen in **Figure 2**.

On average across the six model architectures, the AUROC scores for t-SNE splits are ~9% lower than random stratified split scores. For AXL models, switching from stratified random splitting to t-SNE splitting results in a lower mean AUROC score across the six architectures from 0.945 to 0.872. BRD4 models behave similarly, decreasing from 0.942 to 0.832. Combining these data with average-maximum Tanimoto similarity scores reported above in **Table 2**, we see that a 27% increase in test-train similarity led to ~7% increase in AUROC for AXL and a 51% increase in test-train similarity led to ~11% increase in AUROC for BRD4, for the stratified random method relative to the t-SNE method. Thus, we can see that as similarity increases, AUROC scores increase. These results indicate that t-SNE splitting increases the dissimilarity between training and testing sets and poses a more rigorous challenge to the models. In doing so, the t-SNE split demonstrates its utility in creating a more rigorous test set. Moreover, the empirical evidence presented in **Table 3** demonstrates that the average-maximum Tanimoto similarity score calculated in **Table 2** correlates well with test set difficulty.

Contrary to the stratified random split's inflated model performance[12], the t-SNE split provides a more accurate reflection of model performance on novel chemistries due to the chemical dissimilarity between its sets. The use of our t-SNE split allows for the proper evaluation of model performance in chemical space that exists beyond the data it was trained on, demonstrating extrapolative abilities. The t-SNE data split serves as a pressure test for these models. While model architecture is often selected and benchmarked on validation metrics alone, we show that AUROC scores are dependent upon splitting technique. It becomes clear that employing only the stratified random split confounds model performance and fails to provide granular differentiation between model ability, especially for unseen chemical space. As seen in **Table 3**, if the “best” performing model were selected through metric

evaluation on the stratified random split, RF would be nominated as the best model. However, as demonstrated by the t-SNE split, both ChemPrint and SVM perform better when trained and tested on dissimilar chemistry, showcasing their ability to generalize beyond their training data.

Incorporating the t-SNE data split allows for the proper evaluation and optimization of ChemPrint, thereby enabling it to excel beyond the data it was initially trained on. ChemPrint has the best AUROC score of all models when tested on unseen chemistry, further demonstrating its potential use in zero-shot applications. This allows for the efficient discovery of active compounds with drug-like characteristics from novel chemical space in prospective hit-finding studies. This success is likely due to the Mol-GDL architecture at the core of ChemPrint, which leverages adaptive molecular embeddings rather than conventional models, which rely on static fingerprints, limiting their extrapolative potential.

2.3 Adaptive Molecular Embeddings: Quality Assessment

To elucidate the underlying adaptive molecular embedding process of ChemPrint, we plotted the t-SNE of model latent embeddings at each epoch to observe what changes were taking place. For the AXL ChemPrint model, **Figure 3** displays the evolution of chemical space representations of our adaptive molecular embeddings visualized through t-SNE for the AXL discovery program across 1000 epochs.

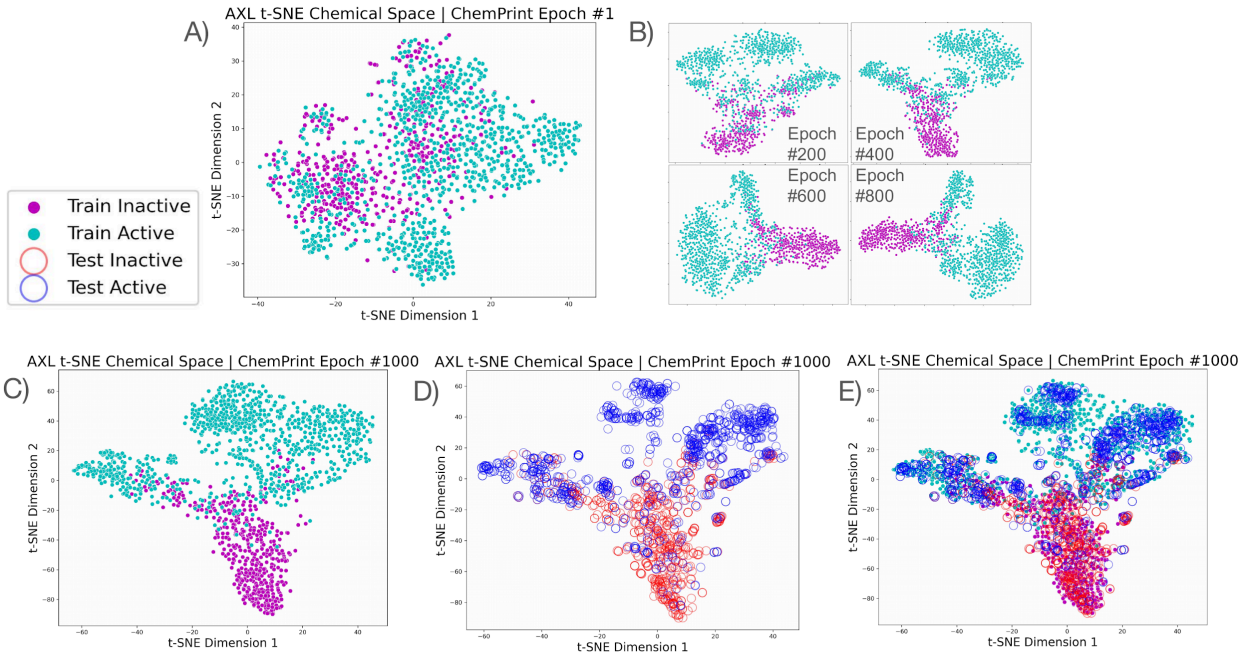

Figure 3: t-SNE visualizations of model latent embeddings of AXL t-SNE split data during ChemPrint's training cycles. Displayed are the embeddings at every 200th epoch, starting at (**A**) epoch 1 (initialization), (**B**) during training, (**C**) epoch 1000 (training), (**D**) epoch 1000 (inference), and (**E**) epoch 1000 (training and inference). Each point is representative of a molecule. (**A-C**) contain active (teal) and inactive (magenta) training compounds, while (**D-E**) contain active (blue) and inactive (red) inference/test compounds. Larger and fully labeled versions of (**B**) can be found in **Supplemental Figure S1**.

The t-SNE visualizations demonstrate that at training initiation (epoch 0), activity classes are largely intersecting embeddings, with no obvious pattern evident (**Figure 3A**). This suggests that no chemical or

structural information has yet been learned; there is no separability between active and inactive compounds and clear relationships between compounds are not present. However, as training epochs progress (**Figure 3B**), the latent representations generated by ChemPrint's convolutional layers demonstrate an evolution in embedded learning on molecular structure and activity. A clear separation of active compounds and inactive compounds is observed in later epochs. Specifically, there is prominent clustering of active and inactive compounds, with each cluster bound by structural and chemical similarities. With few errors in assignments, the active and inactive clusters also serve as powerful predictors of activity for unseen compounds in the respective ligand spaces by assigning the label of the nearest cluster to the point in consideration (**Figure 3C-E**). From this analysis, it is clear that ChemPrint has learned how to create powerful maps of chemical space relative to compounds capable of binding to and having activity against AXL.

Furthermore, **Figure 3** demonstrates the human interpretability of ChemPrint and its learning process. This allows drug developers to understand model performance in a sophisticated paradigm that moves beyond metrics like AUROC and similar. This is a quality that is unique to ChemPrint versus the comparative baseline models, as they learn from static embeddings. This quality of ChemPrint allows for Human-In-The-Loop (HITL) interpretation and Informed Inductive Bias (IIB) and removes reliability concerns associated with alternative black-box AI drug discovery approaches.

2.4 Compound Nomination: ChemPrint Inference Predictions

To test ChemPrint's ability when allowed to train on the full complement of curated AXL and BRD4 datasets, we deployed an ensemble of 100 independently trained ChemPrint models. These production ensembles were created for each target, AXL and BRD4, and evaluated with each across our known chemical entity (KCE) inference library containing ~17,000 compounds. These compounds were selected for having in-human safety and tolerability data, typically arising from Phase 1 and 2 clinical trials. All compounds from our KCE library with predicted activity against AXL and BRD4 were priority ranked and then subsequently screened by our Freedom To Operate (FTO) pipeline (see methods) for novelty to target, novelty to indication, and purchasability. These filters resulted in 44 compounds that were then nominated, purchased, and evaluated *in vitro* against their respective targets.

2.5 Compound Validation: *In Vitro*

The 44 molecules predicted by ChemPrint were tested *in vitro* for enzymatic inhibition via an AXL kinase activity assay and a BRD4 (BD1 + BD2) bromodomain activity assay (see methods). We observed that 20 of the 44 predicted molecules displayed inhibition at <20 μM resulting in a 45.5% hit rate (**Table 4**).

Table 4: Contains IC50 values of compound inhibitory effects on AXL and BRD4 activities.

Target	Drug Name	IC50 (μM)
AXL	Compound 1	0.21
AXL	Compound 2	0.26
AXL	Compound 3	1.9
AXL	Compound 4	4.6

AXL	Compound 5	4.8
AXL	Compound 6	5.3
AXL	Compound 7	5.6
AXL	Compound 8	6.1
AXL	Compound 9	~20
AXL	Compound 10	17% Inhibition @ 10 μM
AXL	Compound 11	27% Inhibition @ 20 μM
AXL	Compound 12	41% Inhibition @ 20 μM
BRD4	Compound 13	0.57
BRD4	Compound 14	16.8
BRD4	Compound 15	~20
BRD4	Compound 16	36% Inhibition @ 20 μM
BRD4	Compound 17	24% Inhibition @ 20 μM
BRD4	Compound 18	22% Inhibition @ 20 μM
BRD4	Compound 19	30% Inhibition @ 20 μM
BRD4	Compound 20	28% Inhibition @ 20 μM

To quantify the chemical novelty of these findings, we conducted a similarity analysis between our validated hits and the training data used to discover them. Using ECFP-4 fingerprints, we conducted a similarity search, recording the maximum Tanimoto similarity between each *in vitro* hit among both active and inactive training set compounds (**Figure 4**).

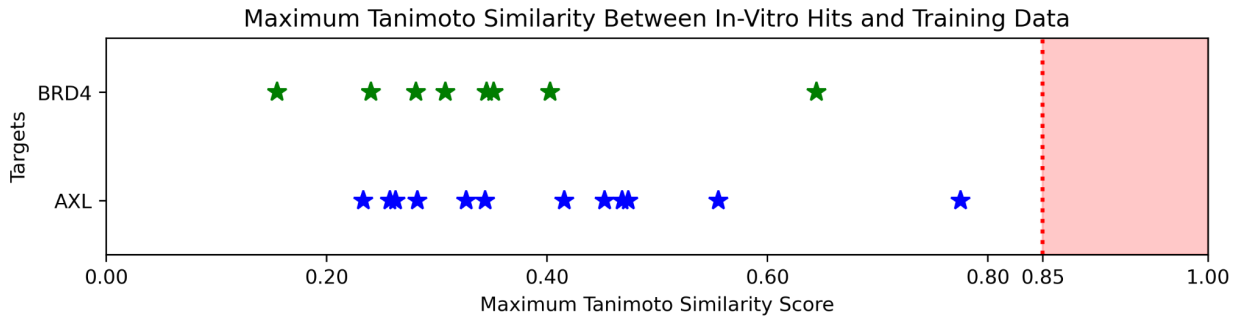

Figure 4: Plot of our *in vitro* validated hits for both AXL and BRD4 displaying their maximum Tanimoto similarity score to respective training datasets used to discover them.

The chemical dissimilarity of our lab validated discoveries is demonstrated in **Figure 4**, where the Tanimoto similarity scores range from 0.16 to 0.78 with a mean of 0.40 for AXL and 0.34 for BRD4. A Tanimoto similarity score above 0.85 is a common industry standard threshold for obviousness of similarity, beyond which two small molecules are expected to exhibit similar bioactivities [22, 23]. The principle of molecular similarity suggests that molecules similar to potent ligands are likely to also be potent, while those similar to inactive ligands are likely to be inactive as well. Conversely, predicting the

efficacy of a molecule with no close resemblance to any previously tested compound remains a challenge in the field of computer-aided drug design [7, 24]. A simple Tanimoto similarity search of our inference set against known active compounds would have failed to identify these hits. The chemical dissimilarity of our validated discoveries highlights the finely tuned nature of our GALILEO AI drug discovery platform and the zero-shot capability of ChemPrint models. Together, they are able to make novel discoveries beyond the capabilities of traditional techniques and standard industry approaches.

3. DISCUSSION

The results from our study highlight the substantial advancements in drug discovery facilitated by our GALILEO AI drug discovery platform and its zero-shot Mol-GDL model, ChemPrint. By utilizing prospective benchmarking and employing adaptive molecular embeddings, our methodology does not merely enhance predictive accuracy; it also facilitates a more nuanced comprehension of the molecular interactions that are fundamental to identifying viable therapeutic agents. Model Medicines has harnessed the power of AI to achieve an exceptional 45.5% hit rate in its oncology program targeting AXL and BRD4, a 75% increase over average rates reported by industry leaders [3].

ChemPrint stands out by employing Mol-GDL, an advancement that ensures the preservation of vital chemical and structural information throughout the AI discovery process. Our approach makes use of rigorous, prospective validation of models on the most difficult data sets that solve for real-world therapeutic discovery, rather than vanity technical metrics. During ChemPrint's training and validation process, we employ optimization techniques to reduce overfitting that create highly specialized adaptive molecular embeddings, affirming its effectiveness in prospective drug discovery studies. This advancement is critical to making discoveries that can reach the clinic, especially considering conventional drug discovery methods' high failure rates and resource intensiveness.

4. CONCLUSION

Our study introduces GALILEO's unique approach, combining prospective data segmentation and ChemPrint’s adaptive molecular embeddings. Our technology stack has been validated through our oncology program targeting AXL and BRD4. Our findings underscore the model's potential to significantly accelerate the drug discovery process, highlighting the critical role of innovative AI technologies in advancing pharmaceutical research. GALILEO and ChemPrint effectively bridge the gap between the promised efficiency of AI and practical therapeutic discovery. Our platform, as presented, addresses some of the most difficult challenges in AI-driven drug discovery, offering a promising path forward in the pursuit of more efficient and effective drug development processes that will ultimately save lives.

5. METHODS

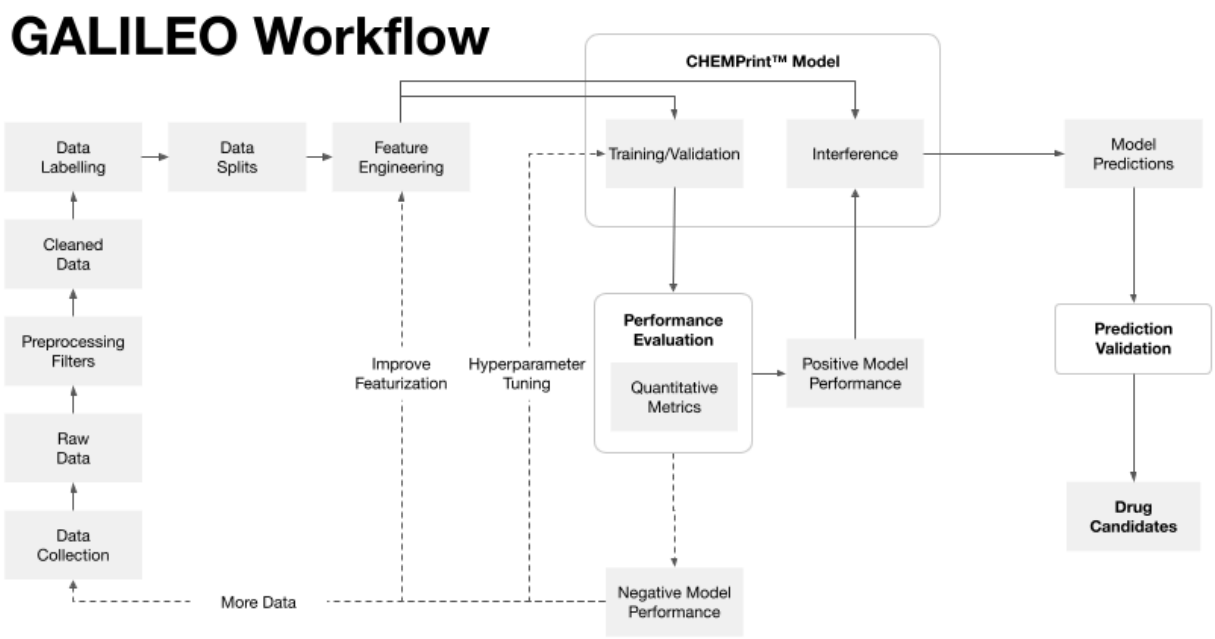

Figure 5: Workflow of ChemPrint within our GALILEO AI drug discovery pipeline as detailed in Methods.

5.1 Data Sets and Preparation for Prospective Discovery

Data Curation

Our GALILEO drug discovery pipeline was used to curate protein-ligand interaction data for AXL and BRD4. GALILEO identified primary source literature (**available in supporting documents**), extracted protein-specific QSAR data that was checked by our experts and filtered for undesired characteristics, such as Pan Assay Interference Compounds (PAINS) [25], molecular weight extrema, and duplicate compounds. Based on the distribution of activity values in the training sets, a potency cutoff of 1 μM was applied to separate active from inactive compounds in each set. This cutoff created a balance between active and inactive compounds and minimized data exclusion.

Train-Test Split

To split our data into training and test sets, we utilized t-SNE to project ECFP-4 encoded molecule data into 2-dimensions. The dimensionally reduced data preserves similarity relationships between points by minimizing the divergence between probability distributions of pairwise similarity from the original space and new embedding space. We used the line y = 0 to separate training (above) and test (below) data. This decision boundary creates chemically distinct training data and testing data, while also creating our desired ~50-50% (within 2%) train-test split that maintains class balance. Code Availability: GiHub repository for the t-SNE splitting method is archived (here).

Split Evaluation

The Tanimoto coefficient [15] is a common metric to evaluate the chemical similarity between molecules (represented as bit-vectors), and is defined as:

$$T(m_1, m_2) = \frac{|m_1 \wedge m_2|}{|m_1 \vee m_2|}$$

where $|m_1 \wedge m_2|$ is the number of common 1's in the bit-vectors and $|m_1 \vee m_2|$ is the number of indices where either bit-vector has a 1. This value ranges from 0 to 1, with a value of 0 indicating completely distinct representations and 1 indicating identical representations. Using this metric, we can approximate the chemical diversity of the dataset split via the average of the maximum Tanimoto coefficient of each test molecule when measured against all training points, i.e.:

$$\frac{1}{|\mathcal{D}_{test}|} \sum_{m_{test} \in \mathcal{D}_{test}} \max\{T(m_{test}, m_{train}) : m_{train} \in \mathcal{D}_{train}\}$$

5.2 ChemPrint Architecture

ChemPrint leverages mol-GDL to learn the adaptive embeddings derived from its training data to make predictions when given novel inference sets. This model architecture effectively captures the intricate relationships between training features and proficiently generalizes to small molecules it has not seen before. The input data takes the form of a geometric graph, a molecular representation that encapsulates the structural information of each datapoint, to which one-hot encoded atoms types are passed in as features. This graph-based learning enables us to encode and leverage the inherent dependencies and relationships among the data, making it particularly well-suited for tasks involving complex structures.

Our model architecture encompasses an end-to-end Graph Convolutional Network (GCN) with a Multilayer Perceptron (MLP) module to facilitate positive and negative classification. Use of select normalization, activation, pooling, and dropout layers are used to construct ChemPrint. A generalized workflow of our model architecture is presented in **Figure 6**. This combination of networks stabilizes the model and the parameters are fine-tuned to avoid overfitting. Thus, enabling our ChemPrint module to effectively capture molecular features to make accurate predictions.

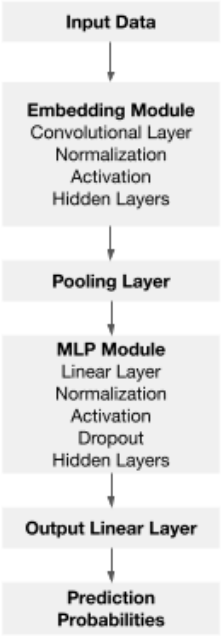

Figure 6: Illustration of ChemPrint's generalized architecture.

5.3 Baseline Models

ChemPrint was benchmarked against a selection of widely utilized machine learning algorithms. In order for a molecule to be evaluated by a classifier, it must first be processed into a vectorized numerical representation, or embedding that preserves as much structural and chemical information as possible. We utilized the general hashed radial fingerprint, ECFP-4 (radius of 2) encoded in 2048 bits to preprocess the data used to train and test the baseline models. Logistic regression (LR), random forest (RF), k-nearest neighbors (KNN), and support vector machine (SVM) were implemented using the scikit-learn library [19]. The extreme gradient boosting (XGB) was built using the XGBoost library [20].

5.4 Performance Evaluation

Area Under the Receiver Operating Characteristics (AUROC)

The Receiver Operating Characteristic (ROC) curve is a plot of a classifier's true positive rate (sensitivity) versus its false positive rate (1-specificity) as the classification threshold is varied from 0 to 1. The Area Under the Receiver Operating Characteristics (AUROC) curve [19] establishes a baseline confidence in the model performance through statistical relationships between the model's predicted probabilities of activity and true binding status. Setting a threshold at 0 results in all the molecules being classified as active, leading to both true positive and false positive rates of 1. Conversely, setting a threshold of 1 results in all molecules being classified as inactive, leading to both true and false positive rates of 0. AUROC reflects the likelihood of the classifier assigning a higher probability of a positive example (active molecule) compared to a negative example (inactive molecule). A perfect classifier will have an AUROC of 1.0 while a random classifier will have an AUROC of 0.5. Unlike accuracy, AUROC is useful regardless

of data imbalance between classes, making it one of the most common metrics used to benchmark machine learning models. AUROC was calculated using the scikit-learn library[19] for each replicate (n=10) of every model-splitting method pair.

Molecular Embeddings

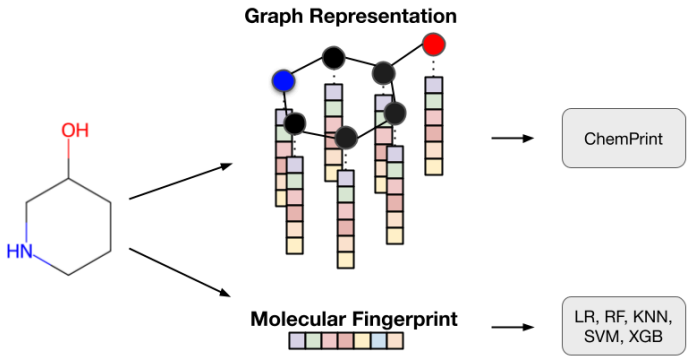

Figure 7: Two methods of molecular embedding: graph representation and molecular fingerprinting, i.e. ECFP. A GCN can operate directly on molecular structure with a graph representation whereas the standard fingerprint is a static embedding that is bounded by a predetermined encoding.

ChemPrint learns optimal, real-valued, differentiable, property-specific, molecular embeddings by leveraging its GCN instead of the non adaptive ECFP fingerprinting method (used by the LR, RF, KNN, SVM, XGB models). GCN latent space molecular embeddings are engineered and optimized through end-to-end back propagation. In comparison to embeddings such as ECFP, ChemPrint's embeddings are adaptive and self optimizing for a specific molecular property. **Figure 7** depicts the graphical representation of the adaptive embedding used as input for our model as compared to the stagnant molecular fingerprint leveraged by baseline models.

For our analysis of ChemPrint's embedding space, we captured the model's learned embeddings at progressive epochs during the training process (see **Figure 3**). We then projected these embeddings down to 2 dimensions using t-SNE, and plotted all points colored according to activity and distinguished them as training or test points. For each plot, we noted the quality of clustering within each activity class and the level of separability between classes as strong indicators of embedding quality based on preservation of molecular property relationships.

5.5 ChemPrint Compound Nomination

Known Chemical Entity (KCE) Library

Our known chemical entity (KCE) inference set contains ~17,000 curated compounds that have been thoroughly characterized and have a documented history in clinical trials. For each compound in this inference set, pertinent information such as synonyms/identifiers, ADMET properties, clinical trial details, current approval statuses, and purchasability are recorded.

Freedom to Operate Pipeline (FTO)

As part of the GALILEO™ platform of AI drug discovery tools, each predicted active compound from our in-silico discovery pipeline is investigated for freedom to operate (FTO) using data miners and web search tools. Non-obviousness and novelty of the predicted compounds are verified using these tools.

The data miners make use of our exhaustive set of synonyms and identifiers for each compound to perform in depth target and indication searches. These searches are performed across the six databases ACM, Arxiv, Biorxiv, Medrxiv, Pubmed, and Scopus. Primary source material is returned when both the synonym/identifier and indication or synonym/identifier and target is found in the same publication. The identified literature is then subjected to our natural language processing (NLP) and large language model (LLM) packages to ensure that followup manual review relies on proper relevance and filtration of inconsequential co-mentions of compound and target in the literature. The manual verification process is conducted on all compounds flagged as potentially non-novel by our pipeline.

5.6 Enzyme Inhibition Assays

All *in vitro* experiments were conducted by BPS Bioscience Inc.

AXL Kinase Assay

The assay was performed using ADP-Glo Kinase assay reagents (Promega). It measures kinase activity by quantitating the ADP amount produced from the enzymatic reaction. The luminescent signal from the assay is correlated with the amount of ADP present and is directly correlated with the amount of kinase activity. The compounds were diluted in 10% DMSO and 2.5 µl of the dilution was added to a 25 µl reaction so that the final concentration of DMSO is 1% in all of reactions. All of the enzymatic reactions were conducted at 30 ºC for 45 minutes. The 25 µl reaction mixture contains 40 mM Tris, pH 7.4, 10 mM MgCl2, 0.1 mg/ml BSA, 1 mM DTT, 10 µM ATP, kinase substrate and the enzyme. After the enzymatic reaction, 25 µl of ADP-Glo reagent was added and incubated for 45 - 60 minutes at room temperature followed by another 45 mins incubation with 50 µl of kinase detection mixture. Luminescence signal was measured using a BioTek Synergy 2 microplate reader.

BRD4 (BD1 + BD2) Bromodomain Assay

The assay was performed by TR-FRET technology using recombinant bromodomains (BD1 and BD2) and the endogenous BET Ligand. The TR-FRET signal from the assay is correlated with the amount of Ligand binding to the bromodomain. The compounds were diluted in 100% DMSO. They were diluted 20-fold in 5% DMSO in Reaction Buffer and 2 µl of the dilution was added to a 20 µl reaction so that the final concentration of DMSO is 0.5% in all of reactions. The binding reaction was conducted at room temperature. The 20 µl reaction mixture in Assay Buffer contains the protein, the indicated amount of the inhibitor, acetylated BET ligand (or non-acetylated ligand for negative control wells), and the reaction dyes. The reaction mixture incubated for 120 min prior to reading the TR-FRET signal. Fluorescence signals for both the donor and acceptor dyes were measured using a Tecan Infinite at an excitation of 340 nm and emissions at 620 nm and 665 nm.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the support and valuable discussions from numerous consultants, advisors, partners and former employees. Special thanks are due to Dr. David Garvey and Dr. Tushar Menon for their consultation and thorough review of this manuscript. The authors acknowledge the contributions of ChatGPT 4 [26] in editing and refining this manuscript's language to improve coherence and fluency.

7. REFERENCES

1. D. Sun, W. Gao, H. Hu, and S. Zhou, "Why 90% of clinical drug development fails and how to improve it?," Acta Pharmaceutica Sinica B, vol. 12, no. 7. Elsevier BV, pp. 3049–3062, Jul. 2022. doi: 10.1016/j.apsb.2022.02.002.
2. A. Schuhmacher, M. Hinder, A. von Stegmann und Stein, D. Hartl, and O. Gassmann, "Analysis of pharma R&D productivity – a new perspective needed," Drug Discovery Today, vol. 28, no. 10. Elsevier BV, p. 103726, Oct. 2023. doi: 10.1016/j.drudis.2023.103726.
3. Schrödinger, Inc. (2023). *Dramatically improving hit rates with a modern virtual screening workflow*. Retrieved from https://newsite.schrodinger.com/life-science/learn/white-papers/dramatically-improving-hit-rates-modern-virtual-screening-workflow/
4. A. S. M. Z. Rahman et al., "A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery," PLOS Computational Biology, vol. 18, no. 10. Public Library of Science (PLoS), p. e1010613, Oct. 13, 2022. doi: 10.1371/journal.pcbi.1010613.
5. R. S. K. Vijayan, J. Kihlberg, J. B. Cross, and V. Poongavanam, "Enhancing preclinical drug discovery with artificial intelligence," Drug Discovery Today, vol. 27, no. 4. Elsevier BV, pp. 967–984, Apr. 2022. doi: 10.1016/j.drudis.2021.11.023.
6. J. M. Stokes et al., "A Deep Learning Approach to Antibiotic Discovery," Cell, vol. 180, no. 4. Elsevier BV, pp. 688-702.e13, Feb. 2020. doi: 10.1016/j.cell.2020.01.021.
7. Variational AI, & Rolfe, J. T. (2024, February 8). *Applicability domains are common in QSAR but irrelevant for conventional ML tasks.* Variational AI. Retrieved from https://variationalai.substack.com/p/applicability-domains-are-common
8. C. Shen, J. Luo, and K. Xia, "Molecular geometric deep learning," Cell Reports Methods, vol. 3, no. 11. Elsevier BV, p. 100621, Nov. 2023. doi: 10.1016/j.crmeth.2023.100621.
9. S. Bhalla and D. E. Gerber, "AXL Inhibitors: Status of Clinical Development," Current Oncology Reports, vol. 25, no. 5. Springer Science and Business Media LLC, pp. 521–529, Mar. 15, 2023. doi: 10.1007/s11912-023-01392-7.
10. M. M. Mita and A. C. Mita, "Bromodomain inhibitors a decade later: a promise unfulfilled?," British Journal of Cancer, vol. 123, no. 12. Springer Science and Business Media LLC, pp. 1713–1714, Sep. 29, 2020. doi: 10.1038/s41416-020-01079-x.
11. D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," Journal of Chemical Information and Modeling, vol. 50, no. 5. American Chemical Society (ACS), pp. 742–754, Apr. 28, 2010. doi:

10.1021/ci100050t.

12. E. N. Feinberg, E. Joshi, V. S. Pande, and A. C. Cheng, “Improvement in ADMET Prediction with Multitask Deep Featurization,” Journal of Medicinal Chemistry, vol. 63, no. 16. American Chemical Society (ACS), pp. 8835–8848, Apr. 14, 2020. doi: 10.1021/acs.jmedchem.9b02187.
13. E. N. Feinberg et al., “PotentialNet for Molecular Property Prediction,” ACS Central Science, vol. 4, no. 11. American Chemical Society (ACS), pp. 1520–1530, Nov. 02, 2018. doi: 10.1021/acscentsci.8b00507.
14. C. Zhu, Y. Wei, and X. Wei, “AXL receptor tyrosine kinase as a promising anti-cancer approach: functions, molecular mechanisms and clinical applications,” Molecular Cancer, vol. 18, no. 1. Springer Science and Business Media LLC, Nov. 04, 2019. doi: 10.1186/s12943-019-1090-3.
15. J. Hu, D. Pan, G. Li, K. Chen, and X. Hu, “Regulation of programmed cell death by Brd4,” Cell Death & Disease, vol. 13, no. 12. Springer Science and Business Media LLC, Dec. 20, 2022. doi: 10.1038/s41419-022-05505-1.
16. L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," in Journal of Machine Learning Research, vol. 9, no. Nov, pp. 2579-2605, 2008.
17. P. Willett, J. M. Barnard, and G. M. Downs, “Chemical Similarity Searching,” Journal of Chemical Information and Computer Sciences, vol. 38, no. 6. American Chemical Society (ACS), pp. 983–996, Jul. 21, 1998. doi: 10.1021/ci9800211.
18. D. Polykovskiy et al., “Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models,” Frontiers in Pharmacology, vol. 11. Frontiers Media SA, Dec. 18, 2020. doi: 10.3389/fphar.2020.565644.
19. Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), pp.2825–2830.
20. Chen, T. & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY, USA: ACM, pp. 785–794. Available at: http://doi.acm.org/10.1145/2939672.2939785.
21. A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” Pattern Recognition, vol. 30, no. 7. Elsevier BV, pp. 1145–1159, Jul. 1997. doi: 10.1016/s0031-3203(96)00142-2.
22. R. D. Brown and Y. C. Martin, “The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding,” Journal of Chemical Information and Computer Sciences, vol. 37, no. 1. American Chemical Society (ACS), pp. 1–9, Jan. 01, 1997. doi: 10.1021/ci960373c.
23. S. Jasial, Y. Hu, M. Vogt, and J. Bajorath, “Activity-relevant similarity values for fingerprints and implications for similarity searching,” F1000Research, vol. 5. F1000 Research Ltd, p. 591, Apr. 28, 2016. doi: 10.12688/f1000research.8357.2.
24. G. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, “Molecular Similarity in Medicinal Chemistry,” Journal of Medicinal Chemistry, vol. 57, no. 8. American Chemical Society (ACS), pp. 3186–3204, Nov. 11, 2013. doi: 10.1021/jm401411z.

25. J. B. Baell and G. A. Holloway, “New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays,” Journal of Medicinal Chemistry, vol. 53, no. 7. American Chemical Society (ACS), pp. 2719–2740, Apr. 08, 2010. doi: 10.1021/jm901137j.
26. OpenAI. "ChatGPT-4." OpenAI, 2023. [Online]. Available: https://openai.com/chatgpt.