JOURNAL OF
EMERGING INVESTIGATORS

Article

Validating DTAPs with large language models: A novel approach to drug repurposing

Ethan Curtis[1], Dean Curtis[1]
[1] St. Michaels University School, Victoria, British Columbia, Canada

SUMMARY

In the face of escalating costs and lengthy timelines associated with traditional drug discovery, our study introduces a novel approach aimed at enhancing the drug repurposing process through the integration of computational models. Specifically, we explore the potential to enhance drug target affinity predictors (DTAPs)—tools that predict how well a drug binds to its target—by integrating them with advanced large language models (LLMs), such as GPT-4 and Llama-2-70b. We hypothesized that this synergy between DTAPs and LLMs would significantly improve the accuracy of identifying suitable drug-target interactions, a crucial step in repurposing existing drugs for new medical uses. Employing a rigorous comparative analysis, we tested the efficacy of traditional DTAPs against a specialized dataset focused on psychotropic drugs and their interactions with the sigma-1 receptor, an area ripe with repurposing opportunities. We then assessed how the integration of these DTAPs with LLMs affected prediction accuracy. The results showed a marked improvement in binary prediction accuracy, especially when DTAPs were combined with GPT-4. The implications of our findings are significant, suggesting that the fusion of DTAPs with LLMs could revolutionize the process of drug repurposing. This integrated approach offers a faster, more cost-effective pathway to drug development, streamlining the identification of new therapeutic applications for existing drugs. Our study not only validates the hypothesis of enhanced performance through integration of LLMs with DTAPs but also sets the stage for a new era in pharmacology, where the combination of advanced AI techniques can lead to breakthroughs in treatment discovery.

INTRODUCTION

The field of drug discovery is rapidly evolving due to advancements in computational biology and the analysis of interactions between targets and drug compounds (1). The drug discovery process, crucial for developing new medications, is notably long and expensive, typically spanning over a decade and costing approximately $2.6 billion per successful drug approval (2, 3). Given the high cost and a 90% failure rate in clinical trials, drug repurposing has become an increasingly favored approach (4). This strategy repackages existing drugs for new therapeutic purposes, considerably reducing both the time and costs associated with drug development (5). Notably, 30–40% of drugs recently approved by the Food and Drug Administration originated from repurposing efforts, highlighting the significant role of repurposing in the pharmaceutical industry's shift towards more efficient drug development methods (5).

Despite its potential, modern drug repurposing is still a difficult process and challenged by the ever-growing catalog of repurposable drugs and thousands of articles and clinical trial results with contrasting conclusions (6). One possible remedy to these challenges is the introduction of deep learning techniques. Deep learning is a type of artificial intelligence (AI) characterized by its ability to learn from large datasets, recognize complex patterns, and make data-driven predictions that can be effectively integrated with drug repurposing to process, analyze, and contrast the vast array of structural and literary data available across the internet (7). Traditional deep-learning approaches in drug repurposing have mainly focused on drug-target interactions, using AI algorithms to predict the level of affinity a drug has for a target based on its structural patterns (7). These deep learning algorithms, known as Drug-Target Affinity Predictors (DTAPs), require manual validation in which researchers evaluate the performance of predictors by finding studies in which a relationship between the drug and target has already been established (8,9). This process is slow, inefficient, and a bottleneck in the deep learning drug candidate prediction process. (8)

Large language models (LLMs) are deep learning models at the forefront of AI innovation. These models possess the capacity to both analyze and generate human-like text (10). Although there have not yet been extensive efforts to harness LLMs within the drug-discovery domain specifically, similar endeavors exist within the broader category of scientific research. Projects like Coscientist, an AI system based on GPT-4, demonstrate the application of these models beyond conventional boundaries (10). Coscientist can design experimental structures using data from the internet and its own custom database, showcasing the versatility of LLMs in automating and innovating within scientific research (10).

Drug discovery is a complex and time-consuming process that could potentially benefit from the application of LLMs. One crucial aspect of drug discovery is predicting drug-target interactions, which is often done using DTAPs. These tools use various computational methods to estimate how strongly a drug molecule might bind to a specific target protein. Among these methods, Convolutional Neural Networks (CNNs), a class of deep learning algorithms particularly effective at pattern recognition, have shown promising results in modeling molecular interactions (29). Based on this success, we hypothesized that a model combining two CNNs would outperform other architectures, including those using Morgan fingerprinting (Morgan) and Amino Acid Composition (AAC)

https://doi.org/10.59720/24-103

algorithms, in predicting drug-target binding affinity.

In this study, we aim to enhance drug discovery by integrating LLMs with DTAPs to improve the accuracy and efficiency of drug-target interaction predictions. To demonstrate our approach, we focused on the sigma-1 receptor as our initial target. This receptor has gained significant attention in drug research due to its role in neuroprotection and memory enhancement (23). Recent studies have elucidated its function as a chaperone protein involved in signal transduction, neurite outgrowth, and neuroplasticity, as well as its ability to promote autophagy—a critical cellular process for clearing damaged components (23). This multifaceted role makes the sigma-1 receptor a promising target for neuroprotective therapies. Drugs that target this receptor are being explored for their potential to treat neurological disorders such as Alzheimer's disease, depression, and schizophrenia, making it a key focus for drug repurposing efforts (23). We specifically focused on determining high-affinity sigma-1 receptor ligands, as these compounds show the most promise for therapeutic applications.

Inspired by advancements in LLM applications, our study aimed to determine whether LLMs can be effectively used as "robot researchers" to streamline the drug discovery process, particularly in the context of the sigma-1 receptor. We propose a novel approach that combines the strengths of LLMs with DTAPs to enhance the accuracy and efficiency of drug-target interaction predictions. Our research explores the potential of integrating LLMs with DTAPs to not only augment the accuracy of drug repurposing predictions but also optimize research efficiency by reducing manual verification requirements. We hypothesized that this combined approach would improve both the practicality and objective accuracy of drug-target affinity predictions. Our findings demonstrate that integrating LLMs with DTAPs significantly enhanced prediction accuracy, with our combined GPT-4 and Morgan-AAC model achieving an F1 score of 0.9474 in binary classification tasks, while also providing valuable confidence metrics for potential drug candidates.

RESULTS

In this study, we focused on analyzing drug affinity to the sigma-1 receptor using IC_{50} (half-maximal inhibitory concentration) values. The IC_{50} is a measure of a drug's potency, representing the concentration of a compound needed to inhibit a biological process by half (27). We tested three standalone DTAPs: a model combining two CNNs, a model combining Morgan with AAC, and a model combining Daylight fingerprints (Daylight) with AAC. These models offered complementary approaches to molecular representation: direct structural learning through CNNs, circular fingerprint analysis through Morgan and linear fingerprint analysis through Daylight. We evaluated them separately from our LLM-augmented models using two performance outcomes: Exact affinity (IC_{50} value) prediction and relative (binary) affinity prediction. To test these outcomes, we used two separate datasets, the first of which was comprised of 46 drugs, each with a known exact sigma-1 IC_{50} value. We used this dataset to test the performance of our models in predicting exact affinity. The second dataset was composed of 32 drugs, 19 with high sigma-1 affinity and 13 with low or no sigma-1 affinity. We used this dataset to evaluate our models in predicting relative binary affinity.

Drug Name	IC50 Score	Pred IC50
A: CNN + CNN		
Benztropine	65	62.35
Thioridazine	286	82.46
Lomerizine	37	86.37
Thiothixene	353	116.74
Trifluoperazine	125	118.84
Flunarizine	28	132.03
Sertraline	260	137.43
Haloperidol	73	138.20
Bepridil	365	141.50
Clomiphene	195	142.85
B: Daylight + AAC		
Indatraline	737	7.69
Siramesine	17	14.00
Butaclamol	343	15.57
Perphenazine	104	20.55
Prochlorperazine	232	21.34
Sertraline	260	27.70
Quinidine	480	28.18
Fluspirilene	380	29.47
Flupentixol	70	29.95
Thiothixene	353	30.75
C: Morgan + AAC		
Indatraline	737	2.56
Butaclamol	343	3.91
Astemizole	43	46.99
Haloperidol	73	54.00
Siramesine	17	77.84
Thiothixene	353	84.77
Lomerizine	37	86.37
Protriptyline	307	99.08
Thioridazine	286	113.16
Ifenprodil	5.5	113.51

Table 1: Comparative Analysis of 10 Lowest Predicted IC_{50} Psychotropic Drugs Across Three Models. Actual and predicted IC_{50} values for the ten drugs with the lowest predicted IC50 scores using **A)** a dual convolutional neural network (CNN + CNN) approach, **B)** a combination of Daylight fingerprinting and amino acid composition (Daylight + AAC) methods, and **C)** the Morgan algorithm with amino acid composition (Morgan + AAC) analysis. Each section lists drug names, experimental IC_{50} scores, and model-predicted IC_{50} values, selected from a dataset of 46 psychotropic drugs.

https://doi.org/10.59720/24-103

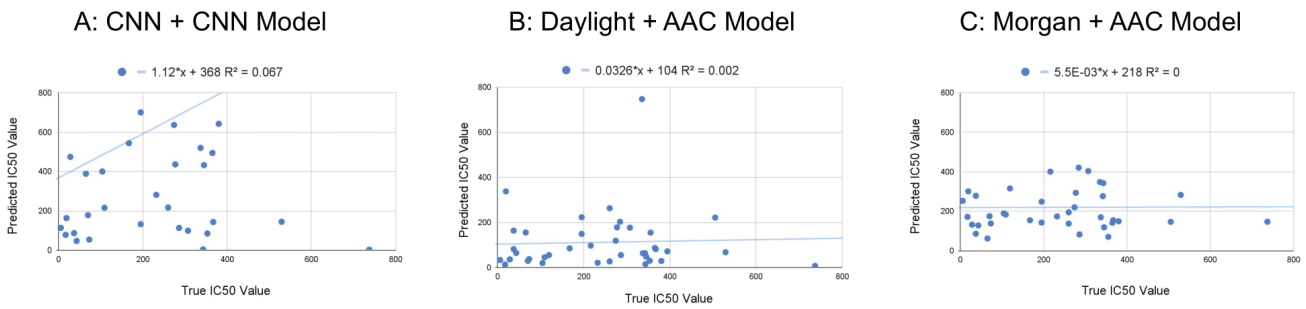


Figure 1: Comparative Analysis of IC_{50} Prediction Models for Psychotropic Drugs. Correlation between actual and predicted IC_{50} values using **A)** a dual convolutional neural network approach, **B)** a combination of Daylight fingerprinting and amino acid composition methods, and **C)** the Morgan algorithm alongside amino acid composition analysis. Each section represents a distinct computational strategy's predictions against the true values, across a dataset of 46 psychotropic drugs.

Prior to testing our standalone models, we trained each of the DTAPs on the BindingDB drug database (20, 24). We split the dataset into three sections: a training set consisting of 70% of the total database, a validation set consisting of 10% of the total database, and a test set consisting of 20% of the total database, so we could evaluate and reduce errors across each epoch. In total, we trained the dataset over 200 epochs to minimize loss value. First, we evaluated the standalone DTAPs' ability to predict exact IC_{50} values. This testing was crucial for establishing thresholds that would later be used in binary classification, as precise numerical predictions helped us determine meaningful cutoff points between high and low affinity compounds. For our binary affinity prediction testing, we began by evaluating the standalone DTAPs. Due to the imbalance in our dataset (19 high affinity vs. 13 low affinity drugs), we opted to use F1 scores instead of accuracy percentages for our evaluation metrics. F1 scores provide a more balanced measure of performance for imbalanced datasets by considering both precision and recall, thus avoiding potential bias towards the majority class that can occur with simple accuracy calculations (34).

For the LLM component of our study, we tested two different models: GPT-4 and llama-2-70b. We chose to evaluate only binary affinity prediction performance with these LLMs, as their architecture is not suited for extracting precise numerical predictions like IC_{50} values from text (35). For each LLM, we knowledge-embedded a database consisting of 50 articles related to sigma-1 ligands. We then prompt-engineered each model and provided it with the same list of 32 drugs used in our DTAP evaluations to determine binary affinity classifications for each compound. We processed the LLMs' natural language outputs into binary classifications and calculated F1 scores using the same methodology applied to the DTAPs, enabling direct performance comparisons across all models using a standardized metric. To leverage both prediction methods, we combined outputs from the DTAPs and LLMs through a logical framework that generated an integrated verdict with a corresponding confidence assessment.

The analysis of the three standalone DTAPs' performance at predicting the exact IC_{50} values for psychotropic drug affinity towards the sigma-1 receptor revealed notable insights into the capabilities and limitations of current computational approaches (**Table 1**). The model combining Morgan and AAC analysis emerged as the most effective, with a mean standard error (MSE) of 59.167 μM, likely due to its proficient use of circular fingerprints to capture the nuanced molecular structures critical for understanding drug-target interactions (**Figure 1C**). Contrary to our initial hypothesis, the model combining two CNNs did not perform as well, with an MSE of 632.696 μM, suggesting that the direct application of CNNs, renowned for their success in image recognition, might not translate seamlessly to the intricate field of drug-target affinity without further domain-specific adaptations (**Figure 1A**). The Daylight and AAC model's significantly poorer performance, with an MSE of 79.093 μM, underscores the challenges faced by descriptor-based models in capturing the complex interactions specific to psychotropic drugs and the sigma-1 receptor (**Figure 1B**).

Following our evaluation of exact affinity prediction, we assessed the three models' performance in binary affinity prediction. This approach simplifies the prediction task by classifying drugs as either high or low affinity, rather than predicting precise IC_{50} values. To achieve this, we established a threshold IC_{50} value of 750 nM. This threshold served to dichotomize our dataset: high affinity compounds (IC_{50} < 750 nM) were considered to have potential therapeutic benefit, while low affinity compounds (IC_{50} > 750 nM) were likely to be less effective for our target. We chose the 750 nM threshold based on observed patterns in our models' exact affinity predictions. Our models consistently predicted IC_{50} values below this 750 nM mark for established agonists (compounds known to bind to and activate the receptor), indicating that this threshold may effectively separate compounds with significant binding potential from those with weaker interactions. This binary classification approach offers several advantages. It mitigates the impact of high variability in exact IC_{50} predictions, aligns well with the practical needs of early-stage drug discovery where the primary goal is often to identify promising candidates rather than determine precise binding affinities, and potentially enhances the overall accuracy and relevance of our predictions by focusing on a more robust binary outcome.

Our dataset consisted of 32 drugs, comprising 19 drugs with high sigma-1 affinity and 13 drugs with low sigma-1 affinity, using our established threshold of 750 nM to differentiate between high and low affinity compounds (**Table 2-4**). Among the evaluated models, the one integrating Morgan with AAC emerged as the most accurate, achieving an F1 score of 0.9231, followed by the CNN-based model with an F1 score of 0.8108 and the model combining Daylight and

https://doi.org/10.59720/24-103

Drug Name	Predicted IC50 Score (nM)	True Affinity	Predicted Affinity
Escitalopram	109.16	High	High
Panamesine	110.69	High	High
Furosemide	111.80	Low	High
Doxycycline	118.02	High	High
Trifluoperazine	118.84	High	High
Triflupromazine	126.28	High	High
Sertraline	137.43	High	High
Haloperidol	138.20	High	High
Carbamazepine	142.53	Low	High
Fluspirilene	149.37	High	High
Ondansetron	149.56	High	High
Fluoxetine	152.00	High	High
Chlorpromazine	153.57	High	High
Psilocybin	155.90	Low	High
Oxazepam	164.14	Low	High
Pimozide	169.47	High	High
Flupentixol	175.00	High	High
Cetirizine	177.53	Low	High
Fluphenazine	182.94	High	High
Clomipramine	248.19	High	High
Ifenprodil	252.53	High	High
Imipramine	282.66	High	High
Butaclamol	342.15	High	High
Amitriptyline	400.03	High	High
Aspirin	444.14	Low	High
Gabapentin	490.61	Low	High
Ibuprofen	757.07	Low	Low
Allopurinol	828.02	Low	Low
Lacosamide	966.08	Low	Low
Diphenhydramine	1081.14	Low	Low
Metoclopramide	1266.44	Low	Low
Acetaminophen	1451.45	Low	Low

Table 2: Binary Affinity Prediction for High Affinity and Low Affinity Compounds Using CCN + CCN Model. Prediction results for 32 drugs (19 high sigma-1 affinity drugs, 13 low affinity sigma-1 drugs) based on IC_{50} threshold values (high affinity < 750 nM, low affinity > 750 nM) using CNN + CNN model (MSE = 632.696 μM. All models were trained on the BindingDB database over 200 epochs with a 70% training, 20% validation, and 10% testing split.

Drug Name	Predicted IC50 Score (nM)	True Affinity	Predicted Affinity
Butaclamol	3.91	High	High
Panamesine	11.01	High	High
Haloperidol	54.00	High	High
Ifenprodil	113.51	High	High
Triflupromazine	121.07	High	High
Clomipramine	131.91	High	High
Imipramine	144.29	High	High
Escitalopram	151.22	High	High
Chlorpromazine	177.16	High	High
Flupentixol	178.08	High	High
Fluoxetine	209.16	High	High
Fluphenazine	215.88	High	High
Sertraline	216.45	High	High
Trifluoperazine	432.63	High	High
Ibuprofen	436.67	Low	High
Aspirin	504.14	Low	High
Pimozide	519.83	High	High
Fluspirilene	642.21	Low	High
Metoclopramide	789.14	Low	Low
Furosemide	842.45	Low	Low
Ondansetron	883.93	Low	Low
Doxycycline	1221.50	Low	Low
Diphenhydramine	1383.63	Low	Low
Allopurinol	1594.32	Low	Low
Oxazepam	1596.02	Low	Low
Amitriptyline	1612.40	Low	Low
Cetirizine	1628.15	Low	Low
Psilocybin	1699.63	Low	Low
Gabapentin	1741.44	Low	Low
Lacosamide	1971.79	Low	Low
Carbamazepine	2018.97	Low	Low
Acetaminophen	3175.64	Low	Low

Table 3: Binary Affinity Prediction for High Affinity and Low Affinity Compounds Using the Morgan + AAC Model. Prediction results for 32 drugs (19 high sigma-1 affinity drugs, 13 low affinity sigma-1 drugs) based on IC_{50} threshold values (high affinity < 750 nM, low affinity > 750 nM) using Morgan + AAC model (MSE = 59.167 μM). All models were trained on the BindingDB database over 200 epochs with a 70% training, 20% validation, and 10% testing split.

AAC with an F1 score of 0.6857 (**Figure 2A**).

Following our evaluation of the DTAP models, we turned our attention to the performance of LLMs in predicting binary affinity for sigma-1 receptor ligands. We focused on two state-of-the-art LLMs: GPT-4 and Llama-2-70b. GPT-4 demonstrated exceptional performance, achieving an F1 score of 1 for its definitive “High” and “Low” predictions. This model employed a conservative strategy, labeling uncertain cases as "Unknown," which resulted in high precision for identifying promising drug repurposing candidates (**Table 5**). GPT-4's approach of only declaring an outcome when highly confident based on our dataset of drug-target agonists significantly enhanced its accuracy and reliability, particularly in distinguishing high-affinity drugs. In contrast, Llama-2-70b adopted a more exploratory approach to prediction. It achieved an F1 score of 0.8108 when discounting its "Unknown" verdicts, indicating a broader but slightly less precise prediction strategy (**Figure 2B**). Llama-2-70b was more inclined to make definitive predictions, only classifying 3 cases as "Unknown" compared to GPT-4's 11 "Unknown" predictions. This difference in approach highlights the trade-off between precision and coverage in prediction tasks.

When incorporating the "Unknown" verdicts into the final performance assessment, GPT-4 and Llama-2-70b achieved F1 scores of 0.7917 and 0.8000, respectively (**Figure 2B**). These scores reflect the models' overall performance, balancing their ability to make accurate predictions with their willingness to classify uncertain cases. The distinct approaches and complementary strengths of these two LLMs—GPT-4's high precision and Llama-2-70b's broader coverage—motivated our decision to incorporate both models in the subsequent phase of our study, where we combined LLM predictions with traditional drug-target affinity predictors.

Our integrated system employs a logic framework prioritizing LLM verdicts, particularly leveraging GPT-4's high accuracy, while incorporating "Unknown" LLM verdicts through confidence ratings. This integration improved the accuracy and interpretability of binary drug affinity predictions across 32 distinct drugs (**Table 6**). Our framework categorizes each drug's predicted affinity and assigns a confidence level based

https://doi.org/10.59720/24-103

Drug Name	Predicted IC50 Score (nM)	True Affinity	Predicted Affinity
Butaclamol	15.57	High	High
Sertraline	27.70	High	High
Fluspirilene	29.47	High	High
Flupentixol	29.95	High	High
Escitalopram	32.40	High	High
Ifenprodil	33.89	High	High
Haloperidol	38.05	High	High
Chlorpromazine	40.04	High	High
Fluphenazine	45.98	High	High
Trifluoperazine	48.66	High	High
Ondansetron	52.30	High	High
Doxycycline	61.07	Low	High
Cetirizine	62.43	Low	High
Pimozide	63.76	High	High
Fluoxetine	64.75	High	High
Imipramine	68.12	High	High
Panamesine	68.44	High	High
Metoclopramide	74.98	Low	High
Triflupromazine	82.98	High	High
Amitriptyline	97.67	High	High
Allopurinol	100.98	Low	High
Lacosamide	147.07	Low	High
Clomipramine	149.45	High	High
Furosemide	155.47	Low	High
Aspirin	183.07	Low	High
Diphenhydramine	267.97	Low	High
Carbamazepine	300.19	Low	High
Ibuprofen	326.34	Low	High
Oxazepam	658.04	Low	High
Psilocybin	784.40	Low	Low
Gabapentin	997.65	Low	Low
Acetaminophen	1254.45	Low	Low

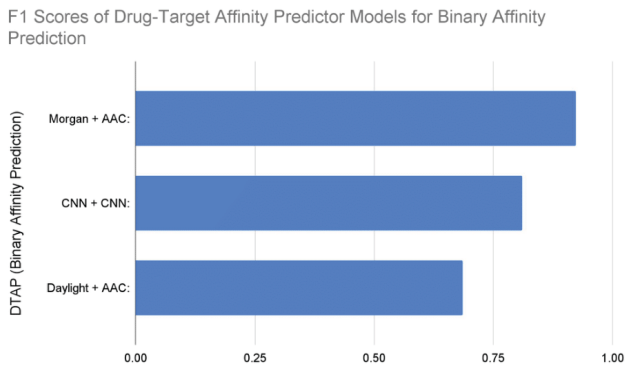
Table 4: Binary Affinity Prediction for High Affinity and Low Affinity Compounds Using the Daylight + AAC Model. Prediction results for 32 drugs (19 high sigma-1 affinity drugs, 13 low affinity sigma-1 drugs) based on IC_{50} threshold values (high affinity < 750 nM, low affinity > 750 nM) using Daylight + AAC model (MSE = 79.093 μM). All models were trained on the BindingDB database over 200 epochs with a 70% training, 20% validation, and 10% testing split.

on consistency across the integrated results. "Confident" predictions indicate unanimous agreement between the LLM and affinity model, while "speculative" labels highlight discrepancies, pointing to potential uncertainties. The GPT-4 and Morgan AAC combination achieved an F1 score of 0.9474, with incorrect predictions occurring only when the models disagreed, reflecting GPT-4's precise and cautious strategy (**Figure 2B**). The Llama-2-70b and Morgan AAC integration showed an F1 score of 0.8421, with "speculative" scores applied to all discrepancies except Ibuprofen, indicating a wider but slightly less precise exploratory scope (**Figure 2B**). These results demonstrate how integrating LLM insights with traditional drug-target affinity models can enhance prediction accuracy. The combination of GPT-4's analytical precision, Llama-2-70b's broad predictive capacity, and the Morgan + AAC model's structural insights creates a framework that could refine drug discovery and repurposing processes.

DISCUSSION

We evaluated the efficacy of DTAPs in determining exact IC_{50} values. Notably, the model that integrates Morgan and AAC algorithms exhibited the smallest MSE, contrary to our initial hypothesis which favored the dual CNN model. The model combining Morgan and AAC algorithms outperformed our hypothesized dual CNN model, highlighting important considerations for deep learning approaches in drug discovery. This finding suggests that model performance may be highly dependent on the specific properties and mechanisms of action of different drug classes. The complex interactions between psychotropic drugs and neurotransmitter systems may be better captured by Morgan fingerprinting's ability to represent detailed molecular structures than by CNN-based pattern recognition. This observation has broader implications for the application of DTAPs and LLMs in drug discovery, suggesting that model architecture selection should carefully consider the unique characteristics of the drug class being studied. Future research could focus on developing molecular descriptors and neural network architectures that can better represent these drug-specific complexities, perhaps through

A:



B:

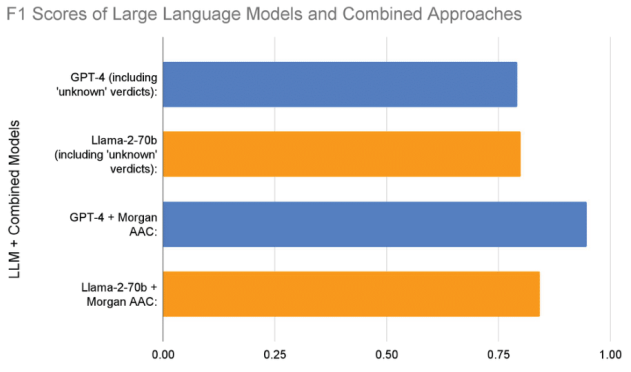


Figure 2: Comparative Analysis of F1 Scores for DTAPs and LLMs. F1 scores of various approaches for predicting drug-target affinity using **A)** traditional prediction models: Morgan + AAC, CNN + CNN, and Daylight + AAC, and **B)** LLMs and combined approaches: GPT-4, Llama-2-70b (both including "unknown" verdicts), GPT-4 + Morgan AAC, and Llama-2-70b + Morgan AAC. Each bar represents the model's F1 score. In **B)**, the blue bars represent F1 scores associated with GPT-4 while the orange bards represent the F1 scores associated with Llama-2-70b.

https://doi.org/10.59720/24-103

Drug Name	True Affinity	Predicted GPT-4 Affinity	Predicted Llama-2-70b Affinity
Butaclamol	High	Unknown	High
Panamesine	High	Unknown	High
Haloperidol	High	High	High
Ifenprodil	High	High	High
Triflupromazine	High	High	High
Clomipramine	High	High	Unknown
Imipramine	High	High	High
Escitalopram	High	High	High
Chlorpromazine	High	High	High
Flupentixol	High	High	High
Fluoxetine	High	High	High
Fluphenazine	High	High	High
Sertraline	High	High	High
Trifluoperazine	High	High	High
Ibuprofen	Low	Unknown	High
Aspirin	Low	Unknown	High
Pimozide	High	High	High
Fluspirilene	Low	Unknown	Low
Metoclopramide	Low	Unknown	Low
Furosemide	Low	Low	Low
Ondansetron	Low	Low	Low
Doxycycline	Low	Low	High
Diphenhydramine	Low	Unknown	Unknown
Allopurinol	Low	Unknown	Low
Oxazepam	Low	Unknown	Low
Amitriptyline	Low	Unknown	High
Cetirizine	Low	Low	High
Psilocybin	Low	Low	Low
Gabapentin	Low	Low	High
Lacosamide	Low	Unknown	High
Carbamazepine	Low	Low	Low
Acetaminophen	Low	Low	Unknown

Table 5: Binary Affinity Classifications for Sigma-1 Ligands. The table displays binary affinity predictions—categorized as high, low, or unknown—for 32 drugs evaluated by the GPT-4 and Llama-2-70b models. The predictions were derived from the analysis conducted by the GPT-4 and Llama-2-70b language models, which were supplemented with information from 50 scientific articles focused on sigma-1 ligands for drug repurposing, extracted from Google Scholar searches.

the integration of domain-specific knowledge from relevant scientific literature.

All predicted exact values were below the 750 nM range when processing sigma-1 agonists. This outcome facilitated the calibration of the models to interpret values exceeding this threshold as indicative of "LOW" sigma-1 affinity. In terms of binary affinity accuracy, the model that combined Morgan fingerprinting architecture with AAC outperformed others, achieving an F1 score of 0.9231.

During the evaluation phase of our DTAPs, we observed that compounds like Haloperidol were consistently predicted as high-affinity binders across all three model architectures (CNN + CNN, Morgan + AAC, Daylight + AAC). This consistent prediction across models is encouraging, as compounds like Haloperidol, a well-documented sigma-1 receptor antagonist (30), were reliably identified by all models. The fact that all three models consistently identified compounds with established binding features demonstrates their reliability in recognizing structural characteristics that contribute to receptor affinity. While this consistency validates our models' performance, future work could explore their ability to identify novel compounds with different structural motifs that might also enable effective receptor binding.

In binary classification, GPT-4 achieved an F1 score of 1, excluding “unknown” verdicts, which reduced to 0.7917 when “unknown” verdicts were included. We implemented a confidence-based weighting system combining LLMs with drug-target affinity models, yielding improved F1 scores of 0.9474 and 0.8421 for GPT-4 and Llama-2-70b, respectively. These results demonstrate the efficacy of hybrid models in enhancing predictive accuracy while balancing the need for manual verification. This approach not only improves overall prediction accuracy but also provides researchers with valuable meta-information about the reliability of each prediction, which could guide decision-making processes and help prioritize candidates for further investigation.

A significant concern that warrants further discussion is the possibility of contamination in our LLM-based approach. Data contamination occurs when training data inadvertently includes test or evaluation data, or when the model has been exposed to the material it's meant to predict or evaluate (25). Given the vast size of datasets used to train models like GPT-4, ensuring absolutely no contamination is challenging. As noted by Bommasani, *et al.*, the risk of test set contamination in LLMs is a pressing issue that requires careful consideration and mitigation strategies (25). To address this, we propose exploring the development of our own LLM, specifically trained on a curated and verifiably clean dataset. This approach, while resource-intensive, could potentially provide a more controlled environment for our predictions and reduce the risk of contamination.

The dataset creation and management process presented a significant bottleneck in our methodology, primarily due to the labor-intensive nature of manual data extraction and the potential for variability depending on the receptor of interest. To address this, we advocate for the automation of data collection through the utilization of application programming interfaces from established scientific databases such as PubMed and Google Scholar. This approach would not only streamline the dataset compilation process but also allow for the creation of more dynamic and scalable datasets. The current lack of automation in the validation processes of LLM and drug-target affinity models also results in a time-consuming and manual verification procedure. To mitigate this, the development of machine learning pipelines capable of automating data processing, model feeding, and result generation is recommended. This could be complemented by user-friendly interfaces or software solutions that automatically produce combined results tables, perform data visualization, and identify key findings.

It's important to note that our study primarily relied on IC_{50} values as the output format for our drug-target affinity

https://doi.org/10.59720/24-103

Drug Name	True Affinity	GPT-4 + Affinity Model Verdict	GPT-4 + Affinity Model Confidence Verdict	Llama-2-70b + Affinity Model Verdict	Llama-2-70b + Affinity Model Confidence Verdict
Butaclamol	High	High	Confident	High	Confident
Panamesine	High	High	Confident	High	Confident
Haloperidol	High	High	Confident	High	Confident
Ifenprodil	High	High	Confident	High	Confident
Triflupromazine	High	High	Confident	High	Confident
Clomipramine	High	High	Confident	High	Speculative
Imipramine	High	High	Confident	High	Confident
Escitalopram	High	High	Confident	High	Confident
Chlorpromazine	High	High	Confident	High	Confident
Flupentixol	High	High	Confident	High	Confident
Fluoxetine	High	High	Confident	High	Confident
Fluphenazine	High	High	Confident	High	Confident
Sertraline	High	High	Confident	High	Confident
Trifluoperazine	High	High	Confident	High	Confident
Ibuprofen	Low	High	Speculative	High	Confident
Aspirin	Low	High	Speculative	High	Confident
Pimozide	High	High	Confident	High	Confident
Fluspirilene	Low	High	Speculative	Low	Speculative
Metoclopramide	Low	Low	Speculative	Low	Confident
Furosemide	Low	Low	Confident	Low	Confident
Ondansetron	Low	Low	Confident	Low	Confident
Doxycycline	Low	Low	Confident	High	Speculative
Diphenhydramine	Low	Low	Speculative	Low	Speculative
Allopurinol	Low	Low	Speculative	Low	Confident
Oxazepam	Low	Low	Speculative	Low	Confident
Amitriptyline	Low	Low	Speculative	High	Speculative
Cetirizine	Low	Low	Confident	High	Speculative
Psilocybin	Low	Low	Confident	Low	Confident
Gabapentin	Low	Low	Confident	High	Speculative
Lacosamide	Low	Low	Speculative	High	Speculative
Carbamazepine	Low	Low	Confident	Low	Confident
Acetaminophen	Low	Low	Confident	Low	Speculative

Table 6: Integrated Affinity Classifications and Confidence Levels for Sigma-1 Ligands. Affinity predictions from GPT-4 and Llama-2-70b models were merged with the Morgan + AAC binary outcomes, showcasing high, low, or unknown affinity classes and confident or speculative confidence ratings for 32 distinct drugs. Affinity assessments from Tables 3 and 2 were combined, applying a logical framework to categorize each drug's predicted affinity and derive a confidence level based on the consistency and clarity across the integrated dataset results.

https://doi.org/10.59720/24-103

models. While IC_{50} is a widely used metric in pharmacology, it has limitations. IC_{50} values can vary significantly based on experimental conditions and substrate concentrations, making direct comparisons between different studies challenging. Additionally, IC_{50} measurements don't provide information about the mechanism of inhibition or the binding kinetics that could be crucial for drug efficacy in vivo (31). Future iterations of this experiment should consider incorporating additional pharmacological parameters to provide a more comprehensive view of drug-target interactions. These could include binding affinity (Ki), efficacy (Emax), and on/off rates. By expanding the range of parameters, we could potentially improve the accuracy and applicability of our models in diverse drug discovery scenarios.

Ultimately, our findings demonstrate that the integration of LLMs with DTAPs significantly improves the practical usability and accuracy of drug repurposing efforts. The combined model, leveraging both the computational strengths of LLMs and the specific insights from DTAPs, outperformed the individual components in binary classification accuracy. This synergy was further enhanced by the introduction of confidence verdicts, which allowed for a more nuanced interpretation of results, distinguishing between “confident” and “speculative” predictions. The enhanced performance and practicality of the combined models underscore the potential of leveraging LLMs in streamlining the identification and validation of new therapeutic uses for existing drugs.

MATERIALS AND METHODS

Dataset Development and Curation

Dataset 1 comprised 46 psychotropic drugs with known exact sigma-1 receptor IC_{50} values. We curated this dataset as a subset of drugs featured in the comprehensive review by Cobos et al. on sigma-1 receptor ligands (26). The selection process prioritized drugs with well-documented sigma-1 receptor interactions and precise IC_{50} measurements, ensuring a robust foundation for evaluating our models' ability to predict exact affinity values. This dataset represents a diverse range of psychotropic compounds, including antidepressants, antipsychotics, and anxiolytics, providing a comprehensive spectrum of sigma-1 receptor agonists.

To assess our models' capacity for binary affinity prediction, we constructed dataset 2, consisting of 32 drugs: 19 high-affinity high sigma-1 affinity drugs and 13 compounds with low or negligible sigma-1 affinity. The high-affinity sigma-1 drugs were randomly selected from dataset 1. To complement these, we conducted an extensive literature review to identify compounds with established negligible sigma-1 affinity. The compounds were primarily sourced from two comprehensive reviews: a study by Cobos et al. on sigma-1 receptor ligands (26), and Hayashi's work on the sigma-1 receptor as a target for neuropsychotherapeutic drugs (12).

DeepPurpose Framework and Model Implementation

We employed the DeepPurpose framework, a Python-based computational tool optimized for drug discovery applications, particularly drug-target interaction predictions (9). DeepPurpose offers a comprehensive suite of machine learning and deep learning models, along with sophisticated data preprocessing tools suited for various molecular and biological data types.

The preprocessing stage involved normalizing chemical compounds and encoding protein sequences to conform to the models' input requirements. We leveraged the extensive compound-protein interaction data from the BindingDB database (20) to enhance our models' predictive capabilities. This process ensured standardized and compatible inputs across all model architectures.

We implemented three distinct model architectures provided by DeepPurpose. The Morgan Fingerprint with Amino Acid Composition (Morgan+AAC) model combines Morgan Fingerprints for detailed drug structure representation with Amino Acid Composition for quantitative protein analysis (13, 14). Convolutional Neural Networks (CNNs) were employed for both drug and target structure analysis, excelling at extracting complex spatial and hierarchical features suitable for understanding intricate relationships between chemical compounds and biological targets (15). The Daylight Fingerprint with Amino Acid Composition (Daylight+AAC) model utilizes Daylight Fingerprints to transform complex molecular structures into detailed binary representations, coupled with AAC for protein target analysis (14, 16).

The training dataset, derived from BindingDB, was partitioned into 70% for training, 10% for validation, and 20% for testing, adhering to standard machine learning practices to ensure a balanced approach between model training and evaluation (21). Each model underwent an intensive training regimen spanning 200 epochs. We closely monitored performance and accuracy metrics throughout the training process to optimize learning outcomes and mitigate potential overfitting.

LLM Integration and Evaluation

To explore the potential of LLMs in drug discovery, we integrated two state-of-the-art LLMs into our workflow: GPT-4 and Llama-2-70b (32,33). We embedded a curated dataset of 50 sigma-1-specific scientific articles within these LLMs to provide them with specialized knowledge in the domain. The knowledge base was constructed from the top 50 articles on "Sigma-1 Receptor ligands" from Google Scholar, ranked by citation count. To incorporate this scientific literature, we implemented a PDF embedding process using LangChain and Pinecone. This process involved splitting PDF contents into manageable chunks, generating vector embeddings using both LLMs, storing these embeddings in Pinecone, and performing similarity searches to retrieve relevant information. We employed these prompt engineering techniques combined with PDF embedding to enhance the performance of GPT-4 and Llama-2-70b for determining drug affinities to the sigma-1 receptor. Both models were prompted with the instruction: "Determine the level of affinity each drug has to the sigma-1 receptor by outputting a verdict: “High Affinity”, “Low Affinity”, or “Unknown” for each of the listed drugs. This approach allowed us to leverage prompt engineering and efficient information retrieval from current scientific literature to guide the models in categorizing drug affinities without the need for fine-tuning.

For our final experiments, we integrated the LLMs with the Morgan AAC DTAP model. We developed a Python script to synthesize the outputs from the LLM and the Morgan AAC DTAP. This script processed the verdict from the LLM (High Affinity, Low Affinity, or Unknown) and the binary prediction from the Morgan AAC DTAP (High or Low Affinity). The script

https://doi.org/10.59720/24-103

then applied a logic system to produce a confidence verdict (Confident or Speculative) and a final affinity verdict (High or Low Affinity). The logic system operated as follows: If the LLM verdict matched the DTAP prediction, the confidence was set to "Confident," and this agreed-upon affinity became the final verdict. When the LLM verdict was "Unknown", the confidence was set to "Speculative" and the final verdict defaulted to the DTAP prediction. If the models contradicted each other, the confidence was set to "Speculative" and the final verdict defaulted to the LLM prediction. This approach allowed us to leverage the strengths of both the LLMs and the DTAP, providing not only a final affinity prediction but also a measure of confidence in that prediction.

Received: March 3, 2024
Accepted: August 12, 2024
Published: March 2, 2025

REFERENCES

1. Sliwoski, Gregory, *et al.* "Computational Methods in Drug Discovery." *Pharmacological Reviews,* vol. 66, no. 1, 31 Dec. 2013, pp. 334-395. https://doi.org/10.1124/pr.112.007336.
2. Mullard, A. "New Drugs Cost US $2.6 Billion to Develop." *Nature Reviews Drug Discovery,* vol. 13, no. 1, Nov. 2014, p. 877. https://doi.org/10.1038/nrd4507.
3. Brown, Dean G., *et al.* "Clinical Development Times for Innovative Drugs." *Nature Reviews Drug Discovery,* vol. 21, 10 Nov. 2021, pp. 793-794. https://doi.org/10.1038/d41573-021-00190-9.
4. Sun, Duxin, *et al.* "Why 90% of Clinical Drug Development Fails and How to Improve It?" *Acta Pharmaceutica Sinica B,* vol. 12, no. 7, 2022, pp. 3049-3062. https://doi.org/10.1016/j.apsb.2022.02.002.
5. Krishnamurthy, Nithya, *et al.* "Drug Repurposing: A Systematic Review on Root Causes, Barriers and Facilitators." *BMC Health Services Research,* vol. 22, no. 1, 29 Jul. 2022, p. 970. https://doi.org/10.1186/s12913-022-08272-z.
6. Begley, C. Glenn, *et al.* "Drug Repurposing: Misconceptions, Challenges, and Opportunities for Academic Researchers." *Science Translational Medicine,* vol. 13, no. 612, 2021, pp. eabd5524. https://doi.org/10.1126/scitranslmed.abd5524.
7. Askr, Heba, *et al.* "Deep Learning in Drug Discovery: An Integrative Review and Future Challenges." *Artificial Intelligence Review,* vol. 56, no. 7, 2023, pp. 5975-6037. https://doi.org/10.1007/s10462-022-10306-1.
8. Paul, Debleena, *et al.* "Artificial Intelligence in Drug Discovery and Development." *Drug Discovery Today,* vol. 26, no. 1, Jan. 2021, pp. 80-93. https://doi.org/10.1016/j.drudis.2020.10.010.
9. Huang, Kexin, *et al.* "DeepPurpose: A Deep Learning Library for Drug-Target Interaction Prediction." *Bioinformatics,* vol. 36, no. 22-23, Dec. 2020, pp. 5545-5547. https://doi.org/10.1093/bioinformatics/btaa1005.
10. Boiko, D.A., *et al.* "Autonomous Chemical Research with LLMs." *Nature,* vol. 624, 2023, pp. 570-578. https://doi.org/10.1038/s41586-023-06792-0.
11. Vela, José Miguel. "Repurposing Sigma-1 Receptor Ligands for COVID-19 Therapy?" *Frontiers in Pharmacology,* vol. 11, 9 Nov. 2020, pp. 582310. https://doi.org/10.3389/fphar.2020.582310.
12. Hayashi, Teruo. "Sigma-1 Receptor: The Novel Intracellular Target of Neuropsychotherapeutic Drugs." *Journal of Pharmacological Sciences,* vol. 127, no. 1, 2015, pp. 2-5. https://doi.org/10.1016/j.jphs.2014.07.001.
13. Capecchi, A., *et al.* "One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome." *Journal of Cheminformatics,* vol. 12, 2020, p. 43. https://doi.org/10.1186/s13321-020-00445-4.
14. Khojasteh, Hakimeh, *et al.* "Improving Prediction of Drug-Target Interactions Based on Fusing Multiple Features with Data Balancing and Feature Selection Techniques." *PLOS ONE,* vol. 18, no. 8, 3 Aug. 2023, pp. e0288173. https://doi.org/10.1371/journal.pone.0288173.
15. Vaz, Joel Markus, and S. Balaji. "Convolutional Neural Networks (CNNs): Concepts and Applications in Pharmacogenomics." *Molecular Diversity,* vol. 25, no. 3, 2021, pp. 1569-1584. https://doi.org/10.1007/s11030-021-10225-3.
16. "Daylight Theory: SMARTS – A Language for Describing Molecular Patterns." *Daylight Theory Manual,* Daylight Version 4.9, Laguna Niguel, CA, 2011.
17. Zhao, Yilun, *et al.* "DocMath-Eval: Evaluating Numerical Reasoning Capabilities of LLMs in Understanding Long Documents with Tabular Data." *arXiv preprint,* 2023. https://arxiv.org/abs/2311.09805.
18. Cheung, Diana. "Meta Llama 2 vs. OpenAI GPT-4: A Comparative Analysis." *Medium,* 21 Mar. 2024.
19. Jiang, Mingjian, *et al.* "A Deep Learning Method for Drug-Target Affinity Prediction Based on Sequence Interaction Information Mining." *PeerJ,* vol. 11, 11 Dec. 2023, pp. e16625. https://doi.org/10.7717/peerj.16625.
20. Liu, Tiqing, *et al.* "BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities." *Nucleic Acids Research,* vol. 35, Database issue, 2007, pp. D198-201. https://doi.org/10.1093/nar/gkl999.
21. Gholamy, Afshin, *et al.* "Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation." Feb. 2018.
22. Sharma, M. "Leveraging GPT-4 for PDF Data Extraction: A Comprehensive Guide." 11 Mar. 2024.
23. Iyer, Prasanth Mani, *et al.* "The Emerging Role of the Sigma-1 Receptor in Autophagy: Hand-in-Hand Targets for the Treatment of Alzheimer's." *Expert Opinion on Therapeutic Targets,* vol. 25, 10 Jun. 2021. https://doi.org/10.1080/14728222.2021.1939681.
24. BindingDB. (n.d.). https://www.bindingdb.org/
25. Bommasani, Rishi, *et al.* "On the Opportunities and Risks of Foundation Models." *arXiv preprint,* 2021. https://arxiv.org/abs/2108.07258.
26. Cobos, E. J., *et al.* "Pharmacology and Therapeutic Potential of Sigma1 Receptor Ligands." *Current Neuropharmacology,* vol. 6, no. 4, 2008, pp. 344-66. https://doi.org/10.2174/157015908787386113.
27. Vela, José Miguel. "Repurposing Sigma-1 Receptor Ligands for COVID-19 Therapy?" *Frontiers in Pharmacology,* vol. 11, 9 Nov. 2020, pp. 582310. https://doi.org/10.3389/fphar.2020.582310.
28. Aykul, S., and E. Martinez-Hackert. "Determination of Half-Maximal Inhibitory Concentration Using Biosensor-Based Protein Interaction Analysis." *Analytical Biochemistry,* vol.

https://doi.org/10.59720/24-103

508, 1 Sep. 2016, pp. 97-103. https://doi.org/10.1016/j.ab.2016.06.025.
29. Damiani, Elisabetta, *et al.* "How Reliable Are in Vitro IC50 Values? Values Vary with Cytotoxicity Assays in Human Glioblastoma Cells." *Toxicology Letters,* vol. 302, 1 Mar. 2019, pp. 28-34. https://doi.org/10.1016/j.toxlet.2018.12.004.
30. Luedtke, Robert R., *et al.* "Neuroprotective Effects of High Affinity Sigma 1 Receptor Selective Compounds." *Brain Research,* vol. 1441, 31 Dec. 2011, pp. 17-26. https://doi.org/10.1016/j.brainres.2011.12.047.
31. Damiani, Elisabetta, *et al.* "How Reliable Are in Vitro IC50 Values? Values Vary with Cytotoxicity Assays in Human Glioblastoma Cells." *Toxicology Letters,* vol. 302, 1 Mar. 2019, pp. 28-34. https://doi.org/10.1016/j.toxlet.2018.12.004.
32. OpenAI. "GPT-4 Technical Report." *arXiv preprint,* 15 Mar. 2023. https://doi.org/10.48550/arXiv.2303.08774.
33. Touvron, Hugo, *et al.* "Llama 2: Open Foundation and Fine-Tuned Chat Models." *arXiv preprint,* 18 Jul. 2023. https://doi.org/10.48550/arXiv.2307.09288.
34. Christen, Peter, *et al.* "A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives." *ACM Computing Surveys,* vol. 56, no. 3, Oct. 2023, pp. 1-24. https://doi.org/10.1145/3606367.
35. Mukherjee, Prasoon. "Large Language Models Are Not a Solution for Precise Data Extraction in Banking." Finextra Research, 25 Oct. 2023. https://www.finextra.com/community-article/large-language-models-are-not-a-solution-for-precise-data-extraction-in-banking