Assay2Mol: Large Language Model-based Drug Design Using BioAssay Context

Yifan Deng[1, 2], Spencer S. Ericksen[3], Anthony Gitter[1,2,4]
[1] Department of Computer Sciences, University of Wisconsin-Madison
[2] Morgridge Institute for Research
[3] Drug Development Core, Small Molecule Screening Facility,
University of Wisconsin Carbone Cancer Center, University of Wisconsin-Madison
[4] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison
{yifan.deng,sericksen}@wisc.edu
gitter@biostat.wisc.edu

Abstract

Scientific databases aggregate vast amounts of quantitative data alongside descriptive text. In biochemistry, molecule screening assays evaluate candidate molecules' functional responses against disease targets. Unstructured text that describes the biological mechanisms through which these targets operate, experimental screening protocols, and other attributes of assays offer rich information for drug discovery campaigns but has been untapped because of that unstructured format. We present Assay2Mol, a large language model-based workflow that can capitalize on the vast existing biochemical screening assays for early-stage drug discovery. Assay2Mol retrieves existing assay records involving targets similar to the new target and generates candidate molecules using in-context learning with the retrieved assay screening data. Assay2Mol outperforms recent machine learning approaches that generate candidate ligand molecules for target protein structures, while also promoting more synthesizable molecule generation.

1 Introduction

Early-stage drug development and target validation typically involve a search through chemical space for drug-like molecules that perturb a target of interest, usually a protein activity connected to a disease condition. For a new target, the search starts with assay development and scaling for high-throughput testing in order to experimentally screen a pre-defined chemical library for active molecules. From data obtained on screens of a target of interest (or related targets), computational models can learn structure-activity relationships that inform selection of new molecules for testing.

Public screening data repositories like PubChem (Kim et al., 2024) and ChEMBL (Mendez et al., 2018) possess great value in this regard. PubChem now contains 1.77 million assay records (BioAssay records) comprising ∼300 million bioactivity outcomes across ∼250,000 protein targets and ∼2,000 cell lines[1]. BioAssay records are richly annotated with chemicals tested, target genes, pathways, proteins, cell lines, publications, patents, related BioAssay records, tabular molecule testing results, text descriptions of assay format, protocols, and relevance to disease states.

Sorting through these repositories for the data pertinent to an arbitrary target is a daunting task. Scientists need workflows that can rapidly identify relevant BioAssay records based on their associated text, extract key textual and tabular chemical testing data comprising molecule structures paired with experimental activity outcomes, and apply this information to models capable of learning structure-activity relationships to recommend new molecules for testing.

Given the extensive descriptive text components in BioAssay records, leveraging natural language processing capabilities becomes crucial for efficient retrieval and interpretation. Large language models (LLMs), with their advanced ability to process and interpret unstructured text, are well-suited for assessing BioAssay relevance by extracting key experimental details and identifying meaningful activity patterns. LLMs have demonstrated great ability in different kinds of tasks including translation, multi-round conversation, and so on. LLMs are also adept in biology- and chemistry-related tasks, for example, molecule property prediction, and text-guided molecule generation. LLMs support in-context learning, which extends to scientific domains. GPT-3 (Brown et al., 2020) highlights the ability of LLMs to adapt to new tasks with minimal examples. In text-guided molecule design, GPT-4 outperforms other fine-tuned models with few-shot in-context learning (Zhao et al., 2024; Guo et al., 2023). This raises the question of whether LLMs

[1] https://pubchem.ncbi.nlm.nih.gov/docs/statistics

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27326–27351
November 4-9, 2025

can support different strategies for molecule design. Beyond generating molecules that satisfy specified property constraints, can LLMs navigate unstructured text within public BioAssay records and then use it as context to design molecules with functional properties, such as protein inhibition?

We propose Assay2Mol to maximize the use of BioAssay records with LLMs. Given input such as protein target descriptions, phenotypic data, or other textual information, Assay2Mol retrieves relevant BioAssays and leverages in-context learning to generate molecules with the desired biological activity (Figure 1). Our contributions can be summarized as follows:

- We introduce Assay2Mol, an LLM-based drug design workflow that retrieves relevant PubChem BioAssay data for a given query and then learns to generate candidate molecules from this assay context.
- Unlike structure-based drug discovery (SBDD) methods, Assay2Mol does not require protein structures or even sequences. It can even generate candidate active molecules for cell-based and phenotypic assays and endpoints (e.g., tumor shrinkage, cardiotoxicity, QT interval prolongation).
- Because Assay2Mol relies on LLMs that include molecules in their pretraining data, some output molecules are more like "retrieval" rather than *de novo* "generation". This increases the chemical plausibility and synthetic accessibility of generated molecules.

2 Related work

2.1 Large Language Models in biochemistry

LLMs are an alternative to biomolecular sequence- or structure-based models for learning structure-activity or structure-property relationships. Multimodal molecule and text generation (Edwards et al., 2021; Pei et al., 2024; Deng et al., 2025; Zhao et al., 2024; Fang et al., 2023), in-context learning for chemistry (Fifty et al., 2023; Jablonka et al., 2024; Nguyen and Grover, 2025; Moayedpour et al., 2024; Schimunek et al., 2025), and LLM agents (M. Bran et al., 2024; McNaughton et al., 2024; Swanson et al., 2025; Gao et al., 2025; Liu et al., 2025; Wei et al., 2025) are at the frontier of this interface of LLMs and biomolecules. Recent reviews (Zhang et al., 2024; Mirza et al., 2025; Ramos et al., 2025; Wang et al., 2025; Alampara et al., 2025) provide broader coverage of this expansive area.

2.2 BioAssay data mining

BioAssay databases like PubChem and ChEMBL are valuable resources for data mining and have been used to train many machine learning models. Sharma et al. (2024) developed a data mining pipeline that compiled and processed 8,415 OXPHOS-related BioAssays from PubChem, identified major OXPHOS inhibitory chemotypes, and trained effective OXPHOS inhibitor classifiers. MolecularGPT (Liu et al., 2024) constructed an instruction tuning dataset by collecting three-shot examples from ChEMBL. MBP (Yan et al., 2023) created a multi-task pretraining dataset with labels from BioAssays to address label inconsistencies and data scarcity. Seidl et al. (2023) proposed CLAMP, a dual-encoder architecture combining chemical structure and natural-language descriptions of BioAssays, trained with a contrastive objective to enable zero- and few-shot activity prediction. TwinBooster (Schuh et al., 2024) is a zero-shot model that integrates molecule structures and BioAssay descriptions for molecular property prediction. Schoenmaker et al. (2025) demonstrated that incorporating assay-aware embeddings derived from ChEMBL assay descriptions can reduce variance in bioactivity data and improve proteochemometric modeling performance. Han et al. (2025) extracted structured BioAssay data from ChEMBL, PubChem, and literature to train AutoML-based models for 11 ADMET properties, without employing any natural language information. Smit et al. (2025) developed manual and AI-based methods to improve BioAssay annotations in ChEMBL, including automated extraction of experimental methods and refined classification. It increased the reusability of bioactivity data, supporting more reliable downstream modeling.

2.3 Structure-based ligand design

Molecule design approaches play a role in the hit finding and lead optimization tasks of early-stage drug discovery. The goal of hit finding is to identify pharmacologically active molecules for a target of interest to serve as starting points for development. Given availability of a 3D structure model for a protein target, structure-based virtual screening methods can be applied for hit finding, such as large-scale docking of pre-enumerated molecule libraries. Alternatively, *de novo* design approaches (Bohacek et al., 1999; Spiegel and Durrant, 2020) have re-

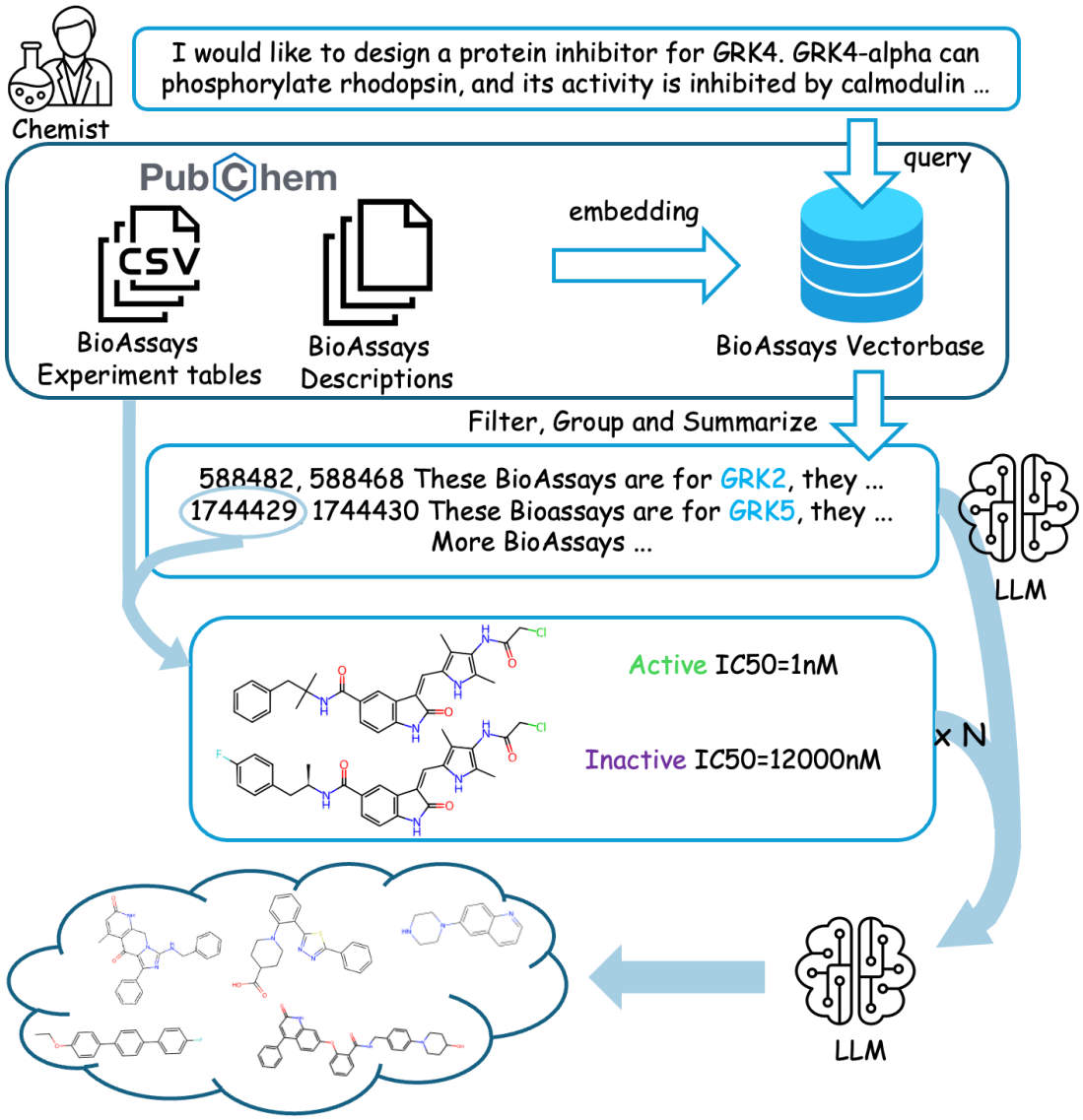

Figure 1: The Assay2Mol workflow. A chemist provides a target description, which is used to retrieve BioAssays from the pre-embedded vector database. After filtering for relevance, the BioAssays are summarized by an LLM. The BioAssay ID is then used to retrieve experimental tables. The final molecule generation prompt is formed by combining the description, summarization, and selected test molecules with associated test outcomes, enabling the LLM to generate relevant active molecules. Icons are from Flaticon.com and svgrepo.com

emerged with the advent of deep learning-based generative methods adapted to build molecules with enhanced affinity for a target structure of interest. A variety of generative approaches have been explored, including Conditional Variational Autoencoder (cVAE) (Ragoza et al., 2022), flow-based models (Shi et al., 2020; Jiang et al., 2024; Qu et al., 2024; Cremer et al., 2025), diffusion models (Guan et al., 2023a; Schneuing et al., 2024; Guan et al., 2023b), and Generative Pretrained Transformers (GPTs) (Wu et al., 2024).

3 Assay2Mol framework

3.1 Motivating example

Before designing a general algorithm, we begin with a proof of concept to assess whether an LLM generates relevant molecules and how BioAssay context affects generation. Using the UniProt (The UniProt Consortium, 2023) protein description of GRK4 (UniProt P32298; PDB 4YHJ) as input, we prompt ChatGPT 4o to generate five molecules. Next, we use the same protein description to search for related BioAssays (see Section 3.3) and retrieve

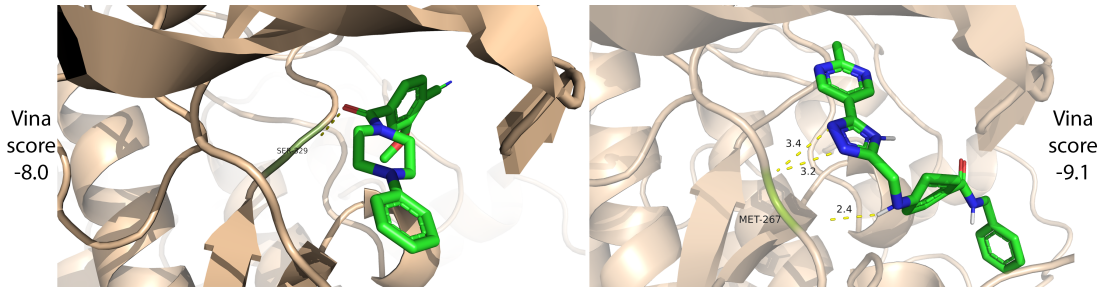

Figure 2: Docked binding poses of generated molecules without (left) and with (right) BioAssay context. With BioAssay context, ChatGPT 4o generates a molecule with three hydrogen bonds to the GRK4 pocket residue MET-267, improving the docking score.

four BioAssays related to GRK4: 775998[2] (human GRK2), 1315729 (human GRK2), 775996 (human GRK5), 1315749 (bovine GRK2). We generate another five molecules, this time providing the retrieved BioAssay data, including the experimental tables, as additional context to ChatGPT 4o. The average AutoDock Vina (Eberhardt et al., 2021) score for the five molecules without BioAssay context is -7.44 (expressed in units of kcal/mol). The average Vina Dock score for the four molecules with BioAssay context is -8.48 (one was invalid). Lower scores reflect better structural complementarity between the generated molecule and the protein target. Docked poses for the top scoring molecule from each group are shown in Figure 2. This small pilot study supports the proposition that the BioAssay context can improve molecule generation and motivates a full exploration of that problem setting.

3.2 Problem definition

Given a text description p for a target protein or phenotype, we want to retrieve the most relevant BioAssays b. Then, based on the experimental results associated with the BioAssays b, we want to generate molecules that produce the desired target response or activity. Below we use "query protein", though it could be a target protein or phenotype.

3.3 BioAssay retrieval

The BioAssay retrieval stage is similar to Retrieval Augmented Generation (RAG) (Lewis et al., 2020). For a query description, we extract keywords with an LLM and obtain a protein description embedding $\mathbf{p} \in \mathbb{R}^d$. We use the OpenAI text embedding tool (Neelakantan et al., 2022) and obtain an embedding for BioAssay record i in json format, recorded as $\mathbf{b}_i \in \mathbb{R}^d$ (Douze et al., 2025). We use cosine similarity to calculate the similarity between the query protein description and the set of available PubChem BioAssays and then select the top-k related BioAssays:

$$\mathcal{I}_k = \text{arg top-}k \left\{ \frac{\mathbf{p} \cdot \mathbf{b}_i}{\|\mathbf{p}\| \|\mathbf{b}_i\|} : i = 1, 2, \ldots, N \right\}$$
$$\{\mathbf{b}_i : i \in \mathcal{I}_k\}. \quad (1)$$

In contrast to RAG, we do not use the retrieved BioAssays as context directly. Instead, we download data tables of these BioAssays based on their Assay ID (AID) and perform further filtering:

- To ensure fair comparisons, BioAssays directly involving the query protein, identified by matching UniProt IDs, are excluded.
- Many BioAssays are derived from literature and involve only a single or few molecules. We prioritize BioAssays with larger data tables using a filter threshold min_mol_num, which removes BioAssays with fewer molecules tested.
- More shots (examples) will usually lead to better performance with in-context learning. However, the context length of an LLM is limited. Thus, we set max_assay_num to limit the number of BioAssays retrieved.
- Sometimes the query protein has no relevant BioAssays. In such cases, the retrieved BioAssays with the highest cosine similarity would be uninformative. To further assess relevance of the retrieved BioAssays, we also use an LLM to determine whether the retrieved BioAssays are relevant to the query protein.

[2] https://pubchem.ncbi.nlm.nih.gov/bioassay/775998

3.4 Counterscreen BioAssay

BioAssay records are sometimes grouped by project. Projects contain independent records for primary screens (often high-throughput screens) and confirmatory screens (secondary re-testing of hits from the primary screen). Of particular interest are counterscreens, which are designed to detect false positives such as pan-assay interference compounds (PAINS) (Baell and Nissink, 2018) or assess hit specificity by testing for activity on an off-target or "anti-target," an undesirable, sometimes related target. Active molecules in counterscreens are undesirable and should be avoided or used as negative training instances. Therefore, we use the LLM to summarize the retrieved BioAssays and identify whether the BioAssay record represents a counterscreen assay (Appendix A.9.1).

3.5 Layered contextual analysis

The LLM sequentially processes the set of relevant BioAssay records to build an input prompt for molecule generation. The workflow processes each BioAssay in three steps:

1. **Summarization of BioAssay findings:** The LLM generates a concise summary that captures the purpose, methodology, and key results of each BioAssay. Additionally, the LLM states the apparent relationship between the BioAssay and the query protein, considering how the BioAssay's findings may inform the design of new molecules. If the BioAssay is a counterscreen, its active molecules should be avoided.

2. **Presentation of tabular experimental data:** For each BioAssay, the experimental data are presented in a table describing the SMILES notation of each molecule and its activity result (Active, Unspecified, or Inactive). Most experimental data tables also include measured pharmacodynamic parameters expressed using standard types (e.g., IC_{50}, K_i, K_d, percent inhibition), relations (e.g., $<, =, >$), values, and units (e.g., μM, %).

3. **Molecule selection:** If active molecules are identified in the data table, we next check the number of actives. If the number of actives exceeds N_{mol}, we randomly sample N_{mol}. Otherwise, we list all active molecules. To maintain class balance, we randomly sample N_{mol} molecules from the combined unspecified and inactive categories.

   If there are no active molecules, we include all molecules unless they exceed $2 \cdot N_{\text{mol}}$, in which case we instead randomly sample $2 \cdot N_{\text{mol}}$. Accordingly, we increase min_mol_num to $2 \cdot N_{\text{mol}}$ to account for the lack of actives.

An example of the BioAssays summarization is in Appendix A.8.

3.6 Molecule generation

Given the context of the query protein description and BioAssay summaries and data tables, Assay2Mol uses the LLM to generate molecules in batches of 10. Details are in the prompt in Appendix A.9.2. The full Assay2Mol workflow is shown in Figure 1.

4 Experiments

We evaluate Assay2Mol in two settings. First, we compare Assay2Mol with SBDD methods for generating candidate protein-binding molecules. Second, we examine Assay2Mol's ability to manage multiple objectives by generating molecules that bind a query protein and avoid cardiotoxicity.

4.1 Generating binders for target proteins

Dataset. CrossDocked2020 (CrossDocked for short) is a common dataset for SBDD (Francoeur et al., 2020) that allows us to assess how BioAssay context compares to protein structure context for generating candidate protein binders. Previous methods refined the original 22.5 million docked protein binding complexes by isolating those with poses < 1 Å RMSD from native (crystallographic poses) and sequence identities $< 30\%$ from the original dataset. They used 100,000 complexes for training and 100 novel complexes as references for testing. Assay2Mol does not require additional training. We select 100 complexes from the training set to develop our pipeline and then evaluate on the test set. As input prompts for protein targets, we use the descriptions returned from PubChem from queries of the PDB ID of each protein. For proteins whose PDB ID cannot be found in PubChem, we use the UniProt mapping tool to convert the PDB ID into the UniProt ID, which is then used to query the PubChem protein webpage. When this approach fails, we manually collect information

Table 1: Experimental results on the CrossDocked dataset. The best results are shown in **bold**, and the second-best results are underlined.

Model	Vina Dock (↓)		High Affinity (↑)		QED (↑)		SA (↑)		Diversity (↑)		Size	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
Reference	-7.117	-6.905	-	-	0.476	0.468	0.728	0.740	-	-	22.75	21.50
FDA drugs	-7.027	-7.169	0.440	0.380	0.567	0.564	0.760	0.758	0.792		24.47	
ZINC 30 atoms	-7.855	-8.088	0.607	0.735	0.576	0.577	0.737	0.736	0.655		30.0	
CVAE	-6.114	-6.118	0.103	0.026	0.390	0.419	0.591	0.580	0.655	0.666	19.97	20.19
AR	-6.751	-6.707	0.459	0.340	0.505	0.499	0.635	0.634	0.698	0.703	17.78	17.54
Pocket2Mol	-7.200	-6.815	0.601	0.593	0.574	0.579	0.754	0.760	0.741	0.781	17.84	16.53
TamGen	-7.475	-7.775	0.526	0.645	0.559	0.559	0.771	0.759	0.747	0.745	23.13	23.29
TargetDiff	-7.788	-7.964	**0.683**	0.634	0.474	0.485	0.584	0.571	0.717	0.714	24.44	24.64
Gemma-3-27B	-7.050	-7.024	0.416	0.281	0.700	0.711	0.860	0.868	0.757	0.765	19.34	18.94
GPT 4o	-7.198	-7.257	0.432	0.294	**0.789**	**0.803**	**0.870**	**0.878**	0.767	0.767	19.70	19.64
DeepSeekV3	-7.230	-7.170	0.443	0.241	0.743	0.756	0.855	0.867	**0.771**	**0.772**	18.96	19.00
Assay2Mol (Gemma-3-27B)	**-8.064**	**-8.280**	0.610	**0.732**	0.585	0.606	0.821	0.834	0.742	0.611	26.59	26.65
Assay2Mol (GPT 4o)	-7.796	-7.881	0.548	0.576	0.600	0.630	0.790	0.801	0.542	0.547	25.90	25.59
Assay2Mol (DeepSeekV3)	-7.861	-7.936	0.557	0.634	0.616	0.647	0.813	0.820	0.593	0.608	24.46	24.26
Assay2Mol (DeepSeekV3 <30%)	-7.826	-7.925	0.562	0.673	0.594	0.619	0.814	0.820	0.609	0.628	24.49	24.73

Table 2: Average improvement over randomly sampled FDA drugs grouped by LLM-estimated relevance of the retrieved BioAssays. The value represents the increase in the docking score, measured in kcal/mol.

Model	High (39%)		Medium (42%)		Low (7%)		No (12%)		Overall	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
TargetDiff	0.838	0.802	0.701	0.777	0.669	0.696	0.771	1.052	0.761	0.796
Gemma-3-27B	0.196	0.170	-0.050	-0.145	0.197	0.293	-0.390	-0.535	0.023	0.034
GPT 4o	0.331	0.228	0.116	0.118	0.396	0.302	-0.289	-0.122	0.171	0.130
DeepSeekV3	0.379	0.311	0.079	0.070	0.429	0.300	-0.067	-0.098	0.203	0.159
Assay2Mol (Gemma-3-27B)	1.277	1.124	1.037	1.121	0.535	0.770	0.554	0.606	1.037	1.069
Assay2Mol (GPT 4o)	1.061	1.046	0.732	0.741	0.223	0.517	0.269	0.151	0.769	0.777
Assay2Mol (DeepSeekV3)	1.042	0.634	0.842	0.921	0.599	0.579	0.267	0.273	0.834	0.849

about the protein from the literature listed on its PDB homepage (Burley et al., 2023).

Methods. Before evaluating Assay2Mol on CrossDocked, we selected its hyperparameters using nine proteins from the CrossDocked training set. We set max_assay_num to 10, N_{mol} to 8, and max_mol_size to 45. We filter out molecules greater than max_mol_size from the input context in the BioAssay data table, which helps control the size of the generated molecules. We test Assay2Mol with three LLMs: Gemma-3-27B (Gemma Team et al., 2025), DeepSeekV3 (DeepSeek-AI et al., 2024), and GPT 4o (OpenAI et al., 2024). Details are in Appendix A.2. To assess whether data leakage affects the results, we run Assay2Mol with an additional filter with DeepSeekV3. After retrieving each BioAssay, we compute the sequence identity between its associated protein and the query protein with MMseqs2 (Steinegger and Söding, 2017). If the sequence identity exceeds 30%, the BioAssay will be discarded and replaced by the next candidate until max_assay_num BioAssays are collected. This result is denoted as Assay2Mol (DeepSeekV3 <30%) in Table 1.

We compare Assay2Mol against the following SBDD methods: CVAE (Ragoza et al., 2022), AR (Luo et al., 2021), Pocket2Mol (Peng et al., 2022), GraphBP (Liu et al., 2022), TamGen (Wu et al., 2024), and TargetDiff (Guan et al., 2023a). Among these, TamGen generates SMILES, whereas the other methods generate 3D molecule conformations. We obtain the previously generated TamGen results from its repository. Docking scripts and other methods' results come from the TargetDiff repository. We use the existing generated molecules from files linked in these repositories directly for evaluation. When rerunning the docking and evaluation scripts, some results may differ from those reported in previous papers due to differences in computing environments. We prioritized comparing to these SBDD methods that had readily available generated molecules or scripts.

The Gemma-3-27B, GPT 4o, and DeepSeekV3

methods are an Assay2Mol ablation that assesses how much of the molecule generation capability comes from the BioAssay context. These methods generate molecules with an LLM using the protein description alone, as shown in Appendix A.9.4.

Metric. We use **Vina Dock** (Eberhardt et al., 2021) to score the binding complementarity of a molecule, or the strength of interaction, with a protein target. **High Affinity** indicates the percentage of generated molecules that outperform the reference molecules in **Vina Dock**. Quantitative Estimate of Drug-likeness (**QED**) is a metric that combines multiple molecular properties (e.g., molecular weight, logP, hydrogen bond donors) into a single value between 0 and 1, with higher values indicating more drug-like compounds (Bickerton et al., 2012). Synthetic Accessibility (**SA**) score estimates synthetic feasibility based on fragment contributions observed in known molecules and structural complexity penalties (Ertl and Schuffenhauer, 2009). We use the normalized SA score between 0 and 1, where 0 is most difficult to synthesize. QED and SA are computed with RDKit (Landrum, 2016). **Diversity** is quantified as the average pairwise Tanimoto distance between Morgan fingerprint of the generated molecules. The Vina score is correlated with the number of atoms in the molecule (Weller and Rohs, 2024), so we also track the **Molecule Size** as the number of heavy atoms. In order to demonstrate the influence of molecule size on docking score, we randomly sample 100 molecules with 30 heavy atoms from ZINC20 (Irwin et al., 2020) and add this baseline to Table 1. We also discuss the molecule **validity** rate and **price** for different LLMs in Appendix A.3.

Metrics are calculated for each protein target. First, we compute the average metrics of the generated molecules for each corresponding protein. Then, we calculate the mean and median scores across all 100 proteins. As an additional baseline that can highlight protein-specific biases in docking scores, we randomly sample 100 FDA-approved drugs.

Results. Most of the existing SBDD methods such as CVAE, AR, and Pocket2Mol do not perform well on average. This may be partially due to the size of the molecules they generate. The random sample of large, irrelevant ZINC molecules with 30 heavy atoms shows that they produce better docking scores than many of the generative models.

All versions of Assay2Mol consistently outperform the best SBDD method, TargetDiff, in average docking scores (Table 1). Assay2Mol (Gemma-3-27B) produces better docking scores than the other Assay2Mol variants, though it generates larger molecules, making it appealing as a locally-run open weights model that can process BioAssay context. The performance of Assay2Mol (DeepSeekV3) remains stable after removing proteins with more than 30% sequence identity. This indicates that the observed performance is not attributable to potential data leakage from the most closely related proteins, but rather reflects the model's ability to generalize.

Beyond improved docking scores, Assay2Mol generates molecules with relatively high synthetic accessibility and QED scores, benefiting from LLMs' molecule generation capability. However, the GPT 4o and DeepSeekV3 versions perform poorly in terms of molecular diversity. We find that LLMs tend to generate similar molecules within a group when context molecules are provided. In extreme cases in our preliminary testing that generated 100 molecules per batch, the model incrementally added a single carbon atom to the molecular backbone each time. The current setting of 10 molecules per batch alleviates this issue and helps balance computational costs and molecule diversity. It is surprising that the three LLMs in the Assay2Mol ablation can generate high-quality molecules in a zero-shot setting, outperforming some SBDD models. These molecules also exhibit desirable drug-likeness and synthetic accessibility characteristics (QED, SA). However, without guidance from the BioAssay context, the generated molecules tend to be smaller in size and exhibit less favorable docking scores than Assay2Mol. We also compute the similarity between generated molecules and context molecules to demonstrate that the LLMs learn from context rather than merely making minor modifications to existing molecules. More details are shown in Appendix A.4.

A more detailed examination of the mouse protein TFPI provides context to Assay2Mol's performance on the CrossDocked dataset (Appendix A.6). It reveals that not all BioAssays with similar embeddings to the target protein query are biologically relevant. This is why we use an LLM to estimate the relevance of the retrieved BioAssays to the query protein. For each query protein, we analyze the top 10 BioAssays after filtering. We run GPT 4o and DeepSeek-V3 and then aggregate their results. We define x as (relevant BioAs-

says)/(total BioAssays). This categorizes the proteins into four groups: high relevance ($x \geq 0.7$), medium relevance ($0.4 < x < 0.7$), low relevance ($0.1 < x \leq 0.4$), and no relevance ($x \leq 0.1$). The results of different groups are shown in Figure 4. As expected, Assay2Mol performs worst on the no relevance group compared with TargetDiff, and it is consistent with our small-scale evaluation of low-similarity BioAssays (Figure 9).

To check the accuracy of the LLMs' relevance evaluation, we manually evaluate the BioAssays retrieved for 25 targets and find that GPT 4o is fairly accurate in its relevance assessment but DeepSeek-V3 struggles (Table 4). DeepSeek-V3 incorrectly assesses the relevance for all ten BioAssays for six different targets. We perform the manual assessment by reading through the assay description and abstract for each of the 10 retrieved AIDs for each target protein (based on PDB code) and confirming whether the assay readout would correspond with a biochemical interaction of the tested compounds with a protein of the same family as the target protein. Also, where applicable we also confirmed the desired functional activity, inhibition or activation. In cases where it was not clear, the AID was labeled as not relevant.

Because docking score distributions differ across proteins, directly averaging scores as in Table 1 may fail to capture meaningful improvements over baseline methods. To account for protein-specific effects, we dock 100 FDA-approved drugs to the 100 CrossDocked test proteins to establish a baseline score distribution. We then compute the improvement in docking score for each generated molecule with respect to these protein-specific background distributions. The improvement of Assay2Mol decreases as BioAssay relevance decreases, while TargetDiff demonstrates relatively uniform performance across all groups, consistent with its reliance solely on protein structure (Table 2). GPT 4o outperforms Assay2Mol (GPT 4o) in the low relevance group, suggesting that the inclusion of irrelevant BioAssay context may mislead the LLM for molecule generation. LLMs without context perform poorly in the no relevance group, indicating that these proteins are possibly understudied and underrepresented in existing databases and their pretraining data. SBDD is a more suitable strategy in such cases. These observations reinforce the validity of the Assay2Mol concept, suggesting that the LLM benefits from assay context to guide molecule generation. Also, most of the proteins in the CrossDocked test set fall into the "High" and "Medium" groups (81%), indicating Assay2Mol's practical utility for potential targets of interest.

4.2 Specificity and counterscreen with hERG

KCNH2, also known as hERG, is a voltage-gated potassium ion channel that plays a crucial role in cardiac repolarization. Blocking hERG channels with drugs can lead to prolonged QT intervals, potentially causing severe cardiac arrhythmias or sudden death. It has been one of the most frequent adverse side effects leading to drug failure (Sanguinetti and Tristani-Firouzi, 2006). As a result, evaluating a molecule's interaction with hERG is a critical step in drug development to ensure safety. In the PubChem database, there are many BioAssays using hERG as a counterscreen (Garrido et al., 2020). We ask:

1. Can Assay2Mol accurately interpret the counterscreen context to reduce the generated molecule's affinity for hERG?
2. Can the generated molecules retain high affinity to the original target protein?

We selected proteins GRK4 (PDB: 4YHJ), HD8 (PDB: 4RN0), and CD38 (PDB: 3DZH) from the CrossDocked test set. These proteins serve as potential targets for cancer treatments or antibiotics. Since the corresponding molecules may potentially be developed as drugs taken by humans, it is crucial to evaluate their potential hERG-related interactions to assess safety. After we generate molecules for these proteins using the same methods as in the CrossDocked evaluation, we use the description of hERG[3] to search for related BioAssays, in the same manner used to formulate the context described in Section 3. After we prepare the whole context, we append generated molecules from the previously selected proteins and ask the LLM to optimize these molecules to reduce binding affinity toward hERG. The prompt is shown in Appendix A.9.5. We use ADMETlab 3.0 (Fu et al., 2024) to predict the hERG score. To avoid circular reasoning, we verify that no molecules in the hERG generation context were present in the ADMETlab 3.0 hERG training set.

We examine the shift in Vina docking scores and predicted hERG scores for the original generated molecules versus those optimized to minimize hERG interaction (Figure 3) The average hERG

[3] https://pubchem.ncbi.nlm.nih.gov/gene/3757

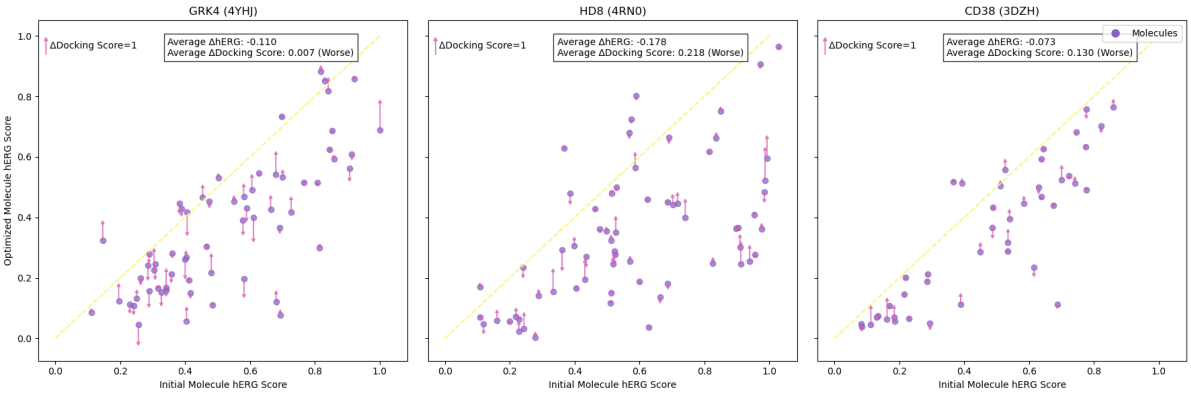

Figure 3: Change in predicted hERG score and docking score between initial and optimized molecules for three proteins. The up arrow indicates the docking score decreases and the down arrow indicates it increases. The length of the arrow (top-left) serves as a scale bar, representing an increase of 1 score unit (kcal/mol) from Vina Dock.

score of molecules for 4YHJ decreases from 0.503 before optimization to 0.393 after optimization. For comparison, the average hERG score of 2,965 FDA-approved drugs is 0.284. This reduction in hERG score indicates lower predicted cardiotoxicity, making the generated molecules more suitable for further study, while docking scores remain largely unaffected, demonstrating that in-context learning with a counterscreen enhances specificity without compromising affinity.

5 Discussion and conclusions

Our initial version of Assay2Mol can successfully query PubChem with text descriptions, retrieve relevant BioAssays, and generate molecules based on the text and screening data from those relevant BioAssays. Assay2Mol shows how to use LLMs to make better use of decades' worth of valuable, unstructured chemical screening data in PubChem. Our approach generates molecules with docking scores comparable to SBDD methods and advances controllable natural language-driven molecule design. There are many opportunities to expand on the core Assay2Mol framework, making it more robust and building new capabilities to make it more relevant for the difficult multi-property optimization required for actual drug discovery campaigns (van den Broek et al., 2025). For example, we observed that the initial embedding-based query and BioAssay similarity calculations are not sophisticated enough to capture the complexity of biological regulation in pathways (Appendix A.6) along with other limitations in Section 6.

6 Limitations

There are several opportunities to improve how LLMs are used within Assay2Mol. Having LLMs directly assess the relevance of the retrieved BioAssay text guards against many irrelevant matches but is imperfect (Table 4). We continue to explore methods to enhance LLM interpretation of the desired activity of generated molecules with respect to target function (i.e., activation, inhibition, allosteric regulation). A related limitation is that the current version of Assay2Mol cannot properly process conditional text queries, such as molecules that inhibit proteins A, B, and C but not D and E. This limitation is again related to the initial embedding-based similarity calculations, which could possibly be addressed with enhancement to the Assay2Mol relevance processing step. Our hERG example shows that Assay2Mol can be used for conditional molecule generation, for example inhibiting GRK4 but not hERG, if these steps are run sequentially. Furthermore, there is potential for improvement in both construction of text prompts and the LLMs used within the Assay2Mol framework. Our BioAssay relevance assessment evaluation showed that GPT 4o matched manual assessments much more closely than DeepSeek-V3 (Table 4), so the choice of LLM can impact the overall results. Finally, we have not yet evaluated the sensitivity of Assay2Mol to different LLM prompting strategies or optimized the prompts.

LLMs are pre-trained on a large-scale text corpus, which includes a substantial number of molecules. As a result, when generating molecules, LLMs may not always create new molecules. Instead, they tend to "retrieve" molecules or recom-

bine patterns from molecules they have trained on. Repurposing existing molecules for new targets or modifying existing molecules means that they tend to exist in certain databases, be available for purchase, or be more feasible to synthesize chemically. However, for researchers whose primary goal is to push beyond known chemical space, this characteristic might be regarded as a limitation rather than an advantage.

Most of the LLM-based steps in Assay2Mol are implemented using both closed and open weights LLMs with the goal of having a fully open weight version of Assay2Mol. However, currently the BioAssay embeddings are generated solely using the OpenAI text embedding API. An alternative implementation and evaluation with open weights embedding models remains for future work. Even the open weights LLMs Assay2Mol uses are not fully open source and do not have their training data available. This lack of training data and what biochemistry data are included makes it challenging to fully interpret the Assay2Mol-generated molecules and their limitations. If it is generating novel molecules as opposed to retrieving molecules from the LLM training set, general caveats about generative molecular design apply to those outputs (Walters, 2024).

The evaluations of the Assay2Mol generated molecules rely entirely on other computational assessments of molecule quality. These are not a substitute for actual wet lab assays. Vina Dock energies are not true binding affinities, and its scoring function has known biases and limitations (Xu et al., 2022). For instance, we found that randomly sampled ZINC molecules with 30 heavy atoms produced docking scores comparable to Assay2Mol (Table 1), illustrating the known relationship between docking score and molecule size (Weller and Rohs, 2024). In addition, the hERG scores are computed with an existing regressor that has reasonably good but imperfect performance (Fu et al., 2024).

Software availability

The code is available at `https://github.com/gitter-lab/Assay2Mol` under the MIT License and archived at `https://doi.org/10.5281/zenodo.15871304`. The datasets are available at `https://doi.org/10.5281/zenodo.15867214` under the CC BY 4.0 license.

Acknowledgments

This research was supported by National Institutes of Health award R01GM135631.

References

Nawaf Alampara, Anagha Aneesh, Martiño Ríos-García, Adrian Mirza, Mara Schilling-Wilhelmi, Ali Asghar Aghajani, Meiling Sun, Gordan Prastalo, and Kevin Maik Jablonka. 2025. General purpose models for the chemical sciences. *arXiv:2507.07456*.

Jonathan B. Baell and J. Willem M. Nissink. 2018. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. *ACS Chemical Biology*, 13(1):36–44. Publisher: American Chemical Society.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. 2012. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98. Publisher: Nature Publishing Group.

Regine Bohacek, Colin Mcmartin, Peter Glunz, and Daniel H. Rich. 1999. Growmol, A De novo Computer Program, and its Application to Thermolysin and Pepsin: Results of the Design and Synthesis of a Novel Inhibitor. In Donald G. Truhlar, W. Jeffrey Howe, Anthony J. Hopfinger, Jeff Blaney, and Richard A. Dammkoehler, editors, *Rational Drug Design*, volume 108 of *The IMA Volumes in Mathematics and its Applications*, pages 103–114. Springer, New York, NY.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165 [cs].

Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Henry Chao, Li Chen, Paul A Craig, Gregg V Crichlow, Kenneth Dalenberg, Jose M Duarte, Shuchismita Dutta, Maryam Fayazi, Zukang Feng, Justin W Flatt, Sai Ganesan, Sutapa Ghosh, David S Goodsell, Rachel Kramer Green, Vladimir Guranovic, Jeremy Henry, Brian P Hudson, Igor Khokhriakov, Catherine L Lawson, Yuhe Liang, Robert Lowe, Ezra Peisach, Irina Persikova, Dennis W Piehl, Yana Rose, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, Brinda Vallat, Maria Voigt, Ben Webb, John D Westbrook, Shamara Whetstone, Jasmine Y Young, Arthur Zalevsky, and Christine Zardecki. 2023. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined

PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research*, 51(D1):D488–D508.

Julian Cremer, Ross Irwin, Alessandro Tibo, Jon Paul Janet, Simon Olsson, and Djork-Arné Clevert. 2025. FLOWR: Flow Matching for Structure-Aware De Novo, Interaction- and Fragment-Based Ligand Generation. *arXiv:2504.10564*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. DeepSeek-V3 Technical Report. *arXiv:2412.19437*.

Yifan Deng, Spencer S. Ericksen, and Anthony Gitter. 2025. Chemical Language Model Linker: Blending Text and Molecules with Modular Adapters. *Journal of Chemical Information and Modeling*, 65(17):8944–8956. Publisher: American Chemical Society.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The Faiss library. *arXiv:2401.08281*.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. 2021. AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Ertl and Ansgar Schuffenhauer. 2009. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv:2306.08018*.

Christopher Fifty, Jure Leskovec, and Sebastian Thrun. 2023. In-Context Learning for Few-Shot Molecular Property Prediction. *arXiv:2310.08863*.

Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. 2020. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *Journal of Chemical Information and Modeling*, 60(9):4200–4215. Publisher: American Chemical Society.

Li Fu, Shaohua Shi, Jiacai Yi, Ningning Wang, Yuanhang He, Zhenxing Wu, Jinfu Peng, Youchao Deng, Wenxuan Wang, Chengkun Wu, Aiping Lyu, Xiangxiang Zeng, Wentao Zhao, Tingjun Hou, and Dongsheng Cao. 2024. ADMETlab 3.0: an updated comprehensive online admet prediction platform enhanced with broader coverage, improved performance, api functionality and decision support. *Nucleic Acids Research*, 52(W1):W422–W431.

Bowen Gao, Yanwen Huang, Yiqiao Liu, Wenxuan Xie, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. 2025. PharmAgents: Building a Virtual Pharma with Large Language Model Agents. *arXiv:2503.22164*.

Amanda Garrido, Alban Lepailleur, Serge M. Mignani, Patrick Dallemagne, and Christophe Rochais. 2020. herg toxicity assessment: Useful guidelines for drug design. *European Journal of Medicinal Chemistry*, 195:112290.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C. J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *arXiv:2503.19786*.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 2023a. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. 2023b. DecompDiff: Diffusion Models with Decomposed Priors for Structure-Based Drug Design. In *Proceedings of the 40th International Conference on Machine Learning*, pages 11827–11846. PMLR. ISSN: 2640-3498.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Herim Han, Bilal Shaker, Jin Hee Lee, Sunghwan Choi, Sanghee Yoon, Maninder Singh, Shaherin Basith, Minghua Cui, Sunil Ahn, Junyoung An, Soosung Kang, Min Sun Yeom, and Sun Choi. 2025. Employing automated machine learning (AutoML) methods to facilitate the in silico ADMET properties prediction. *Journal of Chemical Information and Modeling*, 65(7):3215–3225.

John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. 2020. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169. Publisher: Nature Publishing Group.

Yuanyuan Jiang, Guo Zhang, Jing You, Hailin Zhang, Rui Yao, Huanzhang Xie, Liyun Zhang, Ziyi Xia, Mengzhe Dai, Yunjie Wu, Linli Li, and Shengyong Yang. 2024. PocketFlow is a data-and-knowledge-driven structure-based molecular generative model. *Nature Machine Intelligence*, 6(3):326–337. Publisher: Nature Publishing Group.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. 2024. PubChem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525.

Greg Landrum. 2016. RDKit: Open-source cheminformatics software.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. 2022. Generating 3D Molecules for Target Protein Binding. In *Proceedings of the 39th International Conference on Machine Learning*, pages 13912–13924. PMLR. ISSN: 2640-3498.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2025. DrugAgent: Automating AI-aided Drug Discovery Programming through LLM Multi-Agent Collaboration. *arXiv:2411.15692*.

Yuyan Liu, Sirui Ding, Sheng Zhou, Wenqi Fan, and Qiaoyu Tan. 2024. MolecularGPT: Open Large Language Model (LLM) for Few-Shot Molecular Property Prediction. *arXiv:2406.12950*.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. 2021. A 3D Generative Model for Structure-Based Drug Design. In *Advances in Neural Information Processing Systems*, volume 34, pages 6229–6239. Curran Associates, Inc.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

Andrew D. McNaughton, Gautham Krishna Sankar Ramalaxmi, Agustin Kruel, Carter R. Knutson, Rohith A. Varikoti, and Neeraj Kumar. 2024. Cactus: Chemistry agent connecting tool usage to science. *ACS Omega*, 9(46):46563–46573.

John H. McVey. 1999. Tissue Factor pathway. *Best Practice & Research Clinical Haematology*, 12(3):361–372.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. 2018. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Mehrdad Asgari, Juliane Eberhardt, Amir Mohammad Elahi, Hani M. Elbeheiry, María Victoria Gil, Christina Glaubitz, Maximilian Greiner, Caroline T. Holick, Tim Hoffmann, Abdelrahman Ibrahim, Lea C. Klepsch, Yannik Köster, Fabian Alexander Kreth, Jakob Meyer, Santiago Miret, Jan Matthias Peschel, Michael Ringleb, Nicole C. Roesner, Johanna Schreiber, Ulrich S. Schubert, Leanne M. Stafast, A. D. Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. 2025. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17(7):1027–1034. Publisher: Nature Publishing Group.

Saeed Moayedpour, Alejandro Corrochano-Navarro, Faryad Sahneh, Shahriar Noroozizadeh, Alexander Koetter, Jiri Vymetal, Lorenzo Kogler-Anele, Pablo Mas, Yasser Jangjou, Sizhen Li, Michael Bailey, Marc Bianciotto, Hans Matter, Christoph Grebner, Gerhard Hessler, Ziv Bar-Joseph, and Sven Jager. 2024. Many-Shot In-Context Learning for Molecular Inverse Design. *arXiv:2407.19089*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and Code Embeddings by Contrastive Pre-Training. *arXiv:2201.10005*.

Tung Nguyen and Aditya Grover. 2025. LICO: Large Language Models for In-Context Molecular Optimization. In *The Thirteenth International Conference on Learning Representations*.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson,

Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. GPT-4o System Card. *arXiv:2410.21276*.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. BioT5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1216–1240, Bangkok, Thailand. Association for Computational Linguistics.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. 2022. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17644–17655. PMLR. ISSN: 2640-3498.

Yanru Qu, Keyue Qiu, Yuxuan Song, Jingjing Gong, Jiawei Han, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. 2024. MolCRAFT: structure-based drug design in continuous parameter space. In *Proceedings of the 41st International Conference on Machine Learning*.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. 2022. Generating 3D molecules conditional on receptor binding sites with deep generative models.

Chemical Science, 13(9):2701–2713. Publisher: The Royal Society of Chemistry.

Mayk Caldas Ramos, Christopher J. Collison, and Andrew D. White. 2025. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 16(6):2514–2572. Publisher: The Royal Society of Chemistry.

Michael C. Sanguinetti and Martin Tristani-Firouzi. 2006. hERG potassium channels and cardiac arrhythmia. *Nature*, 440(7083):463–469. Publisher: Nature Publishing Group.

Johannes Schimunek, Sohvi Luukkonen, and Günter Klambauer. 2025. MHNfs: Prompting In-Context Bioactivity Predictions for Low-Data Drug Discovery. *Journal of Chemical Information and Modeling*, 65(9):4243–4250. Publisher: American Chemical Society.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom Blundell, Pietro Lio, Max Welling, Michael Bronstein, and Bruno Correia. 2024. Structure-based Drug Design with Equivariant Diffusion Models. *arXiv:2210.13695*.

Linde Schoenmaker, Enzo G. Sastrokarijo, Laura H. Heitman, Joost B. Beltman, Willem Jespers, and Gerard J.P. van Westen. 2025. Toward Assay-Aware Bioactivity Model(er)s: Getting a Grip on Biological Context. *Journal of Chemical Information and Modeling*, 65(13):7013–7023. Publisher: American Chemical Society.

Maximilian G. Schuh, Davide Boldini, and Stephan A. Sieber. 2024. Synergizing chemical structures and bioassay descriptions for enhanced molecular property prediction in drug discovery. *Journal of Chemical Information and Modeling*, 64(12):4640–4650.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. 2023. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *Proceedings of the 40th International Conference on Machine Learning*.

Sejal Sharma, Liping Feng, Nicha Boonpattrawong, Arvinder Kapur, Lisa Barroilhet, Manish S. Patankar, and Spencer S. Ericksen. 2024. Data mining of PubChem bioassay records reveals diverse OXPHOS inhibitory chemotypes as potential therapeutic agents against ovarian cancer. *Journal of Cheminformatics*, 16(1):112.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. 2020. GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation. *arXiv:2001.09382*.

Ines Smit, Melissa F. Adasme, Emma Manners, Sybilla Corbett, Nicolas Bosc, Hoang-My-Anh Do, Andrew R. Leach, Noel M. O'Boyle, and Barbara Zdrazil. 2025. Enhancing Bioassay Annotations in ChEMBL with Artificial Intelligence. *ChemRxiv*.

Jacob O. Spiegel and Jacob D. Durrant. 2020. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *Journal of Cheminformatics*, 12(1):25.

Martin Steinegger and Johannes Söding. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028. Publisher: Nature Publishing Group.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. 2025. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, pages 1–8. Publisher: Nature Publishing Group.

The UniProt Consortium. 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

Remco L. van den Broek, Shivam Patel, Gerard J. P. van Westen, Willem Jespers, and Woody Sherman. 2025. In Search of Beautiful Molecules: A Perspective on Generative Modeling for Drug Design. *Journal of Chemical Information and Modeling*. Publisher: American Chemical Society.

Pat Walters. 2024. Generative Molecular Design Isn't As Easy As People Make It Look.

Ziqing Wang, Kexin Zhang, Zihan Zhao, Yibo Wen, Abhishek Pandey, Han Liu, and Kaize Ding. 2025. A Survey of Large Language Models for Text-Guided Molecular Discovery: from Molecule Generation to Optimization. *arXiv:2505.16094*.

Ziming Wei, Yasha Ektefaie, Andrew Zhou, Dereje Negatu, Bree Aldridge, Thomas Dick, Michael Skarlinski, Andrew D. White, Samuel Rodriques, Sepideh Hosseiniporgham, Inna Krieger, James Sacchettini, Marinka Zitnik, and Maha Farhat. 2025. Fleming: An AI Agent for Antibiotic Discovery in Mycobacterium tuberculosis. *bioRxiv*, page 2025.04.01.646719.

Jesse A. Weller and Remo Rohs. 2024. Structure-Based Drug Design with a Deep Hierarchical Generative Model. *Journal of Chemical Information and Modeling*, 64(16):6450–6463. Publisher: American Chemical Society.

Kehan Wu, Yingce Xia, Pan Deng, Renhe Liu, Yuan Zhang, Han Guo, Yumeng Cui, Qizhi Pei, Lijun Wu, Shufang Xie, Si Chen, Xi Lu, Song Hu, Jinzhi Wu, Chi-Kin Chan, Shawn Chen, Liangliang Zhou, Nenghai Yu, Enhong Chen, Haiguang Liu, Jinjiang Guo, Tao Qin, and Tie-Yan Liu. 2024. TamGen: drug design with target-aware molecule generation through a chemical language model. *Nature Communications*, 15(1):9360. Publisher: Nature Publishing Group.

Min Xu, Cheng Shen, Jincai Yang, Qing Wang, and Niu Huang. 2022. Systematic Investigation of Docking Failures in Large-Scale Structure-Based Virtual Screening. *ACS Omega*, 7(43):39417–39428. Publisher: American Chemical Society.

Jiaxian Yan, Zhaofeng Ye, Ziyi Yang, Chengqiang Lu, Shengyu Zhang, Qi Liu, and Jiezhong Qiu. 2023. Multi-task bioassay pre-training for protein-ligand binding affinity prediction. *Briefings in Bioinformatics*, 25(1):bbad451.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, Kehua Feng, Xiang Zhuang, Zeyuan Wang, Ming Qin, Mengyao Zhang, Jinlu Zhang, Jiyu Cui, Tao Huang, Pengju Yan, Renjun Xu, Hongyang Chen, Xiaolin Li, Xiaohui Fan, Huabin Xing, and Huajun Chen. 2024. Scientific Large Language Models: A Survey on Biological & Chemical Domains. *arXiv:2401.14656*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Xin Chen, and Kai Yu. 2024. ChemDFM: Dialogue Foundation Model for Chemistry. *arXiv:2401.14818*.

A Appendix

A.1 Potential risks

Like many other molecule generation tools, Assay2Mol could be considered a dual-use technology. Its primary intended application is in early stage drug discovery as a strategy to sift through unstructured experimental BioAssay descriptions in PubChem. However, BioAssays in PubChem used for counterscreening molecules measure toxicity, so it could be possible to use Assay2Mol to generate toxic molecules. Existing generative models can also be trained or guided with this public toxicity data, so this is a risk of a broad class of models not Assay2Mol specifically.

Assay2Mol generates molecules tailored to activity against the target protein or phenotype. Because the prompts are constructed in a structured manner using BioAssay context, it is less likely to be able to generate some classes of harmful molecules like explosives compared to generative models that take free text descriptions of molecule properties as input. LLMs perform the actual molecule generation, and we have not formally studied how the BioAssay context expands or reduces risk relative to the baseline LLMs. We anticipate that Assay2Mol's structured prompts restrict the types of molecules the LLMs can generate, and some of the generated molecules are actually retrieved existing molecules.

A.2 Hyperparameter and model selection

For the Assay2Mol hyperparameter selection, we choose nine proteins from the randomly sampled 100 proteins from the training set: 1RQP, 3ANT, 4A9S, 4I29, 4XE6, 4Y2R, 5ACC, 5AI4, 5PNX. We manually inspect their retrieved BioAssays and make sure they are relevant. We also include the reference molecule and label it as "CrossDocked". Assay2Mol invloves three hyperparameters: max_assay_num, N_{mol}, and max_mol_size, making it time-consuming to perform grid-search. To simplify the process, we fix two hyperparameters when optimizing the remaining one.

First, we set N_{mol} to 5, impose no limit on max_mol_size, and vary max_assay_num across [1, 5, 10, 15, 20, 30]. According to Figure 5, we observe that $max\_assay\_num = 10$ yields strong performance. More than 10 BioAssays leads to longer context and increases computational cost without significant gains.

After max_assay_num is fixed to 10, we vary N_{mol} across [1, 3, 5, 8, 12] and find $N_{mol} = 8$ yields the best results, as shown in Figure 6. Finally, we fix max_assay_num to 10 and N_{mol} to 8 and examine max_mol_size's influence on the generated molecules. We do not explicitly prompt the LLM to generate molecules of a specific size. The generated molecules' size and the docking score both correlate with max_mol_size, as shown in Figure 7. Therefore, we set max_mol_size to different values as a way to approximately control the generated molecules' sizes. Additional figures used to guide the hyperparameter tuning are available in our GitHub repository.

For LLM selection, we first select GPT 4o as the base model, a closed weights model. Then, we select DeepSeekV3, an open weights model, so that Assay2Mol does not rely solely on proprietary LLMs. Additionally, we incorporate Gemma-3-27B, a high-performing open weights model that can be locally deployed on a single GPU. For the motivating example, we use the ChatGPT 4o web version. For GPT 4o, we use the version "chatgpt-4o-latest" (as of May 2025) for Assay2Mol experiments and version "gpt-4o-2024-11-20" for LLM BioAssay relevance assessment (Table 4), which was completed earlier. For DeepSeekV3, we use the version "DeepSeek-V3-0325" for Assay2Mol and "DeepSeek-V3-1226" for relevance assessment (Table 4). For Gemma-3-27B, we run the 4-bit bitsandbytes quantized version provided by unsloth[4] on a single RTX 5090 GPU.

A.3 LLM molecule validity and price

Though DeepSeekV3 outperforms GPT 4o in Table 1, DeepSeekV3 generates far more invalid molecules than GPT 4o. We define validity as the percentage of unique generated SMILES that are parsable by RDKit. We list the validity and API price of different LLMs in Table 3.

A.4 Similarity analysis

Given the molecules provided in the BioAssay context, one natural question arises: Does the LLM simply copy and slightly modify the active molecules, rather than generating novel ones? To address this question, we dock the context molecules against five proteins (4AAW, 4YHJ, 14GS, 2V3R, 4RN0) and analyze the similarity between the generated molecules and the context molecules. We select context molecules

[4] https://huggingface.co/unsloth/gemma-3-27b-it-unsloth-bnb-4bit

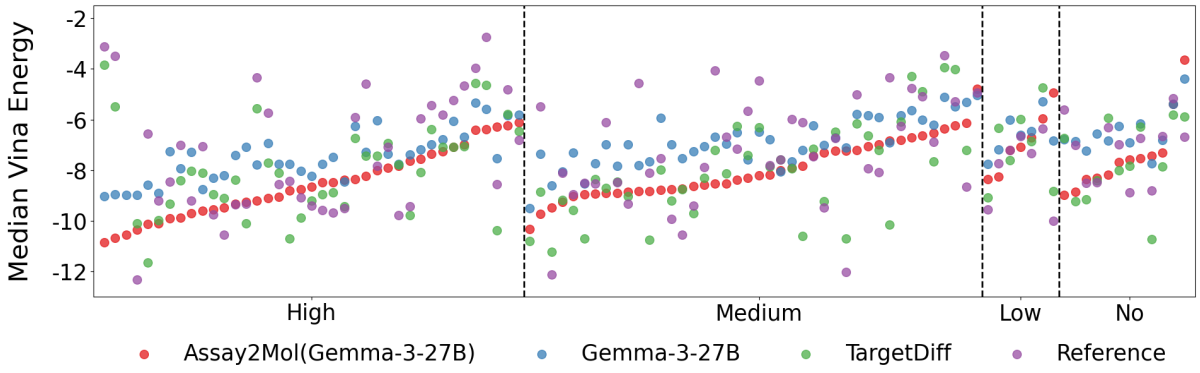

Figure 4: Distribution of docking scores of different relevance level groups.

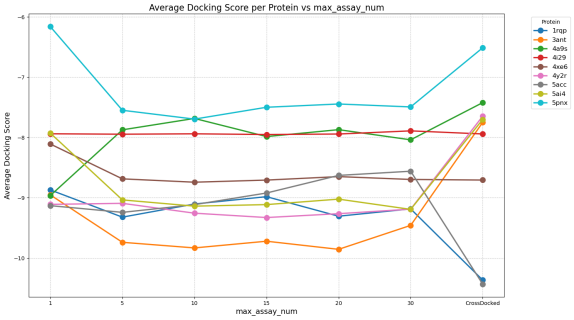

Figure 5: Average docking score under different max_assay_num

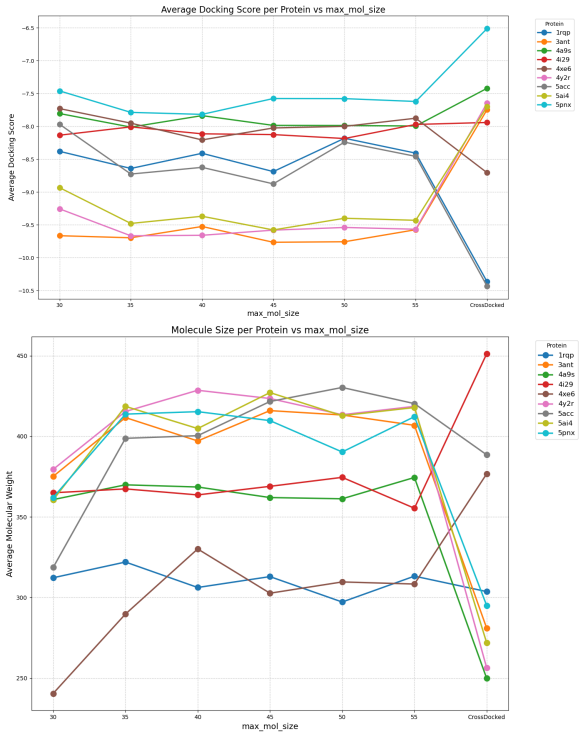

Figure 7: Average docking score and molecule weight under different max_mol_size

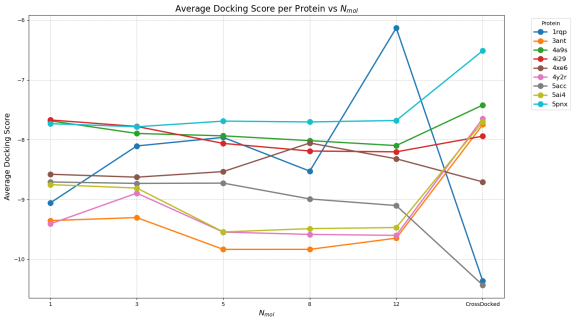

Figure 6: Average docking score under different N_{mol}

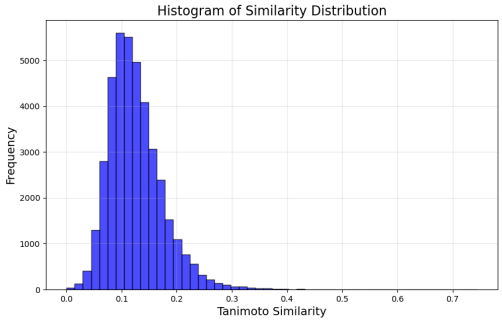

Figure 8: Similarity distribution between generated molecules and high docking score context molecules.

LLM	Validity	Input Price	Output Price
Gemma-3-27B	84.32	N/A	N/A
DeepSeekV3	75.84	$0.07/$0.27	$1.1
GPT 4o	94.33	$2.5/$5	$20

Table 3: Validity of generated molecules and API price of different LLMs. The price is for 1M input/output tokens. The numbers in input price represent the cost of generation with a cache hit and cache miss, respectively. The price for Gemma-3-27B is not applicable because it is run on a local machine.

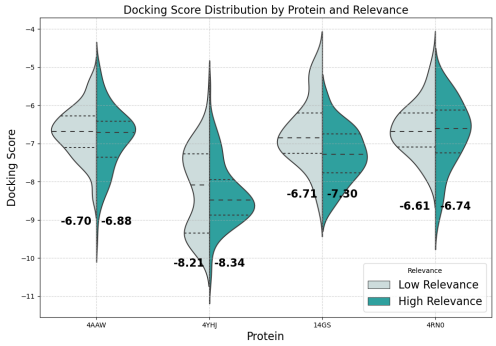

Figure 9: Distribution of docking scores across four proteins using high- and low-relevance BioAssays as context. The average docking score is shown on the top of each plot.

with docking scores more favorable than the reference molecules and consider them to be high-scoring context molecules. Then, we compute the Tanimoto similarity between all pairs of the high-scoring context molecules and the generated molecules using Morgan fingerprints as features (Figure 8). Most generated molecules are dissimilar to the high-scoring context molecules, suggesting that the LLM is learning from the context rather than simply making slight modifications to the context molecules.

A.5 BioAssay context relevance experiment

Table 2 already shows that Assay2Mol performs worse on proteins that do not have relevant BioAssays. We further investigate the impact of context relevance by substituting relevant BioAssays with irrelevant ones. Specifically, we select four proteins for which relevant BioAssays can be retrieved: PDB IDs 4AAW, 4YHJ, 14GS, and 4RN0. During the retrieval stage, we replace the BioAssays with those having relevance scores (measured in cosine similarity) ranging from 0.4 to 0.5, scores much lower than the relevant BioAssays, while keeping the rest of the Assay2Mol pipeline unchanged. The result is shown in Figure 9. Although high-relevance BioAssays consistently outperform low-relevance BioAssays on average, the difference is not particularly pronounced. We suspect this is because GPT 4o also demonstrates the capability to generate molecules solely based on protein descriptions.

To investigate further, we analyzed the context molecules from the irrelevant BioAssays selected by GPT 4o as templates and examined their docking scores toward the query protein. Across the targets 4AAW, 4RN0, 14GS, and 4YHJ the average value of the docking scores are -6.61, -6.04, -5.91, and -6.69, respectively. Even when the BioAssays were irrelevant to the query protein, these context molecules can exhibit moderately high docking scores, making it less surprising that molecules generated from that context yield comparable docking scores.

However, we noticed that the lower tails of the violin plots representing high-relevance molecules tend to extend longer than those of the low-relevance group (Figure 9), indicating that the best docking scores are more frequently found among high-relevance molecules. When using docking for hit identification, researchers focus on the top-scoring molecules as opposed to the entire distribution because these are most likely to be enriched for actual active molecules. Thus, we extended the experiment to 18 proteins and only focused on the molecules with the 10 best docking scores per model. We found that in 11 out of 18 proteins, the high-relevance group performed better than both the low-relevance group and the without context group, which is the GPT 4o baseline (Figure 10). In five proteins (4YHJ, 3DAF, 3DZH, 1COY, 2PQW), the high-relevance group performed similarly to other groups. In two proteins (2JJG, 3GS6), the high-relevance group performed worse than other groups.

These results suggest that in some cases, molecules derived from low-relevance BioAssays can still achieve favorable docking scores either by chance or due to biases in the docking software. Nonetheless, molecules from high-relevance BioAssays generally demonstrate superior performance when focusing on the tail of the docking score distribution.

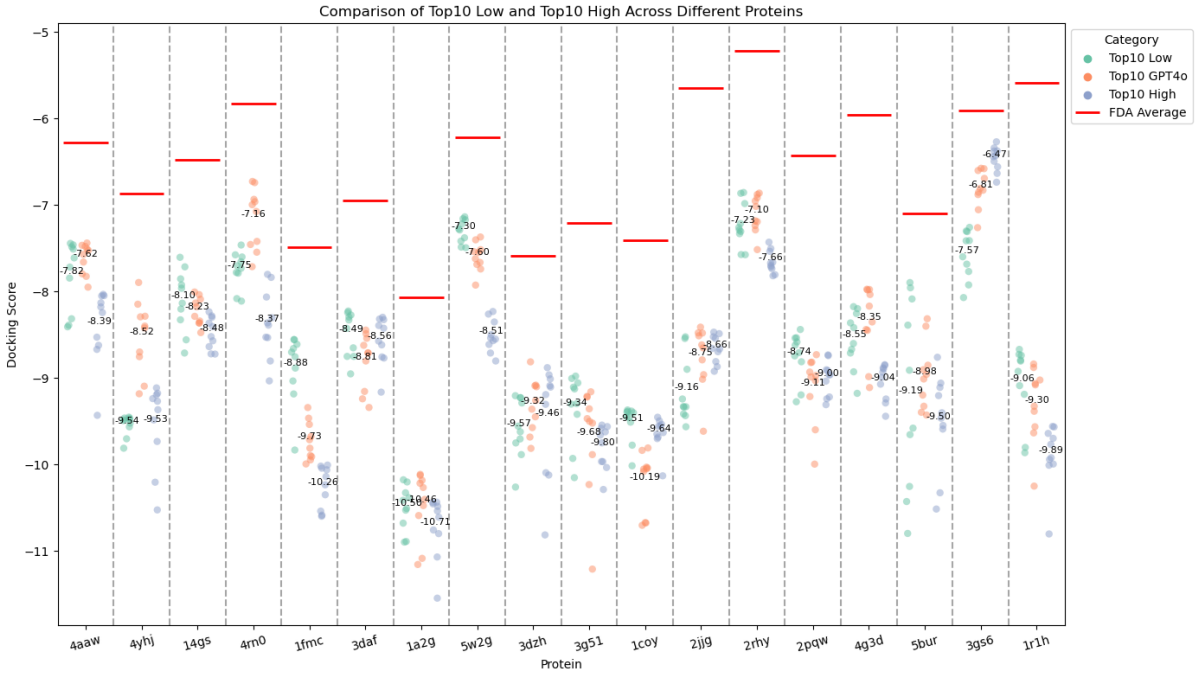

Figure 10: Distribution of the top 10 docking scores from molecules with high- and low-relevance BioAssays as context for different proteins.

A.6 Antagonistic protein pair

During the exploration of BioAssay retrieval, we found a scenario that is beyond the current capabilities of Assay2Mol. We tested Assay2Mol with recent, real drug targets by selecting the most recently FDA-approved drug of 2024[5], concizumab-mtci. This drug is an inhibitor of the human Tissue Factor Pathway Inhibitor (TFPI)[6] protein. Using the protein description of TFPI from PubChem, we performed the BioAssay retrieval task. The top-matching BioAssay[7] related to Tissue Factor (TF) and Coagulation Factor VII (FVII). Part of the retrieval result is shown in Table 5. TFPI, TF, and FVII are key regulators of the extrinsic coagulation pathway, which is responsible for initiating blood clotting (McVey, 1999). TF initiates coagulation, FVII promotes the process, and TFPI inhibits it, maintaining the dynamic balance of the coagulation system. TF inhibitors are designed to treat hypercoagulability and thrombosis. TFPI inhibitors are used for treating hemophilia A and B. Since TF and TFPI have antagonistic functions, an incorrect selection of BioAssays leads to misaligned molecule design. When designing an inhibitor for TFPI, retrieving BioAssays targeting TF directs the molecule design in the opposite direction. Among the 300 retrieved BioAssays, only four directly relate to TFPI, and their target protein is mouse TFPI. The majority are associated with TF, the functional opposite of TFPI.

Although reasoning LLMs (for example, GPT o1 and DeepSeek R1) recognize the antagonistic relationship between the retrieved BioAssays and the query protein—allowing them to discard most mismatched results—the process is not completely accurate. Some of the 296 TF BioAssays are still incorrectly labeled as relevant to TFPI, representing false positives. The LLMs still learn from inhibitors of the opposing protein, leading to failure in molecule design. While the CrossDocked dataset contains no such cases, developing a more robust retrieval pipeline remains a priority for future work.

A.7 Software

We use the hERG Blocker classifier from the ADMETlab 3.0 web server (Fu et al., 2024) to predict the hERG score, which is in the range [0,1].
We use AutoDock Vina (Eberhardt et al., 2021) v1.5.7 as the docking software. We use the original ligand center from the protein-ligand complex as the docking center, set the box size to 20 Å in each

[5] https://www.fda.gov/drugs/novel-drug-approvals-fda/novel-drug-approvals-2024

[6] https://pubchem.ncbi.nlm.nih.gov/protein/P10646

[7] https://pubchem.ncbi.nlm.nih.gov/bioassay/1381622

Relevance	Target	GPT 4o errors	GPT 4o accuracy (%)	DeepSeek-V3 errors	DeepSeek-V3 accuracy (%)	GPT 4o accuracy – DeepSeek-V3 accuracy (%)
high	5W2G	1	90	1	90	0
high	3G51	4	60	10	0	60
high	1COY	3	70	9	10	60
high	2JJG	5	50	9	10	40
high	2RHY	1	83	3	50	33
high	2PQW	0	100	2	67	33
high	4G3D	5	50	5	50	0
medium	4AAW	4	60	6	40	20
medium	4YHJ	0	100	1	90	10
medium	14GS	1	90	9	10	80
medium	4RN0	1	90	4	60	30
medium	1FMC	0	100	10	0	100
medium	3DAF	0	100	10	0	100
medium	1A2G	0	100	10	0	100
medium	3DZH	0	100	8	20	80
medium	5BUR	0	100	10	0	100
low	1R1H	0	100	7	30	70
low	5B08	0	100	3	70	30
low	5I0B	1	90	4	60	30
low	3KC1	0	100	4	60	40
low	1D7J	1	90	4	60	30
no	2Z3H	0	100	0	100	0
no	2V3R	0	100	0	100	0
no	3B6H	8	20	10	0	20
no	4P6P	0	100	0	100	0
total	25	35	86	139	43	43

Table 4: Manual review of LLM BioAssay relevance assessment by a single expert computational chemist (author S.S.E.). All targets have 10 BioAssays retrieved except for 2RHY and 2PQW, which have 6 each. The reason is there are only 6 BioAssays remained after filtering.

BioAssay	Score	Target	Relationship
1381622	0.8284	Tissue factor, Coagulation factor VII	Opposite
51307	0.8245	Tissue factor, Coagulation factor VII	Opposite
397857	0.8233	Tissue factor, Coagulation factor VII	Opposite
385437	0.8202	Tissue factor	Opposite
385438	0.8194	Tissue factor	Opposite
385435	0.8190	Tissue factor	Opposite
385436	0.8188	Tissue factor	Opposite
1871821	0.8183	Coagulation factor VII	Opposite
360094	0.8182	Tissue factor	Opposite
385434	0.8180	Tissue factor	Opposite
		⋮	
1620662	0.7941	Tissue factor pathway inhibitor	Correct
1620663	0.7903	Tissue factor pathway inhibitor	Correct
		⋮	
1476906	0.7888	Coagulation factor XII	Opposite
1775843	0.7888	Carboxypeptidase B2	Opposite
304051	0.7888	Carboxypeptidase B2	Opposite
52023	0.7888	Coagulation factor X	Opposite
1426475	0.7888	Coagulation factor X	Opposite
1775842	0.7888	Carboxypeptidase B2	Opposite
212091	0.7887	Tissue factor	Opposite
362212	0.7887	TISSUE: Plasma	Opposite
1382994	0.7886	Coagulation factor X	Opposite
1657762	0.7886	TISSUE: Plasma	Opposite

Table 5: BioAssay retrieval result using the TFPI description as input. The Score is measured with cosine similarity. Only four out of 300 BioAssays are related to TFPI while others have the opposite function in the pathway.

dimension (x, y, z), and set the exhaustiveness to 32.

A.8 Assay summarization example

Summarized BioAssay Example

This BioAssay measures the inhibition of GRK5-mediated phosphorylation of rhodopsin in bovine rod outer segment membranes under white light conditions. It evaluates the efficacy of cyclic peptide inhibitors derived from the HJ loop of GRK2, providing insights into their potency and selectivity.
This is the activity data table. Each line has the SMILES, followed by activity type (active, inactive or unspecified) and the experimental value.
`[C@H](C(=O)N1)CCCCN))...(N)N)CC(C)C`
Unspecified Inhibition <5%
`CC[C@H](C)[C@@H](C(=O)O)...NC(=O)CN`
Unspecified Inhibition <5%

This BioAssay evaluates the inhibition of recombinant human GRK5 expressed in Sf9 insect cells, measuring the decrease in phosphorylation of urea-washed bovine rod outer segments in the presence of Gbetagamma subunits and [gamma-32P]-ATP.
This is the activity data table. Each line has the SMILES, followed by activity type (active, inactive or unspecified) and the experimental value.
`CC[C@H](C)[C@@H](C(=O)N...(CCCCN)N`
Active IC50 = 2100 nM
`C1=CC=C(C=C1)/C=C\2/C3=CC=CC=C3C(=O)N2`
Inactive IC50 = 60000 nM

A.9 Prompts

A.9.1 BioAssay summarization prompt

Prompt for BioAssays Summarization

Instruction: You are an expert in **BioAssay analysis** and **data extraction**. Your task is to carefully analyze the provided BioAssay JSON data and extract structured key information, including:
1. **BioAssay Summarization** – A concise summary of what this assay measures and its scientific purpose.
2. **Assay Type** – The experimental technique used (e.g., **Enzymatic Inhibition, Fluorescence Assay, SPR, Radioligand Binding**).
3. **Summary of Observations** – Important scientific insights derived from the BioAssay, including key patterns in activity, structural features affecting activity, and notable findings.
Step-by-step extraction process:
- Parse the **"descr"** section of the JSON, identifying key information about the assay. If this BioAssay is a counterscreen assay, set "CounterScreen" to "True".
- Identify the **Assay Type** by analyzing the **"name"** field.
- Extract **scientific insights** from the description and comments to create the **Summary of Observations**.
- Generate a **concise and informative summary** of the BioAssay, keeping scientific accuracy and relevance.
Output Format Return the extracted data in the following structured format:

```
json
{
  "BioAssay_Summary": "A brief but
       complete summary of what
      this assay is measuring and
      why it is important.",
  "Assay_Type": "The experimental
      method used (e.g., Enzymatic
      Inhibition, Fluorescence, SPR
      , etc.)",
  "Summary_of_Observations": "
      Scientific insights, key
      findings, and notable trends
      from the BioAssay.",
  "CounterScreen": "True" if the
      BioAssay is identified as
      CounterScreen against Query
      Protein, else "False".
}
```

Query Protein:
{Protein Description}
BioAssay JSON
{BioAssay JSON}

A.9.2 Generation prompt

Molecule Generation Prompt

Role: AI Molecular Generator and BioAssay Analyst

Profile
- **Author**: LangGPT
- **Version**: 1.1
- **Language**: English
- **Description**: An AI model specialized in analyzing BioAssay results, understanding protein-ligand interactions, and generating high-affinity molecules based on experimental data.

Skills
- Understanding **protein-ligand interactions** from experimental BioAssay data.
- Interpreting **BioAssay results** and extracting meaningful insights.
- Learning from **high-affinity molecules** in BioAssay data to generate new molecules.
- Ensuring **high binding affinity and specificity**, while avoiding **Pan-assay interference compounds (PAINS)**.
- Generating drug-like molecules that align with known **active reference compounds**.

Rules
1. **Carefully analyze** the input description of the **protein**, **BioAssay**, and **experimental results**.
2. Identify **high-affinity molecules** (low IC50/Kd values) from the **BioAssay data** as **reference molecules**.
3. Use reference molecules to **learn key functional groups and molecular scaffolds**.
4. Focus on **specificity rather than only high docking scores**.
5. **Each generated molecule should be enclosed within [BOS] and [EOS]**.
6. **Each SMILES should be numbered from 1 to 10, with one per line.**
7. Avoid **PAINS compounds** and prioritize **drug-likeness**.
8. **Do not blindly maximize molecular size**, as larger molecules may have artificially high docking scores but poor specificity.

Workflows
Step 1: Understand the BioAssay and Its Relation to the Query Protein
- The BioAssays may or may not be related to the **Query Protein**, please identify the correct Query Protein first.
- The type of **assay method** used (e.g., enzymatic, fluorescence, cell-based). - How the **assay measures protein-ligand interaction**. - The **affinity measurements** (e.g., IC50, Kd, Ki).
- Extract **key active molecules** from BioAssay results. (e.g. IC50, Ki, Kd<100nM) - Identify molecular features that contribute to **high binding affinity**.

Step 2: Learn from Active Molecules and Think Step by Step
- Extract **key functional groups** and **molecular scaffolds** from high-affinity reference molecules.
- Avoid **PAINS compounds** and prioritize **specificity**.
- Ensure molecules remain within a **reasonable drug-like chemical space**.
- Optimize molecular properties for **binding affinity and selectivity**.

Step 3: Generate 10 High-Affinity Molecules
- Use the **active reference molecules** as a learning guide, and use the **low binding affinity molecules** as negative samples.
- Each generated molecule should be optimized for **binding affinity and specificity**.
- The output format must follow this structure:
- Each **SMILES string should be enclosed in [BOS] and [EOS]**.
- Each SMILES should be **numbered from 1 to 10**, with each on a separate line.
- Avoid **PAINS compounds** and prioritize **drug-likeness**.
- Avoid generating molecules that are too large

Step 4: Justify the Molecular Selection
- Explain how the **reference molecules** influenced the molecular design.
- Describe how the **assay results** guided molecular modifications.

- Justify why these molecules should have **high binding affinity and specificity**.

—

Output Format:
1. BioAssay Understanding & Analysis

- Step-by-step reasoning about the BioAssay, its setup, and its relevance to the query protein

2. Selected Reference Molecules from BioAssay
- List of highly active molecules from BioAssay used as reference.

3. Generated Molecules
`[BOS] SMILES_1 [EOS]`
⋮
`[BOS] SMILES_10 [EOS]`

4. Justification for Molecular Selection
- Explanation of how reference molecules influenced design choices, ensuring specificity and affinity while avoiding PAINS.

Query Protein
{Protein Description}

BioAssays
{Assay Content}

A.9.3 Relevance assessment prompt

Relevance Assessment Prompt

Instruction:
You are an expert in **BioAssay analysis** and **data extraction**. Your task is to carefully analyze the provided BioAssay JSON data and extract structured key information, and decide whether the protein studied in the BioAssay is related to the Query Protein, with broader consideration of similarity:
- If the BioAssay can help protein inhibitor design toward the **Query Protein**, set "Relevant": "True".
- **Use a broad definition of similarity**, including:
- Same **protein family** (e.g., GRK4 and GRK2 are both **G protein-coupled receptor kinases**).
- Structural similarity (e.g., homologous domains, catalytic site conservation).
- Functional similarity (e.g., overlapping substrates).
Output Format

```
json
{
  "Relevant": "True"  // If the
      target in this BioAssay
      shares protein family,
      structure, function, or
      pathway with the {Query
      Protein, else set to "False".
       Only set to "False" if there
       is no meaningful similarity.
}
```

Query Protein
{protein description}
BioAssay JSON
{BioAssay content}

A.9.4 Ablation study generation prompt

Ablation study prompt

I would like to design drug-like small molecules with high binding affinity to a specific protein. `{protein_description}` Think step by step, generate 10 unique SMILES strings, try to generate diverse molecules instead of making slight change on current molecule. Each molecule should have '[BOS]' at the beginning and '[EOS]' at the end. Each SMILES should be numbered from 1 to 10 and on a separate line.

A.9.5 Molecule optimization prompt

Molecule optimization prompt

To enhance molecular specificity and minimize off-target effects, we aim to reduce potential activity against the hERG channel.
{hERG description}
{hERG BioAssays}
Given the retrieved BioAssays for the hERG channel and the associated activity data table, identify molecular features commonly associated with low activity as favorable and those associated with high activity as

undesirable. Using this information, optimize the following ten candidate SMILES strings to reduce their likelihood of interacting with the target. {Input SMILES}
The output should follow the same format: ten optimized SMILES strings, each enclosed in [BOS] and [EOS], with numbering from 1 to 10.