mCLM: A Modular Chemical Language Model that Generates Functional and Makeable Molecules

Anonymous authors
Paper under double-blind review

Abstract

Despite their ability to understand chemical knowledge, large language models (LLMs) remain limited in their capacity to propose novel molecules with desired functions (e.g., drug-like properties). In addition, the molecules that LLMs propose can often be challenging to make, and are almost never compatible with automated synthesis approaches. To better enable the discovery of functional small molecules, LLMs need to learn a new molecular language that is more effective in predicting properties and inherently synced with automated synthesis technology. Current molecule LLMs are limited by representing molecules based on atoms. In this paper, we argue that just like tokenizing texts into meaning-bearing (sub-)word tokens instead of characters, molecules should be tokenized at the level of functional building blocks, i.e., parts of molecules that bring unique functions and serve as effective building blocks for real-world automated laboratory synthesis. This motivates us to propose mCLM, a modular Chemical-Language Model that comprises a bilingual language model that understands both natural language descriptions of functions and molecular blocks. mCLM front-loads synthesizability considerations while improving the predicted functions of molecules in a principled manner. Experiments on 430 FDA-approved drugs showed that mCLM is capable of significantly improving chemical functions critical to determining drug potentials. mCLM, with only 3B parameters, also achieves improvements in synthetic accessibility relative to 7 other leading generative AI methods including GPT-5. When tested on 122 out-of-distribution medicines using only building blocks/tokens that are compatible with automated modular synthesis, mCLM outperforms all baselines in property scores and synthetic accessibility. mCLM can also reason on multiple functions and iteratively self-improve to rescue drug candidates that failed late in clinical trials ("fallen angels"). *

1 Introduction

Small molecules—the class of chemical matter primarily built from carbon atoms bonded together—can perform a wide range of important functions in human society (Zhang et al., 2025). These include essentials like promoting health by acting as medicines (Zheng et al., 2024; Edwards et al., 2024b; Singhal et al., 2023; Thirunavukarasu et al., 2023; Xiao et al., 2024a), converting energy by functioning as key components in solar cells (Nguyen et al., 2024b; Lv et al., 2021; Li et al., 2023b; 2024b; Si et al., 2024), and achieving sustainability by serving as inherently recyclable products. These functions also include many nice-to-haves that drive substantial economic growth, including colorants, flavorings, perfumes, cosmetics, coatings, quantum dots and insect repellants.

The traditional approach for small molecule synthesis is highly artisanal, and thus slow and expensive: (1) It is unfriendly to automation: machines are not good at performing thousands of reaction types, with each run under thousands of possible conditions and using millions of possible starting materials. (2) It leads to an undemocratized landscape in drug discovery. With development costs averaging around $1.3 billion per drug, only economically advanced countries can afford to invest in such high-risk research (Kneller, 2010). It also excludes potential breakthroughs from the developing world and inhibits research that could prevent diseases prevalent there. Moreover, participation in the process of

*All codes, data, and models will be released publicly upon publication.

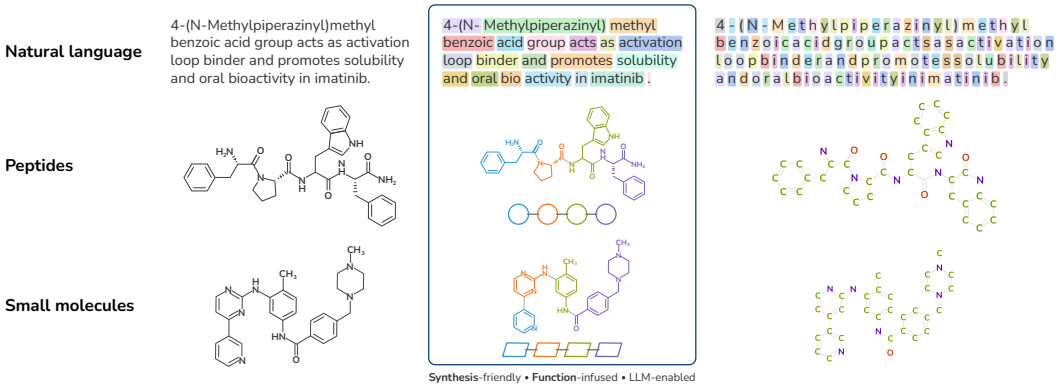

Figure 1: mCLM adopts a modular chemical vocabulary, which uses synthesis robot-friendly molecular building blocks as tokens together with natural language tokens. Compared with using natural language names, SMILES strings, or holistic embeddings for whole molecules, this building block level of tokenization better enables the prediction of compounds with improved properties, and guarantees automated synthesizability a priori, thus building a direct link between the digital and physical worlds. This approach stands to substantially enable AI-guided discovery of new small molecules with targeted functions.

molecular innovation currently requires access to highly trained experts in chemical synthesis. As a result, there are no effective drugs for many diseases such as Parkinson and Amyotrophic Lateral Sclerosis. (3) Most importantly, there are many instances of known commercial drugs or materials that have well-documented limitations that have remained unaddressed. For example, Imatinib is an important anticancer drug that works well in most parts of the body, but it only poorly penetrates the blood-brain barrier (Takayama et al., 2002; Senior, 2003; Coutre et al., 2004; Isobe et al., 2009). This means it may effectively treat cancer at its primary site, yet fail to prevent fatal brain metastasis. In such cases, more blood-brain-barrier penetrant molecules that retain all of the other favorable properties of the current drug are desired, but making such selective modifications in functional properties can be very challenging. As another example in the materials domain, 350+ million people in Asia and the Pacific have only limited access to electricity, and 150 million people still have no access at all. Organic photovoltaic (OPV) molecules are expected to be more economical and environment-friendly alternatives to current solar cells. However, current commercialized OPV devices either achieve less than 10% energy conversion efficiency—significantly lower than traditional silicon solar cells—or have stability far short of 10 years (Solak and Irmak, 2023).

An alternative block chemistry approach for small molecule synthesis has recently emerged (Gillis and Burke, 2007; Woerly et al., 2014; Li et al., 2015a; Lehmann et al., 2018; Trobe and Burke, 2018; Blair et al., 2022a; Angello et al., 2022; Wang et al., 2024; Strieth-Kalthoff et al., 2024a). Block chemistry is iterative carbon-carbon bond-forming chemistry that machines can do. It iteratively assembles small molecules from prefabricated building blocks using chemistry that is simple and general and thus readily automated. Akin to automated DNA, RNA, and peptide synthesis platforms, a major strength of this block-based approach is that it can access billions of novel small molecules with high degrees of functional potential using only a few automation-friendly reactions and a bounded set of pre-fabricated function-infused building blocks. We specifically recognized that this block-based approach provided an opportunity to create a new modular language for chemistry. The idea that molecules and their synthesis may be best understood from the perspective of a "chemical language" dates back to 2014 (Cadeddu et al., 2014), where structural fragments and functional groups play the roles of "chemical words". This view is supported by multiple observations aligned with natural language: (1) they can be decomposed and reassembled, (2) they exhibit ambiguity—the same building block can perform vastly different functions depending on the chemical context, and (3) they possess significant diversity, as many different structures can lead to the same function. This linguistic parallel suggests the potential to train a large language model specifically for molecules by linearizing their structures into modular sequences.

A common representation for such an approach is the Simplified Molecular Input Line Entry System (SMILES; Weininger, 1988), where atoms are denoted by one- or two-character symbols (e.g., C

for carbon, Br for bromine, and F for fluorine), rings are represented with numbers, and branches are indicated using parentheses. However, such atom-level tokenization strategies (e.g., SMILES (Weininger, 1988) or SELFIES (Krenn et al., 2020)) resemble character-level natural language models, which struggle to generalize effectively. Unlike proteins, which have a fixed vocabulary of 20 amino acids, small molecules exhibit an open vocabulary when each atom is treated as a token, as illustrated in Figure 1. It also causes severe restrictions in practicality because many of the new structures proposed by LLMs based on atom-level tokenization are not practically synthesizable in the laboratory. This approach has created a major gap between what is now possible in silico and what is possible in the physical world.

Furthermore, the SMILES representation can obscure critical structure information: two atoms that are direct neighbors in the molecular graph may be distantly separated in the SMILES string. Even with recent efforts to align SMILES and natural language description (Edwards et al., 2022a; Pei et al., 2024; Ahmad et al., 2022), integrate chemical properties and functional groups (Nguyen et al., 2025) into SMILES, incorporate 3D geometric information (Fu et al., 2025; Li et al., 2025a), and employ graph neural networks to capture molecular graphs and chemical reaction contexts (Wang et al., 2022a), these representations still fail to encapsulate functional knowledge that is often described only in natural language literature because the inherent properties and functions of molecules are hidden in their structure, composition, and interaction.

Therefore, our goal is not to make all drug-like molecules, rather, our goal is to figure out how to make the right ones, better, faster, stronger. Unlike machines, human scientists are inherently "multilingual," seamlessly navigating diverse modalities—from natural language and scientific figures in literature to complex scientific data such as molecular structures in knowledge bases. In contrast, most prior work on molecule discovery trains large language models on a single modality. Moreover, Human scientists "think before they talk," grounding their reasoning in deeply reflective and deliberate reflection and critical evaluation to generate new hypotheses. Current models lack this critical thinking capacity, limiting their ability to contribute meaningfully to discovery. In particular, the human body is a highly complex, interconnected system, and drug discovery is essentially a multi-objective optimization problem. It involves balancing factors such as drug absorption, first-pass metabolism, bioavailability, distribution, protein binding, and clearance. However, improving one property often comes at the expense of another. Many promising drug candidates have failed in the final stages of FDA approval due to an unacceptably high risk of drug-induced liver injury.

Against this backdrop, we argue that there is a correctable fundamental mismatch between the way LLMs work and the way chemists traditionally synthesize and study small molecules. Reducing tokenization granularity to the level of individual letters is counterproductive, as it complicates meaning extraction and increases the likelihood of generative AI hallucinate words that don't exist. Analogous limitations are inherently linked to atom-based tokenization. And as a result, unfortunately, much of the generative AI research for scientific discovery within the computer science community does not extend to hypothesis verification in the physical world. To bridge this gap, in this paper, we propose a novel LLM by drawing inspiration from the scientific discovery process itself. We aim to develop a science-inspired large language model that follow three principles: (1) "Observe" - acquire, represent and integrate knowledge from multiple data modalities; (2) "Think" – think critically to generate hypotheses; and (3) "Propose" – verify hypotheses through the Physical World. We aim to teach computers to speak two complementary languages: one that represents molecular building blocks (i.e., subgraph structures) indicative of specific functions and compatible with automated modular assembly, and another that describes these functions in natural language (Figure 2). Unlike existing approaches that add such knowledge as a post hoc step, we develop a function- and synthesis-aware modular chemical language model (mCLM). Inspired by bilingual speakers who frequently "code-switch" (naturally and often switch between their two languages within the same message; Poplack, 2013), we propose a novel neural encoder that integrates molecular structure and natural language. mCLM incorporates both function- and synthesis-related knowledge into the small molecule tokenization process a priori. First, we tokenize small molecules at the level of building blocks (graph substructures) that are able to predict function and are, by design, flexible, multi-scale, fully compatible with automated modular small molecule synthesis. We use graph neural networks to encode each building block. We then extract natural language sentences from the literature that describe the molecule's functions and chemical reactions and synthesis constraints of various molecules, and seamlessly insert the encoded building blocks alongside the corresponding entity names to form the training data, as illustrated in Figure 2. By reasoning on such functional

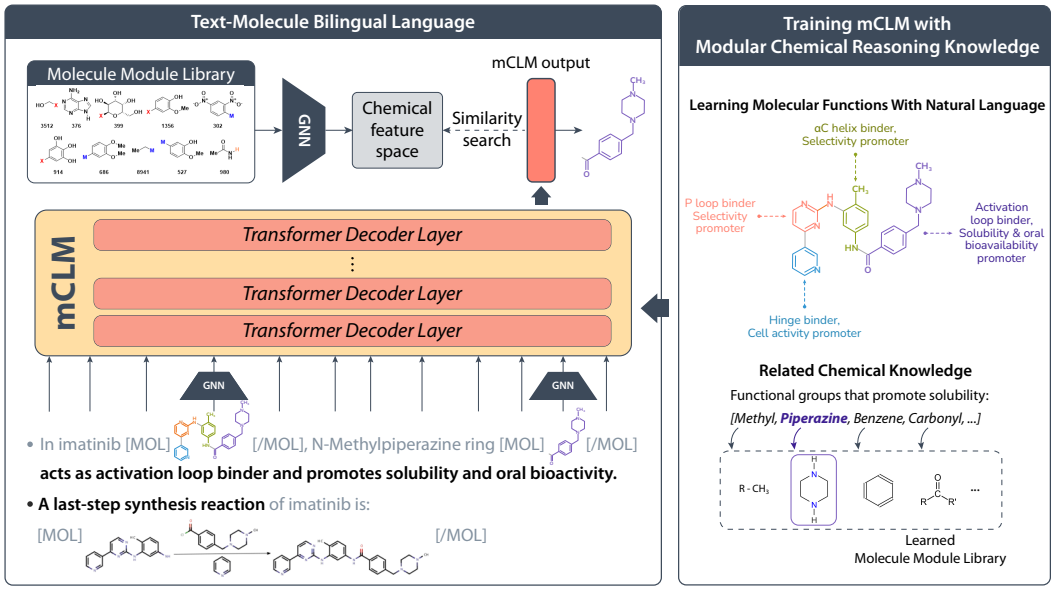

Figure 2: An overview of the mCLM. It is a multimodal chemical-language model that tokenizes molecules into synthesis robot-friendly *building blocks*, thus creating a direct link between the digital and physical worlds. After being trained on datasets consisting of properties, functions and synthesis data, the mCLM can conduct critical chemical reasoning through an iterative refinement process.

building blocks, mCLM guarantees to generate efficiently synthesizable molecules thanks to recent progress in block-based chemistry, while also improving the functions of molecules in a principled manner. During model inference, mCLM enables these functional modules to be predictably and automatically assembled into molecules with desired functions.

Compared to previous work, mCLM has multiple potential advantages of encoding and generating molecules at a modular level:

1. **Synthesis efficiency**: proposed molecules can be faster and more broadly makeable because the process is, by design, simple, iterative, general, and machine-friendly. This can enable rapid iterative drug and material development via seamless integration with automated lab experiments.
2. **Multimodal understanding**: resembling the mechanism of natural language tokenization, molecular modularization provides a more natural interface to align with word representations to form a more powerful multimodal representation.
3. **Deliberate reasoning**: leveraging LLMs' instruction following capability and wide-scale pretraining, mCLM is able to iteratively refine molecules based on knowledge of molecular functions.

2 The Modular Chemical Language Model

In natural language modeling, tokenization identifies common substrings (such as words and subwords), which carry richer semantic information than sequences of characters. Similarly in nature, most small molecules are composed primarily of connected *building blocks* (Lehmann et al., 2018; Trobe and Burke, 2018). There is likewise a high degree of inherent modularity in many medicines and materials (Ertl and Schuhmann, 2019; Arkan et al., 2020; Andrews et al., 1984; Vitaku et al., 2014).To leverage the phenomena in chemistry and borrow the spirit from natural language modeling, we propose mCLM, a multimodal model that jointly encodes and understands natural language and molecules based on synthesis-friendly building blocks instead of atoms. In Section 2.1, we introduce the concept of molecular building blocks as a chemical "vocabulary" and describe the tokenization process to obtain the library of building blocks. Then in Section 2.2 we describe the mCLM architecture and its training. Finally, we introduce the reasoning mechanism of mCLM which refines molecule design over multiple iterations in Section 2.3.

2.1 A FUNCTION-INFUSED AND SYNTHESIS-FRIENDLY VOCABULARY

In this work, we propose to leverage a chemical vocabulary V of synthesis-friendly building blocks. This approach guarantees capacity for automated iterative assembly first reported in 2015 (Li et al., 2015b) and, since then, demonstrated in multiple experimental campaigns including Strieth-Kalthoff et al. (2024b) and Angello et al. (2024). The blocks can be chemically connected using predefined synthesis rules in a short period. Briefly, akin to language models developed for peptides/proteins, small molecules can be assembled automatically from makeable building blocks. These building blocks are often highly associated with chemical functions, such as binding to protein targets, modulating enzyme activity, or affecting involved metabolic processes. This will enable rapid and iterative proposal of new small molecules, automated synthesis of those small molecules, and generation of the corresponding functional data on demand. In contrast to SMILES strings which break molecule structure during graph linearization (e.g., separating physically adjacent subgraphs such as the two carbons in molecule C(N)C), our representation is linear in the physical world.

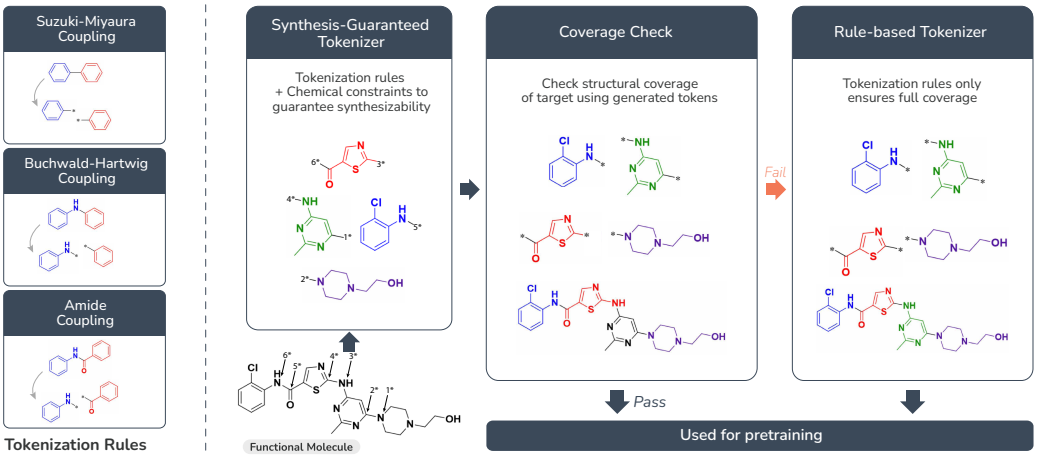

Figure 3: An overview of the tokenization process. A functional molecule is first processed by the synthesis-guaranteed tokenizer to produce a set of building blocks compatible with automated modular synthesis. These blocks are then evaluated via a structure coverage check to determine whether they fully reconstruct the original molecule. If coverage is complete, the blocks are used directly for pretraining. Otherwise, the molecule is reprocessed using a rule-based tokenizer to ensure full representation for training purposes.

During dataset construction, we combine two tokenization strategies to balance coverage and synthetic feasibility: a synthesis-guaranteed tokenizer and a rule-based tokenizer (Figure 3). The synthesis-guaranteed tokenizer disconnects the molecule only at bonds that can be formed by a predetermined small set of reactions that can be performed in an automated manner. Specifically, it only permits bond disconnections that correspond to reactions compatible with state-of-the-art automated synthesis platforms. These three bonds are: amide coupling, Suzuki-Miyaura coupling, and Buchwald-Hartwig coupling (see Figure 3 tokenization rules; Tyrikos-Ergas et al., 2025). The disconnection rules are adapted from well-established principles in block-based and automated chemistry (Blair et al., 2022b; Li et al., 2015b). Qualitative examples of tokenization are included in the appendix F, and the full rules and tokenizer code will be released with documentation. Importantly, the tokenizer is modular: as new automated reactions are validated in practice, their corresponding disconnection rules can be readily incorporated, expanding the makeable chemical space without requiring fundamental architectural changes. Unlike traditional retrosynthesis tools that prioritize maximum coverage or human-designed heuristics, this tokenizer is guided by the operational constraints of block chemistry. It also ensures that resulting blocks are free from functional group conflicts that would interfere with downstream synthesis. This conservative approach leads to a set of molecular tokens that are guaranteed to work under defined synthesis protocols, enabling seamless transition from model-generated output to real-world synthesis. When the synthesis-guaranteed tokenizer cannot fully cover a molecule—due to chemical incompatibilities or reaction constraints—we fall back on a rule-based tokenizer to ensure coverage of a larger variety of molecules. The rule-based tokenizer also breaks molecules along the same automated machine-friendly bonds as the above tokenizer. However, no

other rules are applied beyond specifying a minimum size for blocks. This rule-based tokenizer is used only during training to support learning on diverse molecular structures while maintaining consistency with synthesis logic. We note that the vocabulary growth is analogous to the shift from character- level to subword-level tokenization in natural language, where larger but semantically meaningful tokens significantly improved performance. These reusable, synthesis-friendly building blocks are derived from real data distributions. We collected approximately 8 million molecules from a wide range of public sources (Appendix G), of which approximately 6.8 million were tokenizable using our retrosynthesis-aware tokenization approach (see Figure 3). This process resulted in a vocabulary of 200,000 unique building blocks (see Table 6). An example of how the block sequence can reconstruct the molecule structure is also provided in Appendix F.

2.2 Chemical-Language Modeling

Figure 2 illustrates the architecture of mCLM. The model is a sequential generative model that processes molecule and text sequences in a unified manner. It adopts a Transformer architecture, which is well-suited for handling sequential data and allows for using pre-trained language models as a backbone. After tokenizing the molecules as mentioned in the last section, we encode each building block using graph neural networks (GNNs; Edwards et al., 2024a; Sprueill et al., 2024; Gasteiger et al., 2021) before passing them through an adaptor module. These representations are then concatenated with natural language embeddings at positions where molecule entity names appear. This results in a form of "code-switched" language which blends molecular structures with natural language descriptions. The feature sequence is then fed into a Transformer decoder-only architecture, which predicts the next token based on previous tokens. This allows for pre-training the model on a large corpus of multi-modal data, enabling it to learn and "talk about" the relationships between the different modalities and their respective representations.

We train mCLM on top of open-source pretrained large language models (we adopt Qwen2.5 Yang et al. (2024) as the starting point), allowing us to leverage their natural language understanding and reasoning capabilities without incurring the computational cost of training from scratch. Furthermore, pretrained natural-language LLMs have been exposed to a broad range of scientific information, making them a strong foundation for our curriculum. They help the model acquire general scientific knowledge before progressing to domain-specific training on molecular structures and drug-related information. As the training objective, we adopt a unified categorical cross-entropy (CCE) loss applied to both natural language and molecular tokens:

$$\mathcal{L} = H(P(\mathbf{x}), P_\theta(\mathbf{x})), \quad \text{logit}\left(P_\theta(v \mid \mathbf{x}_{1\cdots i-1})\right) = \begin{cases} \mathbf{c}_i^\top \mathbf{e}_v, & v \in \mathcal{V}_{\text{natural language}} \\ \mathbf{c}_i^\top f_\psi\left(\text{GNN}_\phi(v)\right), & v \in \mathcal{V}_{\text{molecular building block}} \end{cases}$$

Specifically, the cross-entropy loss is computed between the ground truth distribution $P(\mathbf{x})$ and the model distribution $P_\theta(\mathbf{x})$ over the combined vocabulary of natural language and molecular building blocks. The model generates logits for the next token v by computing the dot product between the contextual representation $\mathbf{c}_i$ with token embeddings. $\mathbf{c}_i$ is produced by the Transformer given the previous tokens $\mathbf{x}_{1\cdots i-1} = [X_1, \cdots, X_{i-1}]$). For natural language tokens, the embedding $\mathbf{e}_v$ is directly taken from the pretrained natural language model. The embedding for molecular building blocks is computed by passing the building block's graph through a GNN, followed by a linear adapter function f_ψ to project it into the same embedding space. This formulation enables joint training over both modalities using a single loss function. We train on approximately 1 million examples consisting of the 1,000 most frequent molecular building blocks, sampled from a new dataset we constructed. This dataset contains over 36 million molecular instructions, 1,000 tasks, and 8 million molecules. More details on the training procedure are described in Appendix E. Notably, the architecture of mCLM is compatible with larger vocabularies, and we anticipate future scaling studies as computational resources permit. mCLM is able to generalize to molecules composed of unseen building blocks. During training, the model learns to associate chemical structure with function via instruction-based pretraining. During inference, new blocks can be directly incorporated without retraining.

2.3 Critical Chemical Reasoning

Chemical reasoning over molecules often involves optimizing multiple functions, such as toxicity, bioactivity, and binding affinity. Therefore, it is not a straightforward task to propose an ideal molecule structure, especially with only a single attempt. Optimizing one function may lead to

trade-offs in others. For example, many drugs with higher potency were rejected due to increased toxicity to patients. To address this, we propose a reasoning process that allows the model to refine its own generated molecules and iteratively improve their desired functions. At each iteration, we evaluate the properties and identify one that still requires improvement, and mCLM proposes a modification of the molecule targeting this property. This process is repeated until a maximum number of iterations is reached. This process is summarized in Appendix Algorithm 1.

3 Experimental Evaluation

3.1 Creating Oracle Models for Evaluation

To evaluate the performance of our generated molecules, we construct oracle models focused on Absorption, Distribution, Metabolism, Excretion, Toxicity (ADMET) property prediction. We select 6 tasks from the Therapeutics Data Commons (TDC) benchmark (Huang et al., 2021): **AMES** (mutagenicity), **BBBP** (blood-brain barrier permeability), **CYP3A4** inhibition (metabolism), **DILI** (drug-induced liver injury), **HIA** (human intestinal absorption), and **PGP** (P-glycoprotein substrate classification). The detailed training procedure is described in Appendix D.2.3.

3.2 Improving FDA-Approved Drugs with Out-of-Vocabulary Blocks

In practice, real-world drug discovery is often driven by the optimization of known molecules, as existing approved drugs offer a more direct and promising path to safe and effective new drugs. Therefore, we evaluate mCLM's drug proposal capability in improving FDA-approved drugs. Specifically, we apply the mCLM to improve the 6 properties for all FDA-approved drugs consisting of synthesis-guaranteed blocks. This amounts to 122 molecular structures and 153 unseen blocks. We confine the output vocabulary of mCLM to 582 synthesis-guaranteed blocks. Most of these drugs (120/122) contain blocks that were not present in the 1,000-block training vocabulary, presenting an opportunity to examine how the mCLM works on out-of-distribution molecules. Results in Table 1 show that improvements are achieved for all properties[†]. This shows that mCLM is not limited to a fixed vocabulary but can generalize beyond its training coverage. We also compare our approach against a wide range of strong text-based molecule editing baselines, including MoleculeSTM (Liu et al., 2022), FineMolTex (Li et al., 2025b), GPT-4o, GPT-5, Gemini-2.5-Flash (Gemini-2.5-F), LDMol (Chang and Ye, 2024), and Claude 3.5 Haiku (Claude-3.5-H). We conducted additional comparisons with two relevant baselines: DGAE (Boget et al., 2024) (a discrete graph autoencoder using vector quantization) and HierVAE (Jin et al., 2020) (a hierarchical VAE for molecular graph generation). The more detailed results and a technical comparison with vector-quantized methods are enclosed in Appendix H. Despite the fact that *all* of the baseline models are allowed to generate without synthesis guarantees, their average improvements fall behind mCLM.

Table 1: Average pharmacokinetic and toxicity properties of FDA drugs composed of synthesis-guaranteed blocks, as well as their proposed modifications. (↓: lower is better, ↑: higher is better). Green = better than FDA, Red = worse, Light green bold = best overall per column.* Since HierVAE does not accept natural language input, a unique HierVAE was trained for each property.

Model	AMES (↓)	BBBP (↑)	CYP3A4 (↓)	DILI (↓)	HIA (↑)	PGP (↓)	Avg. Improv.
FDA Drug	47.8	61.4	2.1	60.1	98.96	64.6	0.00 %
MoleculeSTM	47.1	63.4	2.2	59.3	98.76	64.1	0.31 %
FineMolTex	47.5	66.0	2.4	59.5	98.84	**64.0**	-0.73 %
LDMol	49.0	63.5	2.5	57.5	99.0	67.6	-3.07 %
HierVAE*	48.2	66.4	1.2	53.1	99.7	64.3	10.5 %
GPT-4o	46.0	72.1	2.2	60.8	**99.3**	65.2	2.45 %
GPT-5	49.1	70.2	2.4	61.2	99.2	65.0	-0.82 %
Claude-3.5-Haiku	49.3	72.2	2.2	58.6	98.3	65.7	1.64 %
Gemini-2.5-Flash	45.0	72.5	1.8	58.1	99.1	64.5	6.98 %
mCLM (Ours)	**44.4**	**85.2**	**1.4**	**53.7**	98.99	64.4	**15.0 %**

The key to expediting the drug creation process is to discover potent molecular candidates that are simultaneously synthesis-friendly. While mCLM shows strong property editing results, its key benefit lies in its synthesis-friendly nature. We assessed the synthesizability of generated molecules by computing synthetic accessibility (SA scores) (Ertl and Schuffenhauer, 2009) as a quick heuristic

[†]We also test a different distribution of 430 non-synthesis-guaranteed FDA drugs in Appendix A.2.

(see Table 3). Then, as a more rigorous assessment, we consider Allchemy, which is the state-of-the-art retrosynthesis software (Wołos et al., 2022; Strieth-Kalthoff et al., 2024c). Allchemy is computationally expensive, but it evaluates synthesizability to the best of publicly available human chemical knowledge. For example, it finds synthetic routes for 98.1% of the FDA-approved molecules. Moreover, to our knowledge it is the only retrosynthesis software for which many of the proposed routes have been reduced to practice in the lab with physical experimentation. These studies have been published in top tier journals (Mikulak-Klucznik et al., 2020). We select the top 3 natural language models by SA score (mCLM, MolSTM, and FineMolTex) and randomly sample 200 generated molecules from each to be assessed by Allchemy on a supercomputing cluster. In this experiment, we also consider HierVAE and two ablations: mCLM using SMILES strings instead of a GNN (No GNN), and mCLM using BRICS tokens instead of our tokenizer (No Synth. Tokenizer). Quite interestingly, HierVAE produces very good SA scores but poor retrosynthesis-based synthesizability scores. We speculate this is because SA score prefers simpler molecules or those with simpler substructures, whereas retrosynthesis tools evaluate actual synthetic feasibility rather than only based on structural patterns; this exemplifies the limitations of SA score and the necessity of our new metrics.

Table 2: Synthetic accessibility (SA) (Ertl et al., 2009), validity, and retrosynthetic results across baselines. Synthesizability is the percent of valid molecules where a retrosynthetic route was found. Makeability is the overall percent of generations which can be synthesized (Makeability =Valid $\times$ Synth.). *We used AiZynthFinder as the retroysynthesis tool for this result and will replace it in the camera-ready with Allchemy.

Model	SA (↓)	Validity (%)	Synthesizability (%)	Makeability (%)
FDA	2.70	**100.0**	98.11	98.11
MoleculeSTM	2.64	93.80	91.03	85.39
FineMolTex	2.58	94.20	90.15	84.96
mCLM (Ours)	2.43	**100.0**	**98.23**	**98.23**
mCLM (No GNN)*	2.97	49.64	13.5	6.70
mCLM (No Synth. Tokenizer)*	3.09	**100.0**	30.0	30.0
HierVAE*	2.35	99.7	67.3	67.2
DGAE*	4.46	100.0	24.2	24.2

As shown in Table 2, molecules proposed by mCLM are 100% valid (syntactically correct) and 98.2% synthesizable, which is superior even to the FDA drugs. In contrast, MoleculeSTM outputs are valid only 93.9% of the time, and among those, only 90.3% are predicted to be synthesizable using exhaustive retrosynthesis search. Out of MoleculeSTM-generated molecules, only 84.8% can be made ($93.9 \times 0.903 = 84.8$).

3.3 Case Study: Multi-step Reasoning to Resurrect the "Fallen Angels"

There are many new drug candidates that almost reach FDA approval but fall short for various reasons when tested in clinical trials. For example, Evobrutinib is a Bruton's tyrosine kinsase (BTK) inhibitor that went through clinical trial as a drug for relapsing Multiple Sclerosis. However, the FDA placed a partial clinical hold on Phase III trials in April 2023 after two patients showed signs of drug-induced liver injury (Montalban et al., 2024). TNG348 is a USP1 inhibitor designed for treating BRCA1/2-mutant and HRD cancers, but it failed in phase 1/2 clinical trials due to liver abnormalities (Inc, 2024; Simoneau et al., 2025). These "fallen angels" represent tremendous opportunities for impactful engagement of the AI/chemistry interface, because much is known about the strengths of each of these small molecules, and it is also known why they fell short. Fixing such fallen angels is a high leverage opportunity for the function-infused mCLM to contribute.

Figure 4 shows an application of the mCLM, without synthesis restrictions on the vocabulary, to these two molecules. For both, the initial step is to optimize DILI, the reason the drugs failed in clinical trials. Following that, the mCLM fixes other properties which were made worse in the previous attempt (PGP for Evobrutinib and BBBP for TNG248). For good measure, another property of each molecule is then improved. Notably, at each step, the mCLM only makes minor modifications to each drug of roughly 1 building block. While the mCLM shows promising results for repairing these drugs, it is worth noting that drug discovery is a many-objective optimization problem. While we are able to generate molecules with improved toxicity relative to Evobrutinib and TNG348, as well as other properties, yet other important properties may still have been compromised. Future work may

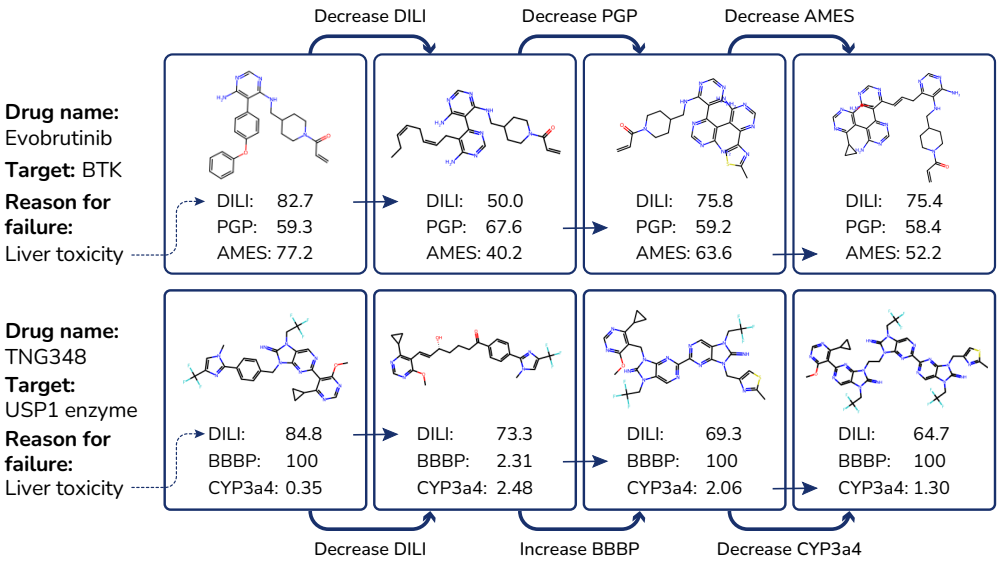

Figure 4: Examples of fallen angel property modification.

want to investigate longer reasoning chains across a wider variety of properties. For comparison, we show a (less-successful) version of this experiment using MoleculeSTM A.3.

3.4 Library Design: Identifying Functionally-Important Blocks

As a by-product, mCLM is also able to identify blocks which are preferred for certain functions. This can help inform virtual screening campaign design (e.g., which 10 blocks should we use to get the best chemical space for BBBP?), and it can also be useful for stimulating scientific inquiry. As an example, we calculated the most frequent modifications preferred by the mCLM to improve DILI in FDA drugs (Figure 5). Notably, modification 3 resembles a newly discovered modification for reducing amphotericin B toxicity by human scientists in Maji et al. (2023).

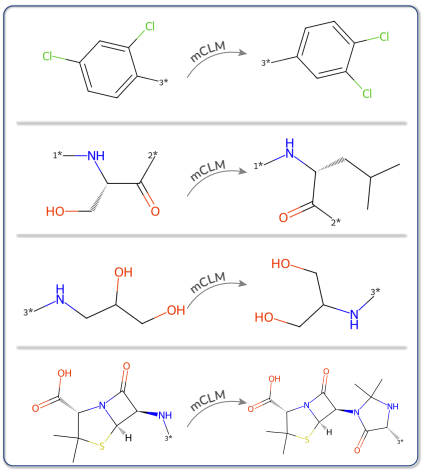

Figure 5: Four of the most frequent mCLM-proposed modifications to improve DILI in FDA-approved drugs.

4 Related Work

Protein language models (Lin et al., 2023; Hayes et al., 2025; Wang et al., 2025b; Madani et al., 2023; Xiao et al., 2024b; Su et al., 2024; Buehler and Buehler, 2024; Wang et al., 2025a; Zhu et al., 2024; Ferruz et al., 2022) have been highly successful because proteins/peptides are tokenized at the level of amino acid building blocks (which are akin to the words in a sentence). This makes it possible to associate the sequences of these building blocks with the fold and function of proteins, and then the corresponding building blocks can be assembled into targeted sequences.

Inspired from the successes of protein LLMs, LLMs have been increasingly applied in computational chemistry (Zhang et al., 2024a;b; Fang et al., 2023; Livne et al., 2024; Pei et al., 2023; Zhao et al., 2024; 2023; Yu et al., 2024; Ye et al., 2025; Liu et al., 2023b; Li et al., 2024c; Liu et al., 2024a; Taylor et al., 2022). Recent work has further explored cross-modal modeling of molecule and language (Edwards et al., 2022b; Liu et al., 2022; Li et al., 2025b; Chang and Ye, 2024). General-purpose LLMs such as GPT-4o (OpenAI, 2025a), GPT-5 (OpenAI, 2025b), Gemini-2.5-Flash (Google DeepMind, 2025), and Claude-3.5-H (Anthropic, 2025) show multimodal reasoning capabilities. However, most molecular language models still rely on atom-level (Wang et al., 2019; Ahmad

et al., 2022; Zhou et al., 2023; Xia et al., 2023) or Byte-Pair Encoding on SMILES (Chang and Ye, 2024), and thus they often produce invalid tokens, misalign with chemical structures, and generate unsynthesizable molecules. To address these limitations, other approaches incorporate functional groups and fragments into molecular representations (Li et al., 2023a; Nguyen et al., 2024c; Han et al., 2023; Jin et al., 2024; Nguyen et al., 2024a; Wang et al., 2023; Zhang et al., 2020; 2021), capturing both atomic- and group-level information. However, these methods tend to break molecules at bonds that are difficult or even impossible to form through chemical reactions. In contrast, our mCLM decomposes small molecules into function-infused and synthesis-friendly building blocks and integrates synthesis constraints early in training, akin to the way peptide/protein LLMs tokenize at the amino acid level.

Several recent papers propose chemically meaningful tokenization schemes. SAFE (Noutahi et al., 2024) modifies SMILES representations to avoid fragment splits but still inherits core issues of line notations as we described in Section 1, including validity and synthesis challenges. Group-SELFIES (Cheng et al., 2023) extends SELFIES with group-level tokens and can in principle represent our synthesis-friendly blocks. If it is used in a line notation form, however, standard LLM tokenizers will not ensure validity or synthesizability. Reasyn (Lee et al., 2025) is contemporary to our work. It focuses on synthesizable molecule projection using reinforcement learning, requiring the model to predict reactions between pairs of blocks. Unlike Reasyn, mCLM does not require the prediction of reactions, since one is guaranteed to exist between any blocks. Their task and training methodology may be relevant for future mCLM models. We also note recent efforts in learning discrete molecule representations with VQ-like mechanisms (Guo et al., 2025; Ha et al., 2025; Boget et al., 2024). However, these approaches do not guarantee synthesis compatibility and generally lack alignment with natural language. In Section H, we include comparisons with DGAE and HierVAE as additional baselines.

5 Conclusions and Future Work

We have developed a function- and synthesis-aware modular Chemical-Language Model (mCLM), as the first attempt to jointly model natural language sequences with modular chemical language. By design, the mCLM only generates chemical building blocks that can be iteratively assembled on robotic small molecule synthesis platforms, enabling the rapid creation of novel molecules with desired functions, all accessible by non-specialists. In the future we aim to scale mCLM to larger backbones, incorporate richer multimodal knowledge related to physical and chemical properties from 2D/3D molecular structures, protein-ligand complexes, cell lines and nucleic acid sequences to further enable the mCLM to reason on biological activity, protein docking knowledge, and individuals' genetic profiles. We plan to extend chemical reasoning to additional aspects such as filling in unspoken knowledge gaps, thinking outside of the box, System 2 thinking for counterfactual reasoning and plausibility prediction, and resolving conflicting claims. We also plan to leverage more physical constraints from simulation tools and chemical and reaction knowledge bases. In the long term, we envision the mCLM as part of a comprehensive, multi-agent, human-in-the-loop autonomous laboratory, structured around iterative cycles of reasoning, proposal, synthesis, physical testing, feedback, and reasoning to enable never-ending self-improvement and co-evolvement with human scientists.

Reproducibility Statement

In section 3.1 and D.2.3, we describe the process and resources of developing oracle models for evaluation. The synthesis-guaranteed tokenizer is described in further detail in Section F. Sections C and D explain the data collection process and the statistics of the training dataset. Section G lists all source datasets. Model training details are described in Section E. Evaluation data from FDA-approved drugs are described in Section 3.2, which also lists the baselines we used for comparison. The full data and resources will be released upon publication.

REFERENCES

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

Allchemy. URL https://allchemy.net/.

PR Andrews, DJ Craik, and JL Martin. Functional group contributions to drug-receptor interactions. *Journal of medicinal chemistry*, 27(12):1648–1657, 1984.

Nicholas H. Angello, Vandana Rathore, Wiktor Beker, Agnieszka Wołos, Edward R. Jira, Rafał Roszak, Tony C. Wu, Charles M. Schroeder, Alán Aspuru-Guzik, Bartosz A. Grzybowski, and et al. Closed-loop optimization of general reaction conditions for heteroaryl suzuki-miyaura coupling. *Science*, 378(6618):399–405, Oct 2022. doi: 10.1126/science.adc8743.

Nicholas H Angello, David M Friday, Changhyun Hwang, Seungjoo Yi, Austin H Cheng, Tiara C Torres-Flores, Edward R Jira, Wesley Wang, Alán Aspuru-Guzik, Martin D Burke, et al. Closed-loop transfer enables artificial intelligence to yield chemical knowledge. *Nature*, 633(8029):351–358, 2024.

Anthropic. Claude-3.5-H large language model. https://www.anthropic.com/, 2025. Accessed: 2025-09-22.

Emre Arkan, Eyup Yalcin, Muhittin Unal, M Zeliha Yigit Arkan, Mustafa Can, Cem Tozlu, and Serafettin Demic. Effect of functional groups of self assembled monolayer molecules on the performance of inverted perovskite solar cell. *Materials Chemistry and Physics*, 254:123435, 2020.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7:1–13, 2015.

Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.

Daniel J. Blair, Sriyankari Chitti, Melanie Trobe, David M. Kostyra, Hannah M. Haley, Richard L. Hansen, Steve G. Ballmer, Toby J. Woods, Wesley Wang, Vikram Mubayi, and et al. Automated iterative csp3–c bond formation. *Nature*, 604(7904):92–97, Feb 2022a. doi: 10.1038/s41586-022-04491-w.

Daniel J Blair, Sriyankari Chitti, Melanie Trobe, David M Kostyra, Hannah MS Haley, Richard L Hansen, Steve G Ballmer, Toby J Woods, Wesley Wang, Vikram Mubayi, et al. Automated iterative c sp 3–c bond formation. *Nature*, 604(7904):92–97, 2022b.

Yoann Boget, Magda Gregorova, and Alexandros Kalousis. Discrete graph auto-encoder. *Transactions on Machine Learning Research*, 2024.

Fabio Broccatelli, Richard Trager, Michael Reutlinger, George Karypis, and Mufei Li. Benchmarking accuracy and generalizability of four graph neural networks using large in vitro adme datasets from different chemical spaces. *Molecular Informatics*, 41(8):2100321, 2022.

Eric L Buehler and Markus J Buehler. X-lora: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design. *APL Machine Learning*, 2(2), 2024.

Andrea Cadeddu, Elizabeth K. Wylie, Janusz Jurczak, Matthew Wampler-Doty, and Bartosz A. Grzybowski. Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. *Angewandte Chemie International Edition*, 53(31):8108–8112, 2014. doi: https://doi.org/10.1002/anie.201403708. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201403708.

Jinho Chang and Jong Chul Ye. Ldmol: A text-to-molecule diffusion model with structurally informative latent space surpasses ar models. *arXiv preprint arXiv:2405.17829*, 2024.

Austin H Cheng, Andy Cai, Santiago Miret, Gustavo Malkomes, Mariano Phielipp, and Alán Aspuru-Guzik. Group selfies: a robust fragment-based molecular string representation. *Digital Discovery*, 2023.

Philipp le Coutre, Karl-Anton Kreuzer, Stefan Pursche, Malte von Bonin, T Leopold, Gökben Baskaynak, Bernd Dörken, Gerhard Ehninger, Oliver G. Ottmann, Andreas Jenke, Martin Bornhäuser, and Eberhard Schleyer. Pharmacokinetics and cellular uptake of imatinib and its main metabolite cgp74588. *Cancer Chemotherapy and Pharmacology*, 2004. doi: 10.1007/s00280-003-0741-6.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 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, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.47. URL https://aclanthology.org/2021.emnlp-main.47.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proc. The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP2022)*, 2022a.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates, 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.26.

Carl Edwards, Ziqing Lu, Ehsan Hajiramezanali, Tommaso Biancalani, Heng Ji, and Gabriele Scalia. Molcap-arena: A comprehensive captioning benchmark on language-enhanced molecular property prediction. *arXiv preprint arXiv:2411.00737*, 2024a.

Carl Edwards, Aakanksha Naik, Tushar Khot, Martin Burke, Heng Ji, and Tom Hope. Synergpt: In-context learning for personalized drug synergy prediction and drug design. In *Proc. 1st Conference on Language Modeling (COLM2024)*, 2024b.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. L+M-24: Building a dataset for Language+Molecules @ ACL 2024. In Carl Edwards, Qingyun Wang, Manling Li, Lawrence Zhao, Tom Hope, and Heng Ji, editors, *Proceedings of the 1st Workshop on Language + Molecules (L+M 2024)*, pages 1–9, Bangkok, Thailand, 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.langmol-1.1. URL https://aclanthology.org/2024.langmol-1.1.

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

Peter Ertl and Tim Schuhmann. A systematic cheminformatics analysis of functional groups occurring in natural products. *Journal of natural products*, 82(5):1258–1263, 2019.

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

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *ArXiv preprint*, abs/2306.08018, 2023. URL https://arxiv.org/abs/2306.08018.

Noelia Ferruz, Steffen Schmidt, and Birte H¨ocker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Cong Fu, Xiner Li, Blake Olson, Heng Ji, and Shuiwang Ji. Fragment and geometry aware tokenization of molecules for structure-based drug design using language models. In *Proc. The Thirteenth International Conference on Learning Representations (ICLR2025)*, 2025.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802, 2021.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020.

Eric P Gillis and Martin D Burke. A simple and modular strategy for small molecule synthesis: Iterative suzuki−miyaura coupling of b-protected haloboronic acid building blocks. *Journal of the American Chemical Society.*, 129(21), 2007. ISSN 0002-7863.

Google DeepMind. Gemini-2.5-F large language model. https://www.deepmind.com/, 2025. Accessed: 2025-09-22.

Shuhan Guo, Yatao Bian, Ruibing Wang, Nan Yin, Zhen Wang, and Quanming Yao. Unified molecule-text language model with discrete token representation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 9205–9213, 2025.

Sumin Ha, Jun Hyeong Kim, Yinhua Piao, and Sun Kim. Mv-clam: Multi-view molecular interpretation with cross-modal projection via language model. *arXiv preprint arXiv:2503.04780*, 2025.

Shen Han, Haitao Fu, Yuyang Wu, Ganglan Zhao, Zhenyu Song, Feng Huang, Zhongfei Zhang, Shichao Liu, and Wen Zhang. Himgnn: a novel hierarchical molecular graph representation learning framework for property prediction. *Briefings in Bioinformatics*, 24(5):bbad305, 2023.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Tango Therapeutics Inc. Tango therapeutics announces discontinuation of tng348 program, May 2024. URL https://ir.tangotx.com/news-releases/news-release-details/tango-therapeutics-announces-discontinuation-tng348-program.

Y. Isobe, K. Sugimoto, A. Masuda, Y. Hamano, and K. Oshimi. Central nervous system is a sanctuary site for chronic myelogenous leukaemia treated with imatinib mesylate. *Internal medicine journal*, 2009. doi: 10.1111/j.1445-5994.2009.01947.x.

Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. Large language models on graphs: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Wengong Jin, Regina Barzilay, and Tommi S. Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4839–4848. PMLR, 2020. URL http://proceedings.mlr.press/v119/jin20a.html.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1): D1373–D1380, 2023.

R. Kneller. The importance of new companies for drug discovery: origins of a decade of new drugs. In *Nat Rev Drug Discov 9, 867–882 (2010)*, 2010.

Clayton W Kosonocky, Claus O Wilke, Edward M Marcotte, and Andrew D Ellington. Mining patents with large language models demonstrates congruence of functional labels and chemical structures. *arXiv preprint arXiv:2309.08765*, 2023.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Saee Paliwal, Weili Nie, and Arash Vahdat. Rethinking molecule synthesizability with chain-of-reaction. *arXiv preprint arXiv:2509.16084*, 2025.

Jonathan W. Lehmann, Daniel J. Blair, and Martin D. Burke. Towards the generalized iterative synthesis of small molecules. *Nature Reviews Chemistry*, 2(2), Feb 2018. doi: 10.1038/s41570-018-0115.

Biaoshun Li, Mujie Lin, Tiegen Chen, and Ling Wang. Fg-bert: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Briefings in Bioinformatics*, 24(6):bbad398, 2023a.

Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, May 2024a. URL https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/.

Jin Li, Naiteng Wu, Jian Zhang, Honghui Wu, Kunming Pan, Yingxue Wang, Guilong Liu, Xianming Liu, Zhenpeng Yao, and Qiaobao Zhang. Machine learning-assisted low-dimensional electrocatalysts design for hydrogen evolution reaction. *Nano-Micro Letters*, 2023b. doi: 10.1007/s40820-023-01192-5.

Jin Li, Meisa Zhou, Honghui Wu, Lifei Wang, Jian Zhang, Naiteng Wu, Kunming Pan, Guilong Liu, Yinggan Zhang, Jiajia Han, Xianming Liu, Xiang Chen, Jiayu Wan, and Qiaobao Zhang. Machine learning-assisted property prediction of solid-state electrolyte. *Advanced Energy Materials*, 2024b. doi: 10.1002/aenm.202304480.

Junqi Li, Steven G Ballmer, Eric P Gillis, Seiko Fujii, Michael J Schmidt, Andrea M E Palazzolo, Jonathan W Lehmann, Greg F Morehouse, and Martin D Burke. Synthesis of many different types of organic small molecules using one automated process. *Science.*, 347(6227), 2015a. ISSN 0036-8075.

Junqi Li, Steven G Ballmer, Eric P Gillis, Seiko Fujii, Michael J Schmidt, Andrea ME Palazzolo, Jonathan W Lehmann, Greg F Morehouse, and Martin D Burke. Synthesis of many different types of organic small molecules using one automated process. *Science*, 347(6227):1221–1226, 2015b.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3d molecule-text interpretation in language models. *ArXiv preprint*, abs/2401.13923, 2024c. URL https://arxiv.org/abs/2401.13923.

Xiner Li, Limei Wang, Youzhi Luo, Carl Edwards, Shurui Gui, Yuchao Lin, Heng Ji, and Shuiwang Ji. Learning to generate 3d molecules via language models with geometry-aware tokenization. In *Proc. 2025 International Conference on Machine Learning (ICML2025)*, 2025a.

Yibo Li, Yuan Fang, Mengmei Zhang, and Chuan Shi. Advancing molecular graph-text pre-training via fine-grained alignment. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 1589–1599, 2025b.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Gang Liu, Michael Sun, Wojciech Matusik, Meng Jiang, and Jie Chen. Multimodal large language models for inverse molecular design with retrosynthetic planning. *arXiv preprint arXiv:2410.04223*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure-text model for text-based retrieval and editing. *ArXiv preprint*, abs/2212.10789, 2022. URL https://arxiv.org/abs/2212.10789.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*, 2023b.

Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, Alán Aspuru-Guzik, et al. nach0: multimodal natural and chemical languages foundation model. *Chemical Science*, 15(22):8380–8389, 2024.

Steven A Lopez, Edward O Pyzer-Knapp, Gregor N Simm, Trevor Lutzow, Kewei Li, Laszlo R Seress, Johannes Hachmann, and Alán Aspuru-Guzik. The harvard organic photovoltaic dataset. *Scientific data*, 3(1):1–7, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chade Lv, Xin Zhou, Lixiang Zhong, Chunshuang Yan, M. Srinivasan, Z. Seh, Chuntai Liu, Hongge Pan, Shuzhou Li, Yonggang Wen, and Qingyu Yan. Machine learning: An advanced platform for materials development and state prediction in lithium-ion batteries. *Advances in Materials*, 2021. doi: 10.1002/adma.202101474.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

Rishikesh Magar, Yuyang Wang, Cooper Lorsung, Chen Liang, Hariharan Ramasubramanian, Peiyuan Li, and Amir Barati Farimani. Auglichem: data augmentation library of chemical structures for machine learning. *Machine Learning: Science and Technology*, 3(4):045015, nov 2022. doi: 10.1088/2632-2153/ac9c84. URL https://dx.doi.org/10.1088/2632-2153/ac9c84.

Arun Maji, Corinne P Soutar, Jiabao Zhang, Agnieszka Lewandowska, Brice E Uno, Su Yan, Yogesh Shelke, Ganesh Murhade, Evgeny Nimerovsky, Collin G Borcik, et al. Tuning sterol extraction kinetics yields a renal-sparing polyene antifungal. *Nature*, 623(7989):1079–1085, 2023.

Barbara Mikulak-Klucznik, Paulina Gołębiowska, Alex A. Bayly, and et al. Computational planning of the synthesis of complex natural products. *Nature*, 588(7836):83–88, 2020. doi: 10.1038/s41586-020-2855-y.

Xavier Montalban, Karolina Piasecka-Stryczynska, Jens Kuhle, Pascal Benkert, Douglas L Arnold, Martin S Weber, Andrea Seitzinger, Hans Guehring, Jamie Shaw, Davorka Tomic, Yann Hyvert, Danielle E Harlow, Martin Dyroff, and Jerry S Wolinsky. Efficacy and safety results after >3.5 years of treatment with the bruton's tyrosine kinase inhibitor evobrutinib in relapsing multiple sclerosis: Long-term follow-up of a phase ii randomised clinical trial with a cerebrospinal fluid sub-study. *Multiple Sclerosis Journal*, 30(4-5):558–570, 2024. doi: 10.1177/13524585241234783. URL https://doi.org/10.1177/13524585241234783. PMID: 38436271.

Thao Nguyen, Kuan-Hao Huang, Ge Liu, Martin D Burke, Ying Diao, and Heng Ji. Farm: Functional group-aware representations for small molecules. *arXiv preprint arXiv:2410.02082*, 2024a.

Thao Nguyen, Tiara Torres-Flores, Changhyun Hwang, Carl Edwards, Ying Diao, and Heng Ji. Glad: Synergizing molecular graphs and language descriptors for enhanced power conversion efficiency prediction in organic photovoltaic devices. In *Proc. 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)*, 2024b.

Thao Nguyen, Tiara Torres-Flores, Changhyun Hwang, Carl Edwards, Ying Diao, and Heng Ji. Glad: Synergizing molecular graphs and language descriptors for enhanced power conversion efficiency prediction in organic photovoltaic devices. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4777–4785, 2024c.

Thao Nguyen, Kuan-Hao Huang, Ge Liu, Martin D. Burke, Ying Diao, and Heng Ji. Farm: Functional group-aware representations for small molecules. In *Proc. NAACL2025 Workshop on AI and Scientific Discovery: Directions and Opportunities*, 2025.

Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan SC Lim, and Prudencio Tossou. Gotta be safe: a new framework for molecular design. *Digital Discovery*, 3(4):796–804, 2024.

OpenAI. GPT-4o: Large language model. https://openai.com/, 2025a. Accessed: 2025-09-22.

OpenAI. GPT-5 large language model. https://openai.com/, 2025b. Accessed: 2025-09-22.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*, 2023.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *ArXiv preprint*, abs/2402.17810, 2024. URL https://arxiv.org/abs/2402.17810.

Shana Poplack. "sometimes i'll start a sentence in spanish y termino en español": Toward a typology of code-switching. *Linguistics*, 51(s1):11–14, 2013.

Benjamin Sanchez-Lengeling, Jennifer N Wei, Brian K Lee, Richard C Gerkin, Al'an Aspuru-Guzik, and Alexander B Wiltschko. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685*, 2019.

K. Senior. Gleevec does not cross blood-brain barrier. *The Lancet. Oncology*, 2003. doi: 10.1016/s1470-2045(03)01050-7.

Zhan Si, Deguang Liu, Wan Nie, Jingjing Hu, Wei Wang, Tingting Jiang, Haizhu Yu, and Yao Fu. Data-based prediction of redox potentials via introducing chemical features into the transformer architecture. *Journal of Chemical Information and Modeling*, 2024. doi: 10.1021/acs.jcim.4c01299.

Antoine Simoneau, Charlotte B Pratt, Hsin-Jung Wu, Shreya S Rajeswaran, Charlotte Grace Comer, Sirimas Sudsakorn, Wenhai Zhang, Shangtao Liu, Samuel R Meier, Ashley H Choi, et al. Characterization of tng348: a selective, allosteric usp1 inhibitor that synergizes with parp inhibitors in tumors with homologous recombination deficiency. *Molecular Cancer Therapeutics*, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

E. Solak and E. Irmak. Advances in organic photovoltaic cells: a comprehensive review of materials, technologies, and performance. *RSC Advances*, 2023. doi: 10.1039/d3ra01454a.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *ArXiv preprint*, abs/2402.10980, 2024. URL https://arxiv.org/abs/2402.10980.

Felix Strieth-Kalthoff, Han Hao, Vandana Rathore, Joshua Derasp, Théophile Gaudin, Nicholas H Angello, Martin Seifrid, Ekaterina Trushina, Mason Guy, Junliang Liu, Xun Tang, Masashi Mamada, Wesley Wang, Tuul Tsagaantsooj, Cyrille Lavigne, Robert Pollice, Tony C Wu, Kazuhiro Hotta, Leticia Bodo, Shangyu Li, Mohammad Haddadnia, Agnieszka Wołos, Rafał Roszak, Cher Tian Ser, Carlota Bozal-Ginesta, Riley J Hickman, Jenya Vestfrid, Andrés Aguilar-Granda, Elena L Klimareva, Ralph C Sigerson, Wenduan Hou, Daniel Gahler, Slawomir Lach, Adrian Warzybok, Oleg Borodin, Simon Rohrbach, Benjamin Sanchez-Lengeling, Chihaya Adachi, Bartosz A Grzybowski, Leroy Cronin, Jason E Hein, Martin D Burke, and Alán Aspuru-Guzik. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science*, 384 (6697):eadk9227, May 2024a.

Felix Strieth-Kalthoff, Han Hao, Vandana Rathore, Joshua Derasp, Théophile Gaudin, Nicholas H Angello, Martin Seifrid, Ekaterina Trushina, Mason Guy, Junliang Liu, et al. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science*, 384(6697):eadk9227, 2024b.

Felix Strieth-Kalthoff, Sara Szymkuc, Karol Molga, Alán Aspuru-Guzik, Frank Glorius, and Bartosz A Grzybowski. Artificial intelligence for retrosynthetic planning needs both data and expert knowledge. *Journal of the American Chemical Society*, 146(16):11005–11017, 2024c.

Dagmar Stumpfe, Huabin Hu, and Jurgen Bajorath. Evolving concept of activity cliffs. *ACS omega*, 4(11): 14360–14368, 2019.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *ICLR*, 2024.

Xiaoyu Sun, Nathaniel J. Krakauer, Alexander Politowicz, Wei-Ting Chen, Qiying Li, Zuoyi Li, Xianjia Shao, Alfred Sunaryo, Mingren Shen, James Wang, and Dane Morgan. Assessing graph-based deep learning models for predicting flash point. *Mol. Inf.*, 39(6):1900101, feb 2020. doi: 10.1002/minf.201900101. URL https://doi.org/10.1002%2Fminf.201900101.

N. Takayama, N. Sato, S. O'Brien, Y. Ikeda, and S. Okamoto. Imatinib mesylate has limited activity against the central nervous system involvement of philadelphia chromosome-positive acute lymphoblastic leukaemia due to poor penetration into cerebrospinal fluid. *British journal of haematology*, 2002. doi: 10.1046/j.1365-2141.2002.03881.x.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

A. Thirunavukarasu, D. Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and D. Ting. Large language models in medicine. *Nature Network Boston*, 2023. doi: 10.1038/s41591-023-02448-8.

Melanie Trobe and Martin D. Burke. The molecular industrial revolution: Automated synthesis of small molecules. *Angewandte Chemie International Edition*, 57(16):4192–4214, Mar 2018. doi: 10.1002/anie.201710482.

Theodore Tyrikos-Ergas, Sevasti Agiakloglou, Antonio J Laporte, Wesley Wang, Chieh-Kai Chan, Clare E Wells, Christopher K Rakowski, Rachel I Hammond, Jia Qiu, Jonathan D Raymond, et al. Automated iterative nc and cc bond formation. *ChemRxiv*, 2025.

Derek Van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling*, 62(23):5938–5951, 2022.

Edon Vitaku, David T. Smith, and Jon T. Njardarson. Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among u.s. fda approved pharmaceuticals. *Journal of Medicinal Chemistry*, 2014. doi: 10.1021/jm501100b.

Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin Burke. Chemical-reaction-aware molecule representation learning. In *Proc. The International Conference on Learning Representations (ICLR2022)*, 2022a.

Ruheng Wang, Hang Zhang, Trieu Nguyen, Shasha Feng, Hao-Wei Pang, Xiang Yu, Li Xiao, and Peter Zhiping Zhang. Pepthink-r1: Llm for interpretable cyclic peptide optimization with cot sft and reinforcement learning. *arXiv preprint arXiv:2508.14765*, 2025a.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436, 2019.

Wesley Wang, Nicholas H Angello, Daniel J Blair, Theodore Tyrikos-Ergas, William H Krueger, Kameron N S Medine, Antonio J LaPorte, Joshua M Berger, and Martin D Burke. Rapid automated iterative small-molecule synthesis. *Nat. Synth.*, 3(8):1031–1038, May 2024.

Xinyou Wang, Zaixiang Zheng, Fei YE, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=5z9GjHgerY.

Yifei Wang, Shiyang Chen, Guobin Chen, Ethan Shurberg, Hang Liu, and Pengyu Hong. Motif-based graph representation learning with application to chemical molecules. In *Informatics*, volume 10, page 8. MDPI, 2023.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022b.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

David S Wishart, Sagan Girod, Harrison Peters, Eponine Oler, Juan Jovel, Zachary Budinski, Ralph Milford, Vicki W Lui, Zinat Sayeeda, Robert Mah, et al. Chemfont: the chemical functional ontology resource. *Nucleic Acids Research*, 51(D1):D1220–D1229, 2023.

Eric M Woerly, Jahnabi Roy, and Martin D Burke. Synthesis of most polyene natural product motifs using just 12 building blocks and one coupling reaction. *Nature Chemistry.*, 6, 2014. ISSN 1755-4349.

Agnieszka Wołos, Dominik Koszelewski, Rafał Roszak, Sara Szymkuć, Martyna Moskal, Ryszard Ostaszewski, Brenden T Herrera, Josef M Maier, Gordon Brezicki, Jonathon Samuel, et al. Computer-designed repurposing of chemical wastes into drugs. *Nature*, 604(7907):668–676, 2022.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9 (2):513–530, 2018.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*, 2023.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, page 102888, 2024a.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv:2408.11363*, 2024b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. Drugassist: A large language model for molecule optimization. *Briefings in Bioinformatics*, 26(1):bbae693, 2025.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*, 2024.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024a.

Shichang Zhang, Ziniu Hu, Arjun Subramonian, and Yizhou Sun. Motif-driven contrastive learning of graph representations. *arXiv preprint arXiv:2012.12533*, 2020.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, Keir Adams, Maurice Weiler, Xiner Li, Tianfan Fu, Yucheng Wang, Haiyang Yu, YuQing Xie, Xiang Fu, Alex Strasser, Shenglong Xu, Yi Liu, Yuanqi Du, Alexandra Saxton, Hongyi Ling, Hannah Lawrence, Hannes Stärk, Shurui Gui, Carl Edwards, Nicholas Gao, Adriana Ladera, Tailin Wu, Elyssa F. Hofgard, Aria Mansouri Tehrani, Rui Wang, Ameya Daigavane, Montgomery Bohde, Jerry Kurtin, Qian Huang, Tuong Phung, Minkai Xu, Chaitanya K. Joshi, Simon V. Mathis, Kamyar Azizzadenesheli, Ada Fang, Alán Aspuru-Guzik, Erik Bekkers, Michael Bronstein, Marinka Zitnik, Anima Anandkumar, Stefano Ermon, Pietro Liò, Rose Yu, Stephan Günnemann, Jure Leskovec, Heng Ji, Jimeng Sun, Regina Barzilay, Tommi Jaakkola, Connor W. Coley, Xiaoning Qian, Xiaofeng Qian, Tess Smidt, and Shuiwang Ji. Artificial intelligence for science in quantum, atomistic, and continuum systems. In *Foundations and Trends in Machine Learning*, 2025.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*, 2024b.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882, 2021.

Ziqiao Zhang, Bangyi Zhao, Ailin Xie, Yatao Bian, and Shuigeng Zhou. Activity cliff prediction: Dataset and benchmark. *arXiv preprint arXiv:2302.07541*, 2023.

Haiteng Zhao, Shengchao Liu, Ma Chang, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in neural information processing systems*, 36:5850–5887, 2023.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, et al. Chemdfm: Dialogue foundation model for chemistry. *arXiv preprint arXiv:2401.14818*, 2024.

Yizhen Zheng, Huan Yee Koh, Maddie Yang, Li Li, Lauren T. May, Geoffrey I. Webb, Shirui Pan, and George Church. Large language models in drug discovery and development: From disease mechanisms to clinical trials, 2024. URL `https://arxiv.org/abs/2409.04481`.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.

Xiao Zhu, Chenchen Qin, Fang Wang, Fan Yang, Bing He, Yu Zhao, and Jianhua Yao. Cd-gpt: a biological foundation model bridging the gap between molecular sequences through central dogma. *bioRxiv*, pages 2024–06, 2024.

George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 1936.

A Additional Results

A.1 Synthesizability

For the mCLM, all starting materials are the core building blocks, and the same collection of building blocks is used redundantly to make all the proposed structures. For this analysis, we assume that all the building blocks required for the mCLM approach are available. Given a sequence of blocks A - B - C - D in our methodology, we can assemble them with the following steps:

- Coupling reaction (A and B)
- Deprotect reaction (AB)
- Coupling reaction (AB and C)
- Deprotect reaction (ABC)
- Coupling reaction (ABC and D)

Scaling this formula to n building blocks gives a requirement of $2n - 3$ reactions (synthetic steps).

On the other hand, each molecule generated by MoleculeSTM is a unique problem for synthesis, and thus cannot leverage the strengths of the block-based approach. Each generated target requires a customized synthesis solution.

However, it is important to consider an external metric for verifying the synthesis capabilities of the mCLM. In table 3, we compute validity and synthetic accessibility scores for all baselines, both overall and in terms of specific property modifications. We find that mCLM outperforms all baselines (lower SA is better).

Table 3: Synthetic Accessibility (SA) scores with RDKit validity percentages across datasets.

Dataset	FDA		mCLM		MoleculeSTM		FineMolTex		GPT-4o		GPT-5		Gemini-2.5-F		LDMol		Claude-3.5-H	
	SA	Valid (%)	SA	Valid (%)	SA	Valid (%)	SA	Valid (%)	SA	Valid (%)	SA	Valid (%)	SA	Valid (%)	SA	Valid (%)	SA	Valid (%)
AMES	2.70	**100.0**	**2.39**	**100.0**	2.64	94.26	2.52	94.26	3.06	90.98	2.98	72.13	2.92	89.34	2.85	90.16	3.12	93.44
BBBP	2.70	**100.0**	**2.44**	**100.0**	2.67	93.44	2.61	95.90	2.67	88.52	2.87	78.69	2.69	90.16	2.89	94.26	2.90	93.44
CYP3A4	2.70	**100.0**	**2.40**	**100.0**	2.60	95.90	2.63	94.26	2.70	89.34	2.83	85.25	2.75	93.44	2.74	93.44	2.69	88.52
DILI	2.70	**100.0**	**2.38**	**100.0**	2.62	94.26	2.59	93.44	2.73	93.44	2.75	81.15	2.70	90.16	2.80	87.70	2.79	88.52
HIA	2.70	**100.0**	**2.42**	**100.0**	2.73	91.80	2.59	93.44	2.68	87.70	2.82	73.77	2.77	94.26	2.80	90.16	2.86	91.80
PGP	2.70	**100.0**	**2.45**	**100.0**	2.63	93.44	2.59	94.26	2.81	90.16	2.85	86.07	2.79	82.79	2.87	93.44	2.75	90.16
Mean	2.70	**100.0**	**2.41**	**100.0**	2.64	93.85	2.59	94.26	2.78	90.02	2.85	79.51	2.77	90.03	2.83	91.53	2.85	90.98

Next, we select the best 3 models and consider a computationally expensive but rigorous approach. We leverage Allchemy, which is the state-of-the-art retrosynthesis software, to quantitatively measure this. We randomly sampled a representative set of 200 molecules generated by each baseline (and the original FDA molecules). We gave each molecule 30 minutes on a supercomputing cluster (which is quite exhaustive) and didn't apply a price limit for substrates (so this is really quite generous for starting materials). Scores, as reported in Table 2, show much stronger overall performance of the mCLM. Here, validity is the percent of syntactically correct molecules as measured by RDKit, synthesizability is the percent of valid molecules where Allchemy could find a synthetic path, and Makeability is the product of those two metrics. Makeability represents the percent of generated molecules which could be made in the lab.

A.2 Expanded Testing Without Synthesis Guarantees

As a large-scale test, we apply the mCLM to improve all FDA-approved drugs consisting of at least 3 blocks (without synthesis guarantees). Note: this is just a confirmatory test, since the mCLM is intended to be used on the distribution of synthesis-guaranteed molecules (as shown in the main paper). This amounts to 430 molecular structures and 796 unseen blocks. Since most of these molecules (426/430) contain blocks that were not present in the 1,000 blocks used for training, this presents an opportunity to find how the mCLM performs out-of-distribution. Results in Table 4 show that improvement is achieved for 5/6 properties, even though the model has not seen almost half of the blocks in its vocabulary during training. Specifically, the strictest synthesis-guaranteed tokenizer covers 26.7% of compounds in our corpus. For the remainder, a rule-based tokenizer is used. Despite this fallback, mCLM demonstrates strong generalization, as it successfully incorporated GNN-derived features for the unseen modules.

Table 4: Average pharmacokinetic and toxicity properties of FDA drugs with 3 or more blocks and their proposed modifications. Note, these molecules do not have synthesis-guarantees. (↓: lower is better, ↑: higher is better.

Property	AMES Mut. (↓)	BBBP (↑)	CYP3A4 Inhib. (↓)	DILI (↓)	HIA (↑)	PGP (↓)	Mean Improv.
FDA Drug	59.5	37.6	2.0	66.2	93.2	66.0	0.00 %
mCLM	54.0	41.4	**1.2**	**55.5**	**97.6**	68.0	12.87 %

A.3 Applying MoleculeSTM to the "Fallen Angels"

As a baseline comparison, we repeat the fallen angels reasoning experiment using MoleculeSTM (Figure 6). Even though the steps 1 and 2 of MoleculeSTM's modification of TNG348 showed comparable property changes to mCLM, MoleculeSTM encounters molecule validity problems: it generates a syntactically incorrect molecule on step 1 of Evobrutinib and on step 3 of TNG348.

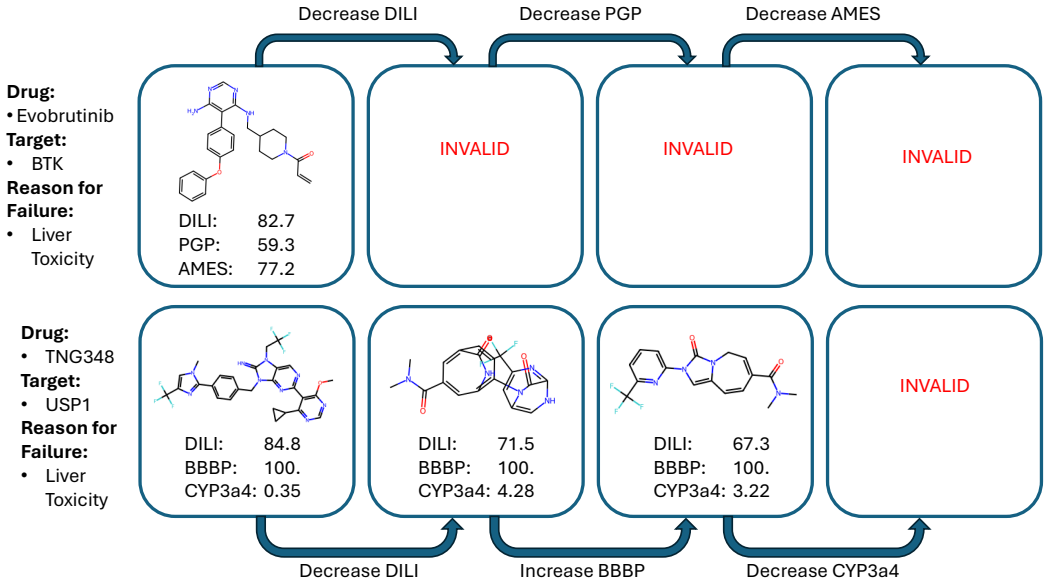

Figure 6: Examples of fallen angel property modification using MoleculeSTM.

B Critical Chemical Reasoning Algorithm

Algorithm 1 Critical Chemical Reasoning for Molecule Design in mCLM

Input: Initial molecule M_0
Output: Refined molecule M^*
Initialize $M \leftarrow M_0$
while True **do**
    Enumerate over functions to improve in M (e.g., metabolism, drug-induced organ injury, blood-brain barrier penetration)
    if No clear objective remains for improvement **then**
        return M
    elseSelect property which requires most improvement.
    end if
    mCLM generates a candidate modification M' by replacing, adding, or removing building blocks in M
    Evaluate M' with respect to desired functions
end while

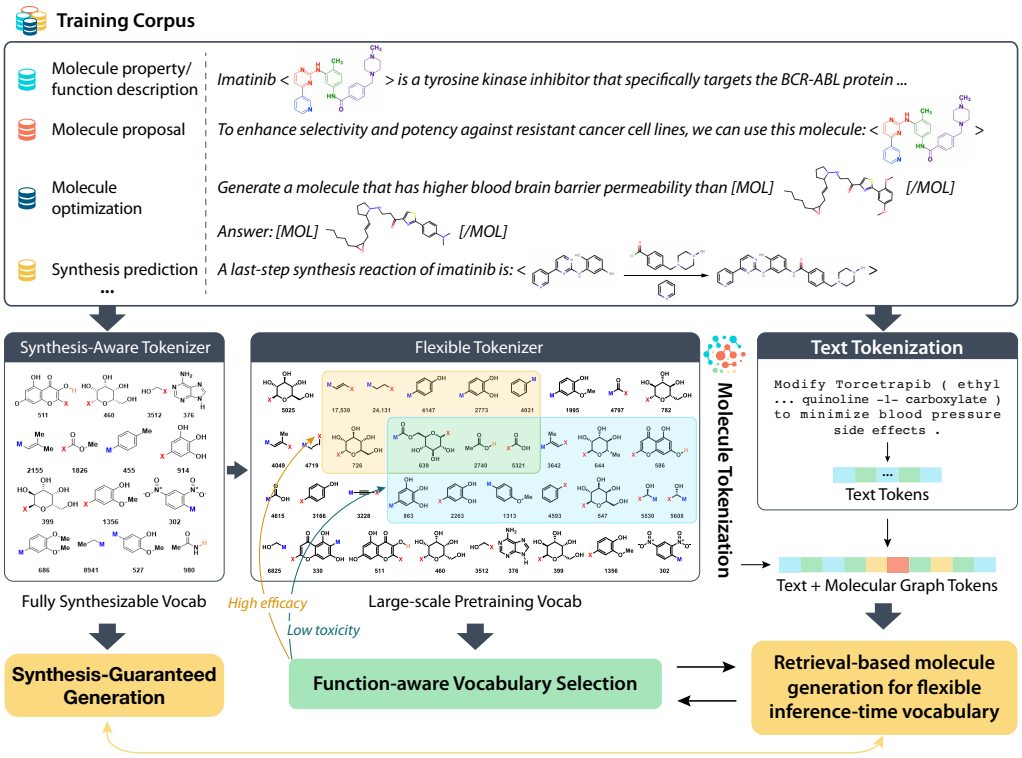

Figure 7: Overview of Data Creation.

C Pretraining Data Mixture

To train the model, our goal is to cover a wide variety of different molecular functions and classes of molecules. To do so, we consider a wide variety of data sources. We use three existing instruction datasets: SMolInstruct (Yu et al., 2024), Tulu-3 (Lambert et al., 2024), and the Biomolecular text portion of Mol-Instructions (Fang et al., 2023). SMolInstruct is used to cover the standard set of chemistry LLM tasks. We supplement this with the non-overlapping portion of Mol-Instructions. Finally, we include Tulu-3 to preserve general reasoning capability.

C.1 Property-Based Synthetic Instruction Data Generation for Contrastive Learning

Additionally, we augment this data mixture with a significant amount of synthetic data created using real-world property datasets. Given our datasets, we create additional question datapoints for both regression and classification tasks, such as "Is this molecule <SMILES> ... </SMILES> blood brain barrier permeable? Yes". We also consider the opposite direction: property to molecule and multiple properties to molecule, and unconditional molecule generation. The templates used for constructing this data were created using GPT-4o; 50 were originally generated and then bad templates were removed by hand. Further, we augmented each data point with a description of the property, also written by GPT-4o.

In addition to these tasks, we also consider the molecule optimization task as described in DrugAssist (Ye et al., 2025) (e.g., given molecule A, improve property X). However, their approach has fundamental limitations due to the reliance on oracle models. These models may be out-of-distribution for molecules within our dataset, and low-quality models may propagate errors into our training data. To address this issue, we consider "activity cliffs" (Stumpfe et al., 2019; Zhang et al., 2023) within existing property datasets to train the model for molecule optimization. Activity cliffs are two molecules which are structurally similar but have large differences in a given property. We opt to base our definition of activity cliff on Murcko scaffolds (Bemis and Murcko, 1996). This methodology, which is widely used in medicinal chemistry, extracts the core subgraph, or scaffold, of a molecule. We found that this approach was much more computationally feasible on large-scale molecule datasets than approaches like fingerprint similarity (Bajusz et al., 2015).

Given nominal data (e.g., toxic, kinase inhibitor, banana smell), we look for a pair of molecules that have the same scaffold but only one molecule has the desired property. Given numerical data (e.g., solubility, power conversion efficiency, HOMO-LUMO gap), we instead look for molecules which have a property difference of 1 standard deviation. This forms the positive-negative same-scaffold portion of our data. We also consider pairs of molecules that have different scaffolds but the same property. Here, our goal is to teach the model to consider function over structure. As an example, we may ask "Propose a molecule with a different structure than [MOL] ... [/MOL] that still demonstrates anticoagulant properties. [MOL] ... [/MOL]". For a select number of property types, we use oracle models instead of ground truth data. Please see Appendix D.2.3 for more information.

C.2 Data Filtering

Due to the large size of our dataset, we filter our data to only keep the most frequent 100k (50k end blocks and 50k mid blocks), 10k (5k and 5k), and 1000 blocks (500 and 500) for experiments. This is because there are far fewer molecule tokens compared to text (each training sample only has up to about 10). Due to Zipf's law (Zipf, 1936), in the full dataset, a single molecule token may only appear in one example in the entire training corpus. Without considerably scaling the compute budget, understanding this additional block is difficult. Overall, this filtering enables us to learn more efficiently and requires less memory resources for training the model, and allows us to better test the model architecture. Please see the section on training details for more information in Appendix E and see Appendix G for source property datasets. Table 5 gives a breakdown data quantity by instruction type for each subset of the data. Table 6 gives a breakdown of the molecule data for each subset of the data. Note that not all molecules in the "Full Data" can be represented with blocks, so we have a subset for "Only Blocks" molecules, and then molecules consisting of the most frequent "Top 100k" blocks, "Top 10k" blocks, and "Top 1000" blocks. Scaffolds are the number of unique Murcko scaffolds (Bemis and Murcko, 1996). "Caps" are end blocks, while "Mid" are middle blocks. "Synthesis-aware" is the number of synthesis-guaranteed blocks out of "Total". Data splits are subdivided into their original source: our constructed instruction data or molecule instructions from SMolInstruct.

Table 5: Dataset categories and their respective sample counts.

Data Type	Full Data	Only Blocks	Top 100k	Top 10k	Top 1000
All	36,641,170	20,910,431	14,823,280	6,195,903	1,054,124
Existing Data Sources	3,117,841	1,208,365	1,049,473	537,674	149,994
SMolInstruct	2,124,738	215,262	56,370	14,242	2,300
Tulu-3	939,343	939,343	939,343	469,672	93,934
Mol-Instructions Biomol. Text	53,760	53,760	53,760	53,760	53,760
Synthetic Real Data (Ours)	21,290,353	9,294,434	7,711,548	3,960,633	705,210
Classification	5,612,215	3,771,038	3,192,583	1,337,619	223,341
Molecule Generation	535,317	316,630	245,034	150,001	27,393
Positive Negative Same Scaffold	3,795,820	113,390	65,566	16,286	2,007
Positive Positive Different Scaffold	660,925	477,976	256,435	191,720	33,579
Property to Molecule	6,676,957	3,125,097	2,878,012	1,639,322	307,780
Multi-Property to Molecule	572,505	556,107	533,647	423,089	78,793
Scaffold+Property to Molecule	705,833	405,497	226,696	80,448	13,078
Regression	2,730,781	528,699	313,575	122,148	19,239
Synthetic Oracle Data (Ours)	12,232,976	10,407,632	6,062,259	1,697,596	198,920
Classification	2,077,327	1,778,874	1,058,754	299,444	35,168
Positive Negative Same Scaffold	72,763	58,645	25,035	4,935	469
Scaffold+Property to Molecule	1,169,407	925,055	453,087	113,049	13,642
Property to Molecule	13,257	13,254	8,841	906	2
Regression	8,900,222	7,631,804	4,516,542	1,279,262	149,639

D Synthetic Data Generation

D.1 Task Formulation for Molecular Design

To train our modular chemical language model (mCLM), we generate synthetic data that reflect realistic scenarios encountered in molecular design. These include tasks relevant to drug discovery and organic photovoltaic (OPV) materials design. Our goal is to equip the model with the ability to understand and reason over molecular representations and natural language prompts across a diverse set of tasks.

Drug Discovery Tasks: In the context of drug discovery, chemists often engage in iterative and multi-objective optimization processes, where molecules are evaluated, modified, or generated based on various physicochemical

Table 6: Breakdown of molecule data by dataset category.

Subset	Source	Molecules				Molecule Tokens			
		Total	Tokenized	Untokenized	Scaffolds	Cap	Mid	Total	Synthesis-Aware
Full Data	SMolInstruct	1,951,205	1,566,030	385,175	510,464				
	Our data	6,160,565	5,270,982	864,212	1,109,562				
	Total	7,994,305	6,787,879	1,178,957	1,537,424				
Only Blocks	SMolInstruct	459,910	459,910	0	145,614	157,204	59,681	216,885	50,788
	Ours	3,830,543	3,830,543	0	705,641	531,472	175,903	707,375	189,770
	Total	4,220,604	4,220,604	0	799,267	598,470	203,810	802,280	214,431
Top 100k	SMolInstruct	146,079	146,079	0	49,743	22,871	14,809	37,680	10,930
	Ours	2,345,519	2,345,519	0	429,960	49,941	49,780	99,721	27,823
	Total	2,458,034	2,458,034	0	457,032	50,000	50,000	100,000	28,220
Top 10k	SMolInstruct	37,686	37,686	0	12,373	4,130	2,534	6,664	2,618
	Ours	821,238	821,238	0	147,524	5,000	4,998	9,998	3,468
	Total	848,674	848,674	0	153,144	5,000	5,000	10,000	3,487
Top 1000	SMolInstruct	4,867	4,867	0	1391	477	359	836	420
	Ours	109,554	109,554	0	19,639	500	500	1,000	432
	Total	112,657	112,657	0	20,101	500	500	1,000	432

and pharmacological properties. These tasks typically involve querying for specific properties, modifying structures to meet certain design criteria, or generating novel candidates that satisfy given constraints. We consider the following tasks that a chemist may perform:

- **Text-to-Molecule Generation:** Generate a molecule from a textual description of its structure or properties.
- **Property Prediction:** Given a molecule and a property of interest, predict whether the molecule possesses the property (binary classification) or the quantitative value of the property (regression).
- **Molecular Optimization:** Modify a given molecule to satisfy or improve a specific property.
- **Scaffold-Constrained Generation:** Given a scaffold and a target property, generate a molecule that satisfies both the structural constraint and the property constraint.

OPV Material Design Tasks: To support broader applications of mCLM beyond drug discovery, we also incorporate tasks relevant to organic photovoltaic (OPV) material design. OPVs are an emerging class of lightweight, flexible materials used for solar energy harvesting, where the power conversion efficiency (PCE) is the primary performance metric. A typical OPV device is composed of a donor molecule and an acceptor molecule. The donor is responsible for absorbing sunlight and generating excitons (electron-hole pairs), while the acceptor facilitates charge separation and electron transport. The chemical compatibility and electronic alignment between the donor and acceptor molecules critically influence the resulting PCE. To enable learning in this domain, we define the following OPV-specific tasks

- **PCE Prediction**: Predict the PCE of a given donor–acceptor pair.
- **Donor/Acceptor Completion**: Given a donor (or acceptor) and a target PCE, generate the complementary component (acceptor or donor) that achieves the desired performance.
- **Constrained Completion with Scaffold**: Generate donor or acceptor molecules that match a given scaffold and achieve a target PCE.

D.2 Instruction Tuning Data Generation

D.2.1 Prompt-Answer Templates

To generate instructional data, we construct a pool of question and answer templates for each task. These templates include multiple paraphrased variants to introduce linguistic diversity and improve the model's generalization capability. During data generation, a question template is randomly sampled from the question set and populated with sample-specific content. Likewise, a corresponding answer template is sampled from the answer set to form a complete prompt–response pair. For example:

- **Question Template:** *"Given [a molecule] and [a property of interest], modify the molecule to achieve [desired property value]."*
- **Answer Template:** *"The [property] of [molecule] is [value]."*

This templating strategy allows us to produce a large number of diverse, semantically equivalent training instances that support instruction tuning across multiple molecular design tasks.

D.2.2 Label Sources for Instruction Tuning Data

To enable diverse and meaningful pretraining for mCLM, we incorporate both experimentally derived and model-generated labels, covering a broad spectrum of molecular properties critical to chemistry, pharmacology, and materials science. This dual-labeling strategy allows the model to learn from abundant low-level molecular descriptors while also reasoning over high-level functional and biological endpoints.

Low-Level Molecular Properties from ChEMBL: For foundational chemical descriptors, we leverage the ChEMBL25 database—a comprehensive bioactivity resource containing approximately 2 million compounds with rich structural and physicochemical annotations. ChEMBL25 serves as an abundant and reliable source of labels for low-level properties that are widely used in cheminformatics pipelines. From this corpus, we select a core set of descriptors that are most informative for molecular design: Hydrogen bond acceptors (HBA), Hydrogen bond donors (HBD), LogP (octanol–water partition coefficient), Molecular weight (MolWt), Number of aromatic rings, Number of rotatable bonds, Topological polar surface area (TPSA). These descriptors are inexpensive to compute and provide critical insights into molecular solubility, permeability, and synthetic feasibility—making them essential for early-stage screening and property-based filtering.

High-Level Molecular Properties via Oracle Labeling: In addition to low-level properties, we aim to expose the model to high-level functional endpoints that capture complex biological phenomena. Such endpoints are central to pharmacokinetics, drug safety, and efficacy, but they are rarely available in large quantities due to the high cost of experimental validation. Consequently, labeled datasets for these tasks are limited in size and diversity. To address this challenge, we employ the oracle ensemble models to generate synthetic labels for a curated set of ADMET tasks.

D.2.3 Ensemble Oracle Model:

To generate high-quality synthetic labels for downstream tasks, we construct oracle models focused on ADMET property prediction. The oracle models we use for ADMET assessment are state-of-the-art, high-performing models trained on experimental benchmark datasets with strong reported correlation to assay measurements, as shown in Table 7 below. This setup is highly meaningful in computational drug discovery (especially compared to rougher and less meaningful QED scores) and is designed to emulate digital screening pipelines used in industry.

We select tasks from the Therapeutics Data Commons (TDC) benchmark [‡] using the following criteria:

- **Relevance to Drug Discovery:** The task must reflect a critical aspect of drug efficacy or toxicity.
- **Predictability:** The task must be reliably predictable using existing models. Specifically, we evaluate all 22 ADMET-related classification tasks in TDC and retain only those where standard models achieve an area under the ROC curve (AUC) greater than 0.80. This ensures the synthetic labels are sufficiently accurate for training purposes.

Based on these criteria, we select six tasks:

- **AMES** (mutagenicity),
- **BBBP** (blood-brain barrier permeability),
- **CYP3A4** inhibition (metabolism),
- **DILI** (drug-induced liver injury),
- **HIA** (human intestinal absorption),
- **PGP** (P-glycoprotein substrate classification).

TDC provides predefined scaffold-based data splits with an 8:1:1 ratio for train, validation, and test sets. This splitting strategy ensures that structurally dissimilar compounds are separated across subsets, encouraging generalization to novel scaffolds.

Although TDC provides leaderboards for these tasks, many top-performing entries lack reproducible code or working implementations. For instance, the authors of one of the top submissions explicitly acknowledge on GitHub that their code is not runnable.[§] Therefore, we opt to use robust foundation models—FARM, ChemBERTA-2, and a GNN—for ensemble learning.

- **FARM** (Nguyen et al., 2024a): A SMILES-based BERT model trained with functional group-aware tokenization.

[‡] https://tdcommons.ai/benchmark/admet_group/overview/

[§] https://github.com/maplightrx/MapLight-TDC

- **ChemBERTA-2** (Ahmad et al., 2022): A large-scale transformer model trained on millions of canonical SMILES sequences.
- **GNN** (Edwards et al., 2024a): A graph neural network trained on molecular graphs with atom- and bond-level features.

To build the ensemble, we use each model as a feature extractor. The extracted features are concatenated and passed through a fully connected layer for final prediction. This ensemble approach is stacking, where multiple base learners feed into a meta-learner. For each task, we select a threshold that maximizes the F1 score on the validation set. This threshold is then used to binarize the predicted logits into class labels. The performance of our ensemble model across the selected tasks is summarized in Table 7.

Table 7: Performance (AUC) of individual models and the ensemble across six selected ADMET tasks.

Model	AMES	Pgp	DILI	BBBP	CYP3A4	HIA
FARM (Nguyen et al., 2024a)	0.88	0.89	0.79	**0.94**	0.88	0.92
GNN (Edwards et al., 2024a)	0.75	0.78	**0.86**	0.79	0.80	0.81
ChemBERTa-2 (Ahmad et al., 2022)	0.86	0.89	0.81	0.93	0.86	**0.99**
Ensemble	**0.89**	**0.91**	0.84	0.93	**0.89**	**0.99**

E TRAINING PROCEDURE

We employed Qwen2.5-3B (Yang et al., 2024) as the starting LLM for building the mCLM. Generally, we followed the training procedure from LLaVa (Liu et al., 2023a; 2024b; Li et al., 2024a). We used a two-layer MLP with PReLU activation (He et al., 2015) as an adapter into the LLM input/output from the GNN. We selected an initial learning rate of 1e-5 for the full model and 1e-6 for the adaptor and LM heads. Further, we used a cosine annealing schedule with a minimum of 1e-6 and 2000 linear warmup steps; AdamW (Loshchilov and Hutter, 2017) optimizer was employed. The model was trained on 4 A100 80GB GPUs in bfloat16 precision.

We found that the model learned the molecule tokens much slower than the text (there is usually a 10x difference in loss value). Molecule tokens are rarer and show up less in the training data. Because of this, we decided to separate the molecule classifier head and the language classifier head. We used a standard autoregressive language modeling loss for both, and we averaged these two losses for the final loss value. The main part of our training experiments focused on minimizing the molecule loss, since the text loss was easy to optimize. Further, we found PEFT (Hu et al., 2021) was not sufficient to adapt to molecules, so full finetuning was required. Roughly 10-50 examples from each synthetic (data source, task) pair were put into a validation set.

To initialize the GNN weights, we employed the MolCLR (Wang et al., 2022b) unsupervised contrastive learning technique. We used AugliChem (Magar et al., 2022) for the augmentations: random atom masking, random bond deletion, and motif removal. One of these augmentation was selected uniformly at random for each data point. The GNN was initialized using a batch size of 128 and lr of 1e-4 with a cosine schedule. The model was trained on all 800k blocks in the full data until convergence on a validation set. We tested embedding dimensions between 16 and 4,096 and found 128 dimensions to be sufficient while minimizing total memory cost. This was necessary because we stored the entire embedding matrix in GPU memory, which was much faster, but consumed about 20GB VRAM. Doing so allowed us to train without the GNN during our pretraining process, which is considerably more efficient. We note the GNN can then be finetuned along with the rest of the mCLM during finetuning to new types of molecules or specific tasks. While we did consider a sampled softmax to train the mCLM, we found this to limit the learning of the model.

For training the mCLM, we used two stages for pretraining. First, we trained for 1 epoch with everything frozen except the adaptor, to allow the adaptor to adjust to the LLM representation space. For the second stage, we trained for 5 epochs with only the GNN embeddings frozen. As discussed in the training data mixture section C, we used the most frequent 1000 building blocks as our vocabulary.

After pretraining, we finetune the mCLM to standardize it's outputs for our experiments. During pretraining, we train for robustness by using a wide variety of responses (e.g., for BBBP prediction we might respond "It is restricted from entering the central nervous system" instead of 'No'). For finetuning, we train with standardized responses for our desired tasks (e.g., "Generate a molecule that has higher blood brain barrier permeability than [MOL] ... [/MOL].", "[MOL] ... [/MOL]". Due to our downstream tasks, we finetuned exclusively on the molecule optimization task for 5 epochs over 100k examples for each property. We trained using the same procedure as the pretraining stage, but we selected the best model using validation loss.

F TOKENIZER DETAILS

F.1 SYNTHESIS-GUARANTEED TOKENIZER DETAILS

The synthesis-guaranteed tokenizer disconnects the molecule only at bonds that can be formed by a predetermined small set of reactions, preferably only those that can be performed in an automated manner. For instance, if amide bond formation is defined as available, the tokenizer will be able to disconnect amide bonds in the molecule of interest. Up to this point, the protocol is synonymous with classical computational retrosynthesis, but there is a fundamental difference. Namely, the sets of reactions suitable for automated synthesis is very limited – in fact, state-of-the-art synthesis machines utilize only three types of disconnections (amide bond formation, as well as Suzuki and Buchwald-Hartwig couplings). This places very stringent requirements on the groups that can be present in the disconnected blocks – for instance, when the disconnection (say, Buchwald-Hartwig coupling) yields an amine functionality on one of the blocks, this block cannot contain any groups that during the anticipated uses of this block would present a synthetic incompatibility. In the most trivial case, the block cannot contain another unprotected amine because after the Buchwald-Hartwig disconnection, the block would feature two amines which, in turn, would present competing reactive sites (in the Buchwald-Hartwig synthesis but also in the formation of amide bonds). Therefore, the tokenizer performs retrosynthetic operations while simultaneously checking if they do not lead to blocks with functional groups presenting competing reactivities. Only disconnections avoiding such problems are allowed. This then guarantees that when the corresponding blocks are used to make other molecules, they give only the selective synthesis outcomes. The tokenization process maintains the integrity of the molecules: each molecule can be reconstructed exactly from the building-block sequence, whether produced by tokenization or generated by the mCLM. Each tokenized building block contains placeholder atoms (e.g., [1*], [2*], [3*]) that encode connection points. These placeholders allow us to preserve the full structural integrity of the original molecule, enabling lossless reconstruction from a sequence of tokens. For example, as shown in Figure 1, the drug imatinib with SMILES string

```
Cc1ccc(cc1Nc2nccc(n2)c3cccnc3)NC(=O)c4ccc(cc4)CN5CCN(CC5)C
```

is tokenized into:

```
[ [3*]c1cccnc1,
  [2*]c1ccnc(N[1*])n1,
  [1*]Nc1ccc(C)c([2*])c1,
  [3*]C(=O)c1ccc(CN2CCN(C)CC2)cc1 ]
```

Following the connection rules embedded in the placeholders, the original molecule can be fully reconstructed. For example, [3*] connects to [2*], [1*] connects to [2*], and so on. This process is fully deterministic and reversible. The placeholder atoms and their positions are also encoded by the GNN encoder, allowing the model to reason about both structural and functional context during generation.

F.2 MOLECULE TOKEN EXAMPLES

As discussed in E, we pretrained our GNN encoder using MolCLR (Wang et al., 2022b) on all $\tilde{8}$00k blocks in our full pretraining dataset. Figures 8-13 shows six examples of random blocks and their 5 nearest neighbors in our dataset. The similarity of the blocks and their neighbors shows the high quality representations which were obtained.

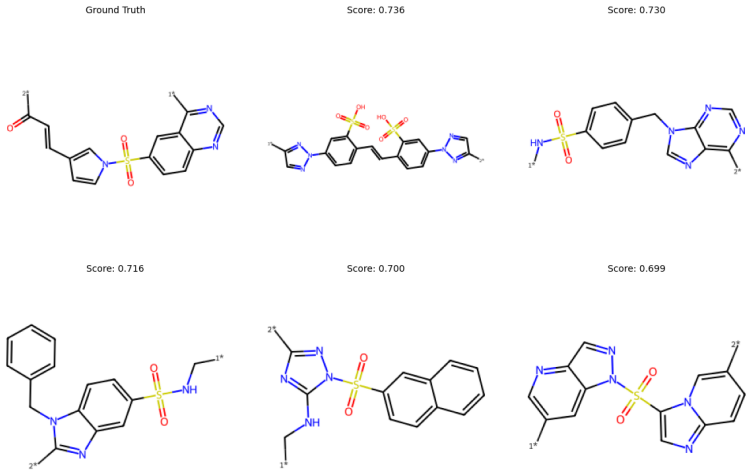

Figure 8: Example token from pretraining dataset. Similarity above each molecule is ECFP4 Fingerprint Tanimoto similarity from the ground truth molecule.

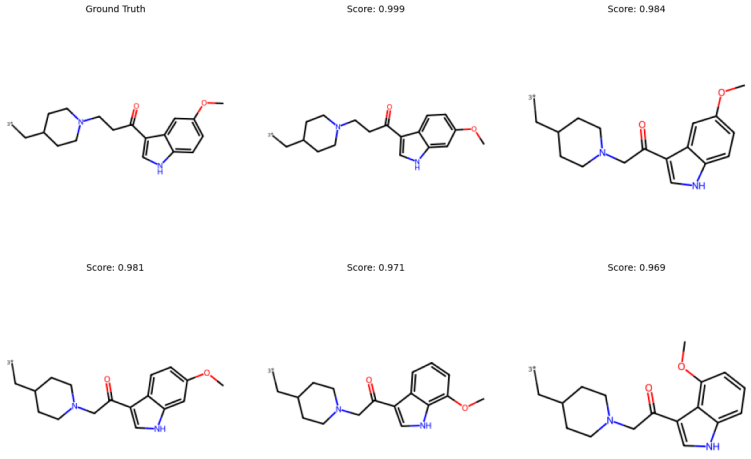

Figure 9: Example token from pretraining dataset. Similarity above each molecule is ECFP4 Fingerprint Tanimoto similarity from the ground truth molecule.

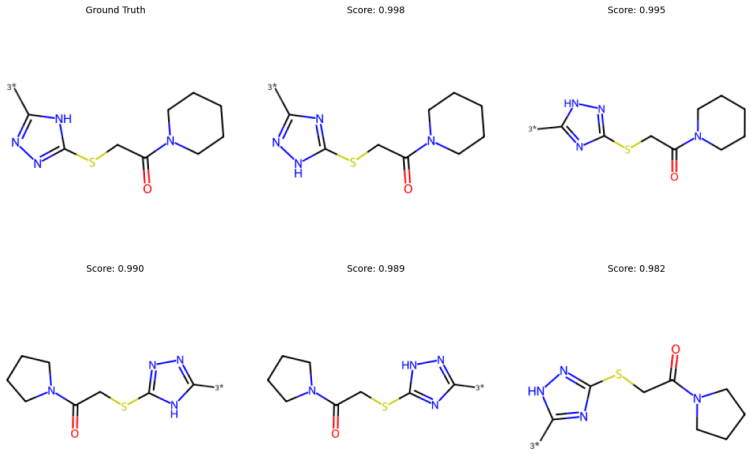

Figure 10: Example token from pretraining dataset. Similarity above each molecule is ECFP4 Fingerprint Tanimoto similarity from the ground truth molecule.

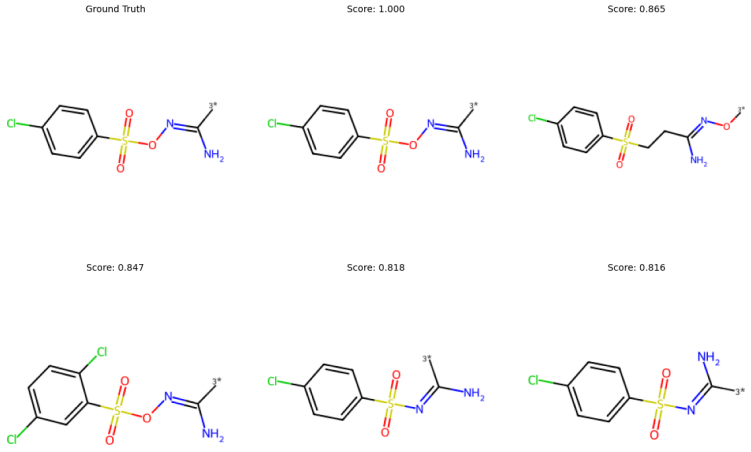

Figure 11: Example token from pretraining dataset. Similarity above each molecule is ECFP4 Fingerprint Tanimoto similarity from the ground truth molecule.

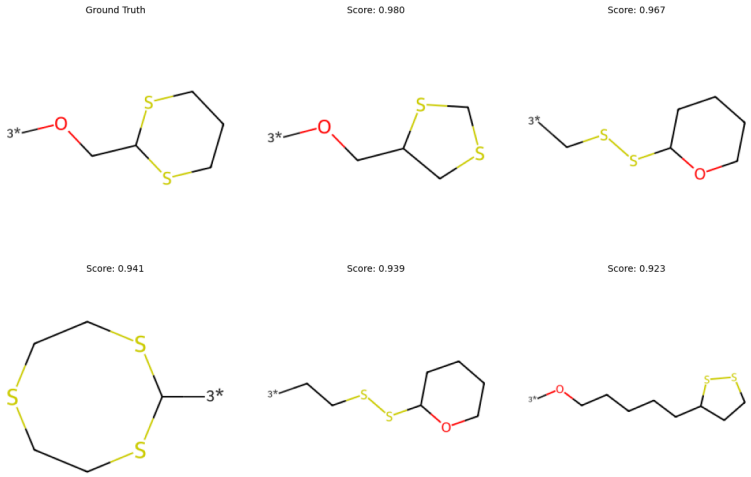

Figure 12: Example token from pretraining dataset. Similarity above each molecule is ECFP4 Fingerprint Tanimoto similarity from the ground truth molecule.

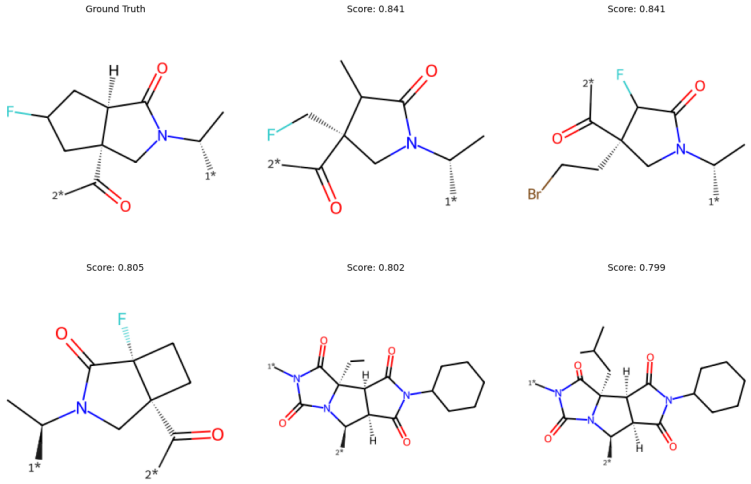

Figure 13: Example token from pretraining dataset. Similarity above each molecule is ECFP4 Fingerprint Tanimoto similarity from the ground truth molecule.

G SOURCE DATASETS AND DATABASES

- Leffingwell odors (Sanchez-Lengeling et al., 2019)
- BACE (Wu et al., 2018)
- Flashpoint (Sun et al., 2020)
- MUV (Wu et al., 2018)
- Tox21 (Wu et al., 2018)
- AMES (Huang et al., 2021)
- Bioavailability (Huang et al., 2021)
- Caco2 (Huang et al., 2021)
- Carcinogens (Huang et al., 2021)
- cav3_t (Huang et al., 2021)
- choline_transporter (Huang et al., 2021)
- clearance hepatocyte (Huang et al., 2021)
- CYP1A2 (Huang et al., 2021)
- CYP2C9 (Huang et al., 2021)
- CYP2D6 (Huang et al., 2021)
- CYP3A4 (Huang et al., 2021)
- DILI (drug induced liver toxicity) (Huang et al., 2021)
- Half_life (Huang et al., 2021)
- hERG (Huang et al., 2021)
- HIA (human intestinal absorption) (Huang et al., 2021)
- Hydration free energy (Huang et al., 2021)
- kcnq2 (Huang et al., 2021)
- LD50 (Huang et al., 2021)
- Lipophilicity (Wu et al., 2018)
- m1_muscarinic (Huang et al., 2021)
- OPV data (Nguyen et al., 2024b; Lopez et al., 2016)
- orexin_receptor (Huang et al., 2021)
- PAMPA_NCATS (Huang et al., 2021)
- Broccatelli (Broccatelli et al., 2022)
- potassium_ion_channel (Huang et al., 2021)
- PPBR (Plasma Protein Binding Rate) (Huang et al., 2021)
- Pubchem logP (Kim et al., 2023)
- SARSCoV2_3CL (Huang et al., 2021)
- SARSCoV2_vitro (Huang et al., 2021)
- serine_threonine_kinase_33 (Huang et al., 2021)
- Skin Reaction (Huang et al., 2021)
- Solubility_AqSolDB (Huang et al., 2021)
- tyrosyl-dna_phosphodiesterase (Huang et al., 2021)
- VDss (volume of distribution at steady state) (Huang et al., 2021)
- Molecule Property Cliff Datasets (30+ datasets) (Van Tilborg et al., 2022)
- Chemical Function (CheF) (Kosonocky et al., 2023)
- ChemFOnt: the chemical functional ontology resource (Wishart et al., 2023)
- Pubchem properties (Kim et al., 2023)
- FreeSolv (Wu et al., 2018)
- QM8 (Wu et al., 2018)

- QM9 (Wu et al., 2018)
- Thermosol (Wu et al., 2018)
- ESOL (Estimated SOLubility) (Wu et al., 2018)
- Lipo (Wu et al., 2018)
- BBBP (Gaulton et al., 2012)
- ClinTox (Wu et al., 2018)
- HIV (Wu et al., 2018)
- SIDER (Wu et al., 2018)
- Forward synthesis (USPTO) (Yu et al., 2024)
- Retrosynthesis (Yu et al., 2024)
- CheBI-20 (Edwards et al., 2021)
- L+M-24 (Edwards et al., 2024c)
- HBA (Gaulton et al., 2012)
- HBD (Gaulton et al., 2012)
- MolWt (Gaulton et al., 2012)
- NumAromaticRings (Gaulton et al., 2012)
- rotatable_bonds (Gaulton et al., 2012)
- TPSA (topological polar surface area) (Gaulton et al., 2012)

H Additional Baselines

We conducted additional comparisons with two relevant baselines: DGAE (Boget et al., 2024) (a discrete graph autoencoder using vector quantization) and HierVAE (Jin et al., 2020) (a hierarchical VAE for molecular graph generation). Note that DGAE's synthesizability scores are computed using a different retrosynthesis tool (AiZynthFinder (Genheden et al., 2020)) and thus are not directly comparable to the Allchemy-based scores in the main text. Besides, HierVAE is trained separately for each property, giving it an unfair advantage in per-property optimization, while mCLM operates via zero-shot instruction prompting. Quite interestingly, we find a significant disagreement between the less-rigorous SAscore and actual retrosynthesis metrics, where mCLM outperforms significantly on the more grounded retrosynthesis score. We speculate this is because SAScore prefers simpler molecules or those with simpler substructures, whereas retrosynthesis tools evaluate actual synthetic feasibility rather than only based on structural patterns. We find that HierVAE is effective for optimizing properties, especially properties that already have good scores, such as CYP3A4 and HIA. Still, mCLM outperforms it on average.

We would also like to explain the technical differences between mCLM and vector-quantized methods, and why we adopt our current approach instead of a vector-quantization technique. Vector-quantized autoencoders (e.g., VQ-VAE) are commonly used to build discrete vocabularies for downstream models. However, in the context of molecular design, these learned codebooks lack chemically grounded guarantees. In contrast, our vocabulary consists of synthesis-verified building blocks, derived from retrosynthesis constraints and domain knowledge, ensuring that all generated molecules are compatible with automation-friendly synthesis. Further, applying vector quantization in this domain risks collapsing chemically distinct structures into the same token, reducing interpretability and limiting extension to new blocks at inference-time. Instead, we use a GNN-based encoder pretrained on chemical properties to produce semantically meaningful embeddings of molecular blocks, which are aligned with a language model via adapter layers. This design allows mCLM to retain both chemical fidelity and generation flexibility, while supporting plug-and-play extensions and instruction-based editing.

Table 8: Synthesizability comparison using AiZynthFinder. Note: these metrics are not directly comparable to the Allchemy-based results in the main paper.

Model	Validity (%)	Synthesizability (%)	Makeability (%)	SA Score (↓)
DGAE (VQ-based)	**100.0**	24.2	24.2	4.46
HierVAE	99.7	67.3	67.2	2.35
mCLM (Ours)	**100.0**	**93.7**	**93.7**	**2.23**

Table 9: Comparison of pharmacokinetic and toxicity property scores between mCLM and HierVAE. Note that HierVAE is fine-tuned separately for each property, whereas mCLM is used in a instruction-following setting.

Model	AMES (↓)	BBBP (↑)	CYP3A4 (↓)	DILI (↓)	HIA (↑)	PGP (↓)	Avg. Improv.
HierVAE	48.2	66.4	**1.2**	53.1	**99.7**	**64.3**	10.5%
mCLM (Ours)	**44.4**	**85.2**	1.4	**53.7**	98.99	64.4	**15.0%**

I Ablation Studies

As detailed in the main text, we performed two ablations of the mCLM block encoding process. First, we trained a SMILES-based version of the mCLM without a GNN (No GNN), still using a Qwen-3B backbone. Second, we conducted a test of the mCLM using blocks extracted from the BRICS algorithm instead of from our tokenization method (No Synth. Tokenizer). These showed significantly degraded property modification capabilities compared to the mCLM.

We also did an ablation study of using Llama-3.2-3B as the backbone for mCLM. We find that, as also noted by others in the community, with the same backbone size, Qwen works better when fine-tuned on downstream tasks than Llama. Specifically, we find evidence that the Llama-3.2-3B-mCLM is overfitting compared with Qwen-2.5-3B-mCLM, evidenced by a lower training loss but higher validation loss (3.702 v.s. 3.620). We also evaluated the property optimization capability of Llama-3.2-3B-mCLM as follows:

Table 10: Comparison of pharmacokinetic and toxicity property scores between using Qwen-2.5-3B and Llama-3.2-3B as backbones.

Model	AMES (↓)	BBBP (↑)	CYP3A4 (↓)	DILI (↓)	HIA (↑)	PGP (↓)	Avg. Improv.
mCLM (Llama-3.2-3B)	48.2	59.4	1.7	59.0	98.1	**63.9**	2.82%
mCLM (Qwen-2.5-3B)	**44.4**	**85.2**	**1.4**	53.7	**98.99**	64.4	**15.0%**
mCLM (No GNN)	50.4	46.0	2.5	**50.2**	97.9	71.7	-7.53%
mCLM (No Synth. Tokenizer)	48.7	55.0	2.9	54.9	98.3	68.3	-19.0%