

Retrieval-Augmented Foundation Models for Matched Molecular Pair Transformations to Recapitulate Medicinal Chemistry Intuition

Bo Pan

Department of Computer Science,
Emory University
Atlanta, GA, USA
bo.pan@emory.edu

Alex Zhu

Department of Computer Science,
Emory University
Atlanta, GA, USA
alex.zhu@emory.edu

Peter Zhiping Zhang

Merck & Co., Inc.
Rahway, NJ, USA
zhiping.peter.zhang@merck.com

Xiang Yu

Merck & Co., Inc.
Rahway, NJ, USA
xiang.yu2@merck.com

Liang Zhao

Department of Computer Science,
Emory University
Atlanta, GA, USA
liang.zhao@emory.edu

Hao-Wei Pang

Merck & Co., Inc.
Rahway, NJ, USA
hao-wei.pang@merck.com

Liying Zhang

Merck & Co., Inc.
Rahway, NJ, USA
liying.zhang@merck.com

Abstract

Matched molecular pairs (MMPs) captures the local chemical edits that medicinal chemists routinely use to design analogs, but existing ML approaches either operate at the whole-molecule level with limited edit controllability or learn MMP-style edits from restricted settings and small models. We propose a variable-to-variable formulation of analog generation and train a foundation model on large-scale MMP transformations (MMPTs) to generate diverse variables conditioned on an input variable. To enable practical control, we develop prompting mechanisms that let the users specify preferred transformation patterns during generation. We further introduce MMPT-RAG, a retrieval-augmented framework that uses external reference analogs as contextual guidance to steer generation and generalize from project-specific series. Experiments on general chemical corpora and patent-specific datasets demonstrate improved diversity, novelty, and controllability, and show that our method recovers realistic analog structures in practical discovery scenarios.

1 Introduction

In drug discovery, a fundamental strategy to optimize lead molecules is *analog design*, which involves medicinal chemists leveraging their intuition to make localized, knowledge-driven edits to existing molecules, instead of designing entirely novel molecules [6, 11, 36]. Machine learning models for molecular optimization have increasingly adopted transformation-based formulations, learning to transform one molecule into another through graph edits or sequence-to-sequence generation [14, 15, 38, 39]. However, most such approaches treat transformations as implicit, molecule-level operations, without distinguishing which edits correspond to chemically meaningful local modifications versus arbitrary global rewrites [32]. In contrast, medicinal chemists typically reason in terms of

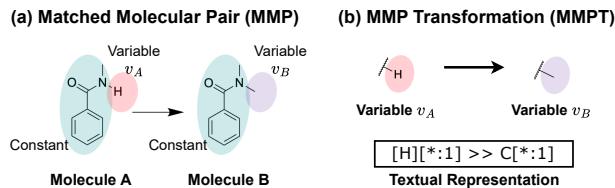


Figure 1: An example of (a) Matched Molecular Pairs (MMP); (b) Matched Molecular Pair Transformation (MMPT) and its textual representation.

matched molecular pairs (MMPs) [2, 6, 10, 11, 13, 28, 33, 48], which are pairs of molecules that differ by discrete and minimal modifications, such as R-group substitutions or core replacements, that preserve synthetic feasibility and support interpretable structure–activity comparisons, as shown in Fig. 1 (a). Abstracting away the constant chemical context and focusing solely on the localized edit yields an **MMPT**, which captures a context-independent medicinal chemistry modification corresponding to a single, well-defined *variable-to-variable* change, as an example shown in Fig. 1 (b). MMPTs directly reflect how chemists use their medicinal chemistry intuition to optimize lead molecules, and provide a principled way to represent and learn transferable medicinal chemistry modifications across different molecular contexts.

While MMPTs have appeared occasionally in machine learning-based molecular optimization, they remain underexploited as a first-class generative representation: prior models are typically limited in scale or trained with specific data [5, 17], or embedded within molecule-level generators where MMP relationship is only weakly enforced [14, 15, 38, 39]. At the same time, recent advances in large-scale chemical corpora [12, 37, 50] and foundation models [1, 3]

make it newly feasible to learn generalizable, MMPT-level priors directly from millions of real data points. This convergence creates a timely opportunity to revisit MMPTs: not as an auxiliary constraint, but as the central abstraction for controllable and scalable analog generation.

Despite their conceptual appeal, realizing MMPT-centric learning within modern ML systems poses several technical challenges. First, many existing analog generation models are optimized for molecule-level similarity rather than explicit localized modifications [14, 38], making it difficult to guarantee that generated candidates differ from the input by a single, well-defined transformation. As a result, users cannot reliably specify where a modification should occur (e.g., which R-group or core), nor ensure that unintended global changes are avoided. Second, prior MMPT-based learning approaches are often constrained to small models, limited transformation vocabularies, or narrowly curated datasets, which restrict their ability to absorb the long-tailed, heterogeneous transformation patterns observed across large chemical corpora. Third, existing controllable methods, whether graph-editing models with fixed operators [5, 17] or prompt-based use of large language models [4, 43, 46], lack mechanisms to learn transferable, transformation-level priors that generalize across scaffolds while remaining interpretable and synthetically plausible. Finally, practical deployment in medicinal chemistry requires models to adapt to user- or project-specific preferences, such as emphasizing rare but relevant modifications or following established series patterns, without costly retraining. Addressing these challenges requires moving beyond molecule-level generators toward scalable, controllable models that operate directly in the MMPT space.

Motivated by the above complementarity and limitations, we aim to build a practical MMPT-centric generation framework that synergizes chemists’ intuition with the ability of modern ML to learn from massive data. First, we train an MMPT foundation model on large-scale transformation data to generate variables conditioned on input variables, enabling high-throughput analog design in a user-controlled edit region. Second, we develop prompting mechanisms that expose substructure-level control to users, allowing them to customize the structural patterns of generated variables. Third, we introduce an MMPT-RAG framework that incorporates external reference datasets: retrieved analogs are organized into structural clusters and used to reweight the generation distribution, improving coverage of infrequent yet chemically meaningful transformations while preserving plausibility.

Our main contributions are summarized as follows:

- We formalize analog generation in the matched molecular pair transformation (MMPT) space, treating analog design as context-independent local edits that can be composed across diverse molecular scaffolds.
- We train a foundation model on large-scale MMPTs extracted from a broad corpus of drug-like molecules, and enable controllable generation via prompting mechanisms that specify desired transformation templates or structural patterns.
- We propose an MMPT-RAG framework that leverages external reference datasets by retrieving structurally related

examples and using them as contextual guidance to steer generation toward user-preferred patterns.

- We validate our approach on three complementary MMPT benchmarks: an in-distribution setting, a within-patent analog expansion setting, and a cross-patent generalization setting. Across tasks, our method consistently improves recall of ground-truth transformations while maintaining high validity and producing non-trivial novel edits.

2 Related Work

2.1 MMP-Based Analog Generation

Most of the existing MMP-based analog design methods explore the setting of molecule-level generation, i.e., generating entire molecules that are similar to the given molecule [14, 15, 18, 38, 39], and thus this stream of methods does not support users specifying a substructure (variable) to edit. Some existing methods also allow users to specify a structure to change, with one stream of work formulating it as a partial molecule generation problem, where the user-provided substructure is fixed, and the model is tasked with completing the remaining molecule, including those VAE-based graph generative models which operate via node generation [17, 18, 23, 49] and auto-regressive generative models that implements it as a token generation task in the SMILES space [24, 31]. With the advancements of large language models (LLMs), some work also explored leverage LLMs’ zero-shot ability to suggest variable replacements [4, 43, 46]. Among these methods, the Lib-INVENT module [24] in REINVENT 4 [24] stands out as a strong baseline with a wide industrial applicability, benefiting from its large training corpus. However, all the above methods are trying to learn the conditioned molecular completion task (constant to variable). To our best knowledge, no existing method tries to directly learn a context-independent variable replacement objective (variable to variable).

2.2 Retrieval-Augmented Generation for Molecule Generation

Recent advances in retrieval-augmented generation (RAG) have been applied with notable success to molecular and materials design. For example, RetMol [44] uses an exemplar retrieval module to guide a pretrained molecular generator by fusing input compounds with retrieved analogues, enabling efficient design of molecules satisfying complex properties without task-specific fine-tuning. [22] propose f-RAG, which retrieves both “hard” and “soft” fragment contexts to steer a fragment-based generative model, thereby improving diversity and design novelty. Structure-based methods such as Rag2Mol [51] and IRDiff [16] integrate retrieval of known ligands or fragments and inject them into 3D generation via autoregressive graph or diffusion models, aligning generation to target binding pockets. [47] further extend this paradigm by introducing READ, an SE(3)-equivariant diffusion model aligned with retrieval of scaffold embeddings to enhance geometric and chemical realism. Finally, [20] demonstrate the flexibility of RAG techniques in materials science, coupling literature retrieval with LLM-based generative reasoning for nanostructured material design.

3 Problem Formulation

3.1 MMPTs as the Generative Unit

In medicinal chemistry, analog design proceeds by modifying an existing compound through a single, localized structural change, such as replacing a substituent, linker, or core, while keeping the remainder of the molecule fixed. This concept is formalized through matched molecular pair transformations (MMPTs). An MMPT is defined as the transformation linking a pair of molecules that share an invariant chemical context, while differing by a single, well-defined variable fragment, as illustrated in Fig. 1. Formally, an MMPT can be represented as $(v_A \rightarrow v_B)$, where v_A, v_B are referred to as the *variables* before and after the transformation. By construction, MMPTs isolate minimal chemical edits that are synthetically feasible and empirically validated through historical discovery efforts. As such, MMPTs constitute the primary unit through which medicinal chemists reason about structure–activity relationships (SAR) and explore local chemical space.

3.2 Problem Definition: MMPT-Centric Analog Generation

Given an input variable v_A , our goal is to generate chemically plausible alternative variables that correspond to valid MMPTs. Concretely, given an input variable v_A , we aim to generate a set of candidate substitutions $\{v_B^{(1)}, v_B^{(2)}, \dots\}$ such that each pair $(v_A \rightarrow v_B^{(i)})$ constitutes a valid MMPT.

This formulation differs fundamentally from molecule-level analog generation, where edits are implicit and entangled across the structure. By operating directly in the MMPT space, the task becomes learning conditional distributions over chemical transformations, rather than over entire molecules.

Although MMPTs provide a natural abstraction for localized analog design, learning to generate MMPTs at scale poses several challenges. First, the space of MMPT is large, sparse, and highly imbalanced: a small number of common substitutions dominate, while many chemically meaningful transformations occur rarely, making it difficult for task-specific or small models to learn transferable priors. Second, models grounded in formal chemical languages preserve precise structural semantics but are inherently rigid and difficult to instruct through user prompts; on the other hand, natural language models support flexible prompting, but often lack explicit structural constraint enforcement and are not specialized for chemical data. It is fundamentally challenging to retain chemically grounded priors while achieving user controllability. Third, practical deployment in medicinal chemistry requires explicit distributional steering, enabling users to bias generation toward preferred or project-specific MMPT patterns without retraining.

4 Methodology

To address the challenges outlined above, we propose a two-component framework centered on MMPTs. First, to address the challenges of the large, long-tailed MMPT search space and the challenge in enabling flexible user control, we introduce MMPT-FM, a foundation model trained directly on large-scale MMPT data. By modeling variable-to-variable transformations in a chemically grounded token space and supporting structured prompting, MMPT-FM learns generalizable transformation-level priors while enabling user-guided

generation. To address the challenge of explicit, project-level controllability without retraining, we introduce MMPT-RAG, a retrieval-augmented generation framework that incorporates external reference analogs as guidance, allowing users to bias generation toward preferred transformation patterns and project-specific needs.

4.1 MMPT-FM: A Promptable MMPT Foundation Model

Here, we construct our foundation model, MMPT-FM, that operates directly in the MMPT space, learning variable-to-variable transformations as the primary generative unit. We further equip it with a prompting mechanism that conditions generation on optional user-specified structural templates, thereby enabling more flexible user control.

4.1.1 Training of MMPT-FM. As described earlier, we train our foundation model on the variables rather than on entire molecules, modeling each MMPT as a localized transformation $v_A \rightarrow v_B$, where both v_A and v_B are represented as SMARTS [9], a textual chemical representation method for substructures. Formally, let Σ denote a finite vocabulary of SMARTS tokens, including atomic symbols (e.g., C, N, O), bond descriptors (e.g., =, #), and other syntactic elements for SMARTS expressions. Each variable is represented as a sequence over Σ as $v_A = (\tau_1^A, \dots, \tau_n^A), v_B = (\tau_1^B, \dots, \tau_m^B)$, where $\tau_i^A, \tau_j^B \in \Sigma$. An MMPT is therefore formulated as a conditional sequence-to-sequence mapping from v_A to v_B , and the model learns the conditional distribution $p(v_B | v_A)$ defined over syntactically valid and chemically plausible SMARTS variables.

To construct the training data, we use MMPDB [8] to extract MMPs from ChEMBL, a large database of drug-like compounds [12]. To enrich for drug-like chemistry, we first filter molecules using the medchem package’s `rule_of_druglike_soft` criterion [30], further retain only compounds with molecular weight ≥ 200 Da, and remove structural alerts using the curated list compiled by Walters [42]. After filtering, 800,714 compounds remain. We then generate MMPs using MMPDB with the maximum variable ratio constrained to 33% (`max-variable-ratio=0.33`), yielding 2.63 million MMPs. For each MMP, MMPDB identifies the shared substructure (the *constant* part) and the differing variable pair (the *variable* parts). Totally, there are approximately 800K distinct MMPTs. We use 90% of these transformations for training and reserve the remaining 10% as a held-out evaluation set.

To model this variable-to-variable translation task, we adopt an encoder–decoder Transformer architecture. The encoder takes as input the variable that the user wishes to replace, and the decoder generates suggested replacement variables. To provide the model with a chemically aware initialization, we initialize from T5Chem [7, 25], a T5-style model pretrained on large-scale chemical tasks, which offers a moderate model size and strong coverage of chemical syntax and semantics.

4.1.2 Prompted Generation with MMPT-FM. In practical medicinal chemistry workflows, analog design is rarely unconstrained; chemists often seek to impose explicit structural conditions, such as preserving a core motif or exploring a specific transformation family, making it essential for an MMPT generation framework to support user-specified structural patterns. Unlike natural language

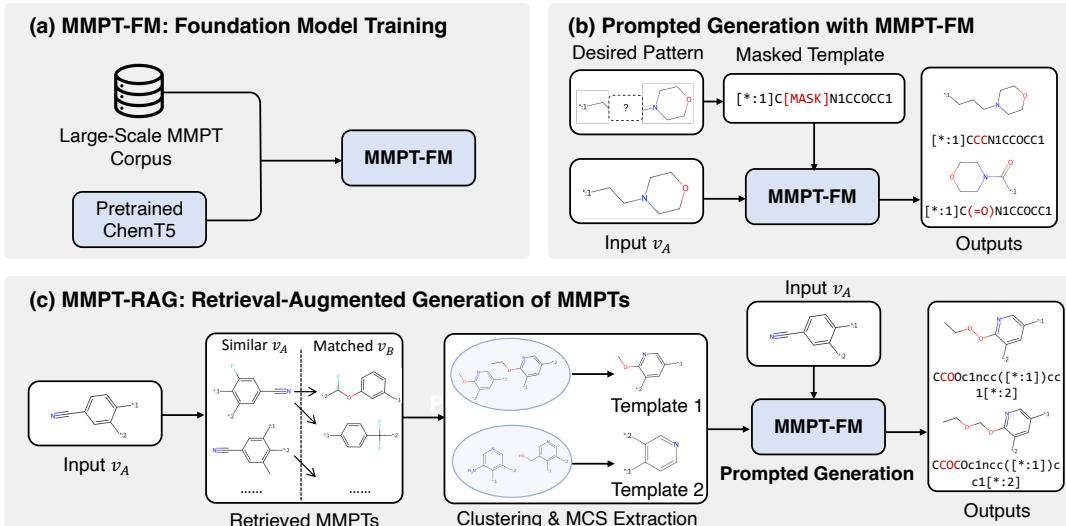


Figure 2: Overview of the proposed MMPT framework. (a) The foundation model (MMPT-FM) is trained on large-scale MMPT data. (b) MMPT-FM supports controllable generation via masked template prompting. (c) MMPT-RAG augments generation with retrieval, clustering, and MCS-based template extraction to guide context-aware transformation generation.

models, where intent can be expressed through appended text instructions, MMPT-FM operates purely in chemical token space, so desired structural patterns cannot be specified as free-form text. To preserve the variable-to-variable formulation while enabling control, we encode user intent as partial structural constraints on the output.

The prompted generation process is illustrated in Fig. 2 (b). Specifically, we formulate prompted generation of MMPT-FM as a constrained sequence completion task, where the user can impose a structural constraint S that defines the preferred chemical substructure to be preserved in the generated v_B , with some undefined positions that the model is asked to complete. Formally, we define a masked template T as $T = (\tau_1, \tau_2, \dots, \tau_L)$, where each $\tau_i \in \Sigma \cup \{[\text{MASK}]\}$. The mapping from a user-specified structural constraint S to a textual template T is operated by preserving the chemical tokens corresponding to the fixed substructure S while replacing the undefined positions with $[\text{MASK}]$ tokens.

The generation objective is to find complete variable sequences v_B that are compatible with T . At inference time, the generator performs approximate inference to produce a set of candidate variables $\mathcal{V}_B = \{v_b^{(1)}, \dots, v_b^{(K)}\}$ that complete the masked positions conditioned on T . These candidates are selected to have high likelihood under the model as

$$\text{PromptGen}(v_A, T, K) = \operatorname{argmax}_{S, |\mathcal{S}|=K} \sum_{v_B^{(i)} \in \mathcal{S}} \log p_\theta(v_B^{(i)} | v_A, T), \quad (1)$$

where θ denotes the parameters of the trained MMPT-FM, and K is the desired number of output variables.

We implement masked infilling via an explicit search over the space of possible span completions, using the model likelihood to score each candidate. To make this search tractable, in practice, we limit the branching factor at each masked position by selecting only the effective number N_{eff} [19] of high-probability tokens, calculated by $N_{\text{eff}} = 2^{H_2(p)}$, where p is the token probability distribution of

generating each token, and $H_2(p) = -\log_2 \sum_i p_i^2$ is the Rényi entropy [35] of order 2 of the token distribution p . Starting from the initial masked template, we conduct a tree search by iteratively filling masked positions with these candidate tokens, and accumulating sequence-level log-likelihoods from the model. Finally, we rank completed candidates by their likelihood and return the top- k infilled fragments as the model’s suggested variable substitutions.

4.2 MMPT-RAG: A RAG Framework for MMPT Generation

MMPT-FM learns a global prior over chemically meaningful MMPTs from large-scale data, whereas practical design often relies on project-specific reference analogs. A principled generator should therefore interpolate between general chemical knowledge and project-specific patterns, rather than overriding one with the other. We realize this idea through MMPT-RAG, which integrates retrieval as guidance (Section 4.2.1) and admits a theoretical interpretation as an explicit distribution shift from model prior to reference dataset distribution (Section 4.2.2).

4.2.1 Workflow. MMPT-RAG guides the generation towards a reference database \mathcal{D} . Let $\mathcal{D} = \{(v_A^{(i)} \rightarrow v_B^{(i)})\}_{i=1}^N$ denote a reference database of MMPTs. Given an input variable v_A , the framework retrieves relevant MMPTs, clusters them to identify representative patterns, and converts each cluster into a structural template. These templates are then used to prompt MMPT-FM via masked infilling, as illustrated in Fig. 2 (c). The workflow proceeds in three steps detailed below.

Step 1: Retrieval with input similarity. To leverage the useful examples from the retrieval dataset, we retrieve variables structurally similar to the query. Let $\psi(\cdot)$ denote an embedding function, which is usually implemented with the Morgan fingerprint [29]. For query v_A , we retrieve top- K nearest neighbors \mathcal{V}_A from the database \mathcal{D}

as

$$\begin{aligned} \mathcal{V}_A &= \text{Retrieve}(v_A; \mathcal{D}) \\ &= \left\{ v_A^{(i)} \mid (v_A^{(i)} \rightarrow v_B^{(i)}) \in \mathcal{D}, i \in \text{TopK} \left(\text{sim}(\psi(v_A), \psi(v_A^{(i)})) \right) \right\}, \end{aligned} \quad (2)$$

where $\text{sim}(\cdot)$ denotes a function to calculate similarity, usually implemented using cosine similarity, yielding candidate contexts $v'_A \in \mathcal{V}_A$. Given a retrieved set of input variables $\mathcal{V}_A = \{v_A^{(i)}\}$, we denote by $\mathcal{V}_B = \{v_B^{(j)}\}$ the set of all variables such that there exists some i, j that $(v_A^{(i)} \rightarrow v_B^{(j)})$ forms an MMPT in \mathcal{D} as

$$\mathcal{V}_B = \{v_B \mid \exists v_A \in \mathcal{V}_A, (v_A, v_B) \in \mathcal{D}\} \quad (3)$$

Step 2: Clustering of Retrieved Examples. To extract representative structural patterns in retrieved \mathcal{V}_B , we perform clustering over the retrieved outputs. We compute embeddings $\phi(v_B)$ for all $v_B \in \mathcal{V}_B$ and partition them into K clusters

$$C_1, \dots, C_K = \text{Cluster}(\{\phi(v_B) : v_B \in \mathcal{V}_B\}), \quad (4)$$

using a clustering algorithm, which is implemented as HDBSCAN [27] in this work.

Step 3: Cluster pattern-prompted generation. For each cluster C_k , we first identify its invariant chemical scaffold. The invariant substructure is defined as the Maximum Common Substructure (MCS) [34] among these retrieved molecules as $S_k = \text{MCS}(C_k)$. In practice, this can be automatically computed using the `rdkit.Chem.MCS.FindMCS` function [21], which finds the largest subgraph common to all variables in C_k . Then we construct the masked template T_k for each MCS S_k via the approach introduced in Section 4.1.2. The resulting T_k serves as the cluster-invariant template, which is further used as the prompt to guide the generation.

Given cluster-specific templates $\{T_k\}_{k=1}^K$ and an (optional) user-specified preference distribution $\tilde{\pi}(x)$, MMPT-RAG generates outputs from each template proportionally to its assigned weight. Formally, the RAG output is defined as

$$\text{RAG}(v_A) = \bigcup_{k=1}^K \text{PromptGen}(v_A, T_k, N_k), \quad (5)$$

where N_k denotes the generation budget allocated to cluster k and $N_k \propto \tilde{\pi}_k(x)$, $\tilde{\pi}(x)$ is the user-specified preferred cluster distribution for each cluster C_k :

$$\tilde{\pi}(x) = (\tilde{\pi}_1(x), \dots, \tilde{\pi}_K(x)), \quad \tilde{\pi}_k(x) \geq 0, \quad \sum_{k=1}^K \tilde{\pi}_k(x) = 1, \quad (6)$$

where $\tilde{\pi}_k(x)$ can encode arbitrary user preferences.

4.2.2 Theoretical Analysis. To analyze the RAG mechanism, we formalize the interaction between the foundation model and the reference set as a Bayesian integration. We show that MMPT-RAG performs a global distribution shift toward the reference dataset while maintaining the knowledge of the foundation model

THEOREM 4.1 (GLOBAL STEERING). *Let $p_\theta(y \mid x)$ be the conditional distribution over variables $y \in \mathcal{V}$ defined by the unconstrained foundation model. Assume that for each cluster k , prompting the model with template T_k (via masked infilling) results in a local distribution $p(y \mid x, T_k)$ that is an adaptive interpolation between the model's prior and the cluster-specific reference $p(y \mid T_k)$:*

$$p(y \mid x, T_k) = (1 - \alpha_k)p_\theta(y \mid x) + \alpha_k p(y \mid T_k), \quad (7)$$

where $\alpha_k \in (0, 1]$ is an adaptive gating factor reflecting the model's adherence to template T_k under context x .

Then, the global RAG distribution defined in (5), $p_{\text{RAG}}(y \mid x)$, satisfies:

$$p_{\text{RAG}}(y \mid x) = (1 - \bar{\alpha})p_\theta(y \mid x) + \bar{\alpha}p_{\text{ref}}^*(y \mid x), \quad (8)$$

where $\bar{\alpha} = \sum_k \tilde{\pi}_k \alpha_k$ and $p_{\text{ref}}^*(y \mid x) = \sum_k \frac{\tilde{\pi}_k \alpha_k}{\bar{\alpha}} p(y \mid T_k)$.

Proof of Theorem 4.1. See Appendix B.

Theorem 4.1 shows that the RAG distribution is a convex interpolation between the original foundation model distribution and a reference set distribution. $\bar{\alpha}$ serves as a weight that quantifies the distribution shift.

5 Experiments

In this section, we evaluate our framework systematically through three progressively difficult tasks, which range from generic in-distribution MMPT generation to the prediction of future analogs in subsequent patents (Section 5.2). Following these main results, we provide a decoupled analysis that examines how the model covers chemical space (Section 5.3.1) and follows user prompts (Section 5.3.2) while also demonstrating the ability of retrieval to align generations with specific project domains (Section 5.3.3). The section concludes with a study of hyperparameter sensitivity (Section 5.3.4) and a qualitative review of specific chemical transformations to illustrate the practical utility of the framework (Section 5.4).

5.1 Experimental Setup

5.1.1 Main Experiment Tasks. We evaluate our framework from the perspective of three progressively more realistic analog-generation tasks, each instantiated with a corresponding dataset.

Task 1: In-distribution MMPT Generation. The first task evaluates whether the model can recover and generate valid and novel local transformations under an in-distribution setting. We instantiate this task using the 10% held-out test split from the ChEMBL MMPT dataset, constructed with the same MMPT extraction pipeline as training but with disjoint MMPTs.

Task 2: Within-Patent Analog Expansion. The second task evaluates MMPT generation within a real-world medicinal chemistry project. We construct this setting using the PMV Pharmaceutical patent dataset (PMV17) [40] with MMPDB [8] to extract MMPTs. This task evaluates whether the model can discover promising variables in a realistic setting.

Task 3: Cross-Patent Follow-up Generation. The third task evaluates whether a model can propose forward-looking MMPTs that may appear in later patents, a realistic and challenging evaluation of temporal medicinal chemistry progression. We construct a patent-to-patent setting by extracting MMPTs linking compounds from PMV17 (from 2017) to those appearing in subsequent patents (PMV21) [41] (from 2021), both derived from patent filings by PMV Pharmaceuticals, Inc.

5.1.2 Compared Methods. To the best of our knowledge, no existing method is explicitly designed to operate in the variable-to-variable MMPT formulation. The only directly comparable baseline in the MMPT space is similarity-based **database retrieval**, which is a non-learning method that returns nearest-neighbor variables from

the reference dataset. To further situate our results within established industrial practice, we additionally include **REINVENT4** (LibINVENT module) [24], a state-of-the-art molecule-level analog generation framework. Although we acknowledge that LibINVENT operates on a different objective by conditioning on a fixed constant scaffold rather than the variable part, we still compare with it to demonstrate that our method can perform better even without the auxiliary information. We report both **MMPT-FM**, which only generates with our foundation model, and **MMPT-RAG**, which denotes the full proposed RAG framework.

5.1.3 Evaluation Metrics. We report a consistent set of metrics across tasks to assess validity, novelty, and recoverability of medicinal-chemistry transformations. **Valid** measures the percentage of generated strings that form chemically valid variables and have the same number of attachment points as the input. **Novel** reports the percentage of generated variables not seen during training. Specifically, **Novel/valid** is calculated by the number of novel and valid variables divided by the number of valid variables; **Novel/all** is calculated by the number of novel and valid variables divided by the number of all outputs. **Recall** measures the percentage of ground-truth variables recovered by the model. For patent-based tasks, we further report **Recall-i** and **Recall-o**, which measure recovery of in-training-set and out-of-training-set transformations, respectively. Among them, Novel and Recall-o are the two most important metrics to evaluate the models' performance since both novelty and ability to learn from prior knowledge are critical in generating analogs by mimicking medicinal chemists' intuition.

5.1.4 Implementation Details. All implementation details are given in Appendix A.

5.2 Main Results

5.2.1 Task 1: In-distribution Evaluation on ChEMBL Table 1a reports the results on Task 1. As expected, database retrieval achieves moderate recall (43.5%) but yields no novel outputs, reflecting the inherent limitation of exact analog lookup. REINVENT4 attains higher novelty (23.0%) but suffers from very low recall (12.7%), indicating that unconstrained molecule-level generation struggles to reproduce specific, localized MMP edits even in an in-distribution setting. In contrast, MMPT-FM substantially improves recall to 67.6% while maintaining high validity. Building on this foundation, MMPT-RAG further boosts recall to 82.1% and achieves the highest novelty (30.1%) among all methods. This improvement confirms that MMPT-RAG is a strong complement for the foundation model itself by leveraging less-represented but still related MMPT patterns. Overall, these results show that MMPT-centric modeling is effective in in-distribution MMPT generation, and that retrieval augmentation further strengthens coverage of valid transformation patterns beyond what can be achieved by learning alone.

5.2.2 Task 2: Within-Patent Analog Expansion on PMV17. Table 1b reports results on Task 2. As in Task 1, database retrieval achieves limited recall (22.7%) and completely fails to recover structurally novel transformations, highlighting that exact lookup is insufficient for realistic series expansion. REINVENT4 exhibits very low recall across all metrics. In contrast, MMPT-FM substantially improves overall recall to 41.36% while achieving strong in-training-set recovery (Recall-i = 52.06%) and non-trivial out-of-training-set recall

Table 1: MMPT generation performance on three tasks.

(a) Task 1: ChEMBL MMPT Dataset					
Method	Recall	Novel/valid	Novel/all	Valid	
Database Retrieval	43.5%	0.0%	0.0%	100%	
REINVENT 4	12.7%	23.0%	5.6%	24.4%	
MMPT-FM (Ours)	67.6%	26.0%	25.8%	99.3%	
MMPT-RAG (Ours)	82.1%	30.1%	29.8%	99.1%	
(b) Task 2: PMV17 MMPT Dataset					
Method	Recall			Novel	
	Total	Recall/i	Recall/o	/valid	/all
Database Retrieval	22.7%	29.4%	0.0%	0.0%	0.0%
REINVENT 4	5.1%	7.1%	0.0%	48.2%	15.7%
MMPT-FM (Ours)	41.4%	52.1%	13.2%	23.0%	22.7%
MMPT-RAG (Ours)	49.2%	62.1%	15.2%	23.7%	23.4%
(c) Task 3: PMV17-PMV21 Cross-Patent Generation					
Method	Recall		Recall/i	Recall/o	
	Database Retrieval	28.57%	57.49%	0.00%	
REINVENT 4	7.36%	12.21%	1.87%		
MMPT-FM (Ours)	43.77%	76.45%	11.48%		
MMPT-RAG (Ours)	46.81%	81.35%	12.99%		

Table 2: Effect of beam size on average validity of MMPT-FM, averaged on the ChEMBL MMPT held-out test set.

Beam	400	600	800	1000	1200
Avg Validity	0.9992	0.9991	0.9989	0.9988	0.9986

(Recall-o = 13.15%). MMPT-RAG further improves performance across all metrics, achieving the highest overall recall (49.21%), the strongest in-training-set recovery (62.08%), and the best out-of-training-set recall (15.24%). The gains in Recall indicate that MMPT-RAG effectively helps guide the generator toward a region which is closer to the PMV17 dataset.

5.2.3 Task 3: Cross-Patent Generation (PMV17 → PMV21) Table 1c reports the results for Task 3. Following the same pattern, database retrieval fails entirely on out-of-training transformations, and REINVENT4 exhibits extremely low recall across all metrics. MMPT-FM substantially improves recall by modeling transformation-level priors, achieving a recall of 43.77% and recovering a large fraction of in-training transformations. By incorporating retrieval-augmented prompting, MMPT-RAG further improves performance across all metrics, achieving the highest overall recall (46.81%), in-training recall (81.35%), and out-of-training recall (12.99%).

5.3 Decoupled Analysis

5.3.1 Evaluation of MMPT-FM's Chemical Space Coverage To assess the fundamental generative capability of MMPT-FM, we evaluate its ability to explore the chemical space compared to Database Retrieval. We utilize a randomly sampled group of 50 unique variables from both ChEMBL and PMV17 to generate candidate sets. The structural distributions are visualized by projecting their Morgan fingerprints using PCA, comparing the spans of FM-generated candidates against retrieved variables. Visual analysis of the PCA

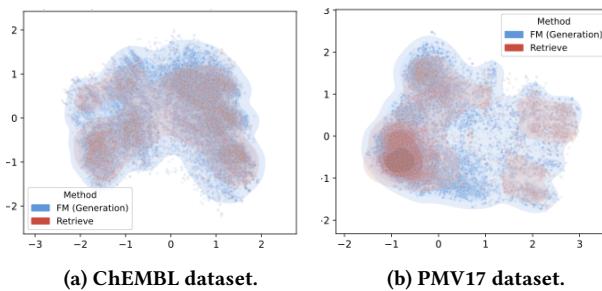


Figure 3: Visualizations of the chemical space explored by our foundation model MMPT-FM (blue) versus Database Retrieval (red) on (a) ChEMBL and (b) PMV17 datasets.

projections (Fig. 3) reveals that MMPT-FM (blue) consistently explores a substantially larger chemical volume compared to Database Retrieval (red), demonstrating superior extrapolation beyond the training distribution.

We further study how beam size affects the validity of MMPT generation on the ChEMBL MMPT held-out test set. As shown in Table 2, validity remains consistently high across all beam sizes, with only a slight decrease from 0.9992 at beam 400 to 0.9986 at beam 1200, which is usually enough for application scenarios. Notably, a beam size of 1200 significantly exceeds the depth typically required for practical medicinal chemistry applications. This indicates that our generator is not validity-bottlenecked by search depth within its intended operational range; increasing beam primarily expands the candidate pool but does not harm chemical correctness.

5.3.2 Evaluation of the Prompted Generation Mechanism of MMPT-FM We evaluate the prompted generation capability of MMPT-FM via a masked infilling task. This task motivates the assessment of whether the model can strictly adhere to user-specified structural templates while proposing chemically plausible completions. We construct a benchmark by randomly sampling 50 unique variables from ChEMBL and PMV17 with lengths exceeding 15 characters. For each variable, three masked versions are generated by masking a consecutive sequence of tokens with a length capped at $\min(\text{half of the output}, 8)$. Here prompted generation is performed using 1,000 beams and a length margin of 7. Results are given in Table 3. At $K = 1$, the model achieves near-perfect validity and high GT recovery (58.0% for ChEMBL, 46.0% for PMV17). By $K = 20$, the model attains near-perfect recall across both datasets. Furthermore, at $K = 200$, the model produces a significant number of unique valid candidates (41.6 for ChEMBL, 31.6 for PMV17), confirming its ability to explore diverse chemical spaces even under rigid structural constraints, which shows the effectiveness of our prompted generation mechanism in generating promising, valid and user-desired variables.

5.3.3 Analysis of Distribution Steering via RAG. To investigate how retrieval augmentation steers the generative process toward target chemical domains, we visualize the global distribution of generated analogs against the reference patent dataset (PMV17). We compare the union of outputs from 50 unique inputs generated via vanilla FM inference versus the MMPT-RAG framework, using the PMV17 dataset as the reference manifold represented by the grey shaded

Table 3: Prompted generation on ChEMBL and PMV17 at different numbers of generations (K). We report Validity, Recall of ground truth, and numbers of generated unique and valid variables (#Uniq).

K	ChEMBL			PMV17		
	Valid (%)	Recall (%)	#Uniq	Valid (%)	Recall (%)	#Uniq
1	100.0	58.0	1.00	98.0	46.0	0.98
10	86.0	96.0	8.60	79.2	90.0	7.92
20	74.1	100.0	14.82	62.9	96.0	12.58
50	48.0	100.0	24.00	39.24	98.0	19.62
100	31.7	100.0	31.70	24.84	98.0	24.84
200	20.8	100.0	41.60	15.79	98.0	31.58

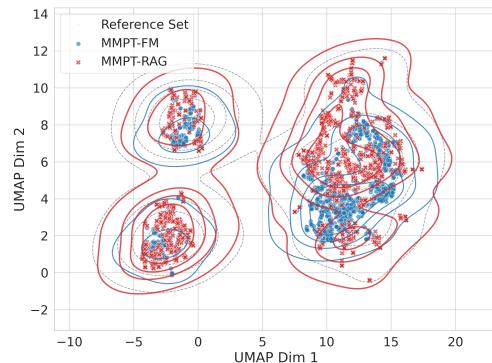


Figure 4: UMAP visualization of MMPT-FM and MMPT-RAG's chemical landscape on PMV17. The grey shaded areas represent the reference dataset's distribution. Compared to FM inference (blue), the MMPT-RAG framework (red) populates structural voids where the foundation model is sparse or absent.

regions in Fig. 4. For clarity, the RAG visualization highlights the additional coverage contributed beyond vanilla FM outputs, illustrating the complementary effect of retrieval augmentation.

As illustrated in Fig. 4, MMPT-RAG (red) expands into multiple structural regions that remain underexplored by the vanilla foundation model (blue). The vanilla FM tends to concentrate in high-probability general regions, leaving several reference clusters sparsely covered. In contrast, retrieval augmentation encourages the model to populate these underrepresented areas, effectively filling distributional gaps. This shift indicates that RAG enhances coverage of project-relevant chemical space by guiding generation toward regions of the reference dataset.

5.3.4 Hyperparameter Sensitivity Analysis. To better understand the behavior of MMPT-RAG, we perform a sensitivity analysis on three hyperparameters that control generation quality: the number of retrieved clusters expanded, i.e., generate using its MCS (default 10), the number of variables generated per cluster (default 50), and the mask-infilling length range used during sequence completion (default = original length before masking ± 7). We vary one hyperparameter at a time while keeping the others fixed: clusters 3, 5, 10, 20, variables per cluster 10, 25, 50, 100, and mask range [1, 3], [1, 5], [1, 7], [1, 9]. Figure 5 shows that moderate increases in all three parameters improve performance, while larger values provide diminishing returns. Expanding more clusters improves

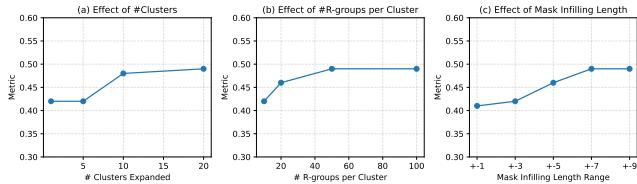


Figure 5: Hyperparameter Study. (a) the number of clusters to expand, (b) the number of variables to generate for each cluster, (c) the range of mask length to fill.

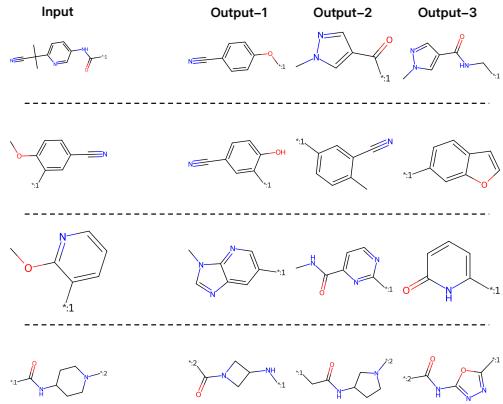


Figure 6: Examples of MMPT-FM generations. In each row, the left structure is the input variable, and the structures on the right are generated outputs.

results by exposing more diverse retrieved patterns and then stabilizes around 10 to 20 clusters. Increasing the number of variables per cluster expands search coverage but saturates near 50 samples. Widening the mask-infilling range consistently helps up to [1, 7], after which additional flexibility yields little gain. Based on these trends, we recommend 10 clusters, 50 variables per cluster, and a mask range of [1, 7] to balance performance and computation.

5.4 Qualitative Evaluation

To better understand the behavior of the proposed model beyond quantitative metrics, we present representative qualitative examples of generated variables. Fig. 6 illustrates variables directly generated from MMPT-FM. For each input variable (left column of each row), the model produces multiple diverse and chemically plausible variants. The generated variables preserve valid valence patterns and maintain realistic functional groups and ring systems. Notably, the model naturally supports multiple attachment points, demonstrating its ability to handle context-dependent transformations and generate structurally coherent edits across different substitution sites.

Fig. 7 shows MMPT-RAG’s generation results for the same input as the first example in Fig. 6. In this case, we first retrieve structurally similar historical variables from the database, then cluster the retrieved examples based on shared substructures. Two representative clusters are shown. For each cluster, we show the Maximum Common Substructure (MCS), which serves as a structural template capturing the dominant transformation pattern within

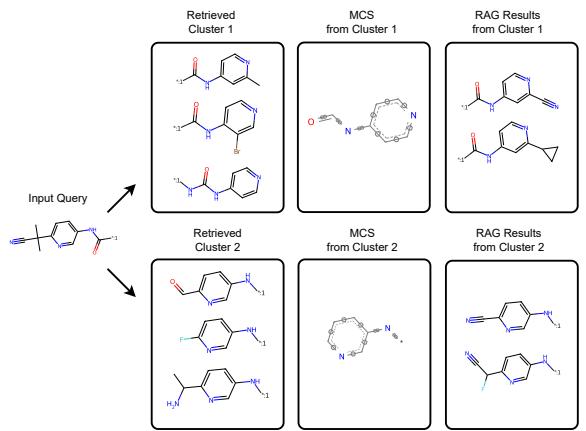


Figure 7: Examples of MMPT-RAG generations. Retrieved variables are clustered, an MCS is extracted per cluster, and generation is conditioned on each cluster’s template (MCS).

that cluster. Conditioned on these cluster-specific templates, the generated outputs reflect the characteristic patterns of their corresponding clusters while remaining chemically valid and diverse. Compared to the standalone foundation model, RAG effectively steers outputs toward specific transformation families while preserving chemical plausibility.

6 Conclusion

In this work, we presented a paradigm shift in generative molecular design by reframing analog generation as a transformation-to-transformation task grounded in Matched Molecular Pair Transformations (MMPTs). Unlike traditional molecule-level approaches that often lack localized control, our framework explicitly models the precise chemical edits that define medicinal chemistry intuition. By training a foundation model (**MMPT-FM**) on large-scale transformation data, we achieved scalable generation of variable substructures that balances chemical plausibility with structural novelty. To address the specific constraints of active drug discovery projects, we introduced **MMPT-RAG**, a retrieval-augmented framework that leverages external reference datasets to steer generation toward relevant, project-specific motifs. Our extensive evaluation on both general chemical corpora and time-split patent series demonstrates that this approach not only improves diversity and validity but effectively recovers prospective ligands in realistic discovery scenarios. Ultimately, this framework operationalizes MMPTs as a first-class generative abstraction, offering a powerful tool to synergize machine learning capability with human expertise.

7 Limitations and Ethical Considerations

Our approach relies on the availability and coverage of large historical transformation datasets, and its performance may vary in underrepresented chemical domains. Our framework is intended for research use, and does not introduce specific ethical concerns.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Stanislav Andreev, Tatjana Pantzar, Ahmed El-Gokha, Francesco Ansideri, Mark Kudolo, Débora Bublitz Anton, Giulia Sita, Jenny Romasco, Christian Geibel, Michael Lämmerhofer, et al. 2020. Discovery and Evaluation of Enantiopure 9 H-pyrimido [4, 5-b] indoles as Nanomolar GSK-3 β Inhibitors with Improved Metabolic Stability. *International Journal of Molecular Sciences* 21, 21 (2020), 7823.
- [3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [5] Binghong Chen, Tianzhe Wang, Chengtao Li, Hanjun Dai, and Le Song. 2021. Molecule optimization by explainable evolution. In *International conference on learning representation (ICLR)*.
- [6] Oh-Hyeon Choung, Riccardo Vianello, Marwin Segler, Niklaus Stiel, and José Jiménez-Luna. 2023. Extracting medicinal chemistry intuition via preference machine learning. *Nature Communications* 14, 1 (2023), 6651.
- [7] Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *International Conference on Machine Learning*. PMLR, 6140–6157.
- [8] Andrew Dalke, Jerome Hert, and Christian Kramer. 2018. mmpdb: An open-source matched molecular pair platform for large multiproperty data sets. *Journal of chemical information and modeling* 58, 5 (2018), 902–910.
- [9] Inc. Daylight Chemical Information Systems. 2019. SMARTS: a language for describing molecular patterns. (2019).
- [10] Alexander G Dossetor, Edward J Griffen, and Andrew G Leach. 2013. Matched molecular pair analysis in drug discovery. *Drug Discovery Today* 18, 15–16 (2013), 724–731.
- [11] Janos Fischer and C Robin Ganellin. 2010. Analogue-based drug discovery. *Chemistry International—Newsmagazine for IUPAC* 32, 4 (2010), 12–15.
- [12] 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. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* 40, D1 (2012), D1100–D1107.
- [13] Ed Griffen, Andrew G Leach, Graeme R Robb, and Daniel J Warner. 2011. Matched molecular pairs as a medicinal chemistry tool: miniperspective. *Journal of medicinal chemistry* 54, 22 (2011), 7739–7750.
- [14] Jiazen He, Eva Nittlinger, Christian Tyrchan, Werngard Czechitzky, Atanas Patronov, Esben Jannik Bjerrum, and Ola Engkvist. 2022. Transformer-based molecular optimization beyond matched molecular pairs. *Journal of cheminformatics* 14, 1 (2022), 18.
- [15] Jiazen He, Hufang You, Emil Sandström, Eva Nittlinger, Esben Jannik Bjerrum, Christian Tyrchan, Werngard Czechitzky, and Ola Engkvist. 2021. Molecular optimization by capturing chemist's intuition using deep neural networks. *Journal of cheminformatics* 13 (2021), 1–17.
- [16] ZhiLin Huang, Ling Yang, Xiangxin Zhou, Chujun Qin, Yijie Yu, Xiawu Zheng, Zikun Zhou, Wentao Zhang, Yu Wang, and Wenming Yang. 2024. Interaction-based retrieval-augmented diffusion models for protein-specific 3d molecule generation. In *Forty-first International Conference on Machine Learning*.
- [17] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2020. Multi-objective molecule generation using interpretable substructures. In *International conference on machine learning*. PMLR, 4849–4859.
- [18] Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi Jaakkola. 2018. Learning multimodal graph-to-graph translation for molecular optimization. *arXiv preprint arXiv:1812.01070* (2018).
- [19] Lou Jost. 2006. Entropy and diversity. *Oikos* 113, 2 (2006), 363–375.
- [20] Nikita A Krotkov, Dmitrii A Sbytov, Anna A Chakhoyan, Polina I Kornienko, Anna A Starikova, Maxim G Stepanov, Anastasiia O Piven, Timur A Aliev, Tetiana Orlova, Mushegh S Rafayelyan, et al. 2025. Nanostructured material design via a retrieval-augmented generation (rag) approach: Bridging laboratory practice and scientific literature. *Journal of Chemical Information and Modeling* (2025).
- [21] Greg Landrum. 2013. RDKit: Open-source cheminformatics. (2013). <https://www.rdkit.org>.
- [22] Seul Lee, Karsten Kreis, Srimukh Veccham, Meng Liu, Danny Reidenbach, Saeed Paliwal, Arash Vahdat, and Weili Nie. 2024. Molecule generation with fragment retrieval augmentation. *Advances in Neural Information Processing Systems* 37 (2024), 132463–132490.
- [23] Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsuk Kim, and Woo Yoon Kim. 2020. Scaffold-based molecular design with a graph generative model. *Chemical science* 11, 4 (2020), 1153–1164.
- [24] Hannes H Loeffler, Jiazen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H Mervin, and Ola Engkvist. 2024. Reinvent 4: modern AI-driven generative molecule design. *Journal of Cheminformatics* 16, 1 (2024), 20.
- [25] Jieyu Lu and Yingkai Zhang. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling* 62, 6 (2022), 1376–1387.
- [26] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836.
- [27] Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.
- [28] NA Meanwell. 2017. A synopsis of the properties and applications of heteroaromatic rings in medicinal chemistry. In *Advances in Heterocyclic Chemistry*. Vol. 123. Elsevier, 245–361.
- [29] Harry L. Morgan. 1965. The generation of a unique machine description for chemical structures—A technique developed at chemical abstracts service. *Journal of Chemical Documentation* 5, 2 (1965), 107–113.
- [30] Emmanuel Noutahi, Hadrien Mary, Kyle M. Kovary, Shawn Whitfield, Julian St-Laurent, Honoré Hounwanou, and Michael Craig. 2025. datamat-io/medchem: Molecular filtering for drug discovery (v2.0.5-alpha). <https://doi.org/10.5281/zenodo.1458893>
- [31] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. 2017. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* 9, 1 (2017), 48.
- [32] Riza Ozcelik, Helena Brinkmann, Emanuele Criscuolo, and Francesca Grisoni. 2025. Generative deep learning for de novo drug design - a chemical space odyssey. *Journal of Chemical Information and Modeling* 65, 14 (2025), 7352–7372.
- [33] George Papadatos, Muhammad Alkarouri, Valerie J Gillet, Peter Willett, Visakan Kadirkamanathan, Christopher N Luscombe, Gianpaolo Bravi, Nicola J Richmond, Stephen D Pickett, Jameed Hussain, et al. 2010. Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of HERG inhibition, solubility, and lipophilicity. *Journal of chemical information and modeling* 50, 10 (2010), 1872–1886.
- [34] John W Raymond, Eleanor J Gardiner, and Peter Willett. 2002. Rascal: Calculation of graph similarity using maximum common edge subgraphs. *Comput. J.* 45, 6 (2002), 631–644.
- [35] Alfréd Rényi. 1961. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, Vol. 4. University of California Press, 547–562.
- [36] Valerie W Shurtliff, Mark E Layton, Craig A Parish, James J Perkins, John D Schreier, Yunyi Wang, Gregory C Adam, Nadine Alvarez, Soheila Bahmanjah, Carolyn M Bahnik-Teets, et al. 2024. Invention of MK-7845, a SARS-CoV-2 3CL protease inhibitor employing a novel difluorinated glutamine mimic. *Journal of Medicinal Chemistry* 67, 5 (2024), 3935–3958.
- [37] Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, et al. 2017. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *Journal of cheminformatics* 9, 1 (2017), 17.
- [38] Alessandro Tibo, Jiazen He, Jon Paul Janet, Eva Nittlinger, and Ola Engkvist. 2024. Exhaustive local chemical space exploration using a transformer model. *Nature Communications* 15, 1 (2024), 7315.
- [39] Emma P Tysinger, Brajesh K Rai, and Anton V Sinit斯基. 2023. Can We Quickly Learn to “Translate” Bioactive Molecules with Transformer Models? *Journal of Chemical Information and Modeling* 63, 6 (2023), 1734–1744.
- [40] Binh Vu, Romyr Dominique, and Hongju Li. 2017. Methods and Compounds for Restoring Mutant p53 Function.
- [41] Binh Vu, Romyr Dominique, Hongju Li, Bruce Fahr, and Andrew Good. 2021. Methods and Compounds for Restoring Mutant p53 Function.
- [42] Pat Walters. 2018. rd_filters. https://github.com/PatWalters/rd_filters/blob/master/rd_filters/data/alert_collection.csv
- [43] Haorui Wang, Marta Skreta, Cher-Tian Ser, Wenhao Gao, Lingkai Kong, Felix Strieth-Kalthoff, Chenru Duan, Yuchen Zhuang, Yue Yu, Yanqiao Zhu, et al. 2024. Efficient evolutionary search over chemical space with large language models. *arXiv preprint arXiv:2406.16976* (2024).
- [44] Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, and Anima Anandkumar. 2022. Retrieval-based controllable molecule generation. *arXiv preprint arXiv:2208.11126* (2022).
- [45] Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* 1, 2 (1989), 270–280.
- [46] Zhenxing Wu, Odin Zhang, Xiaorui Wang, Li Fu, Huifeng Zhao, Jike Wang, Hongyan Du, Dejun Jiang, Yafeng Deng, Dongsheng Cao, et al. 2024. Leveraging language model for advanced multiproperty molecular optimization via prompt engineering. *Nature Machine Intelligence* (2024), 1–11.
- [47] Dong Xu, Zhangfan Yang, Ka-chun Wong, Zexuan Zhu, Jiangqiang Li, and Junkai Ji. 2025. Reimagining Target-Aware Molecular Generation through Retrieval-Enhanced Aligned Diffusion. *arXiv preprint arXiv:2506.14488* (2025).

- [48] Ziyi Yang, Shaohua Shi, Li Fu, Aiping Lu, Tingjun Hou, and Dongsheng Cao. 2023. Matched molecular pair analysis in drug discovery: methods and recent applications. *Journal of Medicinal Chemistry* 66, 7 (2023), 4361–4377.
- [49] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. 2018. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems* 31 (2018).
- [50] Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391* (2024).
- [51] Peidong Zhang, Xingang Peng, Rong Han, Ting Chen, and Jianzhu Ma. 2025. Rag2Mol: structure-based drug design based on retrieval augmented generation. *Briefings in Bioinformatics* 26, 3 (2025).

A Implementation Details

Foundation model training and inference. To obtain the foundation model, we leverage the ChemT5 model [7], an encoder-decoder transformer pretrained on chemical datasets, as our base model. ChemT5 contains approximately 220 million parameters and has been fine-tuned for tasks in cheminformatics. It consists of 12 layers, 12 attention heads, 220M parameters, and processes input sequences up to 512 tokens [7]. We employed standard supervised training to fine-tune all parameters of the base model. Teacher forcing [45] is incorporated to improve training stability. The training was conducted with a batch size of 64 on each device, and a learning rate of 5e-4. We use an early stop strategy with a tolerance of 2 epochs based on the evaluation loss. Utilizing four NVIDIA A6000 GPUs (48 GB each), the training process required approximately 70 hours to complete. During inference with the foundation model, for each input, we use beam search to produce 1000 outputs with a maximum length of 50.

Retriever. We pre-build a nearest-neighbor index with HNSW [26] and query it with cosine distance over the Morgan Fingerprint [29] embedding space. At inference time, we first retrieve at most top 500 nearest input variables, then expand each input into its associated label set. To ensure compatibility with the query, we filter candidates by the number of wild atoms (i.e., attachment points) to match the query variable. We then compute Morgan fingerprints and re-rank retrieved label candidates by Tanimoto similarity between the query and candidate labels, retaining the top set for downstream RAG steps.

Clustering, Template Construction. We cluster the retrieved labels in structure space using a shared-substructure clustering method, where we first compute pairwise similarities between retrieved outputs using the size of the RDKit maximum common substructure (MCS) normalized by the smaller molecule. We then perform agglomerative hierarchical clustering (average linkage) on the corresponding distance matrix, and cut the dendrogram at a similarity threshold of 0.70 to obtain clusters. To avoid overly coarse motifs, any cluster exceeding 10 molecules is recursively split using the same linkage procedure until all clusters satisfy the size constraint. For each retained cluster, we compute a Maximum Common Substructure (MCS), whose resulting SMARTS string serves as the cluster-invariant template.

We further convert this template into a partially specified output constraint by masking atoms outside the invariant scaffold. Concretely, we apply substructure masking to produce a template string where masked spans are denoted by a special masking character. We then convert each masked span into a single <BLANK> placeholder and perform span infilling using the generative model, as introduced in the prompted generation section.

Prompted Generation via Mask Infilling. The infilling search is controlled by 11 maximum new tokens per blank, 200 maximum total candidate continuations, and 200 top candidates scored. The final RAG outputs are the union of valid, RDKit-parsable infilled candidates across templates. We exclude any duplicate sequences.

Implementation of Compared Methods. For Database Retrieval, we first retrieve at most 50 input variables from the reference dataset that are most similar to the test query based on fingerprint similarity. We then collect all corresponding output

variables paired with these retrieved inputs. If the resulting candidate set exceeds 1000 outputs, we retain the 1000 variables that are most similar to the test query to ensure a fair comparison in terms of candidate size. REINVENT4 (LibINVENT) follows a different formulation, generating variables conditioned on a scaffold rather than directly modeling MMPT-style variable-to-variable transformations. To align the setting with our task, we use the constant fragment identified by MMPDB as the scaffold input and generate up to 1000 candidate variables for each query.

For consistency, all methods, including MMPT-FM and MMPT-RAG, generate 1000 candidates per input.

B Proof of Theorem 4.1

PROOF. According to the workflow defined in Eq. 5, the MMPT-RAG framework constructs the global generation as a mixture of cluster-conditioned distributions with weights $\tilde{\pi}_k$. Summing over the variables y :

$$\begin{aligned} p_{\text{RAG}}(y | x) &= \sum_{k=1}^K \tilde{\pi}_k p(y | x, T_k) \\ &= \sum_{k=1}^K \tilde{\pi}_k [(1 - \alpha_k)p_\theta(y | x) + \alpha_k p(y | T_k)] \\ &= p_\theta(y | x) \sum_{k=1}^K \tilde{\pi}_k (1 - \alpha_k) + \sum_{k=1}^K \tilde{\pi}_k \alpha_k p(y | T_k). \end{aligned}$$

Using $\sum \tilde{\pi}_k = 1$ and the definition $\bar{\alpha} = \sum \tilde{\pi}_k \alpha_k$, the first term simplifies to $(1 - \bar{\alpha})p_\theta(y | x)$. The second term, by multiplying and dividing by $\bar{\alpha}$, recovers the effective reference $p_{\text{ref}}^*(y | x)$. Thus:

$$p_{\text{RAG}}(y | x) = (1 - \bar{\alpha})p_\theta(y | x) + \bar{\alpha}p_{\text{ref}}^*(y | x).$$

□