Molecular Optimization Model with Patentability Constraint

Sally Turutov, Kira Radinsky

Technion - Israel Institute of Technology
turutovsally@campus.technion.ac.il, kirar@cs.technion.ac.il

Abstract

In drug development, molecular optimization is a crucial challenge that involves generating novel molecules given a lead molecule as input. The task requires maintaining molecular similarity to the original molecule while simultaneously optimizing multiple chemical attributes. To aid in this process, numerous generative models have been proposed. However, in practical applications, it is crucial for these models not only to generate novel molecules with the above constraints but also to generate molecules that significantly differ from any existing patented compounds. In this work, we present a multi-optimization molecular framework to address this challenge. Our framework trains a model to prioritize both enhanced properties and substantial dissimilarity from patented compounds. By jointly learning continuous representations of optimized and patentable molecules, we ensure that the generated molecules are significantly distant from any patented compounds while improving chemical properties. Through empirical evaluation, we demonstrate the superior performance of our approach compared to state-of-the-art molecular optimization methods both in chemical property optimization and patentability.

Introduction

The process of developing a successful drug is a lengthy and expensive endeavor, typically taking 10 to 15 years and costing around 1 billion dollars. The early stages of drug development involve the discovery, design, and optimization of a lead compound — a chemical entity with desirable drug-like properties. Lead optimization aims to improve the lead compound's properties while maintaining a high degree of similarity. To aid in this process, generative models have been proposed to generate novel molecules with enhanced chemical properties. However, in practical applications, it is crucial for these models not only to prioritize novelty but also to generate molecules that significantly differ from any existing patented compounds. This dual requirement ensures not only chemical improvement but also the potential for patentability, addressing the challenges of real-world drug development.

In this work, our primary focus lies in the domain of molecule optimization while considering the constraint of patentability. This task poses significant challenges due to the unique characteristics of patented molecules. Unlike molecules with similar chemical enhancements that exhibit commonalities and tend to cluster together in the embedding space, patented molecules do not necessarily reside within specific regions. Consequently, traditional gradient-based algorithms encounter difficulties in effectively optimizing molecules under the patentability constraint.

We propose the Molecular Optimization Model with Patentability Constraint (MOMP) to tackle patent-infringement challenges in molecular optimization. Our model utilizes a generative sequence-to-sequence architecture based on the SMILES (Jastrzebski, Leśniak, and Czarnecki 2016) representation to generate molecules with enhanced properties. We propose a multi-cycle encoder-decoder framework. An encoder converts discrete molecules to continuous representations, which are then translated by a translator to embedded destination domain with improved properties. The cycle is completed with a decoder, which translates the continuous representations back into discrete molecules, facilitating property enhancement. An additional cycle is proposed, with a translator trained to convert representations of molecules with enhanced properties into molecules possessing both high property values and low resemblance to patents. We utilize these cycles during training, while during testing, a discrete source molecule is encoded, translated for enhanced properties, and then translated again into an optimized molecule with high property and high patentability. To ensure patentability, we employ a molecular attention mechanism using fingerprints from the domain of patentable molecules. By learning continuous representations of optimized and patentable molecules, MOMP effectively balances the enhancement of molecular properties and significant dissimilarity from existing patented compounds, enabling the generation of promising molecules while adhering to patentability constraints.

We empirically evaluate our proposed model on numerous molecule optimization tasks, demonstrating its ability to maintain similarity and optimize properties while considering patent constraints. Our results show that our model successfully reduces the similarity of optimized molecules to existing patents while still generating highly optimized molecules, thus outperforming the state-of-the-art (SOTA)

models. Additionally, comprehensive ablation experiments provide detailed insights into the effectiveness of our approach and its individual components.

The contributions of this work are threefold: (1) We propose the MOMP algorithm, which effectively addresses patent-infringement challenges in molecular optimization. By jointly optimizing molecular properties and patentability, as measured by the resemblance of molecules to patented compounds, MOMP ensures the generation of molecules that are substantially dissimilar from existing patents while possessing enhanced properties. (2) We integrate the concept of molecular attention, utilizing fingerprints of patentable molecules in the optimization process. This attention mechanism enables our model to reinforce the patentability constraint without compromising the chemical characteristics of the molecules. (3) We conduct comprehensive empirical evaluations on two widely used molecule optimization tasks, showcasing the superiority of our proposed MOMP model over SOTA baselines. To facilitate further research and exploration of the problem, we provide the community with access to our code and data through the following link: https://github.com/SallyTurutov/MOMP.

Related Work

Several strategies were suggested to address the task of molecule optimization, each employing unique representations of molecules. Graph-to-graph optimization methods aim to optimize the score of molecular graphs by converting one graph representation of a molecule into another. JTVAE (Jin, Barzilay, and Jaakkola 2018) interprets an input molecule as composed of sub-graphs selected from a valid component vocabulary and optimizes the property score predictor based on its latent space. Mol-CG (Maziarka et al. 2020) extends JTVAE using a cycle-GAN architecture. CORE (Fu, Xiao, and Sun 2020), an enhanced version of JTVAE, introduces the copy and refine technique to enhance molecular optimization. An alternative line of work, leverages SMILES-based representations for optimization. SMILES representations have been employed to optimize molecules, and the UGMMT (Barshatski and Radinsky 2021) reached SOTA results outperforming all SMILES-based methods and numerous graph-to-graph methods.

Recent advancements focused on going beyond the optimization of a single property. IPCA (Barshatski, Nordon, and Radinsky 2021) generalizes UGMMT for multiple property optimization. MIMOSA (Fu et al. 2021), a graph-to-graph approach, uses GNNs to predict molecular topology and substructure types for generating new molecules.

Our work differs from the aforementioned lines of work by focusing not only on molecular optimization but also on optimization without patent infringement. One of the only works to address optimization while reducing patent infringement was presented by (Turutov and Radinsky 2023). The authors propose a patents-loss that can be incorporated into existing models. The loss function leverages the original model's loss to ensure differentiation from a specific focal patent. In contrast to this method, our MOMP model is an end-to-end generative approach tailored for patentable molecule optimization, yielding superior results.

MOMP Algorithm

The concept of "patent-likeliness" (PL) was introduced by (Turutov and Radinsky 2023) in order to quantify how similar a generated molecule m' is to all existing patents, represented by P_{all}. The PL value serves as a measure to ensure that the generated molecule is significantly different from existing patents, and is defined using the following equation:

$$PL_{P_{all}}(m') = \max_{p \in P_{all}} Sim(m', p) \quad (1)$$

where $Sim(m', p)$ is the Tanimoto similarity of molecule m' and each patent p in the collection of existing patents P_{all}. A molecule m' is patentable if the $PL_{P_{all}}$ value associated with m' is below a predefined threshold, indicating that it does not infringe upon any existing patents.

In our framework, we consider different domains of molecules. Let the domain of molecules be denoted by a capital Latin letter, e.g, X, a molecule taken from this domain by a small Latin letter, x, and their distribution by $p(X)$. Similarly, we denote by $\langle x \rangle$ the embedding vector of a molecule x and a molecule's property by $prop$. For example, if $prop$ is drug-likeness (e.g., QED), then $prop(x)$ is QED value of x. Given an input molecule m, we aim to generate a patentable molecule m' which resembles m and satisfies an enhanced $prop$.

Figure 3 illustrates the MOMP architecture. The algorithm operates through multiple optimization paths. At first, molecules with degraded properties (low $prop$ value) in domain A are transformed into molecules with enhanced properties (high $prop$ value) and high resemblance to patents (high PL value) in domain B. Subsequently, these molecules from domain B are further optimized to achieve enhanced properties (high $prop$ value) and low resemblance to patents (low PL value) in domain C. The paths are joint by an additional cycle, in order to keep the optimized molecule similar to the input molecule going through the path. During optimization, the first path transforms molecules from domain A to domain B, and then the second path transforms molecules from domain B to domain C, producing molecules with both enhanced properties and low resemblance to patents. Moreover, the optimization path from domain A to domain C ensures that the optimized molecule in domain C maintains similarity to the input molecule in domain A.

Molecule-Embedding Translation Network

The Molecule-Embedding Translation Network (METN) plays a critical role in the pre-training phase of the MOMP framework. Initially introduced by (Barshatski and

Algorithm 1 METN Training Algorithm

Input: training set of molecules X

for $epoch = 1, 2, \ldots, E_{METN}$ **do**
  Sample mini-batch $x \in X$
  $x' = De_X(En_X(x))$
  $L = CE(x', x)$
  Minimize L using Adam optimizer
end for

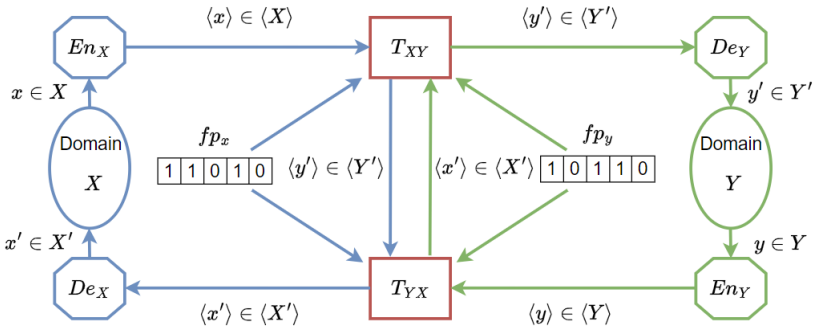

Figure 1: The EETN architecture – Each training path originates from an input molecule in a specific domain and concludes with its optimized counterpart in the same domain. The color of the arrows corresponds to the domain of the input molecules.

Algorithm 2 EETN Training Algorithm

Input: training sets of molecules X, Y and fps fp_x, fp_y

for $epoch = 1, 2, \ldots, E_{EETN}$ **do**
Sample mini-batches $x \in X, y \in Y$
$\langle x \rangle = En_X(x)$
$\langle y' \rangle = T_{XY}(\langle x \rangle, fp_x)$
$\langle x' \rangle = T_{YX}(\langle y' \rangle, fp_x)$
$x' = De_X(\langle x' \rangle)$

$\langle y \rangle = En_Y(y)$
$\langle x' \rangle = T_{YX}(\langle y \rangle, fp_y)$
$\langle y' \rangle = T_{XY}(\langle x' \rangle, fp_y)$
$y' = De_Y(\langle y' \rangle)$

$L = CE(x', x) + \lambda_{XY} \cdot CE(y', y)$
Minimize L using Adam optimizer
end for

Radinsky 2021), it translates molecules between their discrete SMILES representations and continuous embeddings, preparing the latent embedding spaces for each domain X (where $X \in \{A, B, C\}$) within MOMP.

METN comprises an encoder (En_X) and a decoder (De_X). The encoder converts a molecule's SMILES representation ($x \in X$) into a continuous embedding ($\langle x \rangle \in \langle X \rangle$). The decoder reconstructs the continuous embedding back into its corresponding SMILES representation ($x' \in X'$), as described in Algorithm 1. The translation process is guided by the cross-entropy loss (CE), encouraging the similarity between the original molecule x and its reconstructed version x'. This optimization ensures that the embedding accurately captures essential structural information.

Including METN in the pre-training phase aligns the latent embedding spaces for each domain, allowing for effective optimization in subsequent MOMP processes. By capturing meaningful features and dependencies of molecular structures, METN facilitates the generation of high-quality molecules within their respective domains.

Embedding-Embedding Translation Network

The Embedding-Embedding Translation Network (EETN) translates continuous molecule embeddings between distinct domains denoted as X and Y. Working in conjunction with METN, the EETN facilitates the translation of molecule embeddings between these domains.

EETN comprises two inverted translators: T_{XY} and T_{YX}. The translator T_{XY} converts an embedding of a molecule from domain X to an embedding of a molecule in domain Y, while T_{YX} performs the inverse translation from domain Y to domain X.

During training, a molecule $x \in X$ undergoes the following transformations: it is encoded by En_X into its continuous embedding. This embedding then goes through the EETN pipeline: T_{XY} translates it to the embedding of domain Y, followed by T_{YX} which returns it to the embedding of domain X. Finally, the embedding is passed back through METN, specifically De_X, to reconstruct a discrete molecule $x' \in X$. Similarly, a molecule $y \in Y$ follows the path of $En_Y \rightarrow T_{YX} \rightarrow T_{XY} \rightarrow De_Y$. The process is presented in Figure 1 and Algorithm 2. The MOMP model simultaneously trains T_{XY} and T_{YX} using the double-cycle training scheme (He et al. 2016). This coupling between the two translation sequences encourages the distribution of original and reconstructed molecules to be close, ensuring accurate translation between the domains X and Y.

The MOMP model utilizes two instances of EETN: one between domains A and B and another between domains B and C. This dual EETN setup enhances the model's ability to handle translation tasks between multiple domains, while also considering the patent infringement constraint.

Extended Embedding-Embedding Translation Network

We propose an extension to the existing Embedding-Embedding Translation Network (EETN) by incorporating four additional translators: T_{XY}, T_{YX}, T_{YZ}, and T_{ZY}. This Extended-EETN enhances the translation capabilities between multiple domains denoted as X, Y, and Z.

The molecule $x \in X$ undergoes a sequence of transformations within the Extended-EETN pipeline. First, it is encoded by En_X into its continuous embedding representa-

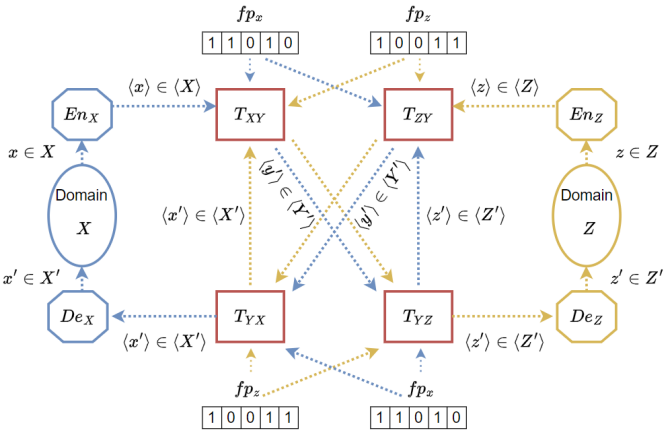

Figure 2: The Extended-EETN architecture – Each training path originates from an input molecule in a specific domain and concludes with its optimized counterpart in the same domain. The color of the arrows corresponds to the domain of the molecules.

Algorithm 3 Extended-EETN Training Algorithm

Input: training sets of molecules X, Z and fps fp_x, fp_z

for $epoch = 1, 2, \ldots, E_{Extended-EETN}$ **do**

Sample mini-batches $x \in X$, $z \in Z$

$\langle x \rangle = En_X(x)$
$\langle y' \rangle = T_{XY}(\langle x \rangle, fp_x)$
$\langle z' \rangle = T_{YZ}(\langle y' \rangle, fp_x)$
$\langle y' \rangle = T_{ZY}(\langle z' \rangle, fp_x)$
$\langle x' \rangle = T_{YX}(\langle y' \rangle, fp_x)$
$x' = De_X(\langle x' \rangle)$

$\langle z \rangle = En_Z(z)$
$\langle y' \rangle = T_{ZY}(\langle z \rangle, fp_z)$
$\langle x' \rangle = T_{YX}(\langle y' \rangle, fp_z)$
$\langle y' \rangle = T_{XY}(\langle x' \rangle, fp_z)$
$\langle z' \rangle = T_{YZ}(\langle y' \rangle, fp_z)$
$z' = De_Z(\langle z' \rangle)$

$L = CE(x', x) + \lambda_{XZ} \cdot CE(z', z)$

Minimize L using Adam optimizer

end for

tion. The translation from domain X to domain Y is performed by T_{XY}, followed by the translation from domain Y to domain Z using T_{YZ}. Subsequently, the molecule is translated back from domain Z to domain Y using T_{ZY} and finally returned to its original domain X through T_{YX}. The resulting embedding is then passed through METN, specifically De_X, to reconstruct a discrete molecule $x' \in X$. Similarly, a molecule $z \in Z$ follows an analogous path within the Extended-EETN pipeline: $En_Z \rightarrow T_{ZY} \rightarrow T_{YX} \rightarrow T_{XY} \rightarrow T_{YZ} \rightarrow De_Z$. The Extended-EETN is illustrated in Figure 2 and described in Algorithm 3.

During training, the Extended-EETN model simultaneously trains the four translation components, T_{XY}, T_{YX}, T_{YZ}, and T_{ZY}, using a dual learning paradigm. The translation tasks from X to Z and from Z to X (primal tasks) and their corresponding inverse translations (dual tasks) create an informative feedback loop, promoting effective translation across multiple domains. This training scheme encourages proximity between the original molecule and its reconstructed version by imposing cross-entropy constraints $CE(x', x)$ and $CE(z', z)$.

In MOMP, the Extended-EETN enables the translation of molecule embeddings between domains A, B, and C. This expands the translation capabilities of MOMP, facilitating accurate transformations of molecules across these domains.

Molecular Attention

To preserve chemical characteristics and reinforce the patentability constraint in the MOMP model, we incorporate molecular fingerprints. Unlike previous approaches, that utilized fingerprints to help the model keep similarity to the lead molecule, we suggest to leverage the fingerprints of non-patented molecules. During inference, the input molecule's fingerprints are combined with the input embedding. During training, to prioritize patentability, we select a molecule ($c \in C$) most similar to x and use its fingerprint (fp_c) along the optimization path. To focus on crucial fingerprint information, we employ an attention mechanism. The input molecule's fingerprint undergoes transformation through a fully-connected layer, followed by a softmax layer, generating a weight vector that highlights an essential part of the embedding for optimization. This molecular attention mechanism ensures efficient optimization while maintaining chemical characteristics.

MOMP End-to-End Architecture

We propose the MOMP architecture, aiming to optimize molecules across domains A, B, and C. The model optimizes an input molecule $m \in A$ to generate an improved molecule $m' \in C$ (Figure 3).

The MOMP architecture consists of three key components. First, the Molecule-Embedding Translation Networks (METNs) pre-train encoders and decoders for each domain individually, enhancing the latent embedding space

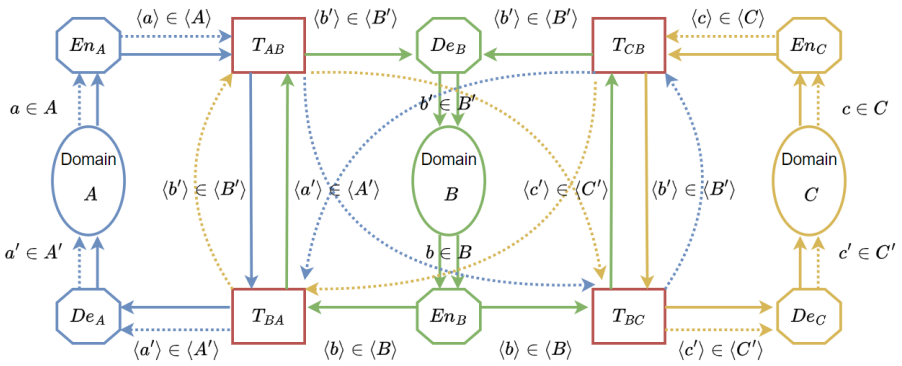

Figure 3: The MOMP architecture – Each training path originates from an input molecule in a specific domain (A, B, or C) and concludes with its optimized counterpart in the same domain. The color of the arrows corresponds to the domain of the input molecules. The cross-entropy (CE) loss function is applied during training to guide the optimization process. Solid arrows represent the training paths of the EETN, while dotted arrows represent the training paths of the Extended-EETN.

Algorithm 4 End-to-End Training Algorithm

Input: A, B C molecule training sets

Pre-train A, B, C with Algorith 1
for $epoch = 1, 2, \ldots, E_{maxTrain}$ **do**
  Sample mini-batches $a \in A$, $b \in B$ and $c \in C$
  Find closest $c_a \in C$ to a and calculate $fp_a = fp(c_a)$
  Find closest $c_b \in C$ to b and calculate $fp_b = fp(c_b)$
  Calculate $fp_c = fp(c)$
  Calculate using Algorithm 2 (EETN $A \rightleftarrows B$ – solid blue and green paths):
    $L_{AB} = CE(a', a) + \lambda_{AB} \cdot CE(b', b)$
  Calculate using Algorithm 2 (EETN $B \rightleftarrows C$ – solid yellow and green paths):
    $L_{BC} = CE(b', b) + \lambda_{BC} \cdot CE(c', c)$
  Calculate using Algorithm 3 (Extended-EETN $A \rightleftarrows C$ – dotted blue and yellow paths):
    $L_{AC} = CE(a', a) + \lambda_{AC} \cdot CE(c', c)$
  $L = L_{AB} + L_{BC} + L_{AC}$
  Minimize L using Adam optimizer
end for

of molecules. Then, for domain translation, we use the Embedding-Embedding Translation Network (EETN) to train translators between A-B and B-C. Finally, an Extended-EETN includes domain B as an intermediate step between A and C. By employing this architecture, we achieve effective translation of molecule embeddings across multiple domains, facilitating the generation of optimized molecules with improved properties while adhering to patentability constraints.

Training and Inference During MOMP training, we pre-train each domain using the METN algorithm (Algorithm 1) to prepare their latent embedding space. Then, we train the EETNs between domain pairs A and B, and B and C, alongside the Extended-EETN enabling translation between domain A and C via domain B (Algorithm 2 and Algorithm 3 respectively). The overall training process is outlined in Algorithm 4 and presented in Figure 3, incorporating pre-training and translation training.

During inference, the trained MOMP model transforms an input molecule m from domain A to an optimized molecule m' in domain C through $En_A \rightarrow T_{AB} \rightarrow T_{BC} \rightarrow De_C$. Additionally, we use the C-fps attention during training, while during inference, we utilize the input-fps to emphasize similarity to the input molecule, ensuring optimized molecules that are sufficiently similar to the input while staying distinct from existing patents.

Experimental Setup

We provide implementation details, as well as the datasets we used and the checkpoints of our trained model.

Implementation Details

Our code and data are available publicly. We employ the Adam optimizer with a learning rate of $3 \cdot 10^{-4}$, a mini-batch size of 32, and set maximum epochs $E_{maxTrain}$ to 12 for QED and 18 for DRD2. The regularization parameters are $\lambda_{AB} = \lambda_{BC} = \lambda_{AC} = 2$. Further implementation details can be found at: https://github.com/SallyTurutov/MOMP.

Metrics

To evaluate our optimization task, we generate $K = 20$ output molecules per input using K random seeds. Evaluation relies on the most similar valid output molecule m' for each input m. Input molecules lacking valid outputs are excluded for accurate metric calculations. Metrics are reported solely for valid molecules:

- **Property:** The average desired property score (QED or DRD2) across all optimized m' molecules, where each molecule's property score falls within the range of $[0, 1]$.
- **Patent Likeness (PL)** The average PL score ($PL_{P_{all}}(m') \in [0, 1]$) measures the similarity of the optimized molecule to existing patents. The PL value of a molecule m' is calculated using Equation 1.

		MOMP	UGMMT		JTVAE		REINVENT		CORE	G2G	Mol-CG	MIMOSA	IPCA
		Original	Original	PatentsLoss	Original	PatentsLoss	Original	PatentsLoss	Original	Original	Original	Original	Original
QED	Property ↑	0.842	0.855	0.817	0.816	0.808	0.819	0.813	0.883	0.890	0.783	0.783	0.764
	PL ↓	0.436	0.510	0.443	0.487	0.453	0.523	0.492	0.525	0.515	0.485	0.460	0.545
	Similarity	0.282	0.365	0.292	0.304	0.248	1.000	1.000	0.362	0.339	0.302	0.821	0.241
	Validity	0.914	0.971	0.879	1.000	1.000	0.992	0.991	1.000	1.000	0.998	0.909	0.922
	Novelty	1.000	0.997	1.000	0.977	0.999	0.999	1.000	0.980	0.979	0.980	0.355	0.993
	Success	**0.272**	0.118	0.232	0.096	0.156	0.040	0.082	0.090	0.086	0.088	0.125	0.053
DRD2	Property ↑	0.746	0.824	0.707	0.340	0.106	0.072	0.073	0.770	0.792	0.382	0.089	0.174
	PL ↓	0.584	0.696	0.630	0.577	0.544	0.470	0.456	0.699	0.697	0.643	0.504	0.545
	Similarity	0.269	0.283	0.248	0.239	0.133	1.000	1.000	0.345	0.333	0.190	0.077	0.241
	Validity	0.981	1.000	0.970	1.000	1.000	0.933	0.934	1.000	0.999	1.000	0.818	0.922
	Novelty	0.892	0.787	0.941	0.991	0.999	0.000	0.000	1.000	1.000	0.992	0.418	0.993
	Success	**0.322**	0.147	0.204	0.103	0.000	0.000	0.000	0.127	0.135	0.051	0.032	0.000

Table 1: Performance Comparison of MOMP and Baseline Models for Various Metrics.

- **Similarity:** The average Tanimoto similarity (Bajusz, Rácz, and Héberger 2015) over Morgan fingerprints (Rogers and Hahn 2010) for all (m', m) pairs, indicating the optimized molecule's similarity to the input molecule. The similarity value is denoted as $Sim(m', m) \in [0, 1]$.
- **Validity:** The proportion of valid optimized molecules, determined using the method proposed by Landrum (2016) for molecule validation.
- **Novelty:** The proportion of optimized molecules m' considered novel, i.e., not present in the training set.
- **Optimization Success (Success):** The proportion of successfully optimized molecules. A molecule m' is successful if it is novel and simultaneously meets criteria for high similarity to the input molecule ($Sim(m', m) > \lambda_s$), high desired property score ($prop > \lambda_{prop}$), and low patent likeness score ($PL(m') < \lambda_{PL}$). Threshold values for these criteria are set based on the property being QED or DRD2: $(\lambda_s, \lambda_{prop}, \lambda_{PL}) = (0.15, 0.7, 0.4)$ for QED and $(0.15, 0.7, 0.6)$ for DRD2

Datasets

We utilized datasets from (Jin et al. 2019). Initially, we partitioned training pairs into distinct domain-specific sets. Domain A encompasses molecules with low property scores (QED$(a) < 0.78$ or DRD2$(a) < 0.02$), while domains B and C comprise molecules with high scores (QED$(c) > 0.85$ or DRD2$(c) > 0.75$). After identifying eligible molecules for B and C, we ranked them by Patent Likeness (PL) values. Half with the lowest PL scores formed domain C, and the rest constituted domain B. We randomly sampled molecules from A to match the size of B and C.

The SureChEMBL dataset (Papadatos et al. 2016) focuses on patent compounds, providing Maximum Common Substructures (MCSs) representing shared core chemical structures within a patent. The PL score assesses the similarity between optimized molecules and these MCSs, offering a pertinent evaluation of patentability.

Baselines

In Section , we present our main results and compare MOMP against several baseline models, both SMILES and Graph-based. We evaluate **UGMMT**, **JTVAE**, and **REINVENT** in both their original versions (Barshatski and Radinsky 2021; Jin, Barzilay, and Jaakkola 2018; Olivecrona et al. 2017) and with the incorporation of patents-loss (Turutov and Radinsky 2023). For these models, we used their original datasets for training and testing. The patents-loss function was applied to these models using all optional variations, and we report the performance of the models with the patents-loss function that achieved the best results.

Additionally, we evaluate **CORE** (Fu, Xiao, and Sun 2020), **G2G** (Jin et al. 2019) and **Mol-CG** (Maziarka et al. 2020), using their original datasets for each property. For **MIMOSA** (Fu et al. 2021), we optimize both the property and (1-PL) using its original pre-trained latent space of the GNNs and original dataset. As for **IPCA** (Barshatski, Nordon, and Radinsky 2021), we use datasets A, B, and C during training for both property and (1-PL) optimization.

Experiments and Results

We begin by comparing our model's ability to optimize successful molecules in comparison to SOTA methods (Main Result Section). Then, we conduct extensive ablation experiments to demonstrate the effect and necessity of key components in our model (Ablation Experiments Section).

Main Result: Molecule Optimization

The results of our experiments are summarized in Table 1 provided. We observe across all experiments that MOMP reaches a significantly higher total success rate. Moreover, the similarity of the generated molecules to the lead molecules is similar to the results reached by the other optimization algorithms. In terms of property optimization, the MOMP model achieves relatively high QED and DRD2 values, while simultaneously achieving the lowest PL value among models that improved the property. Interestingly,

		MOMP	No pre-training	No C-fps	No Extended-EETN	Only A $\rightleftarrows$ C	Only A $\rightleftarrows$ B
QED	Property ↑	0.842	0.814	0.848	0.829	0.843	0.814
	PL ↓	0.436	0.428	0.459	0.438	0.452	0.408
	Similarity	0.282	0.261	0.320	0.302	0.315	0.130
	Validity	0.914	0.570	0.984	0.940	0.977	0.879
	Novelty	1.000	1.000	1.000	1.000	1.000	1.000
	Success	**0.272**	0.217	0.213	0.260	0.225	0.088
DRD2	Property ↑	0.746	0.215	0.711	0.746	0.824	0.789
	PL ↓	0.584	0.433	0.593	0.585	0.598	0.619
	Similarity	0.269	0.206	0.266	0.257	0.201	0.183
	Validity	0.981	0.025	0.977	0.972	1.000	0.950
	Novelty	0.892	0.950	0.907	0.886	0.876	0.868
	Success	**0.322**	0.100	0.269	0.301	0.263	0.212

Table 2: Performance Comparison of MOMP and Ablation Experiments for Various Metrics.

when focusing on single-property optimization, we found that as the QED or DRD2 value increased, so did the PL score, indicating a higher resemblance to existing patents. This trend was observed for both QED and DRD2 properties, suggesting that molecules with higher property values tend to have greater overlap with existing patented compounds. Nevertheless, our MOMP model was able to achieve relatively high property results while maintaining a lower PL score, indicating its effectiveness in balancing property enhancement and patentability constraints.

We draw the reader's attention to the comparison of the MOMP model with multi-property optimization models: IPCA and MIMOSA. Those were specifically trained to optimize numerous targets: the QED or DRD2 property while simultaneously minimizing the PL score. Our results indicate that treating the PL score as an additional target reaches inferior results as compared to a joint optimization as performed in MOMP.

Overall, our empirical evaluation showcases the effectiveness of the MOMP model in optimizing molecules under the patentability constraint. It achieves superior property improvements while minimizing patent infringement, demonstrating its potential as a valuable tool in the drug discovery process.

Ablation Experiments

We conducted ablation experiments on the MOMP model, and the results are summarized in Table 2.

(1) No Pre-training Experiment: To assess the impact of pre-training METNs before the end-to-end model training, we observed that pre-training plays a crucial role in optimizing molecules for validity. This highlights its significance in the holistic optimization process.

(2) No C-fingerprints (fps) Experiment: By using the fps of the input molecule instead of selecting a similar molecule from domain C, we achieved higher similarity between input and optimized molecules. However, this approach showed a trade-off, leading to less effective property optimization and a smaller decrease in PL values. This underscores the importance of utilizing similar molecules from domain C for a more balanced prioritization of property improvement and adherence to patentability constraints.

(3) No Extended-EETN Experiment: Disabling the Extended-EETN pathway, which connects domains A and C, resulted in a noticeable decrease in the similarity between the input and optimized molecules. This highlights the crucial role of the A $\rightleftarrows$ C pathway in preserving the optimized molecule's similarity to the input molecule.

(4) Only A $\rightleftarrows$ C Experiment: Restricting the model to use only one EETN component, connecting domains A and C, led to an increase in PL values. This underscores the necessity of incorporating both EETN components and the Extended-EETN pathway in MOMP architecture for an effective balance between similarity preservation and adherence to patentability constraints during optimization.

(5) Only A $\rightleftarrows$ B Experiment: Utilizing only one EETN component, connecting domains A and B, resulted in a notable decrease in similarity. This reduction can be attributed to the absence of a direct translation path between domains A and B and the presence of domain C fingerprints (C-fps), impacting the overall model performance.

Conclusions

In this research, we propose the Molecule Optimization Model with Patentability Constraint (MOMP), a novel approach to molecule optimization under patentability constraints in drug discovery. MOMP's multi-stage optimization framework enables it to optimize molecules with enhanced properties while avoiding infringement on existing patents. Through empirical evaluation, we demonstrate the superiority of MOMP over SOTA models, achieving improved property optimization while ensuring the non-infringement of existing patents. Our work presents a significant advancement in the field of molecule optimization, offering a practical and efficient solution to the complex problem of patent-constrained optimization in drug discovery. By mitigating patent infringement concerns, MOMP facilitates the acceleration of drug development and contributes to the search for safer and more effective pharmaceutical compounds.

References

Bajusz, D.; Rácz, A.; and Héberger, K. 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7.

Barshatski, G.; Nordon, G.; and Radinsky, K. 2021. Multi-Property Molecular Optimization Using an Integrated Poly-Cycle Architecture. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, CIKM '21, 3727–3736. New York, NY, USA: Association for Computing Machinery.

Barshatski, G.; and Radinsky, K. 2021. Unpaired Generative Molecule-to-Molecule Translation for Lead Optimization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '21, 2554–2564. New York, NY, USA: Association for Computing Machinery.

Fu, T.; Xiao, C.; Li, X.; Glass, L. M.; and Sun, J. 2021. Mimosa: Multi-constraint molecule sampling for molecule optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 125–133.

Fu, T.; Xiao, C.; and Sun, J. 2020. Core: Automatic molecule optimization using copy and refine strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 638–645.

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.-Y.; and Ma, W.-Y. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29.

Jastrzebski, S.; Leśniak, D.; and Czarnecki, W. M. 2016. Learning to smile (s). *arXiv preprint arXiv:1602.06289*.

Jin, W.; Barzilay, R.; and Jaakkola, T. 2018. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *International Conference on Machine Learning*, 2323–2332.

Jin, W.; Yang, K.; Barzilay, R.; and Jaakkola, T. 2019. Learning Multimodal Graph-to-Graph Translation for Molecule Optimization. In *International Conference on Learning Representations*.

Landrum, G. 2016. Rdkit: Open-source cheminformatics software.

Maziarka, ; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Danel, T.; and Warchoł, M. 2020. Mol-CycleGAN: a generative model for molecular optimization. *Journal of Cheminformatics*, 12.

Olivecrona, M.; Blaschke, T.; Engkvist, O.; and Chen, H. 2017. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1): 1–14.

Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; et al. 2016. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic acids research*, 44(D1): D1220–D1228.

Rogers, D.; and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5): 742–754.

Turutov, S.; and Radinsky, K. 2023. Generating Optimized Molecules without Patent Infringement. In *Proceedings of the 32th ACM International Conference on Information and Knowledge Management*, CIKM '23. New York, NY, USA: Association for Computing Machinery.