Gotta be *SAFE*: A New Framework for Molecular Design

Emmanuel Noutahi
Valence Labs
Montreal, QC, Canada
emmanuel@valencediscovery.com

Cristian Gabellini
Valence Labs
Montreal, QC, Canada
cristian@valencediscovery.com

Michael Craig
Valence Labs
Montreal, QC, Canada
michael@valencediscovery.com

Jonathan S.C Lim
Mila-Quebec AI Institute & Valence Labs
Montreal, QC, Canada
jonathan.lim@u.nus.edu

Prudencio Tossou
Valence Labs
Montreal, QC, Canada
prudencio@valencediscovery.com

Abstract

Traditional molecular string representations, such as SMILES, often pose challenges for AI-driven molecular design due to their non-sequential depiction of molecular substructures. To address this issue, we introduce **S**equential **A**ttachment-based **F**ragment **E**mbedding (SAFE), a novel line notation for chemical structures. SAFE reimagines SMILES strings as an unordered sequence of interconnected fragment blocks while maintaining full compatibility with existing SMILES parsers. It streamlines complex generative tasks, including scaffold decoration, fragment linking, polymer generation, and scaffold hopping, while facilitating autoregressive generation for fragment-constrained design, thereby eliminating the need for intricate decoding or graph-based models. We demonstrate the effectiveness of SAFE[1] by training an 87-million-parameter GPT2-like model on a dataset containing 1.1 billion SAFE representations. Through extensive experimentation, we show that our SAFE-GPT model exhibits versatile and robust optimization performance. SAFE opens up new avenues for the rapid exploration of chemical space under various constraints, promising breakthroughs in AI-driven molecular design.

1 Introduction

Molecular design, which consist of constructing molecules with desired characteristics, is a critical task in computational drug discovery. It often necessitates the preservation of certain scaffolds or core chemical substructures, which serve as the backbone for the design process, The motivation for preserving these groups and constraints typically stems from their crucial role in the molecule's biological activity. Nevertheless, incorporating such constraints can be challenging when relying on conventional molecular string representations like the Simplified Molecular Input Line Entry System (SMILES).

[1]Code, data and model available at https://github.com/datamol-io/safe/

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

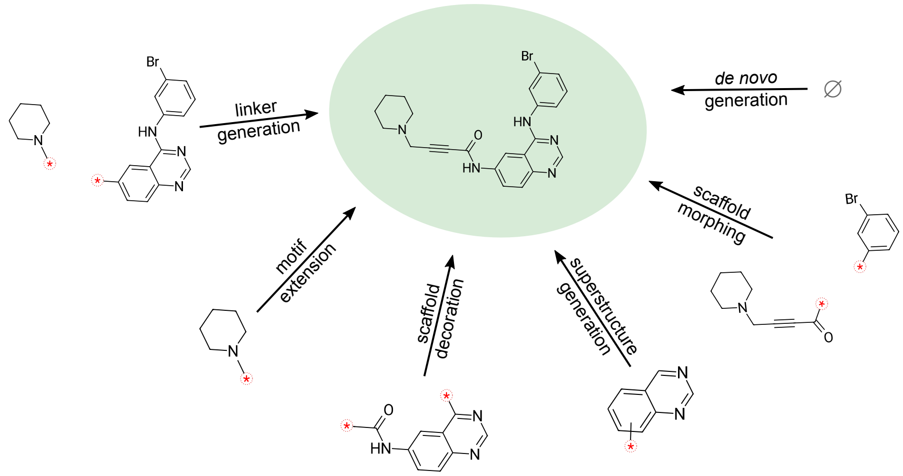


Figure 1: Molecular design tasks that can be performed easily with SAFE

Although SMILES has played a crucial role in chemistry and drug discovery, it is unable to provide a contiguous representation of molecular substructures. This limitation hinders tasks like adding structures to a molecule's scaffold and connecting fragments, limiting its usefulness in improving potential drug candidates, particularly during lead optimization efforts. Addressing these challenges requires the development of an enhanced line notation for molecules, one that can preserve the integrity of molecular scaffolds and fragments while offering flexibility for *de novo* molecular design.

To this end, we introduce Sequential Attachment-based Fragment Embedding (SAFE), a novel line notation for molecules. In contrast to existing methods, SAFE represents molecules as an unordered sequence of fragment blocks. This re-imagines molecular design tasks, transforming them into simpler sequence completion problems. Moreover, SAFE facilitates autoregressive generation, effectively bypassing the need for intricate decoding schemes or graph-based models (see Figure 1 and Table 1). Importantly, despite these novel features, SAFE strings are backward compatible with SMILES parsers, promising an easy integration into existing workflows. Our contributions can be summarized as follow:

- We introduce SAFE, a novel molecular representation compatible with SMILES that represents molecules as a sequence of interconnected fragments.
- We trained a GPT-like generative model with 87.3 million parameters using a dataset of 1.1 billion SAFE strings and demonstrated how this model can be effectively applied to various molecular design tasks, taking advantage of SAFE's unique characteristics.
- We propose a new benchmark inspired by real-world drug discovery challenges to assess generative models' performance in tasks such as scaffold decoration, linker design, and motif extension.

2 Related Works

Molecular line notation representations: The Simplified Molecular-Input Line-Entry System (SMILES) [Weininger, 1988] is the most widely adopted molecular line notation in chemoinformatics due to its simplicity, compactness, and human readability. However, SMILES are not robust to small changes and faces challenges in ensuring the validity and integrity of fragments in deep learning-based molecular design and performs poorly in molecular search and substructure matching tasks. Alternative notations have emerged to address these limitations. In particular, Self-Referencing Embedded Strings (SELFIES) [Krenn et al., 2020, 2022] tackle the robustness and validity issues in deep generative modeling, through a recursive approach, improving performance compared to notations like DeepSmiles [O'Boyle N, 2018] and GenSMILES [Bhadwal et al., 2023]. However, SELFIES come at the cost of simplicity, intepretability and compactness. None of these notations consistently uphold the integrity of scaffolds and fragments essential for most molecular generation tasks. Similarly, the International Chemical Identifier (InCHI), a fixed-length textual representation,

Table 1: Pure generative capabilities of various molecular representations. In the assessment of the inherent generative capabilities of each molecular representation, we employ a marking system: ✓ signifies intrinsic competence, ? indicates the need for intentional engineering, and ✗ suggests unverified capabilities.

Task	SAFE	SMILES	DeepSMILES	GenSMILES	SELFIES	InChi	GRAPHS
De novo design	✓	✓	✓	✓	✓	?	✓
Linker design	✓	?	✗	✗	✗	✗	?
Motif extension	✓	?	✗	✗	?	✗	✓
Scaffold decoration	✓	?	✗	✗	✗	✗	✓
Scaffold morphing	✓	✗	✗	✗	✗	✗	?
Superstructure	✓	✗	✗	✗	✗	✗	✓

ensures global uniqueness and compatibility across chemical databases but is not suitable for molecular generation tasks. In Table 1, we contrast the generative capabilities of various molecular line notations, including SAFE.

Deep generative design: To contextualize our work within the domain of deep generative design, particularly concerning sequence-based and graph-based approaches, we refer interested readers to comprehensive reviews provided in [David et al., 2020, Bilodeau et al., 2022, Du et al., 2022]. Herein, we briefly describe sequence-based and graph-based deep generative models.

In sequence-based methods, initial attempts by Gómez-Bombarelli et al. [2018] focused on generating molecular notations character by character. This approach provided considerable versatility but faced challenges when dealing with fragment-based constraints. Nevertheless, recent advancements have attempted to address this limitation by separately generating scaffolds and side chains [Liao et al., 2023], introducing transformations derived from matched molecular pairs analysis [He et al., 2022], and employing conditional generation [Yang et al., 2021, Bagal et al., 2021].

In the realm of graph-based methods, our work shares similarities with [Jin et al., 2018a, 2020, Maziarz et al., 2021], which uses motifs for molecular graphs but encounter difficulties when extending design to scaffold-based generation, linker-design and generating molecules with unseen building blocks. In particular, these methods, while capable of assembling motifs in a tree-like structure, face challenges when it comes to creating arbitrary cyclic structures with unseen cycles during training.

Constrained molecular design: Notable contributions have emerged in the recent literature on constrained molecular design. Li et al. [2018a] introduced a conditional graph generative model that excels in producing valid molecules while offering the flexibility needed for multi-objective optimization. MolGPT [Bagal et al., 2021], which uses a transformer-decoder architecture for the generation of drug-like molecules, has demonstrated the capacity to conditionally control diverse molecular properties and scaffold designs, highlighting its efficacy in crafting molecules tailored to specific requirements. Furthermore, Multi-Constraint Molecular Generation (MCMG) [Wang et al., 2021], combining conditional transformers, knowledge distillation, and reinforcement learning, has shown the capability to satisfy multiple constraints during the process of molecular generation.

Scaffold-conditioned generation: Under hard scaffold constraints, Lim et al. [2020a] proposed a graph-based model explicitly trained on scaffold and molecule pairs. Under soft scaffold constraints, Li et al. [2018b] have considered the scaffold as part of the input, but their approach does not guarantee its presence in the generated molecules. Arús-Pous et al. [2020] used an iterative conditional training procedure to perform scaffold decoration with an LSTM trained on SMILES. Their work was extended in [Fialková et al., 2021], where a reaction-driven approach for scaffold decoration was proposed. Finally, Langevin et al. [2020] proposed a sampling algorithm that can adapt any SMILES-based auto-regressive model to work with scaffolds. However, being trained on SMILES, their models can neither guarantee validity of generated molecules nor the presence of the input scaffold constraint.

3 SAFE algorithm

In SMILES, ring structures are marked by using digits to identify the opening and closing ring atom, thus denoting a virtual connection between the corresponding atoms. This rule also contributes to the surjectivity of SMILES representation where multiple different SMILES correspond to the same molecular graph. SAFE (Sequence Attachment-based Fragment Embedding) leverages this rule to discover alternative SMILES strings that enforce an order of SMILES characters in which all SMILES tokens belonging to the same molecular fragment are consistently arranged consecutively (see Figure 2). As such, SAFE is a molecular line notation that reimagines SMILES as a collection of connected fragments and remains a valid SMILES representation. Furthermore, the arrangement of fragments within a SAFE string has no impact on the underlying molecular graph, ensuring that common data augmentation techniques for generative models, such as randomization, remain applicable.

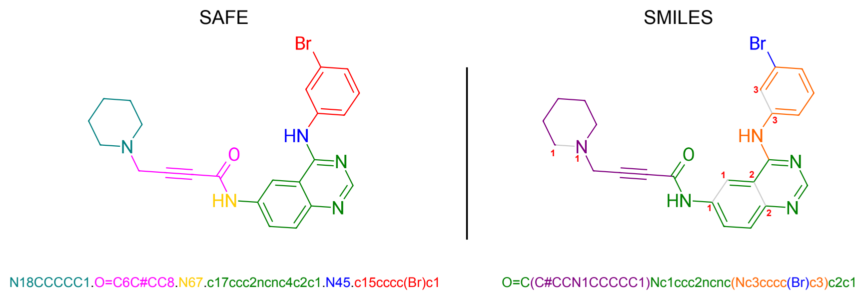


Figure 2: Example of a molecule as a SAFE and SMILES representation. The colored fragments and their corresponding placement in each string show how the ordering of the fragments in the SAFE representation are more easily readable and interpretable than the comparable SMILES string.

3.1 Constructing A SAFE string

The detailed process to convert from SMILES to SAFE is illustrated by Algorithm 1 and Figure 5. It starts by extracting all unique ring digits from the associated molecule and fragmenting it on a desired set of bonds. Our implementation employs the BRICS algorithm (Degen et al. [2008]), although other bond-splitting algorithms, such as Hussain-Rea [Hussain and Rea, 2010], RECAP citeplewell1998recap, or custom patterns, are equally valid. These substructures may represent synthetically accessible building blocks that are common in drug-like compounds. The extracted fragments are then sorted by size and joined with the dot character (**"."**) to mark new fragments in the representation, while preserving their corresponding attachment points. To construct the final SAFE string, we iterate over the numbered attachment points and replace them by novel ring digits to simulate fragment linking. The new ring digits added act as virtual connections between fragments so we obtain a set of linked fragments, as indicated by the presence of the dot character ("."). It's worth noting that, similar to canonicalization in SMILES to produce a single special generic SMILES from multiple valid possibilities, we can achieve a similar outcome by enforcing a decoding order not only on SMILES characters within fragments but also on fragment orders within the final SAFE string.

Algorithm 1 Conversion of SMILES to SAFE Representation

```
procedure ToSAFE(molecule)
    ring_digits ← extract all unique ring digits from molecule
    fragments ← fragment molecule on specified bonds          ▷ We use BRICS bonds here
    Sort fragments in fragments by size in descending order
    fragments_str ← {}
    for each frag in fragments do
        Add smiles of frag to fragments_str
    safe_str ← join all elements in fragments_str with "."
    attach_pos ← extract all attachment points from safe_str
    i ← max (ring_digits) + 1                                 ▷ Find the next possible ring digits
    for each attach in attach_pos do
        Replace attach in safe_str with i
        Increment i by 1
    return safe_str
```

3.2 SAFE facilitates fragment-based design

The inherent sequential block structure of SAFE presents a distinctive advantage for fragment-based design tasks. Traditionally, such endeavors primarily relied on graph-based generative models. However, with a generative model trained on SAFE strings, fragment-based design becomes remarkably straightforward (refer to Figure 1).

Among those, we found the following particularly suitable for SAFE:

- **De novo generation:** which consists of sampling a new sequence from the learned token distribution. It's as straightforward with SAFE as with established SMILES-based auto-regressive models used in molecular generation.
- **Scaffold decoration and motif extension:** which can be framed as sequence completion and new tokens prediction to create novel fragments using SAFE. Starting with an initial sequence corresponding to a scaffold or motif, and marked attachment points for completion, SAFE simplifies this compared to other notations.
- **Linker design and scaffold morphing:** that can also be approached as sequence completion task. Since the order of fragments in a SAFE string doesn't affect the underlying molecular graph, the fragments to be linked can be provided as the initial sequence for a generative model to predict likely tokens for the missing linker.
- **Superstructure generation:** in this setting, the goal is to generate new molecules while adhering to a specified substructure constraint. In the SAFE framework, we achieve this by first generating random attachment points on the substructure to create new scaffolds, followed by scaffold decoration.

4 Experiments

To evaluate the utility of our new molecular line notation, we developed a generative model using a decoder-only transformer architecture. Our aim is to showcase the model's ability, trained on SAFE strings, to generate valid and diverse molecules in *de novo* scenarios. Additionally, we seek to evaluate its effectiveness in practical, real-world scenarios where tasks like scaffold decoration, scaffold morphing, linker design and goal-directed generation are required.

4.1 SAFE-GPT: SAFE generative model

Dataset: We began by constructing a vast chemical dataset comprising over 1 billion unlabeled molecules for pre-training purposes. This dataset was carefully constructed by combining molecules from the ZINC and UniChem libraries [Irwin and Shoichet, 2005, Chambers et al., 2013], resulting in a diverse collection of 1.1 billion SMILES strings. Our dataset spans various molecule types, encompassing drug-like compounds, peptides, multi-fragment molecules, polymers, reagents and non-small molecules, ensuring the wide applicability of our generative model. It stands as the largest and most diverse dataset designed specifically for deep generative molecular design. To convert SMILES strings into SAFE strings, we utilized a combination of BRICS decomposition and a graph partitioning method (Louvain community detection), when BRICS bonds where not available. Molecules that couldn't undergo successful fragmentation were excluded from our dataset. For our experiments we do not use randomization of fragment positions or SMILES ordering due to the already large dataset.

Tokenizer: We trained a BPE tokenizer on the full dataset. As a pre-tokenization step for the inputs, we applied a common regular expression for SMILES syntax [Schwaller et al., 2019]. This process yielded a vocabulary of 1180 tokens, including all special tokens (*EOS, BOS, UNK, MASK, PAD*).

Model architecture: Our SAFE Generative model (SAFE-GPT) is a 87.3M parameters GPT2-like transformer. It comprises 12 layers, each with 12 attention heads per layer, and a hidden state size of 768. All other model parameters adhere to the default settings of GPT-2, as outlined in Hugging Face.

Model training: The SAFE Model was trained using cross-entropy with the next token prediction as training objective. We use the AdamW optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) [Kingma and Ba, 2014], a linear learning rate scheduler with 10000 warmup steps and an initial $lr = 1e-4$. We set the batch size to 100 per GPU and used 2 steps of gradient accumulation and gradient checkpointing. The model was trained on 4 Nvidia A100 GPUs, for a maximum of 1000000 steps (7 days).

4.2 De novo generation results

In *de novo* design, our objective is to generate entirely novel compounds with desirable profiles. Assessing a model's ability to generate valuable compounds in such a setting, even without an optimization objective is crucial, as some models may encounter problems generating valid or sufficiently diverse and novel compounds. To evaluate this, we rely on classical metrics commonly used for generative machine learning models, including compound validity, uniqueness, and internal diversity. These metrics serve to confirm that the model has effectively learned SAFE's grammar and syntax and can generate chemically valid structures.

Table 2: **Molecule generation results on 10K samples** SAFE-GPT performs similarly to generative models trained on the MOSES dataset while producing more diverse molecules.

Model	**Validity@10K**	**Uniqueness@10k**	**Diversity**
SAFE-GPT	0.984*	**1**	**0.878**
CharRNN	0.975	0.999	0.856
VAE	0.977	0.998	0.856
AAE	0.937	0.997	0.856
LatentGAN	0.897	0.997	0.857
JT-VAE	**1**	0.999	0.855
LigGPT	0.900	0.999	0.871
GMT-SELFIES	**1**	**1**	0.870

* SAFE-GPT use a different training dataset that includes non drug-like and challenging molecules.

Table 2 Results: Table 2 presents the results of generating 10,000 molecule samples using SAFE-GPT, which are compared to those of other generative models. Despite differences in training datasets and the incorporation of challenging molecules in SAFE-GPT's training data, it produces molecules with similar levels of validity, uniqueness, and diversity. Notably, SAFE-GPT outperformed other models in terms of uniqueness and diversity. However, when compared to models like JT-VAE [Jin et al., 2018b] and SELFIES-based models such as GMT-SELFIES [Wei et al., 2023], SAFE-GPT exhibited a slightly lower validity score. This could be attributed to the increased complexity of understanding fragment connectivity, represented as pairs of digits, which is a challenge also observed in most SMILES-based representations.

In Figure 6, we show a subset of randomly selected molecules generated with SAFE-GPT. This visual representation offers readers an intuitive sense of the quality and reasonableness of the generated molecules. Furthermore, in Figure 7, we show the distribution of selected molecular properties for the 10,000 generated molecules.

4.3 Performance on fragment-constrained generation

De novo compound generation is only one approach for advancing a drug discovery program. In fact, in many real-world scenarios, generative design involves modifying existing molecules in user-defined ways rather than creating entirely new compounds. This is especially true in later stages of drug discovery, such as hit-to-lead or lead optimization, where well-established structure-activity relationships (SAR) are already in place. Therefore, we examined SAFE's intended capabilities for performing fragment-constrained generative design tasks such as scaffold decoration, scaffold morphing, linker generation, motif extension, and superstructure generation (see subsection 3.2). To facilitate this evaluation, we designed a benchmark that involved working with scaffolds and fragments from 10 existing drugs. Further details about the benchmark design can be found in subsection A.3 in the Appendix.

Table 3 presents averaged validity, diversity, and uniqueness scores for 1000 molecules sampled in each fragment-constrained design task using SAFE-GPT across all drugs. It displays the average Tanimoto distance between the generated molecules to the original drug molecules, along with the average SA score (Synthetic Accessibility Score) [Ertl and Schuffenhauer, 2009]. We observe that SAFE-GPT maintains full validity for all sampled molecules under constraints, while achieving high internal diversity and novelty compared to the original drugs. Moreover, generated molecules exhibit a low SA score, indicating their ease of synthesis. For a visual inspection of sample molecules from each task using Maribavir as the starting molecule, please refer to Table 5 (subsection A.3)

Table 3: Performance on fragment-constrained generative design tasks on 500 molecules sampled

Task	Validity	Diversity	Uniqueness	Distance	SA score
Linker design	1.000±0.000	0.641±0.099	0.887±0.191	0.712±0.097	3.864±0.928
Motif extension	1.000±0.000	0.681±0.089	0.923±0.179	0.772±0.101	3.750±0.651
Scaffold decoration	1.000±0.000	0.571±0.113	0.851±0.162	0.643±0.137	4.017±0.889
Scaffold morphing	1.000±0.000	0.608±0.096	0.717±0.219	0.688±0.113	3.604±0.910
Superstructure	1.000±0.000	0.715±0.059	0.929±0.106	0.812±0.063	3.868±0.919

4.4 Goal-directed generative capabilities

To effectively apply generative approaches in live drug discovery projects, it is essential to incorporate goal-directed generation, guiding generation of novel molecules towards specific properties. Therefore, we follow established methodologies [Lim et al., 2020b, Seo et al., 2023] to assess the model's ability for goal-directed generation using simple molecular properties. More precisely, we optimize toward specific values for key molecular properties, including Topological Polar Surface Area (TPSA), Molecular Weight (MW), Calculated LogP (CLOGP), and Quantitative Estimation of Drug-likeness (QED). To achieve this, we use Proximal Policy Optimization (PPO) [Schulman et al., 2017] with Adaptive KL Penalty to train a policy for generating molecular samples with the targeted property value. A total of 50 steps was performed with a learning rate of 1e-5 (AdamW optimizer) and a batch size of 100. The reward objective used for this optimization was defined as follows:

$$\text{reward}(mol) = \frac{1}{1 + \alpha \cdot |\text{prop}(mol) - \text{target}|}$$

where *prop(mol)* represents the calculated molecular property value for a given sample, *target* signifies the desired target value, and α is set to 0.5.

With the methodology described above, we fine-tuned agents for two target values on each molecular property and evaluated their performance by generating 500 samples from each of them. Notably, all generated samples were valid and unique. The property distribution of these samples is visually presented in Figure 3, where the red line within each plot represents the target value of the molecular property that the agent was optimized towards, and the blue and orange histograms representing the distribution of samples from different agents with distinct goals. The results depicted in Figure 3 demonstrate that the property distribution of the generated molecules, achieved through goal-conditioned optimization using PPO, is notably centered around the respective target values. This outcome indicates the success of our optimization process in aligning the generated molecules distribution with the desired property targets.

4.5 Scaffold-Constrained optimization of CNS penetration of EGFR inhibitors

In this section, we introduce a novel and challenging optimization task aimed at improving the Central Nervous System (CNS) penetration of EGFR Tyrosine Kinase Inhibitors. This optimization task specifically addresses the challenge of CNS metastases in non-small cell lung cancer, a significant concern in cancer treatment [Ahluwalia et al., 2018]. Our objective involves optimizing the CNS-MPO score, a comprehensive metric assessing physico-chemical properties associated with CNS penetration [Wager et al., 2016]. The CNS-MPO score ranges from 0 to 6, with higher scores indicating better desirability, and a score above 4 typically suffices. We introduce additional constraints to our optimization task, requiring that all generated molecules feature a scaffold that has

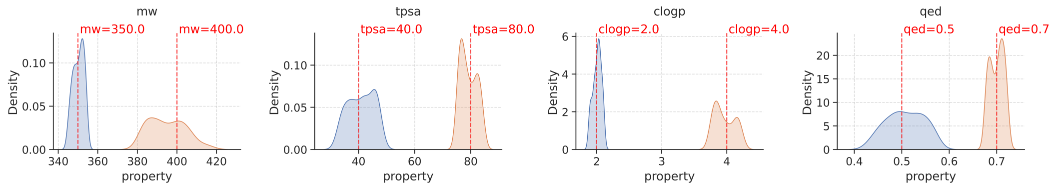


Figure 3: Property distributions of generated molecules, grouped by molecular properties, after goal-conditioned optimization using PPO. The red line in each plot shows the target value the agent was optimized towards.

demonstrated activity against EGFR (see Figure 8). For an in-depth exploration of this topic, please consult subsection A.2 in the Appendix.

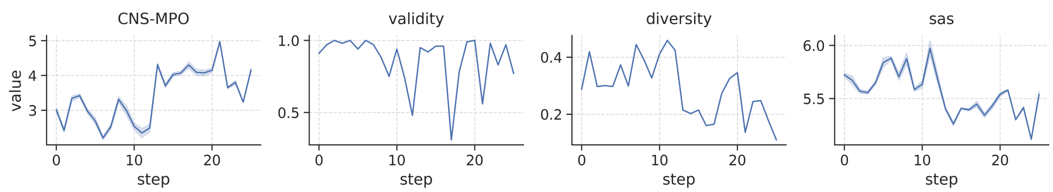


Figure 4: Distribution of CNS-MPO rewards and generative metrics score (validity, internal diversity and SA score) throughout the 25 optimization steps when sampling 100 molecules from the RL agent.

We directly optimize the CNS-MPO score using PPO for 25 steps, and the same training parameters outlined in subsection 4.4.

Figure 4 illustrates the reward distribution obtained by sampling 100 molecules at each optimization iteration. Our findings demonstrate that scaffold-constrained optimization, even when facing challenging metrics, can be efficiently executed with SAFE-GPT using a straightforward optimization algorithm like PPO. As the CNS-MPO policy refines, we observe an expected reduction in the diversity of sampled candidates, while overall validity remains robust. Intriguingly, there's a slight decline in the SA score across iterations, suggesting the presence of synthetically favorable yet optimal compounds within the solution space.

5 Discussion

This work introduces SAFE, a novel molecular representation that maintains compatibility with SMILES parsers while significantly enhancing the versatility and expressive power of molecular design tasks. SAFE represents molecules as sequences of interconnected fragments, providing a new paradigm for molecular description. In the realm of molecular line notations, SAFE emerges as a promising alternative, addressing limitations in existing notations by striking a harmonious balance between simplicity and robustness, making it suitable for a wide range of applications.

Furthermore, we introduce SAFE-GPT, a pioneering generative model comprising 87.3 million parameters, trained on a diverse dataset of 1.1 billion SAFE strings. Through rigorous experimentation, we have demonstrated the model's efficacy across a spectrum of generative and optimization tasks, effectively harnessing SAFE's distinctive attributes.

We postulate that the slightly lower validity observed in SAFE-GPT's generated molecules compared to SELFIES, may be attributed to the increased complexity introduced by SAFE which would require attending to distant token pairs represented as digits. One potential solution would be to enforce phrasal constraints [Post and Vilar, 2018] during sampling. Additionally, in this initial phase of our research, we limited our evaluation to a benchmark set of only 10 drugs for fragment-constrained generation. To provide a more comprehensive assessment of our model's performance, we propose expanding the benchmark set to encompass a significantly larger collection of known drugs and incorporating new generative models designed explicitly to support such tasks.

In future works, we plan to delve deeper into SAFE's performance in multi-property optimization (MPO) scenarios and explore alternative optimization algorithms to further enhance model performance. Additionally, we intend to investigate the integration of a prediction head into the SAFE-GPT architecture, facilitating joint molecular generation and molecular property prediction. Finally, we aim to efficiently scale SAFE-GPT to larger models and datasets, ultimately laying the groundwork for a new generation of foundational models for drug discovery.

Our work brings significant advancements in molecular representation and generative modeling. We believe that these innovations will continue to drive progress in drug discovery, materials science, and other fields where molecular design plays a pivotal role.

A Supplementary Material

A.1 Additional figures

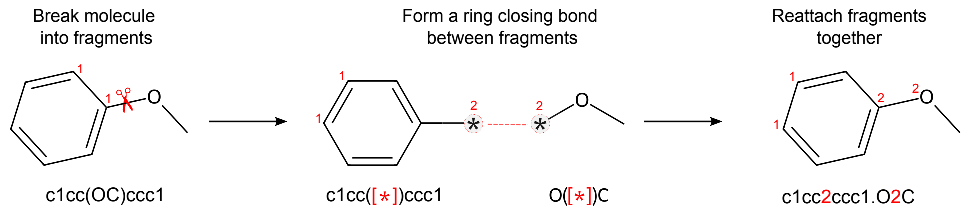


Figure 5: Example encoding of a SMILES string into a SAFE representation. The left panel shows the breaking a bond by the BRICS algorithm. The middle panel shows the addition of attachment points and the ring closing bond connecting the two fragments. The right panel shows the reattached fragments and the final SAFE representation.

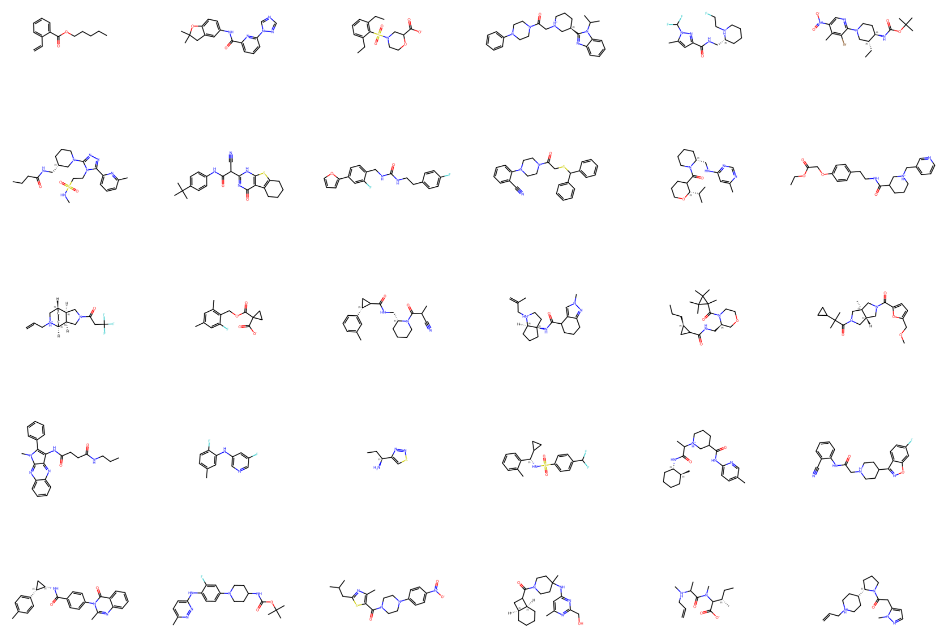


Figure 6: Randomly selected samples of *de novo* generated molecules using SAFE.

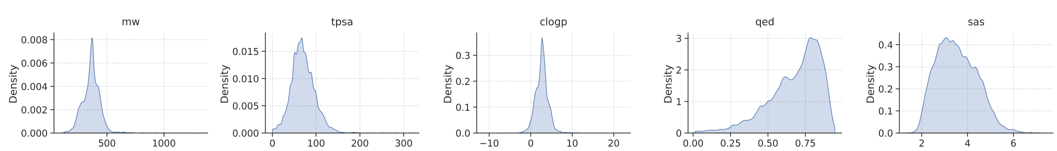
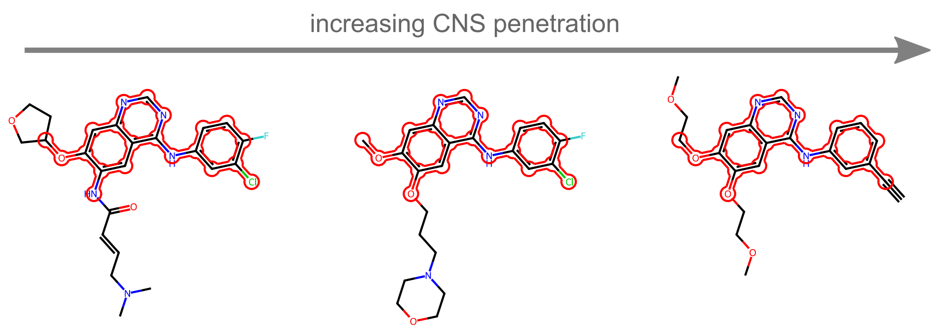


Figure 7: The molecular property distribution for 10,000 molecules generated with SAFE-GPT demonstrates that SAFE-GPT can generate molecules with diverse physicochemical properties, spanning beyond traditional drug-like molecules.

A.2 Optimizing CNS penetration for EGFR inhibitors

Most existing small molecule treatments struggle to effectively penetrate the central nervous system (CNS) due to difficulties in breaching the blood-brain barrier (BBB). Notably, three well-known

EGFR inhibitors (afatinib, gefitinib, and erlotinib), all sharing the same scaffold, exhibit generally low CNS penetration rates, with reported values respectively falling below 1%, in the range of 1%–3%, and in the range of 3%–6%. The ability of a small molecule to penetrate the CNS is often associated with specific physicochemical properties such as CLogD, TPSA, and Molecular Weight. Various scoring systems have been developed to assess this ability. Notably, our findings indicate a correlation between the CNS MPO score [Wager et al., 2016] and the experimental penetration rates for these three EGFR inhibitors.



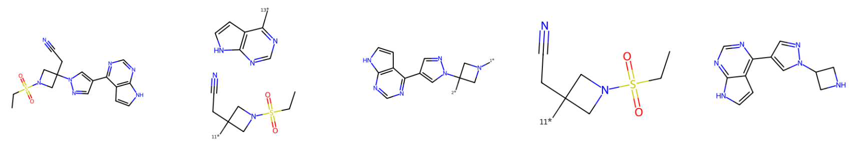
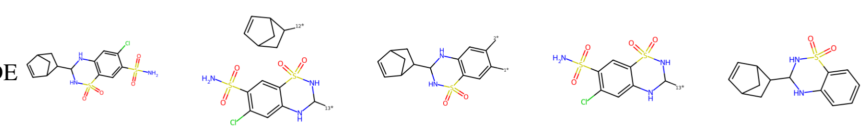
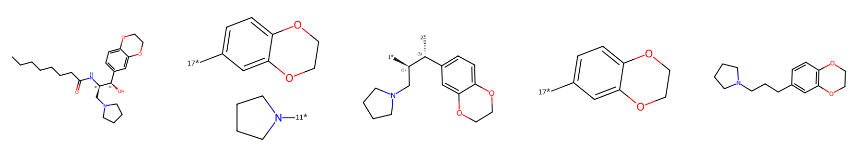
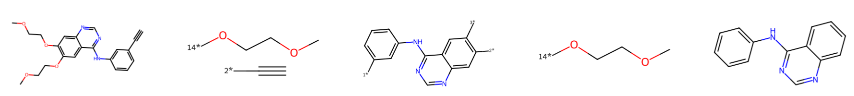
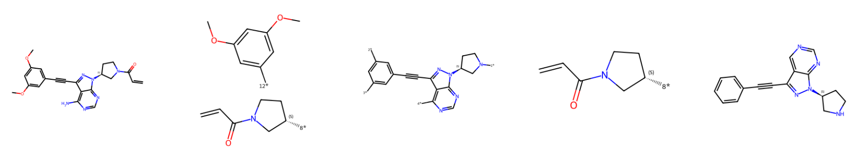
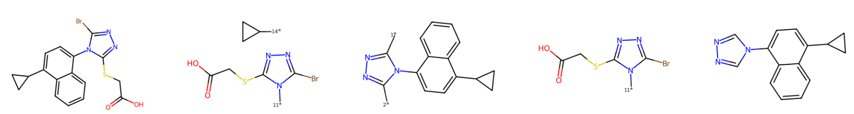
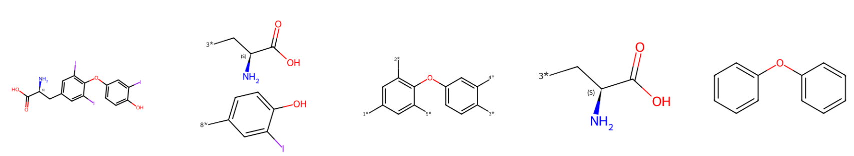
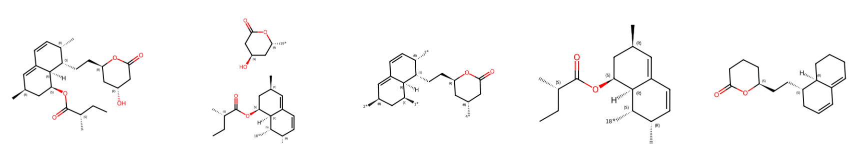
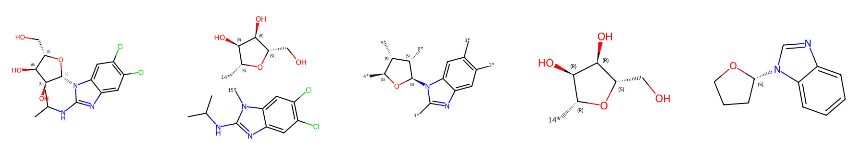
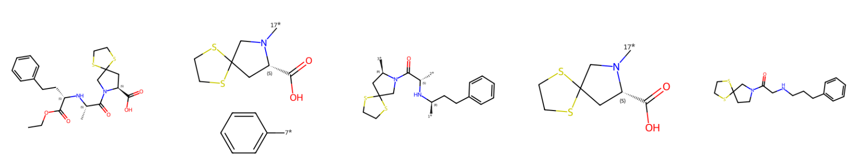
Afatinib	Gefitinib	Erlotinib
CLogD=3.828	CLogD=4.193	CLogD=3.405
TPSA=88.610	TPSA=68.740	TPSA=74.730
CNS-MPO=2.998	CNS-MPO=3.580	CNS-MPO=4.287

Figure 8: Existing EGFR inhibitors and their CNS profile

A.3 Fragment-constrained design results

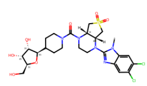
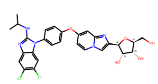
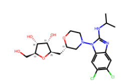
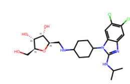
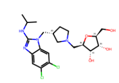
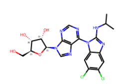
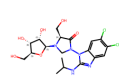
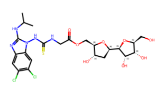
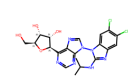
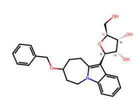
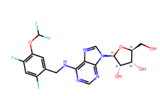
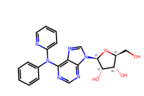
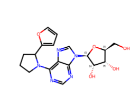
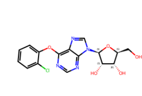
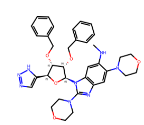
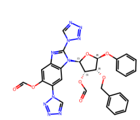
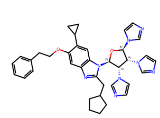
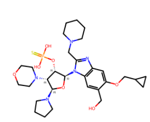
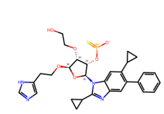
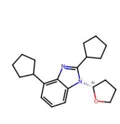
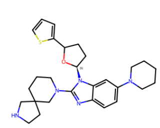
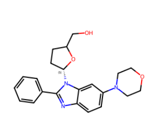
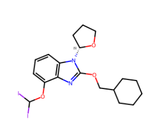
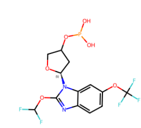
We uses a set of 10 drugs, including **Cyclothiazide**, **Maribavir**, **Spirapril**, **Baricitinib**, **Eliglustat**, **Erlotinib**, **Futibatinib**, **Lesinurad**, **Liothyronine**, and **Lovastatin**. These drugs were chosen as the basis for our fragment-constrained generative design tasks. From each drug, we extracted the main scaffold with attachment points, fragments that serve as side chains, a starting motif, and a core substructure. These components were then respectively used as input for scaffold decoration, linker design / scaffold morphing, motif extension, and superstructure generation, each with its specific objective. The details of the selected drugs and their corresponding inputs for each task can be found in Table 4. It should be noted that linker design and scaffold morphing are two very similar tasks that share the same inputs. In our implementation, the only difference between them lies in the constraints imposed during sampling. For linker design, we employ a constrained beam search to ensure the presence of every fragment in the final molecules. In contrast, for scaffold morphing, new molecules are generated from each fragment with connectivity constraints, after which the scaffold is inferred and linked to the other fragments.

Table 4: List of 10 known drugs and corresponding inputs used by SAFE-GPT for the fragment-constrained benchmark.

Name	Structure	Linker Design*	Scaffold Decoration	Motif Extension	Superstructure
BARICITINIB					
CYCLOTHIAZIDE					
ELIGLUSTAT					
ERLOTINIB					
FUTIBATINIB					
LESINURAD					
LIOTHYRONINE					
LOVASTATIN					
MARIBAVIR					
SPIRAPRIL					

* the linker design and scaffold morphing task share the same input fragments.

Table 5: Examples of generated samples under fragment-constraints for the Maribavir structure

Task	**Generated samples**				
Linker design					
Scaffold morphing					
Motif extension					
Scaffold decoration					
Superstructure					

References

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

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

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. Selfies and the future of molecular string representations. *Patterns*, 3(10), 2022.

Dalke A. O'Boyle N. Deepsmiles: An adaptation of smiles for use in machine-learning of chemical structures. *ChemRxiv. Cambridge: Cambridge Open Engage*, 2018. doi:https://doi.org/10.26434/chemrxiv.7097960.v1. URL `https://chemrxiv.org/engage/chemrxiv/article-details/60c73ed6567dfe7e5fec388d`.

Arun Singh Bhadwal, Kamal Kumar, and Neeraj Kumar. Gensmiles: An enhanced validity conscious representation for inverse design of molecules. *Knowledge-Based Systems*, 268: 110429, 2023. ISSN 0950-7051. doi:https://doi.org/10.1016/j.knosys.2023.110429. URL `https://www.sciencedirect.com/science/article/pii/S095070512300179X`.

Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12(1):1–22, 2020.

Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F. Jensen. Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science*, 12(5):e1608, 2022. doi:https://doi.org/10.1002/wcms.1608. URL `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1608`.

Yuanqi Du, Tianfan Fu, Jimeng Sun, and Shengchao Liu. Molgensurvey: A systematic survey in machine learning models for molecule design. 2022.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Zhirui Liao, Lei Xie, Hiroshi Mamitsuka, and Shanfeng Zhu. Sc2mol: a scaffold-based two-step molecule generator with variational autoencoder and transformer. *Bioinformatics*, 39(1):btac814, 2023.

Jiazhen He, Eva Nittinger, Christian Tyrchan, Werngard Czechtizky, Atanas Patronov, Esben Jannik Bjerrum, and Ola Engkvist. Transformer-based molecular optimization beyond matched molecular pairs. *Journal of cheminformatics*, 14(1):18, 2022.

Lijuan Yang, Guanghui Yang, Zhitong Bing, Yuan Tian, Yuzhen Niu, Liang Huang, and Lei Yang. Transformer-based generative model accelerating the development of novel braf inhibitors. *ACS omega*, 6(49):33864–33873, 2021.

Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9): 2064–2076, 2021.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2323–2332. PMLR, 10–15 Jul 2018a. URL `https://proceedings.mlr.press/v80/jin18a.html`.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *International conference on machine learning*, pages 4849–4859. PMLR, 2020.

Krzysztof Maziarz, Henry Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefl, Marwin Segler, and Marc Brockschmidt. Learning to extend molecular scaffolds with structural motifs. *arXiv preprint arXiv:2103.03864*, 2021.

Yibo Li, Liangren Zhang, and Zhenming Liu. Multi-objective de novo drug design with conditional graph generative model. *Journal of cheminformatics*, 10:1–24, 2018a.

Jike Wang, Chang-Yu Hsieh, Mingyang Wang, Xiaorui Wang, Zhenxing Wu, Dejun Jiang, Benben Liao, Xujun Zhang, Bo Yang, Qiaojun He, et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nature Machine Intelligence*, 3(10):914–922, 2021.

Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim. Scaffold-based molecular design with a graph generative model. *Chemical Science*, 11(4):1153–1164, 2020a. doi:10.1039/c9sc04503a. URL `https://doi.org/10.1039%2Fc9sc04503a`.

Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. Learning deep generative models of graphs. 2018b.

Josep Arús-Pous, Atanas Patronov, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Smiles-based deep generative scaffold decorator for de-novo drug design. *Journal of cheminformatics*, 12(1):1–18, 2020.

Vendy Fialková, Jiaxi Zhao, Kostas Papadopoulos, Ola Engkvist, Esben Jannik Bjerrum, Thierry Kogej, and Atanas Patronov. Libinvent: reaction-based generative scaffold decoration for in silico library design. *Journal of Chemical Information and Modeling*, 62(9):2046–2063, 2021.

Maxime Langevin, Hervé Minoux, Maximilien Levesque, and Marc Bianciotto. Scaffold-constrained molecular generation. *Journal of Chemical Information and Modeling*, 60(12):5637–5646, 2020. doi:10.1021/acs.jcim.0c01015. URL `https://doi.org/10.1021/acs.jcim.0c01015`. PMID: 33301333.

Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3(10):1503–1507, 2008.

Jameed Hussain and Ceara Rea. Computationally efficient algorithm to identify matched molecular pairs (mmps) in large data sets. *Journal of chemical information and modeling*, 50(3):339–348, 2010.

John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

Jon Chambers, Mark Davies, Anna Gaulton, Anne Hersey, Sameer Velankar, Robert Petryszak, Janna Hastings, Louisa Bellis, Shaun McGlinchey, and John P Overington. Unichem: a unified chemical structure cross-referencing and identifier tracking system. *Journal of cheminformatics*, 5(1):3, 2013.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018b.

Lai Wei, Nihang Fu, Yuqi Song, Qian Wang, and Jianjun Hu. Probabilistic generative transformer language models for generative design of molecules. *Journal of Cheminformatics*, 15(1):1–15, 2023.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1:1–11, 2009.

Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim. Scaffold-based molecular design with a graph generative model. *Chemical science*, 11(4):1153–1164, 2020b.

Seonghwan Seo, Jaechang Lim, and Woo Youn Kim. Molecular generative model via retrosynthetically prepared chemical building block assembly. *Advanced Science*, 10(8):2206674, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Manmeet S Ahluwalia, Kevin Becker, and Benjamin P Levy. Epidermal growth factor receptor tyrosine kinase inhibitors for central nervous system metastases from non-small cell lung cancer. *The Oncologist*, 23(10):1199–1209, 2018.

Travis T Wager, Xinjun Hou, Patrick R Verhoest, and Anabella Villalobos. Central nervous system multiparameter optimization desirability: application in drug discovery. *ACS chemical neuroscience*, 7(6):767–775, 2016.

Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. *arXiv preprint arXiv:1804.06609*, 2018.