SAFE setup for generative molecular design

Yassir El Mesbahi
Valence Labs
yassir@valencelabs.com

Emmanuel Noutahi
Valence Labs
emmanuel@valencelabs.com

Abstract

SMILES-based molecular generative models have been pivotal in drug design but face challenges in fragment-constrained tasks. To address this, the Sequential Attachment-based Fragment Embedding (SAFE) representation was recently introduced as an alternative that streamlines those tasks. In this study, we investigate the optimal setups for training SAFE generative models, focusing on dataset size, data augmentation through randomization, model architecture, and bond disconnection algorithms. We found that larger, more diverse datasets improve performance, with the LLaMA architecture using Rotary Positional Embedding proving most robust. SAFE-based models also consistently outperform SMILES-based approaches in scaffold decoration and linker design, particularly with BRICS decomposition yielding the best results. These insights highlight key factors that significantly impact the efficacy of SAFE-based generative models.

1 Introduction

Molecular design is a fundamental task in computational drug discovery, aimed at constructing molecules with desired properties. Recently, deep generative models have emerged as valuable tools for efficiently exploring chemical space and designing novel molecules [8, 18, 6, 23, 11]. Among these, Chemical Language Models (CLMs) which use string-based molecular representations, typically the Simplified Molecular Input Line Entry System (SMILES) [31], have shown robust performance by adapting neural architectures from natural language processing to generate molecular strings [9].

In practical drug design, it is often crucial to either preserve core 'scaffolds' while experimenting with various decoration groups or link several molecular fragments. This is essential for optimizing molecular properties, analyzing structure-activity relationships, and generating intellectual property. However, traditional CLMs require complex adaptations, including new architectures, retraining, custom data processing and sampling algorithms, to handle fragment-constrained tasks.

To address these challenges, the Sequential Attachment-based Fragment Embedding (SAFE) representation was introduced [21]. SAFE simplifies molecular design by treating molecules as unordered sequences of fragment blocks, transforming tasks such as *de novo* generation, fragment linking, and scaffold decoration into simple sequence completion problems, all while remaining compatible with existing SMILES parsers.

Building on the original SAFE work and the generative SAFE-GPT-2 model it introduced, this study investigates how different experimental design choices impact the performance of SAFE-based generative models. Specifically, we assess:

- The influence of bond disconnection algorithms on generative outcomes.
- The impact of dataset size on model performance.
- The performance of various neural architectures.

AI4Mat Workshop @ 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

- The effects of data augmentation via SAFE fragment randomization.

Our goal is to identify the experimental conditions that maximize the generative capabilities of SAFE-based models, thereby providing a comprehensive understanding of the SAFE representation's strengths and limitations in molecular design.

2 Related Works

2.1 Generative Chemical Language Models

Sequence-based generative models, particularly those using molecular line notations like SMILES and SELFIES [14], have grown in popularity in molecular design. These autoregressive models, leveraging advances in natural language processing (e.g., RNNs, transformers, and State Space Models), have demonstrated strong performance in both *de novo* generation and goal-directed molecule design [20, 22, 9]. While SELFIES ensures robustness and validity, it comes at the cost of simplicity, interpretability, and as recent studies suggest, exploration and generalization capabilities compared to SMILES [26]. Additionally, SMILES, with its non-injective nature, allows multiple valid strings for the same molecular graph. This feature, known as SMILES randomization, can be used as data augmentation strategy to enhance model robustness and generalization, especially in data-scarce settings [3, 27, 1, 19].

2.2 Fragment-Constrained Generation with CLMs

The need for fragment-constrained design in drug discovery has driven various adaptations of CLMs. Arús-Pous et al. [2] proposed an encoder-decoder architecture that frames the task as sequence translation. It requires slicing the molecular dataset into scaffolds and attachment groups before training. Initially, this method was used to translate input scaffolds into output decorations, and was later extended in LibINVENT [7] to generate novel scaffold-constrained molecules that follow bespoke chemical reaction rules. Building on this framework, LinkINVENT [10] and SyntaLinker [32] further refined the encoder-decoder architecture. LinkINVENT reverses the translation task by taking a pair of molecular fragments to predict a linker, while SyntaLinker uses a conditional transformer model to translate fragments into fully linked molecules. However, these methods require multiple task-specific architectures, making them less practical due to extensive retraining for each unique chemical design challenge.

More flexible approaches, such as SAMOA [15] and PromptSMILES [28], extend existing SMILES-based CLMs to scaffold decoration and linker design without requiring retraining or custom datasets. SAMOA allows free sampling at attachment/linking points but lacks guarantees for validity or scaffold constraints, while PromptSMILES rearranges SMILES strings iteratively to position attachment points for completion, and can be further complemented by reinforcement learning for fine-tuning towards specific goals.

In contrast, SAFE [21], a sub-grammar of SMILES that reorganizes SMILES strings as unordered sequences of interconnected fragment blocks, simplifies fragment-constrained generation into a sequence completion task. SAFE eliminates the need for complex decoding schemes and has proven effective in both *de novo* and fragment-constrained tasks using a pretrained GPT-2 model. Furthermore, the approach can benefit not only from SMILES randomization but also from a novel form of data augmentation (SAFE randomization), where the order of fragment blocks is randomized to further diversify training data.

3 Experimental Setup

3.1 Building Chemical Language Models

3.1.1 Dataset

We conducted our experiments using the MOSES benchmark dataset [24] provided by TDC [12], consisting of $\sim$1.3M training molecules, $\sim$193k validation, and $\sim$387k test molecules. The dataset was curated from the ZINC clean lead collection by removing molecules with charged atoms; atoms

besides C, N, S, O, F, Cl, Br, H; large cycles, and those failing custom medicinal chemistry filters. For studying the impact of training set size, we downsampled the training data into subsets of 10k and 100k molecules, referred to as MOSES-10k and MOSES-100k, respectively, while the original dataset will be referred to as MOSES-Full.

For each dataset, we converted SMILES into canonical SAFE format using five fragmentation algorithms: Hussain-Rea (HR) [13], BRICS [5], RECAP [16], RDKit's MMPA bond rules (MMPA) [4], and Rotatable bonds (ROTATABLE) [25]. In rare cases where fragmentation failed, resulting molecules were discarded. To evaluate data augmentation effects, we generated a 5x augmented version of MOSES-100k via SAFE fragment order randomization, referred to as MOSES-Augmented. Validation and test splits were consistent across all datasets.

3.1.2 Model

We evaluated four autoregressive CLM architectures: an RNN as a baseline, GPT-2 similar to the original SAFE work, LLaMA [30], and Jamba [17]. GPT-2 uses absolute positional encoding, where fixed positional information is added to token embeddings. In contrast, LLaMA employs Rotary Positional Embeddings (RoPE), which rotates token embeddings based on their relative positions. RoPE improves the model's ability to capture long-range token dependencies by better preserving relationships across varying sequence lengths. Jamba is a novel hybrid Transformer-Mamba model claimed to increase model capacity while keeping active parameter usage low. After performing hyperparameter searches, we trained 92 models in total, across 4 architectures, 3-4 datasets, and 6 representations. Further details on training can be found in section A.2 of the Appendix.

3.2 Evaluation

We assessed model performance in both *de novo* generation and fragment-constrained tasks using standard metrics such as validity, uniqueness, and internal diversity [24, 12, 29], alongside SAFE-specific metrics like fragmentation percentage which measures the proportion of molecular graphs generated with disconnected fragments. For *de novo* generation, we sampled 10,000 molecules across 5 seeds, reporting average performance. In fragment-constrained tasks, we evaluated scaffold decoration using the SureChEMBL (17 scaffolds) and DRD2 (5 scaffolds) benchmarks [15], and both scaffold decoration and fragment linking on the DRUG dataset (10 molecules) proposed in Noutahi et al. [21]. The DRUG dataset serves as a challenging out-of-distribution benchmark due to its divergence from the training MOSES dataset. Any molecules or fragments with tokens not found in the respective vocabularies were excluded from evaluation. We benchmarked SAFE-based models against the closest SMILES-based frameworks, SAMOA and PromptSMILES. For fragment-constrained tasks, benchmarking was conducted using 5,000 samples per constraint, with all methods trained on MOSES-Full.

4 Results

4.1 Effect of Model Architecture

Table 1 compares the generative performance of the four architectures across the MOSES-Full dataset for each representation, and Figure 1 illustrates their overall performance across all datasets. Both show that while smaller RNN models are efficient at generating novel, diverse, and unique molecules, they exhibit limitations in validity and struggle with the complexities of the SAFE representation grammar. By contrast, LLaMA consistently generates the most valid and least fragmented molecules on SAFE strings, highlighting its superior capacity to model the underlying data distribution and grasp the syntax of the SAFE line notation. However on internal diversity, LLaMa models underperformed compared to RNNs. Intuitively, models that excel at fitting the data and can capture its intricate rules are also likely to experience a tradeoff, with a reduction in diversity and novelty. These observations align with previous findings that sampling generalization in CLMs (measured by internal diversity and novelty), does not necessarily correlate with validity [26].

Table 1: Performance comparison of SAFE and SMILES models on MOSES-Full (10k samples, 5 replicates). The best performance within each representation is highlighted in gray, with the overall best performance in red. * denotes results not statistically different from the best ($\alpha = 0.01$).

Representation	Model	↑ Validity (%)	↑ Uniqueness (%)	↑ Novelty (%)	↑ Int.Div (%)	↓ Fragmented (%)
SMILES	Jamba	0.995 ± 0.001	0.997 ± 0.000	1.000 ± 0.000*	0.838 ± 0.002	0.000 ± 0.000*
	LLaMA	0.998 ± 0.001*	0.998 ± 0.001*	1.000 ± 0.000*	0.842 ± 0.001	0.000 ± 0.000*
	GPT-2	0.994 ± 0.001	0.997 ± 0.000	1.000 ± 0.000*	0.834 ± 0.001	0.000 ± 0.000*
	RNN	0.987 ± 0.001	0.999 ± 0.000*	1.000 ± 0.000*	0.835 ± 0.001	0.000 ± 0.000*
SAFE-BRICS	Jamba	0.967 ± 0.001	0.998 ± 0.000	0.806 ± 0.005	0.847 ± 0.000	0.033 ± 0.002
	LLaMA	0.990 ± 0.001	0.997 ± 0.000	0.765 ± 0.004	0.847 ± 0.000	0.014 ± 0.001
	GPT-2	0.970 ± 0.002	0.998 ± 0.000	0.858 ± 0.003	0.847 ± 0.000	0.029 ± 0.001
	RNN	0.938 ± 0.003	1.000 ± 0.000*	0.915 ± 0.003	0.859 ± 0.000	0.105 ± 0.003
SAFE-RECAP	Jamba	0.952 ± 0.004	0.998 ± 0.001	0.783 ± 0.006	0.847 ± 0.000	0.016 ± 0.002
	LLaMA	0.991 ± 0.001	0.997 ± 0.000	0.739 ± 0.005	0.848 ± 0.000	0.005 ± 0.001
	GPT-2	0.978 ± 0.002	0.997 ± 0.001	0.803 ± 0.004	0.847 ± 0.001	0.006 ± 0.001
	RNN	0.883 ± 0.002	1.000 ± 0.000*	0.878 ± 0.001	0.857 ± 0.000	0.039 ± 0.004
SAFE-HR	Jamba	0.942 ± 0.002	0.999 ± 0.000	0.907 ± 0.002	0.852 ± 0.000	0.124 ± 0.004
	LLaMA	0.974 ± 0.001	0.998 ± 0.000	0.868 ± 0.004	0.850 ± 0.001	0.064 ± 0.002
	GPT-2	0.925 ± 0.003	0.998 ± 0.000	0.902 ± 0.003	0.851 ± 0.000	0.088 ± 0.003
	RNN	0.854 ± 0.002	1.000 ± 0.000*	0.989 ± 0.001	0.866 ± 0.000*	0.333 ± 0.005
SAFE-MMPA	Jamba	0.967 ± 0.002	0.998 ± 0.000*	0.888 ± 0.002	0.849 ± 0.000	0.095 ± 0.003
	LLaMA	0.984 ± 0.001	0.998 ± 0.000	0.844 ± 0.002	0.848 ± 0.001	0.049 ± 0.002
	GPT-2	0.968 ± 0.002	0.998 ± 0.000	0.890 ± 0.003	0.849 ± 0.000	0.063 ± 0.002
	RNN	1.000 ± 0.000*	1.000 ± 0.000*	0.967 ± 0.001	0.862 ± 0.000	0.250 ± 0.007
SAFE-ROTATABLE	Jamba	0.959 ± 0.002	0.998 ± 0.000	0.877 ± 0.003	0.850 ± 0.000	0.064 ± 0.002
	LLaMA	0.988 ± 0.001	0.997 ± 0.001	0.813 ± 0.003	0.849 ± 0.001	0.020 ± 0.001
	GPT-2	0.974 ± 0.002	0.998 ± 0.000	0.884 ± 0.003	0.849 ± 0.000	0.041 ± 0.002
	RNN	0.919 ± 0.002	1.000 ± 0.000*	0.954 ± 0.002	0.861 ± 0.000	0.141 ± 0.002

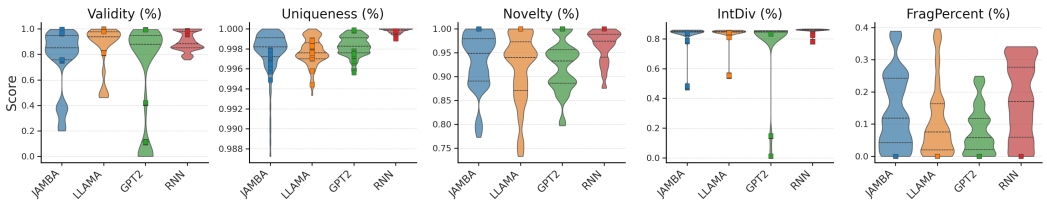

Figure 1: Performance of each architecture across 4 datasets and 6 representations. SMILES-based results (squares) are indicated for reference.

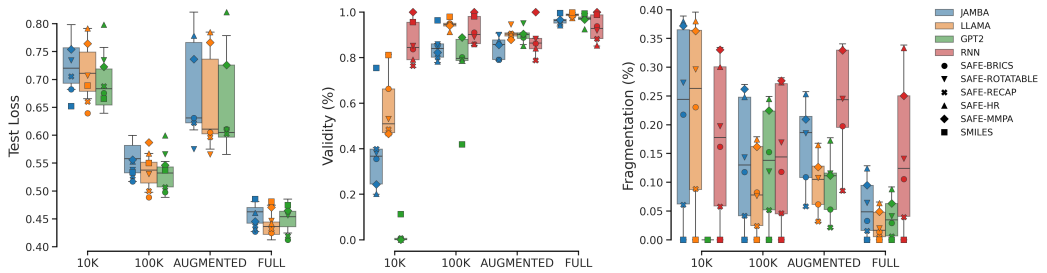

Figure 2: Performance across datasets of different sizes, measured by test loss, validity, and fragmentation percentage. RNNs' test loss results are omitted due to incomparable scale and the use of a different tokenization approach.

4.2 Effect of Dataset Size and Data Augmentation

Figure 2 shows that increasing dataset size generally reduces test loss and improves validity across all models and representations, indicating better generalization and modeling of the data distribution. SAFE models also exhibit a decrease in fragmentation percentage with larger datasets, suggesting an improved understanding of the respective line notation grammar and cross-fragment linking schemes.

Interestingly, RNN models show the least improvement as datasets grow, struggling with SAFE syntax (evidenced by a high fragmentation percentage), despite maintaining high validity. In contrast, GPT-2-based models benefit significantly from increased data, with performance jumping from almost no valid molecules on MOSES-10k to 100% validity on MOSES-Full, alongside gains in internal diversity.

We note, surprisingly, that unlike SAFE-based models, which can generate diverse molecules even in settings with limited training data, SMILES-based models collapse on internal diversity in low-data regimes. For instance, SMILES-GPT-2 shows poor diversity and validity, even on MOSES-100k, compared to their SAFE counterparts (Figure 10).

While SAFE randomization helps maintain average validity and benefits data-hungry models like GPT-2, it does not significantly enhance generalization or improve comprehension of SAFE syntax. Although it is possible that further increasing the percentage of data augmentation could lead to improved performance, results observed on MOSES-Augmented suggest that the primary benefit of SAFE randomization augmentation may lie in extending training time and preventing overfitting. Expanding the amount of novel and diverse structures in the training set might yield similar results while also significantly improving the understanding of SAFE syntax.

Nevertheless, we note that data augmentation via SAFE randomization mitigates the usual trend of reduced novelty as training data size increases (Figure 10). We hypothesize that, akin to SMILES randomization, SAFE randomization helps maintain novelty by exposing the model to alternative molecular configurations during training.

4.3 Effect of Different Fragmentation Algorithms on Performance

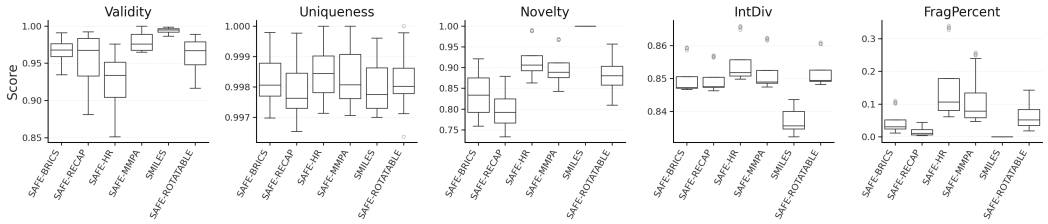

Figure 3: Effect of fragmentation algorithm on generative metrics and on the percentage of fragmented molecules, across all models, on MOSES-Full.

Figure 3 compares the performance of SAFE generative models trained on MOSES-Full across various bond disconnection algorithms, evaluating how each affects their generative capabilities. Indeed, the SAFE grammar can introduce additional challenges compared to SMILES, as models must learn the cross-fragment linking schemes embedded in the grammar.

The results indicate that validity, internal diversity, novelty, and uniqueness are generally high across all bond disconnection methods. However, HR and RECAP models slightly underperform in validity.

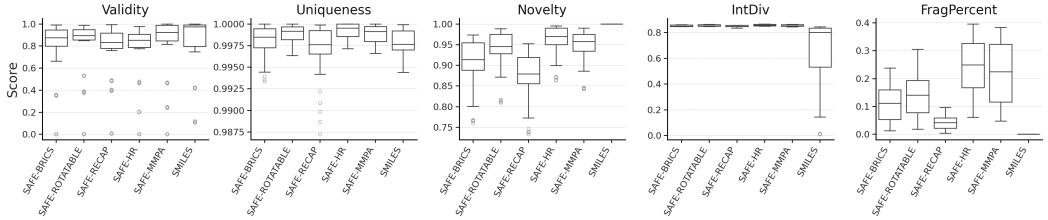

Figure 4: Effect of fragmentation algorithm on performance, across all models, when taking into account all datasets irrespective of the size.

Though all fragmentation algorithms produce diverse and unique molecules, novelty scores are slightly lower than those of SMILES-based models This is likely due to SAFE's more complex grammar and the need to learn specific tokens positions (e.g., '.' and digits) and their associations to maintain

validity, which can, in turn impact generalization. Notably, BRICS and RECAP decompositions result in significantly fewer fragmented molecules compared to HR and MMPA. This is however not because they exploit synthetic accessibility rules but rather, as shown in Figure 7, because BRICS and RECAP bond disconnections produce fewer and larger fragments, leading to simpler grammar for models to learn. Thus, SAFE-BRICS and SAFE-RECAP represent syntaxes that are significantly easier to learn compared to other decompositions, as they require fewer cross-fragment linking tokens, making them excellent compromises for training SAFE-based CLMs. Additionally, on the simple MOSES-Full dataset, these decompositions yielded molecules with the highest QED and SAScore, comparable to or better than those generated by SMILES-based models (see Table 2). These observations are consistent across all datasets, as shown in Figure 4.

4.4 Performance on Fragment-Constrained Molecule Design Tasks

4.4.1 SAFE Outperforms SMILES on Scaffold Decoration

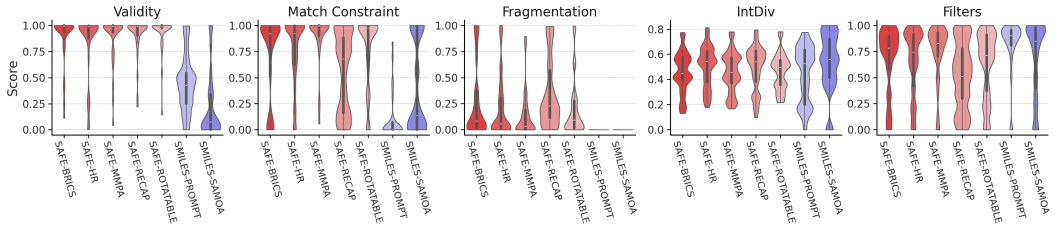

Figure 5: Average performance of SAFE and SMILES sampling algorithms on scaffold decoration tasks (3 benchmarks, 29 scaffolds). SAFE models outperformed SMILES approaches. SMILES-PROMPT refers PromptSMILES, and SMILES-SAMOA to SAMOA.

Using LLaMA, the most stable and robust CLM architecture, we evaluated the performance of various SAFE-based models and two alternative SMILES approaches on standard scaffold decoration benchmarks. We assessed the generated molecules on several criteria: validity, preservation of the input scaffold constraint (`Match Constraint`), fragmentation percentage, internal diversity, and the proportion of molecules passing the medicinal chemistry filters used to curate the MOSES dataset. The latter metric helps gauge how well each algorithm captures implicit rules from the training data distribution when generating fragment-constrained samples. Figure 5 presents the average performance of all scaffold decoration approaches evaluated.

SAFE models consistently outperformed SMILES-based approaches across all metrics. Despite generating fragmented molecules in some cases, SAFE-based methods produced the most valid molecules while preserving the input scaffold, unlike both PromptSMILES and SAMOA. Surprisingly, SAMOA generated the highest number of invalid molecules. Although around 50% of the molecules generated by PromptSMILES were valid, they often failed to respect scaffold constraints. Both SMILES-based algorithms performed particularly poorly on the DRUG dataset (see Figure 10) outright failing on several scaffolds (see Figure 11), suggesting they struggle to generalize to out-of-distribution scaffolds or chemical structures.

Upon closer inspection of molecules generated for the drug Baricitinib under standardized scaffold constraints: `[*]N1CC([*])(n2cc(-c3ncnc4[nH]ccc34)cn2)C1` (see Figure 16), we observed that SAMOA's sampling algorithm rewrote the scaffold by forming new bonds between existing atoms. Due to early stopping during autoregressive generation (either from sampling the `<EOS>` token or reaching the maximum sequence length), SAMOA can produce random and divergent scaffolds lacking significant parts of the input scaffold. While PromptSMILES, which rearranges the scaffold to position the attachment point last, was more robust, it still generated different, though structurally closer, scaffolds.

Among SAFE-based models, RECAP showed the lowest performance on fragmentation percentage and scaffold preservation metrics, additionally struggling to maintain the implicit filtering rules of the training data. Other fragmentation algorithms, especially MMPA, were highly consistent in both standard generative metrics and molecular quality (as measured by SAScore and QED) (see Figure 12). SAFE-HR, meanwhile, produced the most diverse decorations overall.

4.4.2 SAFE Outperforms SMILES on Linker Design

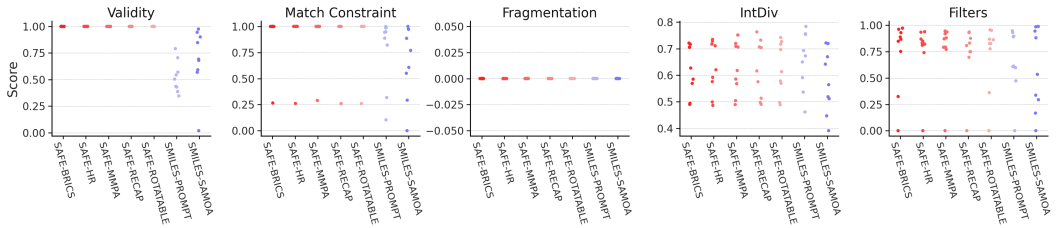

Figure 6: Average performance of SAFE and SMILES algorithms on linker design tasks (9 fragment sets from the DRUG benchmark).

As in the scaffold decoration task, Figure 6 and show that SAFE-based methods outperformed SMILES-based approaches in linker design. SAFE models achieved perfect validity and fragmentation scores, near-perfect substructure constraints preservation, and maintained high QED and SAScore (Figure 14). In contrast, SMILES-based methods often collapsed, failing to maintain the input fragment constraints, as illustrated in Figure 18 and 15. SAFE approaches preserved these constraints, though occasionally sacrificing design diversity, especially on the most challenging inputs.

We hypothesize that the perfect validity of SAFE models can be can be partly attributed to LLaMA's ability to better capture the underlying line notation grammar. This allows for 0-size linking by directly generating ring closures between the two unlinked fragments on difficult samples. Notably, the SAMOA algorithm also exploits this bias, as seen in the linker design examples for Baricitinib (Figure 17).

5 Discussion

Our exploration of SAFE-based generative models demonstrated clear advantages over SMILES-based approaches, specifically on fragment-constrained molecule design tasks. While both approaches performed comparably in pure *de novo* generation, SAFE-based models significantly outperformed SMILES-based methods in scaffold decoration and linker design.

SAFE-LLaMA models, in particular, demonstrated more robust performance across various representations, dataset sizes, and evaluation tasks. We attribute this, in part, to LLaMA's use of Rotary Positional Embedding (RoPE), which likely captures positional dependencies between tokens of different fragments more effectively than GPT-2's absolute positional embeddings. This allowed LLaMA to better learn the nuances of SAFE grammar, resulting in the generation of more valid and connected molecular graphs. We did not find any specific advantage of Jamba over other architectures, but it demonstrated greater robustness on smaller datasets compared to GPT-2. This result likely reflects Jamba's design, which balances model capacity with parameter efficiency, making it less prone to overfitting in small datasets. Further investigation into its performance on datasets with larger, more complex molecules (e.g., polymers) could reveal potential strengths in those contexts.

We found that dataset size played a crucial role, with larger datasets consistently improving performance, especially for transformer models, and most notably for GPT-2. This finding emphasizes the importance of diverse and extensive training data for capturing a broader chemical space, to enable models to generalize more effectively in both *de novo* and fragment-constrained tasks. While SAFE-based data augmentation did not significantly enhance generalization or sampling quality, it did help mitigate the loss of novelty in generated samples as dataset size increased, primarily by preventing overfitting. Investigating the potential synergy between SMILES and SAFE randomization may provide additional insights in future work.

Finally, the choice of fragmentation algorithm also had a significant impact on performance. SAFE-BRICS and SAFE-RECAP, which generate fewer and larger fragments using synthetic accessibility decomposition rules, were easier for models to learn, leading to lower fragmentation rates and improved synthetic accessibility of generated molecules. However, this simplicity came at a cost: models trained on decompositions with more fragments, like HR and MMPA, demonstrated higher

novelty and uniqueness although at the expense of lower validity and increased fragmentation in the generated molecules. Interestingly, SAFE-RECAP performed the worst in scaffold decoration tasks, producing the most fragmented and least diverse molecules (see Figure 11). This may be due to RECAP's rigid bond disconnection rules, limiting bond formation possibilities in fragment-constrained sampling when trained on a dataset like MOSES. By contrast, BRICS which expands the bond disconnection criteria used by RECAP from 11 to 16 by taking into account additional chemical environment surrounding bonds, has led to improved scores. Overall, our results suggest a balance between fragmentation complexity and generalization: fewer fragments simplify grammar but limit exploration of chemical space, especially in fragment-constrained design tasks. The performance of SAFE-ROTATABLE, which serves as a middle ground, supported this trade-off. Nevertheless, considering auxiliary objectives like synthetic accessibility and drug-likeness, we recommend SAFE-BRICS as the most effective representation.

6 Conclusion and Future Works

While this study focused on finding the best training setup for SAFE generative models, several promising avenues for future research are apparent. One interesting direction is ensuring that the distribution of randomized SAFE strings preserves a consistent probabilistic density within the generative model's learned distribution. This could enhance the model's ability to effectively capture the underlying chemical space. Another important challenge is improving stereochemistry handling, which remains an issue for both SMILES and SAFE representations. Addressing these challenges could pave the way for using SAFE strings in representation learning, resulting in new opportunities for predictive and unsupervised tasks in molecular design.

Additionally, a deeper exploration of the robustness of SAFE-based models in optimization, particularly in low-data settings where diversity and novelty are crucial, is needed. While prior work used Proximal Policy Optimization (PPO) for goal-directed optimization of a SAFE-GPT model, other optimization algorithms should be systematically evaluated and benchmarked. Furthermore, while this study did not examine the effects of scaling model parameters, leveraging the novel insights gained here, it would be worthwhile to explore scaling SAFE models trained on large, diverse datasets spanning a wide range of molecular structures, properties, and sizes. This could lead to foundational models with significant utility for generative molecular design.

In conclusion, SAFE-based generative CLMs offer a stable, versatile, and powerful alternative to SMILES for molecular design, with considerable potential to advance applications in drug discovery and material design. We believe future research should focus on experimentally validating these models in practical settings to assess their effectiveness and real-world impact.

References

[1] Josep Arús-Pous, Simon Viet Johansson, Oleksii Prykhodko, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Randomized smiles strings improve the quality of molecular generative models. *Journal of cheminformatics*, 11:1–13, 2019.

[2] 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*, 12:1–18, 2020.

[3] Esben Jannik Bjerrum. Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*, 2017.

[4] Andrew Dalke, Jerome Hert, and Christian Kramer. mmpdb: An open-source matched molecular pair platform for large multiproperty data sets. *Journal of chemical information and modeling*, 58(5):902–910, 2018.

[5] Jorg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3(10):1503, 2008.

[6] Alberga Domenico, Gambacorta Nicola, Trisciuzzi Daniela, Ciriaco Fulvio, Amoroso Nicola, and Nicolotti Orazio. De novo drug design of targeted chemical libraries based on artificial

intelligence and pair-based multiobjective optimization. *Journal of Chemical Information and Modeling*, 60(10):4582–4593, 2020.

[7] 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.

[8] 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.

[9] Francesca Grisoni. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology*, 79:102527, 2023.

[10] Jeff Guo, Franziska Knuth, Christian Margreitter, Jon Paul Janet, Kostas Papadopoulos, Ola Engkvist, and Atanas Patronov. Link-invent: generative linker design with reinforcement learning. *Digital Discovery*, 2(2):392–408, 2023.

[11] Julien Horwood and Emmanuel Noutahi. Molecular design in synthetically accessible chemical space via deep reinforcement learning. *ACS omega*, 5(51):32984–32994, 2020.

[12] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Proceedings of Neural Information Processing Systems, NeurIPS Datasets and Benchmarks*, 2021.

[13] 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.

[14] 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.

[15] Maxime Langevin, Hervé Minoux, Maximilien Levesque, and Marc Bianciotto. Scaffold-constrained molecular generation. *Journal of Chemical Information and Modeling*, 60(12): 5637–5646, 2020.

[16] Xiao Qing Lewell, Duncan B Judd, Stephen P Watson, and Michael M Hann. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of chemical information and computer sciences*, 38(3):511–522, 1998.

[17] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

[18] Daniel Merk, Lukas Friedrich, Francesca Grisoni, and Gisbert Schneider. De novo design of bioactive small molecules by artificial intelligence. *Molecular informatics*, 37(1-2):1700153, 2018.

[19] Michael Moret, Lukas Friedrich, Francesca Grisoni, Daniel Merk, and Gisbert Schneider. Generative molecular design in low data regimes. *Nature Machine Intelligence*, 2(3):171–180, 2020.

[20] Michael Moret, Irene Pachon Angona, Leandro Cotos, Shen Yan, Kenneth Atz, Cyrill Brunner, Martin Baumgartner, Francesca Grisoni, and Gisbert Schneider. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications*, 14(1):114, 2023.

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

[22] Rıza Özçelik, Sarah de Ruiter, Emanuele Criscuolo, and Francesca Grisoni. Chemical language modeling with structured state space sequence models. *Nature Communications*, 15(1):6176, 2024.

[23] Hakime Öztürk, Arzucan Özgür, Philippe Schwaller, Teodoro Laino, and Elif Ozkirimli. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discovery Today*, 25(4):689–705, 2020.

[24] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020.

[25] Christin Scharfer, Tanja Schulz-Gasch, Hans-Christian Ehrlich, Wolfgang Guba, Matthias Rarey, and Martin Stahl. Torsion angle preferences in druglike chemical space: a comprehensive guide. *Journal of Medicinal Chemistry*, 56(5):2016–2028, 2013.

[26] Michael A Skinnider. Invalid smiles are beneficial rather than detrimental to chemical language models. *Nature Machine Intelligence*, 6(4):437–448, 2024.

[27] Michael A Skinnider, R Greg Stacey, David S Wishart, and Leonard J Foster. Chemical language models enable navigation in sparsely populated chemical space. *Nature Machine Intelligence*, 3 (9):759–770, 2021.

[28] Morgan Thomas, Mazen Ahmad, Gary Tresadern, and Gianni de Fabritiis. Promptsmiles: prompting for scaffold decoration and fragment linking in chemical language models. *Journal of Cheminformatics*, 16(1):77, 2024.

[29] Morgan Thomas, Noel M O'Boyle, Andreas Bender, and Chris De Graaf. Molscore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. *Journal of Cheminformatics*, 16(1):64, 2024.

[30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[31] 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.

[32] Yuyao Yang, Shuangjia Zheng, Shimin Su, Chao Zhao, Jun Xu, and Hongming Chen. Syntalinker: automatic fragment linking with deep conditional transformer neural networks. *Chemical science*, 11(31):8312–8322, 2020.

A Appendix

This appendix provides additional details regarding the dataset, model architectures, training configurations, and experiment results, supplementing the findings discussed in the main paper.

A.1 Dataset

Figure 7 shows the distribution of the number of fragments generated by the different bond disconnection algorithms used to create the SAFE versions of the MOSES dataset. Both RECAP and BRICS algorithms resulted in molecules with fewer fragments (average of 5 fragments), while the Hussain-Rea (HR) and MMPA decompositions produced the highest number of fragments on average (9 and 8 fragments, respectively). The distribution of the number of fragments generated by these algorithms plays an essential role in the complexity of learning SAFE representations.

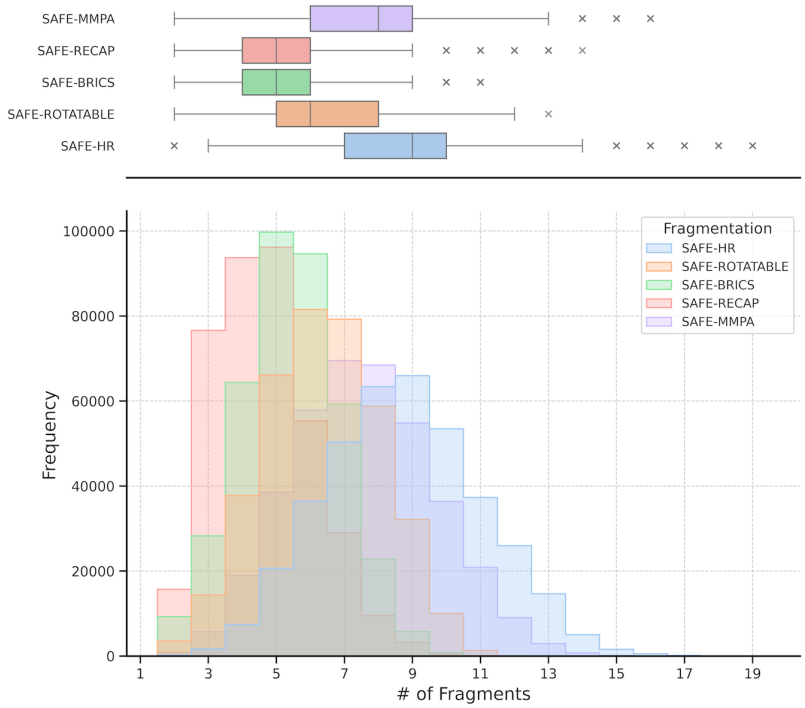

Figure 7: Distribution of the number of fragments for each bond disconnection algorithm.

A.2 Model Architecture and Training

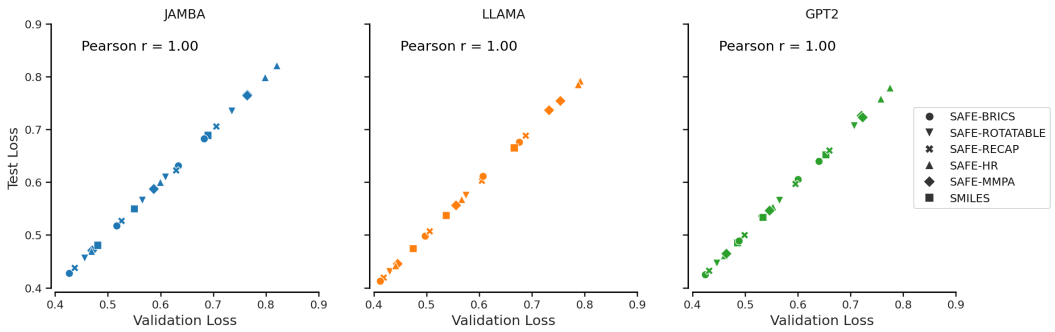

Figure 8: Validation vs Test Loss Across Different Transformer Architectures

For each architecture, we performed a grid search on key hyperparameters such as learning rate, batch size, gradient norm clipping, and model configurations to ensure comparable sizes across all transformer models. We used the codebase provided by Noutahi et al. [21] [1] for training the transformer models, and for RNN (GRU) models, we followed the setup proposed by Thomas et al. [28] using the SMILES-RNN codebase [2].

The models were trained on the various MOSES datasets, and for all transformer models, we used a batch size of 128 and trained them for up to 100,000 steps (5,000 steps on MOSES-10k). Optimization was performed using the AdamW optimizer with a cosine learning-rate scheduler and a 10% warmup. RNN models were trained using an exponential learning-rate decay, with early stopping based on validation loss. We observed overfitting with large RNN models, so their size was kept smaller than that of the other models. Each training run was executed on a single NVIDIA H100/80Gb GPU, utilizing approximately 20% of its capacity. On average, model training took 10 hours.

The token vocabulary was fixed for each representation/fragmentation algorithm, with size ranging from 37 (SMILES) to 53 (SAFE-BRICS) tokens for transformers, and from 26 (SMILES) to 38 (MMPA) for RNNs (the SMILES-RNN codebase use a different tokenization scheme).

Below are the final configurations for each model:

GPT-2 ($\sim$ 10M parameters): 3 layers, 512 hidden size, 8 attention heads, 1024 max position embeddings, vocab size ranging from 37 to 53, GELU activation function, and 0.1 dropout.

LLaMA ($\sim$ 9M parameters): 4 hidden layers, 512 hidden size, 512 intermediate size, 8 attention heads, 1024 max position embeddings, 37 to 53 vocab size, SiLU activation function, 0.1 dropout, and RoPE theta of 500,000.0.

Jamba ($\sim$ 9,4M parameters): 2 layers, 480 hidden size, 512 intermediate size, 8 attention heads, 1024 max position embeddings, 37 to 53 vocab size, SiLU activation function, 0.1 dropout, with Mamba parameters (D-Conv: 4, D-State: 16).

GRU ($\sim$ 4.3M parameters): 3 layers, 512 hidden size, 256 embedding layer size, 0.0 dropout, trained with a batch size of 125, learning rate of 0.001, and 3-10 epochs depending on dataset size.

Figure 8 shows the relationship between validation loss and test loss across the three transformer architectures (Jamba, LLaMA, and GPT-2). Each marker represents a unique configuration using either SMILES or one of the SAFE fragmentations (BRICS, ROTATABLE, RECAP, HR, MMPA). The strong linear correlation (Pearson's r = 1.00) indicates that validation loss reliably predicts test loss, suggesting that the models generalize well from validation to test sets.

A.3 Experiments

A.4 De novo generation

We evaluated the performance of the generative models on *de novo* molecule generation across the MOSES-Full dataset. Table 2 presents a comparison of key molecular properties, including QED, SAScore, cLogP, and Molecular Weight, across 50,000 sampled molecules (5 seeds, 10,000 samples per seed). All architectures and representations mostly performed similarly on these metrics. As expected, SMILES and SAFE representation using fragmentation algorithms that leverage synthetic accessibility rules (SAFE-BRICS, SAFE-RECAP) produced molecules that have marginally better SAScore in average.

A.4.1 Impact of Data Augmentation

Figures 9 and 10 illustrate the impact of dataset size and data augmentation on model performance. These results show that as dataset size increases, model performance usually improve. However, data augmentation does not provide equal benefits for all models. While augmentation helps maintain performance in low-data regimes by exposing models to diverse molecular configurations, its advantages vary depending on the model architecture.

[1] `https://github.com/datamol-io/safe/`

[2] `https://github.com/MorganCThomas/SMILES-RNN/`

Table 2: Comparison of the sampled molecules (10k samples, 5 replicates) for each generative CLM performance trained on MOSES-Full dataset on QED, SAScore, cLogP, and Molecular Weight. The best performance within each representation is highlighted in gray, and the overall best performance across all representations is in red. All architecture and representation performed similarly.

Representation	Model	QED ↑	SAScore ↓	cLogP	Mol. Weight
SMILES	Jamba	0.816 ± 0.090	2.357 ± 0.449	2.505 ± 0.915	307.675 ± 27.445
	LLaMA	0.815 ± 0.090	2.368 ± 0.448	2.493 ± 0.908	308.132 ± 27.110
	GPT-2	0.819 ± 0.088	2.340 ± 0.446	2.483 ± 0.914	306.920 ± 27.339
	RNN	0.807 ± 0.096	2.451 ± 0.467	2.451 ± 0.959	306.517 ± 28.904
SAFE-BRICS	Jamba	0.815 ± 0.093	2.363 ± 0.494	2.572 ± 0.888	304.814 ± 27.848
	LLaMA	0.815 ± 0.092	2.357 ± 0.472	2.572 ± 0.881	305.547 ± 27.510
	GPT-2	0.808 ± 0.097	2.386 ± 0.486	2.642 ± 0.915	311.903 ± 28.573
	RNN	0.802 ± 0.097	2.485 ± 0.581	2.469 ± 0.988	307.134 ± 29.936
SAFE-RECAP	Jamba	0.818 ± 0.092	2.382 ± 0.484	2.545 ± 0.907	305.297 ± 28.945
	LLaMA	0.817 ± 0.092	2.379 ± 0.466	2.544 ± 0.889	306.980 ± 27.653
	GPT-2	0.811 ± 0.095	2.380 ± 0.463	2.565 ± 0.907	312.147 ± 29.510
	RNN	0.801 ± 0.100	2.508 ± 0.518	2.474 ± 0.977	307.473 ± 30.869
SAFE-HR	Jamba	0.801 ± 0.100	2.489 ± 0.612	2.610 ± 0.911	304.228 ± 28.523
	LLaMA	0.807 ± 0.096	2.406 ± 0.545	2.591 ± 0.907	304.863 ± 28.160
	GPT-2	0.803 ± 0.099	2.410 ± 0.549	2.588 ± 0.907	307.660 ± 28.931
	RNN	0.768 ± 0.116	2.980 ± 0.738	2.484 ± 0.986	307.896 ± 29.196
SAFE-MMPA	Jamba	0.809 ± 0.095	2.428 ± 0.583	2.617 ± 0.896	305.708 ± 28.510
	LLaMA	0.811 ± 0.093	2.373 ± 0.528	2.616 ± 0.898	305.985 ± 28.033
	GPT-2	0.805 ± 0.097	2.387 ± 0.542	2.599 ± 0.907	310.175 ± 29.361
	RNN	0.789 ± 0.106	2.794 ± 0.696	2.463 ± 0.977	308.580 ± 30.096
SAFE-ROTATABLE	Jamba	0.812 ± 0.095	2.406 ± 0.537	2.639 ± 0.913	304.755 ± 28.853
	LLaMA	0.815 ± 0.093	2.362 ± 0.496	2.608 ± 0.903	305.078 ± 27.984
	GPT-2	0.810 ± 0.096	2.404 ± 0.511	2.626 ± 0.927	310.809 ± 28.912
	RNN	0.796 ± 0.102	2.690 ± 0.624	2.494 ± 0.994	307.297 ± 30.027

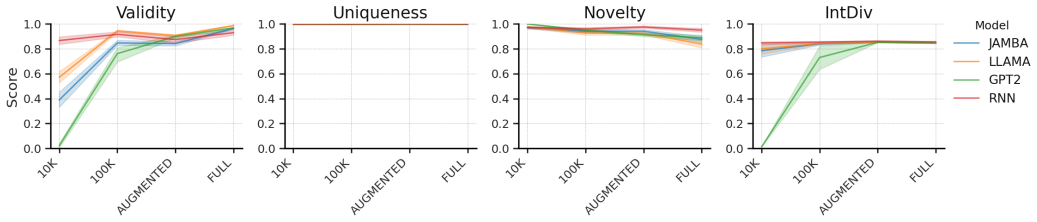

Figure 9: Effect of training data size on architecture performance, aggregated across representations.

A.5 Fragment-Constrained Generation

In this section, we evaluate the performance of various fragment-constrained generation methods. We focus on both scaffold decoration and linker design. The figures provide an in-depth look into how each model performs in preserving structural constraints while generating valid and chemically diverse molecules.

Figure 11 shows the per-dataset performance of each design method in scaffold decoration. Here, SAFE-based models outperform their SMILES counterparts, particularly in maintaining scaffold constraints. Figure 12 visualizes the performance of each method on standard generative metrics, confirming the robustness of SAFE models. In Figure 14, we compare the performance of different fragment-constrained methods on the linker design task. SAFE-based methods achieve near-perfect validity and substructure preservation, outperforming SMILES-based approaches. The detailed breakdown in Figures 13 and 15 further highlights the strengths and limitations of each method.

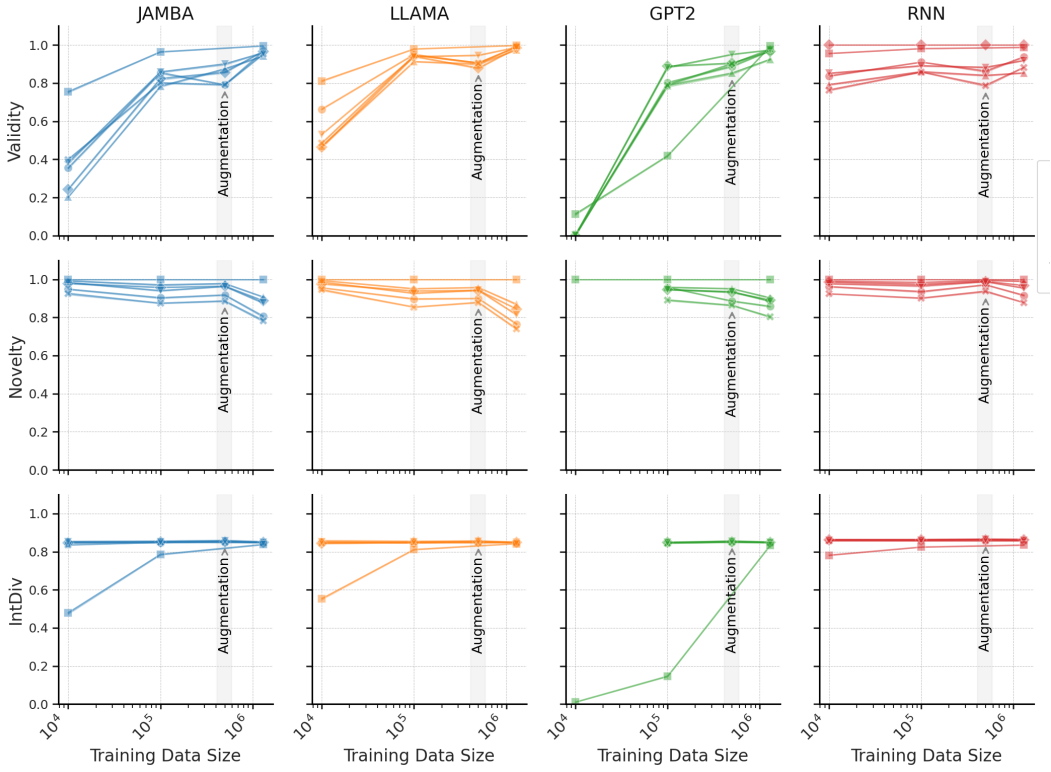

Figure 10: Performance of each architecture across different data sizes and representations.

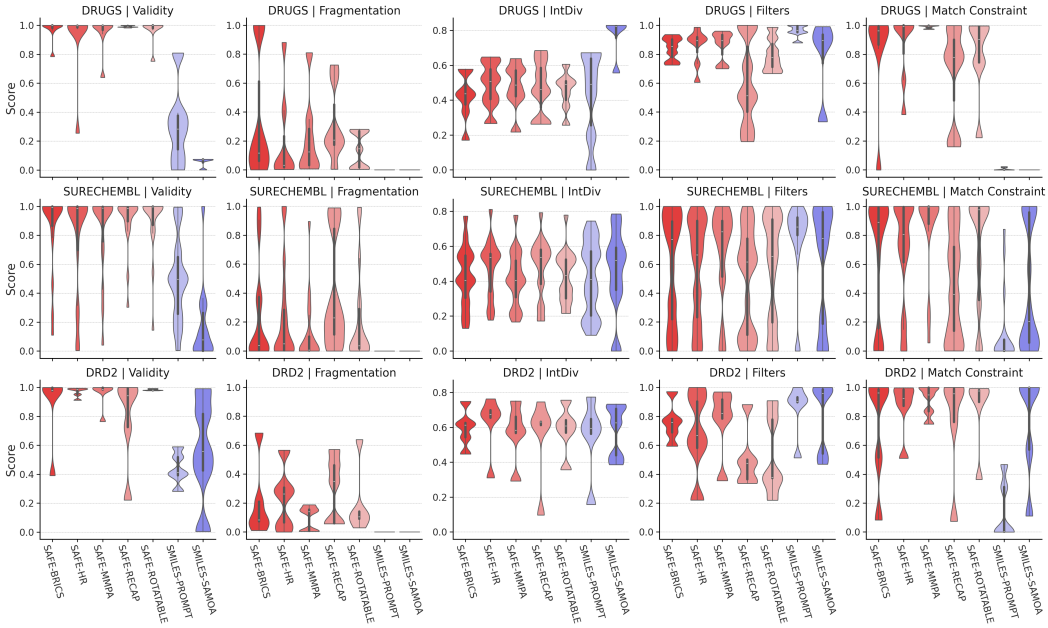

Figure 11: Per dataset performance of each fragment-constrained design method on scaffold decoration.

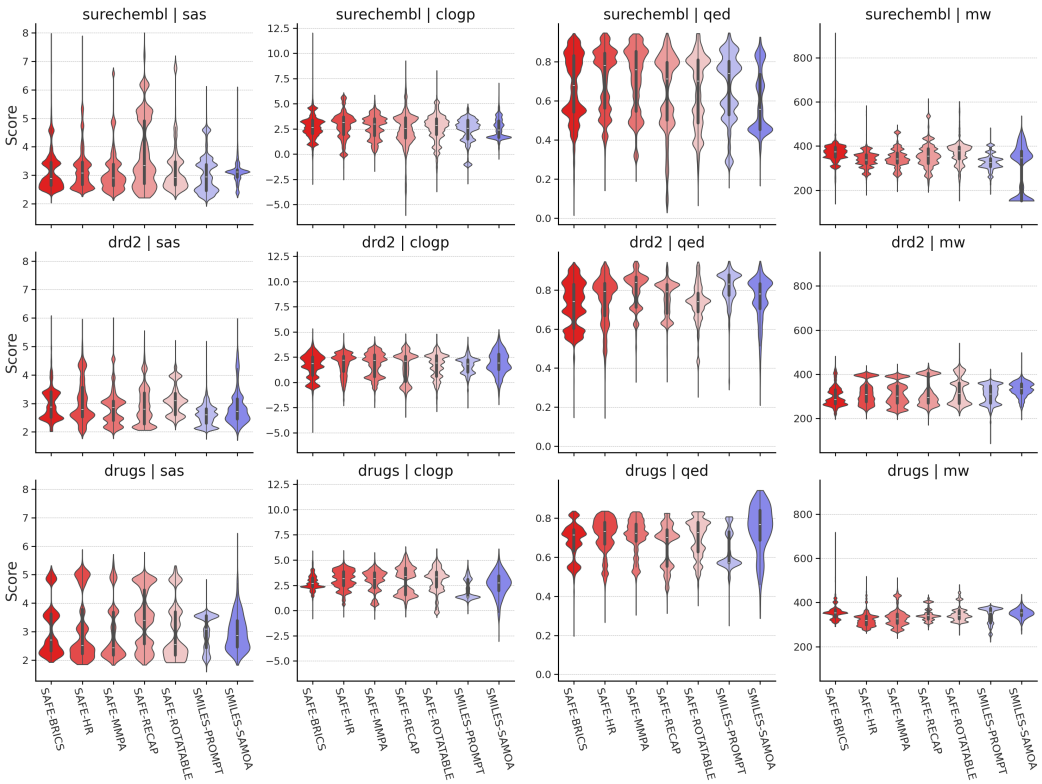

Figure 12: Per dataset standard molecule quality metrics distribution for each fragment-constrained design method on scaffold decoration.

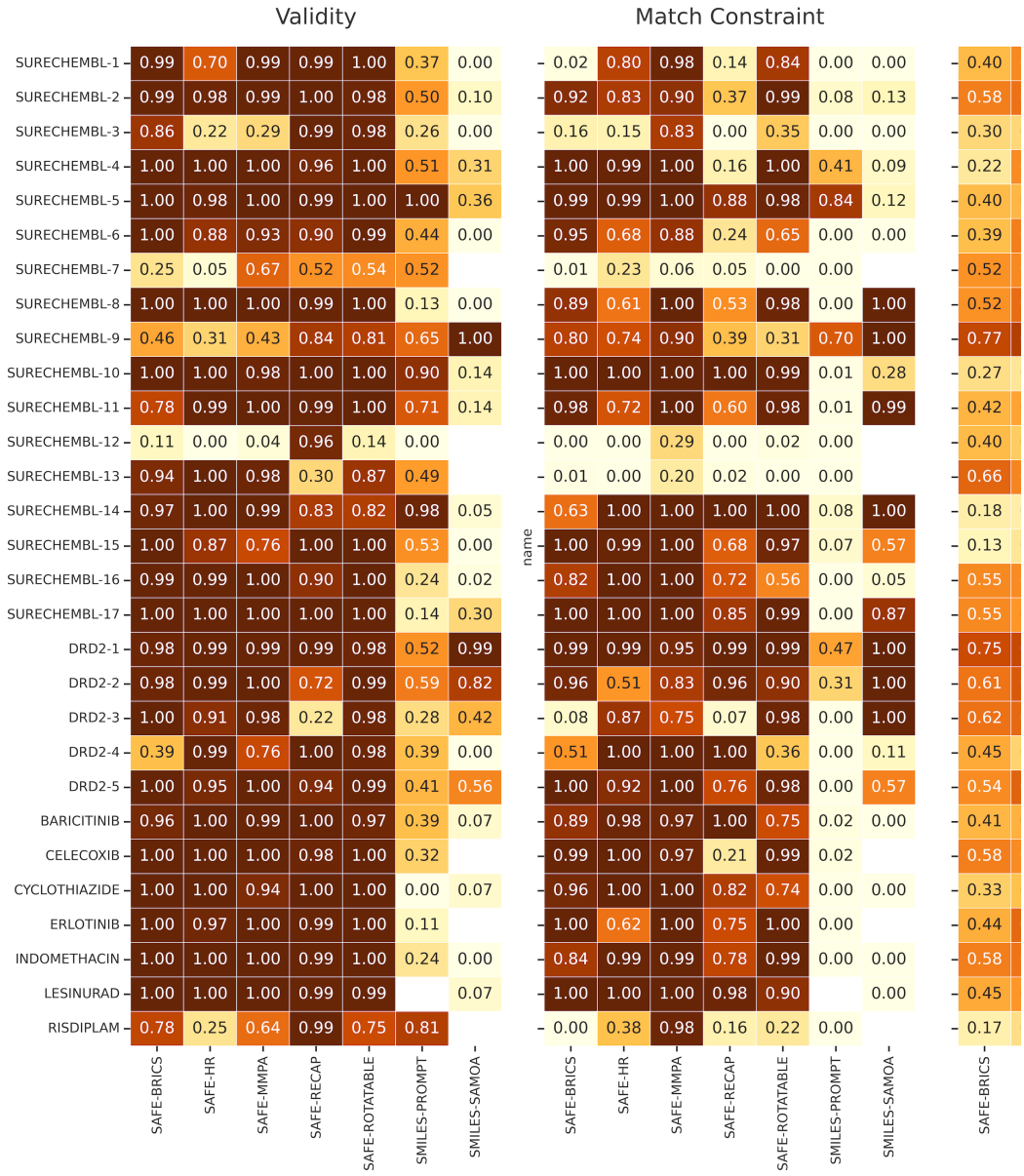

Figure 13: Per-scaffold performance of each fragment-constrained design method on the 3 different benchmarks considered for scaffold decoration.

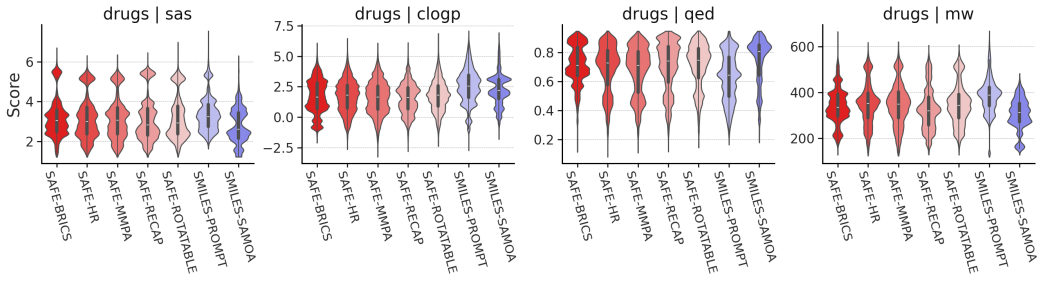

Figure 14: Performance on standard molecule quality metrics for each fragment-constrained design method on the DRUG linker design benchmark.

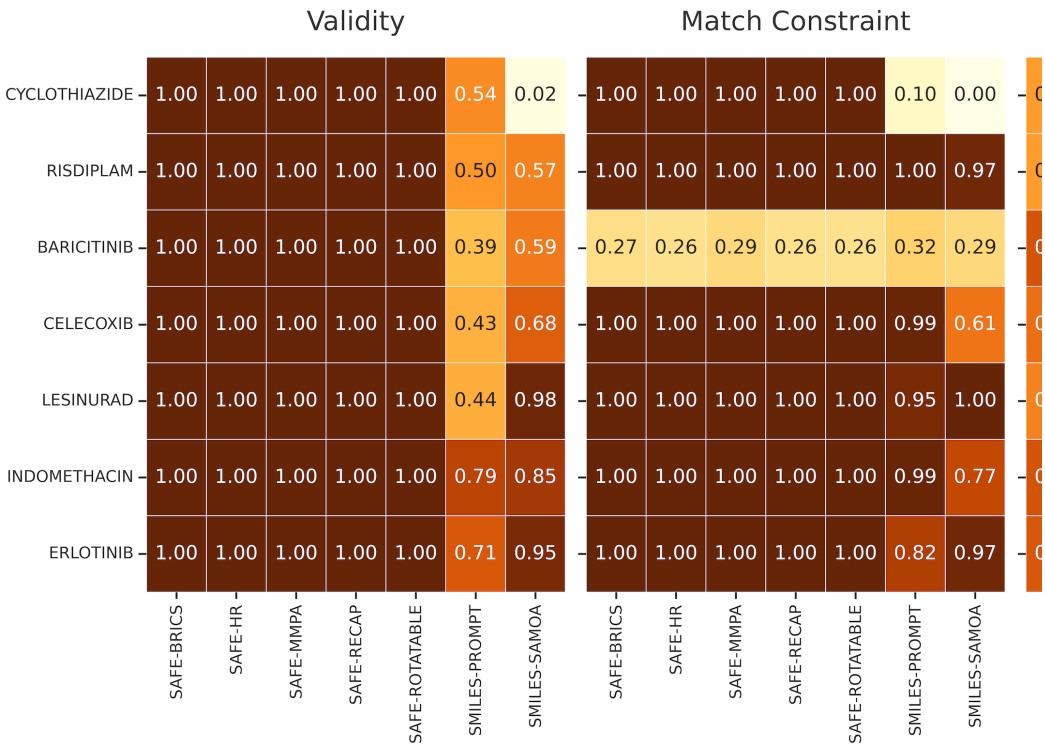

Figure 15: Per-drug performance of each fragment-constrained design method on the DRUG linker design benchmark.

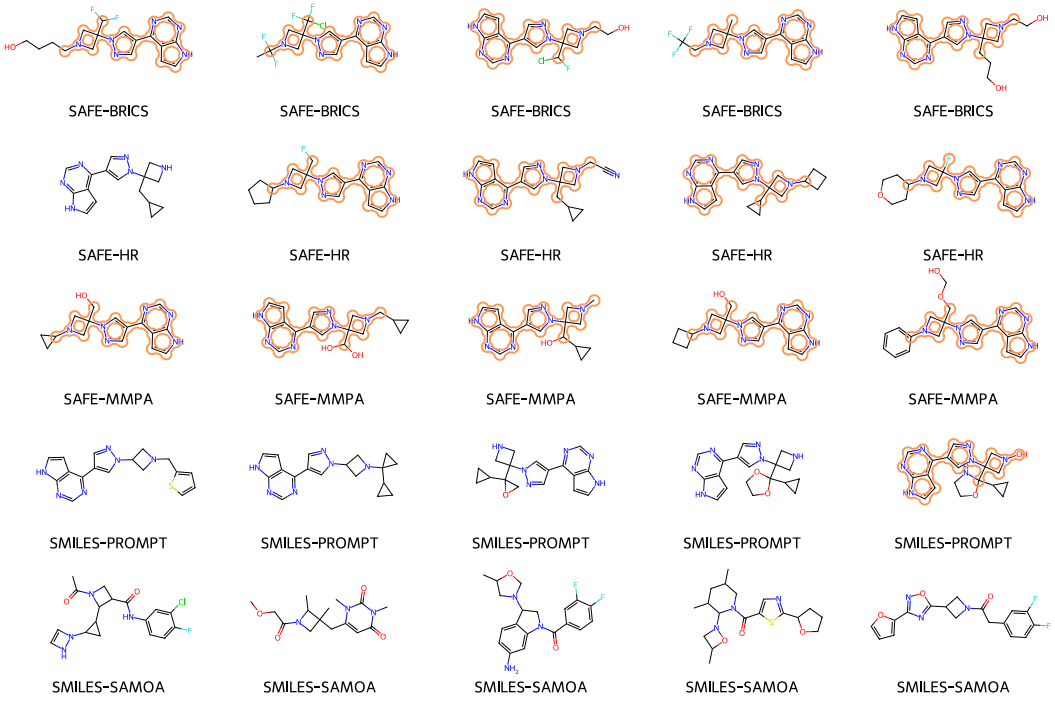

Figure 16: Example of sampled molecules under the fragment-constrained design of novel molecules sharing the same scaffold as the drug Baricitinib `:[*:1]N1CC([*:2])(n2cc(-c3ncnc4[nH]ccc34)cn2)C1`. The core being decorated is highlighted in the figure.

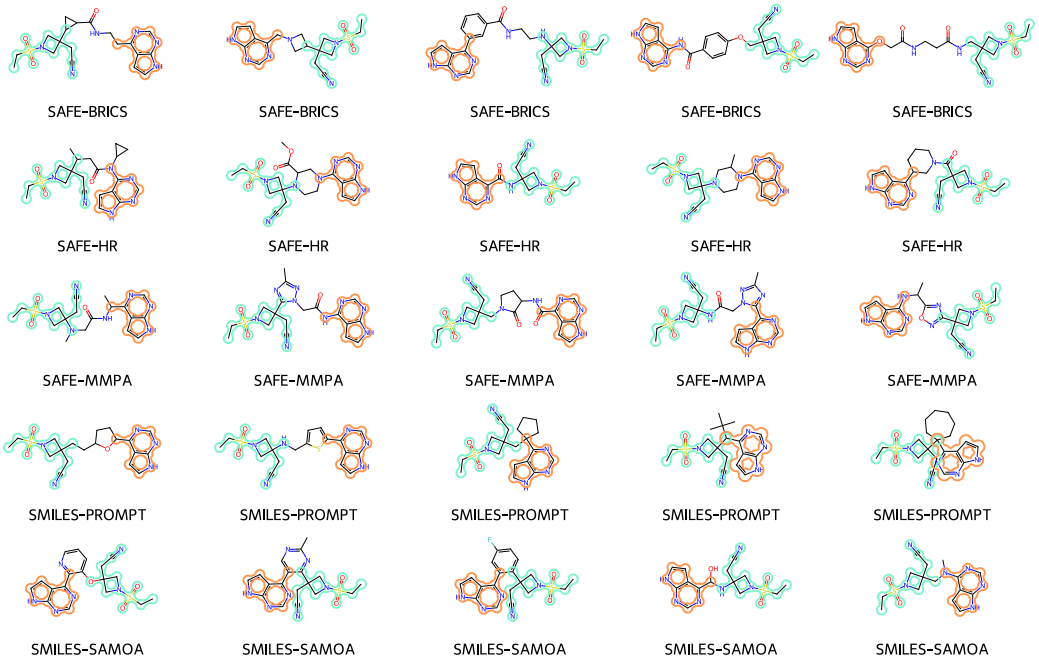

Figure 17: Example of sampled molecules under the fragment-constrained design of novel molecules by linking fragments of the drug Baricitinib: `[*:1]C1(CC#N)CN(S(=O)(=O)CC)C1` and `[*]c1ncnc2[nH]ccc12`. The two fragments linked are highlighted in the figure.

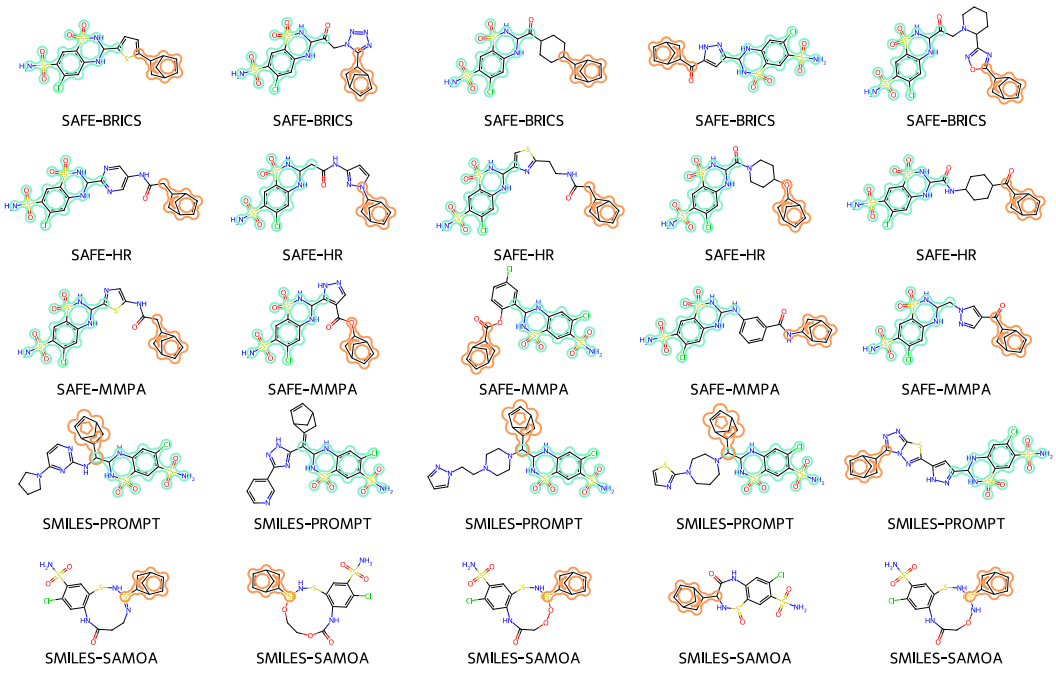

Figure 18: Example of sampled molecules under the fragment-constrained design of novel molecules by linking fragments of drug Cyclothiazide: `[*]C1CC2C=CC1C2`, `[*]C1Nc2cc(Cl)c(S(N)(=O)=O)cc2S(=O)(=O)N1`. The two fragments linked are highlighted in the figure.