Molecular Generation with State Space Sequence Models

Anri Lombard[†], **Shane Acton**[‡], **Ulrich Mbou Sob**[‡], **Jan Buys**[†]
[†] Department of Computer Science, University of Cape Town, South Africa
[‡] InstaDeep
LMBANR001@myuct.ac.za, {s.acton,u.mbousob}@instadeep.com, jbuys@cs.uct.ac.za

Abstract

Molecular generation is a critical task in drug discovery but current approaches often struggle with efficiency and scalability when dealing with complex molecular structures. This paper aims to address these challenges by training and evaluating models for molecular generation using the MAMBA State Space Model architecture. We develop models with 20M and 90M parameters trained on the MOSES and ZINC datasets, respectively, using the Sequential Attachment-based Fragment Embedding (SAFE) representation. We compare MAMBA models against the prevailing Transformer architecture in terms of generation quality and computational efficiency. Our findings suggest that MAMBA models can achieve performance comparable to Transformers in generating valid, unique, and diverse molecules. Generation from both architectures can achieve close to perfect validity and uniqueness scores, although MAMBA models require more conservative sampling parameters or regeneration steps to achieve these results. MAMBA models consistently demonstrates lower perplexity and reduced GPU power consumption (up to 30% reduction) compared to Transformer models. These results indicate that State Space Models may offer a computationally efficient alternative for molecular generation tasks, potentially enabling more efficient processing of larger datasets and complex molecular structures. The efficiency gains of MAMBA models become more pronounced with longer sequences, suggesting that this architecture could enable the modeling and generation of more complex molecules. This capability could significantly expand the scope of AI-driven molecular design in drug discovery.

1 Introduction

The application of artificial intelligence (AI) to molecular design and drug discovery has emerged as a promising approach to accelerate the identification of novel therapeutic compounds [1]. This intersection of AI and chemistry builds upon the remarkable success of sequence modeling techniques in natural language processing (NLP), where models have demonstrated an unprecedented ability to understand and generate human-like text [2]. The parallels between language and molecular structures have inspired researchers to adapt and apply these sequence modeling techniques to the complex task of molecular generation.

Advancements in deep learning architectures, particularly the Transformer model [3], have shown promise in molecule generation compared to older generative approaches [4]. The Transformer's attention mechanism, which allows the model to weigh the importance of different parts of the input sequence dynamically, has proven especially effective in capturing long-range dependencies in both text and molecular structures. However, the quadratic computational complexity of Transformers with respect to sequence length poses challenges for scaling to larger datasets or more complex molecules.

AI for New Drug Modalities at NeurIPS 2024.

This limitation has motivated research into alternative architectures such as State Space Models (SSMs), which offer linear time complexity [5]. SSMs, inspired by control theory and dynamical systems, provide a different approach to capturing sequential dependencies. The MAMBA architecture, a recent innovation in SSMs, has shown promising results in language modeling tasks.

The main contribution of this paper is to evaluate models based on the MAMBA architecture on molecular generation. Our study addresses two critical questions: First, how do State Space Models compare to Transformer-based architectures in generating valid, unique, and diverse molecules using the SAFE representation? Second, can the efficiency of the MAMBA architecture provide advantages in terms of computational resources and training time when applied to larger datasets and model sizes in molecular generation tasks?

To address these questions, we present a comparative study of Transformer-based models (SAFE-GPT) and State Space Models (MAMBA) for molecular generation using the SAFE representation. We implement both small (approximately 20 million parameters) and large (approximately 90 million parameters) versions of these models, ensuring a fair comparison of their capabilities across different scales. The paper makes several key contributions:

1. We provide a comprehensive comparison of Transformer and MAMBA architectures for molecular generation using the SAFE representation across different model sizes.
2. We evaluate the potential of State Space Models as an alternative to Transformers for capturing complex structural information in molecular generation tasks.
3. We assess the computational efficiency advantages of MAMBA-based models, exploring their potential for processing larger molecular datasets and more complex structures.
4. We offer insights into the trade-offs between model architecture, performance, and computational resources, informing future research directions in AI-driven molecular design.

By bridging the gap between cutting-edge sequence modeling techniques and molecular generation, our work contributes to the ongoing efforts to accelerate drug discovery through AI-driven approaches. The results of this study indicate that transitioning molecular generation model architectures from Transformers to MAMBA presents a viable strategy for scaling to larger models or datasets. Furthermore, this shift potentially allows for the modeling of more complex molecules, opening new avenues for exploring previously intractable chemical spaces and accelerating the discovery of novel therapeutic compounds.

2 Background

2.1 Molecular Representations

The choice of molecular representation is crucial for the effectiveness of machine learning models in computational chemistry and drug discovery. The evolution of these representations reflects ongoing efforts to balance chemical validity, informativeness, and computational efficiency. The Simplified Molecular-Input Line-Entry System (SMILES) [6], while widely used, has limitations in maintaining validity during generation tasks [7]. To address these issues, the Self-Referencing Embedded Strings (SELFIES) representation was introduced [7], using a robust grammar to ensure generated strings correspond to valid molecular graphs. Building upon these advancements, the Sequential Attachment-based Fragment Embedding (SAFE) representation [8] combines fragment-based approaches with sequential string representations. SAFE represents molecules as an unordered sequence of interconnected fragment blocks, maintaining high validity rates in generative tasks while capturing meaningful chemical substructures. By fragmenting molecules using methods like BRICS [9] and employing a compact encoding scheme, SAFE offers a balance between chemical relevance and computational efficiency. This approach has shown promise in improving the performance of molecular generation models.

2.2 Molecular Sequence Modeling

Unlike biological sequences such as DNA, RNA, and proteins that have natural sequential representations, modeling arbitrary molecules as sequences requires careful consideration of mapping between molecular graphs and sequential tokens. Early approaches to molecular sequence modeling

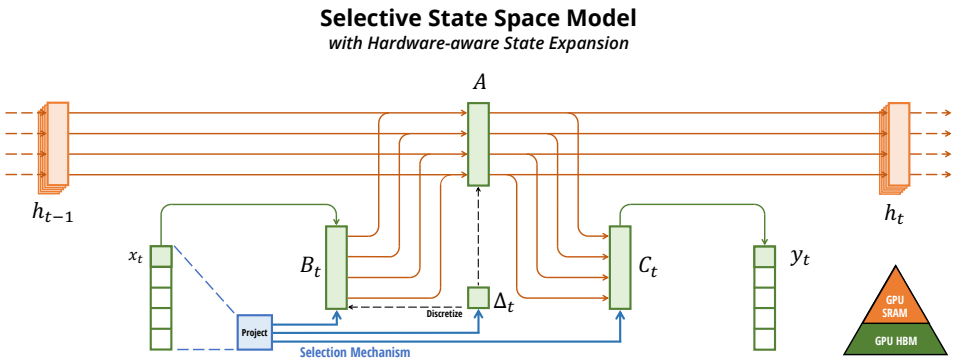


Figure 1: Schematic representation of Mamba's Selective State Space Model architecture, illustrating the selection mechanism, which dynamically focuses on relevant input information, and the hardware-aware state expansion design [5].

adapted methods from natural language processing, starting with n-gram models and recurrent neural networks (RNNs) [10, 11]. These models demonstrated initial success but struggled with maintaining chemical validity and capturing long-range dependencies in complex molecular structures, particularly for molecules with nested branching patterns or multiple cycles. The development of masked language modeling and auto regressive training objectives, combined with attention-based architectures enabled more robust molecular generation [12, 13]. These models showed particular strength in capturing non-local interactions between functional groups and modeling complex stereo chemical relationships.

Recent architectural innovations have focused on improving model efficiency while maintaining chemical validity. This includes specialized architectures that incorporate chemical knowledge through modified attention patterns [4], and the adaption of pre-training strategies from language models to molecular domains [14], enabling models to learn general chemical patterns from large unlabeled datasets before fine-tuning on specific tasks. These advances have progressively improved the ability to model longer sequences and more complex molecular structures, though computational efficiency remains a challenge, particularly for modelling large libraries of drug-like compounds.

2.3 State Space Models and MAMBA

State Space Models (SSMs) offer an alternative approach to sequence modeling by representing sequences as continuous-time dynamical systems. The general form of a discrete-time linear SSM is:

$$x_{k+1} = Ax_k + Bu_k, \tag{1}$$

$$y_k = Cx_k + Du_k, \tag{2}$$

where x_k is the hidden state, u_k is the input, y_k is the output, and A, B, C, and D are learnable parameters.

The MAMBA architecture [5] is a selective state space model, which addresses key limitations of previous SSMs through its selection mechanism, which allows the model to dynamically focus on or ignore specific inputs based on their content. This mechanism is implemented by making several SSM parameters, namely Δ, B, and C, functions of the input:

$$B : (B, L, N) \leftarrow s_B(x) \tag{3}$$

$$C : (B, L, N) \leftarrow s_C(x) \tag{4}$$

$$\Delta : (B, L, D) \leftarrow \tau_\Delta(\text{Parameter} + s_\Delta(x)) \tag{5}$$

Here, $s_B(x)$, $s_C(x)$, and $s_\Delta(x)$ are learnable functions that transform the input x, allowing the model to adapt its behavior based on the content of the sequence. The function τ_Δ is typically chosen to be the softplus function, which ensures that Δ remains positive.

Figure 1 illustrates the key components of the Mamba architecture, namely its selection mechanism and hardware-aware state expansion. The input x is first discretized and then processed through the

selective state space model. The model uses learnable parameters A, B_t, and C_t to update the hidden state h_t and produce the output y_t. The time step size Δ_t is also a learnt parameter and can vary based on the input. This allows the model to adapt its temporal resolution, effectively zooming in or out on different parts of the sequence as needed.

The architecture is designed to enable efficient GPU computation by utilizing both GPU SRAM and HBM (High Bandwidth Memory). This dual-memory approach allows for efficient processing of long sequences, as it balances the need for fast access to recent information with the ability to reference information from much earlier in the sequence when necessary. Its architecture enables Mamba to process sequences more efficiently than Transformers while still capturing complex dependencies. In the context of molecular generation, this suggests potential for processing larger datasets and modelling more complex molecular structures.

3 Methodology

Our study aims to evaluate the efficacy of autoregressive sequence models in molecular generation tasks. We focus on comparing Transformer-based models and State Space Models (SSMs), both implemented as autoregressive sequence models for molecular generation.

3.1 Dataset Preparation

We train models on two datasets: the Molecular Sets (MOSES) dataset and a canonicalized subset of the ZINC database. The MOSES dataset [15], comprising approximately 1.6 million drug-like molecules, serves as our primary benchmark. MOSES offers a representation of the chemical space relevant to drug discovery, with compounds selected based on specific physicochemical properties and synthetic accessibility criteria. We use the original train-validation split to ensure comparability with previous work.

To complement MOSES and assess the scalability of our findings, we incorporate a larger dataset derived from ZINC20 [16]. Specifically, we used a canonicalized subset of 23 million molecules from ZINC.[1] This expanded dataset allows us to investigate whether the trends observed with MOSES persist when applied to a larger and more diverse chemical space. We randomly split the ZINC subset was into training (90%) and validation (10%) sets.

The length distribution of molecular representations varies between datasets. In the MOSES dataset, 90% of molecules have token lengths between 25 and 45 tokens, with a median length of 35 tokens. The ZINC dataset shows greater variation, with 90% of molecules containing between 32 and 62 tokens and a median length of 45 tokens. We set the maximum sequence length to 1024 tokens for all models to ensure complete coverage of all molecules in both datasets while maintaining consistency across model architectures.

We apply identical preprocessing to both datasets. We convert the original SMILES strings into the SAFE (Sequential Attachment-based Fragment Embedding) representation using the SAFE library.[2] The SAFE encoding process involves extracting unique ring digits from the SMILES string, fragmenting the molecule using methods such as BRICS [9], sorting fragments by size, concatenating fragment SMILES strings, and replacing attachment points with new ring digits.

SAFE strings are tokenized with the pre-trained byte-pair encoding (BPE) tokenizer from the SAFE-GPT model.[3] This tokenizer has a vocabulary size of 1 880 tokens and using it ensures ensures consistency with the original SAFE-GPT implementation.

3.2 Model Architectures and Training

We train four models across two architectures: Transformers (SAFE-GPT_Small and SAFE-GPT_Large) and State Space Models (MAMBA_Small and MAMBA_Large). These models were selected to investigate both small (approximately 20M parameters) and large (approximately 90M parameters) variants.

[1] `https://huggingface.co/datasets/sagawa/ZINC-canonicalized`

[2] `https://github.com/datamol-io/SAFE`

[3] `https://huggingface.co/datamol-io/safe-gpt`

Table 1: Performance metrics for molecular generation models. Top: models trained in this study. Bottom: results from previous work. Metrics: Valid@10K, Unique@10K, and Diversity (based on Tanimoto similarity). Higher values indicate better performance for all metrics. Models with $*$ for uniqueness obtain 1.0 for uniqueness after regenerating any molecules that fail SAFE-to-SMILES conversion.

Model	Valid@10K↑	Unique@10K↑	Diversity↑
SAFE-GPT_Large (87M)	00.98	**1**	0.880
Mamba_Large (94M)	0.72*	**1**	0.873
SAFE-GPT_Small (21M)	**1**	0.999	0.864
Mamba_Small (20M)	0.62*	0.999	0.860
GSELFIES-GPT20M [7]	**1**	0.999	**0.887**
GSELFIES-VAE [7]	**1**	0.999	0.859
SELFIES-VAE [7]	**1**	0.999	0.858
GMT-SELFIES [17]	**1**	**1**	0.870
CharRNN [15]	0.975	0.999	0.856
VAE [15]	0.977	0.998	0.856
LatentGAN [15]	0.897	0.997	0.857
JT-VAE [15]	**1**	0.999	0.855
LigGPT [13]	0.900	0.999	0.871

Our SAFE-GPT models reproduce the original SAFE models [8], serving as our Transformer-based baselines. We adapt the original MAMBA implementation[4] and integrate it with the rest of the SAFE library. This ensures that only the architecture is varied while keeping all other aspects of the pipeline constant. Our molecular generation implementation is available at `https://github.com/Anri-Lombard/Mamba-SAFE`.

All models were trained on NVIDIA A100 GPUs. The small models were trained for 10 epochs, while the large models were trained for a fixed number of 250,000 steps, corresponding to approximately 2.4 epochs. Detailed model architecture parameters and training hyperparameters can be found in Appendix A.

3.3 Molecule Generation and Evaluation

We generate molecules as SAFE sequences from the trained models using sampling-based decoding [18]. We use same decoding parameters as in the original SAFE-GPT implementation (which follows the default Hugging Face decoding parameters): a temperature of 1.0, top-p parameter of 1.0, and a top-k parameter of 50. These parameters are used for both the SAFE-GPT and MAMBA models. For evaluation we generated 10,000 molecules from each model (as a single trial per model).

Our evaluation framework encompasses both quantitative measures and qualitative analyses, building upon established metrics widely used in molecular generation literature [15, 12]. We assess Validity, calculated using RDKit [19], which has been established as a key metric for evaluating molecular generation models [17, 7]. Uniqueness and Diversity, first introduced for molecular generation evaluation by [10], assess the model's ability to generate distinct molecular structures and the structural variety within the generated set, respectively.

Validity is calculated as the fraction of molecules successfully converted to valid RDkit molecules, where the generate SAFE strings are first converted to SMILES before being validated by RDkit. While the SAFE-to-SMILES conversion is always successful for SAFE-GPT outputs, we found that for the MAMBA models some generated molecules unexpected fail this conversion step, although all molecules that pass it are successfully converted to RDkit molecules. Therefore, while we report the initial validity which requires passing both validation steps, our generation process includes a retry mechanism that regenerates any molecules that fail the SAFE-to-SMILES conversion step. The rest of the evaluations use the set of molecules obtained after this. See Appendix B for detailed discussion of the two-step validation process.

Uniqueness assesses the model's ability to generate distinct molecular structures. Diversity measures the structural variety within the generated set. It is quantified using the average pairwise Tanimoto

[4] `https://github.com/state-spaces/mamba`

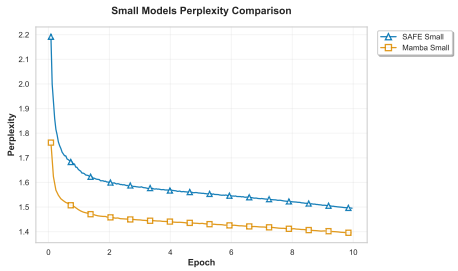


Figure 2: Validation set perplexity of SAFE-GPT_Small and MAMBA_Small models during training on the MOSES dataset.

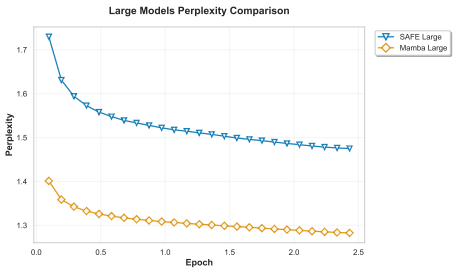


Figure 3: Validation set perplexity of SAFE-GPT_Large and MAMBA_Large models during training on the ZINC dataset.

distance between molecules based on their ECFP4 fingerprint representations [20]. Diversity is calculated using the following equation:

$$\text{Diversity} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (1 - T(m_i, m_j)), \tag{6}$$

where N is the number of molecules, m_i and m_j are molecules, and $T(m_i, m_j)$ is the Tanimoto similarity between their ECFP4 fingerprints.

To gauge how well the models capture the characteristics of drug-like molecules, we compare the distributions of key physicochemical properties between the generated molecules and the training set. These properties, which are crucial in drug discovery [21, 22], include molecular weight, LogP, topological polar surface area (TPSA), number of rotatable bonds, hydrogen bond acceptors and donors, and aromatic rings.

We also evaluate the Quantitative Estimate of Drug-likeness (QED) [23], a composite measure that combines several molecular properties to assess how drug-like a compound is. QED is calculated as the geometric mean of desirability functions for each property:

$$QED = exp\left(\frac{1}{n} \sum_{i=1}^{n} \ln d_i\right), \tag{7}$$

where d_i are the desirability functions for each molecular descriptor.

Additionally we assess the computational resource utilization of the evaluated models by measuring GPU power consumption and GPU utilization throughout training. This enables us to compare the efficiency of each model architecture in terms of energy consumption, hardware utilization, and overall training time, providing insights into their scalability and potential for handling larger datasets or more complex molecular structures.

4 Results

4.1 Model Comparison

Table 1 summarizes the key performance metrics for our models, alongside previously reported results for other molecular generation approaches with a comparable setup. While all models achieve high final validity scores after applying our retry mechanism for invalid generations, there are important differences in their initial generation behavior. The SAFE-GPT models generate molecules that consistently convert from SAFE to SMILES format without errors, achieving perfect (1.0) or close to perfect (0.98) validity directly. In contrast, MAMBA models with default sampling parameters (top-p = 1.0) initially show SAFE-to-SMILES conversion failures in 28-38% of generations, requiring regeneration to achieve perfect validity. Reducing the top-p parameter to 0.9 decreases but does not eliminate these failures. This behavior, which is unexpected given SAFE's design principles, suggests that the representation might be particularly optimized for Transformer architectures.

After accounting for regeneration, uniqueness is consistently high across all models, with large models achieving perfect uniqueness (1.000) and small models reaching near-perfect uniqueness

(0.999). The diversity scores are comparable across the models, with SAFE-GPT_Large achieving the highest score of 0.880, followed closely by MAMBA_Large at 0.873. The small models show slightly lower but still competitive diversity scores. The diversity scores are comparable across the models, with SAFE-GPT_Large achieving the highest score of 0.880, followed closely by MAMBA_Large at 0.873. The small models show slightly lower but still competitive diversity scores.

These results demonstrate a parity in performance between State Space (MAMBA) and Transformer (SAFE-GPT) architectures across all three evaluation metrics for molecular generation. Both small (20M parameters) and large (90M parameters) models achieve high validity and uniqueness scores, regardless of the underlying architecture. This parity extends the applicability of State Space Models, previously shown to be effective in language tasks, to the domain of molecular generation. The comparable diversity scores across all models further demonstrates that SSMs can be as effective as Transformers in exploring vast chemical spaces, a crucial aspect of molecular generation tasks.

4.2 Perplexity Analysis

Figures 2 and 3 reports the perplexity of each of the models across training epochs for small and large models, respectively, as measured on a held-out validation set at intervals throughout training. MAMBA_Small exhibits consistently lower perplexity than SAFE-GPT_Small throughout training, converging to values around 1.4 (compared to around 1.5 for SAFE-GPT_Small). The gap in perplexity is even more pronounced in the larger models, with MAMBA_Large displaying noticeably lower perplexity than SAFE-GPT_Large.

This consistently lower perplexity exhibited by the MAMBA models suggests that they have learned a better probability distribution over the space of possible molecules. This efficiency in modeling could be attributed to the continuous-time dynamics of SSMs, which may be particularly well-suited to capturing the sequential nature of molecular structures, though the interaction between this architecture and the SAFE representation requires further investigation. However, this appears to be in tension with their need for more conservative sampling (decreasing the top-p parameter) or regeneration steps to achieve high validity rates.

4.3 Molecular Property Distributions

To assess how well our models captures the characteristics of drug-like molecules, we analyze the distribution of various molecular properties for the generated compounds, following established evaluation approaches [15, 12]. For small models, we compare the distributions to the MOSES training dataset, while for large models, we compare them to the ZINC dataset. Figures 4 and 5 show the distributions of key molecular properties for small and large models, respectively.

Overall, the distributions of molecular properties for generated molecules closely matched those of their respective training datasets (MOSES for small models, ZINC for large models) across all evaluated models. This consistency was observed for all of the following properties: number of aromatic rings, indicating aromaticity patterns in drug-like molecules; hydrogen bond acceptors (HBA) and donors (HBD), crucial for predicting molecular interactions [24]; LogP, a measure of lipophilicity important for drug-like properties [24]; molecular weight, ensuring generated molecules fall within appropriate size ranges for potential drug candidates [24]; Quantitative Estimate of Drug-likeness (QED), a composite measure of overall drug-like characteristics [23]; number of rotatable bonds, indicating molecular flexibility [22]; and Topological Polar Surface Area (TPSA), representing molecular polarity [25].

Notably, the Mamba model distributions closely align with those of the SAFE-GPT models for both small and large variants. This suggests that the State Space Model architecture can capture the same molecular property characteristics as the Transformer-based model when trained on the same dataset. The ability of both architectures to accurately reproduce these property distributions demonstrates their capability to learn and generate molecules with realistic and diverse properties. This is crucial for the application of these models in drug discovery and molecular design tasks, as it ensures that the generated molecules are likely to possess the physicochemical properties required for potential drug candidates.

Table 2: Computational efficiency metrics for training molecular generation models. Reported values: GPU Utilization (%) and GPU Power Consumption (W) for SAFE-GPT and MAMBA models of different sizes.

Model	GPU Utilization (%)	GPU Power Consumption (W)
SAFE-GPT_Small	60 ± 2	280
Mamba_Small	22 ± 1	190
SAFE-GPT_Large	95 ± 5	360
Mamba_Large	80 ± 15	280

4.4 Computational Efficiency Comparison

Table 2 reports the computational efficiency metrics for training each model, including GPU utilization and GPU power consumption. MAMBA models consistently demonstrates lower GPU power consumption than the SAFE-GPT models. MAMBA_Small consumed approximately 32% less power than SAFE-GPT_Small, while MAMBA_Large consumed about 22% less power than SAFE-GPT_Large. GPU utilization patterns correspondingly reveal that MAMBA_Small has significantly lower GPU utilization (22%) compared to SAFE-GPT_Small (60%). For large models, the gap narrows, with MAMBA_Large utilizing 80% of GPU capacity compared to 95% for SAFE-GPT_Large. This substantial reduction in computational resource requirements could prove crucial for scaling up to larger datasets or more complex molecular structures, potentially enabling the exploration of chemical spaces that were previously computationally infeasible.

The small Mamba model trains slightly longer than the Transformer-based model despite its lower resource utilization ($\tilde{1}0$ hours for MAMBA_Small against $\tilde{8}$ hours for SAFE-GPT_Small). However, this trend reverses for large models. The 90M parameter SAFE-GPT model took approximately 90 hours to train for 250,000 steps, while the equivalent Mamba model completed the same training in 64 hours. This observation suggests that the efficiency advantages of MAMBA models become more pronounced as model size increases, offering significant time savings for large-scale molecular generation tasks.

These efficiency gains observed in MAMBA models, particularly at larger scales, points to the potential to apply the Mamba architecture to train models for more complex molecular structures or to iterate through model designs more rapidly. This could potentially accelerate the drug discovery process and enable more comprehensive explorations of chemical space in materials science applications. These efficiency comparisons are dependent on the training data's sequence lengths. Additionally, the need for regeneration steps with MAMBA models suggests that further architectural optimization might be needed to fully realize their potential with the SAFE representation. We hypothesize that that the efficiency gap will be even wider with longer sequences, due to Mamba's lower algorithmic complexity. However, further investigations are required to verify the scalability of these efficiency gains to larger models or different hardware configurations.

5 Conclusion

This paper empirically validates the efficacy of State Space Models, specifically the MAMBA architecture, for molecular generation. By demonstrating comparable performance to Transformer-based models in generating valid, unique, and diverse molecules, we contribute to the growing body of evidence suggesting that SSMs represent a viable alternative to attention-based architectures across diverse domains. The success of MAMBA models in capturing the intricacies of molecular structures, as encoded in the SAFE representation, underscores the versatility of SSMs. The efficiency advantage demonstrated by MAMBA-based models, evidenced by substantial reductions in GPU power consumption and improved training times for larger models, highlights the theoretical advantage of SSMs in processing sequences in linear time complexity. By offering a balance between generation quality and computational efficiency, MAMBA models enable new possibilities for scaling up molecular generation tasks to handle larger datasets and more complex molecular representations, enabling more efficient exploration of vast chemical spaces.

While our study provides valuable insights into State Space Models for molecular generation, several limitations warrant further investigation. Our analysis was confined to models with 20M and 90M

parameters, leaving the performance characteristics of larger-scale architectures unexplored. Additionally, the observed differences in SAFE-to-SMILES conversion success between MAMBA and Transformer models suggests a need to investigate the interaction between molecular representations and model architectures more thoroughly. Future work should extend this comparison to more extensive models and diverse molecular datasets, particularly focusing on how task performance and computational efficiencies scale with sequence length. Additionally, the discrepancy between MAMBA models' lower perplexity and their need for more conservative sampling parameters to ensure validity requires further examination. Future research should also explore the application of these architectures to more complex tasks such as targeted molecule design or optimization of specific molecular properties. Such advancements could significantly impact drug discovery processes and materials development, potentially enabling the exploration of previously intractable chemical spaces.

Acknowledgments and Disclosure of Funding

Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team.

References

[1] Petra Schneider, W. Patrick Walters, Alleyn T. Plowright, Norman Sieroka, Jennifer Listgarten, Robert A. Goodnow, Jasmin Fisher, Johanna M. Jansen, José S. Duca, Thomas S. Rush, Matthias Zentgraf, John Edward Hill, Elizabeth Krutoholow, Matthias Kohler, Jeff Blaney, Kimito Funatsu, Chris Luebkemann, and Gisbert Schneider. Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery*, 19(5):353–364, May 2020. ISSN 1474-1784. doi: 10.1038/s41573-019-0050-3. URL https://doi.org/10.1038/s41573-019-0050-3.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

[4] Daria Grechishnikova. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific Reports*, 11(1):321, January 2021. ISSN 2045-2322. doi: 10.1038/s41598-020-79682-4. URL https://doi.org/10.1038/s41598-020-79682-4.

[5] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

[6] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, February 1988. ISSN 0095-2338. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005. Publisher: American Chemical Society.

[7] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. *CoRR*, abs/1905.13741, 2019. URL http://arxiv.org/abs/1905.13741.

[8] Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan S. C. Lim, and Prudencio Tossou. Gotta be safe: a new framework for molecular design. *Digital Discovery*, 3:796–804, 2024. doi: 10.1039/D4DD00019F. URL http://dx.doi.org/10.1039/D4DD00019F.

[9] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008. doi: https://doi.org/10.1002/cmdc.200800178. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.200800178.

[10] Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating focussed molecule libraries for drug discovery with recurrent neural networks. *CoRR*, abs/1701.01329, 2017. URL http://arxiv.org/abs/1701.01329.

[11] Connor W. Coley, Wengong Jin, Luke Rogers, Timothy F. Jamison, Tommi S. Jaakkola, William H. Green, Regina Barzilay, and Klavs F. Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.*, 10:370–377, 2019. doi: 10.1039/C8SC04228D. URL http://dx.doi.org/10.1039/C8SC04228D.

[12] Nathan Brown, Marco Fiscato, Marwin H.S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.*, 59(3):1096–1108, March 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00839. URL https://doi.org/10.1021/acs.jcim.8b00839. Publisher: American Chemical Society.

[13] Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. Liggpt: Molecular generation using a transformer-decoder model. *ChemRxiv*, 2021. URL https://api.semanticscholar.org/CorpusID:239738964.

[14] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885, 2020. URL https://arxiv.org/abs/2010.09885.

[15] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez, Sergey Golovanov, Oktay Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alán Aspuru-Guzik, and Alexander Zhavoronkov. Molecular sets (moses): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11, 12 2020. doi: 10.3389/fphar.2020.565644.

[16] Teague Sterling and John J. Irwin. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.*, 55(11):2324–2337, November 2015. ISSN 1549-9596. doi: 10.1021/acs.jcim.5b00559. URL https://doi.org/10.1021/acs.jcim.5b00559. Publisher: American Chemical Society.

[17] Wengong Jin, Regina Barzilay, and Tommi S. Jaakkola. Junction tree variational autoencoder for molecular graph generation. *CoRR*, abs/1802.04364, 2018. URL http://arxiv.org/abs/1802.04364.

[18] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *CoRR*, abs/1904.09751, 2019. URL http://arxiv.org/abs/1904.09751.

[19] Greg Landrum et al. RDKit: Open-source cheminformatics. *Online. http://www. rdkit. org*, 2023.

[20] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50 (5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t. Publisher: American Chemical Society.

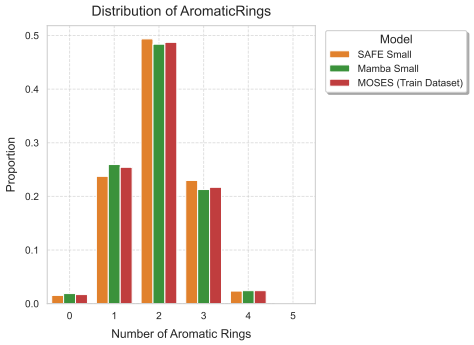
[21] Christopher A. Lipinski. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337–341, 2004. ISSN 1740-6749. doi: https://doi.org/10.1016/j.ddtec.2004.11.007. URL https://www.sciencedirect.com/science/article/pii/S1740674904000551.

[22] Daniel F. Veber, Stephen R. Johnson, Hung-Yuan Cheng, Brian R. Smith, Keith W. Ward, and Kenneth D. Kopple. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.*, 45(12):2615–2623, June 2002. ISSN 0022-2623. doi: 10.1021/jm020017n. URL https://doi.org/10.1021/jm020017n. Publisher: American Chemical Society.

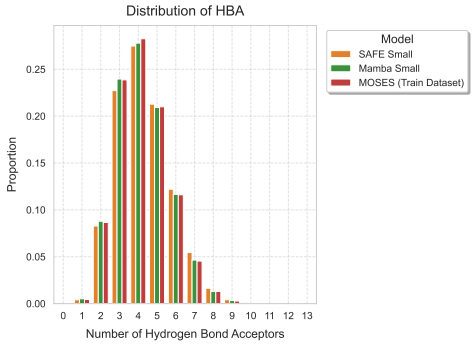
[23] G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, February 2012. ISSN 1755-4349. doi: 10.1038/nchem.1243. URL https://doi.org/10.1038/nchem.1243.

[24] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1):3–25, 1997. ISSN 0169-409X. doi: https://doi.org/10.1016/S0169-409X(96)00423-1. URL https://www.sciencedirect.com/science/article/pii/S0169409X96004231. In Vitro Models for Selection of Development Candidates.

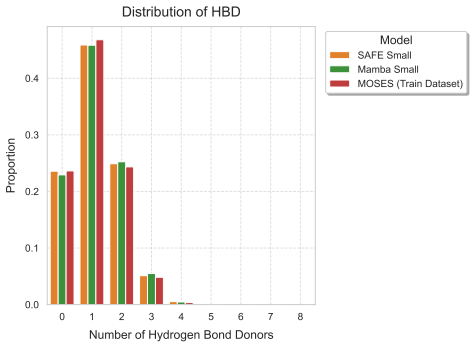
[25] Peter Ertl, Bernhard Rohde, and Paul Selzer. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.*, 43(20):3714–3717, October 2000. ISSN 0022-2623. doi: 10.1021/jm000942e. URL https://doi.org/10.1021/jm000942e. Publisher: American Chemical Society.



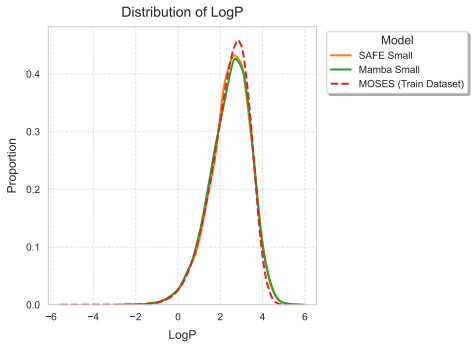
(a) Aromatic Rings



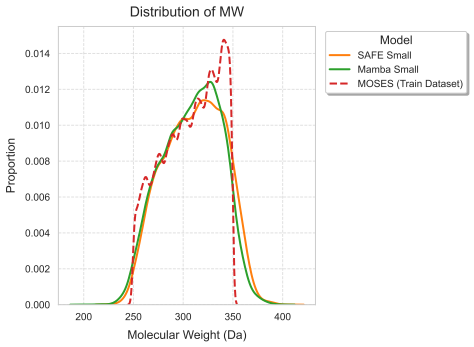
(b) H-Bond Acceptors



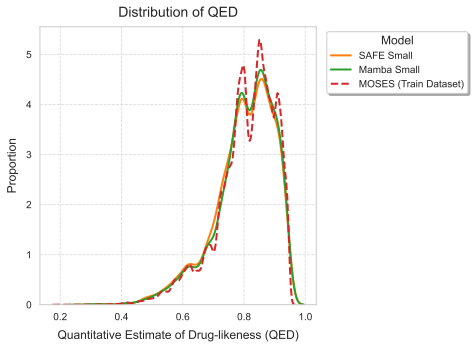
(c) H-Bond Donors



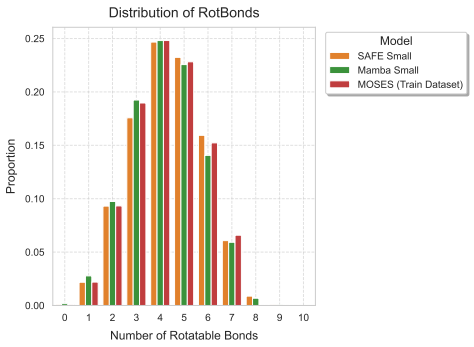
(d) LogP



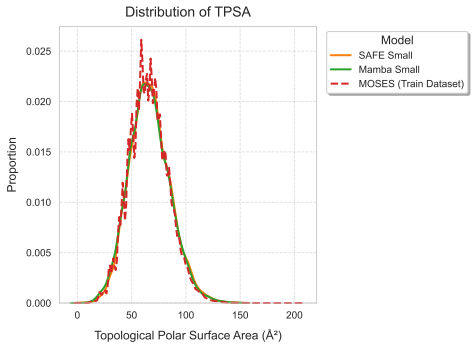
(e) Molecular Weight



(f) QED

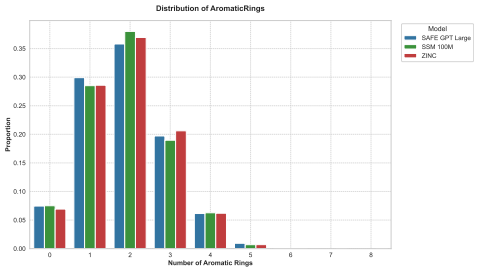


(g) Rotatable Bonds

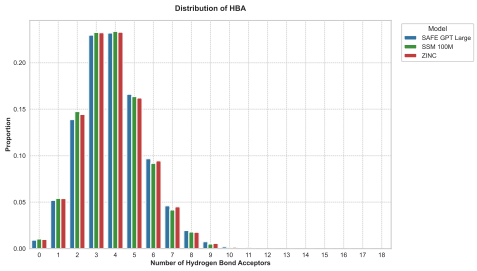


(h) TPSA

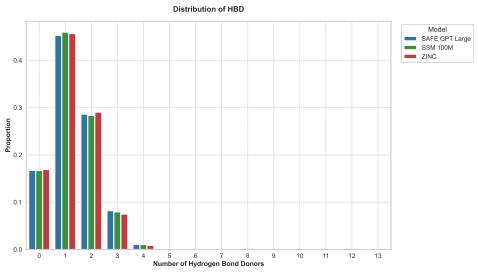
Figure 4: Distributions of key molecular properties for molecules generated by small models (SAFE-GPT_Small and MAMBA_Small, both ∼20M parameters) compared to the MOSES dataset.



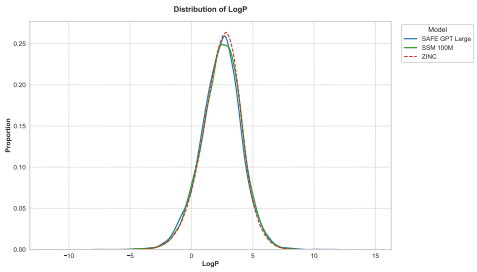
(a) Aromatic Rings



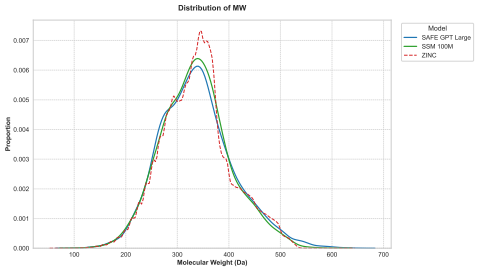
(b) H-Bond Acceptors



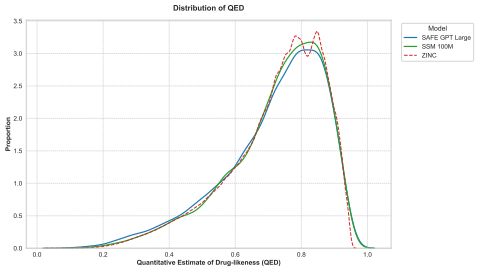
(c) H-Bond Donors



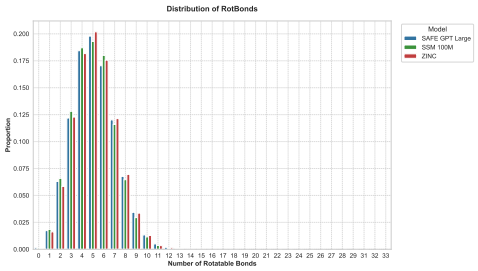
(d) LogP



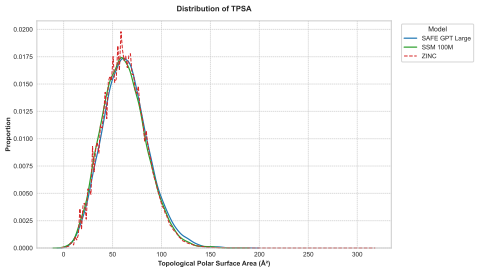
(e) Molecular Weight



(f) QED



(g) Rotatable Bonds



(h) TPSA

Figure 5: Distributions of molecular properties for molecules generated by SAFE-GPT_Large and MAMBA_Large models (∼90M parameters) compared to the ZINC dataset.

A Model Architecture and Training Details

A.1 Model Architecture Parameters

Tables 3 and 4 summarize the key parameters of each model architecture used in our study.

Table 3: SAFE-GPT Model Architecture Parameters

Parameter	SAFE-GPT-Small	SAFE-GPT-Large
Model Type	Transformer	Transformer
Embedding Dimension	512	768
Number of Layers	6	12
Attention Heads	8	12
Max Sequence Length	1024	1024
Dropout Rate	0.1	0.1
Normalization	LayerNorm	LayerNorm

Table 4: MAMBA Model Architecture Parameters

Parameter	MAMBA-Small	MAMBA-Large
Model Type	SSM	SSM
Embedding Dimension	512	768
Number of Layers	6	12
SSM Variant	Mamba2	Mamba2
Max Sequence Length	1024	1024
Dropout Rate	0.1	0.1
Normalization	RMSNorm	RMSNorm
Residual Connections	FP32	FP32

A.2 Model Training Parameters

Table 5 summarizes the key training parameters for both small and large models.

Table 5: Training Parameters for Small and Large Models

Parameter	Small Models	Large Models
Optimizer	AdamW	AdamW
Learning rate	5e-4	1e-4
Warmup steps	20,000	10,000
Weight decay	0.1	0.1
Gradient clipping	1.0	1.0
Batch size (per device)	32	100
Gradient accumulation steps	2	2
Effective batch size	64	200
Training duration	10 epochs	250,000 steps

B Validity Calculation Details

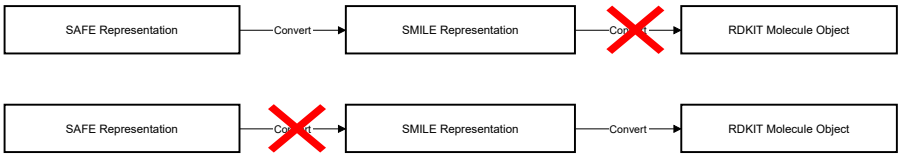


Figure 6: Validity calculation process. Top: Traditional invalidity case where SMILES to RDKit conversion fails, which is what's measured in validity scores. Bottom: SAFE to SMILES conversion failure, which according to SAFE's design should not occur, holds true for Transformer models but was observed with MAMBA models using default parameters.

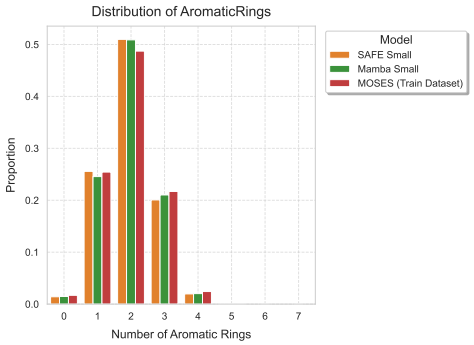
The validity metric measures the success rate of converting SMILES strings to RDKit molecule objects. The SAFE representation is supposed to ensure robust conversion to SMILES strings through its grammatical constraints. Notably, MAMBA models with default sampling parameters often generate SAFE representations that fail to convert to SMILES format, triggering decoder errors. When accounting for these SAFE-to-SMILES conversion failures, the actual validity rates for MAMBA models drops by 28-38%. Reducing the top-p parameter to 0.9 decreases but does not eliminate these failures. This behavior suggests the SAFE representation might be particularly optimized for Transformer architectures. However our implementation includes a retry mechanism that regenerates batches containing invalid molecules, ensuring 100% validity in the SAFE to SMILES conversion step. In the absence of a better explanation of the SAFE-to-SMILES conversion failures we believe this enables fairer comparisons for the rest of the evaluations.

Figures 7 and 8 show the molecular property distributions when using top-p = 0.9 for small and large MAMBA models, respectively. As expected, these distributions slightly deviate from their respective training distributions due to the more conservative sampling strategy, but they demonstrate fewer decoding errors compared to using top-p = 1.0. This trade-off between distribution matching and generation stability warrants further investigation.

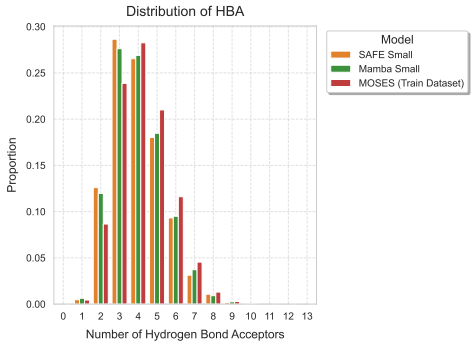
Further research should evaluate SAFE's robustness across different model architectures without the retry mechanism to comprehensively assess its architecture-specific performance.

C Example Molecules

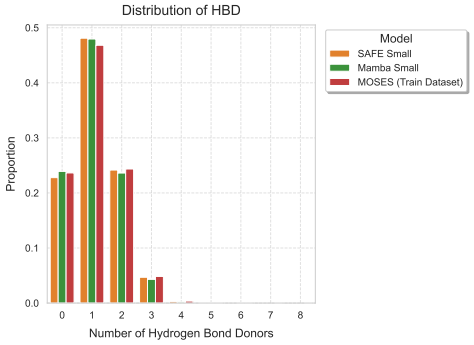
This appendix presents representative molecules generated by each model, showcasing the longest, shortest, most diverse, and highest QED molecules from the 10k generated.



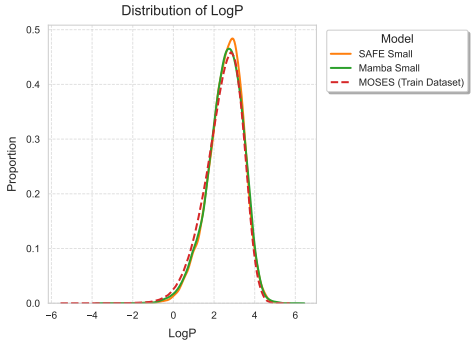
(a) Aromatic Rings



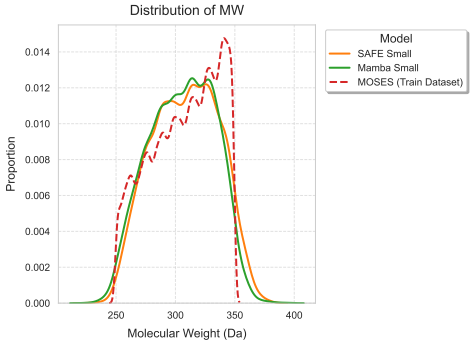
(b) H-Bond Acceptors



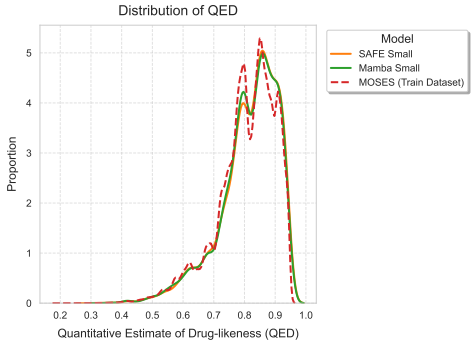
(c) H-Bond Donors



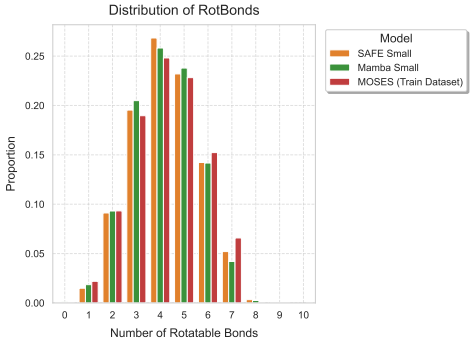
(d) LogP



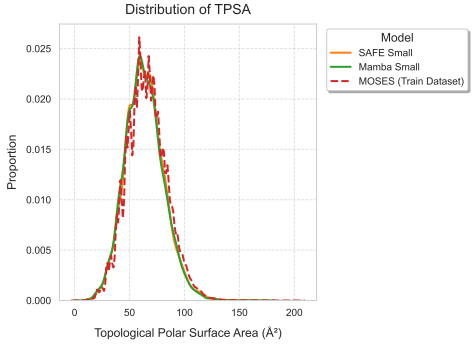
(e) Molecular Weight



(f) QED

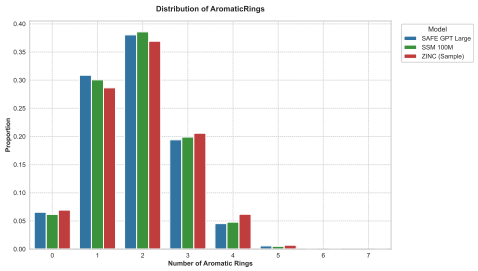


(g) Rotatable Bonds

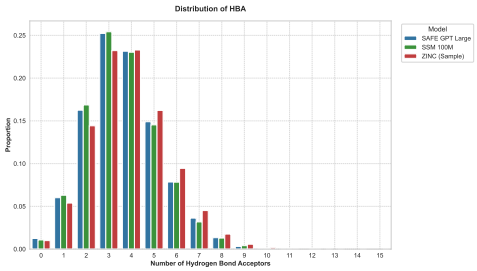


(h) TPSA

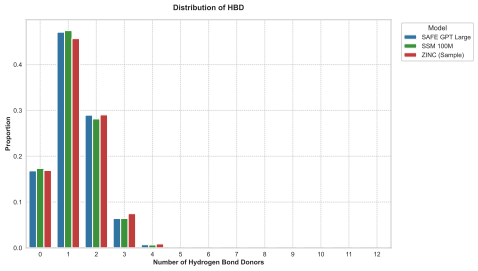
Figure 7: Molecular property distributions using top-p = 0.9 for small MAMBA models. While these distributions show slight deviations from the MOSES dataset compared to top-p = 1.0, they exhibit improved generation stability with fewer decoding errors.



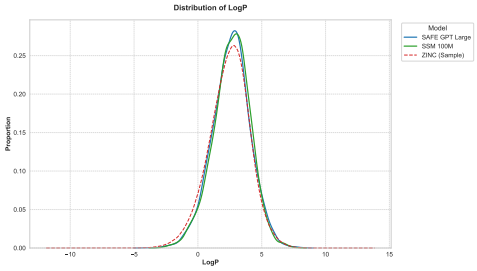
(a) Aromatic Rings



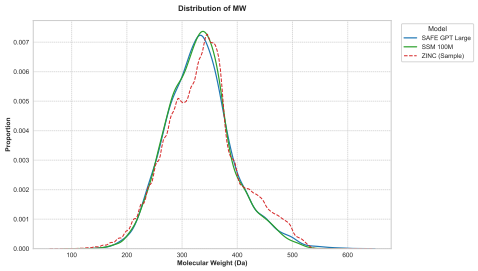
(b) H-Bond Acceptors



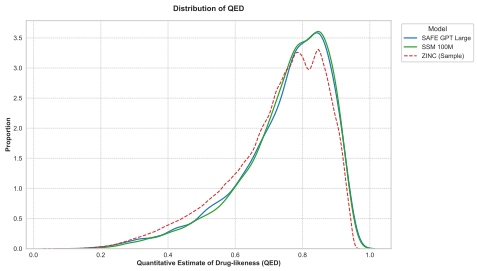
(c) H-Bond Donors



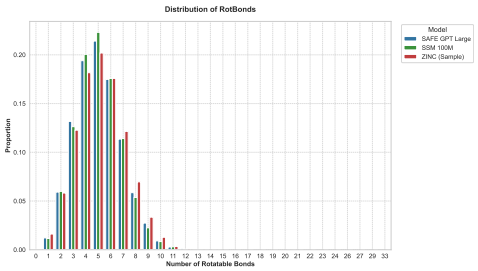
(d) LogP



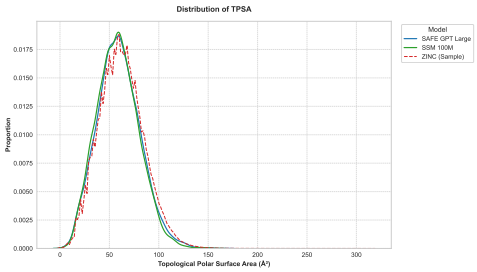
(e) Molecular Weight



(f) QED



(g) Rotatable Bonds



(h) TPSA

Figure 8: Molecular property distributions using top-p = 0.9 for large MAMBA models. While these distributions show slight deviations from the ZINC dataset compared to top-p = 1.0, they exhibit improved generation stability with fewer decoding errors.

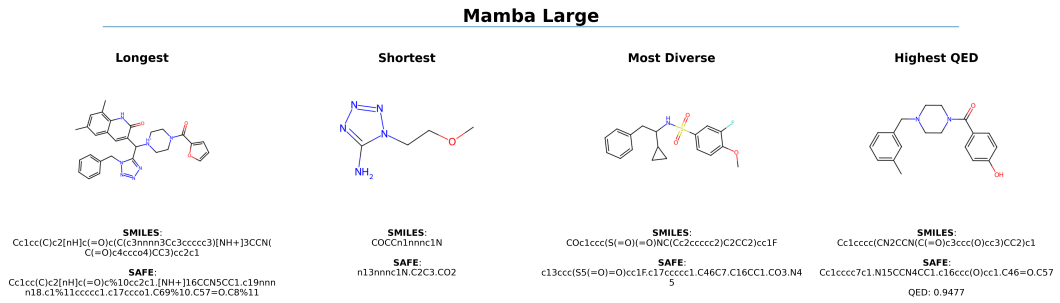


Figure 9: Representative molecules generated by the Mamba_Large model

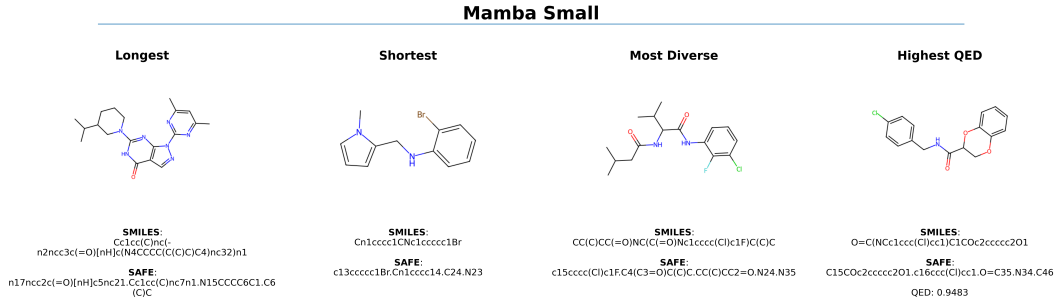


Figure 10: Representative molecules generated by the Mamba_Small model

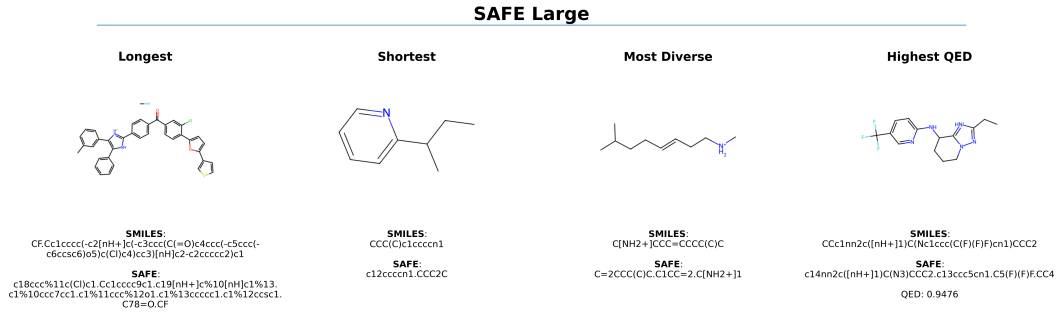


Figure 11: Representative molecules generated by the SAFE_Large model

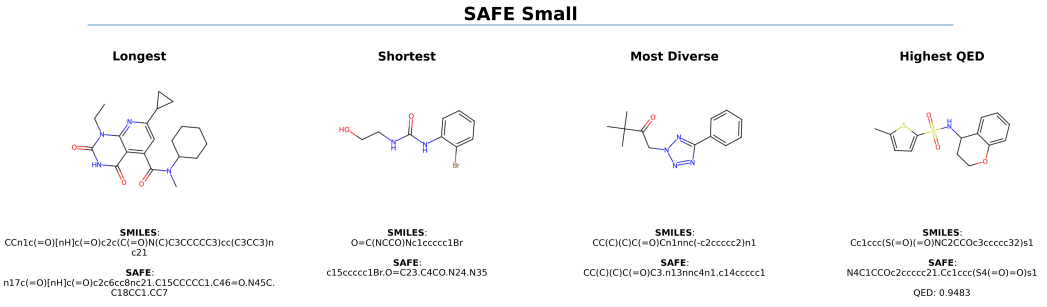


Figure 12: Representative molecules generated by the SAFE_Small model

D Supplemental Materials

To ensure reproducibility and facilitate further research, we provide open access to both the code used in our experiments and the datasets employed in our study.

D.1 Code Availability

The code used to implement and train the models is available in two separate repositories:

D.1.1 MAMBA Model Implementation

The implementation of the State Space Model (MAMBA) architecture adapted for molecular generation tasks is available in a public GitHub repository: `https://github.com/Anri-Lombard/Mamba-SAFE` This repository contains the necessary scripts and configuration files to reproduce our results for the MAMBA models.

D.1.2 SAFE-GPT Model Implementation

The Transformer-based SAFE-GPT model implementation and associated code can be found in the official SAFE repository: `https://github.com/datamol-io/safe` This repository provides the complete implementation of the SAFE-GPT models used in our comparative study.

D.2 Model Weights

Pre-trained model weights for all configurations are available on Hugging Face:

- Large Models (90M parameters):
  - SAFE-GPT: `https://huggingface.co/anrilombard/safe-100m`
  - MAMBA: `https://huggingface.co/anrilombard/ssm-100m`
- Small Models (20M parameters):
  - SAFE-GPT: `https://huggingface.co/anrilombard/safe-20m`
  - MAMBA: `https://huggingface.co/anrilombard/ssm-20m`

D.3 Datasets

We used two primary datasets in our study, both of which are openly accessible:

D.3.1 MOSES Dataset

The Molecular Sets (MOSES) dataset, comprising approximately 1.6 million drug-like molecules, is available at:

`https://github.com/molecularsets/moses`

This dataset serves as our primary benchmark and is widely used in molecular generation tasks.

D.3.2 ZINC Dataset

We used a canonicalized subset of the ZINC database, containing 23 million molecules. This dataset is accessible via Hugging Face:

`https://huggingface.co/datasets/sagawa/ZINC-canonicalized`

This larger dataset allowed us to assess the scalability of our models and findings.

D.4 Usage Instructions

Detailed instructions for using the provided code and datasets to reproduce our experiments are included in the README files of the respective GitHub repositories. These instructions cover environment setup, data preprocessing, model training, and evaluation procedures for both the MAMBA and SAFE-GPT models.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state our main contributions: comparing Transformer and MAMBA architectures for molecular generation using SAFE-GPT representation, evaluating models with 20M and 90M parameters, and assessing computational efficiency. These claims are supported by the results presented in Sections 3 and 4.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Section 4, acknowledging the limited range of model sizes explored (20M and 90M parameters) and the need for further research on larger scales and more complex molecular structures. We also acknowledge the limitation in validity calculation, where to reduce variables we maintained the same validity

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper focuses on empirical comparisons and does not present new theoretical results requiring formal proofs. We provide an overview of existing theoretical foundations in Section 2.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed methodology in Section 3, including dataset preparation, model architectures, training procedures, and evaluation metrics. Appendix A contains model architecture parameters and training details. This information, combined with the open-source code and datasets, enables result reproduction.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to code through two repositories: the MAMBA implementation at https://github.com/Anri-Lombard/Mamba-SAFE and the SAFE-GPT implementation at https://github.com/datamol-io/safe. Pre-trained model weights are available on Hugging Face. The MOSES dataset (https://github.com/molecularsets/moses) and ZINC subset (https://huggingface.co/datasets/sagawa/ZINC-canonicalized) are openly accessible. Instructions for reproducing results are included in the respective repositories.

6. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: We provide detailed methodology in Section 3, including dataset preparation, model architectures, training procedures, and evaluation metrics. Appendix A contains model architecture parameters and training details. This information enables result reproduction.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or statistical significance tests for the main results due to computation and time constraints. We acknowledge this as a limitation of our current study.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify in Section 3 that models were trained on NVIDIA A100 GPUs. Table 2 in Section 4 provides GPU utilization, power consumption, and training times for different model sizes.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research adheres to ethical guidelines. We use publicly available datasets and standard machine learning practices. Our study does not involve human subjects or raise ethical concerns related to data collection or model application.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential impacts in the Introduction and Conclusion. We highlight potential benefits in drug discovery and materials science, and address computational efficiency considerations related to energy consumption in AI research.

11. **SAFE-GPTguards**

    Question: Does the paper describe SAFE-GPTguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our models and data do not pose high risks for misuse. The research focuses on molecular generation for scientific applications and does not involve sensitive information or high-risk AI applications.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the sources of datasets and existing models used in the study, including the MOSES dataset, ZINC database, and SAFE-GPT representation. The use of these resources complies with their respective licenses and terms of use.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not introduce new datasets or pre-trained models. Our research focuses on comparing existing architectures using established datasets.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our research does not involve crowdsourcing or human subjects. The study is computational, focusing on machine learning architectures for molecular generation.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our research does not involve human subjects. The study focuses on computational experiments comparing machine learning models for molecular generation, which does not require IRB approval.