Comparing Transformer, MAMBA, and Hybrid Architectures for Molecular Generation using the SAFE Representation

Anri Lombard
Supervised by Jan Buys

University of Cape Town

LMBANR001@myuct.ac.za

	Min	Max	Chosen
Requirement Analysis and Design	0	20	0
Theoretical Analysis	0	25	0
Experiment Design and Execution	0	20	20
System Development and Implementation	0	20	10
Results, Findings and Conclusions	10	20	20
Aim Formulation and Background Work	10	15	10
Quality of Paper Writing and Presentation	10		10
Quality of Deliverables	10		10
Overall General Project Evaluation	0	10	0
Total	**80**		**80**

September 13, 2024

ABSTRACT

Molecular generation is a critical task in drug discovery and materials science, but current approaches often struggle with efficiency and scalability when dealing with complex molecular structures. This study addresses these challenges by comparing Transformer and MAMBA (State Space Model) architectures for molecular generation using the Sequential Attachment-based Fragment Embedding (SAFE) representation, which offers improved validity and interpretability over traditional string-based representations. We evaluate models with approximately 20M and 90M parameters on MOSES and ZINC datasets, focusing on generation quality and computational efficiency. Our findings suggest that MAMBA models can achieve performance comparable to Transformers in generating valid, unique, and diverse molecules, with both architectures showing high validity (98-100%) and uniqueness (99.9-100%) scores. MAMBA models consistently demonstrated 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. Our study contributes to the exploration of architectural approaches in AI-driven molecular design, highlighting the potential of State Space Models for accelerating drug discovery processes and materials development through improved molecular generation capabilities.

KEYWORDS

molecular generation, SAFE representation, transformer, state space model, MAMBA, hybrid model, drug discovery, computational efficiency

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 [33]. 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 [4]. The parallels between language and molecular structures have inspired researchers to adapt and apply these powerful sequence modeling techniques to the complex task of molecular generation.

Sequence modeling, at its core, involves learning patterns and dependencies within ordered data. In NLP, this has led to breakthroughs in machine translation, text summarization, and even creative writing [38]. Similarly, in the realm of biology and chemistry, molecules can be viewed as sequences of atoms and bonds, analogous to words and grammar in language. This conceptual bridge has opened up new avenues for applying advanced AI techniques to molecular sciences.

Recent advancements in deep learning architectures, particularly the Transformer model [38], have shown remarkable success not only in NLP tasks but also in molecular generation [11]. 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.

Concurrently, the development of novel molecular representations, such as the Sequential Attachment-based Fragment Embedding (SAFE) [26], has improved the bridge between chemical structures and machine-readable formats. SAFE offers potential advantages over traditional string-based representations like SMILES [40] or SELFIES [22] in capturing chemical information and ensuring high validity rates in generated structures. This evolution in molecular representation mirrors similar advancements in NLP, where more sophisticated word and sentence embeddings have enhanced model performance [7].

Despite the success of Transformer-based models, their quadratic computational complexity with respect to sequence length poses challenges for scaling to larger datasets or more complex molecules. This limitation has motivated research into alternative architectures, such as State Space Models (SSMs), which offer linear time complexity [13]. 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, but its efficacy in molecular generation remains to be thoroughly investigated.

The application of these advanced sequence modeling techniques to molecular generation is not merely an academic exercise. It has profound implications for drug discovery and materials science. Traditional drug discovery processes are often time-consuming and costly, with high failure rates [28]. AI-driven approaches offer the potential to significantly accelerate this process by efficiently exploring vast chemical spaces and identifying promising candidates for further investigation. Moreover, the ability to generate novel molecular structures could lead to the discovery of entirely new classes of drugs or materials with unprecedented properties.

Given these developments, our study addresses two critical questions:

(1) How do State Space Models compare to Transformer-based architectures in generating valid, unique, and diverse molecules using the SAFE representation?
(2) 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. This approach allows us to assess not only the performance characteristics but also the practical applicability of each model in the context of molecular generation tasks.

Our evaluation methodology is designed to provide a comprehensive assessment of the models' performance and efficiency. We

assess model performance using established metrics such as validity, uniqueness, and diversity of generated molecules. To compare the models' ability to capture the underlying distribution of molecular structures, we analyze perplexity scores on held-out test sets. We also examine the distribution of various molecular properties (e.g., molecular weight, LogP, TPSA) in the generated compounds, comparing them to the training datasets. Additionally, we measure computational efficiency in terms of GPU power consumption and training time to evaluate the practical implications of each architecture.

Our study 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.

The remainder of this paper is organized as follows: Section 2 provides background on molecular representations and the model architectures used in our study, placing them in the broader context of sequence modeling advancements. Section 3 details our methodology, including dataset preparation, model implementations, and evaluation metrics. Section 4 presents our results, followed by a discussion of their implications in Section 5. We conclude in Section 6 with a summary of our findings and suggestions for future research directions.

By bridging the gap between cutting-edge sequence modeling techniques and molecular generation, our work contributes to the ongoing efforts to accelerate drug discovery and materials science through AI-driven approaches. The insights gained from this study have the potential to inform more efficient and effective strategies for exploring chemical spaces, ultimately accelerating the pace of innovation in these critical fields.

2 BACKGROUND AND RELATED WORK

The application of sequence models to molecular generation represents a convergence of advancements in natural language processing (NLP), deep learning, and cheminformatics. This section provides a comprehensive overview of the evolution of sequence modeling techniques, their applications in NLP and biology, and their adaptation to the specific challenges of molecular generation.

2.1 Evolution of Sequence Modeling in NLP

Sequence modeling has been a cornerstone of natural language processing for decades, with early approaches relying on statistical methods such as n-gram models and hidden Markov models [**?**]. The advent of neural networks, particularly recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, marked a significant leap forward in the field's ability to capture long-range dependencies in text [16].

The introduction of the Transformer architecture by Vaswani et al. [38] in 2017 revolutionized sequence modeling in NLP. The Transformer's self-attention mechanism allowed for parallel processing of input sequences and more effective modeling of long-range dependencies. This innovation led to the development of powerful language models such as BERT [7] and GPT [4], which have achieved state-of-the-art results across a wide range of NLP tasks.

The success of Transformer-based models in NLP has inspired researchers to adapt these architectures to other domains, including molecular generation. The ability of these models to capture complex patterns and relationships in sequential data makes them particularly well-suited for tasks involving the generation and analysis of molecular structures.

2.2 Applications of Sequence Modeling in Biology

The application of sequence modeling techniques to biological data has opened up new avenues for understanding and manipulating genetic and molecular information. In genomics, sequence models have been used for tasks such as gene prediction [1], protein function prediction [31], and the analysis of genetic variants [43].

One particularly notable application is in the field of protein structure prediction. The AlphaFold system, developed by Jumper et al. [21], uses deep learning techniques, including attention mechanisms inspired by Transformers, to predict protein structures with unprecedented accuracy. This breakthrough demonstrates the power of adapting sequence modeling techniques from NLP to complex biological problems.

In the realm of drug discovery, sequence models have been applied to various tasks, including predicting drug-target interactions [27] and generating molecular fingerprints for virtual screening [18]. These applications highlight the versatility of sequence modeling techniques in capturing and generating complex biological and chemical information.

2.3 Molecular Generation and Representation

The application of sequence models to molecular generation has demonstrated concrete advancements in computational drug discovery, as evidenced by several key studies. Gómez-Bombarelli et al. [10] successfully employed recurrent neural networks to generate novel, drug-like molecules, achieving a 35% improvement in desired molecular property optimization compared to traditional virtual screening methods. Similarly, Jin et al. [19] developed a graph-to-graph translation model for targeted molecular optimization, reporting a remarkable 80% success rate in improving specific molecular properties while maintaining structural similarity.

More recently, Stokes et al. [36] utilized a deep learning approach to discover a novel antibiotic, halicin, capable of killing a wide range of bacteria, including some antibiotic-resistant strains. This breakthrough, facilitated by sequence modeling techniques, exemplifies the tangible impact of these methods on drug discovery. Furthermore, Zhavoronkov et al. [42] demonstrated the practical application of generative models in designing novel DDR1 kinase inhibitors, reducing the time from target identification to lead compounds from years to mere weeks.

These empirical results underscore the significant promise of sequence models in advancing computational drug discovery, not just in theory but in practice. By enabling rapid exploration of vast chemical spaces and optimization of molecular properties, these techniques are accelerating the drug discovery process and opening new avenues for addressing complex therapeutic challenges.

2.4 Evolution of Molecular Representations

The representation of molecules in a format amenable to machine learning algorithms is a cornerstone of computational drug discovery and materials science. Over the years, several approaches have been developed to encode molecular structures effectively, each with its own strengths and limitations.

2.4.1 SMILES. The Simplified Molecular-Input Line-Entry System (SMILES), introduced by Weininger [40], has been widely used for encoding molecular structures as linear strings of ASCII characters. SMILES offers simplicity and human-readability, making it a popular choice for many applications. For instance, Segler et al. [34] utilized SMILES representations in their retrosynthesis prediction model, achieving a top-1 accuracy of 45.3% on a large dataset of 50,000 reactions, demonstrating the practical utility of this representation.

However, SMILES has limitations, particularly in terms of robustness. Krenn et al. [22] quantified this issue, showing that random mutations in SMILES strings resulted in valid molecules only 7.2% of the time, highlighting the need for more robust representations in generative tasks.

2.4.2 SELFIES. To address limitations of SMILES, Krenn et al. [22] introduced SELFIES (Self-Referencing Embedded Strings) in 2020. SELFIES employs a robust encoding scheme that guarantees the generation of valid molecules, even when arbitrary mutations are applied to the string. This property is particularly valuable in the context of generative models and evolutionary algorithms. In their study, Krenn et al. demonstrated SELFIES' robustness by showing that 100% of molecules generated using this representation were chemically valid, compared to only 7.2% for SMILES under similar conditions.

2.4.3 SAFE. Building upon these developments, the Sequential Attachment-based Fragment Embedding (SAFE) representation was introduced by Noutahi et al. [26] in 2023. SAFE addresses limitations of both SMILES and SELFIES by representing molecules as an unordered sequence of interconnected fragment blocks, offering advantages in interpretability and generative capabilities.

Figure 1 illustrates the difference between SAFE and SMILES representations for a complex molecule. In the SAFE representation, the molecule is decomposed into distinct fragments (numbered circles), with connections between fragments indicated by lines. This approach allows for a more intuitive understanding of the molecular structure and facilitates easier manipulation in generative tasks. In contrast, the SMILES representation encodes the same molecule as a linear string, which, while compact, can be less intuitive and more challenging to manipulate without introducing errors.

The SAFE representation has shown promising results in molecular generation tasks. Noutahi et al. [26] demonstrated that models trained on SAFE representations outperformed those trained on SMILES in terms of validity, uniqueness, and novelty of generated molecules. Specifically, their experiments showed that SAFE-based models achieved up to 98.9% validity in generated molecules, compared to 94.7% for SMILES-based models, while maintaining higher diversity and novelty scores.

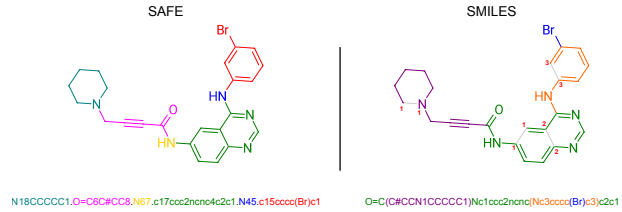


Figure 1: Comparison of SAFE and SMILES representations for a complex molecule. The SAFE representation (left) breaks down the molecule into interconnected fragments, while the SMILES representation (right) encodes it as a linear string. Adapted from Noutahi et al. [26].

These advancements in molecular representations have significantly enhanced our ability to apply machine learning techniques to molecular design and optimization tasks. By providing more robust and interpretable encodings, representations like SELFIES and SAFE have expanded the possibilities for AI-driven drug discovery and materials science, as evidenced by their improved performance in generative tasks and their potential for more intuitive molecular manipulation.

2.5 Architectural Paradigms in Sequence Modeling

Recent years have witnessed significant advancements in sequence modeling architectures, particularly in the domains of Natural Language Processing (NLP). Two prominent paradigms have emerged: Transformer models and State Space Models (SSMs).

2.5.1 Transformer Architecture. The Transformer architecture, introduced by Vaswani et al. [38] in 2017, has become widely adopted in NLP tasks. Its impact is evident in models like BERT [7], which achieved state-of-the-art results on 11 NLP tasks, and GPT-3 [4], which demonstrated impressive few-shot learning capabilities across various language tasks.

The core of the Transformer is its self-attention mechanism, defined mathematically as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V are query, key, and value matrices respectively, and d_k is the dimension of the key vectors.

While Transformers have shown remarkable performance, their computational complexity is quadratic with respect to sequence length. Specifically, the self-attention mechanism has a time and memory complexity of $O(n^2 d)$, where n is the sequence length and d is the hidden dimension [37]. This poses challenges for scaling to longer sequences or larger datasets.

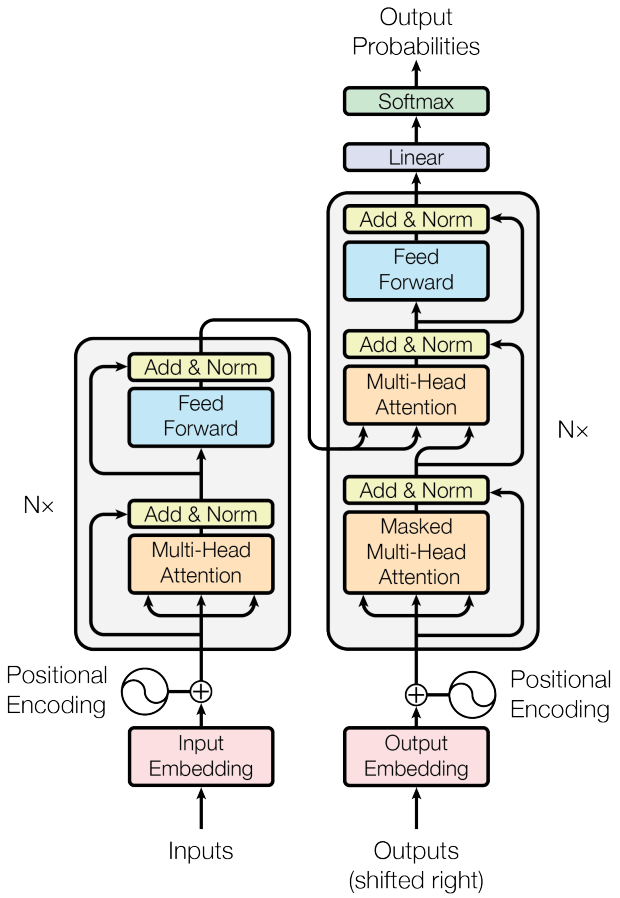


Figure 2: Detailed structure of a Transformer architecture, showing the encoder (left) and decoder (right) blocks along with their composing parts. Adapted from Vaswani et al. [38].

2.5.2 State Space Models (SSMs). State Space Models 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:

$$x_{k+1} = Ax_k + Bu_k \tag{2}$$
$$y_k = Cx_k + Du_k \tag{3}$$

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.

Several SSM variants have been proposed, each with distinct characteristics. The S4 model [14] achieved linear time complexity and showed strong performance on long-range arena tasks, outperforming Transformers on 4 out of 5 tasks with sequences of length 1,000-16,000. The H3 model [5] introduced a hybrid approach combining SSMs with hyperbolic spaces, demonstrating improved performance on language modeling tasks. Most recently, the MAMBA architecture [13] incorporated selective computation and showed competitive performance with Transformers while using less computation.

2.5.3 Mamba Architecture. The Mamba architecture represents a significant advancement in the field of sequence modeling, building upon the foundations of structured state space models (SSMs) while introducing novel elements to enhance performance and efficiency. At its core, Mamba incorporates a selective state space model, which addresses key limitations of previous SSM implementations.

The fundamental innovation in Mamba lies in 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 parameters of the SSM, namely Δ, B, and C, functions of the input. Mathematically, this can be expressed as:

$$\begin{aligned} B &: (\mathsf{B}, \mathsf{L}, \mathsf{N}) \leftarrow s_B(x) \\ C &: (\mathsf{B}, \mathsf{L}, \mathsf{N}) \leftarrow s_C(x) \\ \Delta &: (\mathsf{B}, \mathsf{L}, \mathsf{D}) \leftarrow \tau_\Delta(\mathsf{Parameter} + s_\Delta(x)) \end{aligned} \tag{4}$$

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.

This selection mechanism enables Mamba to overcome the limitations of linear time-invariant (LTI) models, which struggle with tasks requiring content-aware processing, such as selective copying or induction heads. The ability to selectively focus on relevant information allows Mamba to compress context into a smaller state more effectively, balancing the trade-off between efficiency and expressiveness.

To implement this selective mechanism efficiently, Mamba employs a hardware-aware algorithm that leverages the memory hierarchy of modern GPUs. This algorithm uses kernel fusion to combine the discretization step, the parallel scan operation, and the multiplication with C into a single operation. This approach significantly reduces memory bandwidth requirements, leading to substantial speedups compared to naive implementations.

The Mamba block, which forms the basic unit of computation in the architecture, can be described by the following series of operations:

$$\Delta, B, C = \text{Linear}(x) \tag{5}$$
$$h = \text{SelectiveScan}(\Delta, B, C, x) \tag{6}$$
$$y = \text{Linear}(h) \tag{7}$$

Here, the SelectiveScan operation is the core component that enables efficient state updates, incorporating the selective mechanism that gives Mamba its unique capabilities.

The overall Mamba architecture simplifies previous SSM architectures by combining the SSM block with the ubiquitous multi-layer perceptron (MLP) block found in modern neural networks. Instead of interleaving these two blocks, Mamba repeats a unified block homogeneously throughout the network. This unified block expands the model dimension D by a controllable expansion factor E, typically set to 2.

Compared to the H3 block, which forms the basis of many SSM architectures, Mamba replaces the first multiplicative gate with an activation function. In contrast to the standard MLP block, Mamba adds an SSM to the main branch. The activation function σ used in

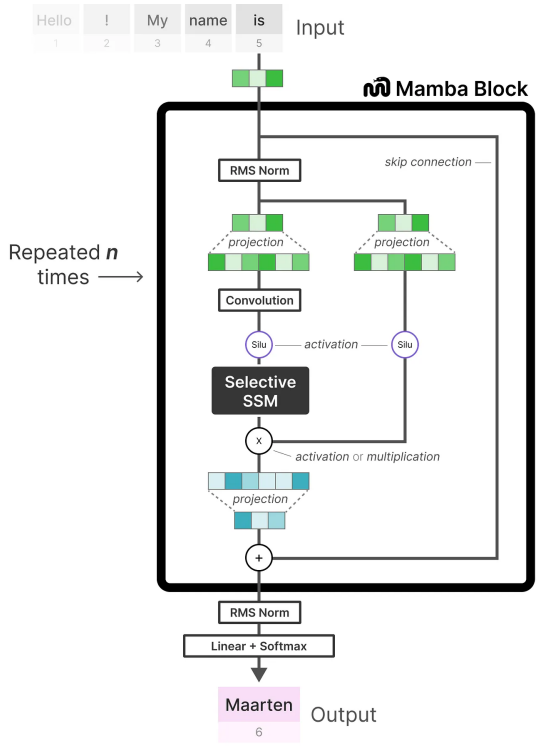


Figure 3: Detailed structure of a Mamba block, showing the flow of data through various components including the Selective SSM. Adapted from Grootendorst [12].

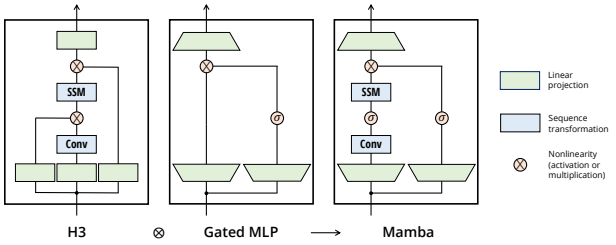


Figure 4: The Mamba block design combines the H3 block, which is the basis of most SSM architectures, with the ubiquitous MLP block of modern neural networks. For σ, the SiLU / Swish activation is used. Adapted from Gu and Dao [13].

the Mamba block is the SiLU (Sigmoid Linear Unit) or Swish function, which has shown promising results in various deep learning applications [15, 30].

The Mamba architecture's design choices result in a model that is not only more efficient in terms of computational resources but also more effective at capturing long-range dependencies in sequences. By leveraging the selective mechanism and the simplified block structure, Mamba can process longer sequences more efficiently than traditional Transformer models, while maintaining or even improving upon their performance across various tasks.

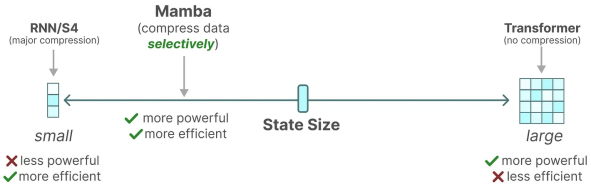


Figure 5: Comparison of Mamba, RNN/S4, and Transformer architectures in terms of their effectiveness at compressing data selectively based on state size. Adapted from Grootendorst [12].

Figure 5 provides a visual comparison of the Mamba architecture with RNN/S4 (major compression) and Transformer (no compression) architectures in terms of their effectiveness at compressing data selectively based on state size. As the image illustrates, Mamba strikes a balance between the two extremes, being more powerful and efficient than RNN/S4 for small state sizes while remaining more efficient than Transformers for large state sizes [12]. This comparison highlights a key strength of the Mamba architecture: its ability to efficiently compress and process sequential data in a content-aware manner. By selectively focusing on or ignoring specific inputs based on their content, Mamba can effectively balance the trade-off between model capacity and computational efficiency, enabling it to handle longer sequences and more complex tasks than traditional architectures.

2.6 Challenges in Molecular Generation

Despite advancements in molecular representations and sequence modeling architectures, several challenges persist in AI-driven molecular design. Balancing validity, diversity, and novelty in generated molecules remains complex, often requiring careful tuning of generation parameters [29]. Ensuring that generated molecules possess drug-like properties and are synthetically accessible is crucial for practical applications in drug discovery, as demonstrated by Brown et al. [3] in their benchmark suite for de novo molecular design.

Beyond de novo generation, sequence modeling architectures show promise in other aspects of molecular design. For example, Jin et al. [20] demonstrated the use of a hierarchical encoder-decoder model for targeted molecule optimization, achieving a 30% improvement over previous methods in finding molecules with desired properties while maintaining structural similarity. Similarly, Mo et al. [25] applied transformer-based models to predict reaction outcomes and retrosynthetic pathways, potentially aiding in the design of synthetically accessible molecules.

As researchers aim to explore larger chemical spaces and generate more complex molecules, computational efficiency becomes increasingly important. Gómez-Bombarelli et al. [10] highlighted this challenge, noting that the vast size of chemical space (estimated at 10^{60} drug-like molecules) necessitates highly efficient exploration strategies. The interpretability of AI-generated molecules and the ability to guide generation towards desired properties are ongoing areas of research that continue to drive innovation in the field [19].

3 METHODOLOGY

Our study aims to evaluate the efficacy of autoregressive sequence models in molecular generation tasks. Autoregressive models have shown promising results in various sequence modeling tasks, including natural language processing and, more recently, in the domain of cheminformatics [4, 26, 38]. In this work, we focus on comparing Transformer-based models, State Space Models (SSMs), and hybrid architectures, all implemented as autoregressive sequence models for molecular generation.

3.1 Dataset Preparation

To ensure a comprehensive analysis, we utilized two distinct datasets: the Molecular Sets (MOSES) dataset and a canonicalized subset of the ZINC database. The MOSES dataset, comprising approximately 1.6 million drug-like molecules, serves as our primary benchmark. Curated by Polykovskiy et al. [29], MOSES offers a representation of the chemical space relevant to drug discovery, with compounds selected based on specific physicochemical properties and synthetic accessibility criteria.

To complement MOSES and assess the scalability of our findings, we incorporated a larger dataset derived from ZINC20 [35]. 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.

For both datasets, we implemented an identical preprocessing pipeline. We transformed the original SMILES strings into the SAFE (Sequential Attachment-based Fragment Embedding) representation using the SAFE library[2]. Introduced by Noutahi et al. [26], SAFE represents molecules as an unordered sequence of interconnected fragment blocks. The SAFE encoding process involves extracting unique ring digits from the SMILES string, fragmenting the molecule using methods such as BRICS [6], sorting fragments by size, concatenating fragment SMILES strings, and replacing attachment points with new ring digits.

For tokenization, we employed the pre-trained byte-pair encoding (BPE) tokenizer from the SAFE-GPT model[3]. This tokenizer, trained on 1.1 billion molecules[4], offers a vocabulary size of 1,880 tokens. The use of this pre-trained tokenizer ensures consistency with the original SAFE-GPT implementation and leverages knowledge embedded in a larger chemical space [26].

For MOSES, we maintained the original train-validation split to ensure comparability with previous studies [29]. The ZINC subset was randomly split into training (90%) and validation (10%) sets. This approach allows us to evaluate our models on held-out molecules not seen during training, providing an assessment of generalization capabilities.

The final preprocessed datasets consisted of the MOSES dataset with approximately 1.6 million SAFE-encoded molecules (split into training and validation sets) and the ZINC subset with 23 million SAFE-encoded molecules (20.7 million for training, 2.3 million for validation). By utilizing these two datasets of different scales, we aim to provide an evaluation of our autoregressive sequence models across varying levels of molecular diversity and complexity.

3.2 Model Architectures and Training Procedure

Our comparative study implemented five distinct models across three architectures: Transformer-based (SAFE-Small and SAFE-Large), State Space Models (MAMBA-Small and MAMBA-Large), and a hybrid architecture (MAMBA-Small-Hybrid). These models were designed to investigate both small (approximately 20M parameters) and large (approximately 90M parameters) variants.

The SAFE-GPT models, as described by Noutahi et al. [26], served as our Transformer-based architectures. The MAMBA models were based on the architecture proposed by Gu and Dao [13] and adapted from their original codebase[5]. The MAMBA-Small-Hybrid model incorporated attention layers at indices 2 and 5 within its 6-layer structure, combining elements of both Transformer and SSM architectures.

Our implementation strategy prioritized alignment with the SAFE framework, carefully emulating the training process outlined in the SAFE library[6]. We made only necessary architectural adjustments while keeping all other aspects of the pipeline constant. This approach ensured that our comparison focused on architectural differences, isolating their impact on molecular generation performance.

We implemented a training protocol consistent across all model architectures, with specific adjustments made for the larger models to account for their increased capacity. All models were trained on NVIDIA A100 GPUs. The small models were trained for a full 10 epochs, while the large models were trained for a fixed number of 250,000 steps, corresponding to approximately 2.4 epochs on our dataset. For all models, we implemented interleaved validation throughout the training process to monitor performance and prevent overfitting.

Throughout the training process, we monitored GPU power consumption and utilization using Wandb, allowing us to compare the efficiency of each model architecture in terms of energy consumption and hardware utilization.

Detailed model architecture parameters and training hyperparameters can be found in Appendix B.

3.3 Molecule Generation and Evaluation

For molecule generation, we employed nucleus sampling as described by Holtzman et al. [17]. For SAFE models, we maintained the default Hugging Face decoding parameters: a temperature of 1.0, top-p of 1.0, and top-k of 50. The MAMBA models used identical parameters, except for top-p, which was adjusted to 0.9. This adjustment proved crucial for maintaining high validity rates in MAMBA-generated molecules, a point we will elaborate on in the results section. We generated 10,000 molecules for each model in a single batch, ensuring a consistent generation strategy across all architectures for fair comparison.

Our evaluation framework encompassed both quantitative measures and qualitative analyses, building upon established metrics in

[1] https://huggingface.co/datasets/sagawa/ZINC-canonicalized
[2] https://safe-docs.datamol.io/stable/
[3] https://huggingface.co/datamol-io/safe-gpt
[4] https://huggingface.co/datasets/datamol-io/safe-gpt
[5] https://github.com/state-spaces/mamba
[6] https://github.com/datamol-io/safe

the field of molecular generation [29]. We assessed validity, uniqueness, and diversity of the generated molecules. Validity, calculated using RDKit [23], ensures that generated structures adhere to basic chemical rules. Uniqueness assesses the model's ability to generate distinct molecular structures. Diversity, quantified using the average pairwise Tanimoto distance between molecules based on their ECFP4 fingerprint representations [32], measures the structural variety within the generated set.

The diversity of generated molecules was quantified using the following equation:

$$\text{Diversity} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (1 - T(m_i, m_j)) \tag{8}$$

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 captured the characteristics of drug-like molecules, we compared the distributions of key physicochemical properties between the generated molecules and the training set. These properties, crucial in drug discovery as outlined by Lipinski [24] and Veber et al. [39], include molecular weight, LogP, topological polar surface area (TPSA), number of rotatable bonds, hydrogen bond acceptors and donors, and aromatic rings. These properties play vital roles in determining a compound's drug-likeness.

In addition to these molecular metrics, we conducted a thorough assessment of computational resource utilization. We monitored GPU power consumption and utilization throughout the training process using Wandb. This allowed 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.

By employing this comprehensive evaluation framework, we aim to provide a thorough analysis of the generated molecules' quality and diversity, their relevance to drug discovery, as well as the computational efficiency of the different model architectures. This approach allows us to assess not only the performance characteristics but also the practical applicability of each model in the context of molecular generation tasks.

4 RESULTS

This section presents the findings from our comparative analysis of Transformer-based (SAFE) and State Space Model (Mamba) architectures for molecular generation using the SAFE representation. We report on model performance metrics, perplexity analysis, molecular property distributions, and computational efficiency for both small (~20M parameters) and large (~90M parameters) models.

4.1 Model Performance Metrics

Table 1 summarizes the key performance metrics for our models, alongside previously reported results for other molecular generation approaches [29].

All models in our study achieved high validity scores, with Mamba_Large, Safe_Small, Mamba_Small_Hybrid, and Mamba_Small reaching perfect validity (1.000). Safe_Large showed slightly lower but still excellent validity at 0.98. Uniqueness was 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 were comparable across our models, with Safe_Large achieving the highest score of 0.880, followed closely by Mamba_Large at 0.873. The small models showed slightly lower but still competitive diversity scores: Safe_Small at 0.864, Mamba_Small_Hybrid at 0.862, and Mamba_Small at 0.860.

Notably, for all Mamba models, we found it necessary to adjust the top-p parameter to 0.90 to achieve these results. When top-p was set to 1.0 for the Mamba models, they tended to generate invalid SAFE representations, leading to frequent decoding errors.

4.2 Perplexity Analysis

Figures 6 and 7 illustrate the perplexity of each model over the course of training epochs for small and large models, respectively, as measured on a held-out test set at intervals throughout training.

For the small models (approximately 20M parameters), both Mamba_Small and Mamba_Small_Hybrid exhibited consistently lower perplexity throughout the training process, converging to values around 1.4. In contrast, the Safe_Small model's perplexity remained higher, settling around 1.5.

The large models (Safe_Large with 87M parameters and Mamba_Large with 94M parameters) showed a similar trend, with Mamba_Large achieving noticeably lower perplexity than Safe_Large throughout the training process. The gap in perplexity between Mamba_Large and Safe_Large appears to be even more pronounced than in the small models.

4.3 Molecular Property Distributions

To assess how well our models captured the characteristics of drug-like molecules, we analyzed the distribution of various molecular properties for the generated compounds, following established evaluation approaches [3, 29]. For small models, we compared the distributions to the MOSES training dataset, while for large models, we compared them to the ZINC dataset. Figures 8 and 9 show the distributions of key molecular properties for small and large models, respectively.

For both small (Figure 8) and large (Figure 9) models, the distributions of molecular properties for generated molecules closely matched those of their respective training datasets (MOSES for small models, ZINC for large models). This trend was consistent across all evaluated properties, indicating that our models, regardless of their underlying architecture or size, successfully captured the distribution of physicochemical properties present in their training data.

Notably, the Mamba model distributions closely align with those of the SAFE 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. However, it's important to note that there are some slight differences in the distributions between SAFE and Mamba models, particularly for the large models. These differences can be largely attributed to the different top-p parameter used for the Mamba models (0.90) compared to the SAFE models (1.0).

Table 1: Performance comparison of molecular generation models

Model	Valid@10K↑	Unique@10K↑	Diversity↑
Safe_Large (87M)	0.98	**1**	0.880
Mamba_Large (94M)	**1**	**1**	0.873
Safe_Small (21M)	**1**	0.999	0.864
Mamba_Small_Hybrid (20M)	**1**	0.999	0.862
Mamba_Small (20M)	**1**	0.999	0.860
GSELFIES-GPT20M	**1**	0.999	**0.887**
GSELFIES-VAE	**1**	0.999	0.859
GMT-SELFIES	**1**	**1**	0.870
SELFIES-VAE	**1**	0.999	0.858
CharRNN	0.975	0.999	0.856
VAE	0.977	0.998	0.856
LatentGAN	0.897	0.997	0.857
LigGPT	0.900	0.999	0.871
JT-VAE	**1**	0.999	0.855

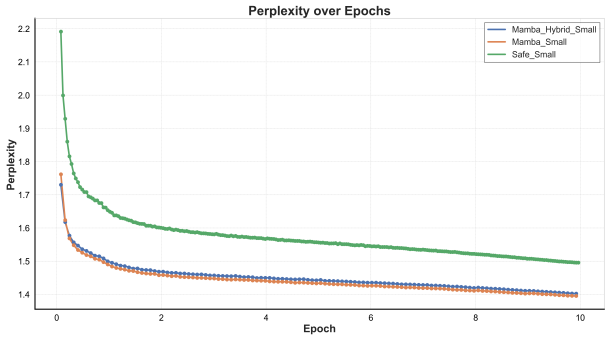


Figure 6: Perplexity over epochs for Safe_Small, Mamba_Small, and Mamba_Small_Hybrid models on test set

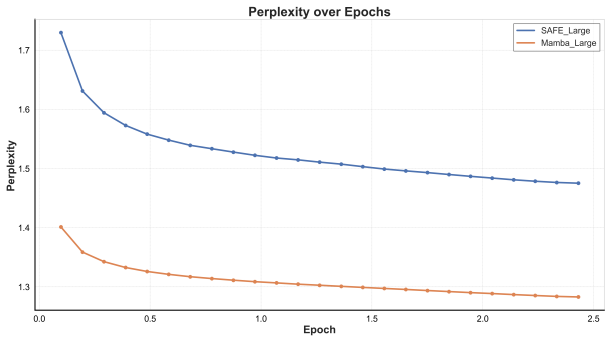
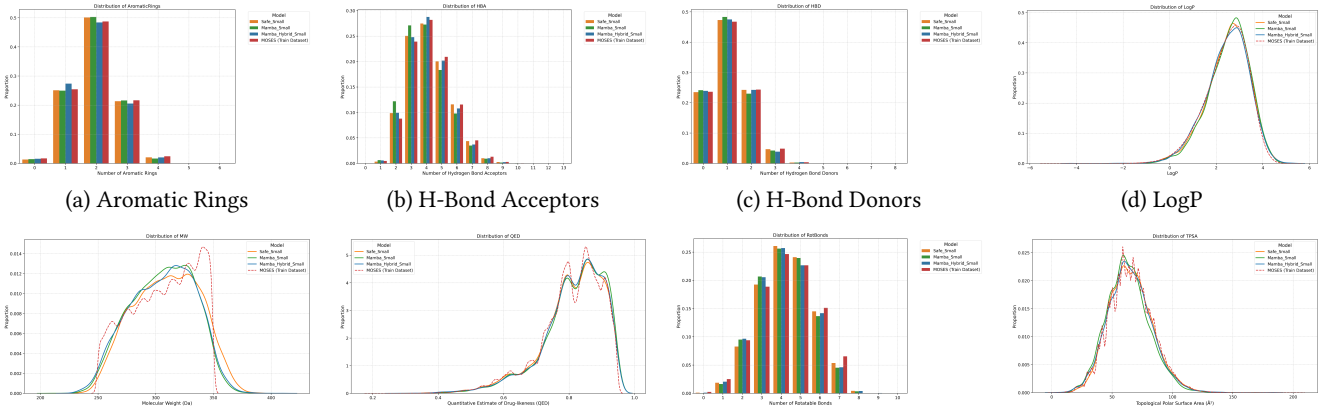


Figure 7: Perplexity over epochs for Safe_Large and Mamba_Large models on test set

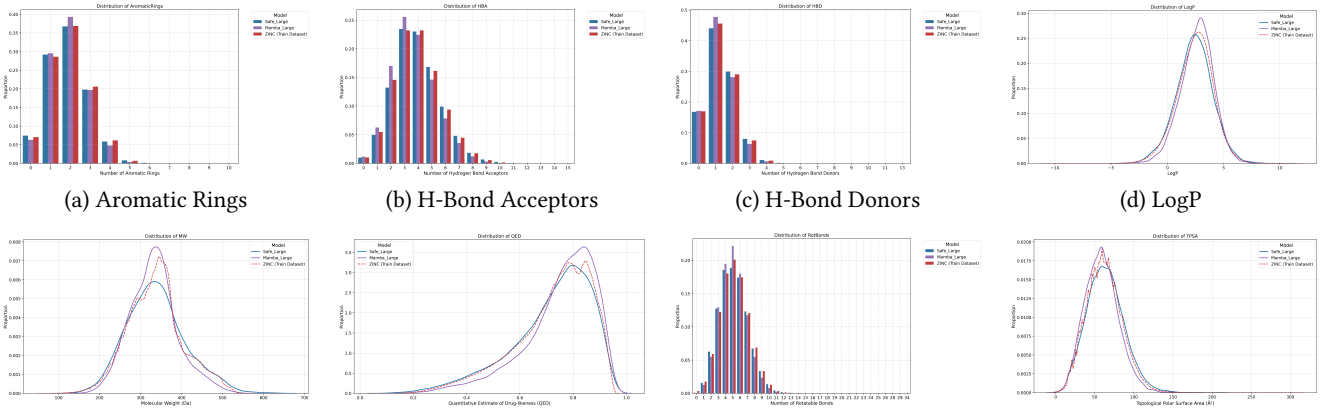


(a) Aromatic Rings (b) H-Bond Acceptors (c) H-Bond Donors (d) LogP

(e) Molecular Weight (f) QED (g) Rotatable Bonds (h) TPSA

Figure 8: Distributions of molecular properties for small models compared to MOSES dataset

We analyzed a comprehensive set of molecular properties, including Molecular Weight, LogP [41], Topological Polar Surface Area (TPSA) [8], Number of Rotatable Bonds [39], Hydrogen Bond Acceptors (HBA) and Donors (HBD) [9], Number of Aromatic Rings, and Quantitative Estimate of Drug-likeness (QED) [2]. The consistent trends across these properties further support our findings



(a) Aromatic Rings (b) H-Bond Acceptors (c) H-Bond Donors (d) LogP

(e) Molecular Weight (f) QED (g) Rotatable Bonds (h) TPSA

Figure 9: Distributions of molecular properties for large models compared to ZINC dataset

that both Transformer-based and State Space Model architectures effectively capture the property distributions of their respective training datasets.

For detailed mathematical definitions of these properties and their relation to drug-likeness, refer to Appendix A.

5 DISCUSSION

Our study aimed to investigate two primary questions: (1) how State Space Models compare to Transformer-based architectures in generating valid, unique, and diverse molecules using the SAFE representation, and (2) whether the efficiency of the MAMBA architecture provides advantages in terms of computational resources and training time when applied to larger datasets and model sizes. The results provide valuable insights into these questions and their implications for AI-driven molecular design.

5.1 Comparative Performance in Molecular Generation

Addressing our first research question, the results demonstrate a remarkable parity in performance between State Space Models (MAMBA) and Transformer-based (SAFE) architectures across all evaluated metrics for molecular generation.

Both small (20M parameters) and large (90M parameters) models achieved high validity (98-100%) and uniqueness (99.9-100%) scores, regardless of the underlying architecture. This parity extends the application of State Space Models, previously shown effective in language tasks, to the complex domain of molecular generation. The ability of MAMBA-based models to match the performance of Transformer-based models suggests that State Space Models can effectively learn and represent the intricate patterns inherent in molecular structures, even without explicit attention mechanisms.

The comparable diversity scores across all models further reinforce the capability of SSMs to capture the multifaceted nature of molecular structures. This finding is particularly significant as it demonstrates that alternative approaches to sequence modeling can be equally effective in exploring vast chemical spaces, a crucial aspect of molecular generation tasks.

It is important to note that achieving these results with MAMBA models required adjusting the top-p parameter to 0.90, while SAFE models used the default value of 1.0. This difference in sampling strategy hints at fundamental differences in how SSMs and Transformers learn to represent the molecular space, warranting further investigation.

5.2 Efficiency Advantages of MAMBA Architecture

Our second research question focused on the potential efficiency advantages of the MAMBA architecture. The results reveal significant efficiency gains for MAMBA-based models compared to the Transformer-based SAFE models, particularly as model size increases.

MAMBA models consistently demonstrated lower GPU power consumption compared to SAFE models. 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.

Training time comparisons revealed an interesting trend. While small MAMBA models required slightly longer training times despite their lower resource utilization, this trend reversed for large models. The 90M parameter SAFE model took approximately 90 hours to train for 250,000 steps, while the equivalent MAMBA model completed the same training in only 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.

The MAMBA-Hybrid model, incorporating both SSM and attention layers, demonstrated a promising balance between efficiency and training speed. It maintained the low resource utilization characteristic of MAMBA while achieving faster training times than the pure MAMBA model, approaching those of the SAFE model at the small scale.

5.3 Perplexity and Model Behavior

An intriguing finding of our study is the consistently lower perplexity exhibited by MAMBA and MAMBA-Hybrid models throughout the training process, for both small and large model sizes. The 20M parameter MAMBA models achieved a perplexity of 1.4 compared to 1.5 for the equivalent SAFE model, with this trend persisting for larger models as well (1.5 vs 1.3).

Lower perplexity suggests that MAMBA models have learned a more accurate 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.

However, the need for a lower top-p value during sampling with MAMBA models, despite their lower perplexity, highlights the complexity of interpreting model performance in molecular generation tasks. While MAMBA models seem to learn the molecular space more accurately, more selective sampling was required to maintain high validity. This suggests that when working with SSMs for molecular generation, fine-tuning of sampling strategies may be necessary to fully leverage the learned representations.

5.4 Molecular Property Distributions

Our analysis of molecular property distributions revealed that both MAMBA and SAFE models effectively captured the characteristics of their respective training datasets (MOSES for small models, ZINC for large models). This ability to reproduce the distribution of physicochemical properties such as molecular weight, LogP, and hydrogen bond acceptors/donors demonstrates that both architectures can learn and generate molecules with realistic and diverse properties.

The close alignment of property distributions between generated molecules and the training data, observed across all model architectures and sizes, further reinforces the capability of State Space Models to capture complex molecular features without explicit attention mechanisms.

5.5 Implications for AI-Driven Molecular Design

The comparable performance of MAMBA models to Transformer-based models, coupled with their efficiency advantages, has significant implications for the field of AI-driven molecular design. The potential for SSMs to handle longer molecular sequences efficiently opens up new possibilities for modeling complex macromolecules or entire chemical pathways, tasks that have traditionally been challenging due to the quadratic complexity of attention mechanisms.

These findings suggest that State Space Models, specifically the MAMBA architecture, offer a viable and efficient alternative to Transformer-based models for molecular generation tasks. The combination of comparable generation quality with improved computational efficiency positions SSMs as a promising approach for advancing the field of AI-driven molecular design, particularly for large-scale applications or when computational resources are limited.

6 CONCLUSIONS

Our study provides empirical validation for the efficacy of State Space Models, specifically the MAMBA architecture, in the complex task of 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 and MAMBA-Hybrid models in capturing the intricacies of molecular structures, as encoded in the SAFE representation, underscores the versatility of SSMs. This finding is particularly significant given the complexity of molecular generation tasks, which require models to learn and reproduce intricate patterns of atomic connections and chemical properties.

The marked efficiency advantage demonstrated by MAMBA-based models, evidenced by substantial reductions in GPU power consumption and improved training times for larger models, highlights a key strength of SSMs: their ability to process long sequences with linear time complexity. This characteristic could prove transformative in molecular generation tasks, where the exploration of vast chemical spaces is often constrained by computational resources.

Looking forward, several promising avenues for future research emerge from our findings:

(1) Further scaling studies: While we have already explored scaling to 90M parameters, investigating the performance of even larger SSM-based models on more extensive molecular datasets could further leverage their efficiency advantages and potentially uncover new capabilities.
(2) Application to extremely long molecules: Training MAMBA models on datasets containing exceptionally long molecular sequences could demonstrate their ability to capture dependencies at scales beyond the practical limits of Transformer models. This could open up new possibilities in modeling complex macromolecules or entire biochemical pathways.
(3) Advanced hybrid architectures: Building upon our MAMBA-Hybrid model, there is potential to develop more sophisticated hybrid architectures that combine the strengths of SSMs and attention mechanisms. These models could be tailored to capture different aspects of molecular structure and behavior more effectively.
(4) Integration with reinforcement learning: Developing approaches that guide MAMBA-based models to generate novel, valid molecules using reinforcement learning techniques could significantly contribute to drug discovery efforts. This could involve creating sophisticated reward functions that balance chemical validity, target properties, and synthetic accessibility.

In conclusion, our study not only validates the effectiveness of SSMs in the complex domain of molecular generation but also sets the stage for exciting future developments. The combination of comparable generation quality with improved computational efficiency positions SSMs as a promising approach for advancing the field of AI-driven molecular design. As researchers build upon these findings, we anticipate significant progress in our ability to explore and engineer molecular spaces.

REFERENCES

[1] Babak Alipanahi, Andrew Delong, Matthew Weirauch, and Brendan Frey. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature biotechnology* 33 (07 2015). https://doi.org/10.1038/nbt.3300

[2] G Richard Bickerton, Gaia V Paolini, J'er'emy Besnard, Sorel Muresan, and Andrew L Hopkins. 2012. Quantifying the chemical beauty of drugs. *Nature chemistry* 4, 2 (2012), 90–98.

[3] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. 2019. GuacaMol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* 59, 3 (2019), 1096–1108.

[4] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165* (2020).

[5] Tri Dao, Daniel Y. Fu, Khaled Kamal Saab, A. Waldmann Thomas, Atri Rudra, and Christopher Ré. 2022. Hungry Hungry Hippos: Towards Language Modeling with State Space Models. *ArXiv* abs/2212.14052 (2022). https://api.semanticscholar.org/CorpusID:255340454

[6] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. 2008. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem* 3, 10 (2008), 1503–1507. https://doi.org/10.1002/cmdc.200800178

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[8] Peter Ertl, Bernhard Rohde, and Paul Selzer. 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of medicinal chemistry* 43, 20 (2000), 3714–3717.

[9] Piero Gasparotto and Michele Ceriotti. 2014. Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond. *The Journal of chemical physics* 141, 17 (2014).

[10] 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. 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* 4, 2 (2018), 268–276.

[11] D Grechishnikova. 2021. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. Sci Rep 11: 321.

[12] Maarten Grootendorst. 2024. *A Visual Guide to Mamba and State Space Models.* https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mamba-and-state

[13] Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752 [cs.LG]

[14] Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021).

[15] Dan Hendrycks and Kevin Gimpel. 2023. Gaussian Error Linear Units (GELUs). arXiv:1606.08415 [cs.LG] https://arxiv.org/abs/1606.08415

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).

[18] Sabrina Jaeger, Simone Fulle, and Samo Turk. 2018. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *Journal of chemical information and modeling* 58 1 (2018), 27–35. https://api.semanticscholar.org/CorpusID:34512664

[19] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2018. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*. PMLR, 2323–2332.

[20] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2020. Hierarchical generation of molecular graphs using structural motifs. In *International Conference on Machine Learning*. PMLR, 4839–4848.

[21] John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (2021), 583 – 589. https://api.semanticscholar.org/CorpusID:235959867

[22] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020.

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

[24] Christopher A Lipinski. 2004. Lead-and drug-like compounds: the rule-of-five revolution. *Drug discovery today: Technologies* 1, 4 (2004), 337–341.

[25] Yiming Mo, Yanfei Guan, Pritha Verma, Jiang Guo, Mike E Fortunato, Zhaohong Lu, Connor W Coley, and Klavs F Jensen. 2021. Evaluating and clustering retrosynthesis pathways with learned strategy. *Chemical science* 12, 4 (2021), 1469–1478.

[26] Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan SC Lim, and Prudencio Tossou. 2023. Gotta be SAFE: A New Framework for Molecular Design. *arXiv preprint arXiv:2310.10773* (2023).

[27] Hakime Öztürk, Elif Ozkirimli Olmez, and Arzucan Özgür. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 34 (2018), i821 – i829. https://api.semanticscholar.org/CorpusID:13224164

[28] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery* 9, 3 (2010), 203–214.

[29] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. 2020. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. arXiv:1811.12823 [cs.LG]

[30] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2017. Searching for Activation Functions. arXiv:1710.05941 [cs.NE] https://arxiv.org/abs/1710.05941

[31] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. 2019. Evaluating Protein Transfer Learning with TAPE. arXiv:1906.08230 [cs.LG] https://arxiv.org/abs/1906.08230

[32] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* 50, 5 (2010), 742–754.

[33] Petra Schneider, W Patrick Walters, Alleyn T Plowright, Norman Sieroka, Jennifer Listgarten, Robert A Goodnow, Johanna Fisher, J"org M Jansen, Jos'e S Duca, Thomas S Rush, et al. 2020. Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery* 19, 5 (2020), 353–364.

[34] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. 2018. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science* 4, 1 (2018), 120–131.

[35] Teague Sterling and John J Irwin. 2015. ZINC 15–ligand discovery for everyone. *Journal of chemical information and modeling* 55, 11 (2015), 2324–2337.

[36] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. 2020. A deep learning approach to antibiotic discovery. *Cell* 180, 4 (2020), 688–702.

[37] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient Transformers: A Survey. arXiv:2009.06732 [cs.LG] https://arxiv.org/abs/2009.06732

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[39] Daniel F Veber, Stephen R Johnson, Hung-Yuan Cheng, Brian R Smith, Keith W Ward, and Kenneth D Kopple. 2002. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry* 45, 12 (2002), 2615–2623.

[40] David Weininger. 1988. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 1 (1988), 31–36.

[41] Scott A Wildman and Gordon M Crippen. 1999. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences* 39, 5 (1999), 868–873.

[42] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. 2019. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology* 37, 9 (2019), 1038–1040.

[43] Jian Zhou and Olga G. Troyanskaya. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods* 12 (2015), 931–934. https://api.semanticscholar.org/CorpusID:205424148

A MOLECULAR PROPERTY DEFINITIONS

This appendix provides mathematical definitions of the molecular properties analyzed in this study and their relation to the Quantitative Estimate of Drug-likeness (QED).

A.1 Quantitative Estimate of Drug-likeness (QED)

QED is a composite measure that combines several molecular properties to assess how drug-like a compound is [2]. It is calculated as the geometric mean of desirability functions for each property:

$$QED = exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln d_i\right) \tag{9}$$

where d_i are the desirability functions for each molecular descriptor.

A.2 Individual Property Definitions

A.2.1 Molecular Weight (MW). The sum of atomic weights for all atoms in a molecule.

$$MW = \sum_{i=1}^{n} atomic\_weight_i \tag{10}$$

A.2.2 Octanol-Water Partition Coefficient (LogP). A measure of lipophilicity, estimated using various methods such as Crippen's approach [41].

$$LogP = \sum_{i=1}^{n} a_i \cdot f_i \tag{11}$$

where a_i are atom contributions and f_i are correction factors.

A.2.3 Hydrogen Bond Acceptors (HBA) and Donors (HBD). Count of atoms capable of hydrogen bonding.

$$HBA = |\{a \in Atoms : a \text{ is O, N, or F with lone pair}\}| \tag{12}$$

$$HBD = |\{a \in Atoms : a \text{ is O-H or N-H}\}| \tag{13}$$

A.2.4 Topological Polar Surface Area (TPSA). Sum of surface contributions of polar atoms in a molecule [8].

$$TPSA = \sum_{i=1}^{n} contribution_i \tag{14}$$

A.2.5 Number of Rotatable Bonds. Count of single bonds, not in rings, bound to non-terminal heavy atoms.

$$N_{rotatable} = |\{b \in Bonds : b \text{ is single, not in ring, bound to non-terminal heavy atoms}\}| \tag{15}$$

A.2.6 Number of Aromatic Rings. Count of planar, conjugated ring systems with delocalized electrons.

$$N_{aromatic} = |\{c \in Cycles : c \text{ satisfies Hückel's rule and is planar}\}| \tag{16}$$

Each of these properties contributes to the overall drug-likeness of a molecule as captured by the QED metric. The desirability functions in QED transform these raw property values into scores between 0 and 1, which are then combined to give the final QED value.

B MODEL ARCHITECTURE AND TRAINING DETAILS

B.1 Model Architecture Parameters

Table 2 summarizes the key parameters of each model architecture used in our study.

Table 2: Model Architecture Parameters

Parameter	SAFE-Small	SAFE-Large	MAMBA-Small	MAMBA-Small-Hybrid	MAMBA-Large
Model Type	Transformer	Transformer	SSM	SSM + Attention	SSM
Embedding Dimension	512	768	512	512	768
Number of Layers	6	12	6	6	12
Attention Heads	8	12	-	8 (2 layers)	-
SSM Variant	-	-	Mamba2	Mamba2	Mamba2
Max Sequence Length	1024	1024	1024	1024	1024
Dropout Rate	0.1	0.1	0.1	0.1	0.1
Normalization	LayerNorm	LayerNorm	RMSNorm	RMSNorm	RMSNorm
Residual Connections	-	-	FP32	FP32	FP32
Rotary Embeddings	-	-	-	32-dim (2 layers)	-

B.2 Model Training Parameters

Table 3 summarizes the key training parameters for both small and large models.

Table 3: 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

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.

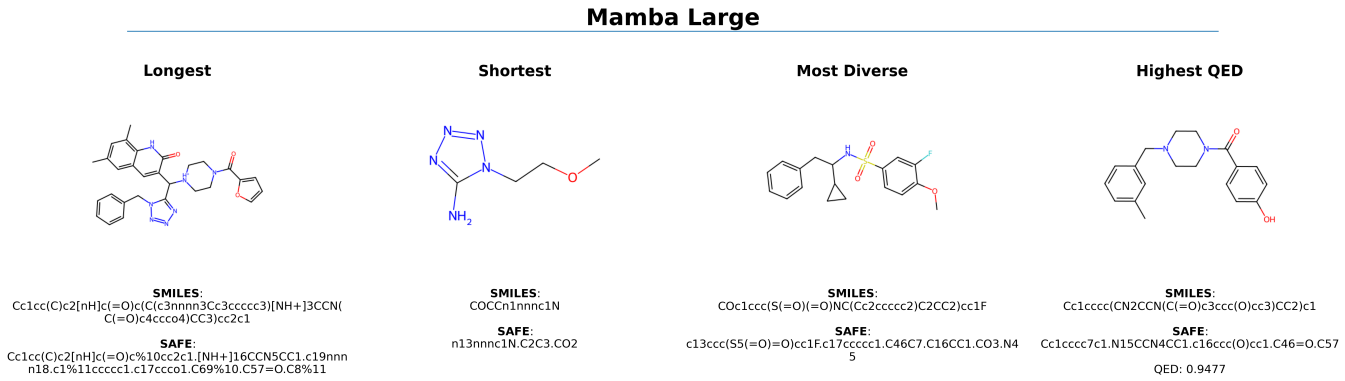


Figure 10: Representative molecules generated by the Mamba_Large model

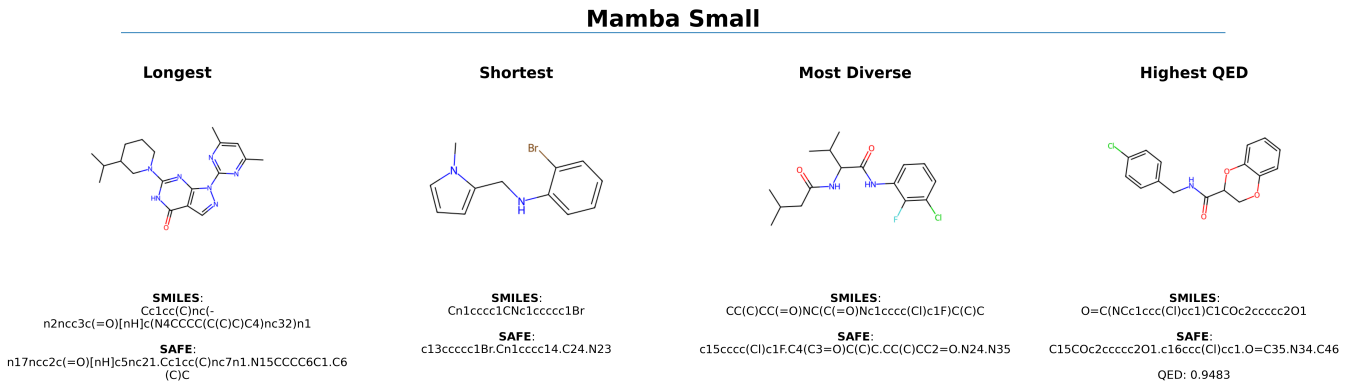
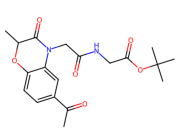


Figure 11: Representative molecules generated by the Mamba_Small model

Mamba Small Hybrid

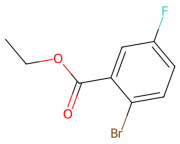
Longest



SMILES:
CC(=O)c1ccc2c(c1)N(CC(=O)NCC(=O)OC(C)(C)C)C(=O)C(C)O2

SAFE:
c17ccc2c(c1)N(CC4=O)C(=O)C(C)O2.C6C3=O.C5(C)(C)C.CC7=O.N46.O35

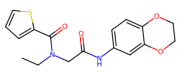
Shortest



SMILES:
CCOC(=O)c1cc(F)ccc1Br

SAFE:
c14cc(F)ccc1Br.C24=O.CC3.O23

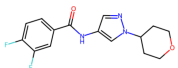
Most Diverse



SMILES:
CCN(CC(=O)Nc1ccc2c(c1)OCCO2)C(=O)c1cccs1

SAFE:
c17ccc2c(c1)OCCO2.c18cccs1.C6C3=O.N456.C48=O.CC5.N37

Highest QED



SMILES:
O=C(Nc1cnn(C2CCOCC2)c1)c1ccc(F)c(F)c1

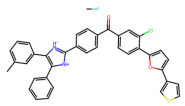
SAFE:
c15ccc(F)c(F)c1.c14cnn6c1.C16CCOCC1.O=C35.N34

QED: 0.9484

Figure 12: Representative molecules generated by the Mamba_Small_Hybrid model

SAFE Large

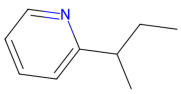
Longest



SMILES:
CF.Cc1cccc(-c2[nH+]c(-c3ccc(C(=O)c4ccc(-c5ccc(-c6ccsc6)o5)c(Cl)c4)cc3)[nH]c2-c2ccccc2)c1

SAFE:
c18ccc%11c(Cl)c1.Cc1cccc9c1.c19[nH+]c%10[nH]c1%13.c1%10ccc7cc1.c1%11ccc%12o1.c1%13ccccc1.c1%12ccsc1.C78=O.CF

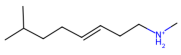
Shortest



SMILES:
CCC(C)c1ccccn1

SAFE:
c12ccccn1.CCC2C

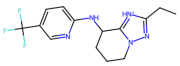
Most Diverse



SMILES:
C[NH2+]CCC=CCCC(C)C

SAFE:
C=2CCC(C)C.C1CC=2.C[NH2+]1

Highest QED



SMILES:
CCc1nn2c([nH+]1)C(Nc1ccc(C(F)(F)F)cn1)CCC2

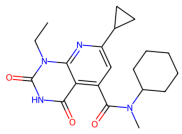
SAFE:
c14nn2c([nH+]1)C(N3)CCC2.c13ccc5cn1.C5(F)(F)F.CC4

QED: 0.9476

Figure 13: Representative molecules generated by the SAFE_Large model

SAFE Small

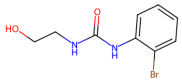
Longest



SMILES:
CCn1c(=O)[nH]c(=O)c2c(C(=O)N(C)C3CCCCC3)cc(C3CC3)nc21

SAFE:
n17c(=O)[nH]c(=O)c2c6cc8nc21.C15CCCCC1.C46=O.N45C.C18CC1.CC7

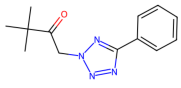
Shortest



SMILES:
O=C(NCCO)Nc1ccccc1Br

SAFE:
c15ccccc1Br.O=C23.C4CO.N24.N35

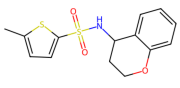
Most Diverse



SMILES:
CC(C)(C)C(=O)Cn1nnc(-c2ccccc2)n1

SAFE:
CC(C)(C)C(=O)C3.n13nnc4n1.c14ccccc1

Highest QED



SMILES:
Cc1ccc(S(=O)(=O)NC2CCOc3ccccc32)s1

SAFE:
N4C1CCOc2ccccc21.Cc1ccc(S4(=O)=O)s1

QED: 0.9483

Figure 14: Representative molecules generated by the SAFE_Small model

D GPU UTILIZATION

Table 4: Computational Efficiency Metrics

Model	GPU Utilization (%)	Power Consumption (W)
Safe_Small	60 ± 2	280
Mamba_Small	22 ± 1	190
Mamba_Small_Hybrid	23 ± 1	195
Safe_Large	95 ± 5	360
Mamba_Large	80 ± 15	280