NovoMolGen: Rethinking Molecular Language Model Pretraining

Kamran Chitsaz [*1 2] **Roshan Balaji** [*3 4 5] **Quentin Fournier** [1 2] **Nirav Pravinbhai Bhatt** [3 4 5 6]
Sarath Chandar [1 2 7 8]

Abstract

Designing *de novo* molecules with desired properties requires efficient exploration of an immense chemical space spanning 10^{23} to 10^{60} potential candidates. Although Molecular Large Language Models (Mol-LLMs) enable scalable exploration using string-based representations, the effects of language modeling practices such as tokenization, model size, and dataset scale on molecular generation performance remain unclear. In this study, we introduce NovoMolGen, a family of transformer-based foundation models pretrained on 1.5 billion molecules, to systematically investigate these key factors. Our analyses demonstrate a weak correlation between standard pretraining metrics and downstream molecular generation performance, highlighting critical differences compared to general NLP models. NovoMolGen achieves state-of-the-art results, outperforming prior Mol-LLMs and specialized generative models in both unconstrained and goal-directed molecule generation tasks.

1. Introduction

The discovery of new drugs for oncology, immunology, and rare or infectious diseases is challenging as exhaustive experimental screening is extremely time- and resource-intensive (Kirkpatrick & Ellis, 2004). Efficient computational strategies are thus essential to explore this vast space and identify novel synthesizable molecules with desired pharmacological properties. Recent progress in deep generative models has transformed molecular design by learning complex structure-property representations from large chemical databases, enabling automated *de novo* lead generation and optimization (Grisoni et al., 2020; Jin et al., 2020a; Podda et al., 2020; Mahmood et al., 2021; Hoogeboom et al., 2022). Notably, various molecular representations have been explored to facilitate *in silico* experimentation, including vector-based (Rogers & Hahn, 2010), graph-based (Lee et al., 2023; Yang et al., 2024), 3D structure-based (Xu et al., 2023; Huang et al., 2023; Zhang et al., 2023), and string-based approaches. Among those, string-based representations, such as SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988), DeepSMILES (O'Boyle & Dalke, 2018), SELFIES (Self-Referencing Embedded Strings) (Krenn et al., 2022), and SAFE (Sequential Attachment Fragment Embedding) (Noutahi et al., 2024), provide a scalable and computationally efficient solution. Their effectiveness is further supported by large-scale databases like GDB-13 (Ruddigkeit et al., 2012) and ZINC (Tingle et al., 2023), which contain billions of molecules in string format. Hence, deep generative models based on string representations have a huge potential to develop scalable solutions for generating diverse, chemically valid, and property-optimized molecules.

Building on this foundation, advancements in Molecular Language Models have further highlighted the effectiveness of string-based representations for automated molecule generation and optimization. For instance, REINVENT (Olivecrona et al., 2017) utilized a Recurrent Neural Network (RNN) to build molecules sequentially, atom-by-atom, in an unconstrained and unrestrained manner. The pretrained model was subsequently fine-tuned using reinforcement learning to generate compounds predicted to be active against a specific biological target. MolGPT (Bagal et al., 2022), inspired by GPT-2, employs an autoregressive decoder trained on SMILES strings to generate structurally valid molecules, treating molecular generation as a sequence prediction task while capturing both the syntax and semantics of SMILES. SMILES-GPT further enhances this approach by incorporating Byte Pair Encoding (BPE) and specialized embeddings, thereby improving representation learning for molecular generation (Adilov, 2021). BindGPT (Zholus et al., 2024) further advances autoregressive architectures, revealing that self-supervised pretraining

[*]Equal contribution [1]Mila – Quebec AI Institute [2]Chandar Research Lab [3]Wadhwani School of Data Science and AI, IIT Madras [4]BioSystems Engineering and Control Lab [5]The Centre for Integrative Biology and Systems medicinE (IBSE) [6]IIT Madras Zanzibar [7]Polytechnique Montréal [8]Canada CIFAR AI Chair. Correspondence to: Nirav Pravinbhai Bhatt <niravbhatt@iitm.ac.in>, Sarath Chandar <sarath.chandar@mila.quebec>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

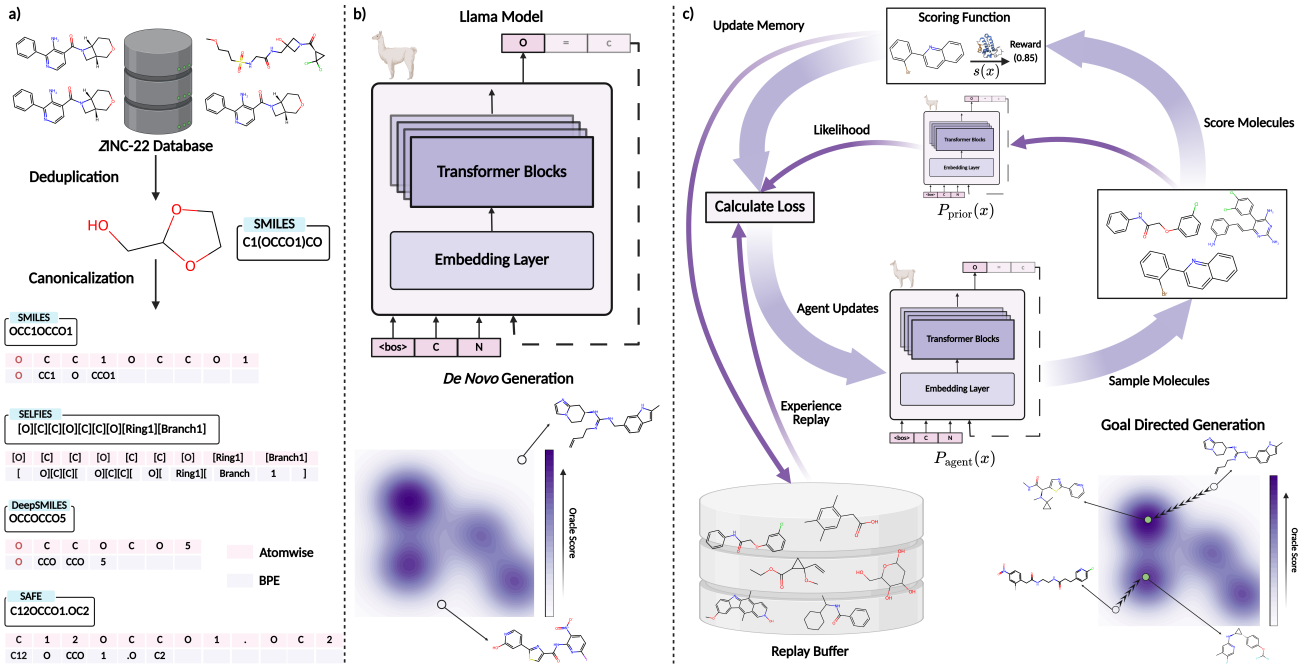


Figure 1. (a) SMILES deduplication, canonicalization, and conversion to SELFIES, SAFE, and DeepSMILES, with Atom-wise and Byte Pair Encoding (BPE) tokenization. (b) Pretraining of NovoMolGen and unconstrained molecular generation along the learned manifold. (c) Reinforcement learning-based fine-tuning for goal-directed molecular design.

enables transformers to implicitly encode spatial molecular features. Beyond autoregressive models, encoder-decoder architectures like BARTSmiles (Chilingaryan et al., 2022) pretrain molecular representations via denoising autoencoding, where the model reconstructs corrupted SMILES strings, enhancing robustness for downstream tasks. MolGen (Fang et al., 2023) further refines this paradigm by adopting SELFIES that guarantees chemical validity by design and self-feedback mechanisms, a reinforcement learning strategy where the model critiques and iteratively edits its outputs to optimize validity and diversity.

Scaling the pretraining of molecular language models has demonstrated strong potential for improving molecular generation and representation learning. SAFE-GPT (Noutahi et al., 2024) exemplifies this trend, training an 87M-parameter GPT-like architecture on 1.1B SAFE strings to improve fragment-based design, enabling scaffold decoration, motif extension, and linker generation. f-RAG (Lee et al., 2024) leverages the SAFE-GPT architecture in combination with a fragment injection module, which suggests additional fragments based on input fragments to complete and generate novel molecules. Building on SMILES-based pretraining, MoLFormer (Ross et al., 2022) employs linear attention and rotary embeddings to pretrain a transformer encoder on 1.1 billion SMILES. GP-MoLFormer (Ross et al., 2024) extends this approach to autoregressive generation, training a 46.8M-parameter decoder on 1.1 billion SMILES and investigating the trade-off between novelty and memorization at extreme scales.

While molecular language models draw inspiration from recent breakthroughs in natural language processing (NLP), directly applying LLM methodologies to molecular generation presents unique challenges. Small molecule representations impose fundamentally different constraints compared to natural languages, with shorter sequence lengths, smaller vocabularies, and highly structured syntax that affects how models capture chemical relationships. Despite progress in the development of string-specific individual models, a fundamental question has remained unanswered: *"How to tailor and optimize language modeling techniques to the unique specificities of small molecules?"*. Frey et al. (2023) examine neural scaling behavior in large chemical models by varying model and dataset sizes; however, their analysis is limited to pretraining loss and does not extend to downstream molecular optimization tasks, which are more indicative of practical performance. Yu et al. (2024) explore molecular generation in the context of a broader chemistry-focused AI assistant, but rely primarily on basic metrics such as validity and fingerprint Tanimoto similarity. Similarly, Özçelik & Grisoni (2024) employ a considerably smaller dataset (∼1.5M molecules) and focus on evaluation metrics such as Fréchet ChemNet Distance (FCD), which,

while informative, offer a narrower assessment of generation quality. So far, no systematic study has examined how architectural decisions and training protocols influence the validity, diversity, and property optimizability of generated molecules. Furthermore, while pretraining improves molecular representations, its impact on downstream task performance remains poorly understood. This highlights the need to investigate how the different aspects of the Mol-LLMs pipeline affect both pretraining efficiency and fine-tuning effectiveness, ensuring that Mol-LLMs generalize well to real-world molecular design challenges.

To bridge these gaps and address open questions in Mol-LLM design, we make the following key contributions:

- We conduct the largest systematic study to date on pretraining molecular foundation models, simultaneously evaluating how molecular representations, tokenization strategies, model scaling, and dataset size influence Mol-LLM performance in *de novo* molecular generation.
- We introduce **NovoMolGen**, a family of transformer-based models pretrained on 1.5 billion molecules, achieving state-of-the-art performance in both unconstrained generation and goal-directed molecular design.
- We investigate the impact of model scaling, finding that while increasing model size shows some trends in goal-directed generation tasks, it does not lead to consistent improvements across all metrics.
- We systematically compare SMILES, SELFIES, SAFE, and DeepSMILES molecular representations with atom-wise and BPE tokenization, revealing that no single representation is universally optimal but that representation choices significantly impact generalization and downstream task performance.
- We find that pretraining efficiency reaches an early plateau, indicating that extended training does not always enhance performance. Surprisingly, in fine-tuning, even our earliest checkpoints outperform strong baselines, demonstrating that pretrained models already capture essential molecular properties.

To facilitate reproducibility and further research, we open-source our models, datasets, and code, establishing a comprehensive benchmark for advancing large-scale pretraining in molecular generation.

2. Methodology

This section details the molecule generation process using Mol-LLMs, which consists of four stages, as illustrated in Figure 1. The process begins with data preprocessing and preparation (§2.1), where molecular datasets are constructed using various representations, including SMILES, DeepSMILES, SELFIES, and SAFE. This is followed by pretraining (§2.2) using transformer models based on the Llama architecture, ranging from 32M to 300M parameters. Subsequently, fine-tuning (§2.3) is performed for task-specific molecule generation, leveraging an oracle to compute rewards. Finally, the models undergo generation and evaluation (§3.1) to assess performance across key metrics relevant to drug discovery.

2.1. Data Preparation

For our experiments, we utilize the ZINC-22 database, the largest publicly available molecular library, encompassing approximately 70 billion synthesizable and commercially available compounds as of September 8, 2024. This database provides extensive coverage of purchasable chemical space, making it a valuable resource for large-scale molecular modeling and generative tasks. The molecules are encoded in the SMILES format and organized based on their heavy atom count, which ranges from 4 to 49. To maintain diversity during pretraining, we employ a random stratified sampling strategy based on heavy atom count to select 1.5 billion molecules from the database. Additionally, to ensure uniqueness and consistency across the dataset, we remove duplicate entries and canonicalize the SMILES strings in the molecular database. To the best of our knowledge, this represents the largest dataset used to date for pretraining in *de novo* molecule generation.

To explore the effect of molecular representation on generation quality, we also transformed the molecules from SMILES into other textual representations, including SELFIES, SAFE, and Deep SMILES. We ensured the dataset's diversity in terms of molecular length and structural similarity, with the average Tanimoto similarity between molecules in a batch kept below 0.5. Details of this analysis, including molecular diversity metrics and visualizations, are provided in Appendix D.

To enable a comprehensive evaluation of the generalization capabilities of our models, we constructed two distinct validation sets. First, a scaffold-based validation set was generated using Bemis-Murcko scaffolds (Bemis & Murcko, 1996), which define core molecular frameworks. This split ensures that scaffolds in the validation set do not appear in the training set, thereby enabling a stringent test of generalization to novel chemical frameworks. This practice is consistent with established protocols in molecular machine learning (Wu et al., 2018; Polykovskiy et al., 2020) and reflects the demands of real-world applications such as lead optimization in drug discovery. Although some structurally similar molecules may fall under different scaffolds, more precise alternatives such as Butina or UMAP-based clustering (Guo et al., 2024) require computing molecular similarities at a scale that is currently infeasible for our dataset.

Following the creation of the scaffold-based set, we con-

structed a second validation set by randomly sampling 10 million molecules from the remaining training data. This random subset captures a chemically diverse yet structurally familiar distribution, allowing us to assess model performance in a setting that more closely resembles the training domain. Together, the scaffold-based and random validation sets provide complementary perspectives on generalization.

2.2. Pretraining

Pretraining begins with an autoregressive decoder-only model, a natural choice for *de novo* molecule generation due to its ability to model the sequential dependencies in representations such as SMILES and SELFIES. These models are trained to predict the next token in a sequence, enabling molecules to be constructed incrementally in a left-to-right manner. This self-supervision encourages the model to learn the syntactic structure and underlying chemical validity of string-based molecular representations during training, enabling the generation of diverse and plausible molecules while efficiently capturing local and global dependencies within large-scale datasets.

In this study, we adopted the Llama architecture (Dubey et al., 2024) for its popularity in large-scale generative tasks. To further explore the impact of molecular representation on generative performance, we employed atom-wise and byte-pair encoding (BPE) tokenization strategies. The atom-wise tokenization (Schwaller et al., 2019) uses a regular expression to split SMILES strings into meaningful units, including atoms, bonds, and functional groups, ensuring the preservation of stereochemistry and bond order. For the BPE tokenization (Kudo, 2018), we trained a tokenizer with a maximum vocabulary size of 500 and a dropout of 0.1 using a subset of 100 million molecules from the training dataset. This approach captures common substructures and recurring patterns in molecular strings, enabling efficient representation of complex chemical structures while balancing granularity and scalability.

Our experiments utilized three model sizes, 32M, 157M, and 300M parameters, to investigate the relationship between model capacity and generative quality. The architectural configurations for each model are summarized in Appendix Table E2. These configurations were selected based on the optimal width-to-depth ratio recommended by Levine et al. (2020), which supports efficient scaling of transformer models. The training was conducted using the FlashAttention library (Dao et al., 2024) integrated within the HuggingFace (Wolf et al., 2019) Trainer framework, enabling efficient handling of large batch sizes and significantly improving training speed.

Each model was trained on 1.5 billion molecules from the ZINC-22 dataset, employing a next-token prediction objective to capture the sequential dependencies within the chemical space. The training process maintained a fixed global batch size of 19,200 molecules (1.2M tokens per gradient step) across 4xA100 GPUs, with optimization performed using the AdamW optimizer and a cosine learning rate scheduler, reaching a peak learning rate of 6×10^{-4}. Additional training details are provided in the Appendix E.

2.3. Fine-Tuning

Fine-tuning in *de novo* molecule generation is crucial for optimizing pretrained models to generate molecules with specific, desirable properties. While pretraining captures general chemical properties and ensures the generation of valid, diverse molecules, fine-tuning is essential to align the model's outputs with specific goals, such as drug-likeness or bioactivity. Reinforcement learning (RL) plays a central role in this process, providing an effective framework for optimizing molecule generation under task-specific constraints. These tasks often face practical limitations, such as a restricted budget of a few thousand oracle calls, making sample efficiency a critical factor in achieving high-quality results.

We found that the REINVENT fine-tuning pipeline (Olivecrona et al., 2017), which utilizes a straightforward reinforcement learning framework for molecular optimization, performs surprisingly well despite its simplicity. Given its effectiveness and ease of integration, we adopted it in our approach, while recognizing that more sophisticated or targeted methods may further enhance performance in future work. The framework includes a fixed pretrained model, $P_{\text{prior}}(x)$, to preserve chemical validity and syntax, and a trainable agent model, $P_{\text{agent}}(x)$, which is optimized to maximize a task-specific reward function, $s(x)$. The primary objective is to minimize the cost function, $J(x)$, defined as:

$$J(x) = \left[\log P_{\text{prior}}(x) - \log P_{\text{agent}}(x) + \sigma \cdot s(x)\right]^2 \quad (1)$$

where σ is a scaling factor that controls the influence of the reward function. This loss balances the dual objectives of adhering to the prior distribution and optimizing for the reward, ensuring that the agent generates molecules that are both valid and aligned with task-specific goals.

Additionally, we use a penalty term, $J_p(x) = \frac{-1}{\log P_{\text{agent}}(x)}$, to discourage the generation of molecules with extremely low likelihoods under the agent's learned distribution. This regularization prevents degenerate solutions, promotes confidence in the agent's predictions, and enhances training stability. The final objective combines these components:

$$J_\theta = J + \lambda \cdot J_p, \quad (2)$$

where λ is a hyperparameter that weights the penalty term. Extensive hyperparameter tuning revealed the critical importance of λ, with its optimal range being significantly

different from that proposed in REINVENT, reflecting the specific requirements of our molecular generation task.

To improve sample efficiency and ensure broader exploration of the chemical space, we incorporate an experience replay mechanism with a molecular memory buffer. The memory buffer stores top-performing molecules, which are periodically revisited and used to reinforce high-quality samples during training. By combining newly generated high-reward candidates with stored experiences, the model avoids overfitting to narrow solutions and ensures efficient utilization of the oracle budget. Further implementation details, including hyperparameter configurations, are provided in Appendix J.

3. Experiments

Our study systematically investigates the impact of four key factors: model size (§3.2), molecule representation (§3.3), tokenization strategy (§3.4), and pretraining data scale (§3.5). To evaluate downstream effectiveness, we fine-tune the models using reinforcement learning for goal-directed molecular design and compare their performance against state-of-the-art baselines.

For each experiment, we tracked the training and validation loss on two validation sets: one created via random sampling and the other using scaffold-based splitting. The final checkpoint used for evaluation was taken at 75,000 steps, corresponding to a full epoch over the training dataset.

3.1. Generation and Evaluation

We evaluated the pretrained models by generating 30,000 molecules using temperature sampling ($T = 1.0$) without top-k or top-p filtering, as shown in Table 1. Generated molecules were assessed using several metrics: Validity, defined as the proportion of chemically plausible structures based on RDKit parsing; Novelty, measuring the fraction of molecules absent from the complete 1.5-billion-molecule training set using canonical SMILES comparison; and Internal Diversity (IntDiv), which quantifies structural diversity within the generated set. We used the Fréchet ChemNet Distance (FCD) to evaluate distributional similarity to drug-like molecules, comparing learned representations of generated and reference molecules. Additionally, we computed Fragment Similarity and Scaffold Similarity to quantify the alignment of fragment distributions based on BRICS fragments and Bemis–Murcko scaffolds, respectively. Similarity to Nearest Neighbor (SNN) quantifies how closely each generated molecule resembles the nearest molecule in the reference set, reflecting the model's capacity to explore novel regions of chemical space. Metrics requiring a reference distribution (e.g., FCD, SNN) were computed using a random subset of 175,000 molecules from the validation dataset. The novelty metric was excluded for all baseline models except GP-MoLFormer, as prior work typically computes novelty with respect to the MOSES dataset rather than the whole training set used in our evaluation, leading to inconsistent comparisons. Most baselines were trained on the smaller MOSES dataset ($\sim$1.5M molecules). In contrast, our models were trained on the significantly larger ZINC dataset (1.5B molecules), affecting the comparability of novelty, SNN, and Scaffold metrics. Accordingly, we report metrics relative to training–validation overlap (*Train*) for consistent evaluation. Detailed descriptions of the evaluation metrics are provided in Appendix G, with additional results available in Appendix H.1.

3.2. Impact of Model Size

To investigate the impact of model size on the performance of pretrained molecule language models, we evaluated models with 32M, 157M, and 300M parameters using atom-wise tokenization and compared their results to the GP-MoLFormer baseline, which shares a similar training pipeline to our method. Unlike GP-MoLFormer's linear attention, our methodology adopts full self-attention mechanisms with a broader systematic study of tokenization and molecular representations and model size to assess their role in *de novo* molecule generation. Our results show comparable performance across most metrics, with high Validity, Fragment, and FCD scores. Notably, our models outperform GP-MoLFormer significantly in Novelty, demonstrating their ability to explore unexplored regions of chemical space.

However, increasing model size from 32M to 300M did not yield significant improvements. Metrics such as SNN, FCD, and Frag remained stable across model sizes, suggesting that smaller models already capture essential chemical properties. These findings indicate that while larger models may offer marginal benefits, smaller models, such as the 32M variant, achieve similar performance with significantly lower computational cost, making them a practical choice for molecule generation.

3.3. Impact of Molecular Representation

To investigate the effect of different molecular representations, we pretrained 32M models on SMILES, SELFIES, SAFE, and DeepSMILES using atom-wise tokenization and evaluated their performance. The results show no significant overall advantage for any specific representation, as each excels in certain metrics while underperforming in others. SELFIES and SAFE inherently enforce chemical validity by design, ensuring that all generated molecules are valid, which explains their perfect Validity scores. However, this strict structure may also limit diversity, as seen in their lower Scaffold Similarity and SNN scores. SAFE achieves the

Table 1. Comprehensive performance metrics for baseline models and **NovoMolGen**. The baselines for CharRNN, VAE, and JT-VAE are sourced from Polykovskiy et al. (2020), while the results for LIMO, MolGen-7B, and GP-Molformer are taken from Ross et al. (2024). **Blue** = best, **Pink** = second-best.

Representation	Tokenization	Model Size	Valid (↑)	IntDiv (↑)	Novelty (↑)	FCD (↓)	SNN (↑)	Frag (↑)	Scaff (↑)
Train	–	–	1.000	1.000	0.0	0.0376	0.4657	0.9999	0.5144
CharRNN (Polykovskiy et al., 2020)	Atom-wise	–	0.9748	**0.856**	–	0.0732	**0.6015**	**0.9998**	**0.9242**
VAE (Polykovskiy et al., 2020)	Atom-wise	–	0.9767	0.855	–	0.0990	**0.6257**	0.9994	**0.9386**
JT-VAE (Jin et al., 2018)	Atom-wise	–	**1.000**	0.855	–	0.3954	0.5477	0.9965	0.8964
LIMO (Eckmann et al., 2022)	Atom-wise	–	**1.000**	0.854	–	26.78	0.2464	0.6989	0.0079
MolGen-7B (Fang et al., 2023)	Atom-wise	7B	**1.000**	0.855	–	0.0435	0.5138	**0.9999**	0.6538
GP-Molformer (Ross et al., 2024)	Atom-wise	46.8M	**1.000**	**0.865**	0.390	0.0591	0.5045	**0.9998**	0.7383
NovoMolGen-SMILES	Atom-wise	32M	**0.999**	0.851	**0.983**	0.0394	0.4657	**0.9999**	0.5309
		157M	**0.999**	0.851	0.981	0.0426	0.4656	**0.9999**	0.5200
		300M	**0.999**	0.851	0.981	0.0419	0.4657	**0.9999**	0.5219
	BPE	32M	**0.999**	0.851	**0.982**	**0.0384**	0.4654	**0.9999**	0.5182
		157M	**0.999**	0.852	0.980	**0.0384**	0.4655	**0.9999**	0.5193
		300M	**0.999**	0.851	0.979	**0.0380**	0.4657	**0.9999**	0.5173
NovoMolGen-SELFIES	Atom-wise	32M	**1.000**	0.852	**0.982**	0.0799	0.4651	0.9996	0.4765
	BPE	32M	**1.000**	0.851	**0.984**	0.0386	0.4649	**0.9999**	0.5105
NovoMolGen-DeepSMILES	Atom-wise	32M	**0.999**	0.851	0.981	0.0400	0.4661	**0.9999**	0.5218
	BPE	32M	0.997	0.851	**0.982**	**0.0380**	0.4655	**0.9999**	0.5165
NovoMolGen-SAFE	Atom-wise	32M	0.957	0.853	0.953	0.9475	0.4562	0.9955	0.4521
	BPE	32M	0.997	0.852	0.976	0.3283	0.4599	0.9972	0.4913

highest Novelty, outperforming other representations, yet performs poorly in FCD, indicating weaker alignment with the training distribution.

While our models generally outperform GP-MoLFormer, it is important to consider the differences in the pretraining datasets. GP-MoLFormer was trained on a combination of ZINC and PubChem, whereas our models were trained solely on ZINC. Although ZINC is one of the largest molecular databases and emphasizes drug-like and synthesizable compounds, PubChem offers a broader and more diverse chemical space. Specifically, PubChem includes a wider variety of exotic moieties, natural products, and compounds with annotated bioactivity data that are largely absent in ZINC. This difference in training data likely contributes to GP-MoLFormer's higher IntDiv scores, despite its lower Novelty. These findings suggest that while molecular representations like SAFE and SELFIES bring specific strengths, there is no one-size-fits-all representation. The choice of representation should align with task-specific goals, and further exploration of hybrid or ensemble approaches may help balance the trade-offs observed across metrics.

3.4. Impact of Tokenization

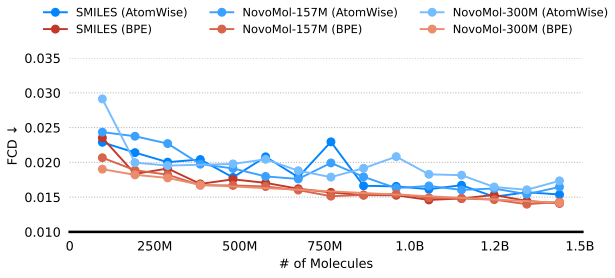
We evaluated the impact of the two most common tokenization strategies, atom-wise and BPE, across different molecular representations. Atom-wise results are depicted in blue shades, while BPE results are in red shades. Despite theoretical differences between the two approaches, such as atom-wise tokenization's interpretability and chemical relevance compared to BPE's larger vocabulary and potential for better generalization, no significant trends emerge in the results across these metrics.

These findings highlight that the choice of tokenization strategy may have minimal impact on overall performance and should instead be guided by practical considerations, such as the specific task requirements or efficiency during training. A similar comparison for different model sizes is provided in the appendix for completeness.

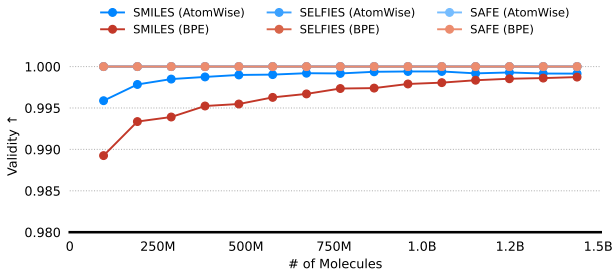
3.5. Progression of Metrics During Pretraining

A central question in Mol-LLM training is whether performance saturates over time and how well training loss correlates with generative quality. To investigate this, we evaluated intermediate checkpoints using multiple metrics, tracking the progression of FCD and Validity during training. Figure 2(a) presents FCD trends for different model sizes (32M, 157M, 300M) and tokenization strategies (atom-wise in blue, BPE in red), while Figure 2(b) shows Validity scores across molecular representations. In both plots, the x-axis corresponds to the number of unique molecules seen during training.

FCD improves modestly over time, with consistent trends across model sizes and tokenizations. Validity remains high throughout and converges to near-perfect values for all molecular formats. These results indicate that performance stabilizes early, with minimal gains from continued training. Moreover, variations in model size, tokenization strategy, and representation format have limited impact un-



(a)



(b)

Figure 2. (a) FCD during training for different model sizes. (b) Validity during training for different molecular representations. Both use atom-wise (blue) and BPE (red) tokenization. The x-axis shows the number of molecules seen during training.

der our experimental conditions. The results suggest that pretrained models quickly learn the core structure of chemical space, and further training primarily refines rather than significantly enhances generative capabilities.

3.6. Goal-Directed Molecular Optimization: PMO Benchmark

To evaluate the effectiveness of NovoMolGen in goal-directed molecular generation, we benchmark our models on the Practical Molecular Optimization (PMO) benchmark (Gao et al., 2022), which evaluates sample-efficient molecular optimization across diverse tasks. PMO includes molecular optimization challenges relevant to drug discovery, encompassing physicochemical properties, biological activity, and multi-property objectives. Each task employs an oracle function to score molecules based on predefined criteria, with a fixed budget of 10,000 function evaluations to ensure a fair comparison.

For each PMO task, we fine-tuned our pretrained models using the simple reinforcement learning-based fine-tuning approach described in Section 2.3. The details of our hyperparameter search procedure are provided in Appendix J. To ensure a fair comparison, we first performed an exhaustive hyperparameter search using the Perindopril_MPO and Zaleplon_MPO tasks, optimizing hyperparameters across three different random seeds per setting. The best hyperparameter configurations were then applied to fine-tune all PMO tasks. We evaluate different models with 32M, 157M, and 300M parameters using both atom-wise and BPE tokenization to assess their impact on optimization performance. For benchmarking, we compare against REINVENT (Olivecrona et al., 2017), which demonstrated the best performance as reported by Gao et al. (2022), and f-RAG (Lee et al., 2024), the current state-of-the-art on PMO benchmark.

In Figure 3, we present the results for NovoMolGen across different model size and tokenization strategies. The heatmap on the left visualizes the performance scores of each model across multiple PMO benchmark tasks, while the bar chart on the right provides the overall aggregate score for each model. Higher values indicate better performance in goal-directed molecular generation. NovoMolGen consistently outperforms both baselines, achieving substantial improvements over REINVENT and surpassing f-RAG across most tasks. Moreover, we observe a clear scaling trend: as model size increases, the total PMO score improves. However, this trend saturates at 300M parameters, where the performance gain becomes marginal. This aligns with our findings in Section 3.2, where all model sizes performed similarly in capturing molecular properties from the training data. Interestingly, as shown in Figure 4, when fine-tuning a small model (32M, atom-wise) with its best hyperparameters, we observe minimal performance improvement across training checkpoints, with even the earliest checkpoint surpassing both REINVENT and f-RAG.

4. Discussion

Our empirical analyses suggest that molecular language models (Mol-LLMs) follow training dynamics that are substantially different from those observed in natural language models. In particular, while the NLP community frequently benefits from large-scale model scaling and extended training times, our intermediate checkpoint evaluations reveal that performance improvements in molecular generative tasks can saturate quickly. Metrics such as FCD and Validity converge to near-optimal values at an early stage of pretraining, and further training yields only minor gains. Strikingly, when fine-tuning on the PMO benchmark, we find that even the earliest checkpoint of our smallest model (32M parameters), pretrained on approximately 100 million molecules and fine-tuned with optimal hyperparameters, already surpasses strong baselines such as REINVENT and f-RAG. This result underscores that extended training, while crucial in some generative paradigms, may not be the driving factor in boosting performance for molecular generation tasks.

These findings highlight that the current landscape of molecular datasets presents key limitations for self-supervised pretraining. Unlike genomic or proteomic sequences, where

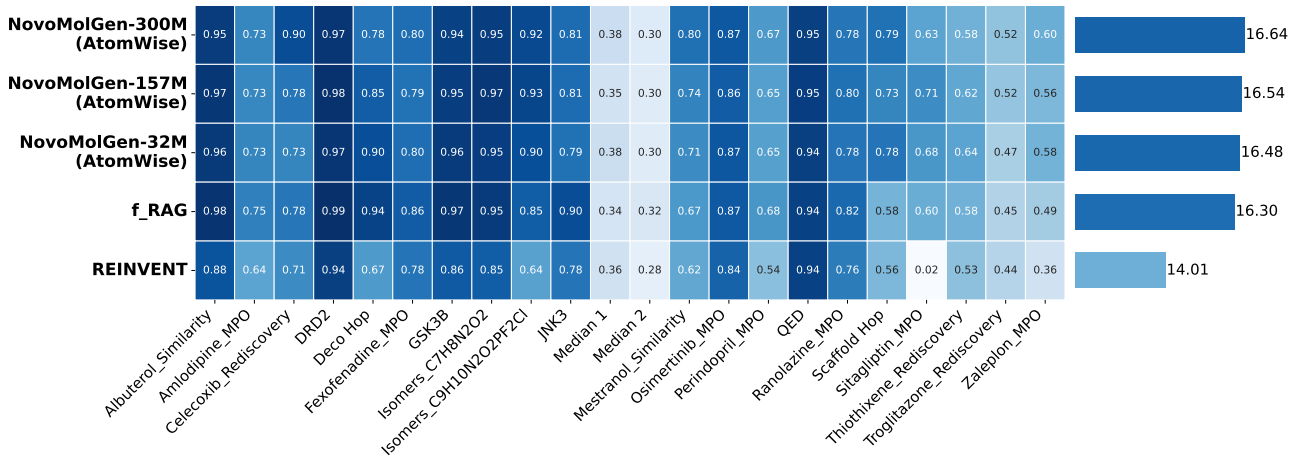


Figure 3. PMO benchmark results for NovoMolGen across different model sizes. The heatmap (left) displays scores for each task, while the bar chart (right) shows total scores, where higher values indicate better performance. **NovoMolGen-300M (Atom-wise)** achieves the highest overall score, outperforming other model variants. Results for REINVENT and f-RAG are taken directly from their respective publications.

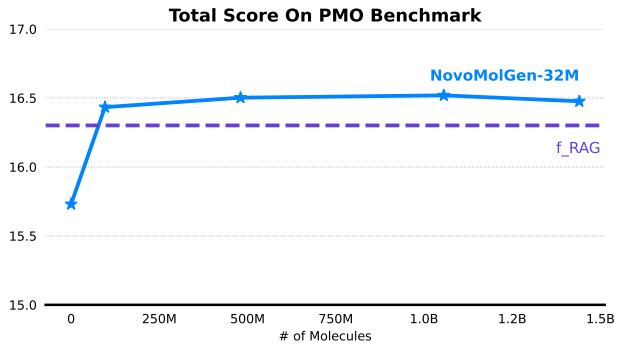


Figure 4. Total PMO score across intermediate checkpoints for NovoMolGen-32M (Atom-wise) with best hyperparameter configuration. Surprisingly, model performance remains stable throughout training, with the earliest checkpoint already surpassing both f-RAG and REINVENT baselines.

evolutionary pressure imparts a rich and functionally meaningful learning signal, datasets of small molecules often lack such inherent contextual information. While natural products such as secondary metabolites reflect evolutionary selection for biological activity and could provide biologically relevant signals, their dataset sizes remain limited (Sorokina & Steinbeck, 2020). In contrast, large-scale chemical libraries like ZINC, which are commonly used for pretraining, are composed of synthetically accessible or commercially available molecules with no inherent selection for biological function. Consequently, these datasets offer only a weak pretraining signal, primarily encouraging the model to learn chemical syntax rather than functional semantics. Given the relative ease of learning syntactic patterns in molecular string representations, relying solely on self-supervised objectives is insufficient. Instead, we believe it would be beneficial to incorporate contextual signals early in training to guide models toward learning biologically relevant features, such as protein-ligand interactions, physicochemical properties, or experimental bioactivity outcomes. Moreover, reinforcement learning can be introduced at earlier stages to align generation with functional objectives. These approaches can provide a "fitness" signal analogous to natural selection, enabling models to capture chemical validity and functional utility.

5. Conclusion

We conducted a broad exploration of how different model sizes, molecular representations, tokenization strategies, and training protocols affect the capabilities of Mol-LLMs. NovoMolGen, which we pretrained on 1.5B molecules in multiple string-based formats, establishes state-of-the-art performance in both unconstrained generation and goal-directed optimization, surpassing the existing Mol-LLMs and specialized generative approaches. Notably, our findings challenge NLP-inspired assumptions about the necessity of extensive training or larger models, suggesting that performance can saturate relatively early. These observations provide a practical framework for building scalable, task-focused molecular foundation models and underscore the need for more demanding benchmarks that capture the true complexity of medicinal chemistry.

Impact Statement

Our study revealed that molecular language models do not need to be large or extensively pre-trained to perform competitively on downstream tasks. Building on this insight, we introduced NovoMolGen, a family of efficient and state-of-the-art molecular language models. Although there are risks of Mol-LLMs misuse in designing harmful compounds, NovoMolGen significantly accelerates drug discovery and represents a crucial step toward making molecular modeling tools faster, cheaper, more inclusive, and sustainable.

References

Adilov, S. Generative Pre-Training from Molecules, September 2021. URL https://chemrxiv.org/engage/chemrxiv/article-details/6142f60742198e8c31782e9e.

Bagal, V., Aggarwal, R., Vinod, P. K., and Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, May 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00600. URL https://doi.org/10.1021/acs.jcim.1c00600. Publisher: American Chemical Society.

Bemis, G. W. and Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, January 1996. ISSN 0022-2623. doi: 10.1021/jm9602928. URL https://doi.org/10.1021/jm9602928. Publisher: American Chemical Society.

Brockschmidt, M. GNN-FiLM: Graph Neural Networks with Feature-wise Linear Modulation. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1144–1152. PMLR, November 2020. URL https://proceedings.mlr.press/v119/brockschmidt20a.html. ISSN: 2640-3498.

Cao, N. D. and Kipf, T. MolGAN: An implicit generative model for small molecular graphs, September 2022. URL http://arxiv.org/abs/1805.11973. arXiv:1805.11973 [stat].

Chilingaryan, G., Tamoyan, H., Tevosyan, A., Babayan, N., Khondkaryan, L., Hambardzumyan, K., Navoyan, Z., Khachatrian, H., and Aghajanyan, A. BARTSmiles: Generative Masked Language Models for Molecular Representations, November 2022. URL http://arxiv.org/abs/2211.16349. arXiv:2211.16349 [cs, q-bio].

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. FLASHATTENTION: fast and memory-efficient exact attention with IO-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 16344–16359, Red Hook, NY, USA, April 2024. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem*, 3(10):1503–1507, 2008. ISSN 1860-7187. doi: 10.1002/cmdc.200800178. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.200800178. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cmdc.200800178.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, November 2002. ISSN 0095-2338. doi: 10.1021/ci010132r. URL https://doi.org/10.1021/ci010132r. Publisher: American Chemical Society.

Eckmann, P., Sun, K., Zhao, B., Feng, M., Gilson, M., and Yu, R. LIMO: Latent Inceptionism for Targeted Molecule Generation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 5777–5792. PMLR, June 2022. URL https://proceedings.mlr.press/v162/eckmann22a.html. ISSN: 2640-3498.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, February 2022. ISSN 2522-5839. doi: 10.1038/s42256-021-00438-4. URL https://www.nature.com/articles/s42256-021-00438-4. Publisher: Nature Publishing Group.

Fang, Y., Zhang, N., Chen, Z., Guo, L., Fan, X., and Chen, H. Domain-Agnostic Molecular Generation with Chemical Feedback. October 2023. URL https://openreview.net/forum?id=9rPyHyjfwP.

Frey, N. C., Soklaski, R., Axelrod, S., Samsi, S., Gomez-Bombarelli, R., Coley, C. W., and Gadepally, V. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305, 2023.

Gao, W., Fu, T., Sun, J., and Coley, C. W. Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization, October 2022. URL http://arxiv.org/abs/2206.12411. arXiv:2206.12411.

Gebauer, N. W. A., Gastegger, M., and Schütt, K. T. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 680, pp. 7566–7578. Curran Associates Inc., Red Hook, NY, USA, December 2019.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for Quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1263–1272, Sydney, NSW, Australia, August 2017. JMLR.org.

Grisoni, F., Moret, M., Lingwood, R., and Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *Journal of Chemical Information and Modeling*, 60(3):1175–1183, March 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00943. URL https://doi.org/10.1021/acs.jcim.9b00943. Publisher: American Chemical Society.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. OLMo: Accelerating the Science of Language Models, June 2024. URL http://arxiv.org/abs/2402.00838. arXiv:2402.00838 [cs].

Guo, Q., Hernandez-Hernandez, S., and Ballester, P. J. Scaffold splits overestimate virtual screening performance. In *International Conference on Artificial Neural Networks*, pp. 58–72. Springer, 2024.

Guo, Z., Guo, K., Nan, B., Tian, Y., Iyer, R. G., Ma, Y., Wiest, O., Zhang, X., Wang, W., Zhang, C., and Chawla, N. V. Graph-based Molecular Representation Learning. volume 6, pp. 6638–6646, August 2023. doi: 10.24963/ijcai.2023/744. URL https://www.ijcai.org/proceedings/2023/744. ISSN: 1045-0823.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, February 2018. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00572. URL https://doi.org/10.1021/acscentsci.7b00572. Publisher: American Chemical Society.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pp. 6840–6851, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant Diffusion for Molecule Generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8867–8887. PMLR, June 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html. ISSN: 2640-3498.

Huang, L., Zhang, H., Xu, T., and Wong, K.-C. MDM: Molecular Diffusion Model for 3D Molecule Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5105–5112, June 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i4.25639. URL https://ojs.aaai.org/index.php/AAAI/article/view/25639. Number: 4.

Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, January 2022. ISSN 2632-2153. doi: 10.1088/2632-2153/ac3ffb. URL https://dx.doi.org/10.1088/2632-2153/ac3ffb. Publisher: IOP Publishing.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

Jin, W., Barzilay, R., and Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2323–2332. PMLR, July 2018. URL https://proceedings.mlr.press/v80/jin18a.html. ISSN: 2640-3498.

Jin, W., Barzilay, D. R., and Jaakkola, T. Multi-Objective Molecule Generation using Interpretable Substructures. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4849–4859. PMLR, November 2020a. URL https://proceedings.mlr.press/v119/jin20b.html. ISSN: 2640-3498.

Jin, W., Barzilay, R., and Jaakkola, T. Hierarchical generation of molecular graphs using structural motifs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pp. 4839–4848. JMLR.org, July 2020b.

Jo, J., Lee, S., and Hwang, S. J. Score-based Generative Modeling of Graphs via the System of Stochastic Differential Equations. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 10362–10383. PMLR, June 2022. URL https://proceedings.mlr.press/v162/jo22a.html. ISSN: 2640-3498.

Kipf, T. N. and Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. February 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Kirkpatrick, P. and Ellis, C. Chemical space. *Nature*, 432(7019):823–823, December 2004. ISSN 1476-4687. doi: 10.1038/432823a. URL https://www.nature.com/articles/432823a. Publisher: Nature Publishing Group.

Kong, L., Cui, J., Sun, H., Zhuang, Y., Prakash, B. A., and Zhang, C. Autoregressive Diffusion Model for Graph Generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17391–17408. PMLR, July 2023. URL https://proceedings.mlr.press/v202/kong23b.html. ISSN: 2640-3498.

Krenn, M., Ai, Q., Barthel, S., Carson, N., Frei, A., Frey, N. C., Friederich, P., Gaudin, T., Gayle, A. A., Jablonka, K. M., Lameiro, R. F., Lemm, D., Lo, A., Moosavi, S. M., Nápoles-Duarte, J. M., Nigam, A., Pollice, R., Rajan, K., Schatzschneider, U., Schwaller, P., Skreta, M., Smit, B., Strieth-Kalthoff, F., Sun, C., Tom, G., Falk von Rudorff, G., Wang, A., White, A. D., Young, A., Yu, R., and Aspuru-Guzik, A. SELFIES and the future of molecular string representations. *Patterns*, 3(10):100588, October 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100588. URL https://www.sciencedirect.com/science/article/pii/S2666389922002069.

Kudo, T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In Gurevych, I. and Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL https://aclanthology.org/P18-1007/.

Kuznetsov, M. and Polykovskiy, D. MolGrow: A Graph Normalizing Flow for Hierarchical Molecular Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8226–8234, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i9.17001. URL https://ojs.aaai.org/index.php/AAAI/article/view/17001. Number: 9.

Le, T., Winter, R., Noé, F., and Clevert, D.-A. Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. *Chemical Science*, 11(38):10378–10389, October 2020. ISSN 2041-6539. doi: 10.1039/D0SC03115A. URL https://pubs.rsc.org/en/content/articlelanding/2020/sc/d0sc03115a. Publisher: The Royal Society of Chemistry.

Lee, S., Jo, J., and Hwang, S. J. Exploring Chemical Space with Score-based Out-of-distribution Generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 18872–18892. PMLR, July 2023. URL https://proceedings.mlr.press/v202/lee23f.html. ISSN: 2640-3498.

Lee, S., Kreis, K., Veccham, S. P., Liu, M., Reidenbach, D., Paliwal, S. G., Vahdat, A., and Nie, W. Molecule Generation with Fragment Retrieval Augmentation. November 2024. URL https://openreview.net/forum?id=56Q0qggDlp&referrer=%5Bthe%20profile%20of%20Arash%20Vahdat%5D(%2Fprofile%3Fid%3D~Arash_Vahdat3).

Levine, Y., Wies, N., Sharir, O., Bata, H., and Shashua, A. Limits to depth-efficiencies of self-attention. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Li, A., Casiraghi, E., and Rousu, J. Chemical reaction enhanced graph learning for molecule representation. *Bioinformatics*, 40(10):btae558, October 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae558. URL https://doi.org/10.1093/bioinformatics/btae558.

Li, S., Zhou, J., Xu, T., Dou, D., and Xiong, H. GeomGCL: Geometric Graph Contrastive Learning for Molecular Property Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4541–4549, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i4.20377. URL https://ojs.aaai.org/index.php/AAAI/article/view/20377. Number: 4.

Luo, S., Guan, J., Ma, J., and Peng, J. A 3D generative model for structure-based drug design. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, pp. 6229–6239, Red Hook, NY, USA, June 2024. Curran Associates Inc. ISBN 978-1-7138-4539-3.

Mahmood, O., Mansimov, E., Bonneau, R., and Cho, K. Masked graph modeling for molecule generation. *Nature Communications*, 12(1):3156,

May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23415-2. URL https://www.nature.com/articles/s41467-021-23415-2. Publisher: Nature Publishing Group.

Maron, H., Ben-Hamu, H., Serviansky, H., and Lipman, Y. Provably Powerful Graph Networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 193, pp. 2156–2167. Curran Associates Inc., Red Hook, NY, USA, December 2019.

Mazuz, E., Shtar, G., Shapira, B., and Rokach, L. Molecule generation using transformers and policy gradient reinforcement learning. *Scientific Reports*, 13(1): 8799, May 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-35648-w. URL https://www.nature.com/articles/s41598-023-35648-w. Number: 1 Publisher: Nature Publishing Group.

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4602–4609, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33014602. URL https://ojs.aaai.org/index.php/AAAI/article/view/4384. Number: 01.

Noutahi, E., Gabellini, C., Craig, M., C. Lim, J. S., and Tossou, P. Gotta be SAFE: a new framework for molecular design. *Digital Discovery*, 3 (4):796–804, 2024. doi: 10.1039/D4DD00019F. URL https://pubs.rsc.org/en/content/articlelanding/2024/dd/d4dd00019f. Publisher: Royal Society of Chemistry.

O'Boyle, N. and Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures, September 2018. URL https://chemrxiv.org/engage/chemrxiv/article-details/60c73ed6567dfe7e5fec388d.

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, September 2017. ISSN 1758-2946. doi: 10.1186/s13321-017-0235-x. URL https://doi.org/10.1186/s13321-017-0235-x.

Özçelik, R. and Grisoni, F. The jungle of generative drug discovery: Traps, treasures, and ways out. *arXiv preprint arXiv:2501.05457*, 2024.

Podda, M., Bacciu, D., and Micheli, A. A Deep Generative Model for Fragment-Based Molecule Generation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2240–2250. PMLR, June 2020. URL https://proceedings.mlr.press/v108/podda20a.html. ISSN: 2640-3498.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Johansson, S., Chen, H., Nikolenko, S., Aspuru-Guzik, A., and Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, 11, December 2020. ISSN 1663-9812. doi: 10.3389/fphar.2020.565644. URL https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2020.565644/full. Publisher: Frontiers.

Rogers, D. and Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t. Publisher: American Chemical Society.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties, December 2022. URL http://arxiv.org/abs/2106.09553. arXiv:2106.09553 [cs, q-bio].

Ross, J., Belgodere, B., Hoffman, S. C., Chenthamarakshan, V., Mroueh, Y., and Das, P. GP-MoLFormer: A Foundation Model For Molecular Generation, April 2024. URL http://arxiv.org/abs/2405.04912. arXiv:2405.04912 [q-bio].

Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, 52(11): 2864–2875, November 2012. ISSN 1549-9596. doi: 10.1021/ci300415d. URL https://doi.org/10.1021/ci300415d. Publisher: American Chemical Society.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583, September 2019. ISSN 2374-7943. doi: 10.1021/acscentsci.9b00576. URL https://doi.org/10.1021/acscentsci.9b00576. Publisher: American Chemical Society.

Sennrich, R., Haddow, B., and Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Erk, K. and Smith, N. A. (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162/.

Simonovsky, M. and Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders, February 2018. URL http://arxiv.org/abs/1802.03480. arXiv:1802.03480 [cs] version: 1.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265. PMLR, June 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html. ISSN: 1938-7228.

Sorokina, M. and Steinbeck, C. Review on natural products databases: where to find data in 2020. *Journal of cheminformatics*, 12(1):20, 2020.

Tazhigulov, R. N., Schiller, J., Oppenheim, J., and Winston, M. Molecular Fingerprints for Robust and Efficient ML-Driven Molecular Generation, November 2022. URL http://arxiv.org/abs/2211.09086. arXiv:2211.09086 [cs].

Tingle, B. I., Tang, K. G., Castanon, M., Gutierrez, J. J., Khurelbaatar, M., Dandarchuluun, C., Moroz, Y. S., and Irwin, J. J. ZINC-22-A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *Journal of Chemical Information and Modeling*, 63(4):1166–1176, February 2023. ISSN 1549-9596. doi: 10.1021/acs.jcim.2c01253. URL https://doi.org/10.1021/acs.jcim.2c01253. Publisher: American Chemical Society.

Wang, Y., Zhao, H., Sciabola, S., and Wang, W. cMolGPT: A Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation. *Molecules*, 28(11):4430, January 2023. ISSN 1420-3049. doi: 10.3390/molecules28114430. URL https://www.mdpi.com/1420-3049/28/11/4430. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, February 1988. ISSN 0095-2338. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005. Publisher: American Chemical Society.

Winter, R., Montanari, F., Steffen, A., Briem, H., Noé, F., and Clevert, D.-A. Efficient multi-objective molecular optimization in a continuous latent space. *Chemical Science*, 10(34):8016–8024, August 2019. ISSN 2041-6539. doi: 10.1039/C9SC01928F. URL https://pubs.rsc.org/en/content/articlelanding/2019/sc/c9sc01928f. Publisher: The Royal Society of Chemistry.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Xiong, J., Xiong, Z., Chen, K., Jiang, H., and Zheng, M. Graph neural networks for automated *de novo* drug design. *Drug Discovery Today*, 26(6):1382–1393, June 2021. ISSN 1359-6446. doi: 10.1016/j.drudis.2021.02.011. URL https://www.sciencedirect.com/science/article/pii/S1359644621000787.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., and Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, August 2020. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.9b00959. URL https://doi.org/10.1021/acs.jmedchem.9b00959. Publisher: American Chemical Society.

Xu*, K., Hu*, W., Leskovec, J., and Jegelka, S. How Powerful are Graph Neural Networks? September 2018. URL https://openreview.net/forum?id=ryGs6iA5Km.

Xu, M., Powers, A. S., Dror, R. O., Ermon, S., and Leskovec, J. Geometric Latent Diffusion Models for 3D Molecule Generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 38592–38610. PMLR, July 2023. URL https://proceedings.mlr.press/v202/xu23n.html. ISSN: 2640-3498.

Yang, S., Hwang, D., Lee, S., Ryu, S., and Hwang, S. J. Hit and lead discovery with explorative RL and fragment-based molecule generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, pp. 7924–7936, Red Hook, NY, USA, June 2024. Curran Associates Inc. ISBN 978-1-7138-4539-3.

Yu, B., Baker, F. N., Chen, Z., Ning, X., and Sun, H. Llasmol: Advancing large language models for chemistry

with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First Conference on Language Modeling*, 2024.

Zang, C. and Wang, F. MoFlow: An Invertible Flow Model for Generating Molecular Graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 617–626, New York, NY, USA, August 2020. Association for Computing Machinery. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403104. URL https://dl.acm.org/doi/10.1145/3394486.3403104.

Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. V. Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 793–803, New York, NY, USA, July 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330961. URL https://dl.acm.org/doi/10.1145/3292500.3330961.

Zhang, O., Zhang, J., Jin, J., Zhang, X., Hu, R., Shen, C., Cao, H., Du, H., Kang, Y., Deng, Y., Liu, F., Chen, G., Hsieh, C.-Y., and Hou, T. ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling. *Nature Machine Intelligence*, 5(9):1020–1030, September 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00712-7. URL https://www.nature.com/articles/s42256-023-00712-7. Publisher: Nature Publishing Group.

Zholus, A., Kuznetsov, M., Schutski, R., Shayakhmetov, R., Polykovskiy, D., Chandar, S., and Zhavoronkov, A. BindGPT: A Scalable Framework for 3D Molecular Design via Language Modeling and Reinforcement Learning, June 2024. URL http://arxiv.org/abs/2406.03686. arXiv:2406.03686 [cs].

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. September 2022. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A. Limitations

This work systematically investigates the effects of molecular string representations, tokenization schemes, and model scaling on *de novo* molecule generation performance. While we explored model scaling in-depth, this analysis was limited to the SMILES representation, which remains the most widely adopted format in molecular generative modeling. However, similar scaling studies for alternative representations such as SELFIES, DeepSMILES, and SAFE were not conducted and remain an avenue for future work.

Additionally, our evaluation focuses exclusively on *de novo* molecular generation. While this task is critical for benchmarking and assessing generative capabilities, its practical utility in real-world drug discovery pipelines is limited. In practice, tasks such as fragment-constrained generation, including scaffold morphing, motif extension, and superstructure generation, are of greater relevance for lead optimization and structure-based design. Future work should extend our framework to evaluate performance on these constraint-driven generative tasks systematically.

B. Broader Impact

Our work demonstrates that NovoMolGen achieves state-of-the-art performance across multiple tasks in *de novo* molecule generation, underscoring its strong potential for real-world applications in drug discovery. While these capabilities present opportunities for accelerating pharmaceutical research, they also raise concerns about potential misuse, such as generating harmful or toxic compounds. To mitigate such risks, safeguards can be implemented at various stages of the generation pipeline. For example, incorporating toxicity-aware objectives in the reward function or applying rigorous post-generation filtering based on toxicity and adverse effect profiles can help prevent the synthesis of hazardous molecules. We encourage the responsible deployment of generative models in chemistry and emphasize the importance of incorporating ethical considerations and safety constraints into future development and usage.

C. Related Work

C.1. Representation of Molecules

Molecules are commonly depicted using structural diagrams, traditionally drawn with pen and paper, to represent bonds and atoms visually. However, in chem-informatics, more advanced representations are needed for the computational processing of molecular structures. In this context, "molecular representations" encompass any encoding of a chemical compound that can be employed for computational exploration of the chemical space. The current approaches to representing molecules can be broadly classified into four types: (i) Vector-based representations, (ii) Graph-based representations, (iii) 3D-based representations, and (iv) String-based representations.

Vector-based: Topological fingerprints, such as Extended Connectivity FingerPrints (ECFP) (Rogers & Hahn, 2010) and Molecular ACCess System (MACCS) (Durant et al., 2002), have traditionally been employed for substructure and molecule similarity searches. These fingerprints encode molecules as a sequence of bits in an identifier list, each denoting the presence or absence of a specific substructure. Although each molecular structure can be deterministically mapped to a fingerprint, the fingerprints are only partially invertible (Le et al., 2020), which limits their applicability in *de-novo* molecule generation (Gómez-Bombarelli et al., 2018; Tazhigulov et al., 2022). Additionally, the fingerprints can be augmented with 2D molecular descriptors such as Molecular Weight, QED score, and Number of Aromatic Rings to help impose specific constraints on the generated molecules and make them more aligned with desired chemical properties or biological activity.

Graph-based: A 2D molecular graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of nodes (atoms) and $\mathcal{E}$ is the set of edges (bonds). The type of atoms and edges can be represented using a feature matrix $\boldsymbol{X}$. Graph Neural Networks (GNNs) have been used to learn the representations of molecules (Kipf & Welling, 2017; Xu* et al., 2018; Xiong et al., 2021) for tasks such as reaction prediction, property prediction and drug discovery. The initial frameworks for learning molecule representation used Message Passing Neural Networks (MPNNs) to compute the atom embedding based on neighbourhood information capturing local interaction effects (Gilmer et al., 2017). Although many variants of GNNs have been proposed (Morris et al., 2019; Maron et al., 2019; Zhang et al., 2019; Brockschmidt, 2020; Xiong et al., 2020; Li et al., 2024), challenges remain in terms of higher-order expressivity, scalability, and computational cost.

3D-based: While using Graph Neural Networks (GNNs) on 2D molecular graphs is convenient and seem to be the obvious choice, the resulting representations often overlook crucial spatial information, such as the spatial direction and torsion between atoms (Guo et al., 2023). Recent advancements in molecule representation learning have focused on integrating

3D coordinate information into 2D molecular graphs (Luo et al., 2024; Fang et al., 2022; Li et al., 2022). Uni-Mol (Zhou et al., 2022) introduced a pretraining framework capable of directly utilizing 3D positions as inputs and outputs. However, a significant challenge in this approach is the existence of multiple low-energy conformations for a given molecule. These conformations are not easily accessible and are particularly difficult to compute, especially for large molecules and the vast chemical space, which spans billions of possible molecules.

String-based: 2D molecular structures can also be encoded as linear notations, which use specialized languages to represent molecular structures and compositions in chemistry. The earliest example of such a molecular language was developed in the 1980s by Weininger (1988). The SMILES (Simplified Molecular Input Line Entry System) notation encodes atoms, bonds, and connectivity patterns using ASCII strings, where atoms are represented by characters (e.g., 'C' for carbon, 'N' for nitrogen) and bonds by special characters (e.g., '-' for a single bond, '=' for a double bond, '#' for a triple bond). However, the syntax rules and restrictive grammar of SMILES can result in many invalid molecules during parsing, even when the string appears to represent a plausible molecular structure. To address some of these limitations, DeepSMILES (O'Boyle & Dalke, 2018) was introduced, which avoids the issue of unbalanced parentheses by using only closing parentheses, where the number of parentheses indicates the branch length. More recently, SELFIES (Self-Referencing Embedded Strings) (Krenn et al., 2022) was developed as a linear notation that is 100% robust; every SELFIES string corresponds to a valid molecule, even for entirely random strings. Additionally, SAFE (Sequential Attachment-based Fragment Embedding) (Noutahi et al., 2024) introduced a framework for fragment-constrained molecule generation tasks while maintaining compatibility with existing SMILES parsers.

String-based molecular representations offer a computationally efficient and scalable approach to exploring the vast chemical space using unlabelled data without relying on additional information such as 3D geometry or complex optimization techniques. Despite their simplicity, these methods capture essential chemical information and structural features, making them valuable tools in computational chemistry and drug discovery.

C.2. Deep Generative Models for De-Novo Molecule Generation

Deep generative models have become a key approach for *de novo* molecule generation, facilitating the discovery of novel compounds by capturing complex patterns within the vast chemical space. Numerous methods have emerged, each focused on different molecular representations and assembly strategies (Olivecrona et al., 2017; Jin et al., 2018; Polykovskiy et al., 2020; Jin et al., 2020b; Bagal et al., 2022; Eckmann et al., 2022; Jo et al., 2022; Irwin et al., 2022; Fang et al., 2023; Lee et al., 2023; Yang et al., 2024; Ross et al., 2024).

Assembly Methods: Early approaches (Gómez-Bombarelli et al., 2018; Jin et al., 2018; Winter et al., 2019; Tazhigulov et al., 2022) primarily utilized Variational Autoencoders (VAEs) to transform SMILES representations into a continuous latent space, followed by sampling and decoding to generate discrete molecular structures. Graph-based generative models have emerged as a natural extension, directly leveraging the molecular graph structure where atoms are represented as nodes and bonds as edges (Cao & Kipf, 2022). GraphVAE (Graph Variational Autoencoder) (Simonovsky & Komodakis, 2018) encodes and decodes molecules using edge-conditioned graph convolutions. MoFlow (Zang & Wang, 2020), a flow-based graph generative model learns invertible mappings between molecular graphs and their latent representations. Graph generation approaches utilize small molecular building blocks such as atoms and their performance degrades significantly for larger molecules. To tackle this problem recent works employ significantly larger and more flexible graph motifs as basic building blocks (Kuznetsov & Polykovskiy, 2021). In parallel, 3D molecule generation has gained substantial attention, particularly through the use of diffusion models. G-Schnet (Gebauer et al., 2019), for instance, utilizes an autoregressive process to iteratively sample atoms and bonds in 3D space. Similarly, inspired by the success of diffusion models in other domains (Sohl-Dickstein et al., 2015; Ho et al., 2020; Kong et al., 2023), Hoogeboom et al. (2022) proposed an equivariant diffusion model for generating novel 3D molecular structures.

Optimization Methods: Molecule optimization involves navigating an immense and complex chemical space, which requires sophisticated algorithms capable of efficiently searching for and generating molecules with optimal characteristics. Several computational approaches have been developed to tackle this challenge, each with distinct mechanisms for exploring the design space and optimizing molecular properties. Reinforcement Learning treats molecule optimization as a sequential decision-making problem. In this context, the state typically represents a partially generated molecule, and actions correspond to modifications at the graph or string level. The reward function is based on the properties of the generated molecules, guiding the model toward desirable outcomes. Bayesian Optimization operates by learning a continuous latent space of molecular representations, optimizing target properties by navigating through this latent space. Genetic algorithms, inspired

by natural evolutionary processes, explore the chemical space through operations such as mutation and crossover applied to a pool of candidate molecules, promoting diversity and exploration. Gradient ascent methods, on the other hand, estimate the gradient of a molecular property across the chemical space and use backpropagation to optimize molecular structures. Hill Climbing is an iterative optimization method with high-performing molecules from previous rounds incorporated into the training data to refine the generative model progressively.

While graph-based and 3D deep generative models have made significant strides in generating molecular structures, recent advances in natural language processing have opened new possibilities for *de novo* molecule generation and optimization. Large Language Models (LLMs) offer a novel approach for navigating the large chemical space, presenting new opportunities for optimizing molecular properties in a scalable and computationally feasible manner. Combining these models with traditional optimization techniques can further enhance the search for *de novo* molecules with desired properties, marking a significant step forward in drug discovery.

C.3. Language Models in Molecule Generation

LLMs can effectively model sequential data and have shown remarkable proficiency in understanding and generating human language (Dubey et al., 2024; Jiang et al., 2023; Groeneveld et al., 2024). These architectures are now being repurposed to explore and generate molecular structures. When treated as sequences of tokens, 1D molecular representations inherently encode chemical information, including 2D bonding topology patterns, while LLMs further enhance this by learning to generate diverse molecular structures. These models leverage unlabelled molecular data from across the chemical space, offering a novel approach to *de novo* molecule generation and broadening the potential for chemical discovery. MolGPT (Bagal et al., 2022) employs a decoder-only transformer architecture, inspired by GPT, to predict SMILES token sequences for molecular generation. Building on these advancements, models like cMolGPT (Wang et al., 2023) have been developed to generate target-specific compounds by incorporating conditional training for property optimization. Taiga (Mazuz et al., 2023) extends this approach by employing a two-stage framework: first, the model treats molecular generation as a language modeling task by predicting the next token in SMILES strings. Subsequently, reinforcement learning (RL) is applied to optimize simple chemical properties such as QED (Quantitative Estimate of Drug-likeness) and logP (Octanol-Water Partition Coefficient). MolGen (Fang et al., 2023), in contrast, utilizes an encoder-decoder BART architecture, focusing on generating chemically valid molecules through the SELFIES notation. It incorporates a chemical feedback mechanism to align generative probabilities with real-world chemical preferences. SAFE-GPT (Noutahi et al., 2024) introduces a new line notation and trains a GPT-like model on 1.1 billion SAFE notations, demonstrating versatile and robust performance in both *de novo* and fragment-constrained molecule generation tasks.

Tokenizers, which convert raw text sequences into tokens, are a critical component of modern language models. In molecular language models, token vocabularies are often constructed using a predefined regular expression developed by Schwaller et al. (2019), which splits SMILES strings into relevant tokens representing each atom (e.g., 'C' for carbon, 'N' for nitrogen). Less commonly, subword tokenization algorithms such as Byte Pair Encoding (BPE) (Sennrich et al., 2016) or Unigram (Kudo, 2018) are employed, sometimes combined with atom-wise pretokenization and BPE. Given that tokenizer design impacts every stage of the modeling pipeline, this study explores the effects of using learned tokenization methods versus hand-crafted approaches.

D. Dataset Diversity

To ensure the diversity of the dataset, we confirmed that it contains a broad range of molecular structures, which is crucial for generating a wide variety of valid, novel, and unique molecules. Figure D5 illustrates the diversity of the data. The left plot shows the distribution of molecular lengths, tokenized using an atom-wise tokenizer, demonstrating variation in molecule size across batches. The right plot displays the Tanimoto similarity between molecule pairs, showcasing the structural diversity in each batch — an essential factor for robust molecular generation.

E. Pretraining Configuration

Our pretraining experiments leverage the computational enhancements of the FlashAttention library (Dao et al., 2024), utilizing its Llama implementation within the HuggingFace Trainer framework [1]. Training was conducted in mixed precision mode using bfloat16 to maximize GPU memory efficiency. We adopted the AdamW optimizer with a learning rate of

[1] https://huggingface.co/docs/transformers/en/main_classes/trainer

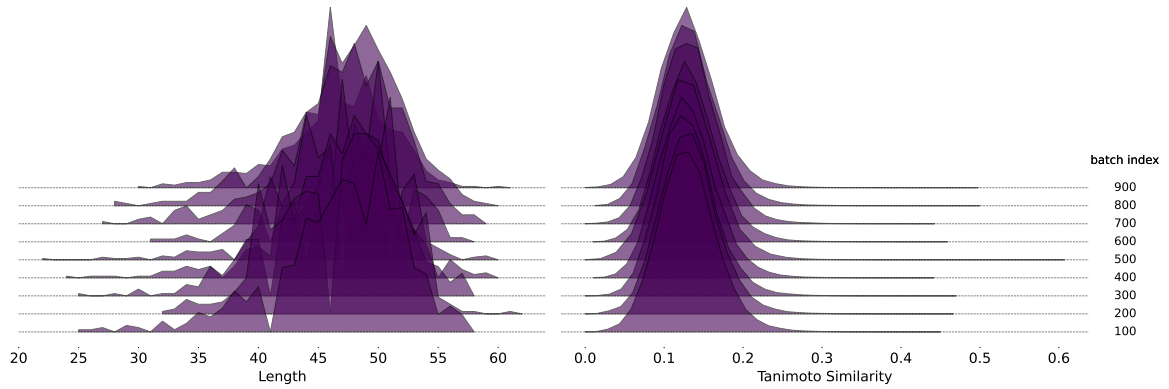


Figure D5. Diversity of the data across batches. The left plot shows the distribution of molecular lengths, tokenized using an atom-wise tokenizer, indicating variation in molecule length within each batch. The right plot shows Tanimoto similarity between molecule pairs, demonstrating structural diversity in each batch.

Table E2. NovoMolGen Configurations

Components	NovoMolGen-32M	NovoMolGen-157M	NovoMolGen-300M
Attention Heads	8	10	12
Hidden Layers	12	24	32
Hidden Size	512	640	768
Intermediate Size	1024	2560	3072

6×10^{-4}, paired with a cosine learning rate scheduler configured with a half cycle. The scheduler includes a warmup phase of 2% of the total training steps, during which the learning rate linearly increases to its peak value of 6×10^{-4} before gradually decaying to a minimum of 6×10^{-5}.

To ensure consistency across experiments, we maintained a fixed global batch size of 19,200 molecules, with gradient accumulation and per-device batch size selected to fit within the memory constraints of 4 NVIDIA A100 GPUs (80 GB each). Weight decay was set to 0.01 to prevent overfitting, and gradient clipping was applied with a maximum gradient norm of 1.0. The AdamW optimizer used $\beta_1 = 0.9$ and $\beta_2 = 0.95$, ensuring stability during training.

F. Training curves

The training curves for various model sizes, tokenization strategies, and molecular representations are shown in Figures F6 to F9. Across model sizes (32M, 157M, 300M), we observe minimal difference in validation loss between random and scaffold-based splits for both the atom-wise (Figure F6) and BPE (Figure F7) tokenizers, suggesting comparable performance under both evaluation strategies. Notably, in Figures F8 and F9, the DeepSMILES representation consistently achieves the lowest validation loss across both splits. Overall, models evaluated on the random split achieve slightly lower losses than those assessed on the scaffold split, indicating a modest challenge in generalizing to unseen scaffolds, although the performance gap remains small.

G. Evaluation Details

G.1. Pretraining Metrics

This section describes the metrics used to assess the performance of our molecule generation model during pretraining, following the metrics outlined in the MOSES benchmark (Polykovskiy et al., 2020). These metrics are computed based on the generated set of molecules from the model, denoted as G, and two reference sets: R_{valid} (ZINC-Random) and R_{test} (ZINC-Scaffold), which correspond to molecules derived from a random split and a scaffold-based split, respectively. All reported metrics are calculated using the subset of G consisting of valid molecules identified through a post-generation

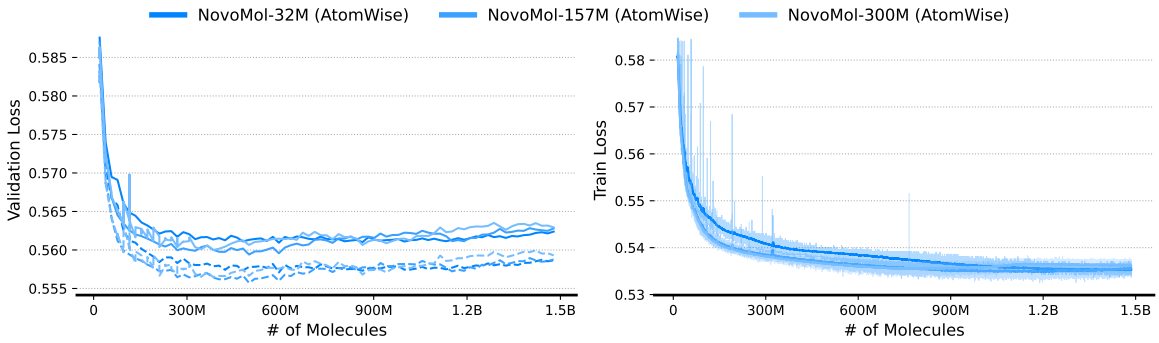


Figure F6. Training and validation loss curves for different model sizes. Solid lines denote performance on randomly split validation sets, while dashed lines indicate results on scaffold-split validation sets. The x-axis shows the number of molecules seen during training.

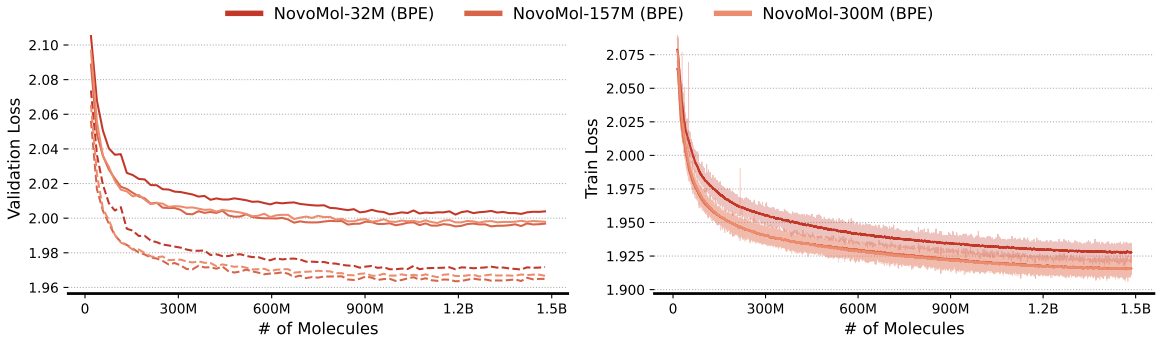


Figure F7. Training and validation loss curves for different model sizes with BPE tokenizer. Solid lines denote performance on randomly split validation sets, while dashed lines indicate results on scaffold-split validation sets. The x-axis shows the number of molecules seen during training.

filtering process except for validity.

1. **Validity**: Validity is determined using RDKit's molecular structure parser, which verifies atomic valency and the consistency of bonds within aromatic rings. The metric ensures that the model adheres to relevant chemical constraints and measures the proportion of valid molecules generated within G. For molecular representations other than SMILES, we convert the generated set to SMILES and assess whether the corresponding decoder successfully decodes the molecule string to SMILES format. This step is essential, as representations that adhere to their respective syntactic rules may still produce chemically invalid molecules.

2. **Novelty**: This metric quantifies the proportion of molecules in G that do not appear in the training dataset. The molecules in G are canonicalized and compared against the training dataset, which comprises 1.5 billion molecules.

3. **Internal Diversity (IntDiv)**: Chemical diversity within the generated set, G, is evaluated using this metric. Higher values indicate better diversity. It is calculated as the average pairwise Tanimoto similarity of Morgan fingerprints for the molecules in G, with values ranging from 0 to 1.

4. **Fréchet ChemNet Distance (FCD)**: Derived from the activations of the penultimate layer of ChemNet, a deep neural network trained to predict the biological activities of drugs, this measure captures the chemical and biological properties of molecules. Activations for canonical SMILES representations of molecules are compared between the generated and reference sets. Lower values indicate better overlap, and the metric is non-negative.

5. **Fragment Similarity (Frag)**: The distribution of BRICS fragments (Degen et al., 2008) is compared between the generated and reference sets. Higher values signify a closer match in fragment distributions, ensuring no fragment is disproportionately overrepresented or underrepresented.

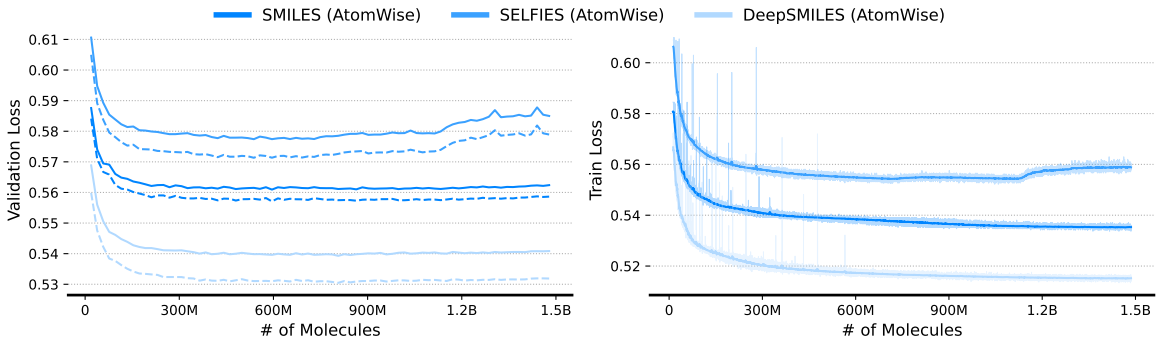


Figure F8. Training and validation loss curves for different Molecule type with atom-wise tokenizer. Solid lines denote performance on randomly split validation sets, while dashed lines indicate results on scaffold-split validation sets. The x-axis shows the number of molecules seen during training.

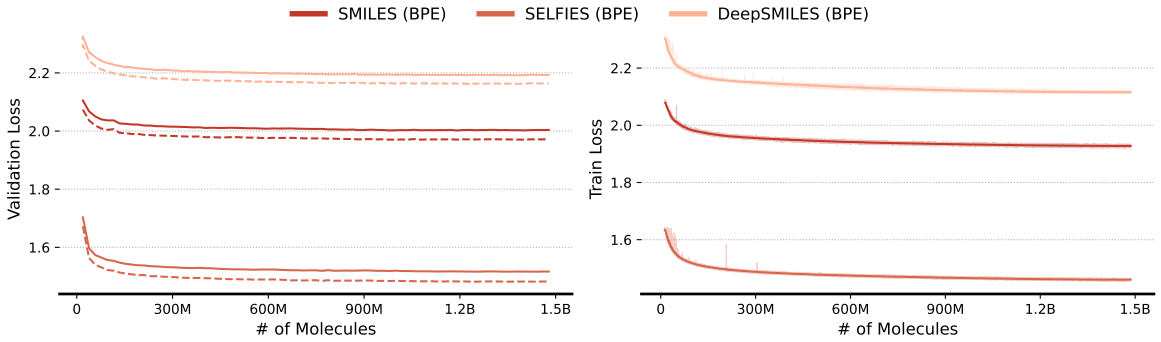


Figure F9. Training and validation loss curves for different Molecule type with BPE tokenizer. Solid lines denote performance on randomly split validation sets, while dashed lines indicate results on scaffold-split validation sets. The x-axis shows the number of molecules seen during training.

6. **Scaffold Similarity (Scaff)**: Similar to Fragment Similarity, this comparison uses Bemis-Murcko scaffolds instead of BRICS fragments to evaluate the resemblance between scaffolds in the generated and reference datasets.

7. **Similarity to Nearest Neighbor (SNN)**: The average Tanimoto similarity between each molecule in the generated set and its nearest counterpart in the reference set is calculated. Lower values suggest the generated molecules are farther from the reference set's manifold, while higher values indicate closer alignment.

H. Additional Analysis and Results

H.1. Pretraining Benchmark

To ensure a fair comparison with baseline models (Polykovskiy et al., 2020; Jin et al., 2018; Eckmann et al., 2022; Fang et al., 2023), we report results for the generation of 30,000 molecules using held-out test sets of 175,000 molecules. All results are averaged over three independent model initialization seeds. Based on the results from Tables H3 to H5, NovoMolGen demonstrates state-of-the-art performance in terms of validity, novelty, and FCD scores. It performs comparably to other baselines on metrics related to fragments and scaffolds. Overall, the SMILES representation, across multiple model sizes and tokenization schemes, yields the best performance, although the differences among them are not substantial. SAFE underperforms in all metrics, with the exception of Internal Diversity. Furthermore, Byte Pair Encoding (BPE) emerges as the preferred tokenization strategy, outperforming atom-wise for SAFE.

Table H3. Performance metrics for baseline models and **NovoMolGen** on Validity, Internal Diversity (IntDiv), and Novelty. Results are reported as mean$_{(\text{std})}$ over three independent model initializations. **Blue** denotes the best performing model, while **Pink** represents the second-best performing model.

REPRESENTATION	TOKENIZATION	MODEL SIZE	Valid (↑)	IntDiv (↑)	Novelty (↑)
Train	–	–	$1.000_{(0.0)}$	$1.000_{(0.0)}$	$0.0_{(0.0)}$
GP-MOLFORMER (Ross et al., 2024)	ATOM-WISE	46.8M	**1.000**	**0.997**	0.390
SMILES	ATOM-WISE	32M	$\mathbf{0.999}_{(\mathbf{0.0001})}$	$0.851_{(0.0001)}$	$\mathbf{0.983}_{(\mathbf{0.0001})}$
		157M	$\mathbf{0.999}_{(\mathbf{0.0001})}$	$0.851_{(0.0001)}$	$0.981_{(0.0001)}$
		300M	$\mathbf{0.999}_{(\mathbf{0.0001})}$	$0.851_{(0.0001)}$	$0.981_{(0.0001)}$
	BPE	32M	$\mathbf{0.999}_{(\mathbf{0.0000})}$	$0.851_{(0.0001)}$	$\mathbf{0.982}_{(\mathbf{0.0001})}$
		157M	$\mathbf{0.999}_{(\mathbf{0.0000})}$	$\mathbf{0.852}_{(\mathbf{0.0001})}$	$0.980_{(0.0001)}$
		300M	$\mathbf{0.999}_{(\mathbf{0.0001})}$	$0.851_{(0.0000)}$	$0.979_{(0.0001)}$
SELFIES	ATOM-WISE	32M	$\mathbf{1.000}_{(\mathbf{0.0000})}$	$\mathbf{0.852}_{(\mathbf{0.0002})}$	$\mathbf{0.982}_{(\mathbf{0.0001})}$
	BPE	32M	$\mathbf{1.000}_{(\mathbf{0.0000})}$	$0.851_{(0.0000)}$	$\mathbf{0.984}_{(\mathbf{0.0001})}$
DEEPSMILES	ATOM-WISE	32M	$\mathbf{0.999}_{(\mathbf{0.0001})}$	$0.851_{(0.0001)}$	$0.981_{(0.0001)}$
	BPE	32M	$0.997_{(0.0001)}$	$0.851_{(0.0001)}$	$\mathbf{0.982}_{(\mathbf{0.0001})}$
SAFE	ATOM-WISE	32M	$0.957_{(0.0012)}$	$\mathbf{0.853}_{(\mathbf{0.0001})}$	$0.953_{(0.0001)}$
	BPE	32M	$0.997_{(0.0007)}$	$\mathbf{0.852}_{(\mathbf{0.0001})}$	$0.976_{(0.0001)}$

H.2. PMO Benchmark

This appendix provides a comprehensive analysis of the PMO benchmark results for **NovoMolGen**, assessing its performance across varying model sizes, tokenization strategies, and intermediate training checkpoints. In Figure H10, the results are presented, highlighting the influence of model size and tokenization approaches on the performance of NovoMolGen. Additionally, Figure H11 tracks the performance progression of NovoMolGen-32M-Atom-wise across intermediate checkpoints, demonstrating the evolution of model performance with the optimal hyperparameter configuration and atom-wise tokenization. The evaluation includes comparisons with the REINVENT and f-RAG baselines, with the mean and standard deviation of 3 independent runs presented in Table H6.

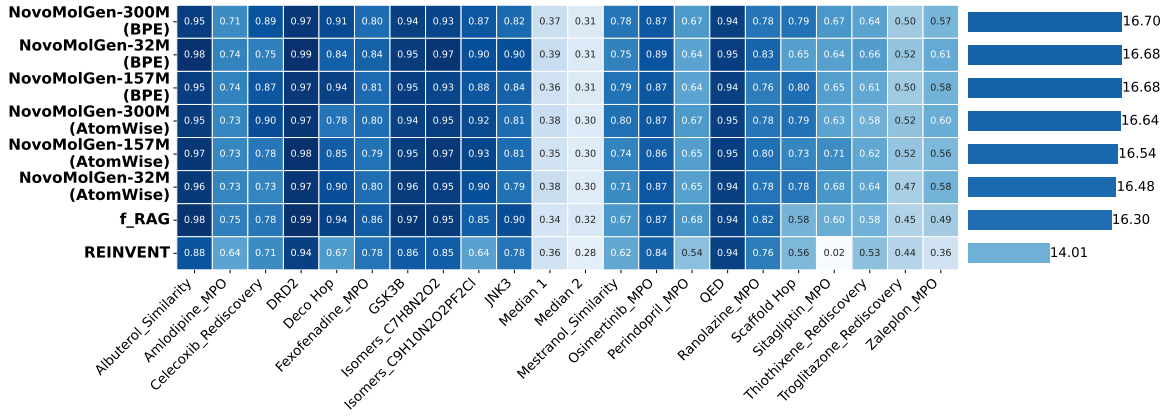


Figure H10. PMO benchmark results for **NovoMolGen-32M** across model sizes and tokenization strategies. The **heatmap** (left) shows normalized scores per task, while the **bar chart** (right) presents total scores (higher is better). Baselines are REINVENT and f-RAG .

I. Property Distributions

The distribution of properties serves as a valuable tool for visually evaluating the generated structures. We present a kernel density estimation of these distributions and calculate the Wasserstein-1 distance to compare the distributions of the

Table H4. Performance metrics for baseline models and **NovoMolGen** on Fréchet ChemNet Distance (FCD) and Similarity to Nearest Neighbor (SNN). Results are presented for both the random test set (Test) and scaffold-split test set (TestSF), with values reported as $\text{mean}_{(\text{std})}$ over three independent model initializations. **Blue** denotes the best performing model, while **Pink** represents the second-best performing model. The baselines for CharRNN, VAE, and JT-VAE are sourced from Polykovskiy et al. (2020), while the results for LIMO, MolGen-7b, and GP-Molformer are taken from Ross et al. (2024).

Representation	Tokenization	Model Size	FCD (↓)		SNN (↑)	
			Test	TestSF	Test	TestSF
Train	–	–	$0.0376_{(0.0002)}$	$0.2392_{(0.01)}$	$0.4657_{(0.0001)}$	$0.4550_{(0.0001)}$
CharRNN (Polykovskiy et al., 2020)	Atom-wise	–	$0.0732_{(0.0247)}$	$0.5204_{(0.0379)}$	$\mathbf{0.6015_{(0.0206)}}$	$\mathbf{0.5649_{(0.0142)}}$
VAE (Polykovskiy et al., 2020)	Atom-wise	–	$0.0990_{(0.0125)}$	$0.5670_{(0.0338)}$	$\mathbf{0.6257_{(0.0005)}}$	$\mathbf{0.5783_{(0.0008)}}$
JT-VAE (Jin et al., 2018)	Atom-wise	–	$0.3954_{(0.0234)}$	$0.9382_{(0.0531)}$	$0.5477_{(0.0076)}$	$0.5194_{(0.007)}$
LIMO (Eckmann et al., 2022)	Atom-wise	–	26.78	–	0.2464	–
MolGen-7b (Fang et al., 2023)	Atom-wise	7B	0.0435	–	0.5138	–
GP-Molformer (Ross et al., 2024)	Atom-wise	46.8M	0.0591	–	0.5045	–
SMILES	Atom-wise	32M	$0.0394_{(0.0002)}$	$\mathbf{0.2386_{(0.0039)}}$	$0.4657_{(0.0002)}$	$0.4551_{(0.0001)}$
		157M	$0.0426_{(0.0012)}$	$\mathbf{0.2225_{(0.0006)}}$	$0.4656_{(0.0002)}$	$0.4551_{(0.0001)}$
		300M	$0.0419_{(0.0004)}$	$0.2446_{(0.0074)}$	$0.4657_{(0.0001)}$	$0.4548_{(0.0001)}$
	BPE	32M	$\mathbf{0.0384_{(0.0008)}}$	$0.2402_{(0.0045)}$	$0.4654_{(0.0003)}$	$0.4550_{(0.0001)}$
		157M	$\mathbf{0.0384_{(0.0008)}}$	$0.2404_{(0.0034)}$	$0.4655_{(0.0003)}$	$0.4547_{(0.0001)}$
		300M	$\mathbf{0.0380_{(0.0003)}}$	$0.2463_{(0.0125)}$	$0.4657_{(0.0003)}$	$0.4550_{(0.0002)}$
SELFIES	Atom-wise	32M	$0.0799_{(0.0031)}$	$0.2480_{(0.0032)}$	$0.4651_{(0.0003)}$	$0.4549_{(0.0002)}$
	BPE	32M	$0.0386_{(0.0006)}$	$0.2436_{(0.0019)}$	$0.4649_{(0.0004)}$	$0.4541_{(0.0002)}$
DeepSMILES	Atom-wise	32M	$0.0400_{(0.0006)}$	$0.2504_{(0.0018)}$	$0.4661_{(0.0001)}$	$0.4554_{(0.0002)}$
	BPE	32M	$\mathbf{0.0380_{(0.0004)}}$	$0.2390_{(0.0048)}$	$0.4655_{(0.0001)}$	$0.4546_{(0.0002)}$
SAFE	Atom-wise	32M	$0.9475_{(0.0042)}$	$1.1979_{(0.0106)}$	$0.4562_{(0.0001)}$	$0.4459_{(0.0003)}$
	BPE	32M	$0.3283_{(0.0045)}$	$0.5506_{(0.0004)}$	$0.4599_{(0.0002)}$	$0.4492_{(0.0003)}$

generated and reference datasets. We use the following properties:

1. **Quantitative Estimation of Drug-likeness (QED)**: A metric derived from medicinal chemistry principles that quantifies the drug-likeness of molecules on a scale from 0 to 1.
2. **Synthetic Accessibility (SA)**: A measure of a molecule's synthesizability, calculated based on the contributions of molecular fragments. The metric ranges from 10 (difficult to synthesize) to 2 (easily synthesizable).
3. **Octanol-Water Partition Coefficient (logP)**: Represents the ratio of a compound's concentration in the octanol phase to its concentration in the aqueous phase in a two-phase octanol/water system, serving as an indicator of solubility.
4. **Molecular Weight (MW)**: Evaluates whether the generated set is biased toward heavier or lighter molecules, computed as the sum of atomic weights.
5. **Topological Polar Surface Area (TPSA)**: Estimated based on functional group contributions from a database of substructures, this metric reflects lipid solubility and molecular polarity. Higher TPSA values indicate reduced absorption and distribution within the body.
6. **Bertz Complexity**: A graph-theoretical measure that quantifies molecular complexity using structural invariants and information-theoretic principles.
7. **Number of Rings (NumRings) and Rotatable Bonds**: Represents the number of independent closed-ring structures and rotatable bonds within a molecule, which are essential for analyzing molecular topology and are commonly used in cheminformatics for compound classification and comparison.

The kernel density estimation plots in Figure I12 indicate that NovoMolGen successfully generates molecules whose property distributions align closely with those of both the training dataset and the scaffold-split dataset. Additionally, a lower Wasserstein-1 distance is observed across all properties, further demonstrating the model's ability to replicate the reference distributions. The training dataset contains a higher proportion of molecules with favorable drug-like properties,

Table H5. Performance metrics for baseline models and **NovoMolGen** on Fragment similarity (Frag) and Scaffold similarity (Scaff). Results are presented for both the random test set (Test) and scaffold-split test set (TestSF), with values reported as mean$_{(\text{std})}$ over three independent model initializations. **Blue** denotes the best performing model, while **Pink** represents the second-best performing model. The baselines for CharRNN, VAE, and JT-VAE are sourced from Polykovskiy et al. (2020), while the results for LIMO, MolGen-7b, and GP-Molformer are taken from Ross et al. (2024).

Representation	Tokenization	Model Size	Frag (↑)		Scaf (↑)	
			Test	TestSF	Test	TestSF
Train	–	–	$0.9999_{(0.00006)}$	$0.9974_{(0.0001)}$	$0.5144_{(0.005)}$	$0.0_{(0.0)}$
CharRNN (Polykovskiy et al., 2020)	Atom-wise	–	$\mathbf{0.9998_{(0.0002)}}$	$0.9983_{(0.0003)}$	$\mathbf{0.9242_{(0.0058)}}$	$\mathbf{0.1101_{(0.0081)}}$
VAE (Polykovskiy et al., 2020)	Atom-wise	–	$0.9994_{(0.0001)}$	$\mathbf{0.9984_{(0.0003)}}$	$\mathbf{0.9386_{(0.0021)}}$	$0.0588_{(0.0095)}$
JT-VAE (Jin et al., 2018)	Atom-wise	–	$0.9965_{(0.0003)}$	$0.9947_{(0.0002)}$	$0.8964_{(0.0039)}$	$\mathbf{0.1009_{(0.0105)}}$
LIMO (Eckmann et al., 2022)	Atom-wise	–	0.6989	–	0.0079	–
MolGen-7b (Fang et al., 2023)	Atom-wise	7B	**0.9999**	–	0.6538	–
GP-Molformer (Ross et al., 2024)	Atom-wise	46.8M	**0.9998**	–	0.7383	–
SMILES	Atom-wise	32M	$\mathbf{0.9999_{(0.0000)}}$	$0.9977_{(0.0001)}$	$0.5309_{(0.0022)}$	$0.0091_{(0.0023)}$
		157M	$\mathbf{0.9999_{(0.0000)}}$	$\mathbf{0.9980_{(0.0001)}}$	$0.5200_{(0.0045)}$	$0.0095_{(0.0036)}$
		300M	$\mathbf{0.9999_{(0.0000)}}$	$0.9973_{(0.0002)}$	$0.5219_{(0.0136)}$	$0.0065_{(0.0009)}$
	BPE	32M	$\mathbf{0.9999_{(0.0000)}}$	$0.9974_{(0.0001)}$	$0.5182_{(0.0126)}$	$0.0098_{(0.0013)}$
		157M	$\mathbf{0.9999_{(0.0000)}}$	$0.9974_{(0.0001)}$	$0.5193_{(0.0099)}$	$0.0114_{(0.0036)}$
		300M	$\mathbf{0.9999_{(0.0000)}}$	$0.9974_{(0.0001)}$	$0.5173_{(0.0094)}$	$0.0131_{(0.0023)}$
SELFIES	Atom-wise	32M	$0.9996_{(0.0000)}$	$0.9977_{(0.0000)}$	$0.4765_{(0.0066)}$	$0.0071_{(0.0005)}$
	BPE	32M	$\mathbf{0.9999_{(0.0000)}}$	$0.9973_{(0.0002)}$	$0.5105_{(0.0133)}$	$0.0061_{(0.0028)}$
DeepSMILES	Atom-wise	32M	$\mathbf{0.9999_{(0.0000)}}$	$0.9974_{(0.0000)}$	$0.5218_{(0.0080)}$	$0.0098_{(0.0044)}$
	BPE	32M	$\mathbf{0.9999_{(0.0000)}}$	$0.9973_{(0.0001)}$	$0.5165_{(0.0008)}$	$0.0070_{(0.0032)}$
SAFE	Atom-wise	32M	$0.9955_{(0.0001)}$	$0.9936_{(0.0002)}$	$0.4521_{(0.0137)}$	$0.0094_{(0.0019)}$
	BPE	32M	$0.9972_{(0.0000)}$	$0.9946_{(0.0002)}$	$0.4913_{(0.0014)}$	$0.0059_{(0.0028)}$

including higher drug-likeness scores (QED > 0.6), greater synthetic accessibility (SA < 4), optimal solubility ($1 <$ logP < 4), and an ideal molecular weight range ($300 <$ MW < 500). By closely matching this distribution, NovoMolGen generates molecules with an increased likelihood of exhibiting drug-like characteristics. Furthermore, the model effectively captures molecular topology, as reflected in the distributions of NumRings, Rotatable Bonds, and Bertz Complexity. These results highlight the potential of NovoMolGen in generating chemically relevant and synthetically accessible molecules suitable for drug discovery applications.

J. Hyperparameter Tuning for Fine-tuning Method

J.1. Hyperparameter Search Setup for PMO Benchmark

To fine-tune our models for goal-directed molecular design tasks, we performed an exhaustive hyperparameter search using the REINVENT-inspired framework. Our experiments focused on two benchmark tasks, **Perindopril_MPO** and **Zaleplon_MPO**, with the overall performance aggregated as the sum of scores across these tasks. Each hyperparameter configuration was evaluated over three different random seeds, and the final score was averaged across seeds to mitigate variability.

The hyperparameter space explored includes the following:

- **Penalty Coefficient** (λ): $[10, 100, 500]$
- **Batch Size**: $[64, 128]$
- **Sigma** (σ): $[1000, 1500, 2000, 2500]$
- **Learning Rate** (lr): $[1 \times 10^{-4}, 5 \times 10^{-4}]$

Initially, we observed that the penalty coefficient (λ) used in the REINVENT implementation (5000) was not optimal for our models due to differences in the likelihood scale of generated molecules. Adjusting this hyperparameter was

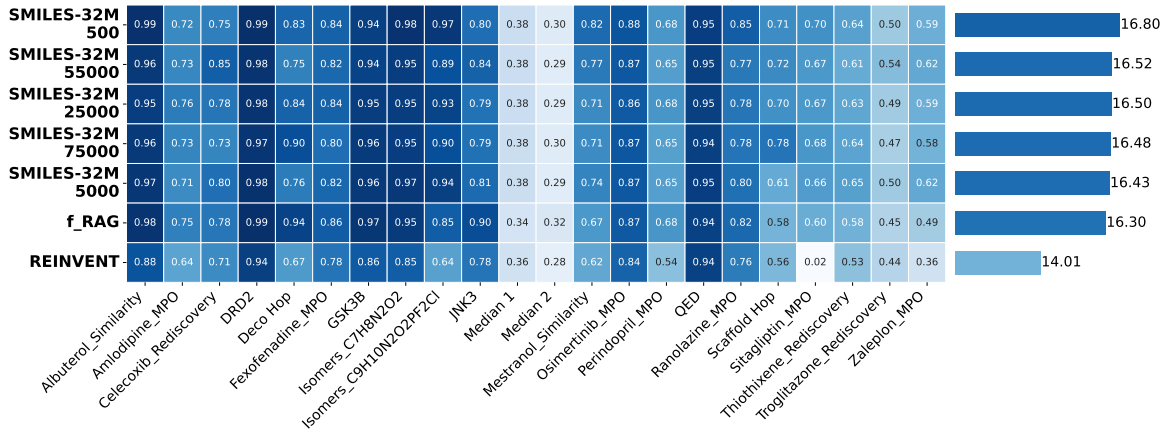


Figure H11. PMO benchmark results for intermediate checkpoints for NovoMolGen-32M (best hyperparameter configuration, atom-wise tokenization).

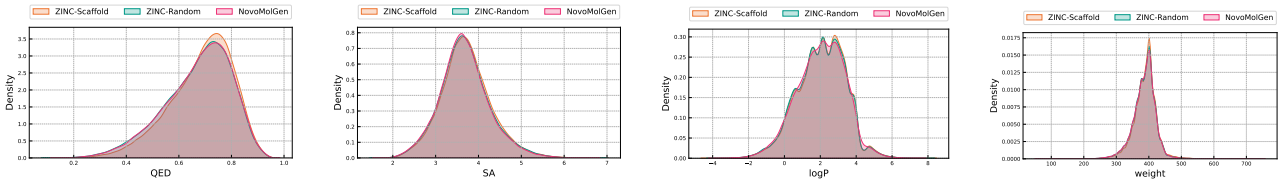
Table H6. PMO top-10 results. The results are the mean and standard deviation of 3 independent runs. Values shown are mean ± std. **Blue** denotes the best performing model, while **Pink** represents the second-best performing model.

Oracle	f-RAG	REINVENT	NovoMolGen-32M (AtomWise)	NovoMolGen-157M (AtomWise)	NovoMolGen-300M (AtomWise)
albuterol_similarity	**0.977 ± 0.002**	0.882 ± 0.006	0.958 ± 0.003	**0.968 ± 0.002**	0.954 ± 0.002
amlodipine_mpo	**0.749 ± 0.019**	0.635 ± 0.035	**0.732 ± 0.024**	0.730 ± 0.037	0.730 ± 0.007
celecoxib_rediscovery	**0.778 ± 0.007**	0.713 ± 0.067	0.732 ± 0.009	0.777 ± 0.132	**0.904 ± 0.022**
drd2	0.936 ± 0.011	0.666 ± 0.044	**0.973 ± 0.001**	**0.983 ± 0.002**	0.972 ± 0.001
deco_hop	**0.992 ± 0.000**	**0.945 ± 0.007**	0.902 ± 0.007	0.853 ± 0.115	0.785 ± 0.130
fexofenadine_mpo	**0.856 ± 0.016**	0.784 ± 0.006	0.802 ± 0.008	0.790 ± 0.028	**0.803 ± 0.018**
gsk3b	**0.969 ± 0.003**	0.865 ± 0.043	**0.955 ± 0.006**	0.949 ± 0.005	0.941 ± 0.012
isomers_c7h8n2o2	**0.955 ± 0.008**	0.852 ± 0.036	0.961 ± 0.001	**0.965 ± 0.002**	0.949 ± 0.005
isomers_c9h10n2o2pf2cl	0.850 ± 0.005	0.642 ± 0.054	0.898 ± 0.032	**0.926 ± 0.024**	**0.915 ± 0.022**
jnk3	**0.904 ± 0.004**	0.783 ± 0.023	0.787 ± 0.014	**0.809 ± 0.066**	0.808 ± 0.045
median1	0.340 ± 0.007	0.356 ± 0.009	**0.380 ± 0.003**	0.354 ± 0.027	**0.379 ± 0.002**
median2	**0.323 ± 0.005**	0.276 ± 0.008	0.300 ± 0.013	0.299 ± 0.001	**0.304 ± 0.019**
mestranol_similarity	0.671 ± 0.021	0.618 ± 0.048	0.712 ± 0.047	**0.739 ± 0.068**	**0.796 ± 0.046**
osimertinib_mpo	0.866 ± 0.009	0.837 ± 0.009	**0.874 ± 0.013**	0.858 ± 0.001	**0.871 ± 0.011**
perindopril_mpo	**0.681 ± 0.017**	0.537 ± 0.016	0.637 ± 0.050	0.649 ± 0.037	**0.668 ± 0.011**
qed	0.939 ± 0.001	0.941 ± 0.000	0.945 ± 0.000	**0.946 ± 0.000**	**0.945 ± 0.000**
ranolazine_mpo	**0.820 ± 0.016**	0.760 ± 0.009	0.784 ± 0.007	**0.797 ± 0.023**	0.779 ± 0.024
scaffold_hop	0.576 ± 0.014	0.560 ± 0.019	**0.783 ± 0.123**	0.735 ± 0.140	**0.791 ± 0.160**
sitagliptin_mpo	0.601 ± 0.011	0.021 ± 0.009	**0.677 ± 0.018**	**0.708 ± 0.038**	0.635 ± 0.043
thiothixene_rediscovery	0.584 ± 0.009	0.534 ± 0.013	**0.635 ± 0.018**	**0.624 ± 0.050**	0.584 ± 0.062
troglitazone_rediscovery	0.448 ± 0.017	0.441 ± 0.032	0.470 ± 0.042	**0.516 ± 0.010**	**0.524 ± 0.058**
zaleplon_mpo	0.486 ± 0.004	0.358 ± 0.062	**0.579 ± 0.023**	0.559 ± 0.013	**0.599 ± 0.000**
Sum	16.30	14.01	16.48	**16.54**	**16.64**

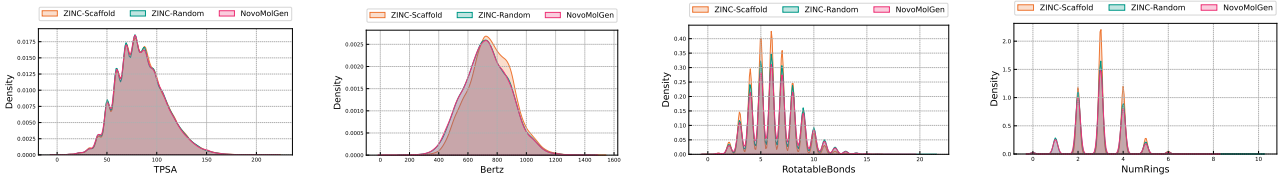
critical to stabilize training and avoid degenerate solutions. In total, 48 unique configurations were tested, resulting in **288 experimental runs** (3 seeds per configuration, 2 tasks) for each model. The Aggregated Score, defined as the sum of the averaged scores for Perindopril_MPO and Zaleplon_MPO, was used as the primary metric for selecting the optimal hyperparameters. We excluded our models from the analysis of the Valsartan_SMARTS task, as no matching patterns were identified in the training dataset, and the model failed to generate any viable molecules conforming to the pattern.

J.2. Results and Analysis

For the **SMILES (atom-wise)** model, we visualized the impact of hyperparameters on the aggregated score using a **Parallel Coordinates Plot** (see Figure J13). In this plot, the most influential hyperparameters are placed to the right, emphasizing their relative importance. We observe that, a higher penalty coefficient (λ) significantly decreases performance, indicating that smaller values are better suited for our model. Also, a higher learning rate consistently improves performance for the SAFE (atom-wise) model.



(a) QED Distribution (b) SA Distribution (c) logP Distribution (d) MW Distribution



(e) TPSA Distribution (f) Bertz Complexity Distribution (g) Rotatable Bonds Distribution (h) NumRings Distribution

Figure I12. Distributions of molecular properties for reference sets (ZINC-Random and ZINC-Scaffold, 175,000 molecules each) and a generated set from NovoMolGen-SMILES-Atom-wise-32M (100,000 molecules). The properties include QED, SA, logP, MW, TPSA, Bertz Complexity, number of rotatable bonds, and number of rings.

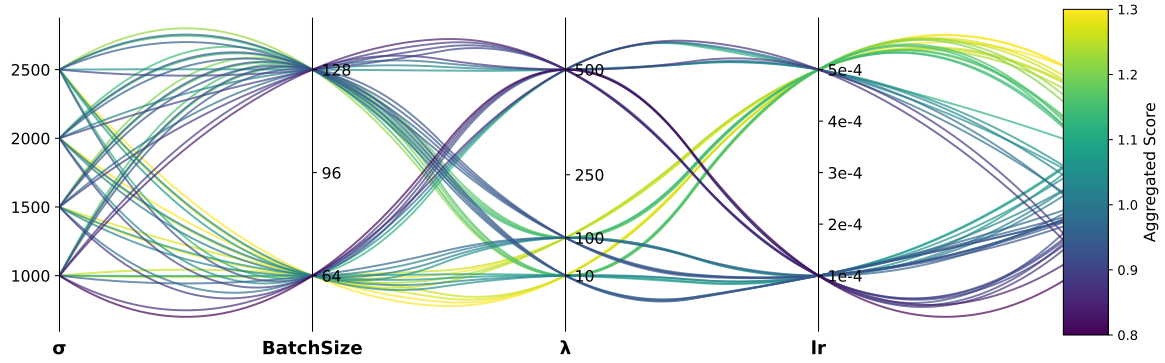


Figure J13. Parallel Coordinates Plot for the **SMILES (AtomWise)** model (32M), showing the importance and effects of hyperparameters on the Aggregated Score. The most influential hyperparameters are positioned on the right side of the plot.

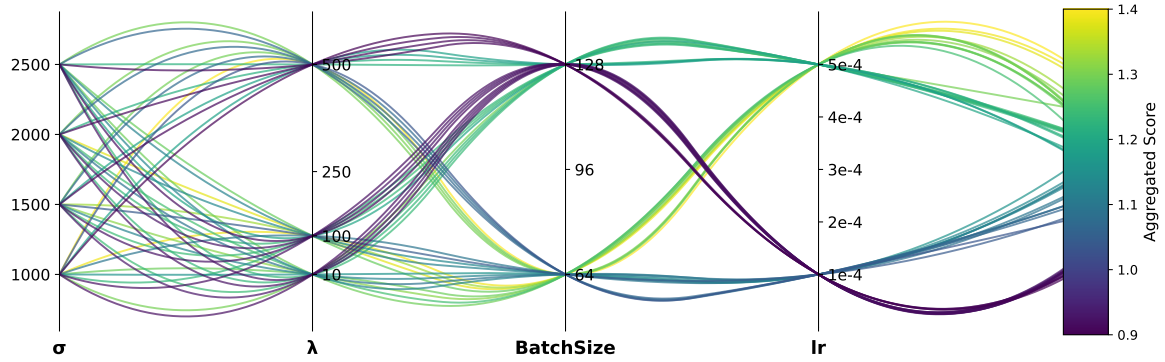


Figure J14. Parallel Coordinates Plot for the **SMILES (BPE)** model (32M).

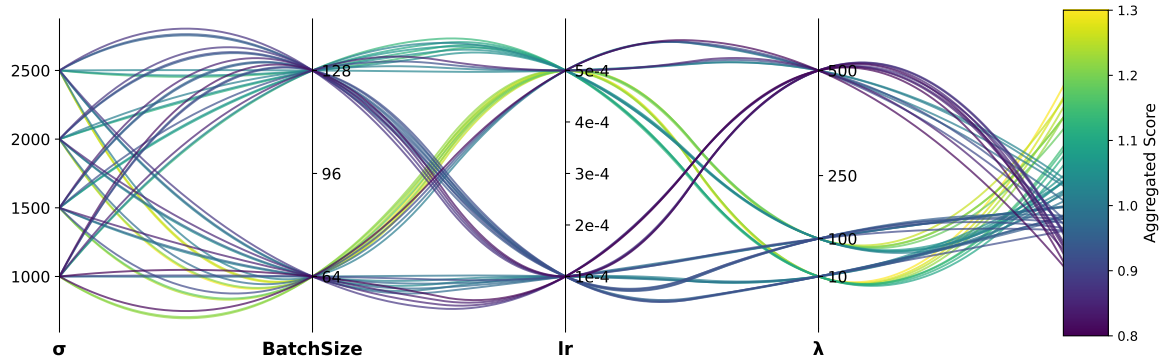


Figure J15. Parallel Coordinates Plot for the **SMILES (AtomWise)** model (157M).

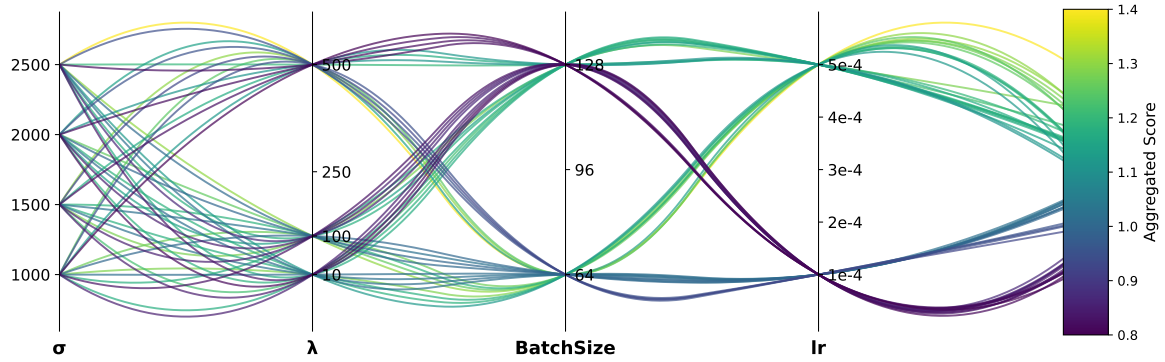


Figure J16. Parallel Coordinates Plot for the **SMILES (BPE)** model (157M).

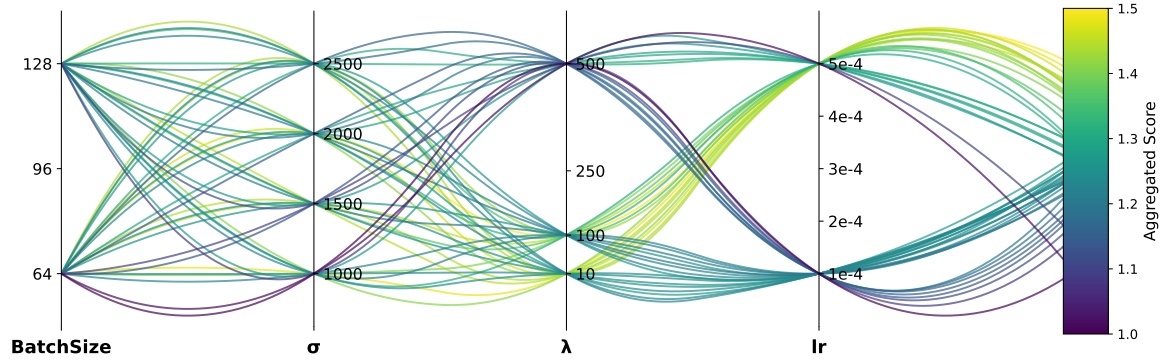


Figure J17. Parallel Coordinates Plot for the **SMILES (AtomWise)** model (300M).

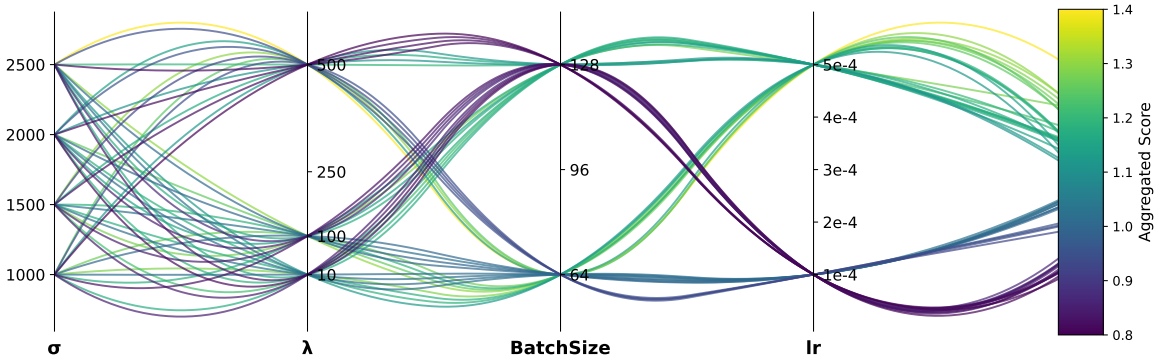


Figure J18. Parallel Coordinates Plot for the **SMILES (BPE)** model (300M).