A Mamba-Based Foundation Model for Chemistry

Emilio Vital Brazil
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
evital@br.ibm.com

Eduardo Soares
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
eduardo.soares@ibm.com

Victor Shirasuna
IBM Research Brazil
São Paulo, SP, Brazil
victor.shirasuna@ibm.com

Renato Cerqueira
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
rcerq@br.ibm.com

Dmitry Zubarev
IBM Research Almaden
San Jose, CA, USA
dmitry.zubarev@ibm.com

Kristin Schmidt
IBM Research Almaden
San Jose, CA, USA
schmidkr@us.ibm.com

Abstract

Large-scale pre-trained foundation models represent a breakthrough in cheminformatics. These methods excel in tasks such as property prediction and molecule generation by learning contextualized representations of input tokens through self-supervised learning on large unlabeled corpora. Most chemical foundation models available are based on the Transformers architecture and its core attention module. The efficacy of self-attention is attributed to its ability to route information densely within a context window, allowing it to model complex data. However, this property brings fundamental drawbacks: an inability to model anything outside of a finite window and quadratic scaling with respect to the window length. Structured state space sequence models (SSMs) have recently emerged as a promising class of architectures for sequence modeling. Mamba is a simplified end-to-end SSM-based neural network architecture without attention or even MLP blocks. This paper introduces a Mamba-based chemical foundational models pre-trained on a curated dataset of 91 million SMILES samples sourced from PubChem, which is equivalent to 4 billion of molecular tokens. These models support different complex tasks, including molecular properties prediction, classification, molecular reconstruction, and synthesis yield prediction. Our experiments across multiple benchmark datasets validate the SSM's capacity of providing state-of-the-art results while is designed for fast inference.

1 Introduction

Large-scale pre-training methodologies for chemical language models (LMs) represent a significant advancement in cheminformatics [1]. These methodologies have shown impressive results in challenging molecular tasks such as predicting properties and generating molecules [2, 3]. The success of these models can be attributed to their ability to learn contextualized representations of input tokens through self-supervised learning on large unlabeled corpora [4].

Most chemical foundation models available are based on the Transformers architecture and its core attention module [5, 6, 7, 8]. The efficacy of self-attention is attributed to its ability to route information densely within a context window [9], allowing it to model complex data [10]. However, this property brings fundamental drawbacks as the inability to model anything outside of a finite window and quadratic scaling with respect to the window length [11]. A substantial amount of research has emerged on more efficient variants of attention to overcome these drawbacks [12].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Structured state space sequence models (SSMs) have recently emerged as a promising class of architectures for sequence modeling [13]. These models can be interpreted as a combination of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [14]. This class of models can be computed very efficiently as either a recurrence or convolution, with linear or near-linear scaling in sequence length. Mamba is a simplified end-to-end SSM-based neural network architecture without attention or even MLP blocks [15, 16]. Mamba enjoys fast inference and linear scaling in sequence length [15].

In this study, we present a novel Mamba-based large foundation model, denoted as O_{SMI}-SSM-336M. Our O_{SMI}-SSM-336M encoder-decoder foundation model was obtained using an efficient encoder SSM-based model aligned with an auto-encoder mechanism pre-trained on a large corpus of 91 million carefully curated molecules from PubChem [17], resulting in 4 billion molecular tokens. Our main contributions are:

- We pre-train a large-scale Mamba-based foundation model for molecules, denoted as O_{SMI}-SSM-336M, on over 91 million molecules carefully curated from PubChem [17], which is equivalent to 4 billion of molecular tokens.
- We perform extensive experimentation on several classification and regression tasks from 11 benchmark datasets, covering quantum mechanical, physical, biophysical, and physiological property prediction of small molecules.
- We evaluate the model's ability to predict chemical reaction yields in synthetic and process chemistry using the Buchwald–Hartwig cross-coupling reaction dataset. Reaction yields refer to the percentage of input materials (reactants) that are converted into output materials (products).
- We evaluate the reconstruction capacity of our O_{SMI}-SSM-336M considering the MOSES benchmarking dataset [18].
- Additionally, we present a study comparing the inference speed of our proposed Mamba-based model with a Transformer-based model in predicting HOMO-LUMO properties for 10 million samples randomly selected from PubChem.

Our results section demonstrates that O_{SMI}-SSM-336M achieves state-of-the-art (SOTA) performance across various tasks, including molecular property prediction, chemical reaction yield prediction, and molecule reconstruction. Furthermore, the findings indicate that the proposed model strikes a balanced equilibrium between inference speed and performance, delivering state-of-the-art results in half the time required by Transformer-based architectures.

2 Overview of the proposed approach

This section provides an overview of the proposed Mamba-based O_{SMI}-SSM-336M foundation model for chemistry. We detail the process of collecting, curating, and pre-processing the pre-training data, along with the token encoding and SMILES encoder-decoder processes. Figure 1 illustrates the general architecture of the base model.

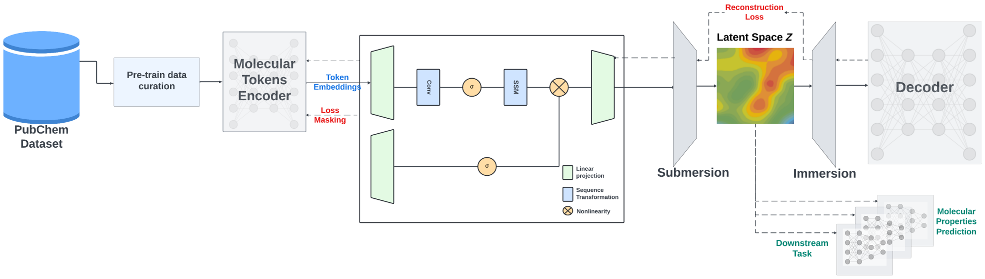

Figure 1: This figure illustrates the general architecture of the base O_{SMI}-SSM-336M model.

2.1 Pre-training Data

The pretraining data was sourced from the PubChem data repository, a public database containing information on chemical substances and their biological activities [17]. Initially, 113 million SMILES strings were collected from PubChem. These molecular strings underwent deduplication and canonicalization to ensure uniqueness [19]. Following this, a molecular transformation process was applied to validate the molecules derived from the unique SMILES strings, resulting in a final set of 91 million unique and valid molecules.

To construct the vocabulary, we utilized the molecular tokenizer proposed by [20]. The tokenization process was applied to all 91 million curated molecules from PubChem, yielding a set of 4 billion molecular tokens. From this output, we extracted 2,988 unique tokens, along with 5 special tokens. In contrast, MoLFormer, which was trained on 1 billion samples with minimal curation, generated a vocabulary of 2,362 tokens using the same tokenization method [2]. This indicates that our curation process led to an enhanced vocabulary model. Detailed statistics of the pre-training dataset are provided in Table 1.

Table 1: Pre-training dataset statistics.

Property	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Number of Atoms	48.95	45.19	1.00	30.00	40.00	53.00	1687.00
Molecular Weight (Daltons)	344.15	137.79	1.01	265.32	330.37	402.47	18838.70
LogP	3.18	2.18	-88.97	2.12	3.29	4.36	59.81
Number of H-Bond Acceptors	4.29	2.62	0	3.00	4.00	5.00	191
Number of H-Bond Donors	1.18	1.48	0	0.00	1.00	2.00	116
Number of Rotatable Bonds	4.79	4.09	0	3.00	4.00	6.00	240
Topological Polar Surface Area	67.81	50.11	0	40.54	61.77	84.22	4201.50
Number of Aliphatic Rings	0.72	1.07	0	0.00	0.00	1.00	54
Number of Aromatic Rings	1.96	1.24	0	1.00	2.00	3.00	32

2.2 Model Architecture

We conduct training for O$_{\text{SMI}}$-SSM-336M employing a Mamba-based encoder for tokens and an encoder-decoder architecture for SMILES embeddings space. The hyper-parameters of the model are detailed in Table 2.

Table 2: O$_{\text{SMI}}$_289M base architecture specificity.

Hidden size	Layers	dt rank	d state	d conv	expand factor	dt min	dt max	dt scale	dt init floor
768	24	auto	16	4	2	0.001	0.1	1.0	1e-4

conv bias	bias	lr start	lr multiplier	Vocab size	# SMILES	# Mol tokens	# Encoder	# Decoder	Total params
True	False	3e-5	1	2993	91M	4B	94M	242M	336M

Mamba models originates from a continuous-time system that maps an input function or sequence $x(t) \in \mathbb{R}^M$ to an output response signal $y(t) \in \mathbb{R}^O$ through an implicit latent state $h(t) \in \mathbb{R}^N$ which can be mathematically formulated using the following ordinary differential equations.

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t) + Dx(t) \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{N \times N}$ and $C \in \mathbb{R}^{O \times N}$ control how the current state evolves over time and translates to the output, $B \in \mathbb{R}^{N \times M}$ and $D \in \mathbb{R}^{O \times M}$ depict how the input influences the state and the output, respectively.

The tokens extracted from SMILES trough the SSM encoder are embedded in a 768-dimensional space. The encoder-decoder layer is designed to process molecular token embeddings, represented as $\mathbf{x} \in \mathbb{R}^{T \times L}$, where T denotes the maximum number of tokens and L represents the embedding space dimension. We limited T at 202 tokens, as 99.4% of molecules in the PubChem dataset contain fewer tokens than this threshold [2].

In encoder-only models, a mean pooling layer is typically employed to represent tokens as SMILES in the latent space [21]. However, this approach is limited by the lack of a natural inversion process for the mean pooling operation. To overcome this limitation, we aim to construct a latent space representation for SMILES by submersing the $\mathbf{x}$ in a latent space, denoted as $\mathbf{z}$, as described in Eq. 2.

$$\mathbf{z} = (\text{LayerNorm}(\text{GELU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)))\,\mathbf{W}_2, \tag{2}$$

where $\mathbf{z} \in \mathbb{R}^L$, $\mathbf{W}_1 \in \mathbb{R}^L$, $\mathbf{b}_1 \in \mathbb{R}^L$, $\mathbf{W}_2 \in \mathbb{R}^{L\times L}$, with L denoting the latent space size (specifically, $L = 768$). Subsequently, we can immerse $\mathbf{z}$ back by calculating Eq. 3.

$$\hat{\mathbf{x}} = (\text{LayerNorm}(\text{GELU}(\mathbf{z}\mathbf{W}_3 + \mathbf{b}_3)))\,\mathbf{W}_4 \tag{3}$$

where $\hat{\mathbf{x}} \in \mathbb{R}^{T\times L}$, $\mathbf{W}_3 \in \mathbb{R}^{L\times L}$, $\mathbf{b}_3 \in \mathbb{R}^L$, $\mathbf{W}_4 \in \mathbb{R}^{L\times T}$. Where T representing the output feature space size (namely, $T = 202$).

A language layer (decoder) is used to process $\hat{\mathbf{x}}$, where it applies non-linearity and normalization, and projects the resulting vector into a set of logits over the vocabulary, which can then be used to predict the next token in the molecular [22]. This architecture serves as a tool for dimensionality reduction and representation learning in the domain of molecular structures.

2.3 Pre-training strategies

Pre-training of O_{SMI}-SSM-336M was performed for 130 epochs through the entire curated PubChem dataset with a fixed learning rate of 3e-5 and a batch size of 128 molecules on a total of 24 NVIDIA V100 (16G) GPUs parallelized into 4 nodes using DDP and *torch run*. It involves two distinct phases: i) Learning of token embeddings through a masking process; ii) Subsequently, the token embeddings are mapped into a common latent space that encapsulates the entire SMILES string. This latent space not only facilitates the representation of the SMILES but also enables the reconstruction of both individual tokens and complete SMILES strings. Consequently, the pre-training process involves two separate loss functions: one for the token embeddings, which is based on the masking process, and another for the encoder-decoder layer, which focuses on the reconstruction of tokens. Two pre-training strategies are employed:

- In phase 1, the token encoder is initially pre-trained using 95% of the available samples, while the remaining 5% is reserved for training the encoder-decoder layer. This partitioning is necessary as the token embeddings may encounter convergence difficulties in the initial epochs, which could adversely affect the training of the encoder-decoder layer.
- In phase 2, once the token embeddings layer has achieved convergence, the pre-training process is expanded to utilize 100% of the available samples for both phases. This approach leads to an enhancement in the performance of the encoder-decoder layer, particularly in terms of token reconstruction.

For encoder pre-training we use the masked language model method defined in [23]. Initially 15% of the tokens are selected for possible learning. From that selection, 80% of the tokens are randomly selected and replaced with the [MASK] token, 10% of the tokens are randomly selected to be replaced with a random token, while the remaining 10% of the tokens will be unchanged.

The adoption of different pre-training strategies has proven instrumental in enhancing the efficiency of our model, as evidenced by improvements observed in the loss functions.

3 Experiments

To evaluate the effectiveness of our proposed Mamba-based model O_{SMI}-SSM-336M, we conducted experiments using a set of 11 datasets sourced from MoleculeNet [24] as demonstrated in Table 3. Specifically, we evaluated 6 datasets for classification task and 5 datasets for regression tasks. To ensure an unbiased assessment, we maintained consistency with the original benchmark by adopting identical train/validation/test splits for all tasks [24]. We also conducted the experiments considered 10 different seeds for all the tests in other to guarantee the robustness of the approach.

We also conducted high-throughput experiments on Pd-catalyzed Buchwald–Hartwig C–N cross-coupling reactions, measuring the yields for each reaction as described in [25]. The experiments

Table 3: Evaluated datasets description

Dataset	Description	# compounds	# tasks	Metric
BBBP	Blood brain barrier penetration dataset	2039	1	ROC-AUC
HIV	Ability of small molecules to inhibit HIV replication	41127	1	ROC-AUC
BACE	Binding results for a set of inhibitors for β – secretase 1	1513	1	ROC-AUC
Clintox	Clinical trial toxicity of drugs	1478	2	ROC-AUC
SIDER	Drug side effect on different organ classes	1427	27	ROC-AUC
Tox21	Toxicity measurements on 12 different targets	7831	12	ROC-AUC
QM9	12 quantum mechanical calculations	133885	12	Average MAE
QM8	12 excited state properties of small molecules	21786	12	Average MAE
ESOL	Water solubility dataset	1128	1	RMSE
FreeSolv	Hydration free energy of small molecules in water	642	1	RMSE
Lipophilicity	Octanol/water distribution coefficient of molecules	4200	1	RMSE

utilized three 1536-well plates, covering a matrix of 15 aryl and heteroaryl halides, four Buchwald ligands, three bases, and 23 isoxazole additives, resulting in a total of 3,955 reactions. We employed the same data splits as in [25] to assess our model's performance with training sets of varying sizes.

To evaluate the reconstruction and decoder capabilities of OSMI-SSM-336M, we utilized the MOSES benchmarking dataset [18], which contains 1,936,962 molecular structures. For the experiments, we adopted the dataset split proposed by [18], dividing it into training, test, and scaffold test sets, comprising approximately 1.6 million, 176,000, and 176,000 molecules, respectively. The scaffold test set includes unique Bemis-Murcko scaffolds that are absent in the training and test sets, allowing us to assess the model's ability to generate previously unobserved scaffolds. Finally, we evaluated the inference speed of OSMI-SSM-336M by predicting HOMO-LUMO properties for 10 million samples randomly selected from PubChem.

4 Results and Discussion

In this section, we present an analysis of the results obtained using OSMI-SSM-336M across various experiments conducted with different versions of the base model. The analysis includes: (i) A comparison between frozen and fine-tuned versions of OSMI-SSM-336M, along with a comparison against state-of-the-art models on various benchmarking datasets for molecular classification and regression tasks; (ii) An evaluation of O$_{\text{SMI}}$-SSM-336M for predicting chemical reaction yields; (iii) An assessment of the Decoder module using the MOSES benchmarking dataset; and (iv) A study comparing the inference speed for predicting HOMO-LUMO properties on 10 million samples randomly selected from PubChem.

4.1 Comparison with SOTA on benchmarking tasks

Results for classification tasks: The analysis evaluates the comparative performance of O$_{\text{SMI}}$-SSM-336M in its fine-tuned and frozen states relative to state-of-the-art algorithms for molecular property classification, as detailed in Table 4.

Table 4: Methods and Performance for the classification tasks of MoleculeNet benchmark datasets

Method	Dataset					
	BBBP	ClinTox	HIV	BACE	SIDER	Tox21
GraphMVP [26]	72.4±1.6	79.1±2.8	77.0±1.2	81.2±0.9	63.9±1.2	75.9±0.5
GEM [27]	72.4±0.4	90.1±1.3	80.6±0.9	85.6±1.1	**67.2±0.4**	78.1±0.1
GROVER$_{\text{Large}}$ [28]	69.5±0.1	76.2±3.7	68.2±1.1	81.0±1.4	65.4±0.1	73.5±0.1
ChemBerta [6]	64.3	90.6	62.2	-	-	-
ChemBerta2 [29]	71.94	90.7	-	85.1	-	-
Galatica 30B [30]	59.6	82.2	75.9	72.7	61.3	68.5
Galatica 120B [30]	66.1	82.6	74.5	61.7	63.2	68.9
Uni-Mol [31]	72.9±0.6	91.9±1.8	80.8±0.3	85.7±0.2	65.9±1.3	79.6±0.5
MolFM [31]	72.9±0.1	79.7±1.6	78.8±1.1	83.9±1.1	64.2±0.9	77.2±0.7
MoLFormer [32]	73.6±0.8	91.2±1.4	80.5±1.65	86.3±0.6	65.5±0.2	80.46±0.2
SMI-TED289M (Frozen Weights) [33]	91.46±0.47	93.49±0.85	80.51±1.34	85.58±0.92	66.01±0.88	81.53 ±0.45
SMI-TED289M (Fine-tuned) [33]	92.26±0.57	**94.27±1.83**	76.85±0.89	**88.24±0.50**	65.68±0.45	81.85±1.42
O$_{\text{SMI}}$-SSM-336M (Frozen)	90.81 ±0.85	86.36 ±0.74	77.04 ±0.64	83.83 ±0.76	63.52 ±0.3	81.42 ±0.8
O$_{\text{SMI}}$-SSM-336M(Fine-tuned)	**92.81 ±0.27**	90.02 ±0.5	**83.14 ±0.34**	86.12 ±0.96	63.17 ±0.75	**83.84 ±0.2**

Table 4 summarizes the performance of various advanced methods across several benchmarking datasets used for molecular classification tasks. OSMI-SSM-336M demonstrates comparative efficacy against Transformer-based approaches, outperforming them in three out of six datasets. Notably, O$_{\text{SMI}}$-SSM-336M with its initial configuration yields results on par with current state-of-the-art methods. Further fine-tuning of O$_{\text{SMI}}$-SSM-336M enhances its performance, indicating its substantial potential for accurate molecular classification and suggesting that additional performance gains may be achieved through further optimization.

Results for regression tasks: Subsequently, we applied O$_{\text{SMI}}$-SSM-336M to the prediction of chemical properties. The performance metrics across five regression benchmarks—QM9, QM8, ESOL, FreeSolv, and Lipophilicity—are presented in Table 5.

Table 5: Methods and Performance for the regression tasks of MoleculeNet benchmark datasets.**Blue** and **Orange** indicates best and second-best performing model, respectively.

Method	Dataset				
	QM9	QM8	ESOL	FreeSolv	Lipophilicity
D-MPNN [34]	3.241±0.119	0.0143±0.0022	0.98±0.26	2.18±0.91	0.65±0.05
N-Gram [35]	2.51±0.19	0.032±0.003	1.074±0.107	2.688±0.085	0.812±0.028
PretrainGNN [36]	-	-	1.100±0.006	2.764±0.002	0.739±0.003
GROVER$_{\text{Large}}$ [28]	-	-	0.895±0.017	2.272±0.051	0.823±0.010
ChemBERTa-2 [29]	-	-	0.89	-	0.80
SPMM [32]	-	-	0.818±0.008	1.907±0.058	0.692±0.008
MolCLR$_{\text{GIN}}$ [37]	2.357±0.118	0.0174±0.0013	1.11±0.01	2.20±0.20	0.65±0.08
Hu et al. [38]	4.349±0.061	0.0191±0.0003	1.22±0.02	2.83±0.12	0.74±0.00
MoLFormer [32]	**1.5894±0.0567**	0.0102	0.880±0.028	2.342±0.052	0.700±0.012
SMI-TED289M [33]	**1.3246±0.0157**	**0.0095±0.0001**	**0.6112±0.0096**	**1.2233±0.0029**	**0.5522±0.0194**
O$_{\text{SMI}}$-SSM-336M(Frozen)	8.9546 ±0.0577	0.0194 ±0.0003	0.8135 ±0.0253	1.6374 ±0.0682	0.746 ±0.0029
O$_{\text{SMI}}$-SSM-336M (Fine-tuned)	2.2175 ±0.3194	**0.0104 ±0.0001**	**0.7222±0.0139**	**1.6288±0.0347**	**0.6048±0.0023**

Results presented in Table 5 indicate that O$_{\text{SMI}}$-SSM-336M achieves performance comparable to state-of-the-art models, securing the second-best results in four of the five regression benchmarks evaluated. This demonstrates the efficacy of the Mamba-based approach in delivering results on par with Transformer-based methods, while also highlighting its robustness across a range of chemical property prediction tasks. The design of O$_{\text{SMI}}$-SSM-336M aims to strike an optimal balance between predictive accuracy and inference efficiency. To exemplify this balance, we provide an analysis comparing the inference time for predicting HOMO-LUMO properties on a dataset of 10 million samples randomly selected from PubChem. This study underscores the model's capability to maintain high prediction accuracy while significantly reducing computational time, thereby offering practical advantages for large-scale chemical property predictions.

Speed inference for HUMO-LUMO properties prediction: To assess the inference speed of the proposed Mamba-based approach, we conducted predictions of HOMO-LUMO properties for 10 million samples randomly selected from PubChem. For comparison, we evaluated the inference time of SMI-TED289M, a Transformer-based model recognized for its state-of-the-art performance. Figure 2 illustrates the superior inference speed of O$_{\text{SMI}}$-SSM-336M compared to SMI-TED289M. Specifically, SMI-TED289M required 20,606.76 seconds for HOMO property predictions and 21,038.43 seconds for LUMO property predictions using a single NVIDIA V100 32GB GPU. In contrast, O$_{\text{SMI}}$-SSM-336M completed HOMO predictions in 9,735.64 seconds and LUMO predictions in 9,823.64 seconds on the same GPU. These results highlight the substantial efficiency gains of the O$_{\text{SMI}}$-SSM-336M model in terms of inference speed.

The Mamba-base approach demonstrates a substantial improvement in efficiency, being approximately 54% faster and reducing GPU usage by 6 hours, while also decreasing CO2 emissions by an average of 0.78 kg equivalent [39]. This reduction in computational resources is crucial for minimizing the environmental impact of machine learning models, which requires significant energy consumption and associated carbon footprints [40].

4.2 Reaction-yield prediction

Previously, we were able to show that the proposed Mamba-based model was able to perform compared to transformer-based methods on single molecule properties prediction. Here, we investigate the Mamba-based approach on chemical reactions. Chemical reactions in organic chemistry are described by writing the structural formula of reactants and products separated by an arrow, representing the chemical transformation by specifying how the atoms rearrange between one or several reactant

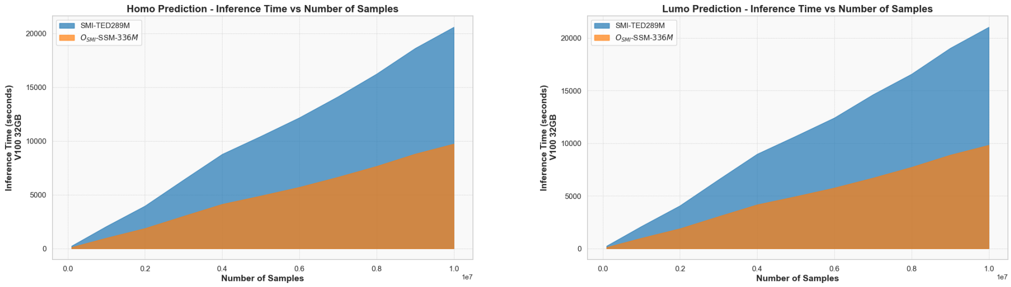

Figure 2: The figure shows the inference speed for O$_{SMI}$-SSM-336M and SMI-TED289M for HOMO-LUMO predictions considering a dataset of 10M samples randomly selected from PubChem and a single NVIDIA V100 32GB GPU.

molecules and one or several product molecules. Predicting outcomes of chemical reactions, such as their yield based on data gathered in high-throughput screening, is an important task in machine learning for chemistry. Fig. 3 the schema for chemical reaction.

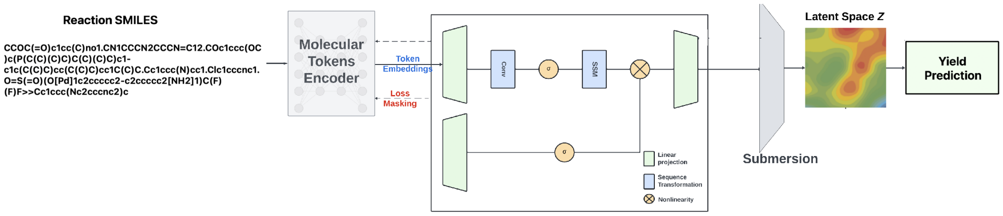

Figure 3: This figure illustrates the schema for chemical reaction yield prediction based on reaction SMILES considering the O$_{SMI}$-SSM-336M model.

We assessed this architecture against state-of-the-art methods using a high-throughput dataset of Buchwald–Hartwig cross-coupling reactions, focusing on predicting reaction yields [25]. This involves estimating the percentage of reactants converted into products. Our evaluation adhered to the schema and data divisions outlined in [25]. Table 6 presents the results for the O$_{SMI}$-SSM-336M model and compares its performance with existing state-of-the-art approaches.

Table 6: Performance of O$_{SMI}$-SSM-336M compared with the state of the art in reaction-yield prediction on experimentally determined yields of Buchwald–Hartwig reactions through HTEs.

Subset/Split	DFT	Yield-BERT	Yield-BERT (Aug)	DRFP	YieldGNN	MSR2-RXN	**Ours**
Rand 70/30	0.92	0.95±0.005	0.97±0.003	0.95±0.005	0.96±0.005	0.94±0.005	**0.9823 ±0.0007**
Rand 50/50	0.9	0.92±0.01	0.95±0.01	0.93±0.01	-	0.93±0.01	**0.982 ±0.0004**
Rand 30/70	0.85	0.88±0.01	0.92±0.01	0.89±0.01	-	0.90±0.01	**0.978 ±0.0013**
Rand 20/80	0.81	0.86±0.01	0.89±0.01	0.87±0.01	-	0.87±0.01	**0.973 ±0.0006**
Rand 10/90	0.77	0.79±0.02	0.81±0.02	0.81±0.01	-	0.80±0.02	**0.952 ±0.0023**
Rand 5/95	0.68	0.61±0.04	0.74±0.03	0.73±0.02	-	0.69±0.03	**0.903 ±0.0043**
Rand 2.5/97.5	0.59	0.45±0.05	0.61±0.04	0.62±0.04	-	0.57±0.05	**0.846 ±0.0044**
Test 1	0.8	0.84±0.01	0.80±0.01	0.81±0.01	-	0.83±0.03	**0.9827 ±0.0002**
Test 2	0.77	0.84±0.03	0.88±0.02	0.83±0.003	-	0.83±0.01	**0.9827 ±0.0005**
Test 3	0.64	0.75±0.04	0.56±0.08	0.71±0.001	-	0.69±0.04	**0.9823 ±0.0012**
Test 4	0.54	0.49±0.05	0.43±0.04	0.49±0.004	-	0.51±0.04	**0.9825 ±0.0008**
Average 1-4	0.69	0.73	0.58±0.33	0.71±0.16	-	0.72±0.15	**0.9826 ±0.0005**

The results presented in Table 6 clearly demonstrate the superiority of the proposed Mamba-based foundation model when benchmarked against state-of-the-art methods, including gradient-boosting and fingerprint-based approaches (DRFP) [41], a DFT-based random forest model (DFT) [41], and transformer-based models like Yield-BERT [42] and its augmented variant, Yield-BERT(aug.) [42], and MSR2-RXN [43]. The performance of the Mamba-based model can be attributed to its pre-training on an expansive dataset of 91 million curated molecules, which provides a robust foundation of chemical knowledge that significantly enhances its predictive capabilities. This pre-training enables the model to achieve high accuracy even with limited training data, as evidenced by its sustained performance when trained on just 2.5% of the available samples—a scenario where task-specific models experience a marked decline in accuracy. To ensure the robustness of our model, we conducted each experiment with 10 different random seeds.

One key observation is the model's robustness across various data splits, particularly in low-resource settings where only a small fraction of the dataset is used for training. This resilience underscores the importance of leveraging large-scale pre-training to encode generalized chemical knowledge, which can then be fine-tuned for specific tasks like reaction yield prediction. In contrast, models that are tailored specifically for a given task tend to overfit to the nuances of the training data and struggle to generalize when the training set size is reduced, highlighting a critical limitation in their design.

Moreover, the robustness of the Mamba-based model extends to its performance on out-of-domain test sets. The ability to generalize well to data distributions that differ from the training set is a crucial aspect of model evaluation, particularly in real-world applications where the diversity of chemical reactions is vast. The Mamba-based model's consistent performance across both in-domain and out-of-domain test sets illustrates the efficacy of pre-training on a diverse and comprehensive dataset, which equips the model with the flexibility to handle a wide range of chemical environments and reaction conditions.

The comparative analysis between the Mamba-based model and other state-of-the-art methods also sheds light on the limitations of traditional approaches like DFT-based models, which, despite their theoretical grounding in quantum chemistry, may not capture the full complexity of reaction mechanisms in practical scenarios. Similarly, while transformer-based models like Yield-BERT and its augmented variant exhibit strong performance, they fall short of the Mamba-based model, particularly in low-data regimes, indicating that the sheer scale and diversity of the pre-training data play a pivotal role in achieving superior results.

These findings underscore the potential of foundation models in chemistry, where pre-training on large, diverse datasets can serve as a powerful paradigm for developing models that are not only accurate but also robust and generalizable. The implications of this work extend beyond reaction yield prediction, suggesting that similar strategies could be applied to other domains within computational chemistry and materials science, where the ability to generalize across diverse datasets is of paramount importance.

4.3 Decoder evaluation over MOSES benchmarking dataset

Next, conducted a comparative evaluation of the O_{SMI}-SSM-336M model against several baseline models for SMILES reconstruction and decoding, using a test set comprising 176,000 molecules. The evaluation metrics, detailed in Table 7, provide a comprehensive view of the model's performance in key areas such as fragment similarity (Frag), scaffold similarity (Scaf), similarity to the nearest neighbor (SNN), internal diversity (IntDiv), and Fréchet ChemNet Distance (FCD).

Table 7: MOSES benchmarking dataset evaluation.

Metric	Frag ↑	Scaf ↑	SNN ↑	IntDiv ↑	FCD ↓
CharRNN [18]	0.9998	0.9242	0.6015	0.8562	0.0732
VAE [18]	0.9984	0.9386	0.6257	0.8558	0.0990
JT-VAE [44]	0.9965	0.8964	0.5477	0.8551	0.3954
LIMO [45]	0.6989	0.0079	0.2464	**0.9039**	26.78
MolGen-7b [46]	0.9999	0.6538	0.5138	0.8617	0.0435
GP-MoLFormer [47]	0.9998	0.7383	0.5045	0.8655	0.0591
O_{SMI}-SSM-336M	**0.9999**	**0.9994**	**0.9960**	0.8561	**0.0025**

The results indicate that O_{SMI}-SSM-336M not only matches but surpasses the performance of state-of-the-art models in generating unique, valid, and novel molecules. Its near-perfect score in the Frag metric highlights its remarkable ability to retain the structural integrity of molecular fragments, a crucial aspect in ensuring the generated molecules remain chemically viable and relevant to real-world applications. This high fragment similarity, coupled with the model's low FCD score, suggests that the distribution of generated molecules closely mirrors that of natural molecules.

In addition to fragment-level accuracy, O_{SMI}-SSM-336M demonstrates superior performance in scaffold similarity (Scaf) and nearest neighbor similarity (SNN). These metrics are particularly important in drug discovery and design, where the preservation of core molecular scaffolds is essential for maintaining biological activity. The model's ability to generate molecules with high scaffold similarity indicates that it can reliably reproduce the core structural features of molecules, which is a requirement for generating candidate compounds that retain their intended biological function.

Another significant finding is the model's performance in internal diversity (IntDiv). While high similarity scores are important, diversity within the generated set is equally crucial, especially in scenarios where a broad exploration of chemical space is required. The O_{SMI}-SSM-336M model achieves a commendable balance, maintaining high similarity metrics while also generating molecules with substantial pairwise dissimilarity. This capability to generate a diverse array of molecules without sacrificing structural integrity makes the model highly valuable for applications in drug discovery, where exploring a wide range of chemical possibilities is often necessary to identify optimal candidates.

Furthermore, when compared to traditional methods such as CharRNN and more advanced approaches like JT-VAE and MolGen-7b, the O_{SMI}-SSM-336M model consistently outperforms across all evaluated metrics. This includes models like LIMO, which, despite its strong internal diversity, fails to match the other metrics, indicating a trade-off in these approaches that O_{SMI}-SSM-336M successfully mitigates. The model's ability to achieve high scaffold similarity while maintaining diverse molecular structures suggests that its pre-training on a large-scale dataset equips it with a broad understanding of chemical space, enabling it to generalize effectively across various molecular configurations.

5 Conclusion

This paper introduces O_{SMI}-SSM-336M, a Mamba-based chemical foundation model pre-trained on a curated dataset of 91 million SMILES samples from PubChem, encompassing 4 billion molecular tokens. The model is designed to achieve a balance between high performance in evaluation metrics and faster inference capabilities.

The efficacy of O_{SMI}-SSM-336M was rigorously assessed across a variety of tasks, including molecular property classification and prediction. The model not only achieved state-of-the-art results but also demonstrated significant efficiency improvements. Specifically, it was approximately 54% faster than existing state-of-the-art Transformer-based approaches, reducing GPU usage by 6 hours and lowering CO2 emissions by an average of 0.78 kg CO2 equivalent [39] during the prediction of HOMO-LUMO gaps for a dataset of 10 million randomly selected samples from PubChem.

We also explored the model's capabilities in predicting chemical reaction outcomes, such as reaction yields based on high-throughput screening data, a critical task in machine learning for chemistry. The consistent performance of the Mamba-based model across both in-domain and out-of-domain test sets underscores the effectiveness of pre-training on a diverse and comprehensive dataset. This pre-training enables the model to adapt to a wide range of chemical environments and reaction conditions. Our comparative analysis revealed that while traditional approaches, such as DFT-based models, are grounded in quantum chemistry, they may not fully capture the complexity of reaction mechanisms in practical scenarios. Similarly, transformer-based models like Yield-BERT and its augmented variant, despite their strong performance, are outperformed by the Mamba-based model, particularly in low-data regimes. This highlights the critical role that large-scale, diverse pre-training data plays in achieving superior results.

Finally, we conducted a comparative evaluation of the O_{SMI}-SSM-336M model against several baseline models for SMILES reconstruction and decoding. The model's performance across diverse metrics demonstrates the importance of leveraging large-scale dataset for pre-training, which can lead to models that not only excel in generating high-quality molecules but also possess the flexibility required to tackle complex challenges in computational chemistry and drug design.

The Mamba-based foundation model presented in this paper offers both flexibility and scalability for a wide range of scientific applications.

References

[1] A. V. Sadybekov and V. Katritch, "Computational approaches streamlining drug discovery," *Nature*, vol. 616, no. 7958, pp. 673–685, 2023.

[2] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, “Large-scale chemical language representations capture molecular structure and properties,” *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.

[3] E. Soares, F. Cipcigan, D. Zubarev, and E. V. Brazil, “A framework for toxic pfas replacement based on gflownet and chemical foundation model,” in *NeurIPS 2023 AI for Science Workshop*, 2023.

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, “On the opportunities and risks of foundation models,” *arXiv preprint arXiv:2108.07258*, 2021.

[5] G. Pesciullesi, P. Schwaller, T. Laino, and J.-L. Reymond, “Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates,” *Nature communications*, vol. 11, no. 1, p. 4874, 2020.

[6] S. Chithrananda, G. Grand, and B. Ramsundar, “Chemberta: large-scale self-supervised pretraining for molecular property prediction,” *arXiv preprint arXiv:2010.09885*, 2020.

[7] N. Janakarajan, T. Erdmann, S. Swaminathan, T. Laino, and J. Born, “Language models in molecular discovery,” *arXiv preprint arXiv:2309.16235*, 2023.

[8] E. Soares, E. V. Brazil, K. F. A. Gutierrez, R. Cerqueira, D. Sanders, K. Schmidt, and D. Zubarev, “Beyond chemical language: A multimodal approach to enhance molecular property prediction,” *arXiv preprint arXiv:2306.14919*, 2023.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[10] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, “Efficient transformers: A survey,” *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.

[11] T. Lin, Y. Wang, X. Liu, and X. Qiu, “A survey of transformers,” *AI open*, vol. 3, pp. 111–132, 2022.

[12] E. Kotei and R. Thirunavukarasu, “A systematic review of transformer-based pre-trained language models through self-supervised learning,” *Information*, vol. 14, no. 3, p. 187, 2023.

[13] A. Gu, K. Goel, and C. Ré, “Efficiently modeling long sequences with structured state spaces,” *arXiv preprint arXiv:2111.00396*, 2021.

[14] J. T. Smith, A. Warrington, and S. W. Linderman, “Simplified state space layers for sequence modeling,” *arXiv preprint arXiv:2208.04933*, 2022.

[15] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.

[16] E. A. Soares, V. Shirasuna, E. A. V. Brazil, R. F. de Gusmao Cerqueira, D. Zubarev, T. Callahan, and S. Capponi, “Molmamba: A large state-space-based foundation model for chemistry,” in *American Chemical Society (ACS) Fall Meeting*, 2024.

[17] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu *et al.*, “Pubchem 2023 update,” *Nucleic acids research*, vol. 51, no. D1, pp. D1373–D1380, 2023.

[18] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov *et al.*, “Molecular sets (moses): a benchmarking platform for molecular generation models,” *Frontiers in pharmacology*, vol. 11, p. 565644, 2020.

[19] E. Heid, J. Liu, A. Aude, and W. H. Green, “Influence of template size, canonicalization, and exclusivity for retrosynthesis and reaction prediction applications,” *Journal of Chemical Information and Modeling*, vol. 62, no. 1, pp. 16–26, 2021.

[20] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, “Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction,” *ACS central science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[21] A. M. Bran and P. Schwaller, “Transformers and large language models for chemistry and drug discovery,” *arXiv preprint arXiv:2310.06083*, 2023.

[22] J. Ferrando, G. I. Gállego, I. Tsiamas, and M. R. Costa-jussà, “Explaining how transformers use context to build predictions,” *arXiv preprint arXiv:2305.12535*, 2023.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[24] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, “Moleculenet: a benchmark for molecular machine learning,” *Chemical science*, vol. 9, no. 2, pp. 513–530, 2018.

[25] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle, “Predicting reaction performance in c–n cross-coupling using machine learning,” *Science*, vol. 360, no. 6385, pp. 186–190, 2018.

[26] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, and J. Tang, “Pre-training molecular graph representation with 3d geometry,” *arXiv preprint arXiv:2110.07728*, 2021.

[27] X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu, and H. Wang, “Geometry-enhanced molecular representation learning for property prediction,” *Nature Machine Intelligence*, vol. 4, no. 2, pp. 127–134, 2022.

[28] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, and J. Huang, “Self-supervised graph transformer on large-scale molecular data,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 559–12 571, 2020.

[29] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, “Chemberta-2: Towards chemical foundation models,” *arXiv preprint arXiv:2209.01712*, 2022.

[30] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, “Galactica: A large language model for science,” *arXiv preprint arXiv:2211.09085*, 2022.

[31] G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang, and G. Ke, “Uni-mol: a universal 3d molecular representation learning framework,” *ChemRxiv preprint*, 2023.

[32] J. Chang and J. C. Ye, “Bidirectional generation of structure and properties through a single molecular foundation model,” *Nature Communications*, vol. 15, no. 1, p. 2323, 2024.

[33] E. Soares, V. Shirasuna, E. V. Brazil, R. Cerqueira, D. Zubarev, and K. Schmidt, “A large encoder-decoder family of foundation models for chemical language,” *arXiv preprint arXiv:2407.20267*, 2024.

[34] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea *et al.*, “Analyzing learned molecular representations for property prediction,” *Journal of chemical information and modeling*, vol. 59, no. 8, pp. 3370–3388, 2019.

[35] S. Liu, M. F. Demirel, and Y. Liang, “N-gram graph: Simple unsupervised representation for graphs, with applications to molecules,” *Advances in neural information processing systems*, vol. 32, 2019.

[36] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, “Strategies for pre-training graph neural networks,” *arXiv preprint arXiv:1905.12265*, 2019.

[37] Y. Wang, J. Wang, Z. Cao, and A. Barati Farimani, “Molecular contrastive learning of representations via graph neural networks,” *Nature Machine Intelligence*, vol. 4, no. 3, pp. 279–287, 2022.

[38] Z. Hu, Y. Dong, K. Wang, K.-W. Chang, and Y. Sun, “Gpt-gnn: Generative pre-training of graph neural networks,” in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 1857–1867.

[39] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, “Quantifying the carbon emissions of machine learning,” *arXiv preprint arXiv:1910.09700*, 2019.

[40] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland, “Risks and benefits of large language models for the environment,” *Environmental Science & Technology*, vol. 57, no. 9, pp. 3464–3466, 2023.

[41] D. Probst, P. Schwaller, and J.-L. Reymond, “Reaction classification and yield prediction using the differential reaction fingerprint drfp,” *Digital discovery*, vol. 1, no. 2, pp. 91–97, 2022.

[42] P. Schwaller, A. C. Vaucher, T. Laino, and J.-L. Reymond, “Prediction of chemical reaction yields using deep learning,” *Machine learning: science and technology*, vol. 2, no. 1, p. 015016, 2021.

[43] M. Boulougouri, P. Vandergheynst, and D. Probst, “Molecular set representation learning,” *Nature Machine Intelligence*, pp. 1–10, 2024.

[44] W. Jin, R. Barzilay, and T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," in *International conference on machine learning*. PMLR, 2018, pp. 2323–2332.

[45] P. Eckmann, K. Sun, B. Zhao, M. Feng, M. K. Gilson, and R. Yu, "Limo: Latent inceptionism for targeted molecule generation," *Proceedings of machine learning research*, vol. 162, p. 5777, 2022.

[46] Y. Fang, N. Zhang, Z. Chen, L. Guo, X. Fan, and H. Chen, "Domain-agnostic molecular generation with self-feedback," *arXiv preprint arXiv:2301.11259*, 2023.

[47] J. Ross, B. Belgodere, S. C. Hoffman, V. Chenthamarakshan, Y. Mroueh, and P. Das, "Gp-molformer: A foundation model for molecular generation," *arXiv preprint arXiv:2405.04912*, 2024.