https://doi.org/10.1038/s44387-025-00009-7

A Mamba-based foundation model for materials

Eduardo Soares[1] ✉, Emilio Vital Brazil[1], Victor Shirasuna[1], Dmitry Zubarev[2], Renato Cerqueira[1] & Kristin Schmidt[2]

We present a novel approach to chemical foundation models, leveraging structured state space sequence models (SSMs) to overcome the limitations of traditional Transformer-based architectures. While Transformers have achieved state-of-the-art results in chemical tasks such as property prediction and molecule generation, their self-attention mechanism is constrained by its inability to model data outside of a finite context window and its quadratic scaling with respect to window length. In contrast, SSMs offer a promising alternative for sequence modeling, enabling the capture of complex patterns and dependencies in molecular structures. Our Mamba architecture, a simplified end-to-end SSM-based neural network, eliminates the need for attention and MLP blocks, allowing for faster inference. We pre-train Mamba on a large, curated dataset of 91 million SMILES samples (equivalent to 4 billion molecular tokens) sourced from PubChem, and evaluate its performance on various benchmark datasets. Our experiments demonstrate the SSM's capacity to provide state-of-the-art results while maintaining fast inference, supporting complex tasks such as molecular property prediction, classification, molecular reconstruction, and synthesis yield prediction. This work advances the state-of-the-art in AI methodology in chemical sciences, offering a promising direction for future research in molecular modeling and discovery.

Large-scale pre-training methodologies for chemical language models (LMs) represent a significant advancement in cheminformatics[1]. These methodologies have demonstrated impressive results in challenging molecular tasks such as property prediction and molecule generation[2]. Their success is largely attributable to the ability of these models to learn rich, contextualized representations of input tokens through self-supervised learning on vast unlabeled corpora[3].

Most chemical foundation models currently available are based on the Transformer architecture and its core attention module[4–6]. The efficacy of self-attention lies in its capacity to densely route information within a fixed context window[7], enabling the modeling of complex chemical data[8]. However, this mechanism is inherently limited by its inability to incorporate information outside the finite context window, and it suffers from quadratic scaling with respect to sequence length[9]. In response, a substantial body of research has been devoted to developing more efficient attention mechanisms[10].

Structured state space sequence models (SSMs) have recently emerged as a promising alternative for sequence modeling[11]. These models combine characteristics of recurrent neural networks (RNNs) and convolutional neural networks (CNNs)[12] to achieve efficient computation via recurrence or convolution, offering linear or near-linear scaling with sequence length. Mamba, a simplified end-to-end SSM-based neural network architecture that eschews both traditional attention and even MLP blocks, exemplifies this approach[13]. Mamba provides fast inference and scales linearly with sequence length, which is especially advantageous for processing the often lengthy SMILES strings encountered in chemical datasets.

SSMs are particularly well-suited for modeling molecular data due to several inherent advantages. First, the near-linear scaling of SSMs enables efficient handling of long SMILES strings—crucial when dealing with extensive chemical databases such as PubChem[14], where a high percentage of molecules exhibit long sequences[15]. Second, the combination of recurrent and convolutional elements allows SSMs to capture both local chemical interactions (e.g., bonding patterns) and global structural dependencies (e.g., overall molecular topology) more effectively than fixed-context approaches like Transformers[16]. This holistic capture of structural nuances underpins robust molecular property prediction, reaction-yield estimation, and molecule reconstruction.

In this study, we present a novel Mamba-based large foundation model, denoted as O_{SMI}-SSM-336*M*. Our encoder-decoder foundation model is trained using an efficient SSM-based encoder aligned with an auto-encoder mechanism on a large corpus of 91 million carefully curated

[1]IBM Research, Rio de Janeiro, Brazil. [2]IBM Research, Almaden, CA, USA. ✉e-mail: eduardo.soares@ibm.com

molecules from PubChem[17], resulting in 4 billion molecular tokens. Our main contributions are as follows:

- We pre-train a large-scale Mamba-based foundation model for molecules, denoted as O_{SMI}-SSM-336*M*, on over 91 million molecules carefully curated from PubChem[17], which is equivalent to 4 billion molecular tokens.
- We perform extensive experimentation across 11 benchmark datasets encompassing quantum mechanical, physical, biophysical, and physiological property prediction of small molecules.
- We evaluate the model's capability to predict chemical reaction yields in both synthetic and process chemistry scenarios using the Buchwald–Hartwig cross-coupling reaction dataset, where reaction yields denote the percentage of reactants converted into products.
- We assess the reconstruction capacity of our model on the MOSES benchmarking dataset[18].
- We present a comparative study of inference speeds, demonstrating that our Mamba-based model outperforms a Transformer-based model in predicting HOMO-LUMO properties for 10 million randomly selected samples from PubChem.

Our results indicate that O_{SMI}-SSM-336*M* achieves state-of-the-art performance across various tasks, including molecular property prediction, reaction-yield estimation, and molecule reconstruction. Importantly, the proposed model strikes an effective balance between inference speed and predictive performance, delivering state-of-the-art results in nearly half the time required by Transformer-based architectures.

Results

Experiments

To evaluate the effectiveness of our proposed Mamba-based model O_{SMI}-SSM-336*M*, we conducted experiments using a set of 11 datasets sourced from MoleculeNet[19] as demonstrated in Table 1. Specifically, we evaluated six datasets for classification task and five 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[19]. We also conducted the experiments considered ten 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 ref. 20. The experiments 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 3955 reactions. We employed the same data splits as in ref. 20 to assess our model's performance with training sets of varying sizes.

To evaluate the reconstruction and decoder capabilities of OSMI-SSM-336*M*, we utilized the MOSES benchmarking dataset[18], which contains 1,936,962 molecular structures. For the experiments, we adopted the dataset split proposed by ref. 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-336*M* by predicting HOMO-LUMO properties for 10 million samples randomly selected from PubChem.

Comparison with SOTA on benchmarking tasks

Results for classification tasks. The analysis evaluates the comparative performance of O_{SMI}-SSM-336*M* in its fine-tuned and frozen states relative to state-of-the-art algorithms for molecular property classification, as detailed in Table 2.

Table 2 summarizes the performance of various advanced methods across several benchmarking datasets used for molecular classification tasks. OSMI-SSM-336*M* demonstrates comparative efficacy against Transformer-based approaches, outperforming them in three out of six datasets. Notably, OSMI-SSM-336*M* with its initial configuration yields results on par with current state-of-the-art methods. Further fine-tuning of O_{SMI}-SSM-336*M* enhances its performance, indicating its substantial potential for accurate molecular classification and suggesting that additional performance gains may be achieved through further optimization[21–25].

Results for regression tasks. Subsequently, we applied O_{SMI}-SSM-336*M* to the prediction of chemical properties. The performance metrics across five regression benchmarks—QM9, QM8, ESOL, FreeSolv, and Lipophilicity—are presented in Table 3.

Results presented in Table 3 indicate that OSMI-SSM-336*M* 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 OSMI-SSM-336*M* 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

Table 1 | 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	41,127	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	133,885	12	Average MAE
QM8	12 excited state properties of small molecules	21,786	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

Table 2 | Methods and performance for the classification tasks of MoleculeNet benchmark datasets

Method	Dataset					
	BBBP	ClinTox	HIV	BACE	SIDER	Tox21
GraphMVP[37]	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[38]	72.4 ± 0.4	90.1 ± 1.3	80.6 ± 0.9	85.6 ± 1.1	**67.2 ± 0.4**	78.1 ± 0.1
$GROVER_{Large}$[39]	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[5]	64.3	90.6	62.2	–	–	–
ChemBerta2[40]	71.94	90.7	–	85.1	–	–
Galatica 30B[41]	59.6	82.2	75.9	72.7	61.3	68.5
Galatica 120B[41]	66.1	82.6	74.5	61.7	63.2	68.9
Uni-Mol[42]	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[42]	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[43]	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)[44]	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)[44]	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_{SMI}-SSM-336*M* (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_{SMI}-SSM-336*M*(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**

Bold values indicate the best results for each task.

Table 3 | Methods and performance for the regression tasks of MoleculeNet benchmark datasets

Method	Dataset				
	QM9	QM8	ESOL	FreeSolv	Lipophilicity
D-MPNN[29]	3.241±0.119	0.0143±0.0022	0.98±0.26	2.18±0.91	0.65±0.05
N-Gram[30]	2.51±0.19	0.032±0.003	1.074±0.107	2.688±0.085	0.812±0.028
PretrainGNN[31]	-	-	1.100±0.006	2.764±0.002	0.739±0.003
$GROVER_{Large}$[23]	-	-	0.895±0.017	2.272±0.051	0.823±0.010
ChemBERTa-2[24]	-	-	0.89	-	0.80
SPMM[27]	-	-	0.818±0.008	1.907±0.058	0.692±0.008
$MolCLR_{GIN}$[32]	2.357±0.118	0.0174±0.0013	1.11±0.01	2.20±0.20	0.65±0.08
Hu et al.[33]	4.349±0.061	0.0191±0.0003	1.22±0.02	2.83±0.12	0.74±0.00
MoLFormer[27]	**1.5894±0.0567**	0.0102	0.880±0.028	2.342±0.052	0.700±0.012
SMI-TED289M[28]	**1.3246±0.0157**	**0.0095±0.0001**	**0.6112±0.0096**	**1.2233±0.0029**	**0.5522±0.0194**
O_{SMI}-SSM-336*M*(Frozen)	8.9546 ±0.0577	0.0194 ±0.0003	0.8135 ±0.0253	1.6374 ±0.0682	0.746 ±0.0029
O_{SMI}-SSM-336*M* (Fine-tuned)	2.2175 ±0.3194	**0.0104 ±0.0001**	**0.7222±0.0139**	**1.6288±0.0347**	**0.6048±0.0023**

Blue and Orange indicates best and second-best performing model, respectively.

inference time of SMI-TED289M, a Transformer-based model recognized for its state-of-the-art performance. Figure 1 illustrates the superior inference speed of OSMI-SSM-336*M* compared to SMI-TED289M. Specifically, SMI-TED289M required 20,606.76 s for HOMO property predictions and 21,038.43 s for LUMO property predictions using a single NVIDIA V100 32GB GPU. In contrast, OSMI-SSM-336*M* completed HOMO predictions in 9735.64 s and LUMO predictions in 9823.64 s on the same GPU. These results highlight the substantial efficiency gains of the O_{SMI}-SSM-336*M* 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 h, while also decreasing CO2 emissions by an average of 0.78 kg equivalent[26]. 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[27].

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 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. Figure 2 shows the schema for chemical reaction.

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[20]. This involves estimating the percentage of reactants converted into products. Our evaluation adhered to the schema and data divisions outlined in ref. 20. Table 4 presents the results for

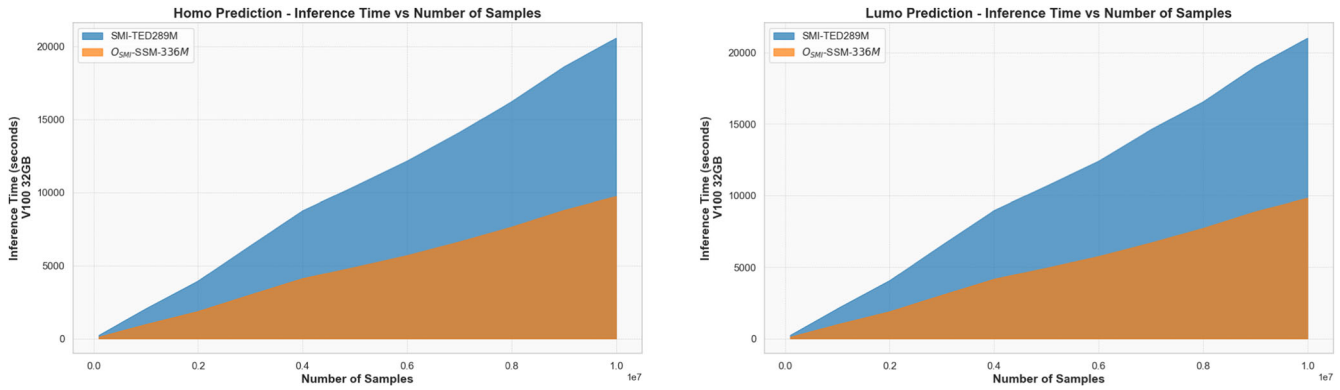

Fig. 1 | The figure shows the inference speed for OSMI-SSM-336*M* and SMI-TED289M for HOMO-LUMO predictions considering a dataset of 10M samples randomly selected from PubChem and a single NVIDIA V100 32GB GPU.

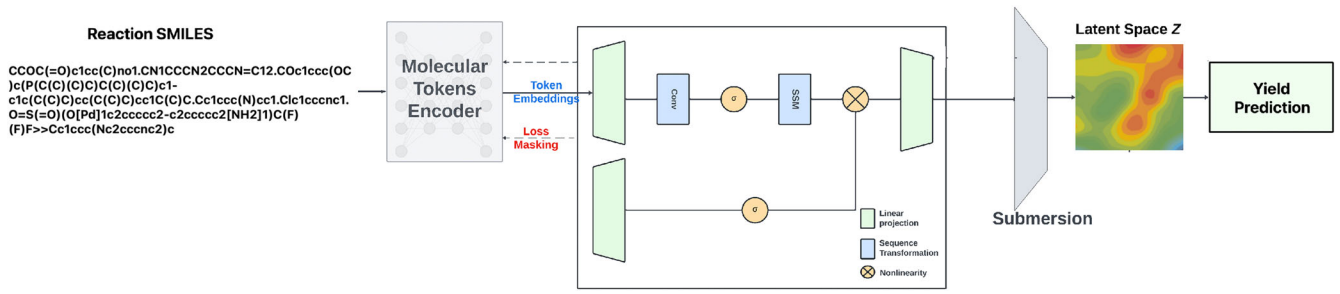

Fig. 2 | This figure illustrates the schema for chemical reaction-yield prediction based on reaction SMILES considering the O_{SMI}-SSM-336*M* model.

Table 4 | Performance of O_{SMI}-SSM-336*M* 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**

Bold values indicate the best results for each task.

the O_{SMI}-SSM-336*M* model and compares its performance with existing state-of-the-art approaches.

The results presented in Table 4 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)[28], a DFT-based random forest model (DFT)[28], and transformer-based models like Yield-BERT[29] and its augmented variant, Yield-BERT(aug.)[29], and MSR2-RXN[30]. 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 ten 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

Table 5 | 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[45]	0.9965	0.8964	0.5477	0.8551	0.3954
LIMO[46]	0.6989	0.0079	0.2464	**0.9039**	26.78
MolGen-7b[47]	0.9999	0.6538	0.5138	0.8617	0.0435
GP-MoLFormer[48]	0.9998	0.7383	0.5045	0.8655	0.0591
O_{SMI}-SSM-336*M*	**0.9999**	**0.9994**	**0.9960**	0.8561	**0.0025**

Bold values indicate the best results for each task.

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.

Decoder evaluation over MOSES benchmarking dataset

Next, conducted a comparative evaluation of the O_{SMI}-SSM-336*M* model against several baseline models for SMILES reconstruction and decoding, using a test set comprising 176,000 molecules. The evaluation metrics, detailed in Table 5, 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).

The results indicate that O_{SMI}-SSM-336*M* 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-336*M* 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-336*M* 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 OSMI-SSM-336*M* 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 OSMI-SSM-336*M* 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.

Discussion

This paper introduces O_{SMI}-SSM-336*M*, 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-336*M* 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 h and lowering CO2 emissions by an average of 0.78 kg CO2 equivalent[26] 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-336*M* 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.

Table 6 | 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	18,838.70
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

Table 7 | O_{SMI}-SSM-336*M* 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

Methods

This section presents an overview of the proposed O_{SMI}-SSM-336*M* foundation model for small molecules. Here, we outline the process of collecting, curating, and pre-processing the pre-train data. Additionally, we describe the token encoder process and the SMILES encoder-decoder process.

Pre-training data

The pre-training 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 using standard procedures implemented in RDKit[31,32]; each SMILES string was standardized, converted to its canonical form, and sanitized to ensure both uniqueness and chemical validity by checking for conformant valence and bonding rules. Following this, a molecular transformation process was applied—incorporating additional chemical sanitization checks—to validate the molecules derived from the unique SMILES strings, resulting in a final curated set of 91 million unique and chemically sound molecules.

To construct the vocabulary, we utilized the molecular tokenizer proposed by ref. 33. 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 2988 unique tokens, along with 5 special tokens. In contrast, MoLFormer, which was trained on 1 billion samples with minimal curation, generated a vocabulary of 2362 tokens using the same tokenization method[2]. This indicates that our comprehensive curation process led to an enhanced and more representative vocabulary model. Detailed statistics of the pre-training dataset are provided in Table 6.

Model architecture

We conduct training for O_{SMI}-SSM-336*M* using a Mamba-based token encoder together with an encoder-decoder architecture to effectively map SMILES sequences into a latent embedding space and back. O_{SMI}-SSM-336*M* design leverages the strengths of SSMs to capture long-range dependencies and process sequences with near-linear scaling. In our approach, the encoder processes input tokens via Mamba blocks, while the decoder reconstructs these embeddings to accurately generate SMILES sequences.

The hyperparameters for the model are specified in Table 7. The hidden size is set to 768 with 24 layers in total. It is important to highlight that the layer count in our Mamba-based model is effectively doubled compared to a Transformer with a similar size. In traditional Transformers, each layer comprises a multi-head attention (MHA) block followed by a multilayer perceptron (MLP) block. In contrast, our architecture implements two distinct Mamba blocks for each such Transformer layer—one handling the functions analogous to the MHA and the other corresponding to the MLP—resulting in a doubled layer count.

Additional parameters—such as dt rank (set automatically), d state of 16, d conv of 4, and an expansion factor of 2—are tuned to balance model expressiveness and computational efficiency. The parameters dt min (0.001), dt max (0.1), dt scale (1.0), and dt init floor (1e−4) govern the dynamics of the continuous-time system underlying our model, ensuring stable training and effective representation learning.

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 Eq. (1):

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

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[34]. 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):

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

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):

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

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[35]. This architecture serves as a tool for dimensionality reduction and representation learning in the domain of molecular structures.

Pre-training strategies

Pre-training of O_{SMI}-SSM-336*M* was performed for 130 epochs on the entire curated PubChem dataset using a fixed learning rate of 3×10^{-5} and a batch size of 128 molecules on 24 NVIDIA V100 (16G) GPUs, distributed across 4 nodes via DDP and *torch run*. The pre-training process is divided into two distinct phases:

- Phase 1: in this initial phase, the token encoder is pre-trained using 95% of the available samples, while the remaining 5% is reserved exclusively for training the encoder-decoder layer. This partitioning is necessary to mitigate convergence difficulties in the early epochs of token embedding learning, which could otherwise adversely affect the training of the encoder-decoder component.
- Phase 2: once the token embeddings have converged, the pre-training is expanded to utilize 100% of the available samples for both phases. This approach enhances the performance of the encoder-decoder layer, particularly in terms of token reconstruction accuracy.

For encoder pre-training, we employ the masked language modeling strategy as defined in ref. 36. Initially, 15% of the tokens are selected for possible prediction; of these, 80% are replaced with the [MASK] token, 10% are substituted with a random token, and the remaining 10% remain unchanged. This strategy facilitates the learning of robust, contextualized representations of the SMILES tokens.

Our approach also incorporates a latent space embedding mechanism that supersedes conventional mean pooling. Instead of aggregating token embeddings via an averaging operation (which lacks a natural inversion process), the token embeddings are transformed into a latent vector z (as detailed in Eqs. (2) and (3) of the "Model architecture" section). This latent representation captures intricate structural nuances of the SMILES strings and supports a reversible mapping, thereby enabling both accurate SMILES reconstruction and effective downstream tasks such as molecular property prediction.

The pre-training procedure is guided by two distinct loss functions. The first loss function, based on the masked language model objective, uses cross-entropy loss to predict the masked tokens. The second loss function governs the reconstruction task and is measured using the mean squared error (MSE) between the original tokens and their reconstructions generated by the encoder-decoder layer. Monitoring these metrics ensures the convergence of the token embeddings and the stability of the latent space representation throughout training.

Data availability

No datasets were generated or analyzed during the current study.

Code availability

All Python codes for training and fine-tuning O_{SMI}-SSM-336*M*, together with Python notebooks for experimental evaluations, are available at https://github.com/IBM/materials/tree/main/models/smi_ssed. Pre-trained model weights can be accessed via our HuggingFace repository at https://huggingface.co/ibm-research/materials.smi_ssed. For other enquiries contact the corresponding authors.

Received: 8 November 2024; Accepted: 3 May 2025;
Published online: 09 June 2025

References

1. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
2. Ross, J. et al. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* **4**, 1256–1264 (2022).
3. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2108.07258 (2021).
4. Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nat. Commun.* **11**, 4874 (2020).
5. Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2010.09885 (2020).
6. Janakarajan, N., Erdmann, T., Swaminathan, S., Laino, T. & Born, J. Language models in molecular discovery. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2309.16235 (2023).
7. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* 30 (2017).
8. Tay, Y., Dehghani, M., Bahri, D. & Metzler, D. Efficient transformers: a survey. *ACM Comput. Surv.* **55**, 1–28 (2022).
9. Lin, T., Wang, Y., Liu, X. & Qiu, X. A survey of transformers. *AI Open* **3**, 111–132 (2022).
10. Kotei, E. & Thirunavukarasu, R. A systematic review of transformer-based pre-trained language models through self-supervised learning. *Information* **14**, 187 (2023).
11. Gu, A., Goel, K. & Ré, C. Efficiently modeling long sequences with structured state spaces. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2111.00396 (2021).
12. Smith, J. T., Warrington, A. & Linderman, S. W. Simplified state space layers for sequence modeling. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2208.04933 (2022).
13. Gu, A. & Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2312.00752 (2023).
14. Patro, B. N. & Agneeswaran, V. S. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2404.16112 (2024).
15. Hähnke, V. D., Kim, S. & Bolton, E. E. Pubchem chemical structure standardization. *J. Cheminformatics* **10**, 1–40 (2018).
16. Waleffe, R. et al. An empirical study of mamba-based language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2406.07887 (2024).
17. Kim, S. et al. Pubchem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).
18. Polykovskiy, D. et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Front. Pharmacol.* **11**, 565644 (2020).
19. Wu, Z. et al. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
20. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in c–n cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
21. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
22. Liu, S., Demirel, M. F. & Liang, Y. N-gram graph: simple unsupervised representation for graphs, with applications to molecules. In *Advances in Neural Information Processing Systems* 32 (2019).
23. Hu, W. et al. Strategies for pre-training graph neural networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1905.12265 (2019).

24. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).
25. Hu, Z., Dong, Y., Wang, K., Chang, K.-W. & Sun, Y. Gpt-gnn: generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1857–1867 (2020).
26. Lacoste, A., Luccioni, A., Schmidt, V. & Dandres, T. Quantifying the carbon emissions of machine learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1910.09700 (2019).
27. Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A. & Sauerland, U. Risks and benefits of large language models for the environment. *Environ. Sci. Technol.* **57**, 3464–3466 (2023).
28. Probst, D., Schwaller, P. & Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital Discov.* **1**, 91–97 (2022).
29. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn. Sci. Technol.* **2**, 015016 (2021).
30. Boulougouri, M., Vandergheynst, P. & Probst, D. Molecular set representation learning. *Nat. Mach. Intell.* **6**, 754–763 (2024).
31. Landrum, G. Rdkit documentation. *Release* **1**, 4 (2013).
32. Heid, E., Liu, J., Aude, A. & Green, W. H. Influence of template size, canonicalization, and exclusivity for retrosynthesis and reaction prediction applications. *J. Chem. Inf. Model.* **62**, 16–26 (2021).
33. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
34. Bran, A. M. & Schwaller, P. Transformers and large language models for chemistry and drug discovery. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2310.06083 (2023).
35. Ferrando, J., Gállego, G. I., Tsiamas, I. & Costa-jussà, M. R. Explaining how transformers use context to build predictions. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.12535 (2023).
36. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics* (2019).
37. Liu, S. et al. Pre-training molecular graph representation with 3d geometry. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2110.07728 (2021).
38. Fang, X. et al. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* **4**, 127–134 (2022).
39. Rong, Y. et al. Self-supervised graph transformer on large-scale molecular data. *Adv. Neural Inf. Process. Syst.* **33**, 12559–12571 (2020).
40. Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta-2: towards chemical foundation models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2209.01712 (2022).
41. Taylor, R. et al. Galactica: A large language model for science. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2211.09085 (2022).
42. Zhou, G. et al. Uni-mol: a universal 3d molecular representation learning framework. *ChemRxiv* (2023).
43. Chang, J. & Ye, J. C. Bidirectional generation of structure and properties through a single molecular foundation model. *Nat. Commun.* **15**, 2323 (2024).
44. Soares, E. et al. A large encoder-decoder family of foundation models for chemical language. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2407.20267 (2024).
45. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, 2323–2332 (PMLR, 2018).
46. Eckmann, P. et al. Limo: latent inceptionism for targeted molecule generation. *Proc. Mach. Learn. Res.* **162**, 5777 (2022).
47. Fang, Y. et al. Domain-agnostic molecular generation with self-feedback. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2301.11259 (2023).
48. Ross, J. et al. Gp-molformer: a foundation model for molecular generation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2405.04912 (2024).

Author contributions

E.S., E.V.B., and D.Z. conceived the computational experiments. E.S. and V.S. carried out the experiments, while E.S., E.V.B., and D.Z. analyzed the results. E.V.B., R.C., and K.S. designed and supervised the project. All authors contributed to and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Eduardo Soares.