

# ChemFM as a Scaling Law Guided Foundation Model Pre-trained on Informative Chemicals

Feiyang Cai<sup>1</sup>, Katelin Zacour<sup>2</sup>, Tianyu Zhu<sup>3</sup>, Tzuen-Rong Tzeng<sup>2</sup>,  
Yongping Duan<sup>4</sup>, Ling Liu<sup>5</sup>, Srikanth Pilla<sup>6</sup>, Gang Li<sup>7</sup>, Feng Luo<sup>1\*</sup>

<sup>1</sup>School of Computing, Clemson University, Clemson, 29634, SC, USA.

<sup>2</sup>Department of Biological Sciences, Clemson University, Clemson, 29634, SC, USA.

<sup>3</sup>Department of Materials Science and Engineering, Clemson University, Clemson, 29634, SC, USA.

<sup>4</sup>Horticultural Research Laboratory, USDA, Fort Pierce, 34945, FL, USA.

<sup>5</sup>College of Computing, Georgia Institute of Technology, Atlanta, 30332, GA, USA.

<sup>6</sup>Center for Composite Materials, University of Delaware, Newark, 19716, DE, USA.

<sup>7</sup>Department of Mechanical Engineering, Clemson University, Clemson, 29634, SC, USA.

\*Corresponding author(s). E-mail(s): [luofeng@clemson.edu](mailto:luofeng@clemson.edu);

## Abstract

Traditional AI methods often rely on task-specific model designs and training, which constrain both the scalability of model size and generalization across different tasks. Here, we introduce ChemFM, a large foundation model specifically developed for chemicals. By conducting a series of scaling experiments, we identify UniChem as the informative molecular database for pre-training the foundation model. ChemFM comprises 3 billion parameters and is pre-trained on 178 million molecules using self-supervised causal language modeling to extract generalizable molecular representations. This model can be adapted to diverse downstream chemical applications using either full-parameter or parameter-efficient fine-tuning methods. ChemFM consistently outperforms state-of-the-art task-specific AI models across all tested tasks. Notably, it achieves up to 67.48% performance improvement across 34 property prediction benchmarks, up to 33.80% reduction in mean average deviation between conditioned and actual properties of generated molecules in conditional molecular generation tasks, and up to 3.7% top-1 accuracy improvement across 4 reaction prediction datasets. Moreover, ChemFM demonstrates its superior performance in predicting antibiotic activity and cytotoxicity, highlighting its potential to advance the discovery of novel antibiotics. Furthermore, we demonstrate that, as a foundation model, ChemFM exhibits strong data efficiency, requiring significantly fewer labeled training samples to achieve state-of-the-art performance. We anticipate that ChemFM will significantly advance chemistry research by providing a foundation model capable of effectively generalizing across a broad range of tasks with minimal additional training.

## 1 Introduction

Over the past decade, artificial intelligence has revolutionized research methodologies across scientific disciplines<sup>1,2</sup>, including chemistry. The prevailing AI-based paradigm in computational chemistry focuses on developing models for specific tasks. For example, deep learning models are often trained using pre-calculated molecular descriptors or fingerprints<sup>3</sup>, molecular graph representations<sup>4</sup>, or serialization format representations<sup>5</sup>. These models excel at tasks such as predicting molecular properties<sup>6,7</sup>, designing and optimizing molecules<sup>8,9</sup>, and forecasting chemical synthesis and retro-synthesis<sup>10,11</sup>. Despite

their advancements, these task-specific models have limitations. First, training a high-performing task-specific model often requires large amounts of high-quality data. Annotating chemical data is typically costly and time-consuming, and may involve extensive laboratory experiments. Second, the models struggle to capture general patterns that reflect the inherent structural dependencies and contextual relationships within molecules, leading to over-fitting and poor generalization to novel molecular features. Furthermore, the diversity of chemical tasks and datasets makes it impractical to annotate comprehensive chemical datasets and train large-scale models for every individual application. However, as suggested by the success in other domains such as computer vision<sup>12</sup> and natural language processing<sup>13,14</sup>, scaling model sizes could unlock new possibilities in chemical AI models.

One promising approach to addressing these challenges is the development of foundation models. These models are pre-trained on large unannotated datasets, often using weakly supervised or unsupervised methods to extract complex, general-domain features, enabling them to be fine-tuned for various downstream tasks with minimal additional training. Initially pioneered in language and image modalities, foundation models have demonstrated substantial performance improvements in other scientific domains, such as retinal imaging<sup>15</sup>, single-cell transcriptomics<sup>16</sup>, and histopathology imaging<sup>17</sup>. Existing efforts to construct large chemical models can be broadly categorized into two directions. The first focuses on pre-training models exclusively on chemical data. Early works, such as developing unified pre-trained models<sup>10,18–20</sup>, have been limited by the scale of both model architectures and pre-training datasets (Table 1). Most of these studies directly used public molecular databases<sup>21,22</sup>, without additional preprocessing or quality assessment. As a result, the datasets often contained redundant or noisy molecules and lacked systematic evaluation of chemical diversity, hindering meaningful scaling analysis and limiting generalization across tasks. Moreover, no prior work has systematically analyzed how large a chemical foundation model should be or how its performance scales with model size, except for a preliminary study by Frey et al.<sup>23</sup> that explored scaling laws during pre-training. However, it did not demonstrate whether scaling leads to improved performance on downstream tasks. The second reframes chemical tasks as natural language problems, fine-tuning large language models to augment chemical knowledge<sup>24,25</sup>. Although this approach benefits from the capabilities of large pre-trained language models, it lacks fundamental alignment between linguistic tokens and molecular representations. As a result, such models struggle even with basic molecule recognition and manipulation tasks<sup>26</sup>, and therefore cannot be expected to outperform chemical-specific models, as has been empirically observed.

In this work, we introduced ChemFM, a 3-billion-parameter foundation model designed for chemicals that can be fine-tuned for various chemical design and property prediction tasks. Our study primarily focuses on two key aspects often overlooked in prior works: (1) analyzing the scaling behavior of chemical foundation models across different model and dataset sizes, and (2) preparing and evaluating large-scale chemical datasets to understand how data diversity and quality affect model performance. By conducting a series of scaling experiments, we demonstrated that UniChem<sup>27</sup>, while being relatively smaller with 178 million molecules, is more diverse and information richer than the much larger ZINC20 database<sup>21</sup> of 1.8 billion molecules. Therefore, we trained ChemFM on SMILES strings<sup>5</sup> from 178 million molecules in UniChem database<sup>27</sup> (Fig. 1a and 1b). By leveraging the paradigm of casual language modeling<sup>28</sup>, ChemFM effectively learned SMILES syntax as well as the molecular internal relationships between atoms and bonds, enabling its adaptation for various downstream tasks (Fig. 2). We first validated ChemFM on 34 property prediction datasets from domains including pharmaceutical, physicochemical, and bioactivity, showing consistent outperformance over existing approaches across all datasets. Moreover, ChemFM demonstrated superior performance for potential antibiotic screening, highlighting its potential to advance real-world drug discovery. ChemFM also exhibited flexibility and versatility in conditional molecular generation tasks. Unlike previous approaches that required training separate models for each condition or condition combination, ChemFM allowed the training of a single unified model capable of handling all variations of condition combinations. The unified model not only achieved strong generative performance but also enabled effective control and matching of flexible desired conditions. Furthermore, we demonstrated that ChemFM can be seamlessly integrated with existing sequence editing-based methods for reaction prediction<sup>11</sup>, resulting in state-of-the-art performance on 4 reaction prediction tasks, including both forward synthesis and retro-synthesis. Furthermore, ChemFM also exhibited remarkable training and data efficiency. By leveraging parameter-efficient fine-tuning<sup>29</sup>, ChemFM can be fine-tuned within a single moderate GPU machine, making it broadly accessible for research applications. ChemFM achieved state-of-the-art results while using significantly less data than

existing task-specific methods. ChemFM can be leveraged for diverse chemical research endeavors and may significantly advance chemistry research.

## 2 Results

### 2.1 Pre-training dataset selection

Traditionally, ZINC20<sup>21</sup> and UniChem<sup>27</sup>, containing 1.8 billion and 178 million molecules, respectively, are suitable candidates for pre-training ChemFM due to their extensive molecular coverage. We first conducted a series of scaling experiments to assess the informativeness of both datasets (Methods and Supplementary Fig. S1.1). The scaling laws of neural models reveals that model performance follows a power-law relationship between model size or dataset size, provided the other is not bottlenecked<sup>30</sup>. Our scaling experiments evaluated models ranging from approximately 10 million to 200 million parameters. The results showed that, for UniChem, the model performance (measured by validation loss) closely followed a power-law scaling trend with model size, with no signs of performance saturation. In contrast, the ZINC20 dataset exhibited performance saturation when the model size reached 60 million parameters. This suggests that the knowledge contained within ZINC20 becomes a bottleneck, limiting further gains from increasing model size. This may be due to the fact that ZINC20 is primarily designed for ligand discovery, enriching commercially available compounds with mainstream medicinal chemistry scaffolds. Many molecules share core structures with minor variations, limiting structural diversity and reducing the dataset’s informativeness for large-scale molecular representation learning. Based on these findings, we selected UniChem as our pre-training dataset, as it offers a broader chemical space and greater structural diversity. Additionally, we experimentally demonstrated that the model pre-trained on the UniChem dataset outperforms the one pre-trained on ZINC20 in downstream tasks (Methods; Supplementary Table S2.6.)

### 2.2 Training of ChemFM

By leveraging self-supervised causal language modeling, we developed two model variants: ChemFM-1B and ChemFM-3B, comprising approximately 970 million and 3.0 billion trainable parameters, respectively. Both models were trained for one epoch on 1.78 billion SMILES strings, augmented from 178 million molecules from UniChem dataset (Methods). Throughout pre-training, the validation perplexity for both ChemFM-1B and ChemFM-3B steadily decreased (Fig. 1c), showing no signs of saturation until processing 818 billion tokens. ChemFM-3B achieved a lower final validation perplexity compared to ChemFM-1B. Moreover, the final validation losses of both models start to deviate from the predicted scaling law (Supplementary Fig. S1.1), suggesting that under the current data regime, further model scaling may lead to diminishing returns as the loss approaches a plateau.

### 2.3 Unconditional molecule generation using pre-trained ChemFM

We evaluated ChemFM-3B in unconditional molecule generation by randomly generating 100,000 molecules and benchmarking validity, uniqueness, novelty, internal diversity, and distribution similarity with the training dataset (Fig. 1d). ChemFM-3B achieved a remarkable validity score of 0.996 without additional constraints during tokenization, model training, or generation. The uniqueness score was perfect (1.0), indicating no duplicate canonical SMILES strings among generated molecules. High internal diversity scores (IntDiv<sub>1</sub> of 0.904 and IntDiv<sub>2</sub> of 0.896) demonstrated the diversity of molecular structures of generations. By comparing various physicochemical descriptors—such as molecular complexity, weight, and structural characteristics—and ECFP fingerprints between the training and generated molecules, we observed that ChemFM faithfully captured the distribution of molecules in the training data without overfitting to a narrow subset (Methods and Supplementary Fig. S1.2 and S1.3). More importantly, over half (55.8%) of the generated molecules were entirely novel, not found in the extensive training dataset, highlighting the potential of these models for exploring chemical space, discovering new molecules, and optimizing molecular structures.

## 2.4 Molecular property prediction

We evaluated the adaptability of ChemFM for molecular property prediction using two widely-used benchmarks: MoleculeNet<sup>31</sup> and ADMET<sup>32</sup>, covering a total of 34 datasets across diverse domains, including pharmaceutical, physicochemical, and bioactivity applications. Across all evaluated datasets from both MoleculeNet and ADMET benchmarks, ChemFM models consistently outperformed existing state-of-the-art methods.

The MoleculeNet benchmark consists of 4 regression datasets (4 properties in total) and 8 classification datasets (189 properties in total) (Supplementary Table S2.8). Comparisons with the methods in the literature for MoleculeNet datasets are often challenging due to varying dataset splitting strategies and random seed choices. To ensure comprehensive evaluation, we compared ChemFM models with different sets of methods using the same splitting methods and random seeds. We first compared the fine-tuned ChemFM-3B models on the standard MoleculeNet datasets against SMILES Transformer<sup>19</sup>, MoleculeNet models<sup>33</sup>, directed message passing neural networks (D-MPNN or Chemprop)<sup>7</sup>, MolMapNet OOTB (MMNB)<sup>6</sup>, and Chemformer<sup>10</sup> (Fig. 3; full comparison results are provided in Supplementary Table S2.1). For classification tasks, ChemFM-3B demonstrated a consistent performance advantage, with improvements in the area under the receiver operating curve (ROC-AUC) of 0.012 on BBBP, 0.034 on BACE, 0.030 on HIV, 0.018 on Tox21, 0.029 on SIDER, and 0.030 on ClinTox. Additionally, ChemFM-3B showed improvements of 0.026 and 0.010 on the MUV and PCBA datasets, respectively, in the area under the precision-recall curve (PRC-AUC). For regression tasks, ChemFM-3B reduced root mean squared errors (RMSE) by 0.039 on ESOL, 0.245 on FreeSolv, 0.010 on Lipophilicity, and 0.024 on PDBbind.

We also compared ChemFM against methods that use different dataset splits, including Pre-train GNNs<sup>34</sup>, ChemBERTa-2<sup>20</sup>, AttentiveFP<sup>35</sup>, 3D InfoMax<sup>36</sup>, Mole-BERT<sup>18</sup>, GraphMVP<sup>37</sup>, and MoleculeSDE<sup>38</sup> (Methods and Supplementary Table S2.2 and S2.3). Across all comparison settings, ChemFM consistently delivered better results than the other methods did. Additionally, we observed that ChemFM-3B generally outperformed ChemFM-1B (Fig. 3) and its non-pre-trained counterpart (Methods and Supplementary Table S2.5), underscoring the benefits of larger model sizes and pre-training.

On the ADMET benchmark, which includes 13 classification and 9 regression datasets (each representing a single property; Supplementary Table S2.10), ChemFM again achieved superior performances across all datasets (quantitative results in Supplementary Table S2.4), with an average improvement of approximately 7.09%. The improvements ranged from a minimum of 0.11% on the DILI dataset to a maximum of 67.48% on the Half\_Life\_Obach dataset.

For a wider application, we also evaluated ChemFM on two additional datasets beyond the MoleculeNet and ADMET benchmarks: an odor prediction dataset<sup>39</sup> and a chromatographic retention time prediction dataset<sup>40</sup>. Across both tasks, ChemFM-3B outperformed the specialized baselines (Methods and Supplementary Table S2.7), with particularly large improvements (31.6% reduction in mean absolute error) on retention time prediction, demonstrating that our approach generalizes effectively to broader chemical property prediction tasks.

## 2.5 Potential discovery of novel antibiotics

Recent study<sup>41</sup> has leveraged multiple Chemprop<sup>7</sup> models to predict antibiotic activity and human cell cytotoxicity, successfully screening over 10 million molecules to identify novel antibiotic candidates with high antibiotic activity and low cytotoxicity. Here, we fine-tuned ChemFM model for the same tasks of predicting antibiotic activity and cytotoxicity across different cell types (Supplementary Fig. S2.1). Specifically, ChemFM has significantly improved the performance with the PRC-AUC values increasing from 0.364 to 0.428 for antibiotic activity, 0.176 to 0.461 for cytotoxicity in human liver carcinoma cells (HepG2), 0.168 to 0.459 for human primary skeletal muscle cells (HskMC), and 0.335 to 0.414 for cytotoxicity in human lung fibroblasts cells (IMR-90).

To further evaluate ChemFM’s predictive capabilities, we applied both ChemFM and Chemprop to an antibiotic library of 1,173 molecules<sup>42</sup>, which consists of real antibiotics but differs significantly from the positive samples in the training dataset. ChemFM labeled 149 molecules as positives, whereas Chemprop labeled only 29 (Supplementary Data 1), suggesting that ChemFM has a higher true positive rate in discovering antibiotics.

These results highlight ChemFM’s potential to significantly improve the screening process for novel antibiotics by providing more accurate predictions for both antibiotic activity and toxicity.

## 2.6 Conditional molecule generation

Conditional molecule generation is critical for designing molecules to meet specific property criteria or incorporate particular scaffold structures. We fine-tuned two separate ChemFM-3B models: one on the GuacaMol<sup>43</sup> dataset for property-based generation and another on the MOSES<sup>44</sup> dataset for scaffold and property-based generation. For each dataset, we considered four continuous properties: octanol-water partition coefficient (logP), synthetic accessibility score (SAS), topological polar surface area (TPSA), and quantitative estimate of drug-likeness (QED). Traditional methods such as cRNN<sup>45</sup> and MolGPT<sup>8</sup> require a separate model for each property combination, resulting in 15 models to cover all four conditions for each dataset. In contrast, by carefully designing the input condition to the model (Methods) ChemFM can handle all combinations within a single unified model.

We first evaluated the property-based generation model trained on the GuacaMol dataset. For each property combination, we generated 10,000 molecules at different sample points and evaluated their validity, uniqueness, novelty, and mean absolute deviation (MAD) between the conditioned and computed properties (Table 2 and Supplementary Fig. S3.1). ChemFM outperformed MolGPT in validity, uniqueness, and novelty across all conditioned properties, whether for individual properties or multiple combined conditions (Table 2; comparison with cRNN<sup>45</sup> is given in Supplementary Table S3.1).

On average, ChemFM achieved improvements of 0.0079 in validity, 0.0104 in uniqueness, and 0.0151 in novelty over MolGPT. Furthermore, ChemFM demonstrated stronger adherence to the desired property values, with an average percentage reduction in MAD across all four properties of 21.19%, ranging from 7.70% for SAS to 33.80% for TPSA.

Next, we evaluated conditional generation based on both scaffold and property on the MOSES dataset. Using the same 5 test scaffolds as MolGPT, we generated 10,000 molecules at each sample point across different scaffold and property combinations (Supplementary Table S3.2 and Supplementary Fig. S3.3). ChemFM consistently outperformed MolGPT across all scaffold and property combinations, generating more valid, unique, and novel molecules, with average improvements of 1.93%, 26.69%, and 26.69%, respectively. Moreover, ChemFM showed a stronger alignment with the desired conditions by: 1) generating more molecules that shared the same scaffold as the conditioned scaffold, with an average improvement of 25.73% over MolGPT, and 2) achieving an average reduction in MAD across all four properties, with reductions of 15.31%, 9.63%, 13.35%, and 1.96% for logP, SAS, TPSA, and QED, respectively.

## 2.7 Reaction prediction

We fine-tuned ChemFM-3B model for both reaction synthesis and retro-synthesis tasks using three USPTO benchmark datasets: USPTO-Full<sup>46</sup>, USPTO-MIT<sup>47</sup>, and USPTO-50K<sup>48</sup>. These datasets, comprising organic chemical reactions extracted from US patents and applications, are widely used for evaluating reaction prediction tasks (Supplementary Table S4.2). We compared ChemFM with existing methods in the literature, employing the same data splitting methods for training and evaluation<sup>11,49</sup>. Table 3 presents a comparison between ChemFM and previous best and second-best performing models, while complete results comparing ChemFM with other methods are available in Supplementary Table S4.1.

For the retro-synthesis task, ChemFM consistently achieved higher top-1, top-3, and top-5 accuracies compared to previous best methods. Our experiments also highlight the training efficiency of the ChemFM foundation model. For instance, on the USPTO-50K dataset, we achieved state-of-the-art results after just one epoch (25,000 steps) of training on the augmented dataset (equivalent to five epochs due to five-fold augmentation), already surpassing the performance of R-SMILES<sup>11</sup>, which used approximately ten times the number of training steps. With additional training, the top-1 accuracy could be further improved (while top-5 accuracy may decrease). Moreover, for USPTO-50K and USPTO-MIT, top-1 accuracies further reach 59.7% and 62.4%, respectively. The top-1 accuracy improvements over the previous best methods were 3.7% for USPTO-50K, 2.1% for USPTO-MIT, and 2.3% for USPTO-Full.

For the reaction synthesis task, we focused on the more challenging setting where reactants and reagents are mixed, evaluating ChemFM on the USPTO-MIT dataset. ChemFM demonstrated competitive performance, surpassing the previous best method (AT<sup>49</sup>) by 0.1% on both top-1 and top-5 accuracies.

## 2.8 Training and data efficiency

ChemFM supported parameter-efficient fine-tuning methods, such as the low-rank adaptation (LoRA) technique<sup>29</sup>, which significantly reduces the number of trainable parameters and GPU memory requirements (Methods). For example, with a LoRA rank of 4 in ChemFM-3B (using 32-bit float precision), the number of trainable parameters is reduced by 460 $\times$ , from 3 billion to 6.5 million. This reduction lowers the GPU memory required during training from 51 GB to 20 GB, making fine-tuning feasible on a single moderate GPU machine. Additionally, the checkpoint size is reduced from 12 GB to 26 MB, allowing for minimal storage for adapters on each dataset.

Furthermore, ChemFM also demonstrates strong data efficiency. We evaluated its performance on a classification task (CYP2D6\_Substrate\_CarbonMangels) and a regression task (Half\_Life\_Obach). In the classification task, ChemFM outperformed previous state-of-the-art methods even when fine-tuned on just 50% of the training data. Similarly, in the regression task, fine-tuning with only 10% of the data was sufficient to achieve state-of-the-art results on the test set (Methods and Supplementary Fig. S2.2). This suggests that ChemFM effectively learns transferable molecular representations, reducing the need for large labeled datasets and making it well-suited for small-data scenarios. This is an important advantage in chemistry research, where collecting and measuring data can be costly and labor-intensive.

## 3 Discussion

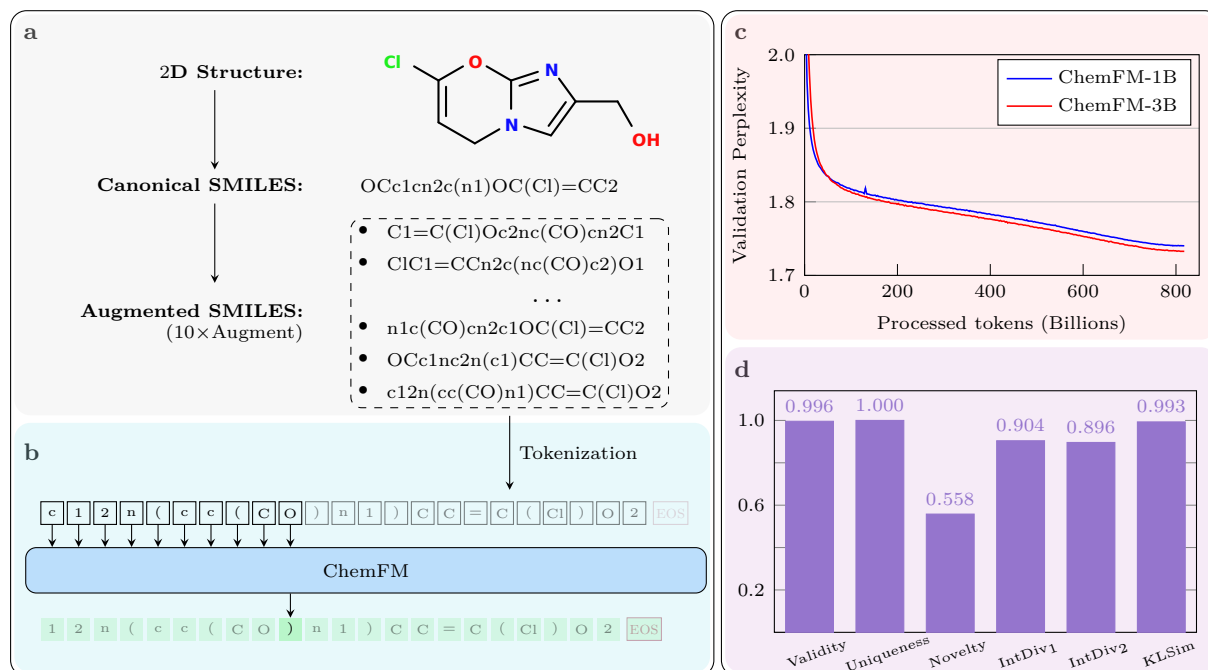
The tasks in computational chemistry are complex and diverse, and training specific models for each task is both resource-intensive and time-consuming. In this work, we introduced ChemFM, a general-purpose foundation model specifically designed for chemicals. By leveraging the causal language modeling framework and extensive self-supervised training on 178 million molecules, ChemFM has successfully learned the molecular structures represented by SMILES, as well as the contextual relationships of atoms and bonds within molecules.

ChemFM effectively characterized the structures of molecules, helping to establish structure-property relationships. Evaluated against 34 molecular property prediction datasets from the MoleculeNet and ADMET benchmarks, ChemFM achieved an average 6.98% improvement over previous state-of-the-art methods. Moreover, in an antibiotic discovery application, ChemFM substantially outperformed models used in a prior study<sup>41</sup>. In the conditional molecule generation, we showed that a single unified ChemFM model can generate molecules given an arbitrary combination of property conditions with high validity, uniqueness, and novelty while precisely matching desired properties or scaffold structures. We further demonstrated ChemFM’s ability to improve both accuracy and computational efficiency in predicting chemical reactions. ChemFM integrated seamlessly with SMILES sequence editing-based methods designed for reaction prediction, such as the root-aligned SMILES (R-SMILES) technique. Requiring fewer training steps, ChemFM consistently achieved higher prediction accuracy than existing models on synthesis and retro-synthesis tasks across USPTO benchmark datasets. Beyond the tasks studied in this work, we believe ChemFM can be effectively extended to broader downstream tasks, including molecular optimization<sup>50,51</sup>, which represent promising directions for future exploration.

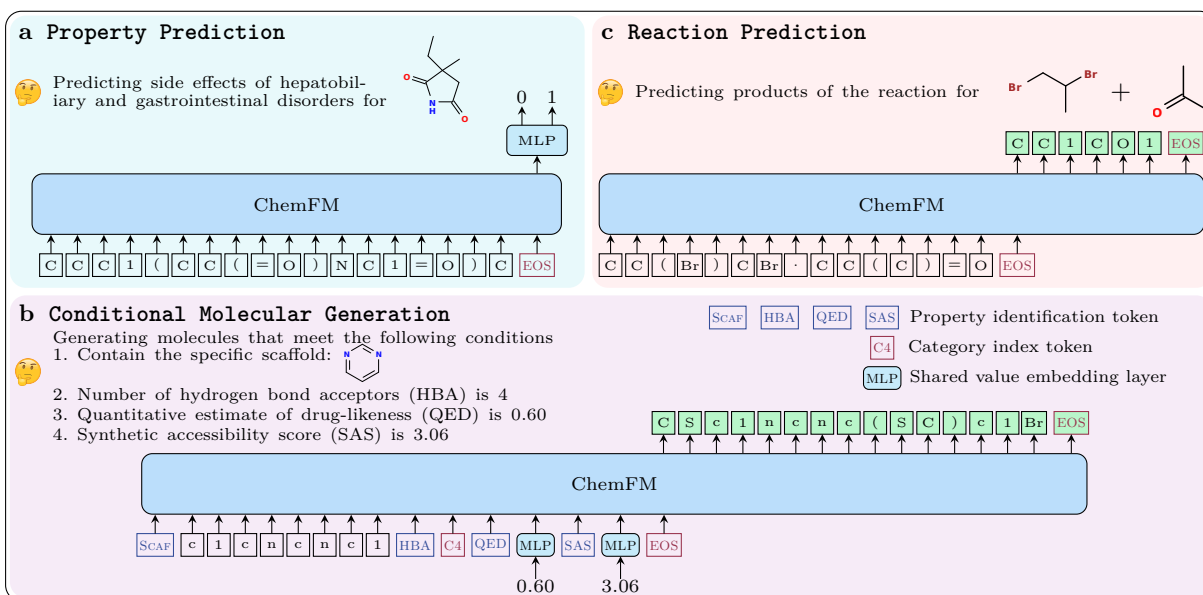
This work has a few limitations. ChemFM was fully trained on the most informative dataset available to our knowledge, which makes the distribution of generated molecules closely mirror the training dataset, limiting exploration of the potentially broader chemical space. While fine-tuning ChemFM is efficient with the LoRA technique, its inference time is not yet comparable to smaller-scale models, particularly when screening large amounts of data. Distilling smaller, cost-efficient models from ChemFM could improve evaluation efficiency.

In conclusion, ChemFM demonstrates its capability as a versatile chemical foundation model, which can efficiently be adapted to diverse tasks and improve upon state-of-the-art performance. The success of ChemFM in unifying various chemical tasks under a single model architecture highlights the capability of foundation models in computational chemistry, potentially significantly advancing drug discovery, molecule optimization, and chemical synthesis planning.

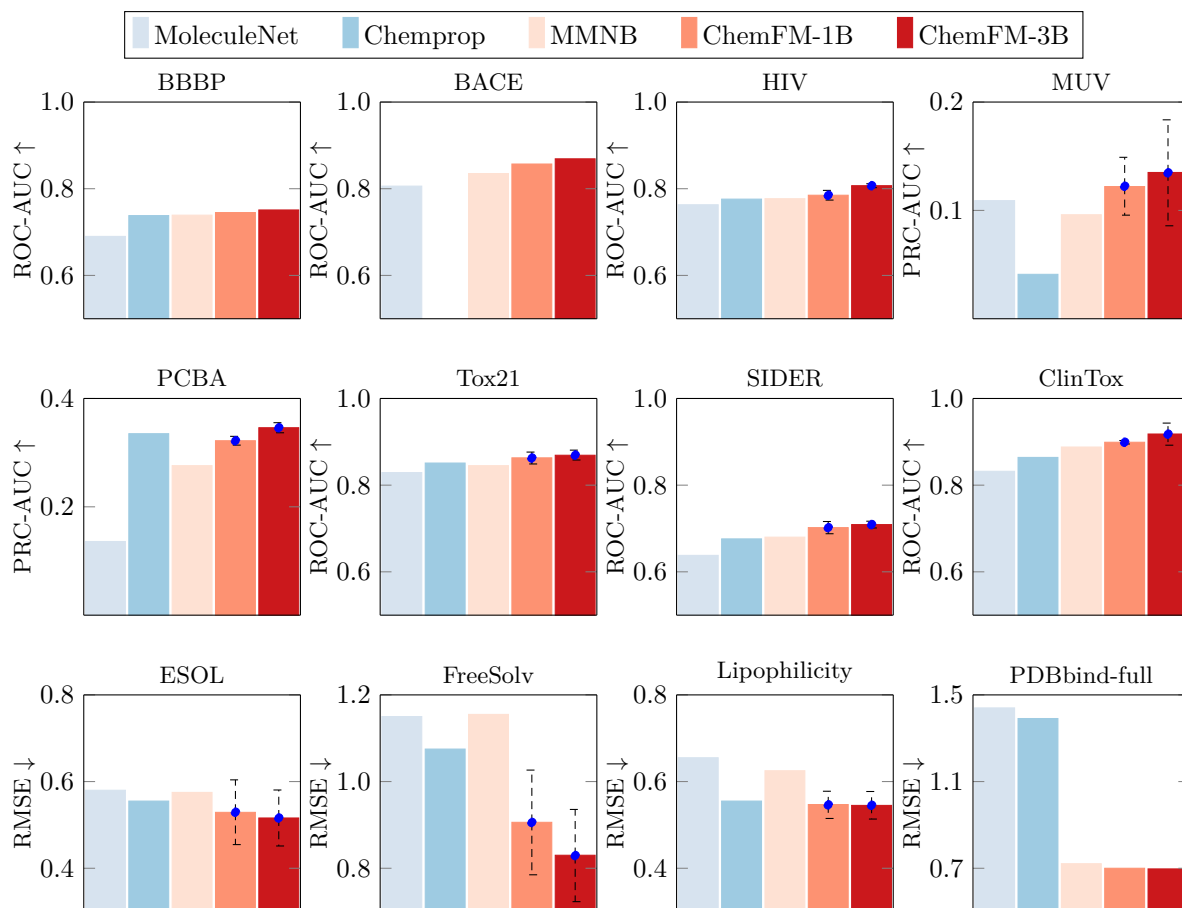
## Figures



**Fig. 1: Pre-training and unconditional molecular generation benchmarking of ChemFM models.** **a**, Pre-processing pipeline for ChemFM’s pre-training dataset. The pipeline starts with 178 million molecules from the UniChem database, initially represented by International Chemical Identifier (InChI)<sup>52</sup>. These InChIs are converted into canonical SMILES strings using RDKit<sup>53</sup>. The SMILES strings are then augmented tenfold through the SMILES enumeration technique<sup>54</sup>, resulting in approximately 1.78 billion SMILES strings for use as the pre-training dataset. **b**, Pre-training process for ChemFM. SMILES strings are segmented, tokenized, and terminated with an end token. These tokens are fed into ChemFM, a causal decoder-only transformer. Pre-training uses self-supervised causal language modeling, where the task is to predict each token based on preceding tokens. **c**, Pre-training performance of ChemFM-1B and ChemFM-3B models, measured by perplexity (exponentiated average negative log-likelihood) on the validation set. Models are trained through 818 billion tokens, slightly exceeding one epoch. **d**, Unconditional generation benchmarking for ChemFM-3B. A total of 100,000 molecules are generated randomly using a temperature setting of 1.0. The validity, uniqueness, and novelty scores of the generated molecules are reported. Additionally, internal diversity metrics (IntDiv<sub>1</sub>, IntDiv<sub>2</sub>) assess the diversity of the generated molecules, while KL similarity (KLSim) evaluates how closely the distribution of generated molecules aligns with that of the training dataset.



**Fig. 2: Illustrations of fine-tuning ChemFM model for downstream tasks.** **a**, Property prediction fine-tuning. During fine-tuning, the SMILES strings of molecules are augmented with a probability of 1.0 and tokenized before input to ChemFM. An MLP layer is added to the final token’s hidden state in the final layer to handle single or multiple regression or classification tasks. For inference, the canonical SMILES is input into ChemFM to predict the desired properties. **b**, Conditional molecular generation fine-tuning. This task is also framed as a sequence-to-sequence problem. The input comprises a sequence of conditions, each initiated by a unique property identification token followed by single or multiple tokens representing the property values. Classification values are encoded as special tokens corresponding to their class indices, continuous values are normalized and mapped into the embedding space using a shared MLP, and scaffolds are represented by their SMILES and tokenized into sequences. During fine-tuning, the target molecules are augmented with a probability of 1.0. **c**, Reaction prediction fine-tuning for both forward synthesis and retro-synthesis. These tasks are approached as sequence-to-sequence problems, where the model predicts the product (or reactant) sequence based on the reactant (or product) sequence. The root-aligned SMILES technique<sup>11</sup> is employed, aligning both sequences using the same root atom and augmenting them by enumerating different atoms as roots.



**Fig. 3: Performance comparison on 12 MoleculeNet<sup>33</sup> benchmark datasets for molecular property prediction.** All methods were evaluated using the *same* datasets, where we employed identical splitting methods and random seeds for data splitting, ensuring that train/validation/test data are the same for each data fold. Results for ChemFM (mean and standard deviation) are reported over three runs with different dataset folds, except for BBBP, BACE, and PDBbind-full, where only a single fold is provided in the original dataset. Values for models other than ChemFM are sourced from MMNB paper<sup>6</sup>. Metrics for classification tasks included ROC-AUC or PRC-AUC, while regression tasks were evaluated using RMSE. An upward arrow (↑) indicates that higher values are better, while a downward arrow (↓) indicates that lower values are better. An empty bar (Chemprop method in the BACE dataset) indicates that the result was not reported in the original paper. Empty standard deviation bars occur when only a single data fold is available.

## Tables

**Table 1: Comparison of representative molecular pre-trained models with ChemFM.**

Model	Pre-training Data	Parameter Size	Pre-training Strategies	Downstream Tasks
Mole-BERT <sup>18</sup>	ZINC15 <sup>55</sup> (2M)	1.86M	Masked atom modeling; contrastive learning	Property prediction
SMILES Transformer <sup>19</sup>	ChEMBL24 <sup>56</sup> (861K)	4.26M	Reconstruction	Property prediction
ChemBERTa-2 <sup>20</sup>	PubChem <sup>22</sup> (77M)	up to 46M	Masked language modeling; multi-task regression	Property prediction
Chemformer <sup>10</sup>	ZINC-15 <sup>55</sup> (100M)	up to 230M	Masked language modeling; SMILES canonicalization	Property prediction; reaction prediction; molecular optimization
ChemFM	UniChem <sup>27</sup> (178M)	up to 3B	Next token prediction	Property prediction; reaction prediction; molecular generation

We compare ChemFM with other representative chemical pre-trained models reported in the literature in terms of model size, pre-training data and scale, pre-training strategies, and downstream tasks. Models that are not open-sourced or whose reported performance could not be reliably validated through our replication efforts are excluded from this comparison. The downstream task performance of ChemFM relative to these pre-trained models is presented in Tables S2.1 (SMILES Transformer and Chemformer on property prediction), S2.3 (Mole-BERT and ChemBERTa-2 on property prediction), and S4.1 (Chemformer on reaction prediction), in Supplementary Information.

**Table 2: Performance comparison for conditional molecule generation on the GuacaMol<sup>43</sup> dataset.**

Property	Model	Validity $\uparrow$	Uniqueness $\uparrow$	Novelty $\uparrow$	Mean average deviation (MAD) $\downarrow$
logP	MolGPT	0.971	0.969	0.947	0.230
	ChemFM-3B	<b>0.981</b>	<b>0.981</b>	<b>0.966</b>	<b>0.182</b>
TPSA	MolGPT	0.971	0.969	0.945	3.562
	ChemFM-3B	<b>0.979</b>	<b>0.979</b>	<b>0.963</b>	<b>2.466</b>
SAS	MolGPT	0.978	0.974	0.941	0.133
	ChemFM-3B	<b>0.986</b>	<b>0.985</b>	<b>0.957</b>	<b>0.126</b>
QED	MolGPT	0.974	0.971	0.940	0.056
	ChemFM-3B	<b>0.982</b>	<b>0.982</b>	<b>0.963</b>	<b>0.045</b>
SAS + logP	MolGPT	0.972	0.963	0.947	0.147/0.253
	ChemFM-3B	<b>0.980</b>	<b>0.975</b>	<b>0.960</b>	<b>0.137/0.195</b>
SAS + TPSA	MolGPT	0.971	0.960	0.944	0.155/3.785
	ChemFM-3B	<b>0.980</b>	<b>0.971</b>	<b>0.956</b>	<b>0.138/2.659</b>
TPSA + logP	MolGPT	0.964	0.958	0.947	3.715/0.243
	ChemFM-3B	<b>0.973</b>	<b>0.970</b>	<b>0.962</b>	<b>2.415/0.184</b>
TPSA + logP + SAS	MolGPT	0.972	0.942	0.931	3.797/0.268/0.180
	ChemFM-3B	<b>0.975</b>	<b>0.946</b>	<b>0.936</b>	<b>2.289/0.191/0.166</b>

Molecules were generated based on desired property values, with a performance comparison between ChemFM-3B, which uses a single model, and MolGPT<sup>8</sup>, which uses 8 separate models. Metrics include validity, uniqueness, novelty, and mean absolute deviation (MAD) between the conditioned and actual properties of the generated molecules. **Bold** values indicate the best performance for each metric. It should be noted that validity, uniqueness, and novelty are computed against the total number of generated molecules, rather than only the valid ones, to more accurately reflect model performance ([Methods](#)).

**Table 3: Performance comparison of ChemFM with the best and second-best models on standard USPTO benchmarks for synthesis and retro-synthesis reaction prediction tasks, showing top-1, top-3, and top-5 accuracies (in percentages).**

Task category	Dataset	Model	Top-1	Top-3	Top-5
Synthesis	USPTO-MIT	Prev. best: AT <sup>49</sup>	90.4	-	96.5
		Prev. second-best: R-SMILES <sup>11</sup>	90.0	95.6	96.4
		ChemFM	<b>90.5</b>	<b>95.7</b>	<b>96.6</b>
Retro-synthesis	USPTO-50K	Prev. best: R-SMILES <sup>11</sup>	56.0	79.0	86.1
		Prev. second-best: Graph2Edits <sup>57</sup>	55.1	77.3	83.4
		ChemFM	58.0	<b>80.0</b>	<b>86.3</b>
		ChemFM*	<b>59.7</b>	79.2	84.2
	USPTO-MIT	Prev. best: R-SMILES <sup>11</sup>	60.3	77.9	82.8
		Prev. second-best: RetroTRAE <sup>58</sup>	60.3	77.9	82.8
		ChemFM	61.6	<b>78.7</b>	<b>83.0</b>
		ChemFM*	<b>62.4</b>	78.5	82.5
	USPTO-Full	Prev. best: RetroXpert <sup>59</sup>	49.4	63.6	67.6
		Prev. second-best: R-SMILES <sup>11</sup>	48.9	66.6	72.0
		ChemFM	<b>51.7</b>	<b>68.0</b>	<b>72.5</b>

The best and second-best models are determined based on top-1 performance. **Bold** values indicate the best performance for each metric. A hyphen “-” indicates that the value was not reported in the original paper. Results for R-SMILES are obtained through our replication using publicly available models<sup>11</sup>. ChemFM\* denotes ChemFM with further pre-training, which achieves better top-1 results but shows a decrease in top-3 and 5 performance.

## 4 Methods

### 4.1 Chemical language modeling

Molecular serialization systems, such as the simplified molecular input line entry system (SMILES)<sup>5</sup> or self-referencing embedded strings (SELFIES)<sup>60</sup>, represent molecules as linear sequences. This linearization enables the use of sequence-based models to effectively model chemical language. Formally, consider a corpus of molecules  $\mathcal{C} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m\}$ . Each molecule  $\mathbf{s}$  is represented as a sequence of tokens (subwords),  $\mathbf{s} = (t_1, t_2, \dots, t_n)$ , using a serialization system. The chemical language model is tasked with computing the joint probability of the sequence:

$$P(\mathbf{s}) = P(t_1, t_2, \dots, t_n).$$

ChemFM extends the principles of causal language modeling<sup>28</sup> by employing a unidirectional transformer decoder, also known as a causal decoder-only transformer, to model chemical language in an autoregressive manner. Within this framework, each token in the sequence is predicted based solely on its preceding tokens, allowing the joint probability of the sequence to be factorized as a Markov chain:

$$P(\mathbf{s}) = \prod_{i=1}^n p(t_i | t_1, \dots, t_{i-1}).$$

By pre-training on a large corpus of molecules, ChemFM learned the syntactic rules of serialization systems and the sequential dependencies inherent in molecular structures. These capabilities in representation learning can then be adapted to a wide range of chemical tasks.

### 4.2 Model architecture

The ChemFM models were based on TinyLlama<sup>61</sup>, a parameter-compact version of the Llama 2 architecture<sup>62</sup>, which employed a causal decoder-only transformer. In this work, we presented two model variations, ChemFM-1B, with 970 million trainable parameters, and ChemFM-3B, with 3.0 billion trainable parameters. These variations differ in the number of hidden layers, the number of attention heads, the dimension of the hidden representations, and the dimension of the multi-layer perceptron (MLP) representations. Detailed architectures for both models are outlined in Supplementary Table S1.1.

### 4.3 Molecule representation and tokenization

We utilized SMILES, a serialization format widely used in computational chemistry<sup>8,11,35</sup>, to represent molecular structures. Molecules were first transformed into SMILES strings using the RDKit library<sup>53</sup>. The resulting SMILES strings were segmented and tokenized using a sub-word tokenizer<sup>63</sup> with a predetermined vocabulary of 266 tokens. The vocabulary includes both uppercase and lowercase representations of the 118 elements from the periodic table, numerical digits from 0 to 9, 19 special symbols as specified by SMILES syntax<sup>5</sup>, and a special end token indicating the termination of a SMILES string. It should be noted that our vocabulary includes a small number of redundant tokens (e.g., lowercase element symbols that rarely or never appear in aromatic form). However, these have no impact on training efficiency or model performance. These tokens form the foundational vocabulary used during the pre-training phase, while additional special tokens are introduced during fine-tuning to address specific task requirements, as detailed in subsequent sections.

### 4.4 Pre-training dataset selection

Public chemical databases like ZINC<sup>21,55</sup>, PubChem<sup>22</sup>, ChemBL<sup>56</sup>, and UniChem<sup>27</sup> contain billions of molecules and are commonly used in pre-training chemical models. For example, Chemformer<sup>10</sup> utilized 100 million molecules randomly sampled from ZINC15, Grover<sup>64</sup> was developed on 11 million molecules sampled from ZINC15 and ChemBL, and MolFormer<sup>65</sup> employed a combination of ZINC15 and PubChem. However, the literature often lacks justification for specific dataset selections.

Considering the large scale of ChemFM and in order to avoid performance saturation, ZINC20 and UniChem (which encompasses most of the molecules from PubChem)—housing 1.8 billion and 178

million molecules respectively—are well-suited candidates for pre-training ChemFM. However, a large dataset alone does not guarantee sufficient information richness. Given the computational intensity of pre-training large models, careful dataset assessment is crucial prior to pre-training.

Therefore, we conducted a series of scaling experiments to evaluate the information content in the UniChem and ZINC20 datasets. The scaling laws of neural language models<sup>30</sup> reveal that model performance strongly depends on the scale of the model’s non-embedding parameters and the dataset size. Empirical evidence shows that performance (as measured by loss) follows a power-law relationship with each of these factors, provided the other is not bottlenecked. Our scaling experiments utilized causal decoder-only transformers from the ChemFM family, with models ranging from approximately 10M to 200M parameters. Each model was trained using cross-entropy loss, with a fixed data budget of 250,000 steps and a consistent batch size of 1,024 across all runs. The detailed architectures for the models used in these experiments are provided in Supplementary Table S1.1. The loss, measured on a validation dataset, was recorded at the end of each run.

For the UniChem dataset, we observed that the validation loss closely followed a power-law scaling with respect to the number of non-embedding parameters, showing no sign of performance saturation as the model size increased. In contrast, the ZINC20 dataset exhibited performance saturation when the model size reached 60M parameters. This suggests that the knowledge contained within ZINC20 becomes a bottleneck, limiting the model’s ability to benefit from increased parameter size.

To directly evaluate the impact of the pre-training dataset, we fully pre-trained a 1B-parameter model on ZINC20 and compared its downstream property prediction performance with the UniChem-pretrained counterpart, with details provided in Section 4.13.

## 4.5 Pre-training details

We selected the UniChem dataset for pre-training ChemFM. Using the SMILES data enumeration technique<sup>54</sup>, we augmented the dataset tenfold, resulting in a final pre-training dataset comprising 1.78 billion molecules, with 90% allocated for training and 10% for validation.

Both model variants were trained using the AdamW<sup>66</sup> optimizer. The learning rate was initially warmed up to  $4 \times 10^{-4}$  over 2,000 steps and then decayed following a cosine scheduler down to  $4 \times 10^{-5}$ . Each sequence was truncated to 512 tokens, and a batch size of 1,024 sequences was utilized. The models were trained for one epoch, processing a total of 818 billion tokens. The two ChemFM variants were pre-trained in a distributed manner on different hardware configurations: ChemFM-1B was trained across eight NVIDIA A100 nodes, each with 2×A100 80GB GPUs, while ChemFM-3B was trained across two NVIDIA HGX H100 nodes, each with 8×H100 80GB GPUs. The pre-training required 23.2 days for ChemFM-1B and 27.6 days for ChemFM-3B. Both pre-trained models and the training codes are publicly available ([Code availability](#)).

## 4.6 Benchmarking unconditional generation for pre-trained models

We evaluated the unconditional generation capability of the pre-trained ChemFM model by generating 100,000 molecules. For each molecule, the generation process began by sampling a start token according to its frequency distribution in the training dataset. The model then autoregressively generated tokens until producing an end token, thus completing the molecule. A temperature of 1.0 was applied to the SoftMax during generation.

We assessed the generated molecules using established metrics from molecule generation benchmarks such as GuacaMol<sup>43</sup> and MOSES<sup>44</sup>. Specifically, we measured the validity, uniqueness, novelty, and internal diversity (IntDiv<sub>1</sub>, IntDiv<sub>2</sub>, and Sphere Exclusion Diversity (SEDiv)). We also compared the distributions of 9 physicochemical descriptors (computed using the RDKit library) between 100,000 generated molecules and 100,000 randomly sampled from the training dataset (Supplementary Fig. S1.2). We quantified the similarity between these distributions by computing Kullback-Leibler (KL) divergence for each descriptor and aggregating them into a final KL similarity (KLSim) score. Additionally, ECFP4<sup>3</sup> fingerprints were computed for both sets, and their 2D t-SNE mapping was visualized to further evaluate how well the generated molecules aligned with the training data (Supplementary Fig. S1.3). We also compared internal diversity between the generated molecules and the pre-training dataset (UniChem) to further evaluate their alignment (Supplementary Table S1.2).

Details on these metrics can be found in the [Supplementary Information](#).

It is worth noting that the ChemFM models were trained on a dataset more than  $100\times$  larger than those used in the GuacaMol and MOSES benchmarks, which limits the validity of direct performance comparisons. This is also why the novelty of molecules generated by ChemFM is lower than values typically reported in the literature. To verify this, we compared ChemFM-3B with MolGPT by generating 100,000 molecules each and computing novelty separately against the GuacaMol and MOSES benchmarks. The results are reported in Supplementary Table S1.3. In this setting, ChemFM achieves higher novelty than MolGPT, confirming that the lower novelty observed with respect to UniChem reflects the scale of the reference database rather than a limitation of the model.

## 4.7 Training objective for property prediction

Fine-tuning ChemFM for supervised molecular property prediction tasks follows the framework of sequence classification and regression in causal language models<sup>28</sup>. Given a labeled dataset  $\mathcal{C}$ , each sample consists of a molecule  $\mathbf{s}$  represented as a SMILES string and its corresponding label set  $\mathbf{y} = (y_1, \dots, y_m)$  for  $m$  prediction tasks. These labels represent either regression or binary classification tasks but do not mix the two, following the settings in the MoleculeNet<sup>33</sup> and ADMET<sup>32</sup> benchmarks. The SMILES string  $\mathbf{s}$  is tokenized into a sequence of tokens  $t_1, t_2, \dots, t_n$ , terminated with a special end token. This tokenized sequence is processed by ChemFM, from which the hidden state  $h_l^n$  from the last layer  $l$  corresponding to the final token  $t_n$  is extracted. A linear layer,  $W_{\mathbf{y}} \in \mathbb{R}^{d_{\text{model}} \times m}$ , where  $d_{\text{model}}$  is the dimension of the model’s hidden representations, is applied to this hidden state, to make the predictions  $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_m)$  for the  $m$  tasks, as shown in Fig. 2a. For regression tasks, the model minimizes the mean square error (MSE) loss over the dataset:

$$\mathcal{L}_{\text{regression}} = \frac{1}{|\mathcal{C}|} \sum_{(\mathbf{s}, \mathbf{y}) \in \mathcal{C}} \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2.$$

For binary classification tasks, the model computes a probability distribution for each task using a Sigmoid activation function, and minimizes the binary cross-entropy loss:

$$\mathcal{L}_{\text{classification}} = -\frac{1}{|\mathcal{C}|} \sum_{(\mathbf{s}, \mathbf{y}) \in \mathcal{C}} \frac{1}{m} \sum_{i=1}^m P_i(y_i | \mathbf{s}),$$

where  $P_i(y_i | \mathbf{s}) = \text{Sigmoid}(\hat{y}_i)$  represents the predicted probability for task  $i$ .

## 4.8 Parameter efficient fine-tuning

Adapting all parameters of ChemFM is resource-intensive, requiring substantial GPU memory and storage. We utilized Low-Rank Adaptation (LoRA)<sup>29</sup>, a popular parameter-efficient fine-tuning technique that reduces the number of trainable parameters by introducing low-rank decomposition matrices for each layer instead of updating the full parameter set. We applied LoRA across all linear layers in the transformer blocks of ChemFM, while freezing the embedding layer, as no task-specific tokens are introduced for molecular property prediction tasks. The prediction head  $W_{\mathbf{y}}$  is fully adapted to predict labels.

The number of trainable parameters is controlled by adjusting the rank  $r$  of the decomposition matrices, and we report the number of trainable parameters for each task in Supplementary Table S2.9, S2.11, and S4.3. For instance, with  $r = 4$  in ChemFM-3B (using 32-bit float precision), the number of trainable parameters is reduced by  $460\times$ , from 3 billion to 6.5 million. This reduces video RAM requirements during training from 51 GB to 20 GB and checkpoint size from 12 GB to 26 MB.

## 4.9 Data pre-processing and training setups for property prediction

SMILES augmentation during training has been shown to improve molecular property prediction<sup>67</sup>. During training, we applied SMILES enumeration<sup>54</sup> with probability  $p = 1.0$ , while during inference, we used only canonical SMILES strings. Though synthesizing results from multiple augmentations could potentially improve performance, this approach was not explored in our experiments.

Through the reduction in GPU memory achieved by using the LoRA technique, both ChemFM-1B and ChemFM-3B can be fine-tuned on a single GPU. While our experiments were conducted on a single H100 80 GB GPU, the fine-tuning process is feasible on more modest hardware setups.

#### 4.10 Experimental setting on MoleculeNet benchmark for property prediction

We began by fine-tuning the ChemFM model on datasets from the MoleculeNet benchmark<sup>33</sup> (dataset descriptions are provided in Supplementary Table S2.8). While many methods have been developed and evaluated on the MoleculeNet benchmark, comparisons between them are often problematic due to variations in dataset splitting strategies and random seed selections across studies. This issue is particularly exacerbated by the “scaffold” split, which partitions molecules based on structural scaffolds. Although this method creates more structurally diverse and challenging train/validation/test folds than random splitting, it can result in significantly different test sets across experiments, complicating cross-study comparisons.

To ensure a comprehensive evaluation of ChemFM, we conducted three distinct sets of comparisons with existing methods, all using the same train/validation/test datasets. We excluded methods that are not open-sourced since verifying their dataset splits is not possible.

**Comparison set 1:** We first fine-tuned both ChemFM-1B and ChemFM-3B models on standard MoleculeNet datasets<sup>33</sup>, as provided in the MoleculeNet paper. Different datasets used different scaffold methods and are detailed in Supplementary Table S2.8. The methods we compared against include SMILES Transformer<sup>19</sup>, MoleculeNet models<sup>33</sup>, Direct Message Passing Neural Networks (D-MPNN or Chemprop)<sup>7</sup>, MolMapNet (MMNB)<sup>6</sup>, and Chemformer<sup>10</sup>. We conducted a random search for the training hyperparameters and LoRA configurations for each dataset, with the selected hyperparameters detailed in Supplementary Table S2.9. *Importantly, hyperparameter tuning was based solely on validation performance, and no tuning was performed on the test datasets.* We evaluated our models across three folds and reported the average performance, along with the corresponding split method and evaluation metrics, in Supplementary Table S2.1.

**Comparison set 2:** We then compared ChemFM with the AttentiveFP method<sup>35</sup>, which used random splitting methods but with different seeds than the standard MoleculeNet benchmark. MMNB provided a direct comparison with AttentiveFP (as shown in Table 2 of the MMNB paper<sup>6</sup>), where MMNB outperformed AttentiveFP on most datasets. Since ChemFM models consistently outperformed MMNB, we can reasonably infer that ChemFM is also superior to AttentiveFP on these datasets. However, on four specific datasets—Tox21, ESOL, FreeSolv, and Lipophilicity—AttentiveFP outperformed MMNB. For these datasets, we conducted a direct comparison by reevaluating AttentiveFP across three folds split by different random seeds and fine-tuning ChemFM using identical data splits. The results are presented in Supplementary Table S2.2.

**Comparison set 3:** The third comparison set focused on methods using a deterministic scaffold split to generate a single fold of train/validation/test sets, including Pretrain GNNs<sup>34</sup>, ChemBERTa-2<sup>20</sup>, 3D InfoMax<sup>36</sup>, Mole-BERT<sup>18</sup>, GraphMVP<sup>37</sup>, and MoleculeSDE<sup>38</sup>. We adopted the same settings as these methods—training on the same data fold with three different random training seeds (which only affect the training procedure like the network weights initialization and network dropout, but not dataset splitting)—and reported the average performance in Supplementary Table S2.3.

It is important to highlight that, despite differences in splitting methods and random seeds, no additional hyperparameter tuning was performed for comparison sets 2 and 3. We reuse the hyperparameters optimized for the standard MoleculeNet datasets (shown in Supplementary Table S2.9). For each dataset, we first fine-tuned the ChemFM-3B model. If ChemFM-3B did not outperform all other methods (specifically, the ESOL dataset in comparison set 2 and the MUV dataset in comparison set 3), we proceeded to fine-tune ChemFM-1B. The results indicated that at least one of our ChemFM models outperformed all other compared methods across the evaluated datasets, even without additional hyperparameter tuning on these specific data folds.

#### 4.11 Experimental setting on ADMET benchmark for property prediction

We compared ChemFM with methods on the leaderboard of the ADMET benchmark<sup>32</sup>, which comprises 22 datasets and provides standard data splits and performance evaluation metrics. The leaderboard

facilitates cross-method comparisons on these datasets. However, not all methods on the leaderboard are evaluated correctly. For example, common reasons for mis-evaluation included optimizing hyperparameters using the test datasets and combining the training and validation datasets for model training (practices that can improve performance, especially in the ADMET benchmarks, where most datasets contain fewer than 1,000 instances). We carefully reviewed the public codes of the methods on the leaderboard and excluded those that were mis-evaluated. The methods and corresponding reasons for exclusion are listed in the Supplementary Table S2.12.

For each dataset, we first conducted a hyperparameter search for the ChemFM-3B model. The adapted ChemFM-3B model outperforms the best models on the leaderboard for 20 out of the 22 datasets, with the exceptions of the Caco2\_Wang and HIA\_Hou datasets. For these two datasets, we then performed a hyperparameter search for the ChemFM-1B model, which can achieve state-of-the-art results. The comparisons between ChemFM and the previous best models are presented in Supplementary Table S2.4, with the hyperparameters used detailed in Supplementary Table S2.11.

#### 4.12 Effect of pre-training on ChemFM performance

To directly assess whether ChemFM’s improvements stem from pre-training rather than simply from model size, we fine-tuned ChemFM-3B on the ADMET dataset with the same hyperparameters as before but initialized the model from scratch (random initialization) instead of using pre-trained weights. The results are reported in the Supplementary Table S2.5 (A similar comparison between pre-training and non-pre-training was conducted for conditional molecular generation, as described in Methods).

As shown, models trained without pre-training perform substantially worse than the pre-trained ChemFM model across all property prediction tasks, demonstrating that the observed improvements are indeed due to large-scale pre-training rather than model size alone.

#### 4.13 Effect of pre-training dataset on ChemFM performance

To further evaluate the impact of the pre-training dataset, we conducted a direct comparison between ChemFM models pre-trained on UniChem and ZINC20. Specifically, we pre-trained a 1B-parameter model on ZINC20 using the same configuration and training steps as the UniChem-pretrained counterpart. Both models were then fine-tuned on the MoleculeNet datasets for molecular property prediction.

The results, summarized in Supplementary Table S2.6, show that the UniChem-pretrained model consistently outperforms the ZINC20-pretrained model on 9 out of 11 datasets, by a large margin, with the remaining two datasets yielding nearly identical performance. This demonstrates that pre-training on the more informative UniChem dataset leads to stronger molecular representations and superior downstream performance.

#### 4.14 Additional experiments on property prediction tasks

We further evaluated ChemFM on two molecular property prediction tasks extending to applications relevant for a broader community of chemists: odor prediction<sup>39</sup> and chromatographic retention time prediction<sup>40</sup>.

For odor prediction<sup>39</sup>, we used a dataset of approximately 5,000 molecules annotated with 138 odor labels, where each molecule may have multiple labels. This is formulated as a multi-label classification task. We compared ChemFM-3B with the open-source reproduction of the original MPNN-based approach (OpenPOM)<sup>68</sup>, ensuring identical 5-fold cross-validation splits (Supplementary Table S2.7).

For chromatographic retention time prediction, we used the METLIN small molecule retention time (SMRT) dataset<sup>40</sup>, which contains experimentally acquired reverse-phase chromatography measurements for up to 80,038 molecules. ChemFM-3B was compared against the baseline regression neural network built on molecular fingerprints and descriptors, using the same 75%/25% train/test random split (Supplementary Table S2.7).

#### 4.15 Experimental setting for potential antibiotics screening

We fine-tuned ChemFM-1B on a dataset used for screening potential antibiotics<sup>41</sup>. This dataset contains 39,312 compounds, with measurements of antibiotic activity based on RN4220 growth inhibition

and cytotoxicity data across three human cell types: liver carcinoma cells (HepG2), primary skeletal muscle cells (HSkMC), and lung fibroblast cells (IMR-90). For a fair comparison, we adhered to the data split protocol from previous work, using 80% of the dataset for training and 20% for testing. While exact train-test splits from the original study were not available, we ensured that active compounds in our train and test sets reflected a similar distribution to the full dataset (1.3% for antibiotic activity, 8.5% for HepG2 cytotoxicity, 3.8% for HSkMC cytotoxicity, and 8.8% for IMR-90 cytotoxicity), consistent with the original paper. We used an empirical hyperparameter setup (detailed in Supplementary Table S2.13) without additional hyperparameter tuning. For each task, in contrast to the previous study, which employed an ensemble of 20 Chemprop models<sup>7</sup>, we trained a single ChemFM model. Evaluation employed bootstrapping, with 100 resampled test sets generated by repeatedly drawing samples of equal size to the original test set. This approach allowed us to compute 95% confidence intervals for the PRC-AUC and capture the variability in precision-recall curves.

We further evaluated both ChemFM and Chemprop on an antibiotic library containing 1,994 real antibiotics<sup>42</sup>. To focus on structurally novel molecules, we deduplicated the library and excluded antibiotics with Tanimoto similarity scores below 0.5 to any known antibiotics in the training dataset, resulting in 1,173 novel molecules. Since both models were trained as classifiers, we applied a threshold of 0.5 to the prediction scores to distinguish positives from negatives. ChemFM labeled 149 molecules as positives, whereas Chemprop labeled only 29. Even when lowering the threshold to 0.4—following the approach used in Wong et al.<sup>41</sup> to identify antibiotic activity hits—Chemprop labeled only 42, still far fewer than ChemFM. Details of this antibiotics dataset and prediction scores from both models are provided in Supplementary Data 1 in a separate file.

#### 4.16 Experimental setting for data efficiency

To evaluate the data efficiency of ChemFM, we conducted experiments on both a regression and a classification task. Specifically, we selected the CYP2D6\_Substrate\_CarbonMangels classification task and the Half\_Life\_Obach regression task. For each task, we randomly sampled 10%, 20%, and 50% of the original training dataset to create reduced training subsets. To ensure robustness, we generated five independent subsets for each ratio. ChemFM-3B was fine-tuned on these subsets and compared against the previous best methods: Chemprop-RDKit<sup>69</sup> for CYP2D6\_Substrate\_CarbonMangels and DeepPurpose<sup>70</sup> for Half\_Life\_Obach. We used the full test set for evaluation. The final reported results for each ratio are the average over five runs, providing a comparison of model performance under different training data constraints (Supplementary Fig S2.2).

#### 4.17 Training objective for conditional generation

Conditional molecular generation tasks aim to produce molecules that meet specified criteria, such as desired molecular properties or structural constraints like scaffold fragments, and can be formalized as a sequence-to-sequence problem, where the goal is to generate a target sequence conditioned on a given input sequence. Let  $\mathcal{C}$  denote a dataset where each instance includes an input sequence,  $\mathbf{s}_i$ , representing the desired molecular characteristics or structural constraints (details on condition representation are provided in the next section), and a corresponding target molecular SMILES sequence,  $\mathbf{s}_o$ . These sequences are tokenized into series of tokens:  $\mathbf{s}_i = (t_{-m}, t_{-m+1}, \dots, t_0)$  for the input and  $\mathbf{s}_o = (t_1, t_2, \dots, t_n)$  for the target sequence. The ChemFM model generates the output sequence autoregressively, conditioned on the input sequence and previously generated tokens, which is illustrated in Fig. 2b. The training objective is to maximize the conditional probability distribution  $P(\mathbf{s}_o|\mathbf{s}_i)$ :

$$P(\mathbf{s}_o|\mathbf{s}_i) = \prod_{i=1}^n p(t_i|t_{-m}, \dots, t_0, \dots, t_{i-1}). \quad (1)$$

#### 4.18 Condition representation for conditional generation

In conditional molecular generation tasks, the input sequence can consist of multiple conditions, each represented by two components: a property name and a property value. A concrete example is shown in the Fig. 2b. The property name serves as a unique identifier and is denoted by a special token indicating the specific molecular property being conditioned upon. The property value can take one of three forms:

**Continuous values** Represented by a special placeholder token. These values are normalized before being processed by the model. They are then mapped into the embedding space through a shared linear layer, which is applied to all real-valued properties, allowing the model to capture the continuous nature of the property.

**Classification values** Encoded as special tokens that correspond to specific class indices. For example, a property "isRing" followed by a classification token "C1" indicates that the molecule should contain a ring structure, where the class index "1" denotes the presence of a ring.

**String representations** Used in cases such as scaffold-conditioned generation, where the scaffold fragment is represented as a SMILES string.

## 4.19 Data pre-processing and training details for conditional generation

We followed the experimental setup of MolGPT<sup>8</sup> to evaluate conditional molecular generation using the GuacaMol<sup>43</sup> and MOSES<sup>44</sup> datasets. The GuacaMol dataset focuses on generation based on molecular properties, while for the MOSES dataset includes both scaffold and molecular properties as generation conditions. For both datasets, we considered four continuous molecular properties: logP, synthetic accessibility score (SAS), topological polar surface area (TPSA), and quantitative estimate of drug-likeness (QED). These properties can be directly computed via RDKit<sup>53</sup>, enabling automatic performance evaluation of conditional molecular generation models. Unlike MolGPT, which requires 15 separate models to cover all property combinations for each dataset, we developed a single unified model for each dataset. This approach allows our models to handle multiple property combinations more flexibly. While it is feasible to train a single model that combines both datasets, we maintained separate models for GuacaMol and MOSES to ensure a fair comparison with MolGPT.

During training, we applied a probabilistic property selection strategy: one property was selected with a probability of 0.1, two properties with 0.2, three with 0.3, and four with 0.4. The order of properties was randomized. Additionally, we used the SMILES enumeration technique<sup>54</sup> with a probability of 1.0 to augment target SMILES strings. Both models underwent full-parameter fine-tuning using the AdamW optimizer with a weight decay of 0.01. The learning rate was initialized at  $6 \times 10^{-4}$ , with a warm-up phase spanning 0.1 epochs, and was decayed using a cosine schedule to a minimum of  $6 \times 10^{-5}$ . Fine-tuning was conducted on an NVIDIA HGX H100 node with 8×80GB GPUs for 10 epochs, using a batch size of 384.

## 4.20 Evaluation details for conditional generation

Our evaluation followed the setup of MolGPT to ensure a fair comparison. For property-based generation (GuacaMol dataset), we evaluated 8 distinct property combinations and compared performance with both cRNN<sup>45</sup> and MolGPT. For each combination, multiple sample points (representing specific property values) were tested, and for each point, we generated 10,000 molecules with the temperature setting to 1.0. The distribution of the generated molecules’ properties across sample points for each combination is presented in Supplementary Fig. S3.1. To assess the basic generation capabilities of the models, we reported the validity, uniqueness, and novelty scores for each property combination. Additionally, to evaluate how well the model adheres to the property conditions, we computed the mean absolute deviation (MAD) between the conditioned property and the computed property. These results are summarized and compared with cRNN and MolGPT in Supplementary Table S3.1. The MolGPT results were obtained by re-running the published checkpoints, while cRNN results are based on our reimplementations and training using the published code, since the original paper did not perform the same experiments as conducted here.

To demonstrate that pre-training benefits conditional molecular generation tasks, we additionally trained a randomly initialized ChemFM-3B model for this task. During these experiments, we observed that the model without pre-training was unstable to train and often diverged, producing almost entirely invalid molecules. After hyperparameter tuning, we were able to obtain a trained model and report the results in Supplementary Table S3.1. The results indicate that simply increasing model size does not yield performance improvements; in fact, the randomly initialized ChemFM-3B often performs worse than the much smaller MolGPT model. These findings confirm that pre-training is crucial for effective conditional molecular generation.

Instead of computing the novelty, and uniqueness scores against on the valid generated molecules, we compute these scores against the total number of generations. This is due to the fact that the standard uniqueness and novelty computation cannot effectively reflect model performance when validity is low. For example, assuming two models generating 10,000 molecules each, one generates 5,000 unique molecules out of 9,000 valid, while the other generates 5,300 unique molecules out of 9,800 valid. Although the second model exhibits better uniqueness performance, calculating uniqueness as a ratio would yield 0.56 for the first model (5,000/9,000) and 0.54 for the second (5,300/9,800).

For scaffold and property-based generation (MOSES dataset), we evaluated the model conditioned on five testing scaffolds and 8 different property combinations. For each sample point, 10,000 molecules were generated, and the distribution of the generated molecules’ properties is presented in Supplementary Fig. S3.3. Here, a valid molecule is defined by two criteria: 1) the SMILES string is syntactically correct and represents a feasible molecular structure, and 2) the scaffold of the generated molecule has a Tanimoto similarity of at least 0.8 to the desired scaffold. Instead of reporting the validity, novelty, and uniqueness scores, we directly presented the counts of valid, unique, and novel molecules generated. We also evaluated the count of molecules that retained the same scaffold as the conditioned scaffold and computed the MAD between the conditioned property and the generated property values. The results of this evaluation, compared with MolGPT, are presented in Supplementary Table S3.2.

In the scaffold-conditioned generation experiments, we observed that although ChemFM substantially outperforms baseline methods in both generation and matching metrics, a non-negligible fraction of generated molecules still failed to include the specified scaffold. Scaffold-constrained generation techniques such as PromptSMILES<sup>71</sup> are fully compatible with ChemFM: by rooting the desired scaffold at the beginning of the SMILES string, PromptSMILES ensures that the generated molecules contain the specified scaffold. Importantly, this approach does not require retraining a scaffold-specific conditioned model and can be directly applied to the property-conditioned ChemFM models for both property- and scaffold-conditioned generation.

## 4.21 Training objective for reaction prediction

We focused on both forward synthesis and retro-synthesis reaction prediction tasks, which leverage the same training objective used in conditional molecular generation as sequence-to-sequence problems. Let  $\mathcal{C}$  denote a reaction dataset, where each instance consists of an input sequence,  $\mathbf{s}_i$ , and a corresponding target sequence,  $\mathbf{s}_o$ . In the forward synthesis task,  $\mathbf{s}_i$  represents the reactants (and possibly includes reagents), while  $\mathbf{s}_o$  denotes the products. In retro-synthesis, the roles are reversed. Both input and target sequences are represented as SMILES strings. When multiple compounds appear in either the reactants or products, they are separated by a predefined delimiter, “,” in the SMILES representation. These sequences are then tokenized into series of tokens:  $\mathbf{s}_i = (t_{-m}, t_{-m+1}, \dots, t_0)$  for the input and  $\mathbf{s}_o = (t_1, t_2, \dots, t_n)$  for the target sequence. The ChemFM model generates the output sequence autoregressively, conditioned on the input sequence and previously generated tokens, as illustrated in Fig. 2c. The training objective is the same as in conditional molecular generation, defined in Eq. (1).

## 4.22 Datasets and pre-processing for reaction prediction

For reaction prediction tasks, we fine-tuned ChemFM-3B on widely-used USPTO-series datasets, including USPTO-50K<sup>48</sup>, USPTO-MIT<sup>47</sup>, and USPTO-Full<sup>46</sup>, commonly employed for benchmarking both forward synthesis and retro-synthesis tasks. Detailed statistics for these datasets are provided in Supplementary Table S4.2. In the forward synthesis task, we focused on the USPTO-MIT dataset with the setting where reactants and reagents are mixed in the input sequence. For retro-synthesis prediction, we conducted experiments on USPTO-50K, USPTO-MIT, and USPTO-Full datasets, focusing on the challenging setting where the reaction class is not provided. For the USPTO-Full dataset, following previous work<sup>11,49</sup>, we removed invalid reactions, such as those containing no products or just single ions as reactants.

Typically, input and output SMILES strings in reaction tasks vary significantly as they are pre-processed independently<sup>49</sup> and no inherent relationship between them is considered. Root-aligned SMILES (R-SMILES)<sup>11</sup>, however, defined a tight, one-to-one mapping between reactant and product SMILES by aligning the same atom as the root in both strings, making them more similar and improving the efficiency of reaction prediction. Following Zhong et al.<sup>11</sup>, we augmented the training data by

enumerating different root atoms, generating  $n$  augmented input-output pairs for each reaction. The augmentation folds for each dataset are determined based on dataset size: USPTO-50K was augmented 20-fold, and USPTO-MIT and USPTO-Full were augmented 5-fold. During inference, test data were also augmented to generate multiple input sequences. We employed beam search to generate  $m$  predictions for each augmented input sequence, yielding  $n \times m$  predictions. Both beam size and the number of generations were set to 10 for all experiments. Final predictions were selected based on the scores of these generations, following the scoring strategy of Zhong et al.<sup>11</sup>.

### 4.23 Training details for reaction prediction

Both input and output sequences are represented as R-SMILES and tokenized, with reactant and product sequences truncated at 512 tokens each. The context length of ChemFM is increased from 512 tokens (used in pre-training) to 1,024 tokens to accommodate these longer sequences. Given the large size of the reaction datasets (e.g., USPTO-Full contains approximately 1 million reactions), fine-tuning was performed on an NVIDIA HGX H100 node with  $8 \times 80$  GB GPUs. We used empirical hyperparameter settings without conducting hyperparameter searches. The detailed hyperparameter settings are provided in Supplementary Table S4.3. We should note that, in forward reaction prediction, we only reported the results with full-parameter fine-tuning. However, LoRA fine-tuning remains a viable option. While its performance was slightly below that of full-parameter fine-tuning, it was still strong and competitive, so we report the latter as the main result.

### 4.24 Comparison with state-of-the-art methods for reaction prediction

We compared the performance of our adapted ChemFM models with various sequence- and graph-based reaction prediction methods reported in the literature, using the same datasets and splits for a fair comparison. Methods that are not open-sourced or cannot be reproduced were excluded from the comparison. For the retro-synthesis task on the USPTO-50K, USPTO-MIT, and USPTO-Full datasets, our model outperformed existing methods by a significant margin (complete results are shown in Supplementary Table S4.1). For forward synthesis on the USPTO-MIT dataset, our initial results were just below the best-reported performance of Chemformer<sup>10</sup>. We observed that Chemformer simplified 903 reactions with multiple products into single-product reactions, which is inconsistent with the original USPTO-MIT dataset. When we excluded this portion of the test data to ensure a fair comparison, the top-1 accuracy of our model reached 91.4%, surpassing previously reported results of Chemformer.

## Data availability

The pre-training datasets for ChemFM are sourced from the UniChem database<sup>27</sup> (<https://ftp.ebi.ac.uk/pub/databases/chembl/UniChem/data/>). The molecular property prediction datasets are derived from the MoleculeNet<sup>33</sup> (<https://github.com/shenwanxiang/ChemBench>) and ADMET<sup>32</sup> ([https://tdcommons.ai/benchmark/admet\\_group/overview](https://tdcommons.ai/benchmark/admet_group/overview)) benchmarks. Datasets for molecular conditional generation tasks are sourced from the GuacaMol<sup>43</sup> database (<https://github.com/BenevolentAI/guacamol>) and the MOSES<sup>44</sup> database (<https://github.com/molecularsets/moses>). Datasets for reaction prediction tasks involving the USPTO series are formatted into Root-aligned SMILES<sup>11</sup> and are available at <https://github.com/otori-bird/retrosynthesis>.

## Code availability

The ChemFM-1B and ChemFM-3B models are publicly available on the Hugging Face Model Hub at <https://huggingface.co/ChemFM>. Additionally, the source code for pre-training and fine-tuning these models, along with the model checkpoints, can be accessed on our GitHub repository at <https://github.com/TheLuoFengLab/ChemFM> and has been archived on Zenodo at <https://zenodo.org/records/17450883>.

## Acknowledgements

This work was supported as part of the AIM for Composites, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences at Clemson University under

award #DE-SC0023389. We would also like to thank Clemson University’s Palmetto Cluster team for their invaluable support with cloud computing resources and maintenance.

## Author contributions

FC designed all models and experiments, performed all experiments and analyses, and drafted the initial version of the paper. KZ, TT, and YD assisted with the potential antibiotic screening testing. TZ, SP, and GL provided suggestions and feedback on the chemical aspects of this research. LL provided suggestions and feedback on the model learning aspects of this research. FL conceived the study and directed and supervised the whole study. All authors contributed to manuscript editing.

## Competing interests

The authors declare no competing interests.

## References

1. Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
2. Messeri, L. & Crockett, M. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
3. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742–754 (2010).
4. Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. Paper presented at the 35th Conference on Neural Information Processing Systems, Montreal, Canada, 7–12 December 2015.
5. Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28**, 31–36 (1988).
6. Shen, W. X. *et al.* Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nature Machine Intelligence* **3**, 334–343 (2021).
7. Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **59**, 3370–3388 (2019).
8. Bagal, V., Aggarwal, R., Vinod, P. & Priyakumar, U. D. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling* **62**, 2064–2076 (2021).
9. Du, Y. *et al.* Machine learning-aided generative molecular design. *Nature Machine Intelligence* 1–16 (2024).
10. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **3**, 015022 (2022).
11. Zhong, Z. *et al.* Root-aligned smiles: a tight representation for chemical reaction prediction. *Chemical Science* **13**, 9023–9034 (2022).
12. Oquab, M. *et al.* DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024).
13. Brown, T. *et al.* Language models are few-shot learners. Paper presented at the 8th International Conference on Learning Representations, Virtual Conference, 26 Apr–1 May 2020.
14. Dubey, A. *et al.* The llama 3 herd of models (2024). Preprint at <https://arxiv.org/abs/2407.21783>.

15. Zhou, Y. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
16. Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nature Methods* 1–11 (2024).
17. Wang, X. *et al.* A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* 1–9 (2024).
18. Xia, J. *et al.* Mole-bert: Rethinking pre-training graph neural networks for molecules. Paper presented at the 11th International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
19. Honda, S., Shi, S. & Ueda, H. R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738* (2019).
20. Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712* (2022).
21. Irwin, J. J. *et al.* Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling* **60**, 6065–6073 (2020).
22. Kim, S. *et al.* Pubchem substance and compound databases. *Nucleic acids research* **44**, D1202–D1213 (2016).
23. Frey, N. C. *et al.* Neural scaling of deep chemical models. *Nature Machine Intelligence* **5**, 1297–1305 (2023).
24. Zhang, D. *et al.* Chemllm: A chemical large language model (2024). Preprint at <https://arxiv.org/abs/2402.06852>.
25. Zhao, Z. *et al.* Chemdfm: Dialogue foundation model for chemistry (2024). Preprint at <https://arxiv.org/abs/2401.14818>.
26. Cai, F. *et al.* Mollangbench: A comprehensive benchmark for language-prompted molecular structure recognition, editing, and generation (2025). Preprint at <https://arxiv.org/abs/2505.15054>.
27. Chambers, J. *et al.* Unichem: a unified chemical structure cross-referencing and identifier tracking system. *Journal of cheminformatics* **5**, 3 (2013).
28. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* Improving language understanding by generative pre-training (2018).
29. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. Paper presented at the 10th International Conference on Learning Representations, Virtual Conference, 25–29 April 2022.
30. Kaplan, J. *et al.* Scaling laws for neural language models (2020). Preprint at <https://arxiv.org/pdf/2001.08361>.
31. Wu, Z. *et al.* Moleculenet: a benchmark for molecular machine learning. *Chemical science* **9**, 513–530 (2018).
32. Huang, K. *et al.* Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. Paper presented at the 35th Conference on Neural Information Processing Systems, Virtual Conference, 6–14 December 2021.
33. Wu, Z. *et al.* Moleculenet: a benchmark for molecular machine learning. *Chemical science* **9**, 513–530 (2018).

34. Hu, W. *et al.* *Strategies for pre-training graph neural networks* (2020).
35. Xiong, Z. *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry* **63**, 8749–8760 (2019).
36. Stärk, H. *et al.* 3d infomax improves gnn for molecular property prediction. Paper presented at the 39th International Conference on Machine Learning, Baltimore, USA, 17–23 July 2022.
37. Liu, S. *et al.* Pre-training molecular graph representation with 3d geometry. Paper presented at the 10th International Conference on Learning Representations, Virtual Conference, 25–29 April 2022.
38. Liu, S., Du, W., Ma, Z.-M., Guo, H. & Tang, J. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. Paper presented at the 40th International Conference on Machine Learning, Honolulu, USA, 23–29 July 2023.
39. Lee, B. K. *et al.* A principal odor map unifies diverse tasks in olfactory perception. *Science* **381**, 999–1006 (2023).
40. Domingo-Almenara, X. *et al.* The metlin small molecule dataset for machine learning-based retention time prediction. *Nature communications* **10**, 5811 (2019).
41. Wong, F. *et al.* Discovery of a structural class of antibiotics with explainable deep learning. *Nature* **626**, 177–185 (2024).
42. MedChemExpress. Antibiotic - medchemexpress. URL <https://www.medchemexpress.com/Targets/antibiotic/antibiotic.html>.
43. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* **59**, 1096–1108 (2019).
44. Polykovskiy, D. *et al.* Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology* **11**, 565644 (2020).
45. Kotsias, P.-C. *et al.* Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature Machine Intelligence* **2**, 254–265 (2020).
46. Lowe, D. Chemical reactions from us patents (1976-sep2016) (2017). URL <https://doi.org/10.6084/m9.figshare.5104873.v1>.
47. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network.
48. Dai, H., Li, C., Coley, C., Dai, B. & Song, L. Retrosynthesis prediction with conditional graph logic network.
49. Tetko, I. V., Karpov, P., Van Deursen, R. & Godin, G. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications* **11**, 5575 (2020).
50. Gao, W., Fu, T., Sun, J. & Coley, C. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in neural information processing systems* **35**, 21342–21357 (2022).
51. Thomas, M., Bou, A. & De Fabritiis, G. Test-time training scaling laws for chemical exploration in drug design. *arXiv preprint arXiv:2501.19153* (2025).
52. Trust, I. The international chemical identifier. URL <https://www.openai.com>.
53. Landrum, G. *et al.* Rdkit: Open-source cheminformatics. <https://www.rdkit.org/>.

54. Bjerrum, E. J. Smiles enumeration as data augmentation for neural network modeling of molecules (2017). Preprint at <https://arxiv.org/abs/1703.07076>.
55. Sterling, T. & Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling* **55**, 2324–2337 (2015).
56. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research* **47**, D930–D940 (2019).
57. Zhong, W., Yang, Z. & Chen, C. Y.-C. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications* **14**, 3009 (2023).
58. Ucak, U. V., Ashyrmamatov, I., Ko, J. & Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nature communications* **13**, 1186 (2022).
59. Yan, C. *et al.* Retroxpert: Decompose retrosynthesis prediction like a chemist. Paper presented at the 34th Annual Conference on Neural Information Processing Systems, Virtual Conference, 6–12 December 2020.
60. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **1**, 045024 (2020).
61. Zhang, P., Zeng, G., Wang, T. & Lu, W. Tinyllama: An open-source small language model (2024). Preprint at <https://arxiv.org/abs/2401.02385>.
62. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models (2023). Preprint at <https://arxiv.org/abs/2307.09288>.
63. Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. Paper presented at the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 Aug 2016.
64. Rong, Y. *et al.* Self-supervised graph transformer on large-scale molecular data. Paper presented at the 34th Annual Conference on Neural Information Processing Systems, Virtual Conference, 6–12 December 2020.
65. Ross, J. *et al.* Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* **4**, 1256–1264 (2022).
66. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Paper presented at the 7th international conference on learning representations, New Orleans, LA, 6–9 May 2019.
67. Born, J. *et al.* Chemical representation learning for toxicity prediction. *Digital Discovery* **2**, 674–691 (2023).
68. Barsainyan, A. A., Kumar, R., Saha, P. & Schmucker, M. Openpom - open principal odor map (2023). <https://github.com/BioMachineLearning/openpom>.
69. Swanson, K. *et al.* ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics* **40**, btae416 (2024).
70. Huang, K. *et al.* DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* **36**, 5545–5547 (2020).
71. Thomas, M., Ahmad, M., Tresadern, G. & De Fabritiis, G. PromptsMiles: prompting for scaffold decoration and fragment linking in chemical language models. *Journal of Cheminformatics* **16**, 77 (2024).

# Supplementary Information

## S1 Supplementary information for pre-training benchmarking

### Definition of benchmark metrics for unconditional molecular generation

**Validity:** The validity score measures the proportion of generated SMILES strings that are valid. A SMILES string is considered valid if it is syntactically correct and represents a feasible molecular structure, such as correct atom valency and consistent bond arrangements in aromatic rings. In our experiments, validity is implicitly checked by the RDKit parser when converting a SMILES string into an RDKit molecule object<sup>41</sup>.

**Uniqueness:** To ensure the model does not collapse into generating a subset of repetitive molecules, we measure the uniqueness score, defined as the proportion of unique SMILES strings among the total generated SMILES strings.

**Novelty:** The novelty score evaluates the model’s ability to explore chemical space by generating new molecules that are not present in the training dataset, which can be defined as the proportion of novel molecules among generations:

**Internal diversity:** Generative model often encounter the issue of mode collapse, where the generated molecules are concentrated in a small region of chemical space. Internal diversity measures how well the model generates diverse molecules by penalizing high similarity between molecules pairs within the generated set, and it is defined as:

$$\text{IntDiv}_p(G) = 1 - \sqrt[p]{\frac{1}{|G|^2} \sum_{m_1, m_2 \in G} S(m_1, m_2)^p},$$

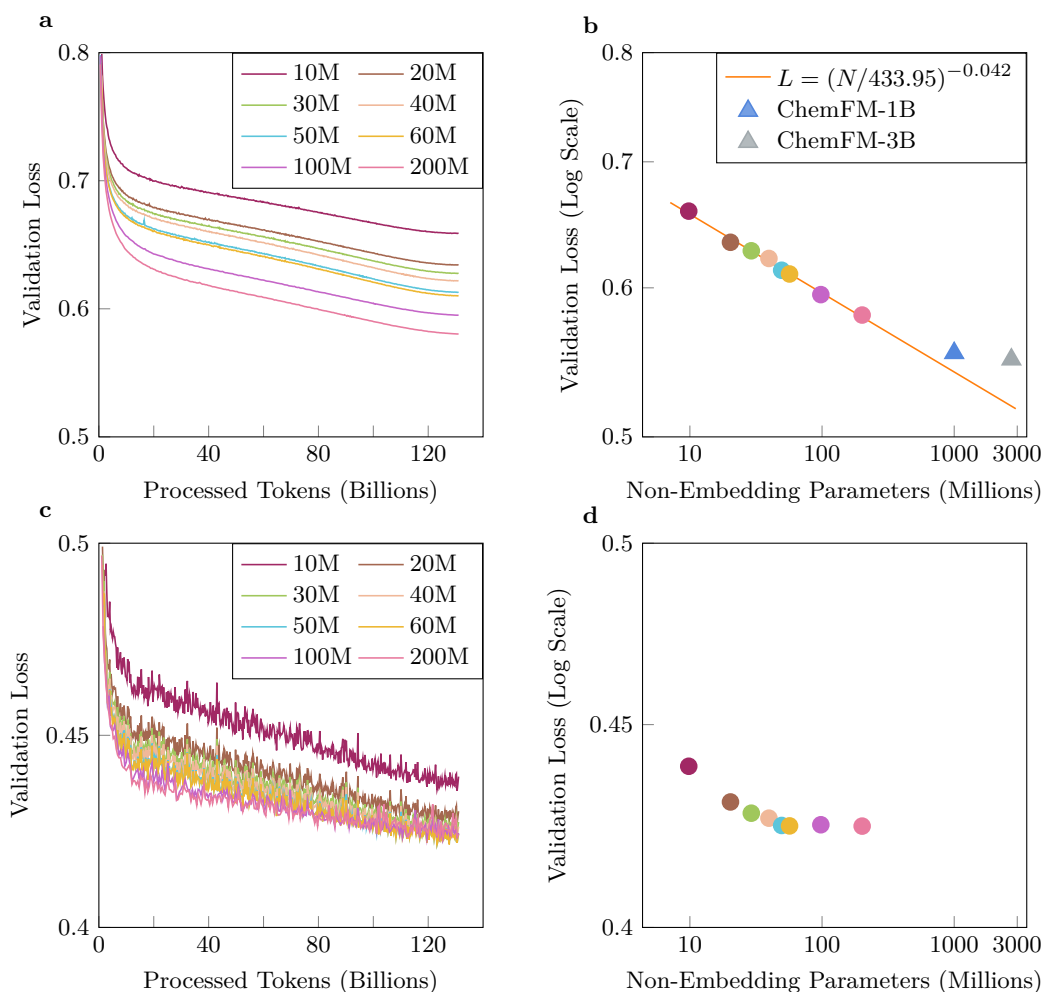
where  $S(\cdot)$  is the Tanimoto Similarity between molecule pair  $m_1$  and  $m_2$  in the generated set  $G$ . We evaluate both  $\text{IntDiv}_1$  and  $\text{IntDiv}_2$  in our experiment.

**Sphere exclusion diversity:** Sphere exclusion diversity is a quantitative measure of molecular diversity derived from the principle of sphere exclusion clustering. In this approach, molecules are embedded in a chosen chemical space (e.g., using structural fingerprints), and representative compounds are sequentially selected while all others within a predefined similarity radius are excluded. This process continues until no unassigned molecules remain, ensuring that the resulting set consists of compounds that are mutually dissimilar by at least the specified threshold. Consequently, sphere exclusion diversity provides an effective means of capturing the breadth of chemical space in a compound library, reducing redundancy, and enhancing coverage of structurally distinct scaffolds. In our evaluation, we employed the MolScore<sup>42</sup> framework with Morgan fingerprints as the representation space, and we set a Tanimoto distance threshold of 0.65.

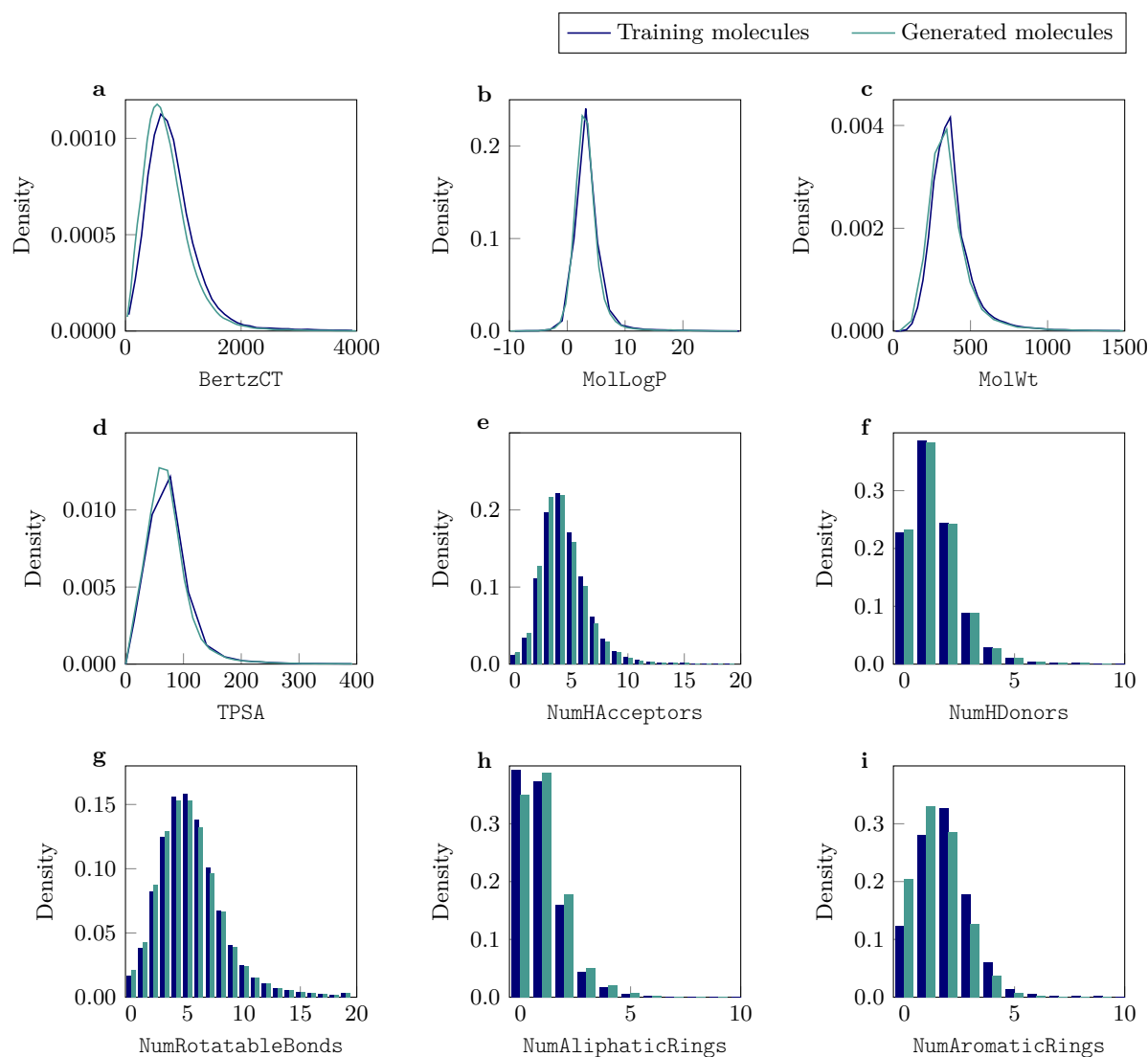
**KL similarity:** Kullback-Leibler similarity (KLSim) measures how closely the distribution of generated molecules matches that of the training dataset. We compare the distributions of various physicochemical descriptors between the generated set  $G$  and the training set  $T$  using KL divergence, and the KL similarity score is computed as:

$$\text{KLSim}(G, T; \text{Desp}) = \frac{1}{|\text{Desp}|} \sum_{i=1}^{|\text{Desp}|} e^{-D_{\text{KL}}(\text{Desp}_i(G), \text{Desp}_i(T))},$$

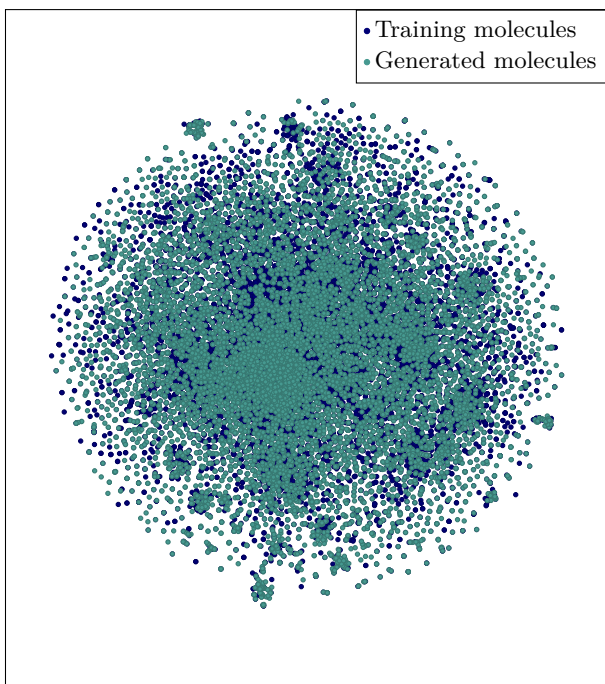
where  $\text{Desp}$  is a set of the descriptors measured using RDKit toolbox, and  $D_{\text{KL}}$  is the KL divergence between the two distributions. We follow the settings established by the GuacaMol<sup>21</sup> benchmarking platform, which measures 9 different molecular descriptors, including molecular complexity, molecular weight, etc. The comparison of these descriptor distributions between the generated set and the training set is illustrated in Supplementary Fig. S1.2. Due to the computational expense of using the entire training dataset, we use a representative subset of 100,000 molecules from the training dataset. Additionally, given the high diversity of the training dataset, a high KL similarity score also suggests that the generated molecules maintain a high degree of diversity<sup>21</sup>.



**Fig. S1.1: Comparison of chemical language model pre-training on the UniChem<sup>1</sup> and ZINC20<sup>2</sup> datasets.** **a, c,** Validation loss trajectories for models trained on the UniChem (a) and ZINC20 (c) datasets using varying model sizes. The models compared here range from approximately 10M to 200M parameters, excluding embeddings. **b,** For the UniChem dataset, the non-embedding parameters ( $N$ ) and validation loss ( $L$ ) closely adhere to an exponential scaling law. However, as model sizes increase to 1B parameters (ChemFM-1B) and further to 3B parameters (ChemFM-3B), the validation loss begins to deviate from the expected power law, suggesting that the performance gains from further increases in parameter size are approaching saturation. **d,** In contrast, for the ZINC20 dataset, validation loss reaches saturation when parameter size exceeds 60M.



**Fig. S1.2: Comparison of physicochemical descriptor distributions between training and generated molecules.** The descriptors were computed for 178 million molecules in the training dataset and 100,000 molecules randomly sampled from the ChemFM-3B model, using RDKit<sup>41</sup>. The descriptors are: **a**, BertzCT, a topological index quantifying molecular complexity; **b**, MolLogP, the octanol-water partition coefficient; **c**, MolWt, molecular weight; **d**, TPSA, topological polar surface area; **e**, NumHAcceptors, number of hydrogen bond acceptors; **f**, NumHDonors, number of hydrogen bond donors; **g**, NumRotatableBonds, number of rotatable single bonds; **h**, NumAliphaticRings, number of aliphatic (non-aromatic) rings; **i**, NumAromaticRings, number of aromatic rings.



**Fig. S1.3: 2D T-SNE visualization of ECFP4 fingerprints<sup>43</sup> for training and generated molecules.** ECFP4 fingerprints were computed using RDKit<sup>41</sup>. To enhance the computational efficiency of the t-SNE mapping, we randomly sampled 10,000 molecules from both the training dataset and the molecules generated by the ChemFM-3B model.

**Table S1.1: Architectures of the ChemFM models.**

Model	$n_{\text{params}}$	$n_{\text{layers}}$	$n_{\text{heads}}$	$n_{\text{ctx}}$ (pre-training)	$d_{\text{model}}$	$d_{\text{ff}}$
ChemFM-1B	970M	22	32	512	2048	5632
ChemFM-3B	3.0B	30	48	512	3072	8640
ChemFM-10M	9.8M	3	8	512	512	1408
ChemFM-20M	20.3M	4	10	512	640	1760
ChemFM-30M	29.2M	4	12	512	768	2112
ChemFM-40M	39.6M	4	14	512	896	2464
ChemFM-50M	49.5M	5	14	512	896	2464
ChemFM-60M	56.8M	5	15	512	960	2640
ChemFM-100M	97.9M	6	18	512	1152	3168
ChemFM-200M	201.1M	10	20	512	1280	3520

ChemFM-1B and ChemFM-3B are the primary models, while ChemFM-10M to ChemFM-200M are the models used for the pre-training dataset selection experiments.  $n_{\text{params}}$  is the actual number of non-embedding trainable parameters,  $n_{\text{layers}}$  is the number of hidden layers in the Transformer decoder,  $n_{\text{heads}}$  is the number of attention heads for each attention layer in the Transformer decoder,  $n_{\text{ctx}}$  is the context length,  $d_{\text{model}}$  is the dimension of the hidden representations, and  $d_{\text{ff}}$  is the dimension of the MLP representations. It should be noted that, for training efficiency, the context lengths during fine-tuning may differ from those used in pre-training, depending on the maximum length of the dataset.

**Table S1.2: Comparison of molecular diversity between ChemFM-3B generated molecules and the UniChem training dataset.**

	IntDiv1	IntDiv2	SEDiv
UniChem	0.903/0.903	0.895/0.894	0.697/0.377
ChemFM-3B	0.904/0.904	0.895/0.896	0.701/0.388

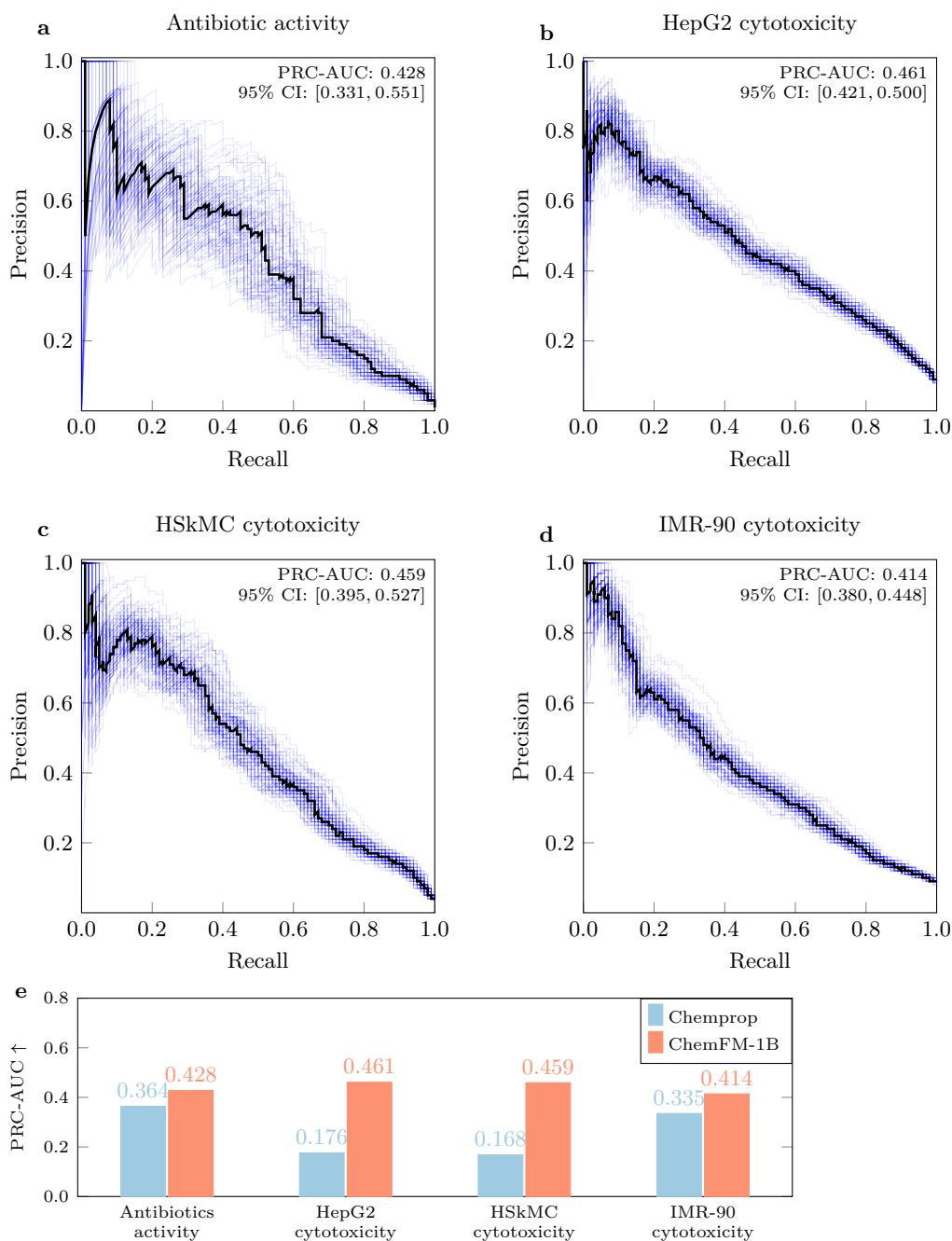
We report three diversity metrics—IntDiv1, IntDiv2, and Sphere Exclusion Diversity (SEDiv) (computed at sample sizes of 10,000/100,000 molecules).

**Table S1.3: Novelty of generated molecules for ChemFM-3B and MolGPT.**

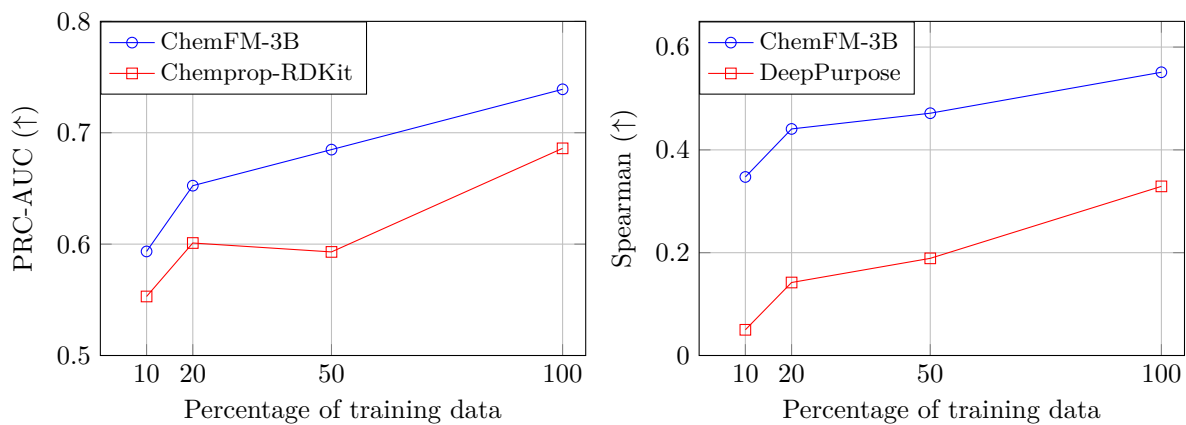
Benchmark	Model	Valid / Unique	Novelty (%)
MOSES	ChemFM-3B	99,596 / 99,585	<b>100.0</b>
	MolGPT	99,737 / 98,865	99.1
GuacaMol	ChemFM-3B	99,596 / 99,585	<b>100.0</b>
	MolGPT	93,254 / 87,439	93.8

Both models generated 100,000 molecules, and novelty was evaluated separately with respect to MOSES and GuacaMol.

## S2 Supplementary information for property prediction tasks



**Fig. S2.1: Precision-recall curve for ChemFM model for predicting antibiotic activity and human cell cytotoxicity.** **a**, Precision-recall curve for antibiotic activity prediction based on *S. aureus* RN4220 growth inhibition. **b**, **c**, **d**, Precision-recall curves for human cell cytotoxicity prediction in human liver carcinoma cells (HepG2) (**b**), human primary skeletal muscle cells (HSkMC) (**c**), and human lung fibroblasts cells (IMR-90) (**d**). The black lines represent precision-recall curves for the ChemFM models, while blue curves with 95% confidence intervals (CI) show the variation from 100 bootstrapping iterations. The datasets are obtained from previous work<sup>20</sup> by screening 39,312 compounds. In the original study, ten Chemprop<sup>5</sup> models were trained and ensemble for each property. Here, we trained a single ChemFM model per property using the same datasets for direct performance comparison. **e**, Comparison of PRC-AUC values between ChemFM and Chemprop models for all four properties.



**Fig. S2.2: Comparison between ChemFM-3B and the previous best method across different percentages of the training dataset. a, CYP2D6\_Substrate\_CarbonMangels classification task, compared with Chemprop-RDKit. b, Half\_Life\_Obach regression task, compared with DeepPurpose.**

**Table S2.1: Performance comparison on 12 MoleculeNet benchmark datasets for molecular property prediction.**

Category	Dataset	Task metric	MoleculeNet <sup>3</sup>	ST <sup>4</sup>	Chemprop <sup>5</sup>	MMNB <sup>6</sup>	CF <sup>7</sup>	ChemFM (1B/3B)
Pharmacokinetic	BBBP	ROC-AUC $\uparrow$	0.690 (Weave)	0.704	0.738	<u>0.739</u>	-	0.745/ <b>0.751</b>
Bioactivity	BACE	ROC-AUC $\uparrow$	0.806 (Weave)	0.701	-	<u>0.835</u>	-	0.857/ <b>0.869</b>
	HIV	ROC-AUC $\uparrow$	0.763 (GC)	<u>0.807</u>	0.729	0.777	-	0.785/ <b>0.807</b>
	MUV	PRC-AUC $\uparrow$	<u>0.109</u> (Weave)	-	0.041	0.096	-	0.122/ <b>0.135</b>
	PCBA	PRC-AUC $\uparrow$	0.136 (GC)	-	<u>0.335</u>	0.276	-	0.322/ <b>0.346</b>
Toxicity	Tox21	ROC-AUC $\uparrow$	0.829 (GC)	0.802	<u>0.851</u>	0.845	-	0.863/ <b>0.869</b>
	SIDER	ROC-AUC $\uparrow$	0.638 (GC)	-	0.676	<u>0.680</u>	-	0.702/ <b>0.709</b>
	ClinTox	ROC-AUC $\uparrow$	0.832 (Weave)	<u>0.895</u>	0.864	<u>0.888</u>	-	0.899/ <b>0.918</b>
Physicochemical	ESOL	RMSE $\downarrow$	0.580 (MPNN)	0.720	<u>0.555</u>	0.575	0.633	0.529/ <b>0.516</b>
	FreeSolv	RMSE $\downarrow$	1.150 (MPNN)	1.650	<u>1.075</u>	1.155	1.230	0.906/ <b>0.830</b>
	Lipop	RMSE $\downarrow$	0.655 (GC)	0.921	<u>0.555</u>	0.625	0.598	0.547/ <b>0.545</b>
Molecular binding	PDBbind	RMSE $\downarrow$	1.440 (GC)	-	1.391	<u>0.721</u>	-	0.700/ <b>0.697</b>

All methods in this table were evaluated using the *same* dataset splits from Shen et al. <sup>6</sup>. **Bold** values indicate the best-performing models, while underline values represent the best performance excluding ChemFM. The ChemFM results are averaged over three runs with different dataset folds, while values for other models are sourced from Shen et al. <sup>6</sup>. Metrics for classification tasks include ROC-AUC or PRC-AUC, while regression tasks are evaluated using RMSE. An upward arrow ( $\uparrow$ ) indicates that higher values are better, while a downward arrow ( $\downarrow$ ) indicates that lower values are better. ST denotes the SMILES Transformer smilestransformer, and CF denotes the Chemformer chemformer. The brackets following each value in MoleculeNet denote the type of model used: Weave (Weave models), MPNN (Message Passing Neural Networks), and GC (Graph Convolutional Models).

**Table S2.2: Performance comparison between AttentiveFP<sup>8</sup> and ChemFM-3B on four MoleculeNet benchmark datasets for molecular property prediction.**

Dataset	Task metric	Split	AttentiveFP <sup>8</sup>	ChemFM-3B
Tox21	ROC-AUC $\uparrow$	Random	$0.848 \pm 0.007$	<b><math>0.861 \pm 0.002</math></b>
ESOL	RMSE $\downarrow$	Random	$0.545 \pm 0.021$	<b><math>0.539 \pm 0.033</math></b>
FreeSolv	RMSE $\downarrow$	Random	$1.145 \pm 0.183$	<b><math>0.675 \pm 0.050</math></b>
Lipophilicity	RMSE $\downarrow$	Random	$0.637 \pm 0.027$	<b><math>0.543 \pm 0.024</math></b>

According to the evaluation in Table 2 of from Shen et al.<sup>6</sup>, MMNB generally outperformed AttentiveFP, while our ChemFM models performed better than MMNB across all benchmarks. Therefore, for most datasets, we did not directly compare ChemFM with AttentiveFP, as it was reasonable to assume ChemFM’s superiority. However, these four datasets represent cases where MMNB performed worse than AttentiveFP, making them essential for a direct comparison. We reran AttentiveFP with three different random seeds and used identical data splits from AttentiveFP to fine-tune our ChemFM-3B model. No hyperparameter tuning was conducted for ChemFM-3B in these experiments; we used the hyperparameters optimized for the standard MoleculeNet benchmark datasets, as presented in Supplementary Table S2.9. **Bold** values indicate the best-performing models.

**Table S2.3: Performance comparison with Pretrain GNNs<sup>9</sup>, ChemBerta-2<sup>10</sup>, 3D InfoMax<sup>11</sup>, Mole-BERT<sup>12</sup>, GraphMVP<sup>13</sup>, and MoleculeSDE<sup>14</sup> on 11 MoleculeNet benchmark datasets for molecular property prediction.**

Dataset	Pretrain GNNs <sup>9</sup>	Chem BERTa <sup>10</sup>	3D InfoMax <sup>11</sup>	Mole-BERT <sup>12</sup>	GraphMVP <sup>13</sup>	MoleculeSDE <sup>14</sup>	ChemFM
BBBP	0.688 ± 0.008	0.728	0.708 ± 0.005	0.719 ± 0.016	0.724 ± 0.016	<u>0.732</u> ± 0.005	<b>0.733</b> ± 0.007
Tox21	0.767 ± 0.004	-	0.749 ± 0.008	0.768 ± 0.005	0.759 ± 0.005	<u>0.768</u> ± 0.003	<b>0.795</b> ± 0.007
ToxCast	0.642 ± 0.005	-	0.635 ± 0.008	0.643 ± 0.002	0.631 ± 0.002	<u>0.652</u> ± 0.003	<b>0.688</b> ± 0.004
SIDER	0.610 ± 0.007	-	0.568 ± 0.021	0.628 ± 0.011	<u>0.639</u> ± 0.012	0.608 ± 0.004	<b>0.650</b> ± 0.014
ClinTox	0.718 ± 0.041	0.563	0.627 ± 0.033	0.789 ± 0.030	0.791 ± 0.028	<u>0.870</u> ± 0.005	<b>0.941</b> ± 0.017
MUV	0.758 ± 0.017	-	0.762 ± 0.014	0.786 ± 0.018	0.777 ± 0.006	<u>0.809</u> ± 0.004	<b>0.812</b> * ± 0.002
HIV	0.773 ± 0.010	-	0.761 ± 0.013	0.782 ± 0.008	0.770 ± 0.012	<u>0.788</u> ± 0.009	<b>0.793</b> ± 0.013
BACE	0.796 ± 0.012	0.799	0.786 ± 0.019	0.808 ± 0.014	<u>0.812</u> ± 0.009	0.804 ± 0.009	<b>0.853</b> ± 0.005
ESOL	-	0.889	<u>0.894</u> ± 0.028	1.015 ± 0.030	1.029 ± 0.033	-	<b>0.844</b> ± 0.032
FreeSolv	-	-	<u>2.337</u> ± 0.227	-	-	-	<b>2.130</b> ± 0.151
Lipop	-	0.798	0.695 ± 0.012	<u>0.676</u> ± 0.017	0.681 ± 0.010	-	<b>0.625</b> ± 0.010

All methods in this table were evaluated using the *same* datasets, with a deterministic scaffold split applied to define the train/validation/test sets. Results are averaged over three random training seeds, except for ChemBERTa-2, whose standard deviation were not reported in the original paper. A horizontal line separates classification and regression tasks: classification tasks are evaluated using ROC-AUC (higher values are better), while regression tasks are evaluated using RMSE (lower values are better). **Bold** values indicate the best-performing models, and underlined values highlight the best performance excluding ChemFM. All ChemFM results were fine-tuned from ChemFM-3B, except those marked with a \*, which were fine-tuned from ChemFM-1B.

**Table S2.4: Performance comparison on 22 ADMET benchmark datasets with the previous best models for each molecular property prediction task.**

Category	Dataset	Task metric	Previous best	ChemFM
Absorption	Caco2_Wang	MAE ↓	0.330 ± 0.024@Chemprop-RDKit <sup>15</sup>	<b>0.322*</b> ± 0.026
	Bioavailability_Ma	ROC-AUC ↑	0.672 ± 0.021@DeepPurpose <sup>16</sup>	<b>0.715</b> ± 0.011
	Lipophilicity_AstraZeneca	MAE ↓	0.467 ± 0.006@Chemprop-RDKit <sup>15</sup>	<b>0.460</b> ± 0.006
	Solubility_AqSolDB	MAE ↓	0.761 ± 0.025@Chemprop-RDKit <sup>15</sup>	<b>0.725</b> ± 0.011
	HIA_Hou	ROC-AUC ↑	0.981 ± 0.002@Chemprop-RDKit <sup>15</sup>	<b>0.984*</b> ± 0.004
	Pgp_Broccatelli	ROC-AUC ↑	0.929 ± 0.006@AttrMasking <sup>17</sup>	<b>0.931</b> ± 0.003
Distribution	BBB_Martins	ROC-AUC ↑	0.897 ± 0.004@ContextPred <sup>17</sup>	<b>0.908</b> ± 0.010
	PPBR_AZ	MAE ↓	7.788 ± 0.210@Chemprop <sup>7</sup>	<b>7.505</b> ± 0.073
	VDss_Lombardo	Spearman ↑	0.561 ± 0.025@DeepPurpose <sup>16</sup>	<b>0.662</b> ± 0.013
Metabolism	CYP2C9_Veith	PRC-AUC ↑	0.777 ± 0.003@Chemprop-RDKit <sup>15</sup>	<b>0.788</b> ± 0.005
	CYP2D6_Veith	PRC-AUC ↑	0.673 ± 0.007@Chemprop-RDKit <sup>15</sup>	<b>0.704</b> ± 0.003
	CYP3A4_Veith	PRC-AUC ↑	0.876 ± 0.003@Chemprop-RDKit <sup>15</sup>	<b>0.878</b> ± 0.003
	CYP2C9_Substrate_CarbonMangels	PRC-AUC ↑	0.400 ± 0.008@Chemprop-RDKit <sup>15</sup>	<b>0.414</b> ± 0.027
	CYP2D6_Substrate_CarbonMangels	PRC-AUC ↑	0.686 ± 0.031@Chemprop-RDKit <sup>15</sup>	<b>0.739</b> ± 0.024
	CYP3A4_Substrate_CarbonMangels	ROC-AUC ↑	0.619 ± 0.030@Chemprop-RDKit <sup>15</sup>	<b>0.654</b> ± 0.022
Excretion	Half_Life_Obach	Spearman ↑	0.329 ± 0.083@DeepPurpose <sup>16</sup>	<b>0.551</b> ± 0.020
	Clearance_Hepatocyte_AZ	Spearman ↑	0.439 ± 0.026@ContextPred <sup>17</sup>	<b>0.495</b> ± 0.030
	Clearance_Microsome_AZ	Spearman ↑	0.599 ± 0.025@Chemprop-RDKit <sup>15</sup>	<b>0.611</b> ± 0.016
Toxicity	LD50_Zhu	MAE ↓	0.606 ± 0.024@Chemprop <sup>7</sup>	<b>0.541</b> ± 0.015
	hERG	ROC-AUC ↑	0.841 ± 0.020@DeepPurpose <sup>16</sup>	<b>0.848</b> ± 0.009
	AMES	ROC-AUC ↑	0.850 ± 0.004@Chemprop-RDKit <sup>15</sup>	<b>0.854</b> ± 0.007
	DILI	ROC-AUC ↑	0.919 ± 0.008@ContextPred <sup>17</sup>	<b>0.920</b> ± 0.012

We used the default metrics and scaffold split methods, with the data pre-split by the benchmark. **Bold** values indicate the best-performing models. Evaluation metrics include ROC-AUC, PRC-AUC, MAE, and Spearman’s rank correlation coefficient, with an upward arrow (↑) indicating that higher values are better and a downward arrow (↓) indicating that lower values are better. All ChemFM results were fine-tuned from the ChemFM-3B model, except those marked with a \*, which were fine-tuned from the ChemFM-1B model.

**Table S2.5: Performance comparison on 22 ADMET benchmark datasets between ChemFM fine-tuned from pre-trained weights and from random initialization.**

Category	Dataset	Task metric	ChemFM (w/o pretraining)	ChemFM
Absorption	Caco2_Wang	MAE ↓	0.390 ± 0.017	<b>0.322*</b> ± 0.026
	Bioavailability_Ma	ROC-AUC ↑	0.575 ± 0.031	<b>0.715</b> ± 0.011
	Lipophilicity_AstraZeneca	MAE ↓	0.779 ± 0.033	<b>0.460</b> ± 0.006
	Solubility_AqSolDB	MAE ↓	0.857 ± 0.019	<b>0.725</b> ± 0.011
	HIA_Hou	ROC-AUC ↑	0.948 ± 0.009	<b>0.984*</b> ± 0.004
	Pgp_Broccatelli	ROC-AUC ↑	0.902 ± 0.004	<b>0.931</b> ± 0.003
Distribution	BBB_Martins	ROC-AUC ↑	0.842 ± 0.023	<b>0.908</b> ± 0.010
	PPBR_AZ	MAE ↓	10.012 ± 0.317	<b>7.505</b> ± 0.073
	VDss_Lombardo	Spearman ↑	0.583 ± 0.038	<b>0.662</b> ± 0.013
Metabolism	CYP2C9_Veith	PRC-AUC ↑	0.694 ± 0.004	<b>0.788</b> ± 0.005
	CYP2D6_Veith	PRC-AUC ↑	0.576 ± 0.008	<b>0.704</b> ± 0.003
	CYP3A4_Veith	PRC-AUC ↑	0.797 ± 0.006	<b>0.878</b> ± 0.003
	CYP2C9_Substrate_CarbonMangels	PRC-AUC ↑	0.349 ± 0.033	<b>0.414</b> ± 0.027
	CYP2D6_Substrate_CarbonMangels	PRC-AUC ↑	0.658 ± 0.030	<b>0.739</b> ± 0.024
	CYP3A4_Substrate_CarbonMangels	ROC-AUC ↑	0.562 ± 0.011	<b>0.654</b> ± 0.022
Excretion	Half_Life_Obach	Spearman ↑	0.327 ± 0.083	<b>0.551</b> ± 0.020
	Clearance_Hepatocyte_AZ	Spearman ↑	0.208 ± 0.053	<b>0.495</b> ± 0.030
	Clearance_Microsome_AZ	Spearman ↑	0.283 ± 0.027	<b>0.611</b> ± 0.016
Toxicity	LD50_Zhu	MAE ↓	0.667 ± 0.026	<b>0.541</b> ± 0.015
	hERG	ROC-AUC ↑	0.820 ± 0.024	<b>0.848</b> ± 0.009
	AMES	ROC-AUC ↑	0.767 ± 0.005	<b>0.854</b> ± 0.007
	DILI	ROC-AUC ↑	0.825 ± 0.011	<b>0.920</b> ± 0.012

**Table S2.6: Performance comparison on 11 MoleculeNet benchmark datasets between ChemFM models pre-trained on UniChem and ZINC20.**

Category	Dataset	Task metric	ChemFM-1B (ZINC20)	ChemFM (1B/3B) (UniChem)
Pharmacokinetic	BBBP	ROC-AUC $\uparrow$	0.739	<u>0.745</u> / <b>0.751</b>
Bioactivity	BACE	ROC-AUC $\uparrow$	0.457	<u>0.857</u> / <b>0.869</b>
	HIV	ROC-AUC $\uparrow$	0.721	<u>0.785</u> / <b>0.807</b>
	MUV	PRC-AUC $\uparrow$	0.033	<u>0.122</u> / <b>0.135</b>
	PCBA	PRC-AUC $\uparrow$	0.209	<u>0.322</u> / <b>0.346</b>
Toxicity	Tox21	ROC-AUC $\uparrow$	0.816	<u>0.863</u> / <b>0.869</b>
	SIDER	ROC-AUC $\uparrow$	0.592	<u>0.702</u> / <b>0.709</b>
	ClinTox	ROC-AUC $\uparrow$	<u>0.900</u>	0.899/ <b>0.918</b>
Physicochemical	ESOL	RMSE $\downarrow$	0.541	<u>0.529</u> / <b>0.516</b>
	FreeSolv	RMSE $\downarrow$	<u>0.905</u>	0.906/ <b>0.830</b>
	Lipophilicity	RMSE $\downarrow$	0.626	<u>0.547</u> / <b>0.545</b>
Molecular binding	PDBbind-Full	RMSE $\downarrow$	0.744	<u>0.700</u> / <b>0.697</b>

**Table S2.7: Performance comparison of ChemFM-3B and baseline methods on odor prediction<sup>18</sup> and chromatographic retention time prediction<sup>19</sup>.**

Task	Metric	Model	Performance
Odor prediction	ROC-AUC $\uparrow$	ChemFM-3B	<b>0.902</b>
		MPNN (OpenPOM)	0.887
Retention time	MAE $\downarrow$	ChemFM-3B	<b>39</b>
		(Fingerprints + Descriptors) + Regression NN	57

Odor prediction results are reported as the mean ROC-AUC over 5-fold random cross-validation, using identical splits as in the baseline. Retention time prediction results are reported as mean absolute error (MAE) in second from a single run using the same 75%/25% train/test random split as in the baseline. **Bold** values indicate the best-performing models.

**Table S2.8: Summary of 12 MoleculeNet benchmark datasets evaluated in this work.**

Category	Dataset	Dataset description	Number of molecules	Number of tasks	Task type	Task metric	Split method
Pharmacokinetic	BBBP	predicts binary labels for blood-brain barrier permeability	2,039	1	C	ROC-AUC	Scaffold
Bioactivity	BACE	predicts binary labels for inhibitors of human $\beta$ -secretase 1 (BACE1)	1,513	1	C	ROC-AUC	Scaffold
	HIV	predicts binary labels for inhibitors of HIV replication	41,127	1	C	ROC-AUC	Scaffold
	MUV	predicts binary labels for a series of bioactivities, which is used for validating virtual screening methods	93,087	17	C	PRC-AUC	Random
	PCBA	predicts binary labels for bioactivities across multiple assays from the PubChem	437,929	128	C	PRC-AUC	Random
Toxicity	Tox21	predicts molecular toxicity with binary labels from assays developed by the Tox21 challenge.	7,831	12	C	ROC-AUC	Random
	SIDER	predicts adverse drug reactions with binary labels based on marketed drugs and their recorded side effects	1,427	27	C	ROC-AUC	Random
	ClinTox	predicts drug toxicity with binary labels comparing FDA-approved drugs and those failing clinical trials due to toxicity	1,427	2	C	ROC-AUC	Random
Physicochemical	ESOL	predicts water solubility with continuous labels measured in log mol/L	1,128	1	R	RMSE	Random
	FreeSolv	predicts hydration free energy with continuous labels measured in k/mol for experimental and calculated free energies	642	1	R	RMSE	Random
	Lipophilicity	predicts lipophilicity with continuous labels for experimental logP values, important for absorption and distribution in drugs	4,200	1	R	RMSE	Random
Molecular binding	PDBbind-Full	predicts continuous labels of experimentally measured binding affinities for protein-ligand complexes	9,880	1	R	RMSE	Time

We used the recommended task metrics and split methods, with additional details and dataset sources provided in Wu et al.<sup>3</sup> and Shen et al.<sup>6</sup>. The task type “C” indicates classification task, while “R” indicates regression task. For classification tasks, the evaluation metrics are primarily Receiver Operating Characteristic Area Under the Curve (ROC-AUC) or Precision-Recall Curve Area Under the Curve (PRC-AUC). For regression tasks, the evaluation is based on Root Mean Squared Error (RMSE).

**Table S2.9: Hyperparameters for fine-tuning MoleculeNet benchmark datasets for molecular property prediction.**

Dataset	BBBP	BACE	HIV	MUV	PCBA	Tox21
Optimizer	AdamW					
LR scheduler	Cosine					
Warmup ratio	0.05					
Minimum LR	$0.1 \times \text{LR}$					
Attention dropout	0.2/0.2	0.4/0.2	0.2/0.0	0.0/0.0	0.0/0.2	0.0/0.0
Learning rate (LR)	$4 \times 10^{-4} / 2 \times 10^{-4}$	$8 \times 10^{-4} / 2 \times 10^{-4}$	$8 \times 10^{-5} / 8 \times 10^{-5}$	$2 \times 10^{-4} / 8 \times 10^{-5}$	$1 \times 10^{-4} / 1 \times 10^{-4}$	$4 \times 10^{-4} / 8 \times 10^{-5}$
Epochs	20/20	20/20	10/10	5/5	5/10	5/5
Batch size	16/16	64/16	8/8	16/8	8/8	32/8
Weight decay	0.01/0.05	0.01/0.01	0.1/0.05	0.1/0.01	0.05/0.01	0.05/0.01
LoRA $\alpha$	1.0					
LoRA rank	16/4	32/32	2/1	1/16	16/8	32/32
LoRA dropout	0.2/0.2	0.6/0.6	0.2/0.4	0.4/0.1	0.4/0.1	0.6/0.4
Number of trainable parameters	12.6M/ 6.5M	25.2M/ 51.9M	1.6M/ 1.6M	0.8M/ 26.0M	12.6M/ 13.0M	25.2M/ 51.9M

Dataset	SIDER	ClinTox	ESOL	FreeSolv	Lipophilicity	PDBbind-Full
Optimizer	AdamW					
LR scheduler	Cosine					
Warmup ratio	0.05					
Minimum LR	$0.1 \times \text{LR}$					
Attention dropout	0.2/0.2	0.2/0.0	0.0/0.0	0.2/0.0	0.0/0.0	0.0/0.2
Learning rate (LR)	$4 \times 10^{-4} / 4 \times 10^{-4}$	$1 \times 10^{-4} / 8 \times 10^{-5}$	$8 \times 10^{-4} / 2 \times 10^{-4}$	$8 \times 10^{-4} / 8 \times 10^{-4}$	$2 \times 10^{-4} / 8 \times 10^{-5}$	$4 \times 10^{-4} / 4 \times 10^{-4}$
Epochs	20/10	20/20	50/50	50/50	20/20	5/5
Batch size	16/8	8/8	32/8	8/8	8/8	64/8
Weight decay	0.05/0.05	0.1/0.1	0.05/0.01	0.1/0.01	0.05/0.01	0.01/0.1
LoRA $\alpha$	1.0					
LoRA rank	8/4	32/16	8/32	4/4	8/32	32/2
LoRA dropout	0.6/0.1	0.1/0.2	0.4/0.6	0.4/0.4	0.4/0.4	0.2/0.1
Number of trainable parameters	6.3M/ 6.5M	25.2M/ 26.0M	6.3M/ 51.9M	3.2M/ 6.5M	6.3M/ 51.9M	25.2M/ 3.2M

For each hyperparameter, the value before the slash corresponds to ChemFM-1B, and the value after the slash corresponds to ChemFM-3B.

**Table S2.10: Summary of 22 ADMET benchmark datasets evaluated in this work.**

Category	Dataset	Dataset description	Number of molecules	Task type	Task metric
Absorption	Caco2_Wang	predicts drug permeability, measured in cm/s, using the Caco-2 cell line as an in vitro model to simulate human intestinal tissue permeability	906	R	MAE
	Bioavailability_Ma	predicts oral bioavailability with binary labels, indicating the rate and extent a drug becomes available at its site of action	640	C	ROC-AUC
	Lipophilicity_AstraZeneca	predicts lipophilicity with continuous labels, measured as a log-ratio, indicating a drug's ability to dissolve in lipid environments	4,200	R	MAE
	Solubility_AqSolDB	predicts aqueous solubility with continuous labels, measured in log mol/L, indicating a drug's ability to dissolve in water	9,982	R	MAE
	HIA_Hou	predicts human intestinal absorption (HIA) with binary labels, indicating a drug's ability to be absorbed into the bloodstream	578	C	ROC-AUC
	Pgp_Broccatelli	predicts P-glycoprotein (Pgp) inhibition with binary labels, indicating a drug's potential to alter bioavailability and overcome multidrug resistance	1,212	C	ROC-AUC
Distribution	BBB_Martins	predicts blood-brain barrier permeability with binary labels, indicating a drug's ability to penetrate the barrier to reach the brain	1,915	C	ROC-AUC
	PPBR_AZ	predicts plasma protein binding rate with continuous labels, indicating the percentage of a drug bound to plasma proteins in the blood	1,797	R	MAE
	VDss_Lombardo	predicts the volume of distribution at steady state (VDss) in L/kg, indicating drug concentration in tissues versus blood	1,130	R	Spearman
Metabolism	CYP2C9_Veith	predicts CYP2C9 inhibition with binary labels, indicating the drug's ability to inhibit the CYP2C9 enzyme involved in metabolism	12,092	C	PRC-AUC
	CYP2D6_Veith	predicts CYP2D6 inhibition with binary labels, indicating the drug's potential to inhibit the CYP2D6 enzyme involved in metabolism	13,130	C	PRC-AUC
	CYP3A4_Veith	predicts CYP3A4 inhibition with binary labels, indicating the drug's ability to inhibit the CYP3A4 enzyme involved in metabolism	12,328	C	PRC-AUC
	CYP2C9_Substrate_CarbonMangels	predicts whether a drug is a substrate of the CYP2C9 enzyme with binary labels, indicating its potential to be metabolized	666	C	PRC-AUC
	CYP2D6_Substrate_CarbonMangels	predicts whether a drug is a substrate of the CYP2D6 enzyme with binary labels, indicating its potential to be metabolized	664	C	PRC-AUC
	CYP3A4_Substrate_CarbonMangels	predicts whether a drug is a substrate of the CYP3A4 enzyme with binary labels, indicating its potential to be metabolized	667	C	ROC-AUC
Excretion	Half_Life_Obach	predicts the half-life duration of a drug, measured in hours, indicating the time for its concentration to reduce by half	667	R	Spearman
	Clearance_Hepatocyte_AZ	predicts drug clearance, measured in $\mu\text{L}/(\text{min} \times 10^6 \text{ cells})$ , from hepatocyte experiments, indicating the rate at which the drug is removed from body	1,020	R	Spearman
	Clearance_Microsome_AZ	predicts drug clearance, measured in $\text{mL}/(\text{min} \times \text{g})$ , from microsome experiments, indicating the rate at which the drug is removed from body	1,102	R	Spearman

Continued on next page

**Table S2.10 Continued: Summary of 22 ADMET benchmark datasets evaluated in this work.**

Category	Dataset	Dataset description	Number of molecules	Task type	Task metric
Toxicity	LD50_Zhu	predicts the acute toxicity of a drug, measured as the dose leading to lethal effects in log kg/mol	7,385	R	MAE
	hERG	predicts whether a drug blocks the hERG channel, which is crucial for heart rhythm, potentially leading to adverse effects	648	C	ROC-AUC
	AMES	predicts whether a drug is mutagenic with binary labels, indicating its ability to induce genetic alterations	7,255	C	ROC-AUC
	DILI	predicts whether a drug can cause liver injury with binary labels, indicating its potential for hepatotoxicity	475	C	ROC-AUC

All datasets are single-task and use the default metrics and scaffold split methods, with the data pre-split by the benchmark. Additional details and dataset sources are provided in Huang et al<sup>44</sup>. Task type "C" refers to classification tasks, while "R" refers to regression tasks. For classification tasks, the primary evaluation metric is Receiver Operating Characteristic Area Under the Curve (ROC-AUC) or Precision-Recall Curve Area Under the Curve (PRC-AUC). For regression tasks, the evaluation is based on Mean Absolute Error (MAE) or Spearman's rank correlation coefficient.

**Table S2.11: Hyperparameters for fine-tuning ADMET benchmark datasets for molecular property prediction.**

Dataset	Caco2_ Wang *	Bioavailability_ Ma	Lipophilicity_ AstraZeneca	Solubility_ AqSolDB	HIA_ Hou *	Pgp_ Broccatelli
Optimizer	AdamW					
LR scheduler	Cosine					
Warmup ratio	0.05					
Minimum LR	$0.1 \times \text{LR}$					
Attention dropout	0.2	0.2	0.0	0.0	0.2	0.2
Learning rate (LR)	$2 \times 10^{-4}$	$4 \times 10^{-4}$	$2 \times 10^{-4}$	$4 \times 10^{-4}$	$2 \times 10^{-4}$	$4 \times 10^{-5}$
Epochs	50	20	20	20	50	50
Batch size	16	16	8	8	8	8
Weight decay	0.1	0.05	0.1	0.1	0.01	0.1
LoRA $\alpha$	1.0					
LoRA rank	16	8	8	8	2	1
LoRA dropout	0.4	0.6	0.6	0.2	0.4	0.4
Number of trainable parameters	12.6M	13.0M	13.0M	13.0M	1.6M	1.6M

Dataset	BBB_ Martins	PPBR_ AZ	VDss_ Lombardo	CYP2C9_ Veith	CYP2D6_ Veith	CYP2A4_ Veith
Optimizer	AdamW					
LR scheduler	Cosine					
Warmup ratio	0.05					
Minimum LR	$0.1 \times \text{LR}$					
Attention dropout	0.2	0.0	0.2	0.0	0.2	0.1
Learning rate (LR)	$4 \times 10^{-4}$	$2 \times 10^{-5}$	$2 \times 10^{-4}$	$2 \times 10^{-5}$	$2 \times 10^{-5}$	$7 \times 10^{-5}$
Epochs	5	20	20	20	20	20
Batch size	16	8	8	16	16	8
Weight decay	0.05	0.01	0.05	0.05	0.1	0.1
LoRA $\alpha$	1.0					
LoRA rank	8	16	16	8	16	16
LoRA dropout	0.4	0.1	0.2	0.6	0.6	0.2
Number of trainable parameters	13.0M	26.0M	26.0M	13.0M	26.0M	26.0M

Dataset	CYP2C9_Substrate_ CarbonMangels	CYP2D6_Substrate_ CarbonMangels	CYP3A4_Substrate_ CarbonMangels	Half_Life_ Obach
Optimizer	AdamW			
LR scheduler	Cosine			
Warmup ratio	0.05			
Minimum LR	$0.1 \times \text{LR}$			
Attention dropout	0.2	0.2	0.4	0.2
Learning rate (LR)	$2 \times 10^{-4}$	$2 \times 10^{-4}$	$1 \times 10^{-4}$	$8 \times 10^{-4}$
Epochs	20	20	20	20
Batch size	16	16	16	16
Weight decay	0.05	0.01	0.05	0.05
LoRA $\alpha$	1.0			
LoRA rank	16	1	32	4
LoRA dropout	0.4	0.4	0.4	0.1
Number of trainable parameters	26.0M	1.6M	52.0M	6.5M

continued on next page

**Table S2.11 Continued: Hyperparameters for fine-tuning ADMET benchmark datasets for molecular property prediction.**

Dataset	Clearance_ Hepatocyte_AZ	Clearance_ Hepatocyte_AZ	LD50_ Zhu	hERG	AMES	DILI
Optimizer			AdamW			
LR scheduler			Cosine			
Warmup ratio			0.05			
Minimum LR			$0.1 \times \text{LR}$			
Attention dropout	0.0	0.0	0.0	0.2	0.2	0.3
Learning rate (LR)	$1 \times 10^{-4}$	$2 \times 10^{-4}$	$4 \times 10^{-4}$	$1 \times 10^{-4}$	$4 \times 10^{-5}$	$4 \times 10^{-4}$
Epochs	20	20	20	20	20	10
Batch size	8	8	16	16	16	16
Weight decay	0.1	0.01	0.1	0.1	0.01	0.01
LoRA $\alpha$			1.0			
LoRA rank	1	2	8	8	4	8
LoRA dropout	0.4	0.1	0.4	0.5	0.4	0.3
Number of trainable parameters	1.6M	3.2M	13.0M	13.0M	6.5M	13.0M

The hyperparameters listed are for the ChemFM-3B model, except those marked with a \*, which are for the ChemFM-1B model.

**Table S2.12: Reasons for excluding certain methods from comparison in ADMET benchmark.**

Method	Reason for Exclusion
ZairaChem	Uses both training and validation datasets for training in all runs <sup>1</sup> .
MapLight series	Uses both training and validation datasets for training in all runs <sup>1</sup> .
Random Forest	Uses both training and validation datasets for training in all runs <sup>1</sup> .
SimGCN	Different hyperparameters are used across different datasets, and no hyperparameter search criteria are found or stated <sup>2</sup> .
CFA	Optimized using the test dataset.
DeepMol (AutoML)	Optimized using the test dataset.
BaseBoosting	Uses both training and validation datasets for training in all runs <sup>1</sup> .
XGBoost	Uses both training and validation datasets for training in all runs <sup>1</sup> .
MolMapNet-D	Uses both training and validation datasets for training in all runs <sup>1</sup> .
Basic ML	Uses both training and validation datasets for training in all runs <sup>1</sup> .
RFStacker	Uses both training and validation datasets for training in all runs <sup>1</sup> .
Innoplexus ADME	Uses both training and validation datasets for training in all runs <sup>1</sup> .
Euclia ML model	Uses both training and validation datasets for training in all runs <sup>1</sup> .
ContextPred <sup>3</sup>	Data information leakage during pre-training <sup>4</sup> .
AttrMasking <sup>3</sup>	Data information leakage during pre-training <sup>4</sup> .
Lantern series <sup>5</sup>	Different hyperparameters are used across different datasets, and no hyperparameter search criteria are found or stated <sup>2</sup> .

1. Lack of variations in the training dataset, and since the dataset sizes are normally less than 1000, including the validation dataset can considerably improve performance.

2. Different hyperparameters are used across different datasets. It could be optimized using the test dataset, but this is not confirmed.

3. The information leakage for ContextPred and AttrMasking occurred only on CYP-based benchmarks.

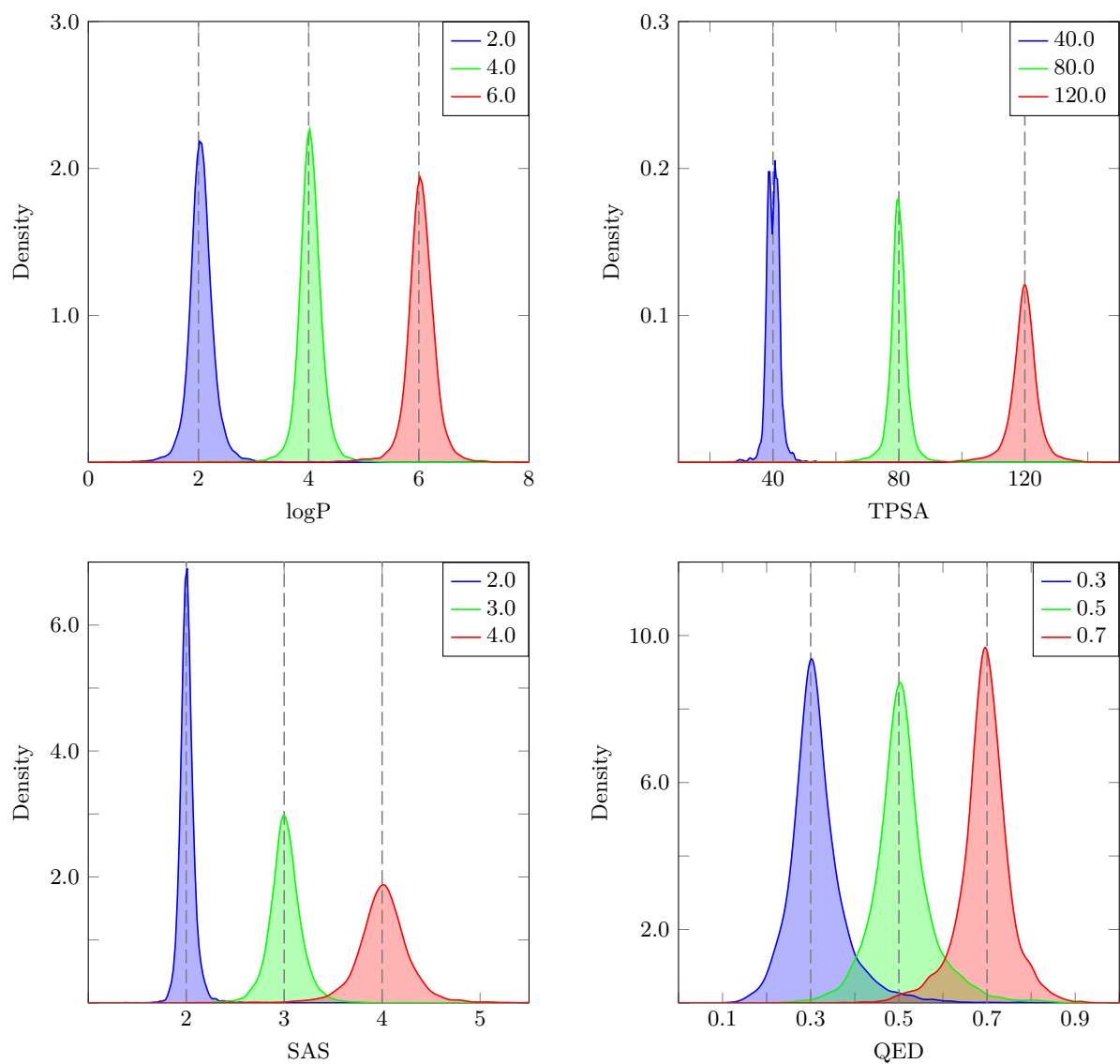
4. The method provider confirmed the data information leakage in this GitHub issue: <https://github.com/mims-harvard/TDC/issues/166>.

5. Lantern series methods are only applied to the BBB\_Martins dataset.

**Table S2.13: Hyperparameters for fine-tuning ChemFM-1B for predicting antibiotic activity and human cell cytotoxicity.**

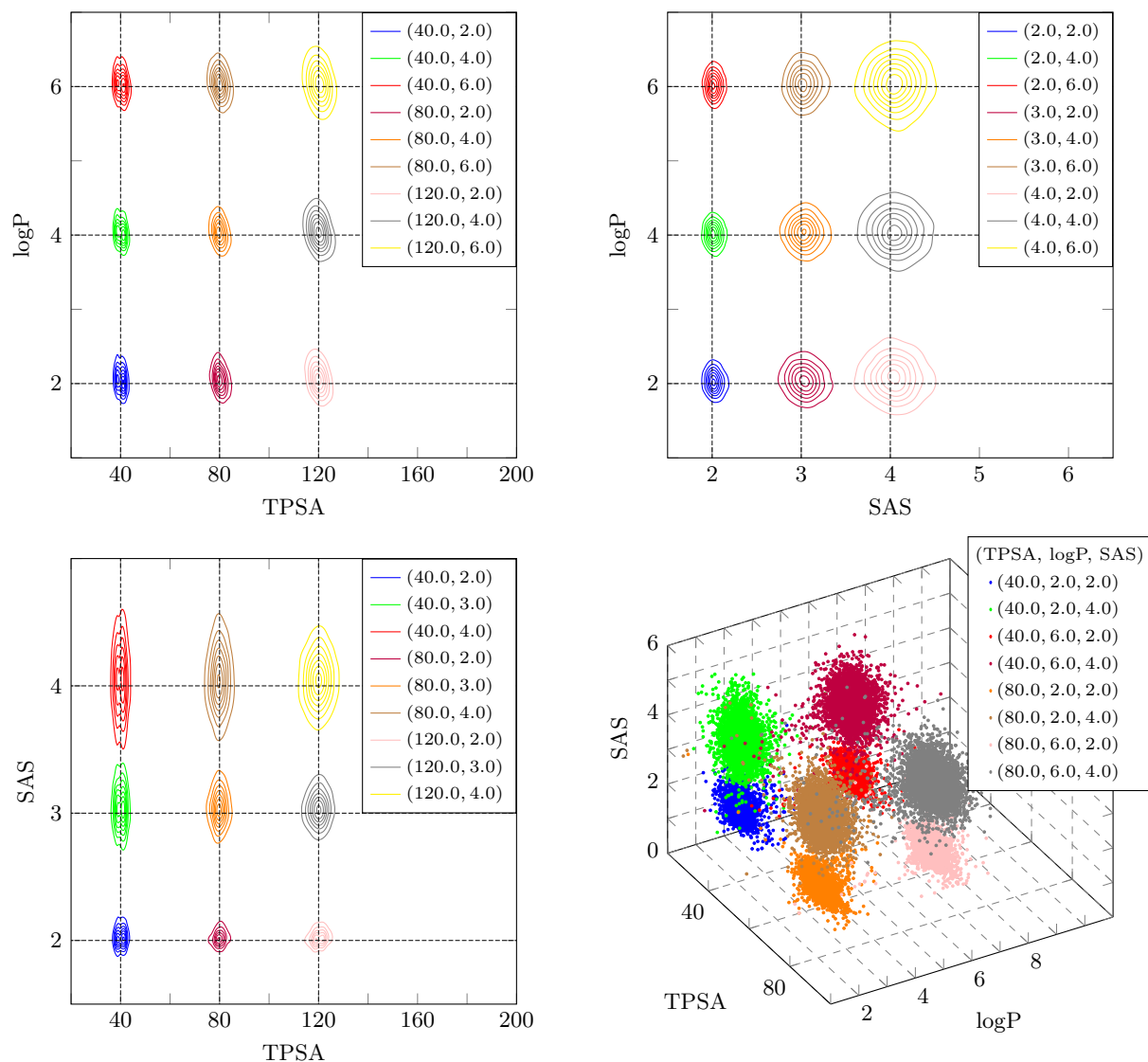
Hyperparameters	Antibiotic activity	HepG2 cytotoxicity	HSkMC cytotoxicity	IMR-90 cytotoxicity
Optimizer			AdamW	
LR scheduler			Cosine	
Warmup ratio			0.05	
Minimum LR			$0.1 \times \text{LR}$	
Attention dropout			0.1	
Learning rate (LR)			$2 \times 10^{-4}$	
Epochs			5	
Batch size			8	
Weight decay			0.05	
LoRA $\alpha$			1.0	
LoRA rank			2	
LoRA dropout			0.1	
Number of trainable parameters			1.6M	

### S3 Supplementary information for conditional molecular generation

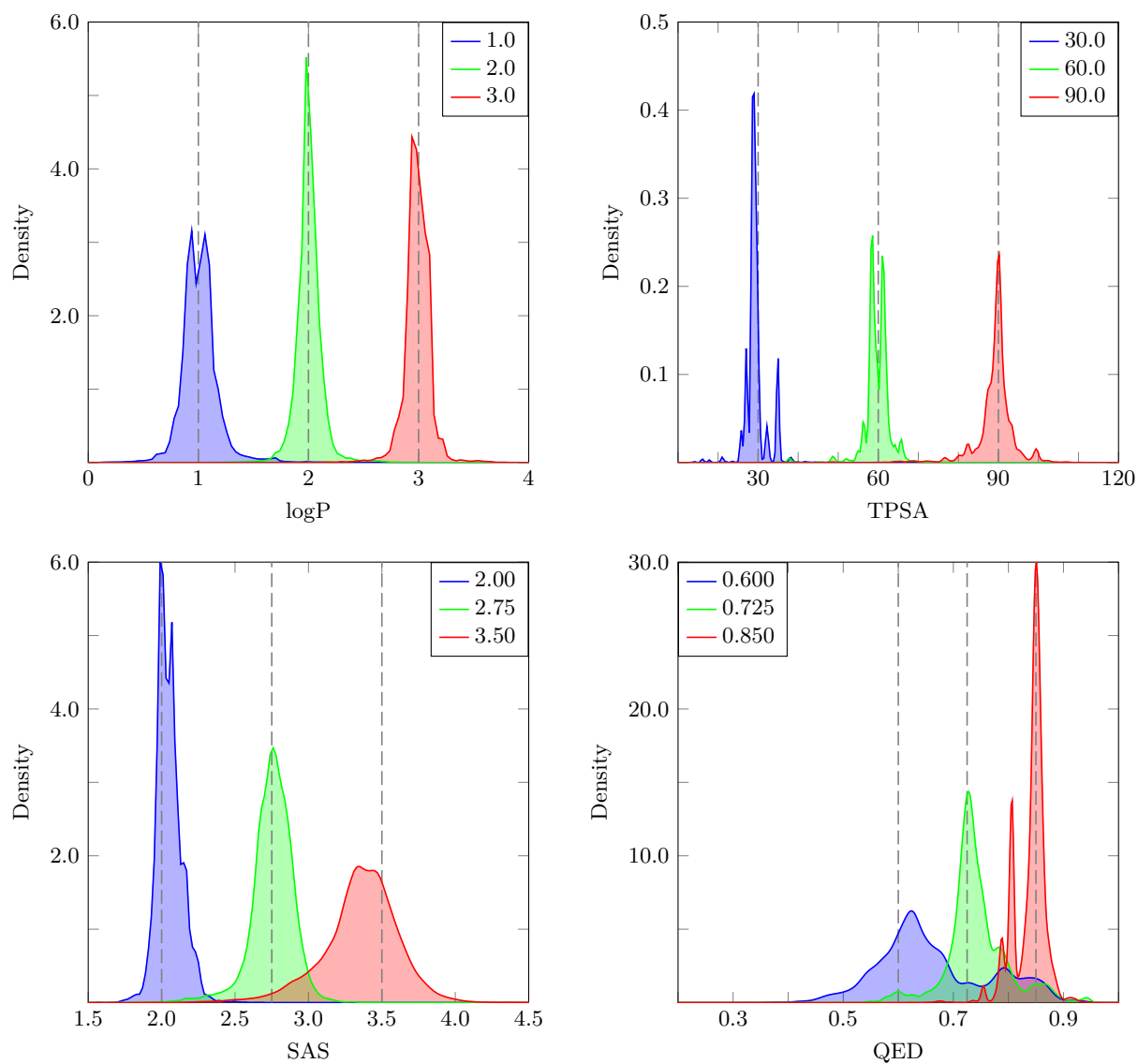


continued on next page

**Fig. S3.1: Distribution of properties for generated molecules on the GuacaMol dataset <sup>21</sup>.** Distribution of molecular properties for generated molecules conditioned on different property combinations. Each subplot corresponds to a specific combination of properties, with the target values indicated in the axis labels.

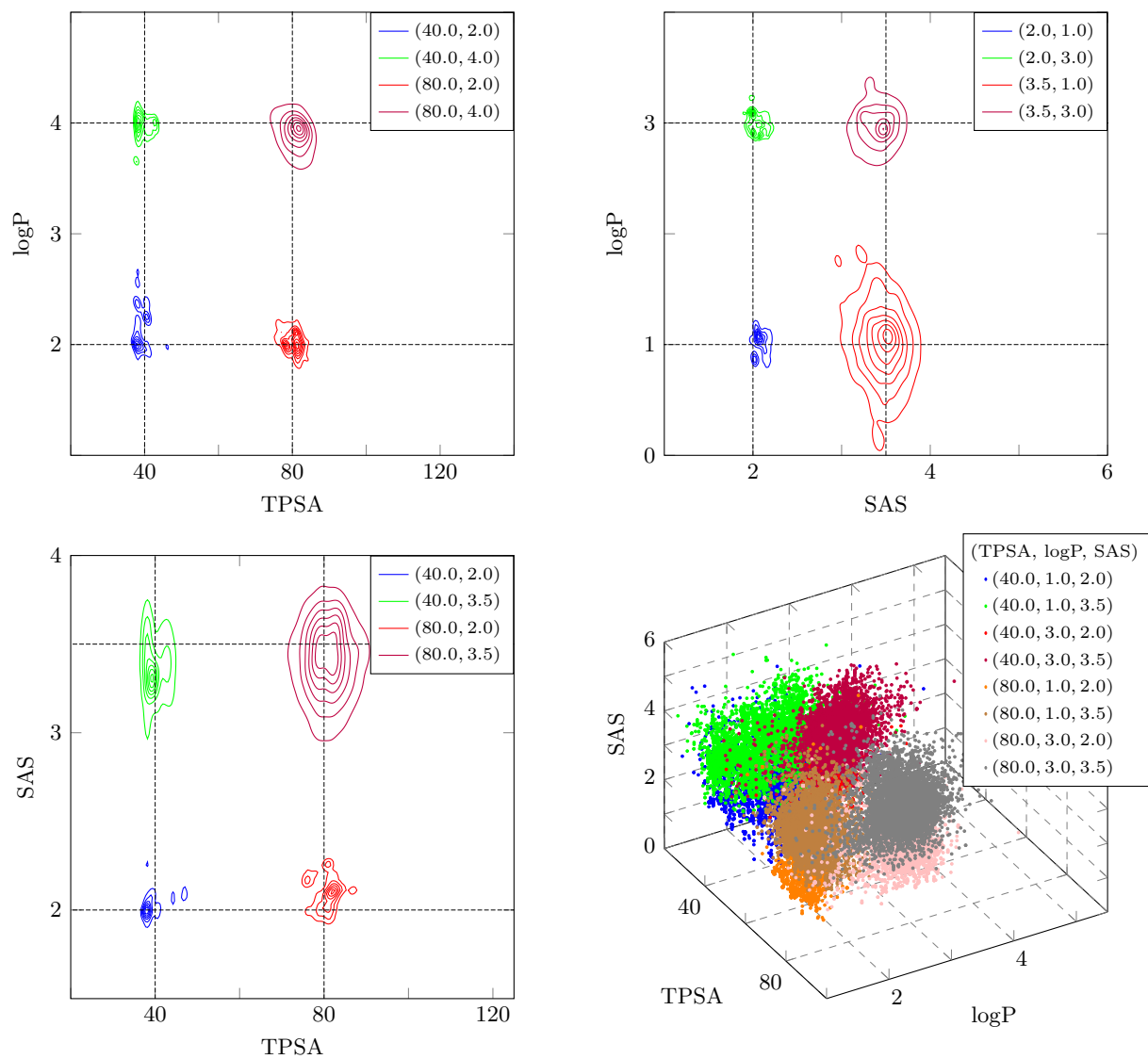


**Fig. S3.2 Continued: Distribution of properties for generated molecules on the GuacaMol dataset<sup>21</sup>.** Distribution of molecular properties for generated molecules conditioned on different property combinations. Each subplot corresponds to a specific combination of properties, with the target values indicated in the axis labels.



continued on next page

**Fig. S3.3: Distribution of properties for generated molecules on the MOSES dataset<sup>24</sup>.** Distribution of molecular properties for generated molecules conditioned on five different scaffold and property combinations. Each subplot corresponds to a specific combination, with the target values indicated in the axis labels.



**Fig. S3.4 Continued: Distribution of properties for generated molecules on the MOSES dataset<sup>24</sup>.** Distribution of molecular properties for generated molecules conditioned on five different scaffold and property combinations. Each subplot corresponds to a specific combination, with the target values indicated in the axis labels.

**Table S3.1: Performance comparison for conditional molecule generation on the GuacaMol<sup>21</sup> dataset.**

Property	Model	Validity $\uparrow$	Uniqueness $\uparrow$	Novelty $\uparrow$	Mean average deviation (MAD) $\downarrow$
logP	cRNN	0.931	0.930	0.926	0.202
	MolGPT	0.971	0.969	0.947	0.230
	ChemFM-3B (w/o pretraining)	0.952	0.952	0.944	0.257
	ChemFM-3B	<b>0.981</b>	<b>0.981</b>	<b>0.966</b>	<b>0.182</b>
TPSA	cRNN	0.923	0.923	0.917	2.629
	MolGPT	0.971	0.969	0.945	3.562
	ChemFM-3B (w/o pretraining)	0.953	0.953	0.942	4.055
	ChemFM-3B	<b>0.979</b>	<b>0.979</b>	<b>0.963</b>	<b>2.466</b>
SAS	cRNN	0.935	0.934	0.922	0.185
	MolGPT	0.978	0.974	0.941	0.133
	ChemFM-3B (w/o pretraining)	0.957	0.956	0.939	0.167
	ChemFM-3B	<b>0.986</b>	<b>0.985</b>	<b>0.957</b>	<b>0.126</b>
QED	cRNN	0.933	0.933	0.925	0.058
	MolGPT	0.974	0.971	0.940	0.056
	MolGPT-3B	0.954	0.954	0.943	0.059
	ChemFM-3B	<b>0.982</b>	<b>0.982</b>	<b>0.963</b>	<b>0.045</b>
SAS + logP	cRNN	0.932	0.931	0.926	0.209/0.201
	MolGPT	0.972	0.963	0.947	0.147/0.253
	ChemFM-3B (w/o pretraining)	0.955	0.953	0.943	0.163/0.271
	ChemFM-3B	<b>0.980</b>	<b>0.975</b>	<b>0.960</b>	<b>0.137/0.195</b>
SAS + TPSA	cRNN	0.923	0.921	0.917	0.218/3.032
	MolGPT	0.971	0.960	0.944	0.155/3.785
	ChemFM-3B (w/o pretraining)	0.959	0.954	0.944	0.172/4.148
	ChemFM-3B	<b>0.980</b>	<b>0.971</b>	<b>0.956</b>	<b>0.138/2.659</b>
TPSA + logP	cRNN	0.903	0.901	0.898	2.845/0.212
	MolGPT	0.964	0.958	0.947	3.715/0.243
	ChemFM-3B (w/o pretraining)	0.951	0.949	0.943	4.036/0.255
	ChemFM-3B	<b>0.973</b>	<b>0.970</b>	<b>0.962</b>	<b>2.415/0.184</b>
TPSA + logP + SAS	cRNN	0.937	0.928	0.922	3.232/0.226/0.261
	MolGPT	0.972	0.942	0.931	3.797/0.268/0.180
	ChemFM-3B (w/o pretraining)	0.965	0.941	0.932	3.626/0.256/0.186
	ChemFM-3B	<b>0.975</b>	<b>0.946</b>	<b>0.936</b>	<b>2.289/0.191/0.166</b>

Molecules were generated according to desired property values, with performance compared across ChemFM-3B, ChemFM-3B without pre-training (both using a single model), and cRNN<sup>22</sup> and MolGPT<sup>23</sup> (each requiring 8 separate models). Metrics include validity, uniqueness, novelty, and mean absolute deviation (MAD) between the conditioned and actual properties of the generated molecules. **Bold** values indicate the best performance for each metric. It should be noted that validity, uniqueness, and novelty are computed against the total number of generated molecules, rather than only the valid ones, to more accurately reflect model performance (Methods).

**Table S3.2: Performance comparison on standard benchmarks for conditional molecule generation on the MOSES dataset<sup>24</sup>.**

Property	Scaffold	Model	Generation count	Valid molecules $\uparrow$	Unique molecules $\uparrow$	Novel molecules $\uparrow$	Same scaffold molecules $\uparrow$	MAD $\downarrow$
logP	(a)	MolGPT	30,000	29,679	7,132	7,132	7,132	0.123
		ChemFM	30,000	29,538	10,117	10,117	10,117	0.087
	(b)	MolGPT	30,000	26,880	15,669	15,669	15,669	0.126
		ChemFM	30,000	27,606	20,011	20,011	19,519	0.097
	(c)	MolGPT	30,000	28,536	10,537	10,537	10,531	0.127
		ChemFM	30,000	29,077	11,398	11,398	11,392	0.077
	(d)	MolGPT	30,000	29,643	8,886	8,886	8,885	0.136
		ChemFM	30,000	29,832	11,559	11,559	11,558	0.093
	(e)	MolGPT	30,000	29,666	2,334	2,334	2,328	0.111
		ChemFM	30,000	29,629	3,521	3,521	3,521	0.084
Total	MolGPT	150,000	144,404	44,558	44,558	44,545	0.125	
ChemFM	150,000	<b>145,682</b>	<b>56,606</b>	<b>56,606</b>	<b>56,107</b>	<b>0.087</b>		
SAS	(a)	MolGPT	30,000	29,253	13,663	13,663	13,663	0.122
		ChemFM	30,000	26,899	16,922	16,922	16,921	0.106
	(b)	MolGPT	30,000	26,719	11,846	11,846	11,843	0.140
		ChemFM	30,000	27,088	15,801	15,801	15,337	0.129
	(c)	MolGPT	30,000	25,269	10,803	10,803	10,791	0.143
		ChemFM	30,000	27,685	14,535	14,534	14,513	0.140
	(d)	MolGPT	30,000	29,011	12,253	12,253	12,253	0.138
		ChemFM	30,000	29,264	15,710	15,710	15,691	0.113
	(e)	MolGPT	30,000	28,540	4,087	4,087	4,065	0.102
		ChemFM	30,000	29,644	5,195	5,195	5,192	0.093
Total	MolGPT	150,000	138,792	52,652	52,652	52,615	0.129	
ChemFM	150,000	<b>140,580</b>	<b>68,163</b>	<b>68,162</b>	<b>67,654</b>	<b>0.123</b>		
TPSA	(a)	MolGPT	30,000	29,809	8,134	8,134	8,132	1.887
		ChemFM	30,000	29,726	10,777	10,777	10,777	1.581
	(b)	MolGPT	30,000	26,509	14,519	14,519	14,518	2.948
		ChemFM	30,000	27,666	17,262	17,259	16,698	2.291
	(c)	MolGPT	30,000	29,522	10,240	10,240	10,240	2.002
		ChemFM	30,000	28,617	10,742	10,742	10,734	1.392
	(d)	MolGPT	30,000	29,655	10,264	10,264	10,263	2.702
		ChemFM	30,000	29,783	11,546	11,546	11,542	2.056
	(e)	MolGPT	30,000	28,716	1,873	1,873	1,867	3.785
		ChemFM	30,000	29,575	3,838	3,838	3,835	3.240
Total	MolGPT	150,000	144,211	45,030	45,030	45,020	2.651	
ChemFM	150,000	<b>145,367</b>	<b>54,165</b>	<b>54,162</b>	<b>53,586</b>	<b>2.114</b>		
QED	(a)	MolGPT	30,000	29,618	12,931	12,931	12,931	0.043
		ChemFM	30,000	29,332	16,723	16,723	16,723	0.044
	(b)	MolGPT	30,000	25,516	15,659	15,659	15,646	0.040
		ChemFM	30,000	27,421	21,597	21,597	21,002	0.036
	(c)	MolGPT	30,000	29,055	13,108	13,108	13,105	0.034
		ChemFM	30,000	28,914	14,643	14,643	14,633	0.033
	(d)	MolGPT	30,000	29,812	12,978	12,978	12,978	0.064
		ChemFM	30,000	29,797	14,368	14,368	14,362	0.043
	(e)	MolGPT	30,000	27,457	2,918	2,918	2,909	0.072
		ChemFM	30,000	29,330	5,127	5,127	5,116	0.095
Total	MolGPT	150,000	141,458	57,594	57,594	57,569	0.051	
ChemFM	150,000	<b>144,794</b>	<b>72,458</b>	<b>72,458</b>	<b>71,836</b>	<b>0.050</b>		

Continued on next page

**Table S3.2 Continued: Performance comparison on standard benchmarks for conditional molecular generation on the MOSES dataset.**

Property	Scaffold	Model	Generation count	Valid molecules $\uparrow$	Unique molecules $\uparrow$	Novel molecules $\uparrow$	Same scaffold molecules $\uparrow$	MAD $\downarrow$
TPSA + logP	(a)	MolGPT	40,000	38,440	8,919	8,919	8,919	3.125/0.238
		ChemFM	40,000	38,845	13,536	13,536	13,536	3.268/0.206
	(b)	MolGPT	40,000	33,172	16,609	16,609	16,578	3.912/0.173
		ChemFM	40,000	36,388	20,940	20,940	20,251	2.767/0.124
	(c)	MolGPT	40,000	38,190	13,901	13,901	13,881	3.492/0.125
		ChemFM	40,000	37,873	13,828	13,828	13,816	2.779/0.110
	(d)	MolGPT	40,000	36,010	11,104	11,104	11,102	4.749/0.165
		ChemFM	40,000	39,085	11,394	11,394	11,391	2.520/0.179
	(e)	MolGPT	40,000	36,122	3,966	3,965	3,942	3.651/0.230
		ChemFM	40,000	35,762	6,950	6,950	6,904	5.104/0.172
Total	MolGPT	200,000	181,934	54,499	54,498	54,422	3.771/0.186	
	ChemFM	200,000	<b>187,953</b>	<b>66,648</b>	<b>66,648</b>	<b>65,898</b>	<b>3.266/0.159</b>	
SAS + logP	(a)	MolGPT	40,000	36,497	11,866	11,866	11,859	0.154/0.238
		ChemFM	40,000	32,753	15,401	15,401	15,401	0.124/0.159
	(b)	MolGPT	40,000	30,249	10,962	10,962	10,958	0.156/0.164
		ChemFM	40,000	35,512	15,937	15,937	15,565	0.166/0.133
	(c)	MolGPT	40,000	36,784	12,745	12,745	12,734	0.157/0.157
		ChemFM	40,000	35,456	14,073	14,073	14,001	0.164/0.141
	(d)	MolGPT	40,000	37,290	11,616	11,616	11,613	0.136/0.192
		ChemFM	40,000	38,417	13,995	13,995	13,966	0.113/0.204
	(e)	MolGPT	40,000	39,243	4,361	4,361	4,262	0.125/0.167
		ChemFM	40,000	38,666	7,059	7,059	7,050	0.120/0.190
Total	MolGPT	200,000	180,063	51,550	51,550	51,426	0.145/0.184	
	ChemFM	200,000	<b>180,804</b>	<b>66,465</b>	<b>66,465</b>	<b>65,983</b>	<b>0.137/0.166</b>	
TPSA + SAS	(a)	MolGPT	40,000	37,901	16,217	16,217	16,214	3.417/0.175
		ChemFM	40,000	34,935	16,994	16,994	16,994	2.990/0.131
	(b)	MolGPT	40,000	32,343	12,808	12,808	12,744	4.633/0.189
		ChemFM	40,000	34,274	15,585	15,585	15,117	4.136/0.181
	(c)	MolGPT	40,000	31,731	13,522	13,522	13,512	3.960/0.181
		ChemFM	40,000	36,445	15,526	15,525	15,482	3.640/0.181
	(d)	MolGPT	40,000	35,947	13,793	13,793	13,793	4.066/0.165
		ChemFM	40,000	38,614	16,790	16,790	16,740	3.343/0.133
	(e)	MolGPT	40,000	39,196	5,170	5,170	5,120	3.292/0.149
		ChemFM	40,000	38,941	6,010	6,010	6,005	3.443/0.119
Total	MolGPT	200,000	177,118	61,510	61,510	61,383	3.840/0.171	
	ChemFM	200,000	<b>183,209</b>	<b>70,905</b>	<b>70,904</b>	<b>70,338</b>	<b>3.504/0.148</b>	
TPSA + logP + SAS	(a)	MolGPT	80,000	66,217	14,699	14,699	14,699	5.125/0.539/0.348
		ChemFM	80,000	56,565	19,454	19,454	19,454	4.334/0.378/0.220
	(b)	MolGPT	80,000	63,655	17,279	17,279	17,257	5.096/0.225/0.227
		ChemFM	80,000	60,128	27,610	27,610	26,850	4.207/0.187/0.258
	(c)	MolGPT	80,000	59,363	16,086	16,086	16,068	5.650/0.256/0.220
		ChemFM	80,000	64,752	22,422	22,422	22,315	4.681/0.285/0.220
	(d)	MolGPT	80,000	50,119	12,081	12,081	12,080	5.145/0.332/0.243
		ChemFM	80,000	67,716	18,970	18,970	18,895	4.337/0.483/0.197
	(e)	MolGPT	80,000	74,433	7,228	7,228	7,111	5.749/0.384/0.233
		ChemFM	80,000	73,882	8,858	8,858	8,787	6.079/0.304/0.196
Total	MolGPT	400,000	313,787	67,373	67,373	67,215	5.370/0.352/0.255	
	ChemFM	400,000	<b>323,043</b>	<b>97,314</b>	<b>97,314</b>	<b>96,301</b>	<b>4.780/0.329/0.217</b>	

Generations were conditioned on five test scaffolds and various property combinations, comparing the performance of ChemFM-3B, which uses a single model, to MolGPT<sup>23</sup>, which uses 8 separate models. Generation performance is reported as the number of valid, unique, and novel molecules. A valid molecule is defined as 1) syntactically correct and 2) having a Tanimoto similarity greater than 0.8 to the conditioned scaffold. For scaffold matching, we report the number of molecules that match the conditioned scaffold and the mean absolute deviation (MAD) between the conditioned and actual property values. The generations are based on five test scaffolds: (a) O=C(Cc1cccc1)NCc1cccc1, (b) c1cnc2[nH]ccc2c1, (c) c1ccc(-c2ccnc2)cc1, (d) c1ccc(-n2nc3cccc32)cc1, and (e) O=C(c1cc[nH]c1)N1CCN(c2cccc2)CC1.

## S4 Supplementary information for reaction prediction tasks

**Table S4.1: Comprehensive comparison of synthesis and retro-synthesis reaction prediction performance across standard USPTO benchmarks.**

Task category	Dataset	Model	Top-1	Top-3	Top-5
Synthesis	USPTO-MIT	WLDN <sup>25</sup>	74.0	-	-
		MT <sup>26</sup>	88.6	93.5	94.2
		AT <sup>27</sup>	<u>90.4</u>	-	<u>96.5</u>
		MEGAN <sup>28</sup>	86.3	92.4	94.0
		R-SMILES <sup>29</sup>	90.0	<u>95.6</u>	96.4
		ChemFM	<b>90.5</b>	<b>95.7</b>	<b>96.6</b>
Retro-synthesis	USPTO-50K	AT <sup>27</sup>	53.5	-	81.0
		MEGAN <sup>28</sup>	48.1	70.7	78.4
		GraphRetro <sup>30</sup>	53.7	68.3	72.2
		GLN <sup>31</sup>	52.5	69.0	75.6
		RetroXpert <sup>32</sup>	50.4	61.1	62.3
		GTA <sup>33</sup>	51.1	-	-
		RetroPrime <sup>34</sup>	51.4	70.8	74.0
		LocalRetro <sup>35</sup>	53.4	77.5	85.9
		Chemformer <sup>7</sup>	54.3	62.3	63.0
		Retroformer <sup>36</sup>	53.2	71.1	76.6
		Graph2Edits <sup>37</sup>	55.1	77.3	83.4
		G <sup>2</sup> Retro <sup>38</sup>	54.1	74.1	81.2
		R-SMILES <sup>29</sup>	<u>56.0</u>	<u>79.0</u>	<u>86.1</u>
		ChemFM	58.0	<b>80.0</b>	<b>86.3</b>
		ChemFM*	<b>59.7</b>	79.2	84.2
	USPTO-MIT	LocalRetro <sup>35</sup>	54.1	73.7	79.4
		RetroTRAE <sup>39</sup>	58.3	-	-
		R-SMILES <sup>29</sup>	<u>60.3</u>	<u>77.9</u>	<u>82.8</u>
		ChemFM	61.6	<b>78.7</b>	<b>83.0</b>
		ChemFM*	<b>62.4</b>	78.5	82.5
	USPTO-Full	AT <sup>27</sup>	46.2	-	-
		MEGAN <sup>28</sup>	33.6	-	-
		GLN <sup>31</sup>	39.3	-	-
		RetroPrime <sup>34</sup>	44.1	59.1	62.8
		LocalRetro <sup>35</sup>	39.1	53.3	58.4
		RetroXpert <sup>32</sup>	49.4	63.6	67.6
		GTA <sup>33</sup>	46.6	-	-
		Substructure <sup>40</sup>	48.2	-	-
		R-SMILES <sup>29</sup>	<u>48.9</u>	<u>66.6</u>	<u>72.0</u>
		ChemFM	<b>51.7</b>	<b>68.0</b>	<b>72.5</b>

This table compares the top-1, top-3, and top-5 accuracies (in percentages) of our ChemFM model against various models from the literature on the USPTO-MIT, USPTO-50K, and USPTO-Full datasets. **Bold** values indicate the best performance, and underlined values represent the best performance other than ChemFM. A hyphen “-” signifies that the result was not reported in the corresponding paper. All results from other methods are obtained directly from their respective publications, except for the R-SMILES results, which were replicated using publicly available models. ChemFM\* denotes ChemFM with further pre-training, which achieve better top-1 results but show a decrease in top-3 and 5 performance.

**Table S4.2: Summary of USPTO benchmark datasets evaluated in this work.**

Dataset	Number of reactions (training)	Number of reactions (validation)	Number of reactions (testing)	Dataset description
USPTO-Full <sup>45</sup>	768,630	96,071	96,023	Largest dataset, containing reactions extracted from patents (1976-2016), covering a wide range of chemical reaction types.
USPTO-MIT <sup>25</sup>	411,685	30,182	40,265	Refined subset of USPTO-Full, with duplicates and erroneous reactions removed. Commonly used for forward reaction prediction benchmarks.
USPTO-50K <sup>46</sup>	40,003	5,001	5,007	Smaller curated dataset of 50K reactions across 10 types, commonly used for benchmarking retro-synthesis prediction tasks.

**Table S4.3: Hyperparameters for fine-tuning ChemFM-3B for reaction prediction tasks.**

Hyperparameters	USPTO-MIT (Synthesis)	USPTO-50K (Retro-synthesis)	USPTO-MIT (Retro-synthesis)	USPTO-Full (Retro-synthesis)
Optimizer			AdamW	
LR scheduler			Cosine	
Warmup ratio			0.05	
Minimum LR			$0.1 \times \text{LR}$	
Attention dropout	0.1	0.1	0.1	0.1
Learning rate (LR)	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$
Epochs	10	10	10	10
Batch size	64	64	64	64
Weight decay	0.01	0.01	0.01	0.01
LoRA $\alpha$	-	1.0	1.0	1.0
LoRA rank	-	32	64	64
LoRA dropout	-	0.1	0.1	0.1
Number of trainable parameters	3.0B	54M	106M	106M

For synthesis prediction using USPTO-MIT, we perform full-parameter fine-tuning, while all retro-synthesis experiments use LoRA. The number of epochs refers to training on the augmented dataset. Due to computational resource constraints, we do not perform hyperparameter searching. Two or three sets of hyperparameters are chosen based on prior experience, and we report the best one. Results were found to be quite stable within a reasonable hyperparameter range. A hyphen “-” indicates that the hyperparameter is not applicable.

## Supplementary References

1. Chambers, J. *et al.* Unichem: a unified chemical structure cross-referencing and identifier tracking system. *Journal of cheminformatics* **5**, 3 (2013).
2. Irwin, J. J. *et al.* Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling* **60**, 6065–6073 (2020).
3. Wu, Z. *et al.* Moleculenet: a benchmark for molecular machine learning. *Chemical science* **9**, 513–530 (2018).
4. Honda, S., Shi, S. & Ueda, H. R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738* (2019).
5. Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **59**, 3370–3388 (2019).
6. Shen, W. X. *et al.* Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nature Machine Intelligence* **3**, 334–343 (2021).
7. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **3**, 015022 (2022).
8. Xiong, Z. *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry* **63**, 8749–8760 (2019).
9. Hu, W. *et al.* Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*.
10. Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712* (2022).
11. Stärk, H. *et al.* 3d infomax improves gnns for molecular property prediction. Paper presented at the 39th International Conference on Machine Learning, Baltimore, USA, 17–23 July 2022.
12. Xia, J. *et al.* Mole-bert: Rethinking pre-training graph neural networks for molecules. Paper presented at the 11th International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
13. Liu, S. *et al.* Pre-training molecular graph representation with 3d geometry. Paper presented at the 10th International Conference on Learning Representations, Virtual Conference, 25–29 April 2022.
14. Liu, S., Du, W., Ma, Z.-M., Guo, H. & Tang, J. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. Paper presented at the 40th International Conference on Machine Learning, Honolulu, USA, 23–29 July 2023.
15. Swanson, K. *et al.* ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics* **40**, btae416 (2024).
16. Huang, K. *et al.* DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* **36**, 5545–5547 (2020).
17. Hu, W. *et al.* Strategies for pre-training graph neural networks. Paper presented at the 10th International Conference on Learning Representations, Virtual Conference, 26 April–1 May 2022.
18. Lee, B. K. *et al.* A principal odor map unifies diverse tasks in olfactory perception. *Science* **381**, 999–1006 (2023).

19. Domingo-Almenara, X. *et al.* The metlin small molecule dataset for machine learning-based retention time prediction. *Nature communications* **10**, 5811 (2019).
20. Wong, F. *et al.* Discovery of a structural class of antibiotics with explainable deep learning. *Nature* **626**, 177–185 (2024).
21. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* **59**, 1096–1108 (2019).
22. Kotsias, P.-C. *et al.* Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature Machine Intelligence* **2**, 254–265 (2020).
23. Bagal, V., Aggarwal, R., Vinod, P. & Priyakumar, U. D. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling* **62**, 2064–2076 (2021).
24. Polykovskiy, D. *et al.* Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology* **11**, 565644 (2020).
25. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network.
26. Schwaller, P. *et al.* Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **5**, 1572–1583 (2019).
27. Tetko, I. V., Karpov, P., Van Deursen, R. & Godin, G. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications* **11**, 5575 (2020).
28. Sacha, M. *et al.* Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling* **61**, 3273–3284 (2021).
29. Zhong, Z. *et al.* Root-aligned smiles: a tight representation for chemical reaction prediction. *Chemical Science* **13**, 9023–9034 (2022).
30. Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. Paper presented at the 35th Conference on Neural Information Processing Systems, Virtual Conference, 6–14 December 2021.
31. Dai, H., Li, C., Coley, C., Dai, B. & Song, L. Retrosynthesis prediction with conditional graph logic network. Paper presented at the 35th Conference on Neural Information Processing Systems, Vancouver, Canada, 8–14 December 2019.
32. Yan, C. *et al.* Retroxpert: Decompose retrosynthesis prediction like a chemist. Paper presented at the 34th Annual Conference on Neural Information Processing Systems, Virtual Conference, 6–12 December 2020.
33. Seo, S.-W. *et al.* Gta: Graph truncated attention for retrosynthesis. Paper presented at the 35th AAAI Conference on Artificial Intelligence, Virtual Conference, 2–9 Feb 2021.
34. Wang, X. *et al.* Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal* **420**, 129845 (2021).
35. Chen, S. & Jung, Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au* **1**, 1612–1620 (2021).
36. Wan, Y., Hsieh, C.-Y., Liao, B. & Zhang, S. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. Paper presented at the 39th International Conference on Machine Learning, Baltimore, USA, 17–23 July 2022.

37. Zhong, W., Yang, Z. & Chen, C. Y.-C. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications* **14**, 3009 (2023).
38. Chen, Z., Ayinde, O. R., Fuchs, J. R., Sun, H. & Ning, X. G 2 retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry* **6**, 102 (2023).
39. Ucak, U. V., Ashyrmamatov, I., Ko, J. & Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nature communications* **13**, 1186 (2022).
40. Fang, L., Li, J., Zhao, M., Tan, L. & Lou, J.-G. Single-step retrosynthesis prediction by leveraging commonly preserved substructures. *Nature Communications* **14**, 2446 (2023).
41. Landrum, G. *et al.* Rdkit: Open-source cheminformatics. <https://www.rdkit.org/>.
42. Thomas, M., O’Boyle, N. M., Bender, A. & De Graaf, C. Molscore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. *Journal of Cheminformatics* **16** (2024).
43. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742–754 (2010).
44. Huang, K. *et al.* Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. Paper presented at the 35th Conference on Neural Information Processing Systems, Virtual Conference, 6–14 December 2021.
45. Lowe, D. Chemical reactions from us patents (1976-sep2016) (2017). URL <https://doi.org/10.6084/m9.figshare.5104873.v1>.
46. Dai, H., Li, C., Coley, C., Dai, B. & Song, L. Retrosynthesis prediction with conditional graph logic network.