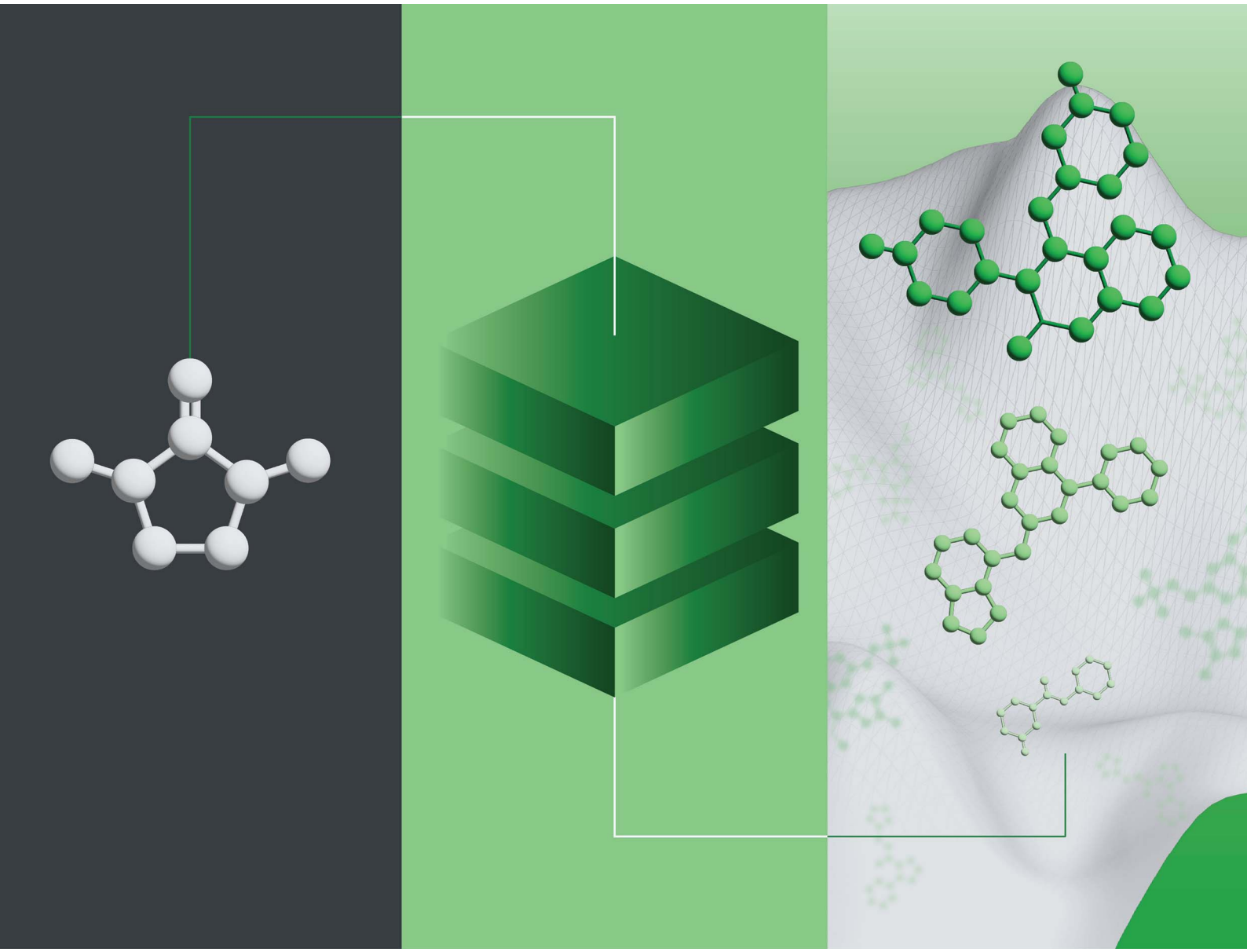
Volume 4
Number 10
October 2025
Pages 2643-3054

Digital Discovery

rsc.li/digitaldiscovery



ISSN 2635-098X

ROYAL SOCIETY OF CHEMISTRY

PAPER
Jerret Ross, Payel Das *et al.*
GP-MoLFormer: a foundation model for molecular generation

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 2684

GP-MoLFormer: a foundation model for molecular generation

Jerret Ross, * Brian Belgodere, Samuel C. Hoffman, Vijil Chenthamarakshan, Jiri Navratil, Youssef Mroueh and Payel Das *

Transformer-based models trained on large and general purpose datasets consisting of molecular strings have recently emerged as a powerful tool for successfully modeling various structure–property relations. Inspired by this success, we extend the paradigm of training chemical language transformers on large-scale chemical datasets to generative tasks in this work. Specifically, we propose GP-MoLFormer, an autoregressive molecular string generator that is trained on more than 1.1b (billion) chemical SMILES. GP-MoLFormer uses a 46.8m parameter transformer decoder model with linear attention and rotary positional encodings as the base architecture. GP-MoLFormer's utility is evaluated and compared with that of existing baselines on three different tasks: *de novo* generation, scaffold-constrained molecular decoration, and unconstrained property-guided optimization. While the first two are handled with no additional training, we propose a parameter-efficient fine-tuning method for the last task, which uses property-ordered molecular pairs as input. We call this new approach pair-tuning. Our results show GP-MoLFormer performs better or comparable with baselines across all three tasks, while producing molecules with higher diversity demonstrating its general utility for a variety of molecular generation tasks. We further report strong memorization of training data in GP-MoLFormer generations, which has so far remained unexplored for chemical language models. Our analyses reveal that training data memorization and novelty in generations are impacted by the quality and scale of the training data; duplication bias in training data can enhance memorization at the cost of lowering novelty. We further establish a scaling law relating inference compute and novelty in generations, and show that the proposed model excels at yielding molecules containing unique scaffolds while generating at $\approx 10^6$ to 10^9 scale.

Received 25th March 2025
Accepted 13th August 2025

DOI: 10.1039/d5dd00122f

rsc.li/digitaldiscovery

Main

Identifying molecules with desirable properties from the vast landscape of possibilities is daunting. As the search space is enormous and high-throughput screening of molecules is costly and time-consuming, this requires a thorough understanding of the chemical data manifold. Recently, similar to what has been experienced in computer vision and natural language processing, deep generative models have made great strides in modeling molecular distributions and sampling new molecules from them in *de novo* or targeted manners. Among those efforts, a significant fraction use string-based representations of molecules as the input; thus, techniques explored in language modeling, such as causal and masked language modeling, are becoming widely used in building molecular deep neural models.

Interestingly, much of the recent performance gains for natural language models have come from training at scale—in terms of the number of parameters and the number of training samples.[1–3] It is reported that larger language models that can memorize training data show improved generalization.[4–6] Furthermore, data that is seen during training many times is memorized more and de-duplication of training data plays a big role in preventing such memorization.[4,5,7,8]

However, less work has been done in understanding the impact of training data scale and its memorization on the performance of generative models of molecules. Specifically, it remains under-explored to what extent a causal large language model of molecules, trained on large-scale (>100m) training data, memorizes its training data and demonstrates such memorization in its generations. In chemical language modeling tasks, molecules in training data originate from publicly available databases such as ZINC[9] and PubChem.[10] It is known that certain molecules as well as certain molecular features are over-represented in those databases[11] but how such training bias is perpetuated by generative chemical language models remains relatively unknown.

An additional dimension of scaling in traditional large language models that has been investigated recently is

IBM Research, Yorktown Heights, NY 10598, USA. E-mail: rossja@us.ibm.com; daspa@us.ibm.com

inference-time compute scaling.[12] It has been shown that with increasing inference compute, performance across multiple tasks can be increased for the same model, as it allows better coverage of the search space. On the other hand, the effect of inference scaling by increasing number of generations is under-explored for molecular generative models.

To bridge these gaps, in this work, we present a family of generative pre-trained molecular foundation models for the unconstrained and targeted generation of novel molecules. These decoder-only models are based on the recently published Molecular Language transFormer (MoLFormer) architecture.[13] We refer to these Generative Pre-trained models as GP-MoLFormer. The base transformer architecture of our GP-MoLFormer consists of ≈47m parameters and uses an efficient linear attention mechanism together with rotary positional encodings—analogous to MoLFormer[13] but using decoder instead of encoder blocks (Fig. 1A). The model is then trained with a causal language modeling objective on a large corpus of 0.65–1.1 billion canonicalized SMILES strings of small molecules from publicly available chemical databases.

We evaluate GP-MoLFormer on an unconditional *de novo* generation task as well as to two targeted molecular design tasks: scaffold-constrained molecular decoration and unconstrained property-guided optimization. For scaffold decoration, we exploit GP-MoLFormer's causal language modeling ability and establish GP-MoLFormer's ability to handle the task without undergoing any task-specific tuning. For the optimization task, we provide a prompt-tuning or soft prompt-learning algorithm that learns from partial orderings of molecules. We name this method pair-tuning (Fig. 1B). Results show that pair-tuning on GP-MoLFormer provides on par or better performance in three different property optimization tasks, namely (i) drug-likeness optimization, (ii) penalized log P optimization, and (iii) optimization of dopamine type 2 receptor binding activity.

We further extensively evaluate quality of GP-MoLFormer-generated molecules, in the light of the training data scale and the bias present in the training data. Experiments reveal significant memorization in *de novo* generations affecting novelty therein. We further analyze how representational bias encoded in the public chemical databases is perpetuated by a generative chemical language model and is reflected in its generation quality. To our knowledge, this is the first report on effect of training data memorization in a generative pre-trained chemical language model. Further, we investigate the effect of inference compute as another scaling dimension by increasing the number of generated samples and establish a inference scaling law relating number of generations with novelty in them. Experiments demonstrate that novelty in *de novo* generations by GP-MoLFormer drops when number of generated samples reaches a scale of ≈1b. Nevertheless, GP-MoLFormer is able to generate novel, unique, diverse, and valid molecules even when the generation pool reaches a size of 10b, while showing consistent memorization of training data.

Our main contributions are:

- We provide a pre-trained, autoregressive, transformer-based SMILES decoder, GP-MoLFormer-Uniq.
- We report the beneficial effects of training this class of models on up to 1.1 billion SMILES, compared to models trained on smaller datasets, by demonstrating higher scaffold-level uniqueness and diversity in GP-MoLFormer generations even when performed at scale, which is attributed to its training data scale and diversity.

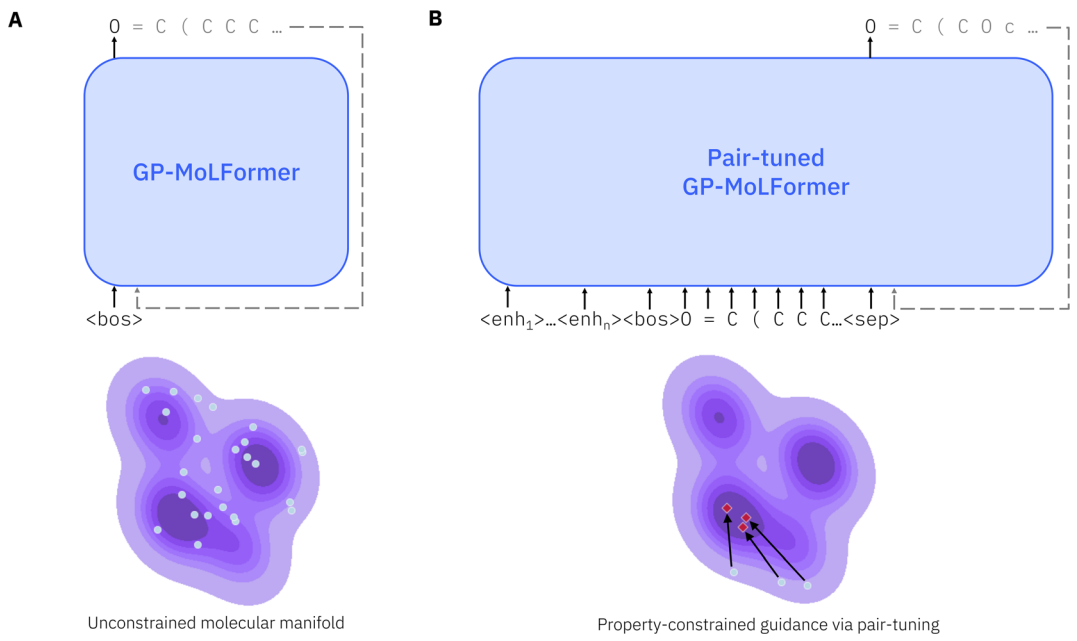


Fig. 1 GP-MoLFormer—a generative pre-trained molecular foundation model. (A) Unconditional generation using GP-MoLFormer. SMILES representations are generated autoregressively and randomly along the learned manifold (purple area). (B) During pair-tuning, a prompt vector is learned, which translates a given molecular representation (light blue dots) to an optimized region of the manifold (red diamonds).

• We provide a parameter-efficient finetuning method, which utilizes property-ranked molecule pairs as input, for property-guided molecule generation and show its effectiveness on three different tasks.

• We further study how training data duplication bias (and therefore training size) affects *de novo* generation and reveal that more duplication significantly reduces novelty in generations.

• We also report a scaling behavior relating inference compute and novelty that follows exponential decay, while showing that GP-MoLFormer can generate a notable fraction of novel SMILES, even when number of generations reaches 10b.

Results and discussion

GP-MoLFormer uses a causal modeling objective of predicting the next token given the context history of prior tokens in the input SMILES strings. For details of model architecture and training, see the Methods section. After pre-training, we quantitatively assess the performance of GP-MoLFormer on *de novo* molecule generation and scaffold-constrained molecule decoration tasks, before applying a novel prompt-tuning algorithm for molecular optimization.

De novo generation of molecules

The task under consideration is to generate random, (syntactically) valid SMILES strings by sampling from the generative model. As downstream optimization may rely on these randomly generated molecules as starting points, the generated distribution must contain novel, diverse, and unique molecules. We also require that the generated distribution resemble the training distribution closely. Here we compare a GP-MoLFormer model trained on 650m unique SMILES, that is a de-duplicated subset of the 1.1b SMILES extracted from ZINC and PubChem, as described in Ross *et al.* (2022).[13] This model variant is referred to as GP-MoLFormer-Uniq, hereafter. Note: all SMILES are converted to canonical form for all tasks (therefore novelty and uniqueness are equivalent to molecular comparisons, *i.e.*, the molecule is not present in the training or generation sets, respectively).

We compare GP-MoLFormer-Uniq with different baseline models, such as the character-level recurrent neural network (CharRNN),[14] SMILES variational autoencoder (VAE),[14] junction tree VAE (JT-VAE),[15] latent inceptionism on molecules (LIMO),[16] and MolGen-7b.[17] Except MolGen-7b, all baselines were trained and tested on datasets from MOSES,[14] whose origin is the ZINC Clean Leads dataset.[9] The size of that training set is 1.6m. MolGen-7b was trained on 100m filtered molecules from ZINC-15 (ref. 9) as detailed in Irwin *et al.* (2022).[18] LIMO and MolGen-7b are trained using an alternative molecular string representation, SELFIES,[19] that guarantees 100% validity of generated molecules. We, in contrast, train GP-MoLFormer-Uniq on SMILES as recent work shows that training a generative language model on SELFIES may hurt model's exploratory ability.[20] All baseline performances are reported on their corresponding test set consisting of 175k molecules (if the original test set was larger, this is a randomly selected subset).

First, we note that GP-MoLFormer-Uniq (and GP-MoLFormer) exhibits excellent validity and uniqueness at standard generation size (30/10k). See SI Table S1 for comparison with the baseline models. At the same time, we argue that these metrics are insufficient to measure generation at scale. Furthermore, as we show later, novelty is dependent on training set size in addition to generation size so models trained on different size datasets are not directly comparable. See the Scaling results section below for further discussion.

Standard metrics for evaluating model-generated molecules are reported in Table 1 for a generation set of 30k molecules. When compared to baselines, GP-MoLFormer-Uniq is equally performant in generating molecules that share high cosine similarity with the corresponding reference molecules at the fragment (Frag) level, consistent with low Fréchet ChemNet

Table 1 Comparison of 30k generations with a held-out test set of size 175k. "MOSES metrics" columns refer to the typical set of generation performance metrics (see Polykovskiy *et al.*[14] for details) computed with respect to the following sets: baseline performances for CharRNN, VAE, and JT-VAE are taken from MOSES.[14] LIMO is reproduced using their random generation model trained on the MOSES data. These models all use the default MOSES test split for reference. The MolGen-7b baseline uses the pre-trained model from Hugging Face[a] with multinomial sampling ($T = 1.0$) and is tested on a random 175k subset of the original test data. GP-MoLFormer-Uniq is tested with respect to a held-out 175k set from its training data. "MoLFormer-based metrics" columns refer to analogous metrics computed using MoLFormer[13] embedding distances instead of Tanimoto similarity. DNN is the average Euclidean distance from generated molecules to the nearest molecule from the test set. IntDiv2 is the average pairwise Euclidean distance between generated molecules. FMD is the Fréchet distance between the MoLFormer embedding distributions. Bold values indicate the best model for a given metric

	MOSES metrics					MoLFormer-based metrics		
	Frag↑	Scaf↑	SNN↑	IntDiv↑	FCD↓	DNN↓	IntDiv2↑	FMD↓
CharRNN[14]	**0.9998**	0.9242	0.6015	0.8562	0.0732	5.735	13.03	0.1515
VAE[14]	0.9984	**0.9386**	**0.6257**	0.8558	0.0990	**5.549**	13.09	0.2531
JT-VAE[15]	0.9965	0.8964	0.5477	0.8551	0.3954	6.312	12.97	1.700
LIMO[16]	0.6989	0.0079	0.2464	**0.9039**	26.78	11.41	13.08	162.0
MolGen-7b[17]	**0.9999**	0.6538	0.5138	0.8617	**0.0435**	6.788	12.58	**0.1237**
GP-MoLFormer-Uniq	**0.9998**	0.7383	0.5045	0.8655	0.0591	6.970	**13.10**	0.1844

[a] https://huggingface.co/zjunlp/MolGen-7b.

Distance (FCD).[21] The scaffold cosine similarity (Scaf) and similarity to the nearest neighbor in the test set (SNN) of GP-MoLFormer-Uniq is comparable to that of baselines for 30k generations. At the same time, GP-MoLFormer-Uniq generates molecules with high internal diversity (IntDiv), *i.e.*, average pairwise dissimilarity. All these metrics are computed using the MOSES[14] framework (we limit our scope to MOSES in this study, although we note that myriad other benchmarks are available for evaluating generative molecular models[22–24]).

We further report analogous metrics computed using MoLFormer[13] embeddings as the chemical features and estimate distances using those embeddings as a measure of similarity (under column MoLFormer-based metrics; see Table 1 caption for details). The trends observed on these metrics further support the fact that GP-MoLFormer-Uniq generates a molecular distribution that is close to the training in terms of fragment and scaffold composition as well as projections to MoLFormer space, while exhibiting high diversity, when compared to baselines.

We also calculated the pairwise Tanimoto similarity between novel and unique generations and molecules from the corresponding 175k sample test set using molecular fingerprints as features. We then report both the average similarity per generated molecule and the maximum similarity per generated molecule over the test set. These results are presented in Table 2. GP-MoLFormer-Uniq is slightly lower than MolGen-7b in both average mean and average maximum similarity, indicating generations are slightly more dissimilar with respect to its test set. LIMO results are much lower than both of these, though, as we see in Table 1, the outputs of this model do not match its test set well, so this is to be expected. Also, LIMO is more suited for property-optimized generation, and therefore we do not include LIMO in the further comparison for *de novo* generations.

Interestingly, Table 1 shows higher internal diversity within the generated molecules for GP-MoLFormer-Uniq. We further extend the internal diversity analysis to the scaffolds present in generated molecules. As shown in Table 3, GP-MoLFormer-Uniq generated scaffolds show more internal diversity than baselines like CharRNN, VAE, and MolGen-7b, suggesting that training on data at scale promotes diversity within generations. As will be shown in later sections, GP-MoLFormer-Uniq also excels at yielding higher number of unique scaffolds, while performing generation at scale. These results reinforce the greater utility of the proposed model, especially when tested at scale. Further analysis of scaffold novelty can be found in SI Fig. S1.

Fig. 2 shows the property distributions of the different test sets, as well as of molecules generated using GP-MoLFormer-Uniq. The generated distribution shows very good reproduction of the corresponding test distribution. Furthermore, while GP-MoLFormer-Uniq's performance is estimated on a held-out test set that is of similar size, we found this test set to be more diverse in terms of number of unique scaffolds present within the set (126k compared to 124k in the ZINC-15 subset and 77k in the MOSES set) and by comparing different property distributions with that of the other baselines. More analyses on how these statistics change with training data variations and generated pool size can be found later (see Discussions).

We also examine the domain adaptation of GP-MoLFormer *via* down-stream finetuning on a set of 36.7m drug-like molecules from PubChem.[10] In Fig. 2, we show results of this finetuned model, referred as GP-MoLFormer-Druglike. The finetuning set contains molecules with QED >0.6.[25] Results show that, the generated molecules undergo a distribution shift in properties as expected. For example, the QED distribution is shifted toward right compared to GP-MoLFormer-Uniq. We also provide examples of *de novo* generated molecules in the SI Fig. S3.

It is important to note that all this analysis is intended simply to show that our model is able to balance reproducing the training distribution, while generating novel, diverse, and unique outputs, both at SMILES and at scaffold level. While there are diminishing returns to be had trying to more closely match the training distribution, we show that increasing the size and diversity of the training data is one way to produce better quality molecules.

Table 2 Average Tanimoto similarity between 30k randomly generated molecules and all respective test molecules (mean) or most similar test molecule (max). Bold values indicate the highest similarity to test molecules

Model	Mean	max
LIMO	0.0905	0.2474
MolGen-7b	**0.1385**	**0.5138**
GP-MoLFormer-Uniq	0.1354	0.4533

Table 3 Average internal scaffold diversity for a random subset of 100k unique scaffolds generated by each model

	GP-MoLFormer-Uniq	MolGen-7b	CharRNN	VAE
IntDiv	**0.855**	0.842	0.840	0.847

Scaffold-constrained decoration

We further subject GP-MoLFormer to the task of scaffold-constrained generation. For this experiment, we utilize the GP-MoLFormer trained on 1.1b SMILES. We first take the five unique scaffolds from the dopamine receptor D_2 (DRD2) active binder dataset validation split.[26] These scaffolds contain between two to four attachment points. We perform a preprocessing step for each unique scaffold, generating every possible randomized SMILES representation of that scaffold. Then we sort the resulting candidates according to the distance of the "*" characters, also known as the attachment point, to the end of the string. For multiple *s, we sum the distances. As an example, C1(=O)N(CCN1*)* would score 2 + 0 = 2 while C1N(*)C(=O)N(*)C1 would score 12 + 3 = 15. If multiple representations are equivalently optimal, we save all of them. During the generation step, we provide the candidates produced in this preprocessing step as input to GP-MoLFormer.

Next, the task is to generate multiple possible candidates for the first attachment point given an input scaffold. First, we

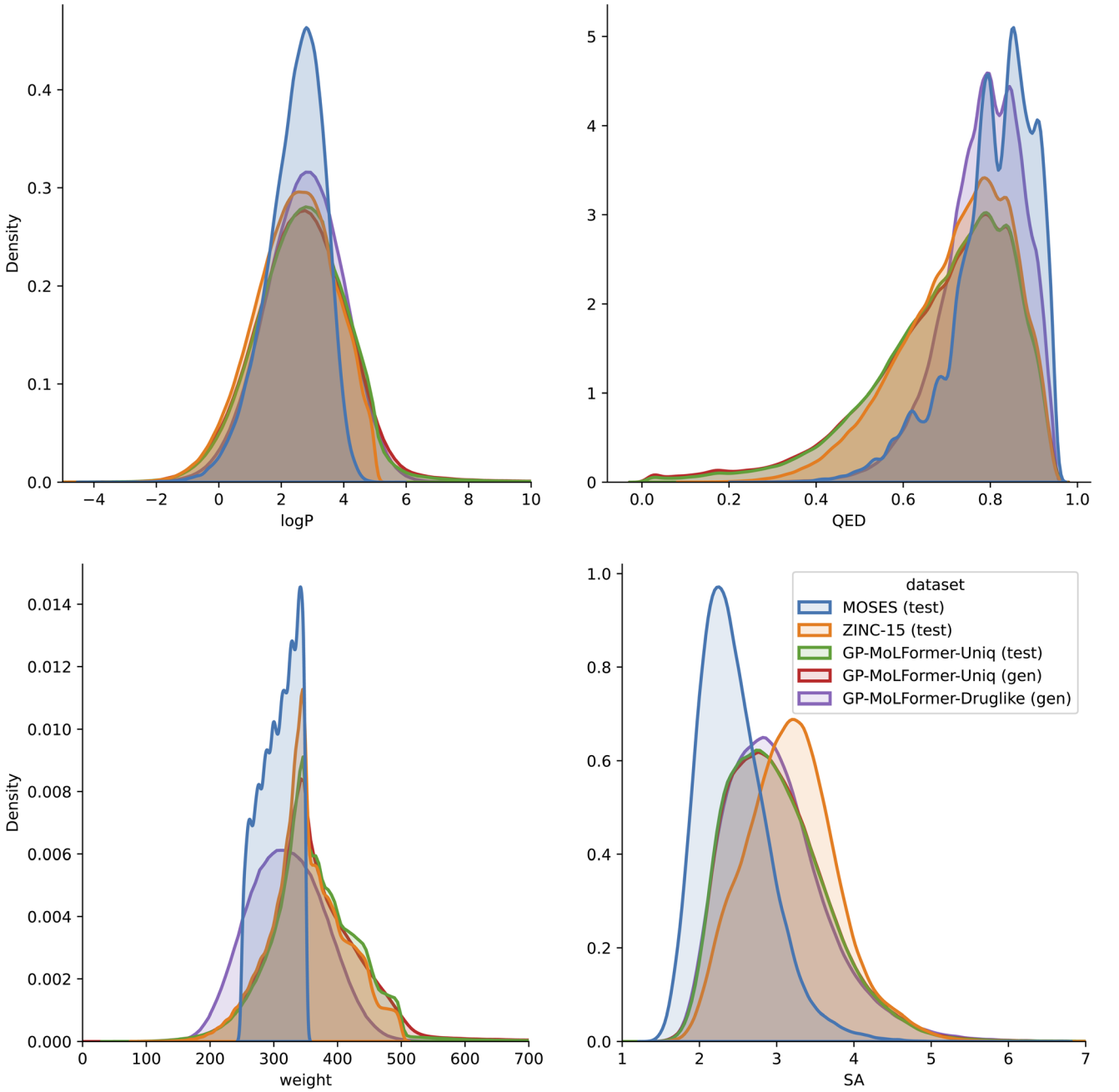


Fig. 2 Property distributions of different test datasets—MOSES, ZINC-15 (MolGen-7b), and GP-MoLFormer-Uniq (ours)—along with generated samples from GP-MoLFormer-Uniq and GP-MoLFormer-Druglike. Clockwise from top left: octanol–water partition coefficient, drug-likeness, synthetic accessibility, molecular weight. Our test distributions are consistently wider (more diverse) than the other baselines. Furthermore, the generated distribution matches the corresponding test distribution almost exactly. In comparison to GP-MoLFormer-Uniq, a density shift toward higher QED values with GP-MoLFormer-Druglike can be observed, as expected.

collect all valid candidates from that generation. Then, we again generate multiple possible candidates for the recently extended scaffolds. This process is repeated until all the attachment points are decorated then we collect all valid molecules generated.

We compare the performance of GP-MoLFormer in terms of generating DRD2 active molecules that will pass the DRD2 binding classifier ($p > 0.5$). For baselines of comparison, we consider our own random generations from GP-MoLFormer, as well as an earlier scaffold-conditioned generation model[26] that was specifically trained for scaffold decoration tasks and was then used to decorate the same scaffolds under investigation here with fragments from ChEMBL. In contrast to this baseline model, GP-MoLFormer has not seen scaffold-constrained generation task during pre-training, nor is it specifically fine-tuned for this purpose. Table 4 shows that GP-MoLFormer

Table 4 Scaffold-constrained generation. Predicted active hits is the percentage of generated molecules that pass the DRD2 binding classifier. Baseline performance is taken from Arús-Pous *et al.* (2020).[26] Bold value indicates the best performing model

	Predicted active hits (%)
Scaffold decorator[26]	3.64
De novo GP-MoLFormer	0.83
Scaffold-conditioned GP-MoLFormer	**4.58**

generates more DRD2 active hits compared to a random baseline of *de novo* generation, as well as a generative model trained on this specific task. Examples of scaffold-decorated molecules using GP-MoLFormer are shown in the SI Fig. S4.

Unconstrained single property optimization

Given GP-MoLFormer demonstrates desirable performance in both novel molecule generation and scaffold-constrained molecular decoration, it makes sense to extend GP-MoLFormer to downstream task settings, where the goal is to generate molecules with a desired property. In light of current LLM adaptation efforts, one obvious path is model tuning (or "fine-tuning"), where all model parameters are tuned during adaptation. This approach often is highly data-hungry. As an alternative, prompt-tuning or "soft prompt" learning has been proposed, which includes an additional n tuneable tokens for each downstream task, which is prepended to the input text.[27] This soft prompt is then trained end-to-end on a labeled dataset, whereas the pre-trained LLM remains frozen. This method has been demonstrated to close the gap in model tuning, even when combined with a smaller LLM[27] and is lower cost compared to full fine-tuning.

We exploit prompt-tuning to introduce a novel means for enabling GP-MoLFormer to tackle property-specific molecular optimization tasks, where the goal is to generate molecules with a specific property value above a pre-defined threshold. Below, we describe the pair-tuning framework and then show that pair-tuning performs well on a set of three tasks. We evaluate pair-tuning using GP-MoLFormer on three property optimization benchmarks, namely drug-likeness (QED) maximization, penalized $\log P$ maximization, and activity maximization for DRD2. The first two properties, QED and penalized $\log P$, are important considerations in drug discovery, and these task shows the ability of a model to optimize salient aspects of a molecule, even if maximization of these properties by themselves is of low utility.[28] The goal of the third task is to increase the binding affinity of a compound to a ligand, in this case, the dopamine D_2 receptor.

Pair-tuning framework. Following a "text-to-text" approach, we formalize the task of generating a (property-)optimal molecule as follows: Given a molecule a, translate it to another molecule b with a more optimal property value where a, b come from domain Ω. This conditional generation task is $P_\theta(b|a)$, where θ is the parametrization of the generative language model. This task is handled *via* learning soft prompts, *i.e.*, prompt-tuning, which is a parameter-efficient task adaptation method for a frozen language model.[27] Specifically, we add a small number of task-specific parameters ϕ_T, such that the conditional task becomes $P_\theta(b|\phi_T, a)$ and is trained through maximizing the probability likelihood of b. Only ϕ_T is updated during gradient backpropagation. This procedure is explained in Algorithm 1.

In this formulation, we do not need absolute property values of the molecules, rather only ordered pairs of molecules are needed. This is to mimic the scenario of many drug and material development tasks, in which two molecules are compared with each other to guide molecular optimization and prioritization, especially for tasks with limited available data. For example, Matched Molecular Pair (MMP) analysis allows the rapid estimation of property differences.[29,30] However, MMP analysis is limited to comparing close molecular derivatives and common molecular derivations, and it can fail to model important chemical contexts. The present formulation of optimizing molecules is free from such constraints and only aims to learn task-specific soft prompts to generate more optimal molecules given a seed molecule.

Algorithm 1 Pair-tuning training

Require: Pre-trained GP-MoLFormer; Number of fine-tuning epochs, m; number of enhancement tokens, n; Pair-Tuning dataset $D = \{(a,b)|b > a\}$ for $a, b \in \Omega$ where $>$ is an order relation defined for a certain property T

1: Append $n+1$ new tokens $\phi_T = \langle enh_1\rangle\langle enh_2\rangle \ldots \langle enh_n\rangle, \langle sep\rangle$ to GP-MoLFormer vocabulary
2: **for** m epochs **do**
3: Prepare training prompts as $\langle enh_1\rangle\langle enh_2\rangle \ldots \langle enh_n\rangle\langle bos\rangle\langle a_1\rangle\langle a_2\rangle \ldots \langle a_i\rangle\langle sep\rangle$ where $a = \langle a_1\rangle\langle a_2\rangle \ldots \langle a_i\rangle$, i.e., tokenized training molecules
4: Compute the cross-entropy (CE) loss conditioned on enhancement tokens ϕ_T and molecule a with the auto-regressive CE loss with target the molecule $b = \langle b_1\rangle\langle b_2\rangle \ldots \langle b_n\rangle\langle eos\rangle$
5: Compute the gradient of the auto-regressive CE loss with respect to enhancement tokens ϕ_T
6: Update enhancement tokens via gradient descent optimizer
7: **end for**

Penalized log *P* optimization. Table 5 shows results of pair-tuning on GP-MoLFormer, as well as of the baselines, in terms of the generated molecules with high penalized log *P*. Penalized log *P* is calculated as log *P* − SA − max(max$_{\text{rings}}$(size) − 6, 0), *i.e.*, log *P* penalized by SA and maximum ring size, if larger than 6. We report pair-tuning performances as a function of two different k, where k is the number of targeted generation attempts per molecule. For $k = 125$, using a test set containing 800 molecules gives a total number of generated molecules of 100k, which is the same used for the baselines. The baselines under consideration are JT-VAE,[15] GCPN,[31] MolDQN,[28] MARS,[32] GraphDF,[33] and LIMO.[16] Penalized log *P* can be artificially inflated simply by generating molecules with increased length, specifically by adding alkyl carbons.[16,32] Many works, *e.g.*, GCPN, MolDQN, and LIMO, avoid this by reporting top property scores given length constraints, *e.g.*, limiting the length up to the maximum molecule size of the ZINC250k dataset.[34] MARS, on the other hand, does not consider such a length constraint. We also report the top 3 scores for pair-tuning with a length constraint (length < 38), added post generation, in Table 5 as the value within parentheses. Compared to the strongest length-constrained baselines, pair-tuning generates molecules with comparably high values. When the length constraint is not considered, pair-tuning still generates molecules with higher but reasonable penalized log *P* values. Note that pair-tuning does not require feedback or evaluation on generations from an additional reward model or a property predictor, nor is the generative model updated during the tuning. We also report top 3 scores for 1m generations ($k = 1000$), which requires less than an hour to generate. Although all the baselines produce molecules with 100% validity due to their methods utilizing SELFIES or graphs, our method's validity is still very high (around 95%) and overall this is negligible compared to the ease of generating additional molecules. Altogether, Table 5 shows that the proposed method can generate molecules with even higher penalized log *P* values, both with and without a length constraint.

QED optimization. As with penalized log *P*, we show results for QED optimization in Table 5 (also see SI Fig. S5 for generated molecules) compared with the same baselines. Again, pair-tuning performances are reported for two different values of k, showing comparable performances with respect to baselines. SI Tables S2 and S3 further demonstrate that pair-tuning with GP-MoLFormer produces higher scoring molecules that also share high diversity as well as high closeness to training distribution, compared to baselines, which is consistent with results in Table 3. To further establish the usefulness of pair-tuning, we also compare it with full fine-tuning of GP-MoLFormer on the high-QED molecules from the same training set. Results show that full fine-tuning of the base model triggers collapse in terms of unique generations. Details are available in SI Table S4.

DRD2 activity optimization. DRD2 activity optimization results are reported in Table 6. Activity scores are calculated using the trained predictor from Olivecrona *et al.* (2017).[36] Average activity scores of the input seed molecules are also shown. Different baseline performances are reported using different test seed molecules, and we are interested in comparing the activity improvement of an experiment with the set of test seed molecules. For pair-tuning, performance reported considers $k = 20$ generations per seed molecule and we use the top 1 for each, similar to the baselines, Mol-CycleGAN[37]—a graph-based generation method that uses a CycleGAN optimization scheme—and Gargoyles,[35] which uses Monte Carlo Tree Search in fragment space to optimize molecular graphs based on an evaluation function. Results show that pair-

Table 6 Performance on the unconstrained DRD2 activity optimization task with respect to the initial value. Baseline performances are reported from Erikawa *et al.* (2023).[35] Bold value indicates the best performing model

	Predicted activity score	Average seed score
Mol-CycleGAN	0.381	0.179
Gargoyles	0.782	0.122
Pair-tuning	**0.844**	0.007

Table 5 Performance on unconstrained penalized log *P* and QED optimization. Pair-tuning is performed using frozen GP-MoLFormer. Baseline performances are taken from Zhou *et al.* (2019)[28] and Eckmann *et al.* (2022)[16] and are reported on 100k generations as per LIMO.[16] For GP-MoLFormer, we set k, the number of targeted generation attempts per molecule, to 125—given a test set of size 800 this results in 100k total generations. Values in parentheses are after *post hoc* length filtering. Bold values indicate the highest property values found (both length-constrained and unlimited)

	Penalized log *P*				QED			
	1 st	2nd	3rd	Validity	1st	2nd	3rd	Validity
JT-VAE	5.30	4.93	4.49	100%	0.925	0.911	0.910	100%
MARS	**45.0**	**44.3**	**43.8**	100%	**0.948**	**0.948**	**0.948**	100%
GRAPHDF	13.7	13.2	13.2	100%	**0.948**	**0.948**	**0.948**	100%
LIMO on z	6.52	6.38	5.59	100%	0.910	0.909	0.892	100%
LIMO	10.5	9.69	9.60	100%	0.947	0.946	0.945	100%
GCPN	7.98	7.85	7.80	100%	**0.948**	0.947	0.946	100%
MolDQN-bootstrap	**11.84**	**11.84**	**11.82**	100%	**0.948**	0.944	0.943	100%
Pair-tuning ($k = 125$)	13.18 (7.12)	12.24 (6.61)	11.51 (6.40)	94.7%	**0.948**	0.947	0.947	94.7%
Pair-tuning ($k = 1000$)	19.59 (9.35)	15.51(8.93)	15.27 (8.64)	94.5%	**0.948**	**0.948**	**0.948**	94.5%

tuning generates molecules with the highest activity improvement with respect to the seed molecules, when compared to baselines. Examples of generated molecules using pair-tuning on GP-MoLFormer are provided in the SI Fig. S6. Out of the 2408 molecules that pass the DRD2 activity threshold, only 7 molecules are seen in the DRD2 training data and only 257 are present in GP-MoLFormer training data. The novel molecules only share about 0.5 Tanimoto similarity to the DRD2 training molecules and about 0.6 Tanimoto similarity to the GP-MoLFormer training molecules, suggesting that GP-MoLFormer can be used for generating novel molecules for downstream property optimization tasks.

Scaling results

Effect of training dataset and generation pool. The generative ability of a chemical language model can be affected by the scale of the training data. Further, bias in training data can also contribute to the quality of generation. To disentangle these factors, we report GP-MoLFormer's performance on two independently varying dimensions: First, we vary the size and quality of pre-training data. For this dimension, we compare GP-MoLFormer trained on 1.1b SMILES extracted from ZINC and PubChem, and GP-MoLFormer-Uniq trained on a 650m de-duplicated subset of that 1.1b SMILES set. Secondly, we analyze the effect of varying the number of generated molecules from 30k to 10b molecules, to study the effect of inference time scaling.

To summarize, GP-MoLFormer is trained on a dataset of 650m–1.1b SMILES, which captures the relative abundance of molecules, as well as the presence of the same molecule in different context, as found in chemical databases and is evaluated on generations up to a scale of billions. This is in contrast to the existing molecular generation benchmarks that report performance metrics for a relatively small 10–30k generations, and to the current generative molecular models that are designed to target a specific distribution of molecules, *e.g.*, synthetic molecules with biological activity or natural products, and are trained on 1–100m samples.[14]

We report in Table 7 the percentage of novel (unseen in training), valid (syntactically correct), and unique (not previously generated) molecules for both GP-MoLFormer and GP-MoLFormer-Uniq, for generation size of 30k to 10b. The results show that the fraction of novel generations stays at a consistent ≈32% for GP-MoLFormer when the number of total generated molecules is below 1b. Novelty in GP-MoLFormer-Uniq is ≈5–8% higher compared to that of GP-MoLFormer for all generation pool sizes. At or beyond 1b generations, the fraction of novel and unique generations drops but still remains significant. Even for 10b generations, GP-MoLFormer is able to generate a significant 16.7% novel molecules while GP-MoLFormer-Uniq is able to generate 21.4% novel molecules. GP-MoLFormer, irrespective of training data, outputs chemically valid SMILES almost all the time. While the percentage of valid molecules drops slightly with increasing generation pool size, it still is over 99% for 10b generations.

Table 7 Novelty, validity, and uniqueness of different numbers of generations for models trained with 650m (GP-MoLFormer-Uniq) and 1.1b (GP-MoLFormer) size training sets

Generation size	Training size = 650m			Training size = 1.1b		
	Novel	Unique	Valid	Novel	Unique	Valid
30k	0.390	0.997	1.000	0.323	0.997	0.997
100k	0.393	0.996	0.999	0.326	0.998	0.998
1m	0.395	0.996	0.999	0.323	0.996	0.997
10m	0.400	0.991	0.996	0.322	0.989	0.997
100m	0.385	0.947	0.996	0.327	0.989	0.997
1b	0.340	0.675	0.996	0.278	0.611	0.997
10b	0.214	0.270	0.996	0.167	0.223	0.997

Additionally, when comparing the 10b molecules generated by the GP-MoLFormer and by the GP-MoLFormer-Uniq model, 67 to 74% of the novel molecules generated by a model are unique to that model (*i.e.*, not in the other model's generated set). This implies that the two models learned separate but overlapping manifolds. This aspect of different coverage of the molecular manifold with different model variants will be investigated further in future work.

This result confirms that (i) GP-MoLFormer trained on a billion of SMILES memorizes training samples, as seen from the high number of exact matches (1 – novelty, which can be up to 60%) with training molecules; and (ii) training memorization becomes less when the training data is de-duplicated, enabling more novel generation. (iii) With scaling of inference compute, novelty in generations reduces, but remains significant, even when evaluated against ≈10b generations. In summary, in all cases studied here, GP-MoLFormer is capable of generating novel, diverse, and valid molecules.

Discussions on the effect of training data bias on generations. Within the 1.1b training set, a notable 45% of SMILES strings were found duplicates. Some of this is due to molecules that appear multiple times as different isomers, which are pre-processed into the same SMILES representation in our canonicalization pipeline. There is also some overlap between the two databases—ZINC and PubChem. This popularity of certain molecules may or may not be related to true factors, such as the molecule being useful in a multitude of applications because of its synthetic ease or lower cost, or a combination of all of those. It is reported that such skewed distribution can also originate from anthropogenic attention bias,[11] where some molecules are studied extensively because scientists themselves or their peers "like" them. Nevertheless, such over-represenation of certain molecules, either originating from human cognitive biases, heuristics, social influences, from rationally made choices, from multiple isomeric instances, or a combination of all of the above, can lead to bias in databases.

Data de-duplication is the first step towards removing such bias, which reduces the concentration of the high density regions of the data manifold. In this case, de-duplication removes isomeric population information as well as repeated molecules across databases. The de-duplicated data is closer to a data manifold that has more homogenized density all over. Training on such a data manifold results in higher novelty in

generations, as found for GP-MoLFormer-Uniq when compared to GP-MoLFormer, as shown in Table 7.

Although many existing molecular generative models trained on a much smaller and much focused datasets have demonstrated near-perfect (100%) novelty in generations, they are for most part not suitable for studying the trade-off between training data memorization and generation novelty. Investigating such phenomenon requires studying a generative chemical (language) model that has been trained on a broader-purpose and much larger dataset at scale. Our experiments attempt to address this under-explored aspect in this study. As shown in Table 7, novelty in GP-MoLFormer generations are lower compared to ≈100% reported by baselines,[14] but still sufficiently high for practical use. When compared with recent baselines, GP-MoLFormer generations are more dissimilar to test molecules (see the earlier sections and Table 1), though GP-MoLFormer's test set is more diverse. And, finally, the low novelty in GP-MoLFormer's generations is reflective of modeling its vast training set that represents the relative usage of molecules in real-world.

Similarly, the present study highlights the importance of studying generated sets of different sizes to obtain a comprehensive view of the quality of generations, particularly when the generative model is trained on data at scale. As GP-MoLFormer-Uniq aims to capture a training data manifold of more uniform density, which is enabled by de-duplicating the training SMILES, we see a 1% rise in novelty as we increase the number of generated samples from 30k to 10m. A similar observation has been reported in image generation[38] and language generation.[7,39] To summarize, novelty in generations is influenced by the support provided by both the training distribution and the generated distribution, and should therefore be assessed relative to the sizes and diversity of those two sets.

These results in Table 7 complement and support earlier efforts focusing on studying scaling behaviors of chemical language models. One such noteworthy effort along this line is Frey *et al.* (2023),[40] where neural-scaling behavior in large chemical models was investigated by studying models with over 1b parameters and a scaling relation following a power law was established between training loss and model parameters. However, the models tested in that work were only pre-trained on datasets of size up to ≈10m data points, which is very small compared to the size of the chemical universe. In Ross *et al.* (2022),[13] the scaling behavior of MoLFormer, which is a transformer-based molecular encoder built using a masked language modeling objective, was studied. That work clearly established the scaling behavior underlying adaptation of a pre-trained model across downstream tasks, in which the number of model parameters was up to 47m while the number of training points considered was >1b. It was shown that a MoLFormer trained on 100m SMILES consistently underperformed across a wide variety of property prediction tasks, including quantum mechanical and physiological, when compared to the model trained on >1.1b SMILES, indicating predictive ability may benefit from such bias in training data. In contrast, the results in Table 7 show that a generative chemical language model trained on cleaner de-duplicated training data produces more novel generations.

We next investigate how these metrics change with varying number of generated and test molecules. Table 8 shows that, with increasing the generated pool size, scaffold similarity with respect to the test molecules becomes >0.9 while SNN reaches >0.5 when compared against 175k held-out test samples. When a larger test set of 1m molecules is used, further increases in both scaffold similarity and SNN are observed. These results imply that, with increasing size and diversity of the training data, the typical metrics used in assessing molecular generative models, such as various similarity measures with respect to a test set, should be carefully analyzed with generation and test sets that are larger in size compared to what is typically used in the field. Note that, even for 1m generations, GP-MoLFormer produces highly diverse molecules.

Presence of unique scaffolds in generations. Table 9 shows the rate at which different models generate new scaffolds as the scale of generation increases. Since comparing novelty with

Table 9 Rate of unique scaffolds generated at different scales of generated molecules. Due to computational constraints, only 10m molecules are analyzed. Bold values indicate the best model at a given scale

	Unique scaffolds			
Gen. mol.	GP-MoLFormer-Uniq	MolGen-7b	CharRNN	VAE
10k	0.839	**0.840**	0.714	0.724
100k	**0.742**	0.723	0.525	0.533
1m	**0.581**	0.550	0.326	0.326
10m	**0.388**	0.343	0.163	0.160

Table 8 Investigation on the interplay between generation set size and test set size on modeling the molecular distribution of GP-MoLFormer-Uniq generations. Columns are the same as Table 1

		MOSES					MoLFormer		
Test size	Gen. size	Frag↑	Scaf↑	SNN↑	IntDiv↑	FCD↓	DNN↓	IntDiv2↑	FMD↓
175k	30k	0.9998	0.7383	0.5045	0.8655	0.0591	6.970	13.10	0.1844
	100k	0.9998	0.8653	0.5045	0.8657	0.0279	6.967	13.10	0.1025
	1m	0.9998	0.9375	0.5040	0.8658	0.0178	6.970	13.11	0.0741
1m	30k	0.9998	0.7702	0.5738	0.8655	0.0646	6.180	13.10	0.1684
	100k	0.9998	0.9026	0.5740	0.8657	0.0331	6.179	13.10	0.0874
	1m	0.9998	0.9786	0.5739	0.8658	0.0227	6.183	13.11	0.0600

respect to different training sets is uninformative, this allows us to compare the capacity of each model to continue generating new structures, independent of the training data size. The data show GP-MoLFormer-Uniq consistently outperforms the other models, demonstrating its superiority as a base generative model.

Scaling law for inference compute. We further attempt to fit a scaling law to the empirical trend observed in Table 7 for novelty with respect to generation size, which reflects the scale of inference compute. For both models, this trend appears to follow an exponential decay of the form:

$$y = ae^{-bx} \tag{1}$$

where y is the novelty, x is the generation size, and a and b are fitted parameters. In practice, b is very close to 0 so we can rewrite this as:

$$y = ae^{-10^{c}x} \tag{2}$$

for ease of reading. These results can be further seen in SI Fig. S7. We observe that the fitted initial value a is higher and decay constant b is lower for GP-MoLFormer, meaning it starts with higher novelty and declines more slowly, when compared to GP-MoLFormer. To our knowledge, this is the first ever investigation of inference scaling of chemical language models.

Conclusion

In this work, we investigate the effect of training data scale, diversity and bias on the downstream performance of a generative chemical language model named GP-MoLFormer, built on top of a recent transformer architecture for chemical language modeling. We show the generality of the proposed GP-MoLFormer architecture on *de novo* generation and on two distinct targeted design tasks, *i.e.*, scaffold-constrained molecular decoration and unconstrained property-guided molecular optimization, whereas the model produces more diverse molecules when compared to the baselines. We further show how bias in training data can induce memorization, and thus impact the novelty of generations. We analyze how the commonly used metrics for comparing models' generations with a held-out test set are affected by the diversity of the training distribution; hence, the sizes of the test set and the generation set should be carefully considered before making a conclusion based on those metrics. To our knowledge, this is the first report demonstrating training data memorization diversity and its impact on the downstream performance of a generative chemical language model pre-trained on billion-scale data. We also investigate effect of inference compute scaling and establish a scaling law between number of generations and novelty in them.

Methods

Model details

The GP-MoLFormer decoder uses the transformer block used in MoLFormer. To avoid the quadratic complexity associated with regular attention computations in vanilla transformers,[41] MoLFormer utilized a base 12-layer transformer architecture with linear attention,[42] wherein each layer has 12 attention heads and a hidden state size of 768. A generalized random feature map[43] for the linear attention was chosen.

To better model positional dependence of tokens within a SMILES string, MoLFormer deviates from using the default absolute position embeddings and instead uses rotary embeddings:[44]

$$\text{Attention}_m(Q,K,V) = \frac{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle v_n}{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle},$$

where Q, K, V are the query, key, and value respectively, and φ a random feature map. The GP-MoLFormer is trained on the next token prediction task using a cross-entropy objective: $L_{\text{LM}}(w_1, \ldots w_n) = \sum_i \log P(w_i | w_{j<i})$. Given the size of the transformer model and the efficient linear attention, GP-MoLFormer takes only around 3 milliseconds for a single forward pass during generation, using a single A100 GPU.

Datasets and tokenization

We used two datasets for pre-training by combining SMILES from the PubChem[45] and the ZINC[9] databases with varying proportions from each. The dataset used in GP-MoLFormer training contains a total of 1.1b SMILES strings; 111m of them are from the PubChem dataset, whereas the larger 1b portion comes from the ZINC database. The Uniq dataset is a de-duplicated version of that 1.1b size dataset, which comprises 650m SMILES. We utilized the tokenizer from Schwaller *et al.* (2019)[46] to construct a vocabulary. All SMILES sequences from both PubChem and ZINC are converted to a canonical format with no isomeric information using RDKit[47] followed by de-duplication (Uniq only) and tokenization. All unique tokens extracted from the resulting output give us a vocabulary of 2357 tokens plus 5 special tokens, resulting in a total of 2362 vocabulary tokens which are used for all pre-trained models considered in this paper, irrespective of pre-training dataset size. The post tokenization sequence length of the molecules range from 1 to just over 2000 tokens. We decide to restrict the sequence length range from 1 token to 202 tokens, special tokens inclusive, to reduce computation time. Since over 99.4 percent of all molecules from our dataset contain less than 202 tokens, we hypothesize that the removal of molecules with more than 202 tokens would be of minimal negative impact on pre-training.

Large-scale training and parallelization

For pre-training, we use the causal language model objective defined in Devlin *et al.* (2019).[48] The training was performed for 4 epochs (for both 1.1b and 650m training dataset sizes) with a fixed learning rate of 1.6×10^{-4} and a batch size of 1600 molecules per GPU on a total of 16 A100 80 GB GPUs over 2 servers connected *via* EDR Infiniband fabric. It should be noted

that as the number of GPUs utilized increased, we found an increase in learning rate was necessary up to a factor of 8. The GP-MoLFormer model was trained for 28.75 hours per epoch, for 115 hours total training time, while the GP-MoLFormer-Uniq model was trained for 19.75 hours per epoch, less than 80 hours total training time.

In order to scale our training to large datasets (>1b data points), we relied on adaptive bucketing of mini-batches by sequence length, as well as parallelization *via* distributed data-parallel training. The combination of linear attention, bucketing, and data parallelism allowed us to reduce the number of GPUs needed from roughly 1000 for quadratic attention with no bucketing to 16.

Pair-tuning

The paired molecule datasets for pair-tuning experiments were taken from Jin *et al.* (2019);[49] The QED paired data used for training consists of 70 644 molecule pairs where the first/seed molecule has a QED value in the range of 0.7–0.8 while the second/target molecule has a QED of 0.9–1.0. The penalized log *P* paired data consists of 60 227 molecule pairs. It should be noted that while the paired datasets were collected such that molecular similarity within the pair is 0.4 and 0.6 for QED and log *P*, respectively, we demonstrate pair-tuning only on unconstrained property optimization tasks—we do not account for similarity preservation. The test set size for both QED and penalized log *P* optimization was 800. For the DRD2 binding optimization task, we used 34 404 molecule pairs from ZINC and Olivecrona *et al.* (2017)[36] for training and a test set of 1000 molecules.[49] For scoring the generated molecules, the bioactivity prediction model from Olivecrona *et al.* (2017)[36] is used; inactive compounds were defined with $p < 0.05$ and actives were with $p > 0.5$.

The vocabulary includes 20 randomly initialized prompt embeddings as well as the <unk> embedding from GP-MoLFormer training. For training, we prepended all 20 prompt embeddings to the <bos> embedding, followed by the embeddings of the first/seed molecule in a specific pair. We then add the <unk> embedding at the end of the first/seed molecule. After the <unk> embedding, we add the embeddings of the target molecule, followed by the <eos> embedding.

For evaluation, we do a forward pass using the following sequence: the first 20 prompt embeddings + the <bos> embedding + the input molecule embeddings + the <unk> embedding. After that, we sample from the token distribution generated by GP-MoLFormer until <eos> is encountered. For all pair-tuning experiments, batch size was set to 35, the learning rate was fixed at 3×10^{-2}, and the number of epochs run was 1000. Each epoch took 6 minutes to complete on a single GPU.

Inference speed

For a batchsize of 1024 and using a single A100 80 GB GPU, inference time of GP-MoLFormer is 4.68 s. The same for MolGen-7b, LIMO, VAE, and CharRNN is 173 s, 0.98 s, 0.29 s, and 13.9 s, respectively.

Author contributions

J. R. and B. B. trained the GP-MoLFormer models. S. H. and J. R. performed pair-tuning experiments. S. H., V. C, J. R., and J. N. conducted the model evaluation experiments. P. D., Y. M., J. R., B. B., V. C, and S. H. contributed to the idea generation, study design, result analyses, and paper writing.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data used for evaluation as well as links to the data used for pre-training are available at **https://github.com/IBM/gp-molformer/** (DOI: **https://doi.org/10.5281/zenodo.16686010**). This repository also contains the scripts used for the evaluation tasks. The model code and weights for the pre-trained GP-MoLFormer-Uniq model are available at **https://huggingface.co/ibm-research/GP-MoLFormer-Uniq** (DOI: **https://doi.org/10.57967/hf/6117**).

Supplementary information contains additional context, descriptions, experiments, and output samples from the model. See DOI: **https://doi.org/10.1039/d5dd00122f**.

Acknowledgements

Authors acknowledge IBM Research for their support.

References

1 J. Kaplan, *et al.*, Scaling laws for neural language models, *arXiv*, 2020, preprint, arXiv:2001.08361, DOI: **10.48550/arXiv.2001.08361**.
2 B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli and A. Morcos, Beyond neural scaling laws: beating power law scaling via data pruning, *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 19523–19536.
3 B. Ghorbani, O. Firat, M. Freitag, A. Bapna, M. Krikun, X. Garcia, C. Chelba and C. Cherry, Scaling Laws for Neural Machine Translation, *International Conference on Learning Representations*, 2022.
4 K. Tirumala, A. Markosyan, L. Zettlemoyer and A. Aghajanyan, Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 38274–38290.
5 N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer and C. Zhang, Quantifying Memorization Across Neural Language Models, *The Eleventh International Conference on Learning Representations*, 2023.
6 N. Kandpal, H. Deng, A. Roberts, E. Wallace and C. Raffel, Large Language Models Struggle to Learn Long-Tail Knowledge, *Proceedings of the 40th International Conference on Machine Learning*, 2023, vol. 202, pp 15696–15707.

7 K. Lee, *et al.*, Deduplicating training data makes language models better, *arXiv*, 2022, preprint, arXiv:2107.06499, DOI: 10.48550/arXiv.2107.06499.
8 A. T. Kalai and S. S. Vempala, Calibrated language models must hallucinate, *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, New York, NY, USA, 2024, preprint, arXiv:2311.14648, pp. 160–171, DOI: 10.1145/3618260.3649777.
9 J. J. Irwin and B. K. Shoichet, ZINC–a free database of commercially available compounds for virtual screening, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.
10 S. Kim, *et al.*, Pubchem substance and compound databases, *Nucleic Acids Res.*, 2016, **44**, D1202–D1213.
11 X. Jia, *et al.*, Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis, *Nature*, 2019, **573**, 251–255.
12 B. Brown, *et al.*, Large language monkeys: Scaling inference compute with repeated sampling, *arXiv*, 2024, preprint, arXiv:2407.21787, DOI: 10.48550/arXiv.2407.21787.
13 J. Ross, *et al.*, Large-scale chemical language representations capture molecular structure and properties, *Nat. Mach. Intell.*, 2022, **4**, 1256–1264.
14 D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik and A. Zhavoronkov, Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models, *Front. Pharmacol.*, 2020, DOI: 10.3389/fphar.2020.565644.
15 W. Jin, R. Barzilay and T. Jaakkola, Junction Tree Variational Autoencoder for Molecular Graph Generation, *Proceedings of the 35th International Conference on Machine Learning*, 2018, vol. 80, pp. 2323–2332.
16 P. Eckmann, *et al.*, Limo: Latent inceptionism for targeted molecule generation, *Proc. Mach. Learn.*, 2022, **162**, 5777.
17 Y. Fang, N. Zhang, Z. Chen, L. Guo, X. Fan and H. Chen, Domain-Agnostic Molecular Generation with Chemical Feedback, *The Twelfth International Conference on Learning Representations*, 2024.
18 R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, Chemformer: a pre-trained transformer for computational chemistry, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 015022.
19 M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045024, DOI: 10.1088/2632-2153/aba947.
20 M. A. Skinnider, Invalid smiles are beneficial rather than detrimental to chemical language models, *Nat. Mach. Intell.*, 2024, 1–12.
21 K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter and G. Klambauer, Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery, *J. Chem. Inf. Model.*, 2018, **58**, 1736–1741.
22 N. Brown, M. Fiscato, M. H. Segler and A. C. Vaucher, Guacamol: benchmarking models for de novo molecular design, *J. Chem. Inf. Model.*, 2019, **59**, 1096–1108.
23 A. Nigam and Tartarus, Platform for Realistic And Practical Inverse Molecular Design, *Adv. Neural. Inf. Process. Syst.*, 2023, **36**, 3263–3306.
24 K. Huang, *et al.*, Artificial intelligence foundation for therapeutic science, *Nat. Chem. Biol.*, 2022, **18**, 1033–1036.
25 J. Lee, I.-S. Myeong and Y. Kim, The Drug-Like Molecule Pre-Training Strategy for Drug Discovery, *IEEE Access*, 2023, **11**, 61680–61687.
26 J. Arús-Pous, *et al.*, Smiles-based deep generative scaffold decorator for de-novo drug design, *J. Cheminf.*, 2020, **12**, 1–18.
27 B. Lester, R. Al-Rfou and N. Constant, *The Power of Scale for Parameter-Efficient Prompt Tuning, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, 2021, pp. 3045–3059, DOI: 10.18653/v1/2021.emnlp-main.243.
28 Z. Zhou, S. Kearnes, L. Li, R. N. Zare and P. Riley, Optimization of molecules via deep reinforcement learning, *Sci. Rep.*, 2019, **9**, 10752.
29 A. G. Dossetter, E. J. Griffen and A. G. Leach, Matched molecular pair analysis in drug discovery, *Drug Discovery Today*, 2013, **18**, 724–731.
30 Z. Yang, *et al.*, Matched molecular pair analysis in drug discovery: methods and recent applications, *J. Med. Chem.*, 2023, **66**, 4361–4377.
31 J. You, B. Liu, Z. Ying, V. Pande and J. Leskovec, Graph convolutional policy network for goal-directed molecular graph generation, in *NeurIPS*, 2018, pp. 6410–6421.
32 Y. Xie, *et al.*, Mars: Markov molecular sampling for multi-objective drug discovery, in *International Conference on Learning Representations*, 2021.
33 Y. Luo, K. Yan and S. Ji, Graphdf: A discrete flow model for molecular graph generation, in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, vol. 139 of Proceedings of Machine Learning Research*, ed. M. Meila and T. Zhang, PMLR, 2021, pp. 7192–7203.
34 R. Gómez-Bombarelli, *et al.*, Automatic chemical design using a data-driven continuous representation of molecules, *ACS Cent. Sci.*, 2018, **4**, 268–276.
35 D. Erikawa, N. Yasuo, T. Suzuki, S. Nakamura and M. Sekijima, Gargoyles: An open source graph-based molecular optimization method based on deep reinforcement learning, *ACS Omega*, 2023, **8**, 37431–37441.
36 M. Olivecrona, T. Blaschke, O. Engkvist and H. Chen, Molecular de-novo design through deep reinforcement learning, *J. Cheminf.*, 2017, **9**, 48.
37 Ł. Maziarka, *et al.*, Mol-cyclegan: a generative model for molecular optimization, *J. Cheminf.*, 2020, **12**, 2.
38 T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen and T. Aila, Improved precision and recall metric for assessing generative models, *Adv. Neural Inf. Process. Syst.*, 2019, **32**, 3927–3936.
39 N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea and C. Raffel, Extracting Training Data from

Large Language Models, *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.
40 N. C. Frey, *et al.*, Neural scaling of deep chemical models, *Nat. Mach. Intell.*, 2023, **5**, 1297–1305.
41 A. Vaswani, *et al.*, Attention is all you need, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 6000–6010.
42 A. Katharopoulos, A. Vyas, N. Pappas and F. Fleuret, Transformers are rnns: Fast autoregressive transformers with linear attention, in *International Conference on Machine Learning*, PMLR, 2020, pp. 5156–5165.
43 K. M. Choromanski, *et al.*, Rethinking attention with performers, in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.
44 J. Su, Y. Lu, S. Pan, B. Wen and Y. Liu, Enhanced transformer with rotary position embedding, *arXiv*, 2021, preprint, arXiv:2104.09864, DOI: **10.1016/j.neucom.2023.127063**.
45 S. Kim, *et al.*, PubChem 2019 update: improved access to chemical data, *Nucleic Acids Res.*, 2018, D1102–D1109.
46 P. Schwaller, *et al.*, Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction, *ACS Cent. Sci.*, 2019, **5**, 1572–1583, DOI: **10.1021/acscentsci.9b00576**.
47 RDKit: Open-source cheminformatics, 2021, **http://www.rdkit.org**, [Online; accessed 28-May-2021].
48 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the NAACL: HLT*, 2019, vol. 1.
49 W. Jin, K. Yang, R. Barzilay and T. Jaakkola, Learning Multimodal Graph-to-Graph Translation for Molecule Optimization, *International Conference on Learning Representations*, 2019.