nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00368-1

Chemical language models enable navigation in sparsely populated chemical space

Michael A. Skinnider[1]✉, R. Greg Stacey[1], David S. Wishart[2,3,4,5] and Leonard J. Foster[1,6]✉

Deep generative models are powerful tools for the exploration of chemical space, enabling the on-demand generation of molecules with desired physical, chemical or biological properties. However, these models are typically thought to require training datasets comprising hundreds of thousands, or even millions, of molecules. This perception limits the application of deep generative models in regions of chemical space populated by a relatively small number of examples. Here, we systematically evaluate and optimize generative models of molecules based on recurrent neural networks in low-data settings. We find that robust models can be learned from far fewer examples than has been widely assumed. We identify strategies that further reduce the number of molecules required to learn a model of equivalent quality, notably including data augmentation by non-canonical SMILES enumeration, and demonstrate the application of these principles by learning models of bacterial, plant and fungal metabolomes. The structure of our experiments also allows us to benchmark the metrics used to evaluate generative models themselves. We find that many of the most widely used metrics in the field fail to capture model quality, but we identify a subset of well-behaved metrics that provide a sound basis for model development. Collectively, our work provides a foundation for directly learning generative models in sparsely populated regions of chemical space.

Chemical space is vast. The number of synthetically accessible organic molecules alone exceeds 10^{60} (ref. [1]). Humans have explored only infinitesimal regions of this vast space over the course of recorded history[2], yet this exploration has yielded an arsenal of molecules that form the basis for much of medical practice. These successes, against overwhelming odds, lead to optimism that more efficient ways of navigating chemical space could help address many of the most pressing challenges facing humanity.

Historically, many approaches to chemical space exploration aimed to enumerate the set of molecules comprising an explicitly defined space[2–8]. More recently, deep generative models have emerged as a powerful tool for chemical space exploration[9]. These models leverage deep neural networks to learn the chemistries implicitly embedded within a set of training molecules. Once trained, these models are capable of stochastically sampling unseen molecules from the target chemical space.

Many of the most successful approaches to generative modelling learn to generate textual representations of molecules, commonly in the simplified molecular-input line-entry system (SMILES) format[10] (Fig. 1a). This strategy allows practitioners to borrow architectures from the field of natural language processing known as recurrent neural networks (RNNs; Fig. 1b)[11–18]. Although alternative approaches have been proposed, such as learning to generate graphs[19,20] or to assemble molecules from substructures[21], systematic benchmarks have not shown these to outperform RNN-based models of SMILES strings[22,23], which we refer to here as chemical language models (CLMs).

CLMs have attracted interest for their potential to generate molecules with arbitrary physicochemical or biological properties on demand, and thereby solve what has been termed the 'inverse design' problem[24]. A major outstanding challenge, however, is that these models are typically seen to require large amounts of training data—on the order of hundreds of thousands to millions of molecules[9]. It is often the case that the chemical space targeted for exploration is not populated by a commensurate number of examples. For example, generative models could be used to suggest plausible structures for unidentified molecules in untargeted metabolomics, but for many species or taxa, only a few thousand metabolites are known. To enable generative modelling in low-data regimes, methods based on reinforcement learning (RL)[13,17,25–27] or transfer learning (TL)[12,15,16,28–31] have been developed. In both paradigms, models are first 'pre-trained' on a large and generic database of chemical structures, and thereafter undergo a second round of 'fine-tuning' meant to guide them into a more restricted chemical space. However, several shortcomings of these approaches have been noted. Both approaches may suffer from mode collapse or catastrophic forgetting[26]. In RL-based approaches, the more powerful generative model may learn to exploit unforeseen deficiencies in the reward function, leading to the generation of unrealistically simple but high-scoring molecules[13,32]. Finally, both strategies yield results that vary depending on the duration of the fine-tuning step, and there is no obvious a priori basis to infer an optimal duration[28].

Ideally, it would be possible to directly learn a generative model from a small number of examples. At present, however, it is unclear what the lower bound might be on the number of molecules needed to learn a robust model. Moreover, despite some pioneering efforts[16,33], it remains unclear whether specific strategies could optimize generative models for the low-data regime. Such strategies might include varying the textual representation of the input molecules, the architecture or hyperparameters of the CLM, the process by which the CLM is trained or strategies for data augmentation.

[1]Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada. [2]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada. [3]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada. [4]Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Alberta, Canada. [5]Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, Alberta, Canada. [6]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia, Canada.
✉e-mail: michael.skinnider@msl.ubc.ca; foster@msl.ubc.ca

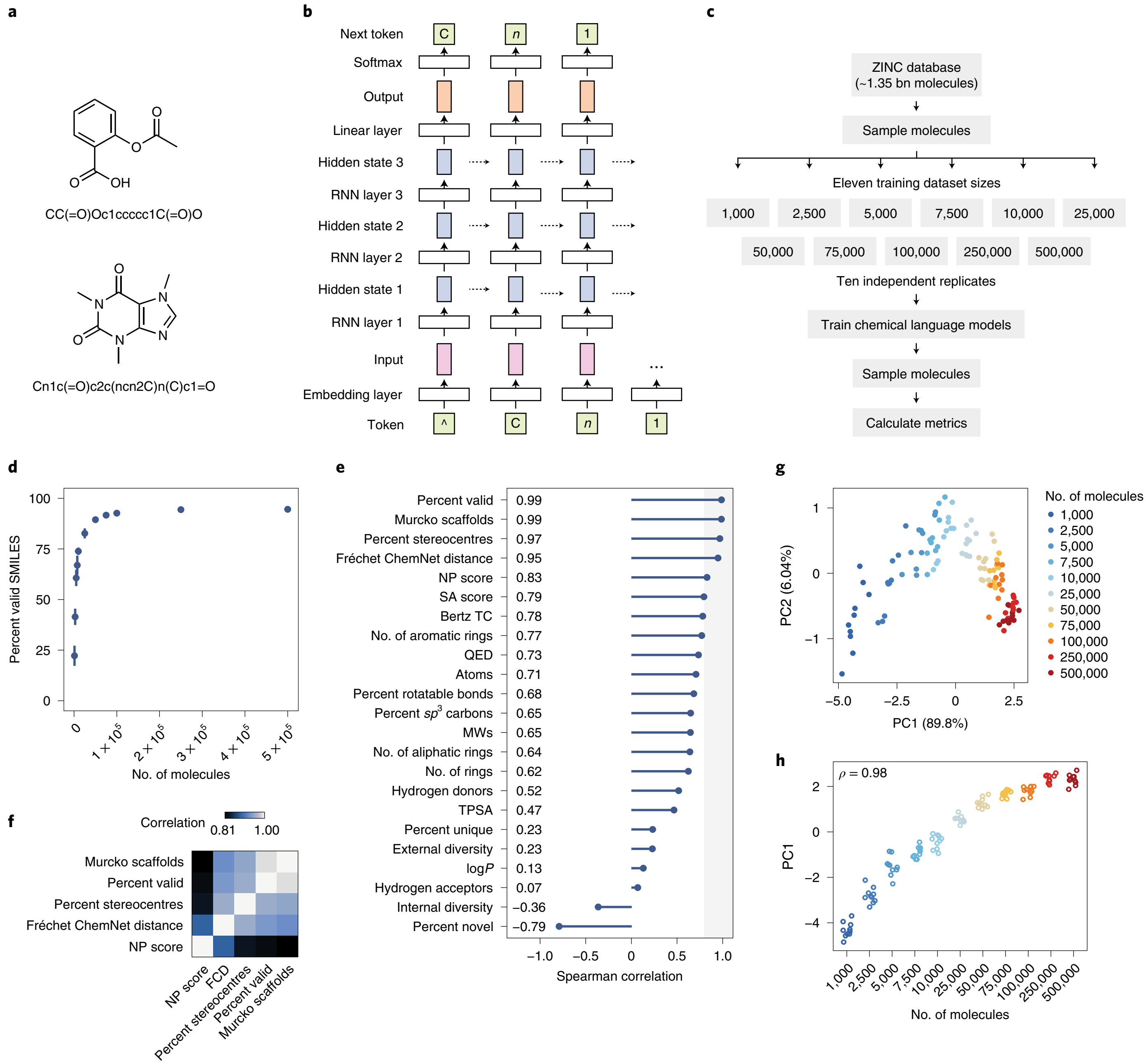


Fig. 1 | Learning generative models of molecules from limited training examples. a, Molecular structures and canonical SMILES representations of two exemplary molecules, aspirin (top) and caffeine (bottom). **b**, Architecture of a three-layer RNN trained to generate SMILES strings. **c**, Overview of the experimental design. **d**, Proportion of valid SMILES generated by CLMs trained on one of varying numbers of molecules sampled from the ZINC database. The means and standard deviations of 10 independent replicates are shown. **e**, Spearman correlations between training dataset size (number of molecules) and each of 23 proposed metrics for the evaluation of CLMs trained on the ZINC database. The shaded area highlights metrics with a rank correlation of ≥0.8 to the training dataset size. NP, natural product-likeness; SA, synthetic accessibility; TC, topological complexity; QED, quantitative estimate of drug-likeness; MWs, molecular weights; TPSA, topological polar surface area. **f**, Matrix of Spearman correlations between the values of the five top-performing metrics across $n=110$ CLMs. **g**, PCA of top-performing metrics for molecules generated by $n=110$ CLMs trained on varying numbers of molecules sampled from ZINC, coloured by the size of the training dataset. **h**, PC1 scores for $n=110$ CLMs trained on varying numbers of molecules sampled from ZINC. Inset text shows the Spearman correlation.

Here, we systematically evaluate the ability of CLMs to learn from limited training data. We find that robust models can be learned from surprisingly few examples. We then identify strategies that reduce the amount of training data required to learn a model of equivalent quality. Conversely, our systematic benchmarks indicate that several of the strategies that have been proposed in the literature for this purpose are ineffective. We demonstrate the application of the principles that emerge from our analysis by training CLMs of bacterial, plant and fungal metabolites, which learn to reproduce highly complex chemical spaces from only thousands of input molecules.

A secondary outcome of our work is that the structure of our experiments provides an opportunity to compare the metrics that are currently used to evaluate generative models themselves.

Surprisingly, we find that many of these metrics fail to capture model quality. However, we identify five metrics that are robustly correlated with the size of the training dataset, and develop a framework to integrate these into a holistic measure of performance.

Results

We initially set out to determine the minimum number of molecules required to train a robust CLM. To this end, we trained models on random samples of between 1,000 and 500,000 SMILES strings from the ZINC database of commercially available compounds[34], then sampled 500,000 SMILES from each trained model (Fig. 1c). We repeated this process 10 times for each sample size.

As an initial assessment, we calculated the proportion of valid SMILES generated by each model, a metric that has been widely used to evaluate generative models of molecules. The proportion of valid molecules increased rapidly as the size of the training dataset increased, from only 6.7% with 1,000 molecules to 69.1% with 25,000 molecules (Fig. 1d). On the other hand, performance saturated rapidly after ~50,000 molecules.

Widely used metrics fail to capture generative model performance. This observation suggested that CLMs can be learned from surprisingly small training datasets. However, the proportion of valid SMILES captures only one aspect of model performance. It is possible that a model has learned to generate valid SMILES strings, but that the resulting molecules bear little resemblance to those in the training set. We therefore sought to achieve a more holistic evaluation of model performance.

We calculated a suite of 23 different metrics that have previously been proposed to evaluate generative models of molecules[16,22,23,33,35–37] (Methods and Extended Data Fig. 1a). In the absence of a 'ground truth', it has been unclear which of these metrics best capture the quality of the underlying model. We reasoned that the structure of our experiment could be used to ascertain the most useful metrics. Specifically, we reasoned that, as the size of the training set increases, so too should measures of model performance. To formalize this notion, we calculated the Spearman rank correlation between the size of the training dataset and the value of each metric. We then compared the 23 metrics based on this correlation.

Surprisingly, we observed enormous variation in the performance of the 23 metrics (Fig. 1e). A handful were strongly correlated to the size of the training dataset, including the proportion of valid molecules, the Fréchet ChemNet distance (FCD)[36] and the Murcko scaffolds of the generated molecules (Extended Data Fig. 1b). However, the majority were at best moderately correlated to this experimental 'ground truth', and a subset exhibited no statistically significant correlation at all (Extended Data Fig. 1c,d). Among these were two of the most widely used metrics in the field: the proportion of unique molecules and the computed octanol-water partition coefficient ($\log P$) of generated molecules. Notably, all models generated unique molecules at a rate exceeding 99%, suggesting that generating diverse molecules is sufficiently easy for CLMs that this metric does not provide a useful benchmark.

Holistic evaluation of generative models of molecules. We sought to integrate information from several top-performing metrics to arrive at a single measure of model performance. However, these metrics are measured on very different scales and exhibit a complex correlation structure (Fig. 1f), precluding a simple averaging procedure. We reasoned that, in the context of this experiment, the size of the training dataset would represent the primary source of variation in the values of these metrics. Consequently, we hypothesized that in a principal component analysis (PCA), models would segregate along the first principal component (PC1) according to the size of the training dataset. This hypothesis was borne out by a PCA of the 110 models trained on samples from the ZINC database (Fig. 1g,h). Notably, integrating information from multiple metrics revealed that model performance continued to improve above the plateau suggested by the proportion of valid molecules. This observation suggests that, as the size of the training set increases, CLMs first learn to produce valid SMILES and only later learn to match the structural and physicochemical properties of the target molecules. Consequently, integrating multiple distinct sources of information is necessary for a holistic evaluation.

Learning CLMs of distinct chemical spaces. We next asked whether the number of molecules required to train a robust CLM would vary as a function of the target chemical space. To test this hypothesis, we repeated our initial experiment, but with molecules sampled from three different databases (Fig. 2a)[7,38,39]. These databases have distinct structural properties, with molecules from COCONUT generally being the most complex, followed by ChEMBL, ZINC and GDB (Fig. 2b).

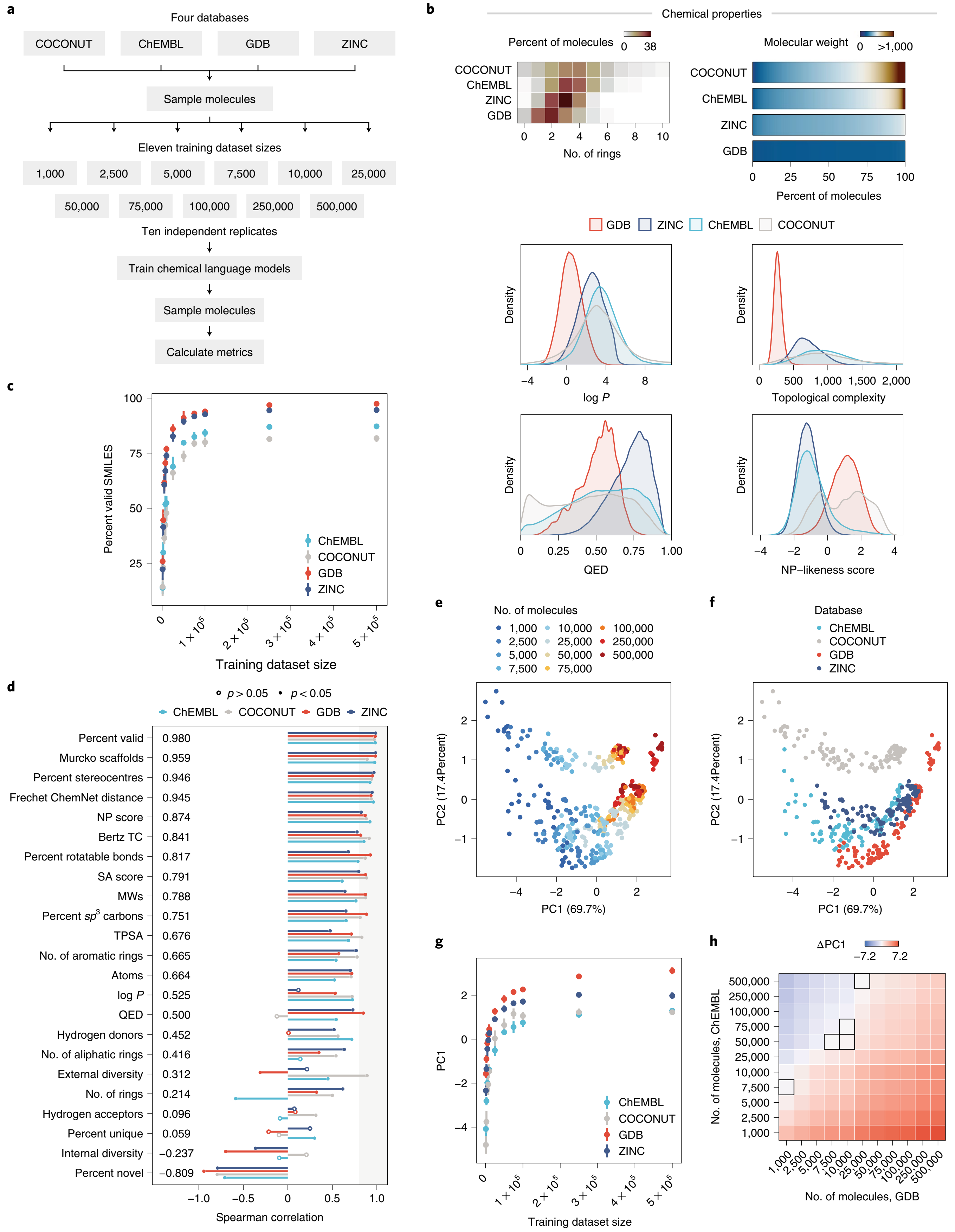
A comparison of the proportion of valid molecules suggested that the minimum number of examples required to learn a robust model depends on the complexity of the target chemical space (Fig. 2c). Models trained on small organic compounds from GDB, for example, always produced a higher proportion of valid SMILES strings than models trained on an equivalent number of molecules from ZINC. By contrast, models of the COCONUT database never produced valid SMILES at a rate exceeding 82%.

We next asked whether the 23 metrics exhibited the same relationship to model performance as observed in ZINC. We confirmed that the majority of proposed metrics were weakly or inconsistently correlated to the experimental ground truth (Fig. 2d and Extended Data Fig. 2). Importantly, however, we found that the same five metrics achieved a rank correlation of $\geq$0.8 in all four databases.

In a combined PCA of all four databases, models separated along PC1 based on the size of the training dataset (Fig. 2d–f and Extended Data Fig. 2d,e). We obtained similar results when performing PCA within each database separately (Extended Data Fig. 3a)

Fig. 2 | Low-data generative models of distinct chemical spaces. a, Overview of the experimental design. **b**, Structural and physicochemical properties of molecules from the four chemical databases analysed in this study. Top left: number of rings per molecule. Top right: molecular weight spectrum of molecules from each database. Centre left: octanol–water partition coefficients ($\log P$)[66]. Centre right: Bertz topological complexities[65] of each molecule. Bottom left: quantitative estimate of drug-likeness (QED) scores[69]. Bottom right: natural product (NP)-likeness scores[68]. **c**, Proportion of valid SMILES generated by CLMs trained on one of varying numbers of molecules sampled from one of four chemical databases. The means and standard deviations of 10 independent replicates are shown. **d**, Spearman correlations between training dataset size (number of molecules) and each of 23 proposed metrics for the evaluation of chemical generative models in four chemical databases. Inset numbers show the mean Spearman correlation. The shaded area highlights metrics with a rank correlation of ≥ 0.8 to the training dataset size. **e**, PCA of top-performing metrics for molecules generated by $n = 440$ CLMs, trained on molecules sampled from four different databases, coloured by the size of the training dataset. **f**, As in **e**, but coloured by the chemical database on which the generative models were trained. **g**, PC1 scores for CLMs trained on varying numbers of molecules sampled from one of four chemical databases. The means and standard deviations of 10 independent replicates are shown. **h**, Mean difference in PC1 scores ($\Delta PC1 = PC1_{GDB} - PC1_{ChEMBL}$) between CLMs trained on varying numbers of molecules sampled from GDB (x axis) or ChEMBL (y axis). Black-outlined squares indicate pairs without statistically significant differences (uncorrected $p > 0.05$, two-sided t-test).

or when withholding one database from the PCA and projecting the withheld models onto the coordinate basis of the other three databases (Extended Data Fig. 3b). These findings indicate that the loadings learned from a PCA of a diverse set of generative models can be applied to unseen models. Performance decreased linearly below 1,000 molecules, suggesting that RL- or TL-based strategies



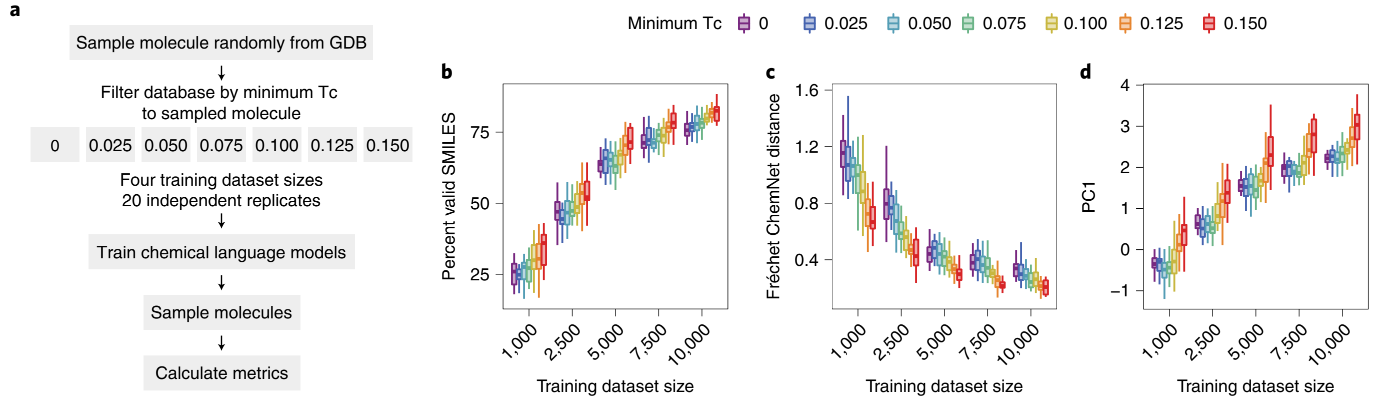


Fig. 3 | Low-data generative models of diverse and homogeneous molecules. **a**, Overview of the experimental design. **b**–**d**, Performance of CLMs trained on samples of molecules from the GDB database with a minimum Tanimoto coefficient (Tc) to a randomly selected 'founder' molecule. Samples with a lower minimum Tc are more diverse, whereas samples with a higher minimum Tc are more homogeneous. **b**, Proportion of valid SMILES generated by CLMs trained on varying numbers of more or less diverse molecules from the GDB database. **c**, Fréchet ChemNet distances of CLMs trained on varying numbers of more or less diverse molecules from the GDB database. **d**, PC1 scores of CLMs trained on varying numbers of more or less diverse molecules from the GDB database.

may remain the only viable options for the smallest training datasets (Extended Data Fig. 4).

A direct comparison of data requirements across chemical spaces revealed unexpectedly large differences in 'data hungriness' (Extended Data Fig. 5). For example, a training dataset of 500,000 molecules was required to learn a model of ChEMBL that was statistically indistinguishable from a model of the GDB learned from only 25,000 examples (Fig. 2h). This observation raises the possibility that results obtained from the GDB database may not be applicable to models of more complex molecules[14,40].

Finally, we asked how the diversity of the sampled molecules impacted model performance in the low-data regime. We trained CLMs on increasingly homogeneous samples from the GDB database. Both individual metrics and PC1 scores indicated that performance decreased as the diversity of the sampled molecules increased (Fig. 3). We observed similar trends in ChEMBL and ZINC (Extended Data Fig. 6). These findings suggest that efforts to learn CLMs from a small number of examples are substantially more likely to succeed in relatively homogeneous regions of chemical space.

Evaluating molecular representations for CLMs. To date, the SMILES format has been the most common textual representation used to train RNNs. However, models trained on SMILES strings often generate a large proportion of invalid molecules, which some have identified as a key limitation[41–44]. Two prominent alternatives to the SMILES format have been proposed. The DeepSMILES variant introduces two modifications to the SMILES syntax to remove long-term dependencies associated with the representation of rings and branches[41]. Self-referencing embedded strings (SELFIES) are an entirely different representation based on a Chomsky type-2 grammar, in which every SELFIES string specifies a valid chemical graph[42].

We trained generative models on SMILES, DeepSMILES and SELFIES representations of molecules from all four databases (Fig. 4a). Inspection of the proportion of valid molecules confirmed that models trained on SELFIES strings did indeed produce valid chemical graphs at a rate of 100% (Fig. 4b and Extended Data Fig. 7a). Surprisingly, models trained on DeepSMILES did not produce valid molecules at a substantially higher rate than ones trained on canonical SMILES.

To investigate how well models trained on each representation learned to match the target chemical space, we again performed PCA (Extended Data Fig. 7b). Surprisingly, we found that models trained to generate SELFIES strings consistently achieved lower PC1 scores than models trained on SMILES or DeepSMILES representations of the same molecules (Fig. 4c and Extended Data Fig. 7c). Inspecting individual metrics corroborated this trend: for example, models trained on SELFIES also had a higher Fréchet ChemNet distance to the training set (Fig. 4d and Extended Data Fig. 7d). For some very small sample sizes ($n \leq 5{,}000$), models trained on SELFIES or DeepSMILES did occasionally achieve higher PC1 scores (Fig. 4e),

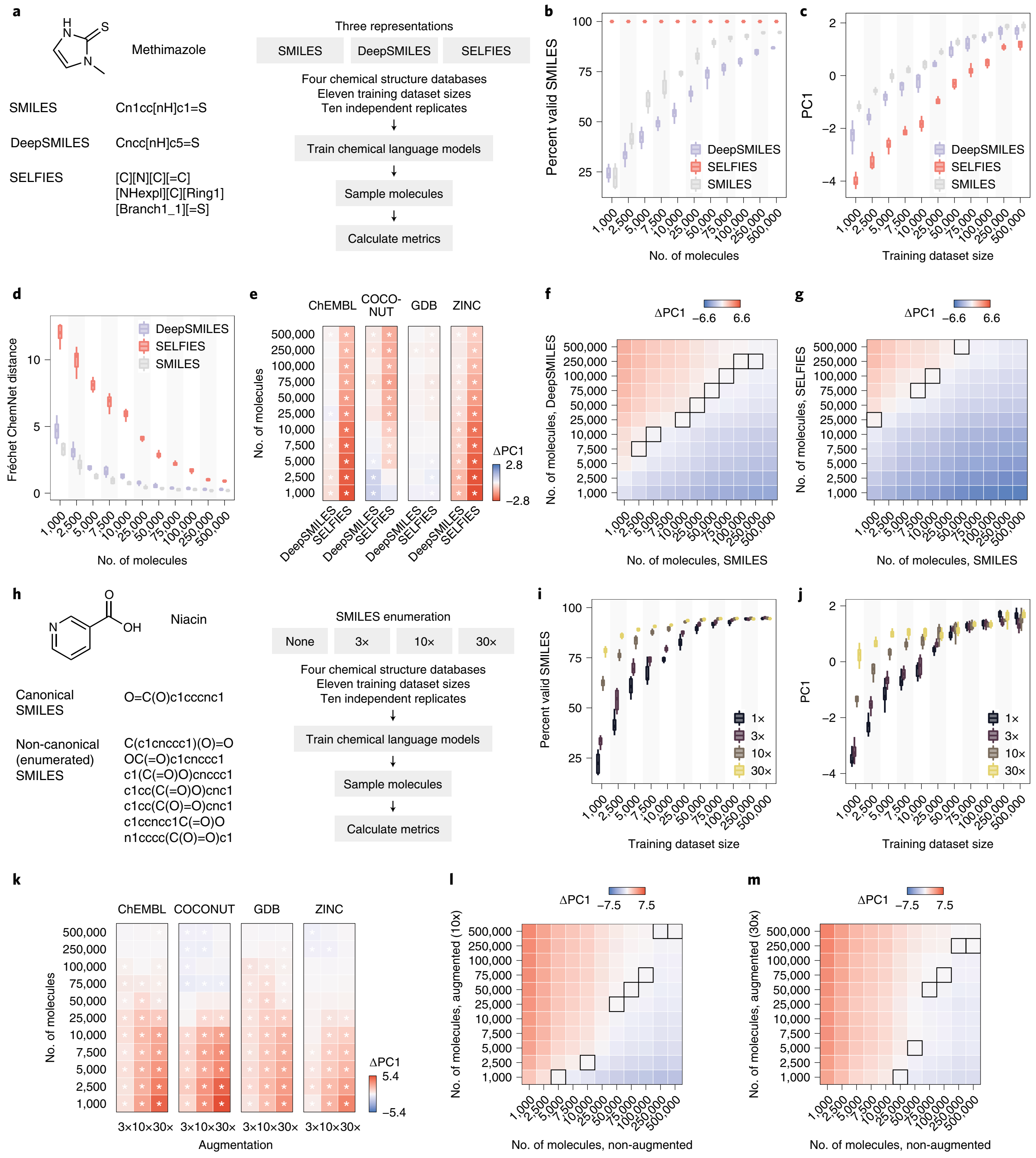
→

Fig. 4 | Alternative molecular representations for low-data generative models. **a**, Left: three string-based molecular representations of an example molecule, the thyroperoxidase inhibitor methimazole. Right: overview of the experimental design. **b**, Proportion of valid SMILES generated by CLMs trained on one of three string representations of molecules from the ZINC database. **c**, PC1 scores of CLMs trained on one of three string representations of molecules from the ZINC database. **d**, Fréchet ChemNet distances of CLMs trained on one of three string representations of molecules from the ZINC database. **e**, Mean difference in PC1 scores (ΔPC1) between CLMs trained on matching numbers of SELFIES or DeepSMILES, as compared to SMILES, from one of four chemical databases. Asterisks indicate statistically significant differences (uncorrected $p < 0.05$, two-sided *t*-test). **f**, Mean difference in PC1 scores between CLMs trained on varying numbers of molecules sampled from ZINC, represented either as DeepSMILES (*y* axis) or SMILES (*x* axis). Black-outlined squares indicate pairs without statistically significant differences (uncorrected $p > 0.05$, two-sided *t*-test). **g**, As in **f**, but with models trained on SELFIES on the *y* axis. **h**, Left: canonical SMILES and seven enumerated non-canonical SMILES for an example molecule, the nutrient and cholesterol-lowering agent niacin. Right: overview of the experimental design. **i**, Proportion of valid SMILES generated by CLMs trained on molecules sampled from the ZINC database after varying degrees of non-canonical SMILES enumeration. **j**, PC1 scores of CLMs trained on molecules sampled from the ZINC database after varying degrees of non-canonical SMILES enumeration. **k**, Mean difference in PC1 scores (ΔPC1) between CLMs trained on non-canonical SMILES with varying degrees of data augmentation from one of four chemical databases, as compared to canonical SMILES. Asterisks indicate statistically significant differences (uncorrected $p < 0.05$, two-sided *t*-test). **l**, Mean difference in PC1 scores between CLMs trained on molecules from the ZINC database represented as canonical SMILES (*x* axis) or non-canonical SMILES after 10× augmentation (*y* axis). Black-outlined squares indicate pairs without statistically significant differences (uncorrected $p > 0.05$, two-sided *t*-test). **m**, As in **l**, but with an augmentation factor of 30×.

but these differences were modest, marginally significant, and inconsistent across chemical spaces. The net result was that substantially more DeepSMILES or SELFIES were required to learn a model of equivalent quality to one trained on SMILES strings (Fig. 4f,g and Extended Data Fig. 7e).

Although the tendency of generative models trained on SMILES strings to produce invalid outputs has been seen as a central limitation of these models, our results suggest that this may actually represent an unrecognized strength. After filtering out these invalid molecules, models trained on SMILES strings matched the target chemical space better than models trained on alternative representations.

Paradoxical effects of data augmentation on CLMs. By convention, each chemical structure possesses a single, 'canonical' SMILES representation. However, hundreds of 'non-canonical' SMILES representations can also be enumerated by varying the order in which the atoms in the molecule are traversed[45] (Fig. 4h).



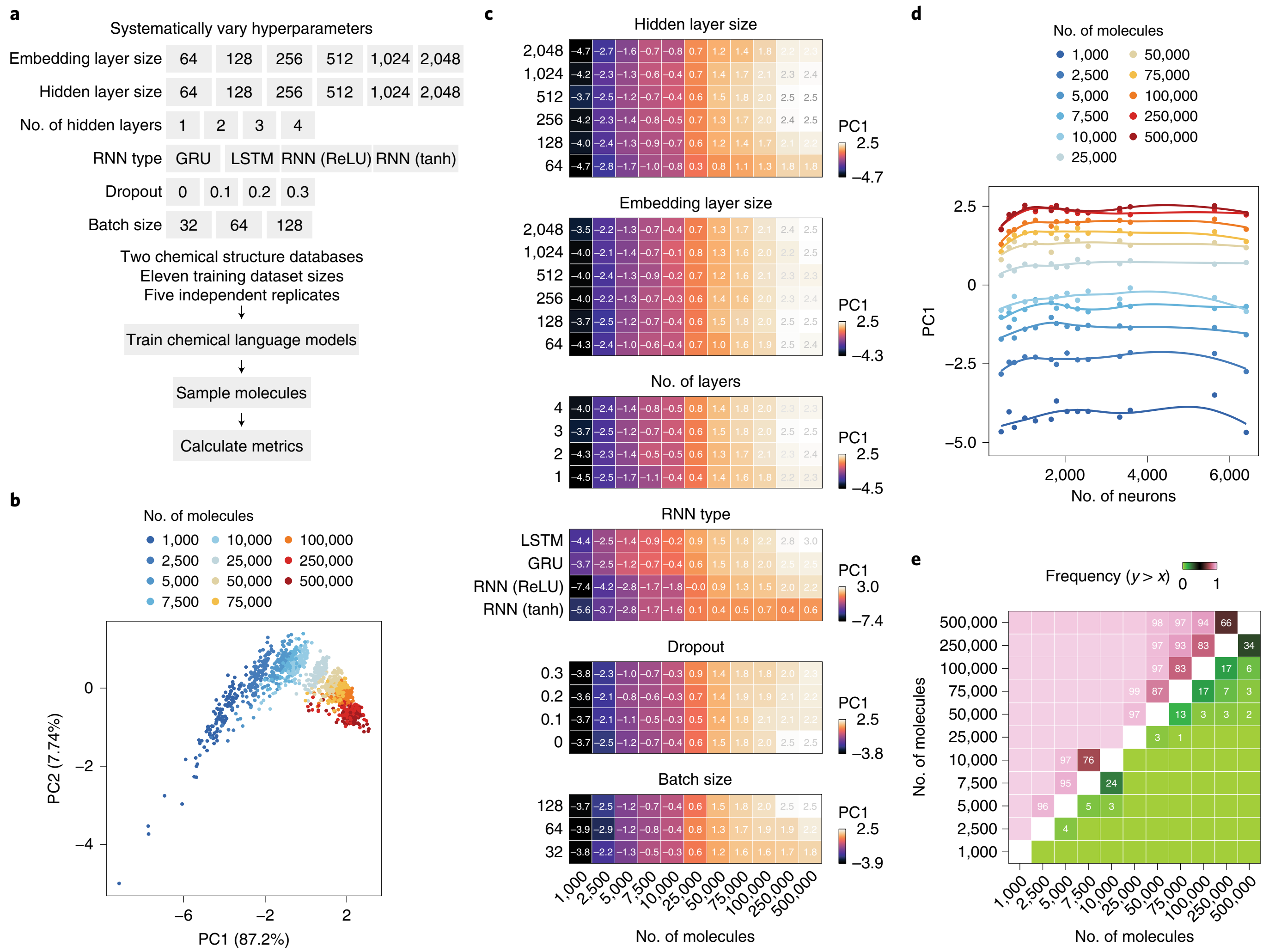


Fig. 5 | Data, not architecture, dictates the performance of low-data generative models. a, Overview of the experimental design. **b**, PCA of top-performing metrics for molecules generated by $n=1{,}210$ CLMs, trained on varying numbers of molecules from the ZINC database with varying model hyperparameters, coloured by the size of the training dataset. **c**, Mean PC1 scores for molecules trained on the ZINC database, as a function of both the number of molecules in the training dataset (*x* axis) and varying hyperparameters (*y* axis). The mean of five independent replicates is shown. **d**, Mean PC1 scores of CLMs as a function of the total number of neurons in the network. Solid lines show local polynomial regression. **e**, Proportion of $n=110$ CLMs with varying hyperparameters, trained on the number of molecules shown on the *y* axis, that outperformed a model without any hyperparameter tuning trained on the number of molecules shown on the *x* axis.

Enumeration of non-canonical SMILES has been employed to learn continuous representations of chemical structures by training sequence-to-sequence models[46,47], and emerging evidence suggests that SMILES enumeration can improve the quality of generative models[16,33].

We tested whether SMILES enumeration could decrease the number of training examples needed to learn a CLM (Fig. 4h). Models trained on enumerated SMILES generated valid molecules at a dramatically higher rate, especially in the smallest training datasets (Fig. 4i and Extended Data Fig. 8a–c). The exception was for the most structurally complex databases, in which very high degrees of data augmentation sometimes appeared to degrade the quality of models learned from large training datasets (Extended Data Fig. 8b).

PCA underscored the context-specific effects of SMILES enumeration (Fig. 4j and Extended Data Fig. 8d,e). Data augmentation had by far the largest effect on models learned from very small training datasets. Conversely, in the largest training datasets, we occasionally observed a negative effect of SMILES enumeration (Fig. 4k). Together, these findings suggest that data augmentation is best reserved for the low-data regime, particularly when modelling structurally complex molecules.

To quantify the improvement in performance attributable to SMILES enumeration, we compared models trained on augmented datasets to non-augmented datasets of varying sizes (Fig. 4l,m). For very small training datasets, data augmentation by a factor of 10 yielded a performance increase on par with quadrupling the number of unique molecules in the training set (Fig. 4l and Extended Data Fig. 8f). Augmentation by a factor of 30 had even more dramatic effects, allowing a model trained on only 5,000 molecules to match the PC1 scores of one trained on 50,000 canonical SMILES (Fig. 4m). However, this improvement in performance was attenuated completely in datasets of 500,000 molecules.

Taken together, these analyses highlight the conflicting impacts of SMILES enumeration. When learning generative models from very small training datasets, data augmentation can dramatically improve performance. On the other hand, our experiments expose a potential for ‘over-enumeration’ in large datasets of structurally

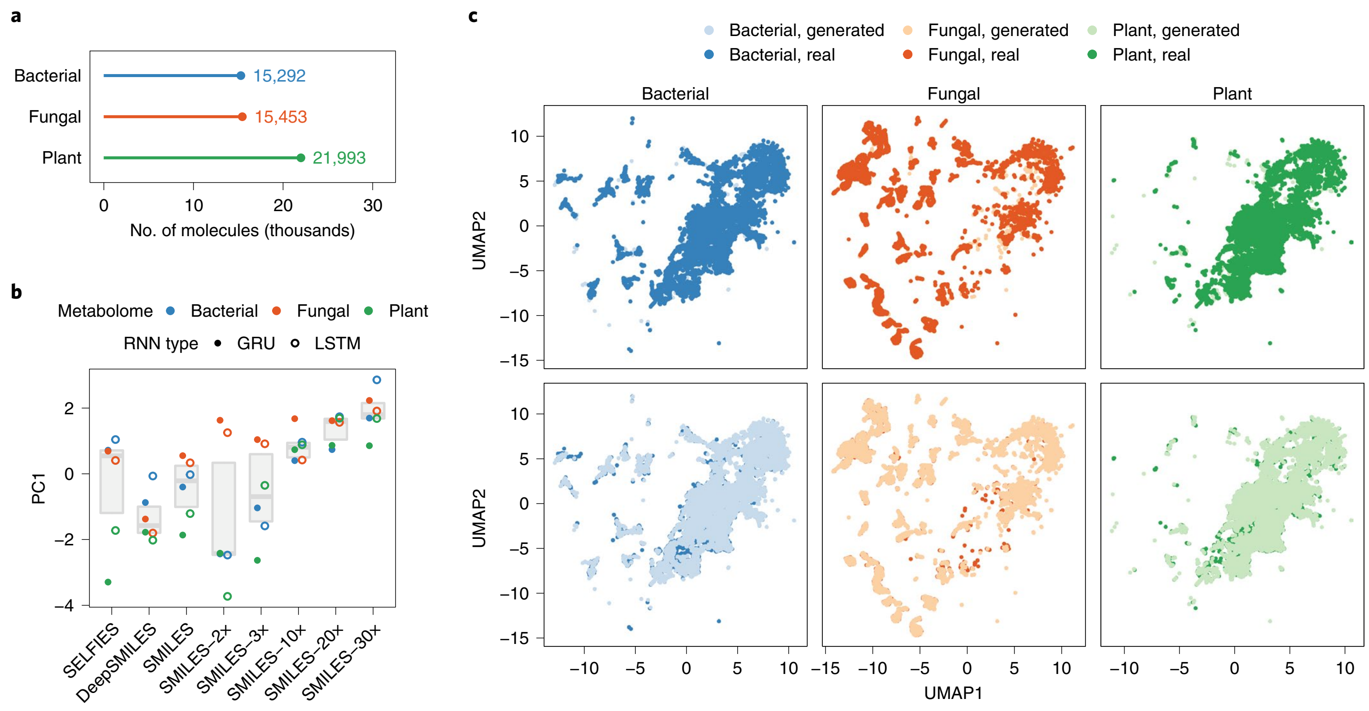


Fig. 6 | Low-data generative models of bacterial, fungal and plant metabolomes. a, Number of bacterial, fungal and plant metabolites used to train CLMs. **b**, PC1 scores of generative models of metabolomes trained with different molecular representations (SMILES, DeepSMILES or SELFIES), data augmentation strategies (non-canonical SMILES enumeration with an augmentation factor of between 2× and 30×) and RNN architectures (GRU or LSTM). **c**, Uniform manifold approximation and projection (UMAP) visualization of the known bacterial, fungal and plant metabolomes and an equal number of hypothetical metabolites sampled at random from generative models. Top: real metabolites superimposed over generated metabolites. Bottom: generated metabolites superimposed over real metabolites.

complex molecules, whereby even low levels of data augmentation can negatively impact performance.

Data, not architecture, dictates model performance in the low-data regime. Our experiments to this point have focused on varying the data provided as input to a CLM. We next asked whether we could optimize the model itself for the low-data regime. To test this possibility, we systematically varied each of six model hyperparameters, training a total of 1,210 models on molecules from the ZINC database (Fig. 5a). These models segregated along PC1 based on the size of the training dataset (Fig. 5b), suggesting that the impact of hyperparameter tuning was small in comparison to the size of the training dataset. To formally quantify this notion, we assessed the impact of each hyperparameter in turn on PC1 (Fig. 5c). For parameters controlling the capacity of the neural network (hidden layer size, embedding layer size and the number of hidden layers), intermediate values typically yielded the best performance. However, even for very small or very large models, the difference was small in comparison to the number of molecules in the training dataset (Fig. 5d). The architecture of the RNN had a somewhat greater effect, with gated recurrent units (GRUs) and long short-term memory networks (LSTMs) achieving roughly identical performance, but 'vanilla' RNNs performing substantially worse. Neither dropout nor batch size markedly affected model performance. We observed concordant results in a second database (Extended Data Fig. 9).

Together, these findings emphasize the importance of the training dataset for CLMs. Across a large grid of hyperparameters, hyperparameter tuning almost never affected performance to a comparable degree as increasing the size of the training dataset (Fig. 5e).

Case study: learning generative models of bacterial, fungal and plant metabolomes. Our experiments elucidated principles for learning CLMs from limited training data. To exemplify these principles, we aimed to learn models of bacterial, fungal and plant metabolomes. Previous work has shown that manual enumeration of hypothetical metabolites can enable the discovery of novel molecules using mass spectrometry[48,49]. Generative models of metabolomes could more efficiently traverse metabolite chemical space[50], and thereby facilitate the identification of unknown metabolites.

We assembled databases of bacterial, fungal and plant metabolites, but these each comprised only 15,000–22,000 molecules (Fig. 6a). These databases are thus far smaller than those typically used to train models of much less complex molecules. With this challenge in mind, we asked whether applying the principles we had elucidated for low-data generative models could allow us to directly model these metabolomes. We selected an LSTM with a high degree of SMILES enumeration as the optimal strategy (Fig. 6b and Extended Data Fig. 10). Despite the limited amount of training data, the optimized models generated molecules whose physicochemical properties closely matched those of the target metabolomes (Supplementary Fig. 1). Moreover, visualizing the chemical space occupied by real and generated metabolites in two dimensions revealed that the generative models almost perfectly reproduced the chemical space of the three target metabolomes (Fig. 6c).

Taken together, these experiments demonstrate that CLMs can directly learn to reproduce even very complex chemical spaces from a small number of training examples. The hypothetical metabolites generated by these models may number among the 'dark matter' of observed but unidentified metabolites in high-throughput metabolomics[51].

Discussion

CLMs have emerged as powerful tools for chemical space exploration. However, these models are widely perceived to require very large training datasets. In this Article, we set out to quantify the minimum number of molecules required to learn a robust CLM and identify strategies to reduce this lower bound. To achieve these goals, we devised a series of systematic benchmarks. In total, we trained almost 8,500 CLMs and evaluated more than four billion generated molecules. The scale of this effort allowed us to comprehensively survey strategies for training and evaluating CLMs in the low-data regime.

We found that robust models can be learned from far less data than has previously been appreciated. However, performance was contingent on the target chemical space, with a larger number of training examples needed to learn models of structurally complex molecules. We evaluated two alternatives to the SMILES format, DeepSMILES and SELFIES, that have been proposed specifically for CLMs. Surprisingly, we found that while models trained on SELFIES strings produced valid molecules at a near-perfect rate, these molecules failed to match the target chemical space as well as those generated by a model trained on SMILES. The most successful strategy we identified to improve generative modelling in the low-data regime involved enumerating multiple non-canonical SMILES for each molecule in the training set. Notably, in contrast to interventions that affected the input data, modifying the architecture, hyperparameters or training strategy of the generative models had little effect on performance. This observation suggests that developing new strategies for molecular representation and data augmentation is likely to present a more fruitful direction for future research than altering the structure of the neural network itself.

Our experiments also allowed us to benchmark the metrics themselves that are used to evaluate generative models. That there is little agreement within the field on how generative models of molecules ought to be evaluated has been noted by several commentators[32,52,53]. The lack of an 'even playing field' for model evaluation hinders comparisons of published models, making it difficult to discern which strategies have been successful and which have not. We found that many widely used metrics were at best weakly correlated to our experimental ground truth. We argue that this calls into question their use in model evaluation. However, we identified a subset of metrics that consistently exhibited strong correlations to this ground truth. Importantly, our data do not allow us to exclude the possibility that the absence of a correlation may reflect shortcomings of the generative models themselves, rather than the metrics under investigation. However, the fact that we identified a number of strong and reproducible correlations to metrics such as the proportion of valid molecules or the Fréchet ChemNet distance supports the general notion that performance should improve as the size of the training dataset increases over several orders of magnitude. We developed a framework to integrate the top-performing metrics using PCA. To enable the integration of these metrics by PCA for newly developed models, we provide an R package, CLMeval, available at https://github.com/skinnider/CLMeval.

A limitation of our analysis is that we focused on a single family of generative models: that is, RNN-based models of textual representations. We were motivated to concentrate on this family of models because they have been arguably the most widely used in the field, and systematic benchmarks have found them to be among the best-performing models on various distribution-learning tasks[22,23]. Future efforts will be needed to understand the performance of other families of deep generative models, including other architectures adapted from natural language processing[40,54,55], as well as graph generative models[56–58], in the low-data regime.

Methods

Input data. Our experiments focused on learning generative models of molecules from four databases of chemical structures: the ZINC database of commercially available compounds[34]; the GDB-13 database, which enumerates all possible small organic molecules containing up to 13 atoms[7]; the ChEMBL database, which contains bioactive small molecules with drug-like properties[38]; and the COCONUT database of natural products[39]. Molecules from the ChEMBL database (version 24.1) were obtained from http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_24_1/chembl_24_1_chemreps.txt.gz. Molecules from the COCONUT database were obtained from the Zenodo upload accompanying the original publication at http://zenodo.org/record/3778405/files/COCONUTapril.zip. A random sample of one million molecules from the GDB-13 database[14] was obtained from the Reymond group website at http://gdbtools.unibe.ch:8080/cdn/gdb13.1M.freq.ll.smi.gz. A random sample of one million molecules was constructed from the ZINC database by first downloading each tranche separately from the ZINC website, then concatenating all 1,669 tranches into a single file and sampling from that file.

For each database, duplicate SMILES and SMILES that could not be parsed by the RDKit were removed. Salts or solvents were removed by splitting molecules into fragments and retaining only the heaviest fragment containing at least three heavy atoms, using code adapted from the Mol2vec package[59]. Charged molecules were neutralized using a list of neutralization reactions provided in the RDKit Cookbook. Molecules with atoms other than Br, C, Cl, F, H, I, N, O, P or S were removed, and molecules were converted to their canonical SMILES representations using the RDKit. Finally, SMILES strings were tokenized, and molecules containing extremely rare tokens (present in less than 0.01% of molecules in the database), as well as SMILES strings longer than 250 characters, were removed. Samples of between 1,000 and 500,000 SMILES were then drawn from the preprocessed databases. A total of 10 independent samples were drawn for each training dataset size. Variation in model performance across samples reflected both the molecules drawn from the chemical structure database and the initialization of the RNN parameters (Supplementary Fig. 2). SMILES strings were subsequently converted to DeepSMILES[60] or SELFIES[42] using versions 1.0.1 and 1.0.2 of the deepsmiles (http://github.com/baoilleach/deepsmiles) and selfies (http://github.com/aspuru-guzik-group/selfies) packages, respectively. Enumeration of non-canonical SMILES was performed using the SmilesEnumerator class available from http://github.com/EBjerrum/SMILES-enumeration, with augmentation factors of 3, 10 or 30. All of the datasets used in this work are available from Zenodo at https://doi.org/10.5281/zenodo.4641960.

We quantified the overlap between the four datasets and evaluated whether the presence of overlapping molecules between databases influenced our results. A substantial overlap was observed between the ChEMBL and COCONUT databases, but removing the overlapping molecules did not markedly affect model performance (Supplementary Fig. 3).

To evaluate the impact of the chemical diversity of the training molecules on model performance, we sampled training sets of between 1,000 and 10,000 molecules with decreasing chemical diversity. These training sets were constructed by selecting a molecule at random from one of the GDB, ChEMBL and ZINC databases, and then computing Tc between the 'founder' molecule and the remainder of the database. The database was then filtered to retain only molecules with a Tc greater than some target minimum. A Tc of zero therefore reflects random selection of molecules across the entire database, whereas an increasing Tc reflects increasing similarity to the 'founder' molecule (that is, decreasing chemical diversity). The maximum Tanimoto coefficient was set as 0.15 for GDB and 0.2 for ChEMBL and ZINC, as these were the highest thresholds at which we could reliably sample 10,000 neighbours for a randomly chosen molecule. The Tc was computed using extended connectivity fingerprint (ECFP) chemical fingerprints[61] with a diameter of six, which were selected due to their excellent performance in chemical similarity search[62,63]. The entire process was repeated 20 times, rather than 10 as done elsewhere in the manuscript, as we observed a greater degree of variability between samples, reflecting the influence of the randomly selected 'founder' molecules on the target chemical space.

Chemical language models. RNNs were trained on samples of 1,000–500,000 molecules from the four chemical structure databases, using code adapted from the REINVENT package (http://github.com/MarcusOlivecrona/REINVENT). SMILES were tokenized by considering individual characters as tokens, except atomic symbols with more than one character (Br, Cl) and environments within square brackets, such as [nH]. SELFIES were tokenized using the split_selfies function from the selfies package. The vocabulary of the RNN then consisted of all unique tokens detected in the training data, as well as start-of-string and end-of-string characters and a padding token. Except where otherwise noted, the architecture of the language models consisted of a three-layer GRU with a hidden layer of 512 dimensions, an embedding layer of 128 dimensions, and no dropout layers. Models were trained using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, with a batch size of 128 (except where otherwise noted) and a learning rate of 0.001, using teacher forcing. Ten percent of the molecules in the training set were reserved as a validation set and used to perform early stopping with a patience of 50,000 minibatches. After completion of model training, a total of 500,000 strings were sampled from each trained model. All of the code used to train CLMs is available from GitHub at http://github.com/skinnider/low-data-generative-models.

To evaluate the impact of the model architecture itself on model quality in the low-data regime, we systematically varied six model hyperparameters. These hyperparameters included the sizes of both the embedding and hidden layers, as well as the total number of hidden layers. We also compared different architectures of RNNs altogether, including GRUs, LSTMs and 'vanilla' RNNs with two different activation functions (tanh and ReLU). In addition, we experimented with adding varying amounts of dropout between each layer. Finally, to gauge whether the manner by which the models were trained could also affect performance, we varied the size of the minibatches used to train the networks[64]. To explore this larger parameter space, we limited our analysis to two of the four chemical databases, ZINC and ChEMBL, and analysed only five replicates per hyperparameter combination instead of 10.

Evaluating model performance. To quantify the performance of the trained models, we implemented Python source code to calculate a suite of 23 metrics that have previously been proposed for the evaluation of generative models of molecules. These metrics were as follows:

- The proportion of valid molecules generated by the model, where 'valid' molecules are those that can be parsed by the RDKit ('% valid').
- The proportion of novel molecules (that is, molecules not found in the training set) generated by the model ('% novel').
- The proportion of unique molecules generated by the model ('% unique').
- The internal diversity[35], defined as the mean Tc between all pairs of molecules generated by the model. Extended connectivity fingerprints[61] with a diameter of 3 and a length of 1,024 bits were used as input to the calculation of Tc. Because calculating the entire matrix of Tanimoto coefficients is prohibitive for very large numbers of molecules, a random sample of 10,000 pairs of molecules was analysed.
- The external diversity[35], defined as the mean Tc between all pairs comprising one molecule generated by the model and one molecule from the training set. Again, a random sample of 10,000 pairs of molecules was analysed rather than computing the entire matrix of Tanimoto coefficients.
- The Fréchet ChemNet distance[36] between the training and generated molecules ('FCD'). The PyTorch implementation available from http://github.com/insilicomedicine/fcd_torch was used to calculate the FCD.
- The Jensen–Shannon distances between the distributions of 17 structural or physicochemical properties, comparing molecules generated by the CLM to the molecules comprising the training dataset. These properties, and their abbreviations used in the figures, were as follows:

- The number of aliphatic rings in each molecule ('No. of aliphatic rings')
- The number of aromatic rings in each molecule ('No. of aromatic rings')
- The total number of rings in each molecule ('No. of rings')
- The proportion of rotatable bonds in each molecule ('percent rotatable bonds')
- The proportion of carbon atoms in each molecule that are sp^3 hybridized ('Percent sp^3 carbons')
- The proportion of atoms in each molecule that were stereocentres ('Percent stereocentres')
- The total proportions of each heavy atom across all molecules in the dataset ('atoms')
- The topological complexity[65] of each molecule ('Bertz TC')
- The number of hydrogen acceptors in each molecule ('No. of hydrogen acceptors')
- The number of hydrogen donors in each molecule ('No. of hydrogen donors')
- The calculated partition coefficient[66] of each molecule ('log *P*')
- The frequencies of Murcko scaffolds[67] of all molecules in the dataset ('Murcko scaffolds')
- The molecular weight of each molecule ('MW')
- The natural product-likeness score[68] for each molecule ('NP score')
- The quantitative estimate of drug-likeness (QED) score[69] for each molecule ('QED')
- The synthetic accessibility (SA) score[70] ('SA score')
- The topological polar surface area[71] of each molecule.

In addition to the Jensen–Shannon distance, we also benchmarked two other measures of differences between property distributions, the Wasserstein distance and Kullback–Leibler divergence, but found JSD was most strongly correlated to the experimental ground truth (Supplementary Fig. 4).

Code used to compute all 23 metrics is available from GitHub at http://github.com/skinnider/low-data-generative-models.

Despite the large number of metrics that have been proposed for the evaluation of generative models of molecules, there is little consensus on which should be used to gauge model quality. We initially evaluated the utility of these metrics themselves by correlating the values of each of the 23 metrics to the size of the training dataset, using the Spearman rank correlation to allow for nonlinear relationships. We reasoned that, because increasing the size of the training dataset from 1,000 to 500,000 molecules would be expected a priori to have a dramatic effect on the performance of a generative model, this analysis could allow us to benchmark the metrics themselves that have been proposed for model evaluation. Five metrics consistently achieved a Spearman correlation of $\geq$0.80 to the size of the training dataset in four different chemical databases (percent valid, FCD, percent stereocentres, Murcko and NP score). To combine information from all five top-performing metrics, while accounting for the covariance between metrics, we performed PCA on the centred and scaled matrix using the R function 'princomp'. The loadings of each model on the first principal component, PC1, were used for model evaluation. To ensure that these scores accurately captured model performance, we also inspected and visualized the proportion of valid molecules generated by each model. Pairwise comparisons of models trained with different input data or different hyperparameters were performed using a two-tailed *t*-test. The complete set of outcomes calculated for all 8,447 CLMs analysed in this study is provided as Supplementary Data 1.

We also investigated the impact of the total number of molecules sampled from the trained model on our conclusions. Throughout the main text, we draw samples of 500,000 molecules from the trained generative models. We sought to draw relatively large samples, despite the increase in computational requirements, to ensure we obtained the most representative results. However, we also investigated whether similar results could be obtained from smaller samples. To this end, we downsampled the samples of 500,000 generated molecules to obtain samples of 1,000, 5,000, 10,000, 50,000 or 100,000 molecules. We found most correlations were robust to the number of molecules sampled, but a subset of property distribution-matching metrics displayed instability with <100,000 molecules sampled, possibly because the sample size affects the number of values that some properties can take on (Supplementary Fig. 5a). The variance across models also decreased as the number of molecules sampled from the model increased (Supplementary Fig. 5b). These findings suggest that practitioners should draw as large a sample as possible during model evaluation.

Generative models of metabolomes. To train generative models of bacterial, fungal and plant metabolomes, we compiled databases of known metabolites from the following sources. Bacterial metabolites were assembled from the *E. Coli* Metabolome Database (ECMDB)[72], the *P. Aeruginosa* Metabolome Database (PAMDB)[73], StreptomeDB[74], NPASS[75] and BioCyc[76]. For the latter two databases, only molecules linked to a bacterial producing organism were retained. Plant metabolites were assembled from the Phenol-Explorer[77], PhytoHub (http://phytohub.eu/), NPASS and BioCyc databases (keeping only metabolites linked to a plant producing organism in the latter two cases). Fungal metabolites were obtained from the Yeast Metabolome Database (YMDB)[78].

We then trained a total of 48 chemical generative models on the three metabolomes. In addition to the input metabolome, we varied the RNN model (comparing LSTM and GRU architectures), the representation (comparing SMILES, DeepSMILES and SELFIES) and performed varying degrees of non-canonical SMILES enumeration (with augmentation factors of 2×, 3×, 10×, 20× or 30×). After inspecting the PC1 scores of all 48 models, as well as the values of individual metrics, we selected the three LSTM networks trained on non-canonical SMILES with the highest augmentation factor for further analysis. To visualize the global chemical space of the real and generated molecules, we computed a continuous, 512-dimensional representation of each molecule using the CDDD package[47] (available from http://github.com/jrwnter/cddd). We then sampled a matching number of real and generated metabolites, and embedded real and generated molecules from all three metabolomes into two dimensions using UMAP[79] (as implemented in the R package 'uwot'), with the following parameters: n_neighbors = 50, alpha = 2 and beta = 1.

Visualization. Throughout the manuscript, boxplots show the median (horizontal line), interquartile range (hinges) and smallest and largest values no more than 1.5 times the interquartile range (whiskers), and error bars show the standard deviation.

Data availability

Input datasets used to train chemical language models are available from Zenodo[80]. Calculated metrics for all 8,447 models discussed in this study are provided as Supplementary Data 1.

Code availability

Code used to train and evaluate chemical language models is available from GitHub at http://github.com/skinnider/low-data-generative-models[81].

Received: 19 January 2021; Accepted: 11 June 2021;
Published online: 19 July 2021

References

1. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
2. Virshup, A. M., Contreras-García, J., Wipf, P., Yang, W. & Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **135**, 7296–7303 (2013).
3. van Deursen, R. & Reymond, J.-L. Chemical space travel. *ChemMedChem* **2**, 636–640 (2007).

4. Lameijer, E.-W., Kok, J. N., Bäck, T. & Ijzerman, A. P. The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. *J. Chem. Inf. Model.* **46**, 545–552 (2006).
5. Pollock, S. N., Coutsias, E. A., Wester, M. J. & Oprea, T. I. Scaffold topologies. 1. Exhaustive enumeration up to eight rings. *J. Chem. Inf. Model.* **48**, 1304–1310 (2008).
6. Fink, T. & Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes and drug discovery. *J. Chem. Inf. Model.* **47**, 342–353 (2007).
7. Blum, L. C. & Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733 (2009).
8. Ruddigkeit, L., van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).
9. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design-a review of the state of the art. *Mol. Syst. Des. Eng* **4**, 828–849 (2019).
10. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).
11. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
12. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
13. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 48 (2017).
14. Arús-Pous, J. et al. Exploring the GDB-13 chemical space using deep generative models. *J. Cheminform.* **11**, 20 (2019).
15. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inform.* **37**, 1700153 (2018).
16. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2**, 171–180 (2020).
17. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
18. Kotsias, P.-C. et al. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2**, 254–265 (2020).
19. Li, Y., Zhang, L. & Liu, Z. Multi-objective de novo drug design with conditional graph generative model. *J. Cheminform.* **10**, 33 (2018).
20. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **9**, 10752 (2019).
21. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *Proc. 35th International Conference on Machine Learning* Vol. 80 (eds Dy, J. & Krause, A.) 2323–2332 (PMLR, 2018).
22. Brown, N., Fiscato, M., Segler, M. H. S. & Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
23. Polykovskiy, D. et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Front. Pharmacol.* **11**, 565644 (2020).
24. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
25. Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G. & Boström, J. Deep reinforcement learning for multiparameter optimization in de novo drug design. *J. Chem. Inf. Model.* **59**, 3166–3176 (2019).
26. Liu, X., Ye, K., van Vlijmen, H. W. T., IJzerman, A. P. & van Westen, G. J. P. An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A2A receptor. *J. Cheminform.* **11**, 35 (2019).
27. Neil, D. et al. Exploring deep recurrent models with reinforcement learning for molecule design. In *Proc. 6th International Conference on Learning Representations* (ICLR, 2018).
28. Amabilino, S., Pogány, P., Pickett, S. D. & Green, D. V. S. Guidelines for recurrent neural network transfer learning-based molecular generation of focused libraries. *J. Chem. Inf. Model.* **60**, 5699–5713 (2020).
29. Gupta, A. et al. Generative recurrent networks for de novo drug design. *Mol. Inform.* **37**, 1700111 (2018).
30. Awale, M., Sirockin, F., Stiefl, N. & Reymond, J.-L. Drug analogs from fragment-based long short-term memory generative neural networks. *J. Chem. Inf. Model.* **59**, 1347–1356 (2019).
31. Merk, D., Grisoni, F., Friedrich, L. & Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Commun. Chem.* **1**, 68 (2018).
32. Renz, P., Van Rompaey, D., Wegner, J. K., Hochreiter, S. & Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discov. Today Technol.* **32–33**, 55–63 (2019).
33. Arús-Pous, J. et al. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminform.* **11**, 71 (2019).
34. Irwin, J. J. & Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).
35. Benhenda, M. Can AI reproduce observed chemical diversity? Preprint at *bioRxiv* https://doi.org/10.1101/292177 (2018).
36. Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. & Klambauer, G. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *J. Chem. Inf. Model.* **58**, 1736–1741 (2018).
37. van Deursen, R., Ertl, P., Tetko, I. V. & Godin, G. GEN: highly efficient SMILES explorer using autodidactic generative examination networks. *J. Cheminform.* **12**, 22 (2020).
38. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
39. Sorokina, M. & Steinbeck, C. Review on natural products databases: where to find data in 2020. *J. Cheminform.* **12**, 20 (2020).
40. Sanchez-Lengeling, B., Outeiral, C., Guimaraes, G. L. & Aspuru-Guzik, A. Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). Preprint at https://doi.org/10.26434/chemrxiv.5309668.v3 (2017).
41. O'Boyle, N. & Dalke, A. DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. Preprint at https://doi.org/10.26434/chemrxiv.7097960 (2018).
42. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **1**, 045024 (2020).
43. Kusner, M. J., Paige, B. & Hernandez-Lobato, J. M. Grammar variational autoencoder. Preprint at https://arxiv.org/pdf/1703.01925.pdf (2017).
44. Dai, H., Tian, Y., Dai, B., Skiena, S. & Song, L. Syntax-directed variational autoencoder for structured data. Preprint at https://arxiv.org/pdf/1802.08786.pdf (2018).
45. Bjerrum, E. J. SMILES enumeration as data augmentation for neural network modeling of molecules. Preprint at https://arxiv.org/pdf/1703.07076.pdf (2017).
46. Bjerrum, E. J. & Sattarov, B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* **8**, 131 (2018).
47. Winter, R., Montanari, F., Noé, F. & Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10**, 1692–1701 (2019).
48. Zhang, Q. et al. Structural investigation of ribosomally synthesized natural products by hypothetical structure enumeration and evaluation using tandem MS. *Proc. Natl Acad. Sci. USA* **111**, 12031–12036 (2014).
49. Johnston, C. W. et al. An automated genomes-to-natural products platform (GNP) for the discovery of modular natural products. *Nat. Commun.* **6**, 8421 (2015).
50. Zheng, S. et al. QBMG: quasi-biogenic molecule generator with deep recurrent neural network. *J. Cheminform.* **11**, 5 (2019).
51. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl Acad. Sci. USA* **112**, 12549–12550 (2015).
52. Vanhaelen, Q., Lin, Y.-C. & Zhavoronkov, A. The advent of generative chemistry. *ACS Med. Chem. Lett.* **11**, 1496–1505 (2020).
53. Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous discovery in the chemical sciences part II: outlook. *Angew. Chem. Int. Ed.* **59**, 23414–23436 (2020).
54. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
55. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* **14**, 3098–3104 (2017).
56. Samanta, B. et al. NEVAE: a deep generative model for molecular graphs. *J. Mach. Learn. Res.* **21**, 1–33 (2020).
57. Mercado, R. et al. Practical notes on building molecular graph generative models. *Appl. AI Lett.* https://doi.org/10.1002/ail2.18 (2020).
58. De Cao, N. & Kipf, T. MolGAN: an implicit generative model for small molecular graphs. Preprint at https://arxiv.org/pdf/1805.11973.pdf (2018).
59. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **58**, 27–35 (2018).
60. O'Boyle, N. & Dalke, A. DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. Preprint at https://doi.org/10.26434/chemrxiv.7097960.v1 (2018).
61. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
62. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.* **8**, 36 (2016).

63. Skinnider, M. A., Dejong, C. A., Franczak, B. C., McNicholas, P. D. & Magarvey, N. A. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. *J. Cheminform.* **9**, 46 (2017).
64. Smith, S. L., Kindermans, P.-J. & Le, Q. V. Don't decay the learning rate, increase the batch size. Preprint at https://arxiv.org/pdf/1711.00489.pdf (2017).
65. Bertz, S. H. The first general index of molecular complexity. *J. Am. Chem. Soc.* **103**, 3599–3601 (1981).
66. Wildman, S. A. & Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).
67. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
68. Ertl, P., Roggo, S. & Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **48**, 68–74 (2008).
69. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
70. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 8 (2009).
71. Ertl, P., Rohde, B. & Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **43**, 3714–3717 (2000).
72. Sajed, T. et al. ECMDB 2.0: a richer resource for understanding the biochemistry of *E. coli*. *Nucleic Acids Res.* **44**, D495–D501 (2016).
73. Huang, W. et al. PAMDB: a comprehensive *Pseudomonas aeruginosa* metabolome database. *Nucleic Acids Res.* **46**, D575–D580 (2018).
74. Moumbock, A. F. A. et al. StreptomeDB 3.0: an updated compendium of streptomycetes natural products. *Nucleic Acids Res* **49**, D600–D604 (2020).
75. Zeng, X. et al. NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* **46**, D1217–D1222 (2018).
76. Karp, P. D. et al. The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* **20**, 1085–1093 (2019).
77. Neveu, V. et al. Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. *Database (Oxford)* **2010**, bap024 (2010).
78. Ramirez-Gaona, M. et al. YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Res.* **45**, D440–D445 (2017).
79. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. Preprint at https://arxiv.org/pdf/1802.03426.pdf (2018).
80. Molecules used to train generative models (Zenodo, 2021); https://doi.org/10.5281/zenodo.4641960
81. Python source code used to train and evaluate generative models of molecules (Zenodo, 2021); https://doi.org/10.5281/zenodo.4642099

Acknowledgements

This work was supported by funding from Genome Canada, Genome British Columbia and Genome Alberta (project nos. 284MBO and 264PRO). Computational resources were provided by WestGrid, Compute Canada and Advanced Research Computing at the University of British Columbia. M.A.S. acknowledges support from a CIHR Vanier Canada Graduate Scholarship, a Roman M. Babicki Fellowship in Medical Research, a Borealis AI Graduate Fellowship, a Walter C. Sumner Memorial Fellowship and a Vancouver Coastal Health–CIHR–UBC MD/PhD Studentship. We thank J. Liigand and F. Wang for helpful discussions.

Author contributions

M.A.S., D.S.W. and L.J.F. designed experiments. M.A.S. and R.G.S. performed experiments. M.A.S. wrote the manuscript. All authors edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-021-00368-1.

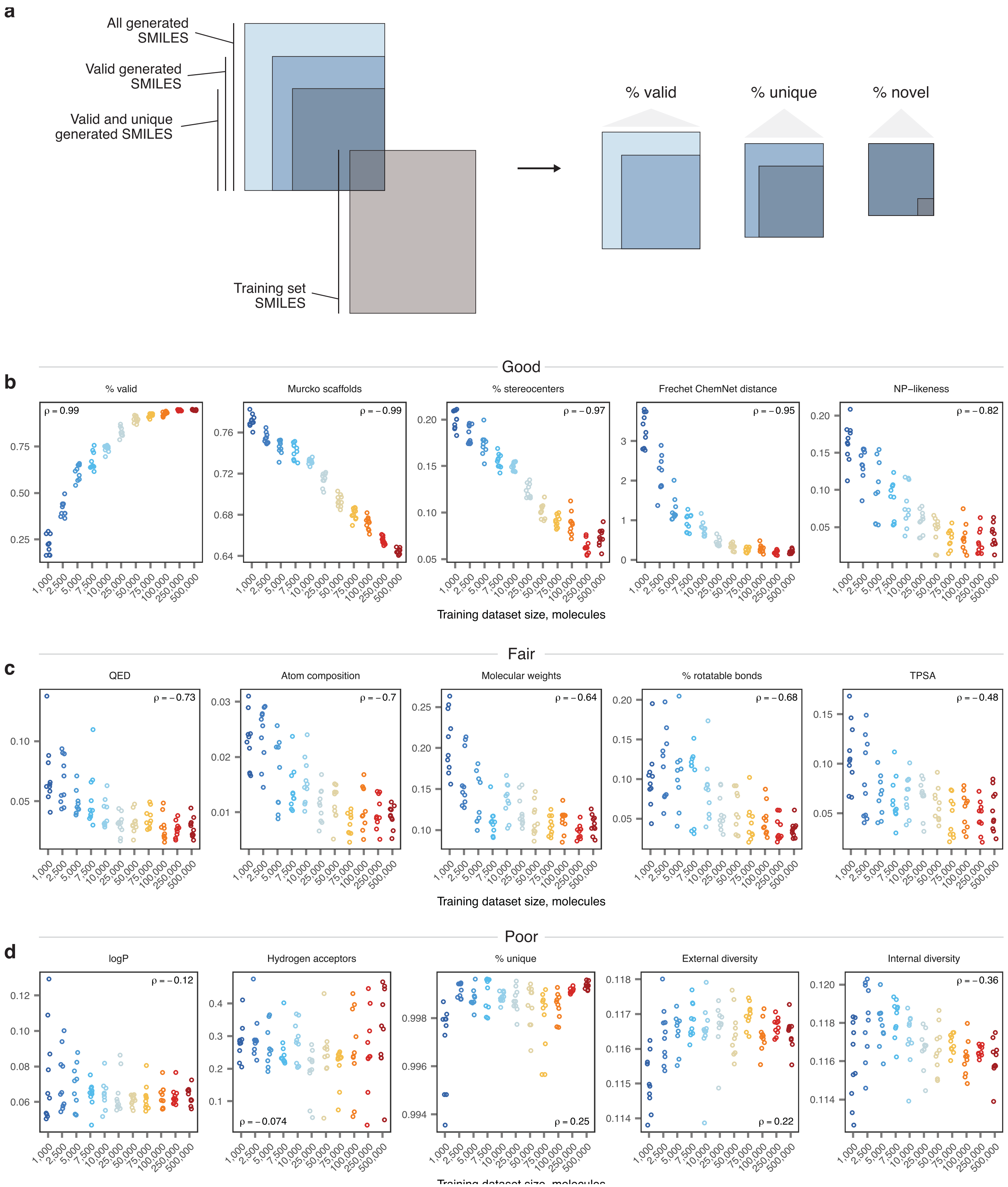
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00368-1.

Correspondence and requests for materials should be addressed to M.A.S. or L.J.F.

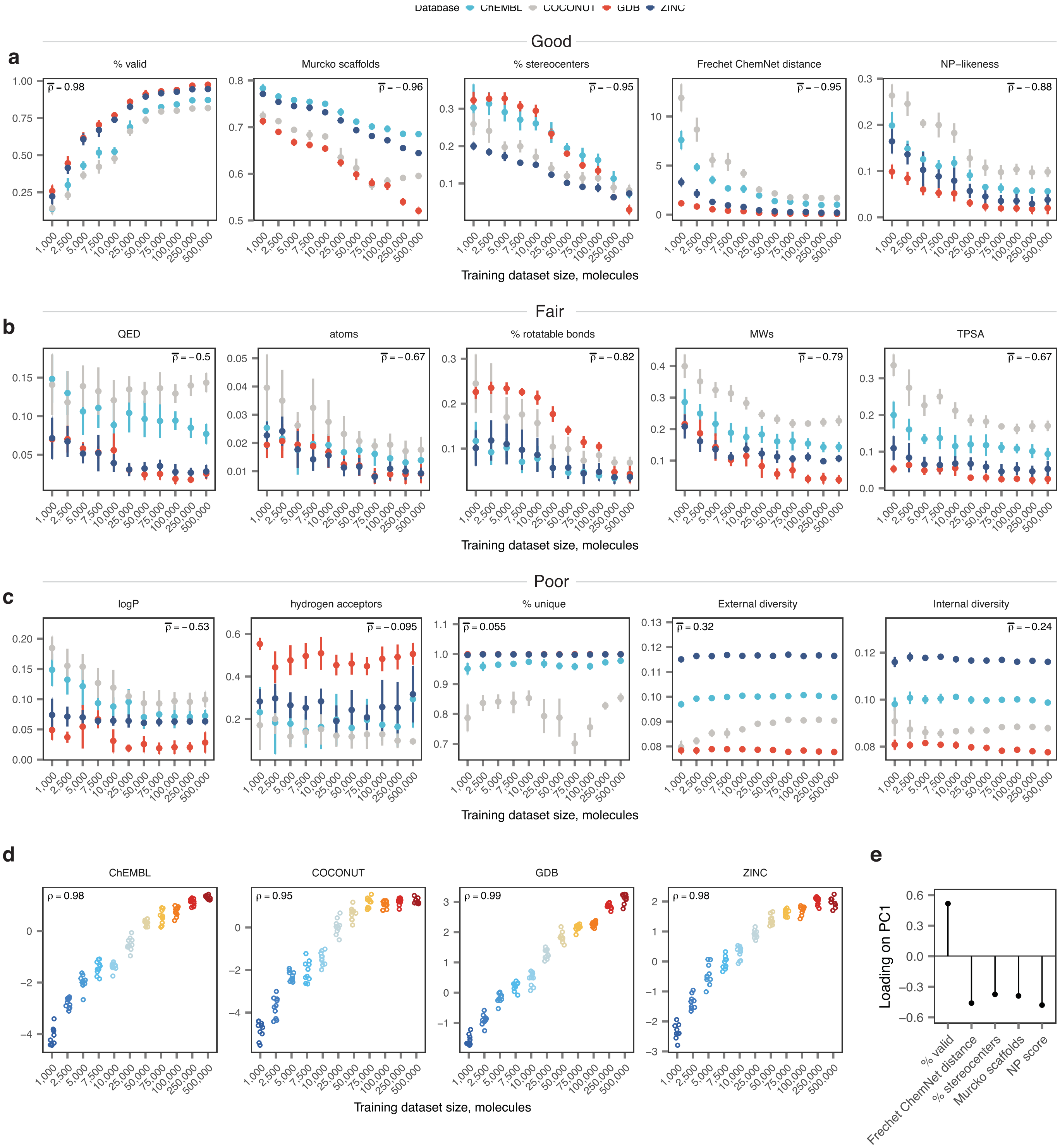
Peer review information *Nature Machine Intelligence* thanks Sebastian Raschka and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

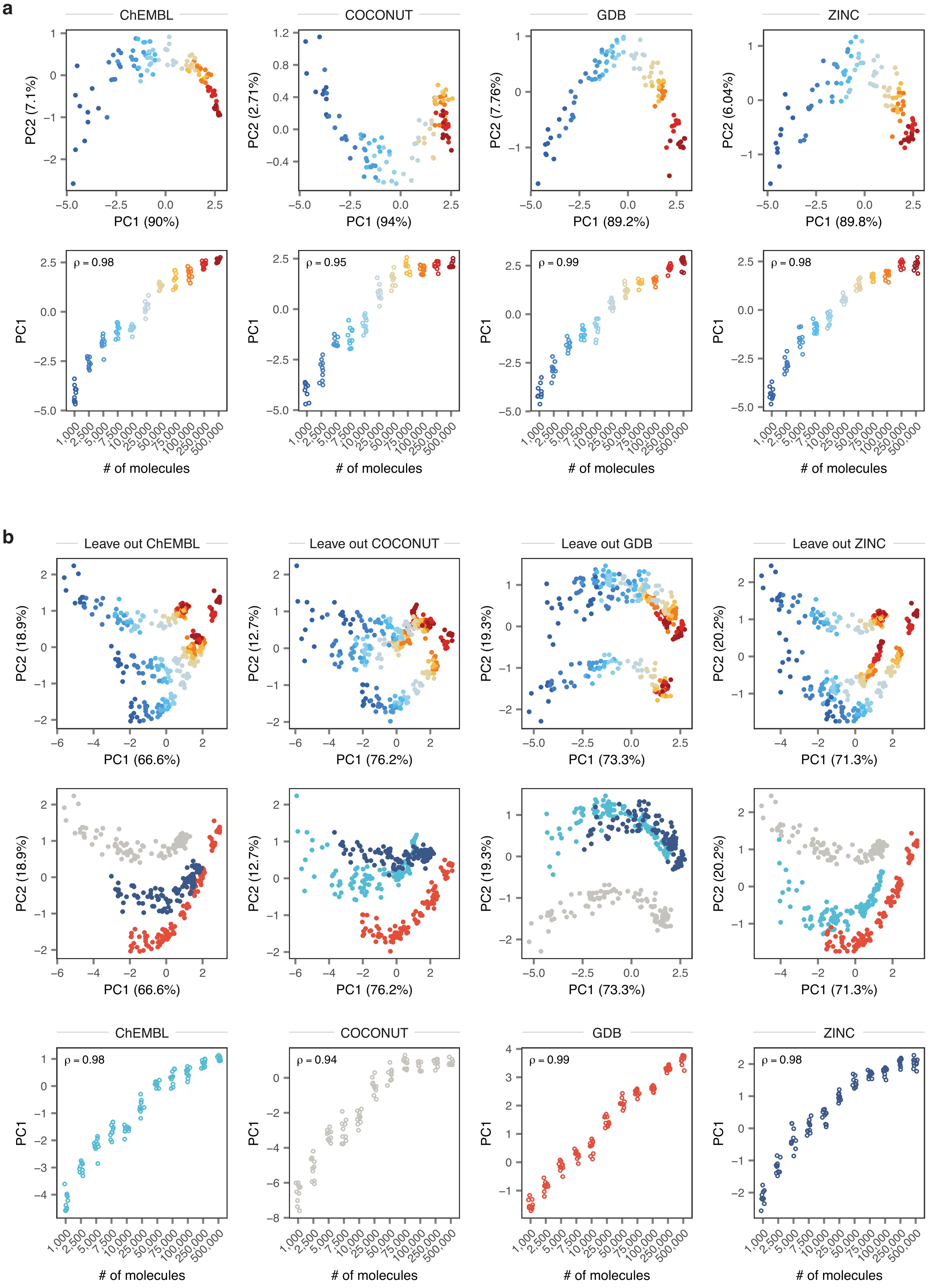
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Evaluating low-data generative models of purchasable chemical space. a, Schematic overview of the '% valid', '% unique', and '% novel' metrics. **b**, Values of the five top-performing metrics with the strongest correlations ($\rho \geq 0.82$) to training dataset size for $n = 110$ generative models trained on varying numbers of molecules from the ZINC database. **c**, Values of five exemplary metrics with moderate to weak correlations ($0.48 \leq \rho \leq 0.73$) to training dataset size for $n = 110$ generative models trained on varying numbers of molecules from the ZINC database. **d**, Values of five exemplary metrics with little or no correlation ($\rho \leq 0.36$) to training dataset size for $n = 110$ generative models trained on varying numbers of molecules from the ZINC database.

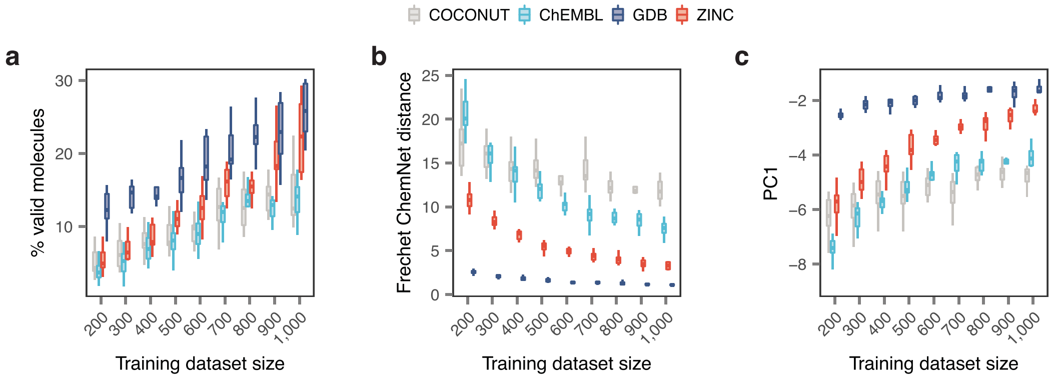


Extended Data Fig. 2 | Evaluating low-data generative models of divergent chemical spaces. a, Values of the five top-performing metrics with the strongest correlations (average rank correlation ≥ 0.80) to training dataset size for *n* = 440 generative models trained on varying numbers of molecules from the ChEMBL, COCONUT, GDB, or ZINC databases. Points and error bars show the mean and standard deviation, respectively, of ten independent replicates. **b**, Values of five exemplary metrics with moderate to weak correlations to training dataset size for *n* = 440 generative models trained on varying numbers of molecules from the ChEMBL, COCONUT, GDB, or ZINC databases. **c**, Values of five exemplary metrics with little or no correlation to training dataset size for *n* = 440 generative models trained on varying numbers of molecules from the ChEMBL, COCONUT, GDB, or ZINC databases. **d**, PC1 scores for *n* = 440 chemical language models trained on varying numbers of molecules sampled from the ChEMBL, COCONUT, GDB, or ZINC databases. Inset text shows the Spearman correlation. **e**, Factor loadings onto the first principal component in a PCA of *n* = 440 chemical language models trained on varying numbers of molecules sampled from the ChEMBL, COCONUT, GDB, or ZINC databases.

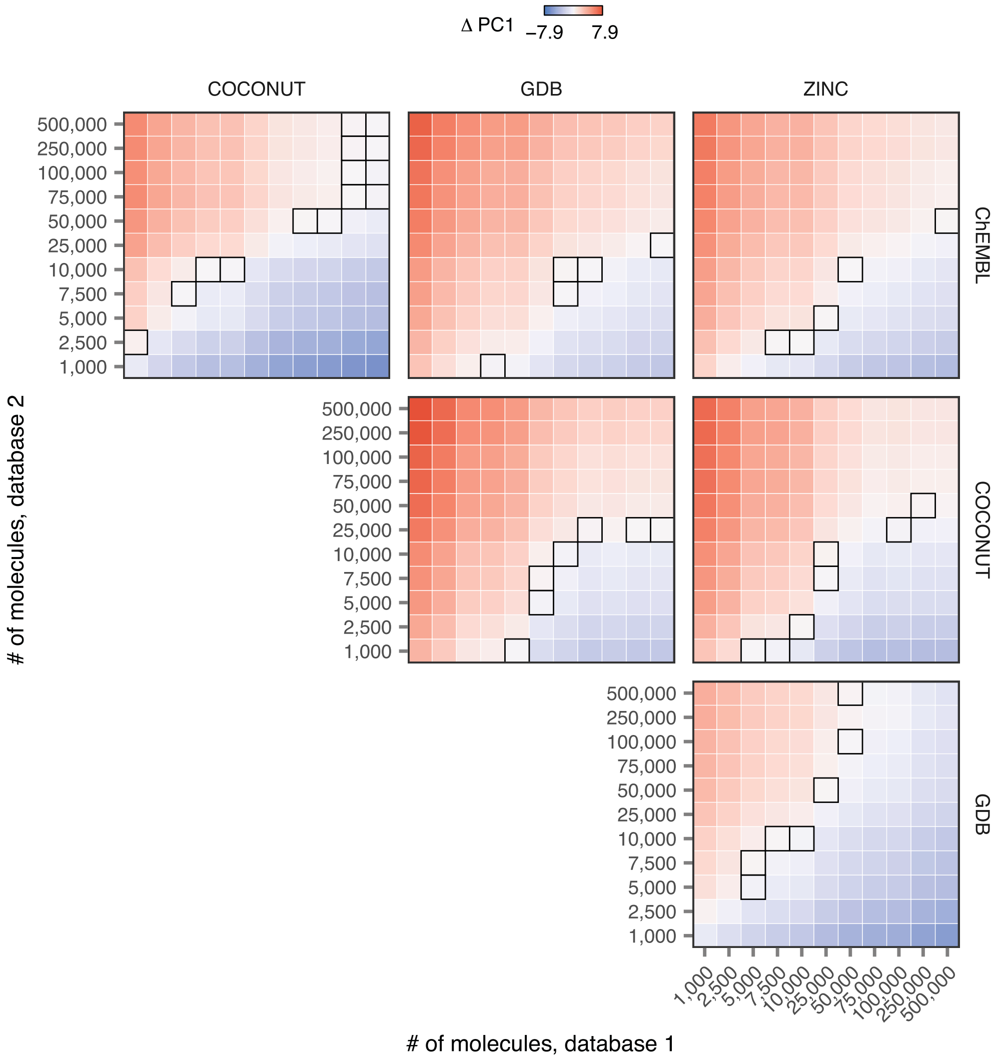


Extended Data Fig. 3 | See next page for caption.

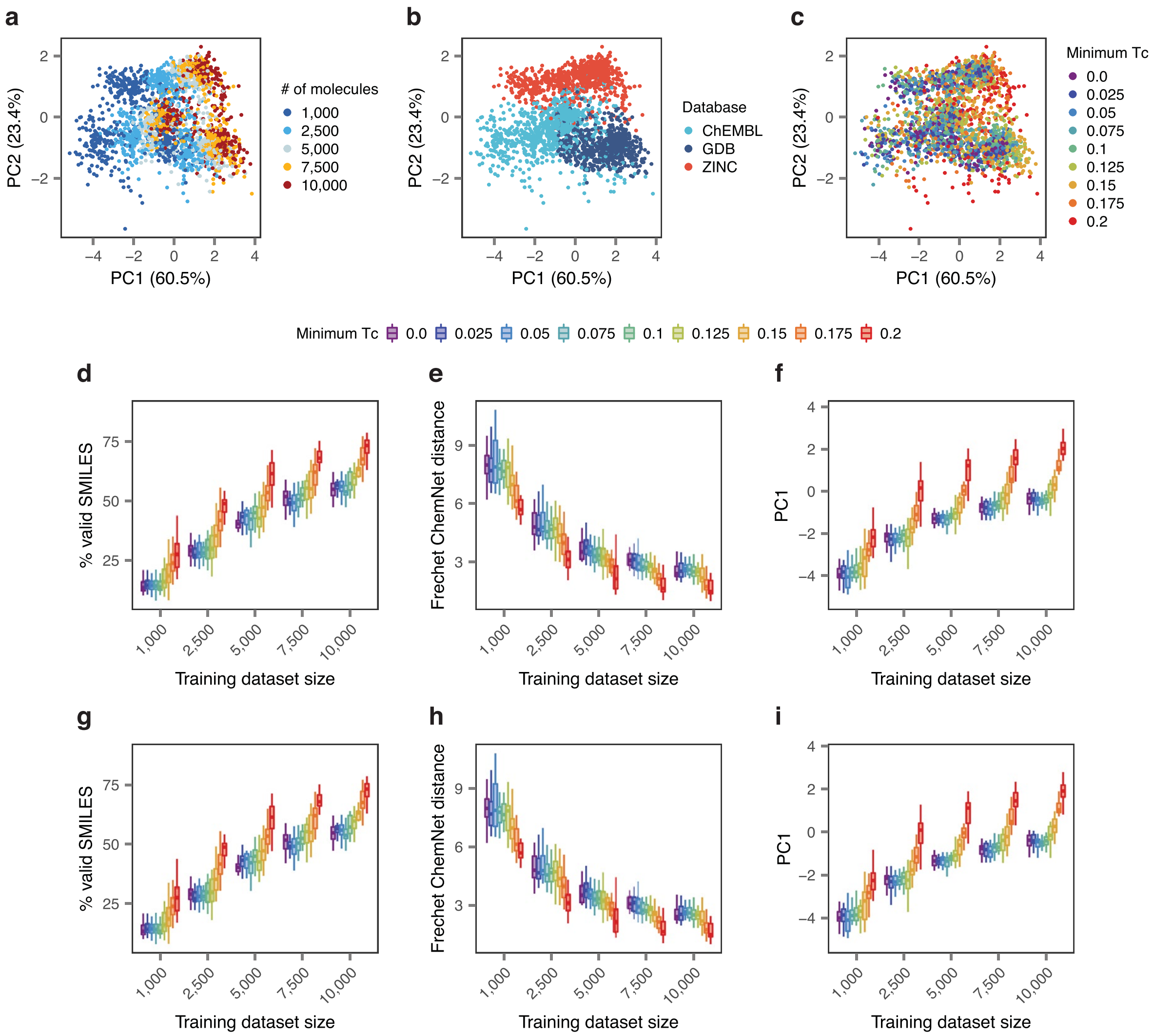
Extended Data Fig. 3 | Robustness of principal component analysis for the evaluation of chemical generative models. a, PCA of top-performing metrics, top, and PC1 scores, bottom, for chemical language models trained on varying numbers of molecules sampled from the ChEMBL, COCONUT, GDB, and ZINC database, with PCA performed separately for each database. Bottom, inset text shows the Spearman correlation. **b**, PCA of top-performing metrics for chemical language models trained on varying numbers of molecules sampled from three of four databases, colored by the size of the training dataset, top, or the chemical database on which the generative models were trained, middle. Bottom, PC1 scores for models trained on the withheld database, projected onto the coordinate basis of the other three databases. Inset text shows the Spearman correlation.



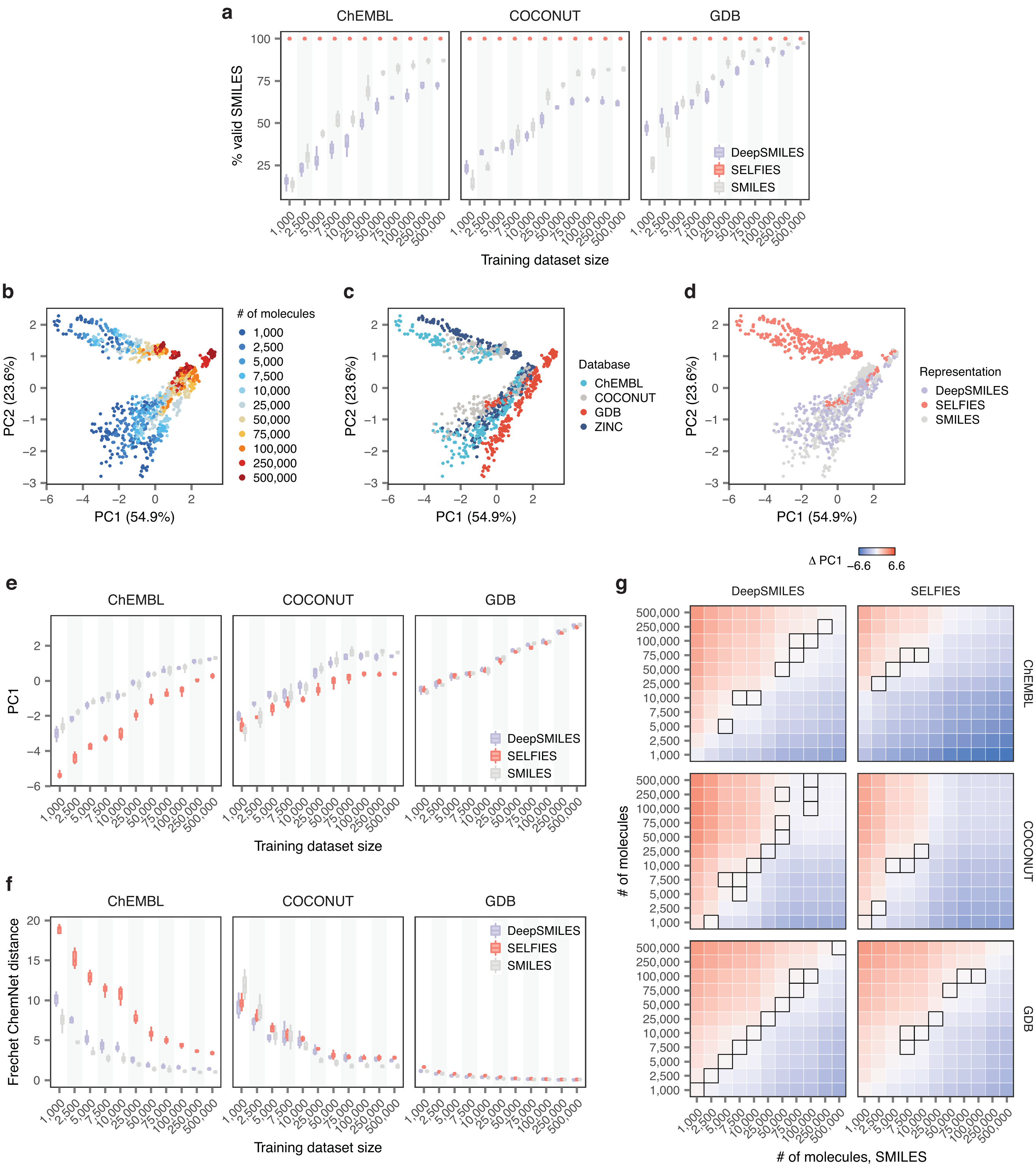
Extended Data Fig. 4 | Learning chemical language models from less than 1,000 examples. a, Proportion of valid SMILES generated by chemical language models trained on samples of between 200 and 1,000 molecules from one of four chemical databases. **b**, Fréchet ChemNet distance of chemical language models trained on samples of between 200 and 1,000 molecules from one of four chemical databases. **c**, PC1 scores of chemical language models trained on samples of between 200 and 1,000 molecules from one of four chemical databases.



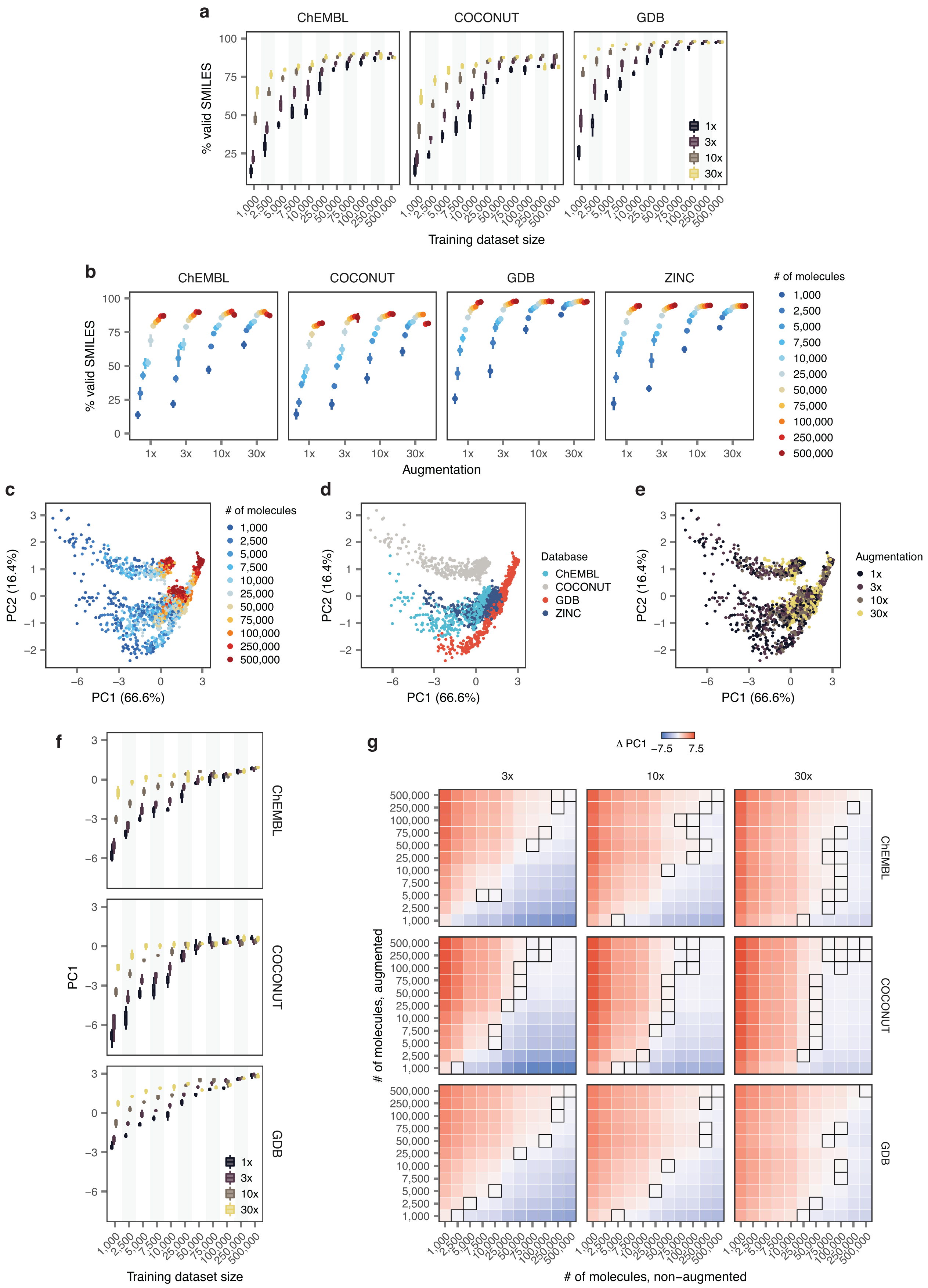
Extended Data Fig. 5 | Training dataset size requirements in different chemical spaces. Mean difference in PC1 scores between chemical language models trained on varying numbers of molecules sampled from each pair of chemical structure databases. Dark squares indicate pairs without statistically significant differences (uncorrected p > 0.05, two-sided t-test).



Extended Data Fig. 6 | Low-data generative models of diverse and homogeneous molecules from the ChEMBL and ZINC databases. a, PCA of top-performing metrics for molecules generated by chemical language models trained on varying numbers of more or less diverse molecules from the GDB, ChEMBL, and ZINC databases, colored by the size of the training dataset. **b**, As in **a**, but colored by the chemical database on which the generative models were trained. **c**, As in **a**, but colored by the diversity (minimum Tanimoto coefficient to a randomly selected 'founder' molecule). **d-i**, Performance of chemical language models trained on samples of molecules from the ChEMBL (**d-f**) and ZINC (**g-i**) databases with a minimum Tanimoto coefficient (Tc) to a randomly selected 'founder' molecule. **d**, Proportion of valid SMILES generated by chemical language models trained on varying numbers of more or less diverse molecules from the ChEMBL database. **e**, Fréchet ChemNet distances of chemical language models trained on varying numbers of more or less diverse molecules from the ChEMBL database. **f**, PC1 scores of chemical language models trained on varying numbers of more or less diverse molecules from the ChEMBL database. **g**, Proportion of valid SMILES generated by chemical language models trained on varying numbers of more or less diverse molecules from the ZINC database. **h**, Fréchet ChemNet distances of chemical language models trained on varying numbers of more or less diverse molecules from the ZINC database. **i**, PC1 scores of chemical language models trained on varying numbers of more or less diverse molecules from the ZINC database.

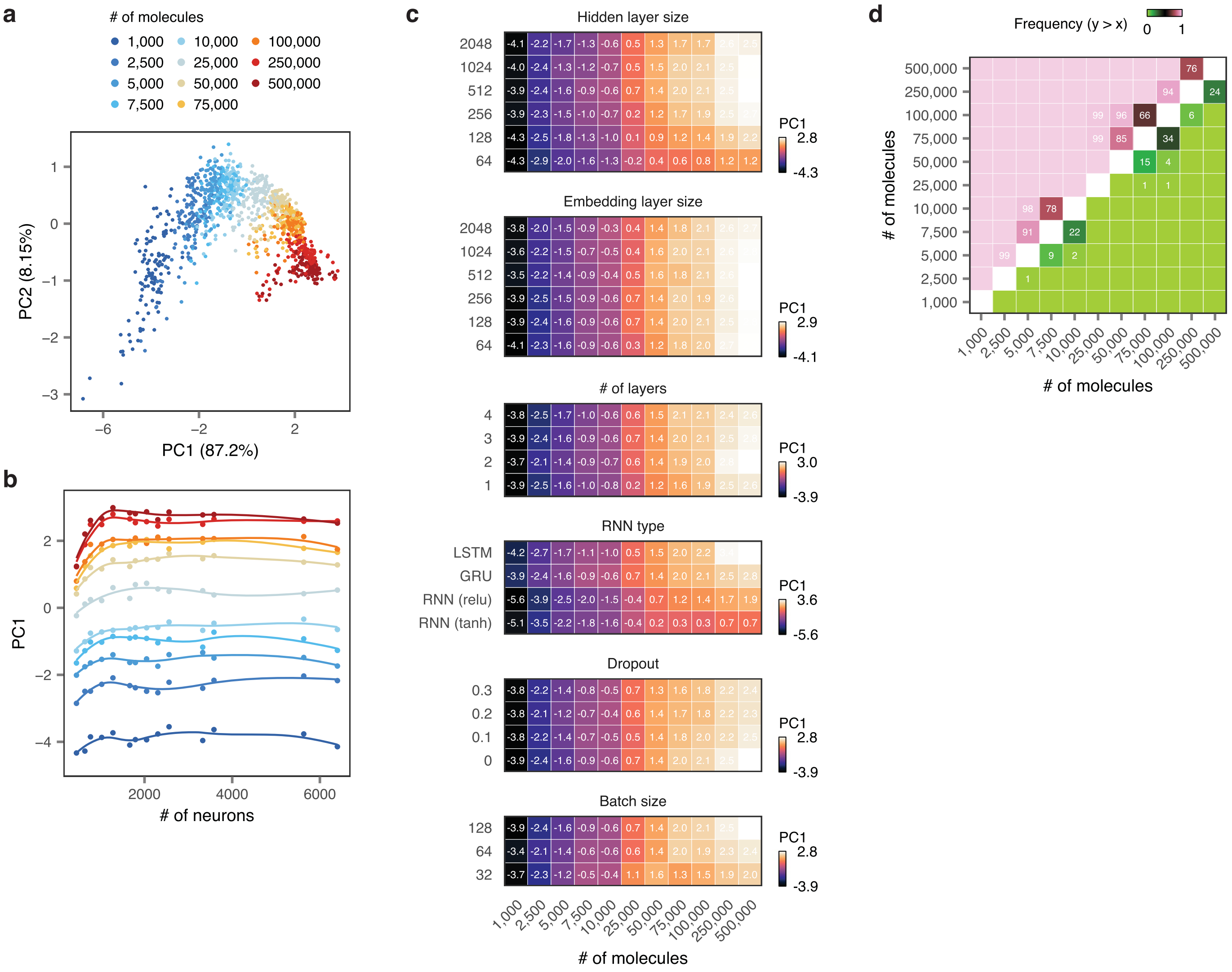


Extended Data Fig. 7 | Evaluating alternative molecular representations for low-data generative models in distinct chemical spaces. a, Proportion of valid SMILES generated by chemical language models trained on one of three string representations of molecules from the ChEMBL, COCONUT, and GDB databases. **b**, PCA of top-performing metrics for molecules generated by *n* = 1,320 chemical language models trained on one of three string representations of molecules from the ChEMBL, COCONUT, and GDB databases, colored by the size of the training dataset. **c**, As in **b**, but colored by the chemical database on which the generative models were trained. **d**, As in **b**, but colored by molecular representation. **e**, PC1 scores of chemical language models trained on one of three string representations of molecules from the ChEMBL, COCONUT, and GDB databases. **f**, Fréchet ChemNet distances of chemical language models trained on one of three string representations of molecules from the ChEMBL, COCONUT, and GDB databases. **g**, Mean difference in PC1 scores between chemical language models trained on varying numbers of molecules sampled from the ChEMBL, COCONUT, and GDB databases, represented either as DeepSMILES or SELFIES, y-axis, or SMILES, x-axis. Dark squares indicate pairs without statistically significant differences (uncorrected $p > 0.05$, two-sided t-test).

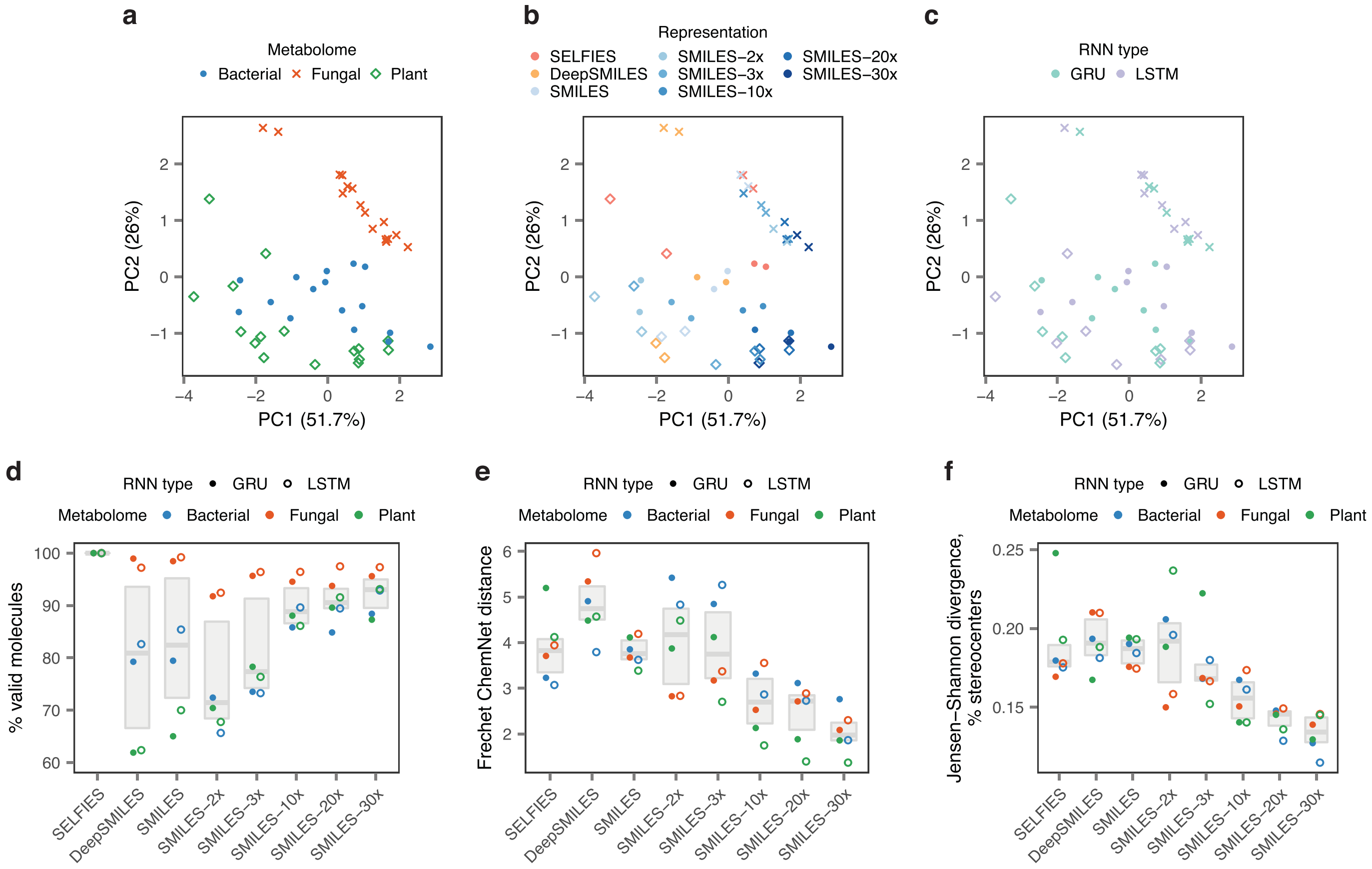


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Data augmentation by non-canonical SMILES enumeration. **a**, Proportion of valid SMILES generated by chemical language models trained on molecules from the ChEMBL, COCONUT, and GDB databases after varying degrees of non-canonical SMILES enumeration. **b**, Data as in **a** and Fig. 3i, but showing the relationship between the size of the training dataset and the proportion of valid SMILES generated by models for each degree of non-canonical SMILES enumeration separately. **c**, PCA of top-performing metrics for molecules generated by $n = 1,760$ chemical language models trained on molecules from the ChEMBL, COCONUT, and GDB databases after varying degrees of non-canonical SMILES enumeration, colored by the size of the training dataset. **d**, As in **c**, but colored by the chemical database on which the generative models were trained. **e**, As in **c**, but colored by the amount of SMILES enumeration. **f**, PC1 scores of chemical language models trained on molecules from the ChEMBL, COCONUT, and GDB databases after varying degrees of non-canonical SMILES enumeration. **g**, Mean difference in PC1 scores between chemical language models trained on molecules from the ChEMBL, COCONUT, and GDB databases represented as canonical SMILES, x-axis, or non-canonical SMILES after varying degrees of data augmentation, y-axis. Dark squares indicate pairs without statistically significant differences (uncorrected $p > 0.05$, two-sided t-test).



Extended Data Fig. 9 | Hyperparameter tuning in the ChEMBL database. a, PCA of top-performing metrics for molecules generated by $n = 1,210$ chemical language models, trained on varying numbers of molecules from the ChEMBL database with varying model hyperparameters, colored by the size of the training dataset. **b**, Mean PC1 scores of chemical language models as a function of the total number of neurons in the model. Solid lines show local polynomial regression. **c**, Mean PC1 scores for molecules trained on the ChEMBL database, as a function of both the number of molecules in the training dataset, x-axis, and varying hyperparameters, y-axis. The mean of five independent replicates is shown. **d**, Proportion of $n = 110$ chemical language models with varying hyperparameters, trained on the number of molecules shown on the y-axis, that outperformed a model without hyperparameter tuning trained on the number of molecules shown on the x-axis.



Extended Data Fig. 10 | Optimizing generative models of bacterial, fungal, and plant metabolomes. a, PCA of top-performing metrics for molecules generated by $n = 48$ chemical language models, trained on bacterial, fungal, or plant metabolomes with varying inputs and hyperparameters, colored by the target metabolome. **b**, As in **a**, but colored by the molecular representation and data augmentation strategy. **c**, As in **a**, but colored by the RNN architecture. **d**, Proportion of valid molecules produced by generative models of metabolomes trained with different molecular representations (SMILES, DeepSMILES, or SELFIES), data augmentation strategies (non-canonical SMILES enumeration with an augmentation factor of between 2x and 30x), and RNN architectures (GRU or LSTM). **e**, As in **d**, but showing the Fréchet ChemNet distance between generated and real metabolites. **f**, As in **d**, but showing the Jensen-Shannon distance of the proportion of stereocenters between generated and real metabolites.