Molecular Fingerprints for Robust and Efficient ML-Driven Molecular Generation

Ruslan N. Tazhigulov, Joshua Schiller, Jacob Oppenheim, Max Winston
EQRx, Inc.
50 Hampshire St., Cambridge, MA 02139, USA
{rtazhigulov, jschiller, joppenheim, mwinston}@eqrx.com

Abstract

We propose a novel molecular fingerprint-based variational autoencoder applied for molecular generation on real-world drug molecules. We define more suitable and pharma-relevant baseline metrics and tests, focusing on the generation of diverse, drug-like, novel small molecules and scaffolds. When we apply these molecular generation metrics to our novel model, we observe a substantial improvement in chemical synthetic accessibility ($\Delta\overline{\mathrm{SAS}}$ = -0.83) and in computational efficiency up to 5.9x in comparison to an existing state-of-the-art SMILES-based architecture.

1 Introduction

Machine learning (ML) driven molecular generation is becoming an industry standard among computational and medicinal chemists for exploration of chemical space and drug design. Since the 1980s, the biopharmaceutical industry has used ML methods on two primary tasks in drug development: generating novel molecules and predicting their properties (e.g. quantitative structure activity relationship, QSAR). The deep learning revolution has enabled these tasks to move from disparate strategies, to models based on the embedding of chemical structures in a high dimensional space, either implicitly or explicitly. We focus on the first task here, as the second task remains limited in applicability by the size of training datasets outside a handful of well-studied molecular sets and properties.

Representation of the chemical structures of the molecules can be split into two major classes: 1D string-based representations, and 2D/3D molecular graphs. While 2D/3D graphical representations can directly capture the connectivity between atoms, graphs are provably more difficult and slow to generate than 1D string sequences as graph generation is known to be an NP-hard problem [1, 2]. The most common type of 1D string representation of molecules is SMILES (Simplified Molecular Input Line Entry System) [3]. This simplistic string-based representation transforms a molecule into a sequence of characters based on the set of predefined atom ordering rules. Using 1D strings for molecular representation has proven beneficial, enabling the reapplication of existing neural network architectures that have been previously developed for natural language processing (NLP) tasks.

In principle, there is hardly any limit on representation of the molecules with SMILES strings. While one molecule can map to its multiple equivalent SMILES representations, typically, only one canonical SMILES representation is used. Arús-Pous et al. [4] showed that augmenting training data with randomized versions of SMILES strings leads to generalization to larger drug-like chemical spaces than solely canonical SMILES (single representation), and increases model performance on generative tasks. A model trained with randomized SMILES was able to generate at least double the amount of unique molecules compared to models based on canonical SMILES.

Although several various model architectures have been developed (VAEs[5, 6], cRNN[7], GAN[8], etc), a lack of meaningful standardized metrics and benchmarks for molecular generation makes it difficult to evaluate and compare them for their actual application in the drug discovery process. For

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

example, despite the increasing prevalence of molecular generation in the pharmaceutical industry, performance in the literature largely continues to be measured by on-the-fly metrics such as reconstruction accuracy. While these on-the-fly metrics are important in the training and testing processes, ultimately metrics capturing computationally-efficient generation of novel chemical matter are more relevant. Given the utility of structural chemical variation in modern drug discovery, capturing unique scaffolds[9] and structural similarity of generated molecules within metrics is vital for the field to advance.

Building on these results, we investigated the use of molecular fingerprints for generating novel chemical matter instead of using SMILES as an input. Unlike SMILES representations, any molecular structure can be deterministically mapped to its fingerprint. Previously it had been assumed that molecular fingerprints are completely uninvertible due to the complex rules of construction and the hash mapping used in implementation. However, Le et al. [10] have shown that circular extended connectivity fingerprints (ECFP) can be at least partially invertible. This presented an opportunity for use in molecular generators since circular molecular fingerprints represent ideal encoders: they are fast to compute, contain a high amount of information, and are related to underlying molecular graphs by design. Given all of these advantages, we sought to build a novel fingerprint-based architecture that could outperform existing state-of-the-art SMILES-based molecular generators.

Here, we set out to benchmark and synthesize existing literature approaches, build best-in-class models, and propose a novel molecular fingerprint-based variational autoencoder (VAE) for molecular generation. In addition to the common on-the-fly metrics used in the literature, we benchmark the implemented models with the computational efficiency of the molecular generation process, determined by the number of structurally unique molecules that a model has generated. Furthermore, in an attempt to highlight the most relevant chemical novelty, we focus on characterizing novel chemical matter with respect to chemical distance from reference target molecules. We thus attempt to define pharma-relevant benchmarks and tests for molecular generation focused on the generation of drug-like novel small molecules and scaffolds.

2 Methods

Molecular ChemVAE: SMILES-based variational autoencoder. In the present work, we trained a molecular ChemVAE model that is descended from the one presented by Gómez-Bombarelli et al. [5]. We applied numerous architecture and training optimizations suggested in the ML molecular generation literature. Among them, we benchmarked train/test reconstruction accuracy (see Appendix A.1), while still focusing on molecular generation capabilities of the models. Note that while the original ChemVAE was trained on ZINC and QM9 datasets [5], we opted for ChEMBL [11], a manually curated database of bioactive molecules with drug-like properties. Our cleaned-up dataset contained 1.83M of unique and valid, from SMILES syntax perspective, molecules. 90 percent of the molecules were used for training, with the 10 percent allocated for testing. For ease of computation, we put a cap of 200 on the maximum length of the SMILES strings. We also implemented the approach discussed by Arús-Pous et al. [4], randomizing the original canonical SMILES strings on the fly. This procedure results in a different SMILES string for the same molecule in each epoch. More details on model parameters and hyperparameters can be found in Appendix A.1.

Molecular FpVAE: Fingerprint / SMILES-based variational autoencoder. Our Fp / SMILES-based variational autoencoder (FpVAE) also used an entire ChEMBL dataset. We obtained circular Morgan fingerprints (non-count version, nBits = 4096, radius = 2), an analogous open-source version of the ECFP fingerprints, for each molecule in the dataset. We have not observed any significant benefits from increasing the number of bits or radius parameter, as Le et al. [10] also observed and indicated for the ECFP fingerprints in their work. The same 90/10 split was used between the training and testing sets. We opted for the canonical SMILES representation of the molecules for training of the FpVAE model. More details on model parameters and hyperparameters can be found in Appendix A.2.

MG-Bridge: SLERP-based navigation in chemical latent space for molecular generation. A high-dimensional chemical latent space generated by the ML models tends to be substantially sparse. Effective sampling of the sparse chemical latent space is becoming a cornerstone of ML-driven molecular generation process. For example, one can imagine that perturbing a latent space vector with a Gaussian noise can potentially produce a new molecule.

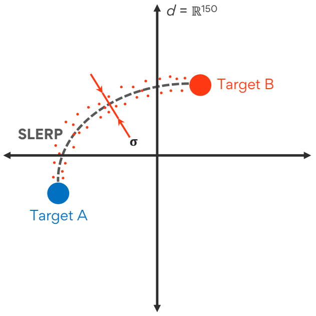

Figure 1: Molecular generation bridge (MG-Bridge).

However, the optimal perturbation noise that allows one to explore the latent space efficiently and produce a substantial number of unique and novel molecules per computational time remains unclear. Moreover, drug discovery researchers typically focus on a few (typically a pair) of active compounds, not just a single compound. Hence, effective techniques that allow one to explore the latent space between multiple active compounds are needed. Gómez-Bombarelli et al. [5] introduced the idea of sampling along the trajectories between a pair of reference molecules in the latent space to generate novel molecules based on linear or spherical linear interpolation (LERP/SLERP). Here, we introduce a molecular generation procedure named as MG-Bridge (Molecular Generation Bridge). The procedure combines SLERP and perturbation of latent space vectors with a Gaussian noise (σ) into one method (Fig. 1, more details in Appendix A.3). MG-Bridge enables efficient molecular generation and exploration of the chemical latent space between a pair of active reference compounds.

3 Results and Discussion

We benchmarked the molecular generation performance of both existing molecular ChemVAE and a new FpVAE models for four reference pairs of drug molecules. Outside of NSAIDs, they reflect recent and current areas of drug development spanning different levels of molecular complexity (Table 1). We applied chemoinformatic filters to the obtained molecular generation sets based on quantitative estimate of drug-likeliness (QED) [12], synthetic accessibility score (SAS) [13], and novel generic Bemis-Murcko (BM) scaffolds [9]. We incorporated these filters in our benchmarks.

Table 1: Reference pair molecules used for benchmarking of VAE models

Class	Target A	Target B
NSAID	Ibuprofen	Naproxen
EGFR	Gefitinib	Erlotinib
VEGFR	Pazopanib	Sunitinib
PI3K	Alpelisib	Inavolisib

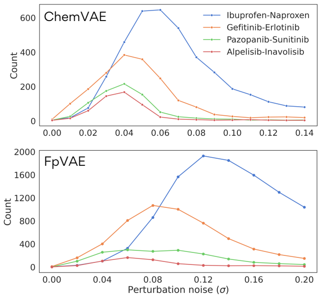

Figure 2: MG-Bridge scan for optimal perturbation noise.

Our results indicate the importance of introducing perturbation in SLERP and using MG-Bridge. We show that SLERP has to be enhanced for the real-world applications, because a pure application of SLERP led to a handful (O(10)) of unique and novel molecules (Fig. 2, $\sigma = 0.0$) across all four reference pairs. Once perturbation is applied, it becomes clear that the effect of perturbation and its size varies for different molecular pairs. The results also indicate different levels of sparsity in the chemical latent space across all pairs and various model architectures. One should note that the novelty–a metric that is typically used to account for molecules not present in the original training set–of molecules generated by the MG-Bridge is >99% for the obtained valid (from SMILES perspective) and unique molecules.

Furthermore, our results show that the FpVAE model significantly outperforms the ChemVAE model in molecular generation computational efficiency[1] by as large as a factor of 5.9, even more notably when physicochemical property (PC filter: QED ≥ 0.4, SAS ≤ 4.0) and novel BM (NBM) scaffold filters are applied (Fig. 3). The PC filter is aimed to filter out the molecules that are not drug-like or hard to synthesize. The NBM filter is aimed to keep the molecules with the generic BM scaffolds not belonging to the original reference molecules, and therefore, it aims to incorporate the chemical structural diversity.

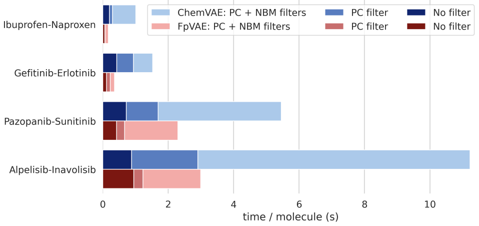

Figure 3: Efficiency of molecular generation and key performance metrics across four reference target pairs.

[1] NVIDIA Tesla T4, 16 CPUs, 60GB RAM

Moreover, by more rigorously examining the physicochemical properties of generated molecules, we demonstrate that the molecular FpVAE outperforms the molecular ChemVAE model. FpVAE is able to generate more sensible, drug-like and synthesizable molecules across all four reference pairs (Fig. 4A). These results imply that the FpVAE learns chemical rules substantially more efficiently than the SMILES-based ChemVAE. Furthermore, they also indicate that circular molecular fingerprints have a better inductive bias compared to SMILES, demonstrating the efficacy of fingerprints for molecular generation. Since the same number of latent space vectors is sampled (with the optimal perturbation noise) for both FpVAE and ChemVAE models, we also show that significantly larger number of unique and valid molecules are generated with the FpVAE model (Fig. 4A).

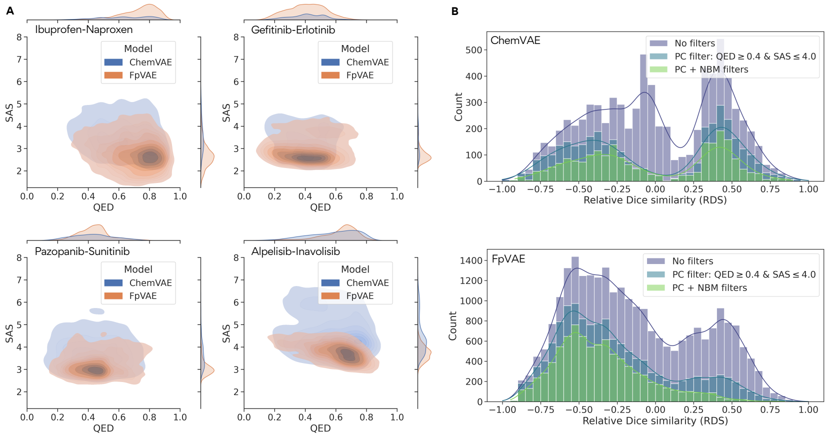

Figure 4: (A) QED-SAS properties computed for the sets obtained with four reference pair molecules. The larger the QED and the smaller the SAS scores (bottom right) indicates better results. (B) RDS computed for the molecular generation sets obtained for EGFR inhibitors (Gefitinib-Erlotinib pair).

We also evaluated structural diversity in the molecular generation sets by computing the relative Dice similarity, $RDS_i = \frac{d_{Bi} - d_{Ai}}{1 - d_{AB}}$, where d is Dice similarity between two Morgan fingerprints corresponding to two different molecules. RDS $\in$ [-1.0, 1.0], where the negative values indicate that a generated molecule is chemically similar to target A, while the positive values indicate that a molecule is chemically similar to target B. The results in Fig. 4B indicate that the molecular FpVAE model is able to generate a substantially larger number of unique and valid molecules compared to ChemVAE for the optimal perturbation noise and with the same number of latent space vectors provided for decoding. Moreover, molecules generated by FpVAE are more drug-like and more likely synthesizable, i.e. more of them pass the PC filter. Furthermore, when it comes to chemical diversity in the generated sets, the FpVAE model is able to generate significantly larger number of molecules with NBM scaffolds. One should note that the most chemically differentiated and diverse molecules lay in the interval of RDS $\in$ (-0.3, 0.3), and the FpVAE model generates a substantially larger number of molecules in this interval. These results also illustrate that the chemical latent space, generated by both ChemVAE and FpVAE models, is not really homogeneous and the sparsity varies, e.g. it is easier to generate more chemically sensible and diverse molecules in the vicinity of gefitinib vs. erlotinib. The sparsity and compactness of the chemical latent spaces will be addressed in the future work.

4 Conclusion

In the present work, we proposed and discussed the performance of a novel variational autoencoder based on the molecular fingerprints (FpVAE). Importantly, we also proposed a set of baseline pharma-relevant benchmarks for comparison of molecular generation performance across various model architectures, methods, molecules, and filters that can be applied for unbiased comparative analysis. These benchmarks put a focus on the number of drug-like, synthesizable, and structurally diverse molecules generated per computational time. We showed that the FpVAE model is able to generate molecules more efficiently compared to the purely SMILES-based ChemVAE model. We also highlighted that the FpVAE generation sets obtained for various reference drug molecule pairs representing different classes outperform the ones obtained with the purely SMILES-based model in chemical synthetic accessibility and structural diversity.

5 Broader Impact

The ability to efficiently generate novel drug-like molecules at scale has the potential to accelerate drug discovery and become a valuable tool in tackling societal challenges such as drug pricing and bringing new medicines at a fraction of the cost of today's leading therapies. This work offers both a novel methodological approach (FpVAE) for more efficient generation of novel drug-like molecules, as well as new suggested benchmarks and metrics to better guide future innovation in molecular generation. The potential negative impact includes the methods and approaches being used for creating and producing harmful molecules.

References

[1] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, and K. F. Jensen, "Generative models for molecular discovery: Recent advances and challenges," *WIREs Computational Molecular Science*, p. e1608, 2022.

[2] J.-G. Sobez and M. Reiher, "Molassembler: Molecular graph construction, modification, and conformer generation for inorganic and organic molecules," *Journal of Chemical Information and Modeling*, vol. 60, no. 8, pp. 3884–3900, 2020.

[3] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[4] J. Arús-Pous, S. V. Johansson, O. Prykhodko, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen, and O. Engkvist, "Randomized smiles strings improve the quality of molecular generative models," *Journal of cheminformatics*, vol. 11, no. 1, pp. 1–13, 2019.

[5] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Central Science*, vol. 4, no. 2, pp. 268–276, 2018.

[6] R. Winter, F. Montanari, F. Noé, and D.-A. Clevert, "Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations," *Chem. Sci.*, vol. 10, pp. 1692–1701, 2019.

[7] P.-C. Kotsias, J. Arús-Pous, H. Chen, O. Engkvist, C. Tyrchan, and E. J. Bjerrum, "Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks," *Nature Machine Intelligence*, vol. 2, no. 5, pp. 254–265, 2020.

[8] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, and A. Aspuru-Guzik, "Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models," 2017. [Online]. Available: https://arxiv.org/abs/1705.10843

[9] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. molecular frameworks," *Journal of Medicinal Chemistry*, vol. 39, no. 15, pp. 2887–2893, 1996.

[10] T. Le, R. Winter, F. Noé, and D.-A. Clevert, "Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures," *Chem. Sci.*, vol. 11, pp. 10 378–10 389, 2020.

[11] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. Magariños, J. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. Radoux, A. Segura-Cabrera, A. Hersey, and A. Leach, "ChEMBL: towards direct deposition of bioassay data," *Nucleic Acids Research*, vol. 47, no. D1, pp. D930–D940, 11 2018.

[12] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, "Quantifying the chemical beauty of drugs," *Nature chemistry*, vol. 4, no. 2, pp. 90–98, 2012.

[13] P. Ertl and A. Schuffenhauer, "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions," *Journal of cheminformatics*, vol. 1, no. 1, pp. 1–11, 2009.

[14] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, and G. Schneider, "Generative recurrent networks for de novo drug design," *Molecular Informatics*, vol. 37, no. 1-2, p. 1700111, 2018.

[15] X. Li and D. Fourches, "Smiles pair encoding: A data-driven substructure tokenization algorithm for deep learning," *Journal of Chemical Information and Modeling*, vol. 61, no. 4, pp. 1560–1569, 2021.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017. [Online]. Available: https://arxiv.org/abs/1711.05101

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We are still in the process of making the code available to the public.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The error bars with respect to the random seed are very small compared with the performance difference of the models.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] ChEMBL and ZINC are both public datasets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] No, the datasets are publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Molecular ChemVAE model: details on parameters and hyperparameters

Data preparation. We used ChEMBL-28 database [11] (named as ChEMBL-1.83M in this work). For our early exploration, we also used a subset of ChEMBL-28 database, namely ChEMBL-567k. For ChEMBL-567k, we only included the active compounds with IC50/EC50 values ≥ 6, similar to Gupta et al. [14], and we restricted the range for the length of included SMILES strings between 10 and 100 for computational purposes. Note that Gómez-Bombarelli et al. [5] used a subset of ZINC dataset (namely ZINC-249k in Table 2) for training of the original ChemVAE model.

We canonicalized the SMILES strings using RDKit package. We treated compounds as racemates, i.e. any information about stereochemistry (represented by "@", "/", and "\ \" tokens) was removed. We also removed all ionic non-bonded components of the original SMILES strings (represented by "."). For ChEMBL-1.83M, we restricted the maximal length to 215 to account for equivalent randomized SMILES representations that are longer than their original canonical SMILES counterparts. The shorter strings were padded with spaces to this same length.

SMILES tokenizer. On top of multiple equivalent representations, SMILES strings can be tokenized in various ways, and more advanced tokenizers have been emerging. In the present work, we relied on the atomwise tokenizer from SMILES Pair Encoding (SPE) package [15]. SPE package is licensed under Apache-2.0 license. SPE aims to augment the widely used atom-level SMILES tokenization by adding chemically explainable and human-readable SMILES substrings as tokens. Case studies showed that SPE could demonstrate superior performances on both molecular generation tasks with regard to novelty, diversity, and ability to resemble the training set distribution [15]. Therefore, our encoding used 193 tokens, which included auxiliary tokens obtained during the randomized SMILES procedure.

Model. The VAE deep network structure of our molecular ChemVAE model was as follows (Fig. 5): the encoder used three 1D convolutional layers of 9, 9, 10 convolution kernels with the kernel sizes of 9, 9, 11, respectively, followed by two fully connected layers with dimensions of 435 and 150. The latter one corresponds to the dimension of the chemical latent space vectors, with the vector values bounded between -1 and 1 by applying tanh activation function. Batch normalization was enforced at the encoder level after each convolution layer. The decoder included three stacked layers of long short-term memory (LSTM) networks[16] with the hidden dimension of 512. The last layers included a fully-connected layer of the vocabulary dimension, followed by the softmax layer with the argmax being applied. The variational (KL) loss was annealed according to sigmoid schedule after 20 epochs, running for a total 100 epochs, and He-initialization was applied to all fully-connected layers. The batch size was set to 256. AdamW optimizer[17] with the learning rate of 0.0001 and weight decay of 0.001 was used for backpropagation, and we also used a learning rate scheduler: the learning rate was decreased by a factor of 0.7 if the token-by-token reconstruction accuracy on the testing set was not improved within the next 5 epochs. Both cross-enthropy and KL losses were averaged over all tokens that combined into a total loss. The best performance model based on the highest token-by-token reconstruction accuracy for the testing set was chosen further for molecular generation tasks.

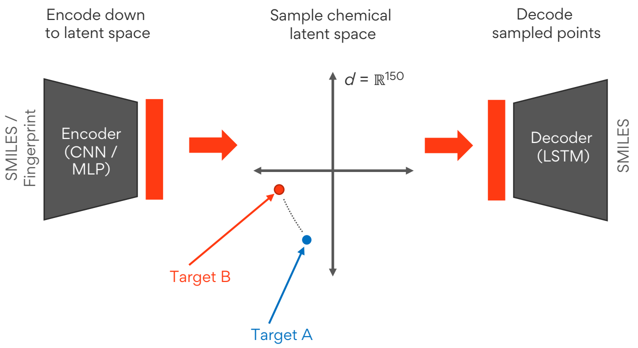

Figure 5: A schematic representation of variational autoencoders trained and applied for molecular generation in the present work.

A.2 Molecular FpVAE model: details on parameters and hyperparameters

SMILES tokenizer. For tokenization of the output, we used the same atomwise tokenizer [15] that we also applied for the molecular ChemVAE model dataset. Encoding on the virtually entire ChEMBL dataset (canonical SMILES) used 184 tokens.

Table 2: Token-by-token accuracy obtained for the test sets of molecular ChemVAE and FpVAE models trained with different datasets

Architecture	Dataset	Token-by-token
ChemVAE	ZINC-249k	0.972
ChemVAE	ChEMBL-567k	0.978
ChemVAE	ChEMBL-1.83M	0.996
FpVAE	ChEMBL-567k	0.878
FpVAE	ChEMBL-1.83M	0.948

Model. The encoder part contained five fully-connected layers with the dimensions of 2048, 1024, 768, and 512, respectively, with the last layer defining the dimension of the latent space, 150 (Fig. 5). Our early results also demonstrated no gain in performance when the dropout for fully-connected layers in the encoder was applied. We set the dropout rates to 0.05, 0.10, 0.15, and 0.20, and similarly to what Le et al. [10] indicated in their work, we have not observed any further improvements from using dropout regularization (Table 3). The decoder part contained three stacked layers of LSTM networks[16] with the hidden dimension of 512. The last layers included a fully-connected layer of the vocabulary dimension, followed by the softmax layer with the argmax being applied. The variational (KL) loss was annealed according to the same sigmoid schedule after 20 epochs, running for a total 200 epochs, and He-initialization was applied to all fully-connected layers. The batch size was set to 256. AdamW optimizer[17] with the learning rate of 0.0005 and weight decay of 0.005 was used for backpropagation, and we also used a learning rate scheduler: the learning rate was decreased by a factor of 0.7 if the token-by-token reconstruction accuracy on the testing set was not improved within the next 5 epochs. Both cross-enthropy and KL losses were averaged over all tokens that combined into a total loss. The best performance model based on the highest token-by-token reconstruction accuracy for the testing set was chosen further for molecular generation tasks.

Table 3: Token-by-token, molecular reconstruction, and Tanimoto accuracy (computed between corresponding Morgan fingerprints) metrics obtained for the test sets of FpVAE models (ChEMBL-567k dataset) trained with different dropout rates in fully connected layers. Numbers in the parentheses are computed for the decoded valid SMILES only

Dropout rate	Token-by-token	Molecule	Tanimoto
0.00	0.878	0.141 (0.382)	0.293 (0.794)
0.05	0.873	0.092 (0.299)	0.232 (0.753)
0.10	0.862	0.055 (0.228)	0.171 (0.710)
0.15	0.851	0.034 (0.187)	0.126 (0.692)
0.20	0.836	0.014 (0.116)	0.075 (0.620)

A.3 MG-Bridge: details on chemical latent space sampling

The coarse-grained MG-Bridge runs were performed as follows: 100 grid points along the SLERP trajectory between molecular target A and target B were perturbed 100 times, resulting in 10000 latent space vectors that were then decoded into SMILES. For each reference molecular pair, we scanned across multiple parameters of the perturbation noise (σ) to obtain an optimal perturbation noise per pair (Fig. 2). Once the optimal noise σ was obtained, we used the optimized value for production MG-Bridge runs: 100 grid points along the SLERP trajectory were perturbed 5000 times, resulting in 500000 latent space vectors that were then decoded into SMILES strings.