MolGene-E: Inverse Molecular Design to Modulate Single Cell Transcriptomics

Anonymous Authors[1]

Abstract

Designing drugs that can restore a diseased cell to its healthy state is an emerging approach in systems pharmacology to address medical needs that conventional target-based drug discovery paradigms have failed to meet. Single-cell transcriptomics can comprehensively map the differences between diseased and healthy cellular states, making it a valuable technique for systems pharmacology. However, single-cell omics data is noisy, heterogeneous, scarce, and high-dimensional. As a result, no machine learning methods currently exist to use single-cell omics data to design new drug molecules. We have developed a new deep generative framework named MolGene-E that can tackle this challenge. MolGene-E combines two novel models: 1) a cross-modal model that can harmonize and denoise chemical-perturbed bulk and single-cell transcriptomics data, and 2) a CLIP-VAE-based generative model that can generate new molecules based on the transcriptomics data. MolGene-E consistently outperforms baseline methods in generating high-quality, hit-like molecules from gene expression profiles obtained from single-cell datasets and gene expressions induced by knockign out targets using CRISPR. This superior performance is demonstrated across de novo molecule generation metrics, including novelty, diversity, uniqueness, and synthesizability. This makes it a potentially powerful new tool for drug discovery.

1. Introduction

Capitalizing on the success of generative Artificial Intelligence (AI) across various domains such as natural language, images, and videos, generative AI techniques have been extensively applied to the generation of small organic compounds targeting a specific disease gene for drug discovery (Zeng et al., 2022). However, this one-drug-one-target paradigm has had limited success in tackling polygenic, multifactorial diseases. Due to the high costs, prolonged development timelines, and low success rates associated with target-based drug discovery, there has been a resurgence of interest in phenotypic drug discovery. As a matter of fact, approximately 90% of approved drugs have been discovered through a phenotype-driven approach (Vincent et al., 2022). Therefore, phenotype-based molecular generation, also known as inverse molecule design, holds promise for the discovery of novel therapeutics aimed at addressing medical needs that conventional target-based drug discovery paradigms have failed to meet.

The effectiveness of phenotype-based drug discovery relies upon the careful selection of an appropriate phenotype readout. Chemical-induced transcriptomics has been embraced as a comprehensive systematic measurement for phenotype drug discovery. The transcriptomic change resulting from chemical exposure can function as a chemical signature for predicting drug responses as well as aid in the elucidation of drug targets and the inference of drug-modulated pathways. This approach has demonstrated successful applications in phenotype drug repurposing (Salame et al., 2022) (Pham et al., 2021). Several deep learning methods have been proposed to leverage chemical-induced bulk gene expression data for inverse molecule design. Notably, MolGAN (Méndez-Lucio et al., 2020a) generates molecules conditioning a generative adversarial network with bulk transcriptomics data. Although it shows promising results, GANs are susceptible to scalability, as we show that their performance drops significantly when trained on higher dimensional data. Furthermore, GANs have a black-box nature, and inferring the relation between the condition (gene expression) and generation (molecules) is quite cumbersome. Another recent work is the GxVAE (Li & Yamanishi, 2024), which employs two joint variational autoencoders to facilitate the extraction of latent gene expression features and using it as a condition to generate molecules using a second VAE. However, GxVAE has not been developed for single-cell transcriptomics data.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2024 AI for Science workshop. Do not distribute.

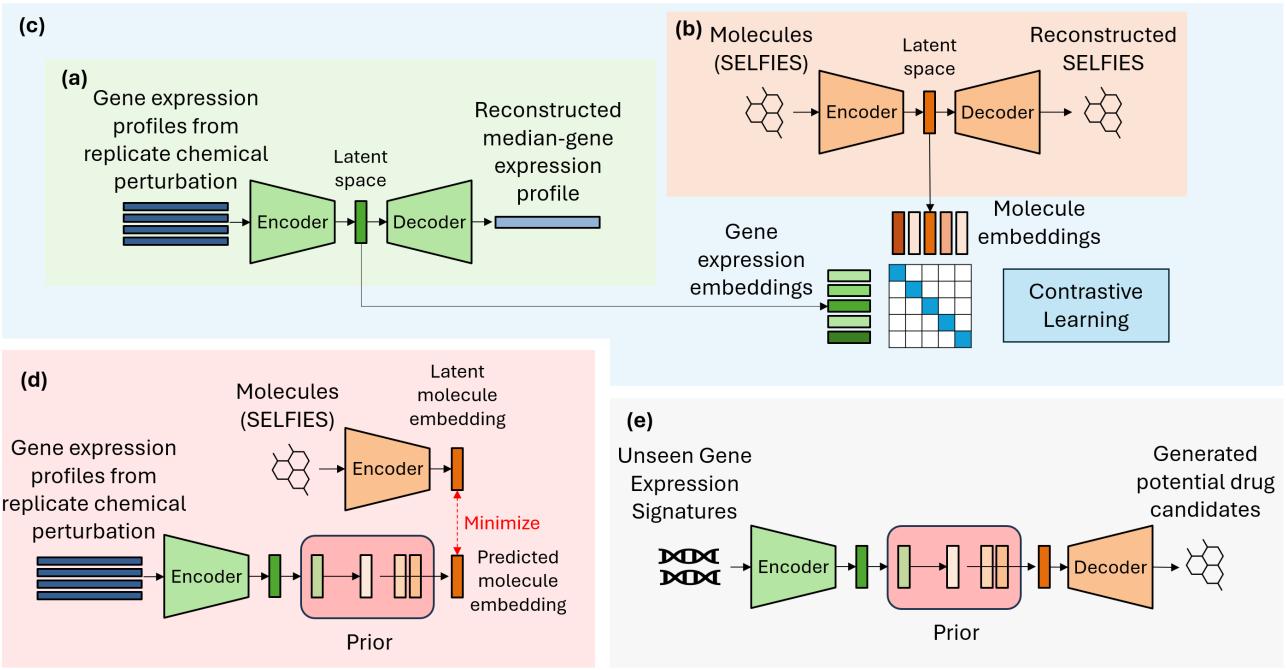

Figure 1. **(a)** A Variational Autoencoder (VAE) denoises the gene expression profiles corresponding to molecule replicates in the batch by reconstructing median gene expression profiles. **(b)** We represent chemical structures via SELFIES and use a pretrained frozen VAE to extract the chemical representations. **(c)** The gene expression encoder is fine-tuned to align gene representations to the chemical representations via a CLIP module. A supervised contrastive learning objective is optimized to maximize the agreement between positive pairs while minimizing the agreement between negative pairs. **(d)** A prior model is trained to map the inferred CLIP embeddings of gene expression to the inferred CLIP embeddings of chemicals. **(e)** Gene expression embeddings mapped to the chemical space are decoded using the SELFIES VAE's decoder to generate novel molecule candidates conditioned on gene expression profiles.

The abundance of single-cell omics data provides new opportunities for phenotype-based drug discovery. single-cell transcriptomics data offer new insights into disease heterogeneity within and across species, illuminating the complexity of pathological processes. An effective therapy often needs to modulate disease etiology at the single-cell level (Han et al., 2022). Furthermore, precise characterization of single-cell chemical transcriptomics is crucial to bridge translational gaps between disease models (e.g., organoids and animals) and human patients, a critical bottleneck in drug discovery (Van de Sande et al., 2023). Nonetheless, there remains a scarcity of methods for leveraging single-cell transcriptomics data in inverse molecule design.

Compared with protein structures that exhibit a relatively clean nature, omics data is plagued by its high-dimensionality and susceptibility to noise, stemming from biological stochasticity and technical artifacts. These complexities pose hurdles for single-cell inverse molecule design, exacerbated by the limited availability of chemical-perturbed single-cell transcriptomics data. LINCS1000 (Subramanian et al., 2017) serves as a comprehensive chemical transcriptomics database, profiling 19,811 chemicals across 77 cell lines. However, this database profiles only 978 landmark genes. Moreover, the gene expression data in LINCS1000 is obtained using a specific imaging technique, leading to significant distributional discrepancies from RNA-seq data. Due to these challenges, no methods exist for inverse molecule design based on single-cell omics data.

To address these challenges, we introduce MolGene-E, a deep learning framework for single-cell molecule generation. The key contributions of MolGene-E are twofold: First, we develop a domain adaptation model that is capable of harmonizing and denoising L1000toRNAseq and Sciplex-3 single-cell chemical transcriptomics data. Second, we design a generative algorithm that leverages contrastive learning to align phenotypic representations to chemical representations, by integrating these components, MolGene-E facilitates the generation of novel molecules with specific phenotypic traits. Extensive evaluations demonstrate that MolGene-E achieves state-of-the-art performance, positioning it as a potentially powerful new AI tool for drug discovery.

2. Related Work

To leverage chemical-induced bulk gene expression data for inverse molecule design, MolGAN (Méndez-Lucio et al., 2020a) exploited gene expression profiles to generate hit-like molecules. We replaced the chemical representation from simplified molecular inline entry systems (SMILES) strings with self-referencing embedded strings (SELFIES) in this model and used the pretrained model for encoding chemicals used by the authors in their subsequent work (Méndez-Lucio et al., 2020b). However, the conditioning networks make use of real and fake conditions, where conditions outside of the training batch are assumed to be fake, which could introduce potential inaccuracies due to the presence of multiple replicates with distinct gene expressions in the L1000 RNA-seq dataset. A more recent work, GxVAE (Li & Yamanishi, 2024), proposed using joint VAEs to extract latent features from gene expressions and using them as conditions for generating molecules utilizing a second VAE for encoding molecules. These approaches result in poorly structured latent space, consequently affecting the model's ability to generate chemically valid molecules. Additionally, these models do not address the incumbent challenge of leveraging highly sparse and out-of-distribution single-cell datasets to generate molecules. To address these challenges, we developed a denoising gene expression encoder, which reconstructs the median gene expression in the case of replicates chemical samples, essentially acting as a denoising gene expression autoencoder. This approach ensures a more meaningful and robust representation of gene expressions, contributing to a better-structured latent space and enhancing the generation of valid and diverse molecular structures.

3. Methods

MolGene-E is an advanced framework designed for the *de novo* generation of molecules with desired biological properties. As shown in Figure 1, the framework involves a meticulous four-step process that integrates diverse data representations and sophisticated machine learning models to align chemical and biological information effectively.

3.1. Denoising VAE for Gene Expression Profiles

In order to manage and interpret the complex data from multiple replicates, MolGene-E employs a Variational Autoencoder (VAE). The VAE is trained with the objective of reconstructing the median gene expression signature from the replicates for each sample (Figure 1a). This approach ensures that the VAE captures the most representative gene expression profile, smoothing out anomalies and focusing on the core response to chemical perturbations. This process enhances the reliability of the gene expression data used in further steps.

3.2. SELFIES VAE for Chemicals

For representing the chemical structures, MolGene-E leverages SELFIES (Self-Referencing Embedded Strings), a robust and versatile chemical representation. These SELFIES representations are encoded using a VAE pretrained model from a VAE model pretrained on ZINC dataset (Gao et al., 2022) (Figure 1b). The use of SELFIES allows for a comprehensive and error-resistant encoding of molecular structures, facilitating seamless integration with machine learning models.

3.3. Alignment of Gene Expressions and Chemical Representations

The critical innovation in MolGene-E lies in aligning the gene expression profiles with their corresponding chemical structures. This is achieved through a CLIP (Contrastive Language-Image Pretraining) module trained with supervised contrastive loss (Figure 1c). The objective of this module is to align the embeddings of phenotypes (gene expression profiles) with the embeddings of the SELFIES representations of the chemicals that caused the perturbations. By doing so, MolGene-E ensures that the biological effects of chemicals are accurately reflected in their encoded representations.

3.4. Mapping Gene Expressions to Chemical Embeddings

To complete the alignment process, MolGene-E employs a Multi-Layer Perceptron (MLP) based prior model (Figure 1d). This model is trained to map the embeddings of gene expression profiles to the embeddings of their corresponding chemical counterparts. The MLP-based prior effectively bridges the gap between biological responses and chemical structures, enabling the generation of novel molecules that can induce desired gene expression changes.

4. Results and Discussion

4.1. Implementation Details

Datasets. The L1000toRNAseq dataset, originally containing 978 landmark genes, was transformed to RNA-seq-like profiles encompassing 23,614 genes using a cycleGAN model as described by (Jeon et al., 2022). The dataset includes gene expression profiles from 221 human cell lines treated with over 30,000 chemical and genetic perturbations, resulting in over 3 million expression profiles. We filtered the data for chemical perturbations with 24-hour infection times and 10 μM dosage for the MCF7 cell line, resulting in 3116 genes with high variance (variance > 0.75). For training MolGene-E we did a 70-15-15 split to get training, validation, and test sets while ensuring there was no

chemical overlap in the data splits.

The Sciplex-3 dataset, sourced from (Srivatsan et al., 2020), and harmonized by (Peidli et al., 2024), includes single-cell transcriptomic profiles of 188 compounds across three cancer cell lines. We focused on the MCF7 cell line, filtering the data to improve quality. The dataset was harmonized using the deep count autoencoder method to impute missing values and align with the L1000toRNAseq dataset using an MLP-based network to remove batch effects. This dataset was only used for inference (molecule generation).

Model Settings. For the denoising VAE, we used hidden layers of sizes [1024, 512, 256] with layernorm and a dropout rate of 0.3. We used a latent dimension of 128. The weight for KL-term of loss was increased linearly from the first to the last epoch. We trained the model for 400 epochs to achieve good performance on the validation set.

For the SELFIES VAE, it maximizes a lower bound of the likelihood (evidence lower bound (ELBO)) instead of estimating the likelihood directly. We used the pretrained model and architecture identical to the one implemented in MOSES (Polykovskiy et al., 2020) to model SELFIES strings. The architecture used a bidirectional Gated Recurrent Unit (GRU) with a linear output layer as an encoder. The decoder was a 3-layer GRU of 512 hidden dimensions with intermediate dropout layers and a dropout rate of 0.2. Training was done with a batch size of 128, utilizing a gradient clipping of 50, KL-term weight linearly increased from 0 to 1 during training. We optimized the model using Adam optimizer with a learning rate of 3e-4.

For the prior model, it is a standard MLP architecture with hidden sizes [1024, 512, 256] and a latent dimension of size 128. The model was trained to minimize a mean squared error loss for the reconstruction of chemical embedding space utilizing the aligned spaces from both modalities. The model used a learning rate of 1e-3 and a batch size of 128.

For the training of MolGene-E, we minimize the self-supervised contrastive learning objective with an approach similar to (Radford et al., 2021). To maximize the similarity between the positive samples and minimize the similarity between negative samples in the learned representation space, we modified the infoNCE (Oord et al., 2018) as per (Khosla et al., 2020) to allow for the data from replicate chemical perturbations in the training batches. MolGene-E was trained for 600 epochs with a batch size of 128 and a learning rate of 1e-4. A projection network with MLP hidden layers [128, 128] was further added to the gene expression encoder and then the model was fine-tuned while optimizing the supervised contrastive loss function $\mathcal{L}_{\text{sup}}$:

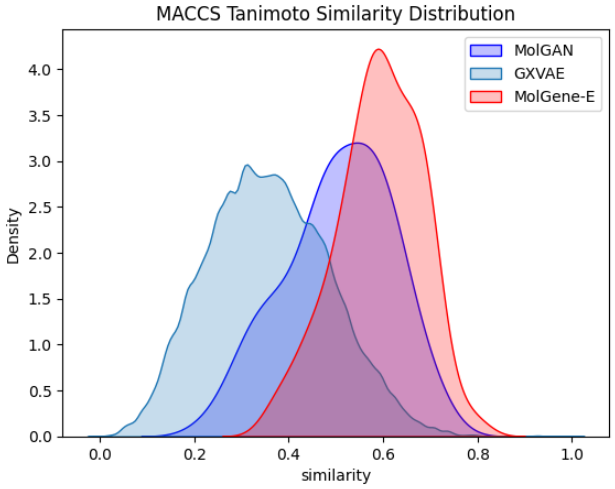

Figure 2. Distributions of MACCS key Tanimoto similarities between molecules generated using gene expression signature induced using CRISPR target knockouts and reference molecules.

Table 1. Evaluation metrics on CRISPR datset.

Model	Validity	Uniqueness	Novelty	Diversity	SA
MolGAN	46.3%	89%	**100%**	**1.0**	**4.14**
GxVAE	92.9%	**91%**	20.7%	0.731	2.87
MolGene-E	**100%**	89%	**100%**	0.99	3.20

$$\mathcal{L}_{\text{sup}} = \sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}, \tag{1}$$

where I is the set of indices of the anchors in the mini-batch, $P(i)$ is the set of indices of replicate gene expressions in the batch i (samples corresponding to same chemical perturbation as i), $A(i)$ is the set of indices of all samples in the mini-batch, excluding i, $\mathbf{z}_i$ denotes the embedding of the i-th sample, τ denotes the temperature parameter that scales the logits. The SELFIES chemical embeddings were used directly from the pre-trained chemical encoder model underscored in the previous section and its parameters were frozen while training. For further details on the training (Algorithm 1,2) and inference (Algorithm 3) process, please refer to Appendix A.

Evaluation Metrics. For performance evaluation, Tanimoto similarity distributions, novelty, and uniqueness between the generated molecules and reference molecules as well as diversity, validity, and synthesizability were computed.

- **Novelty**: We defined as the fraction of generated

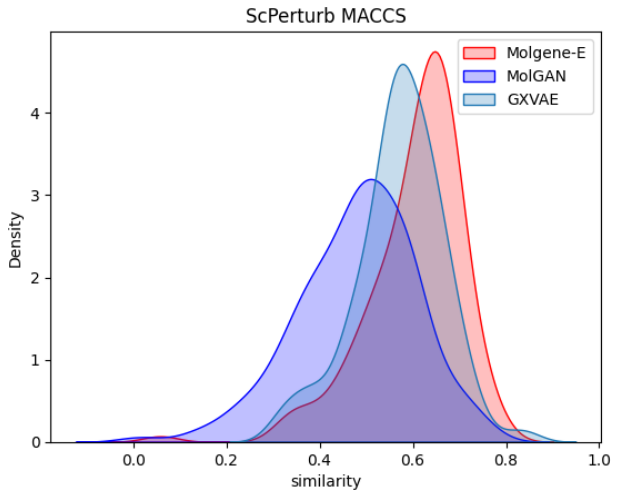

Figure 3. Distributions of MACCS key Tanimoto similarities between reference molecules and molecules generated using gene expression signatures from the Sciplex-3 dataset.

molecules having ≤ 0.4 Tanimoto similarity with the reference molecules.

- **Uniqueness**: The fraction of distinct molecules generated for each input gene expression profile was defined as uniqueness and the mean of uniqueness for all generated molecules corresponding to their reference molecules was reported.

- **Validity**: The validity of generated SELFIES strings was computed using the RDKit library in Python.

- **SA**: The synthesizability and accessibility scores were computed using the RDKit library.

4.2. MolGene-E Improves the Success Rate of Inverse Molecule Design

For molecule generation, reference molecules from the test-split of L1000toRNAseq dataset were considered which had single target knock-outs in the CRISPR chemical perturbation dataset for the MCF7 cell line. Gene expression for profiles for these targets were used for inference from the MolGene-E. As shown in Figure 2, MolGene-E outperforms both baseline models, MolGAN and GxVAE, in terms of average Tanimoto similarities computed using the MACCS keys between the generated and reference molecules. For further evaluation of quality of generated molecules, other metrics such as uniqueness, validity, novelty, diversity and synthesizability were used as listed in Table 1.

Table 2. Evaluation metrics on single cell dataset Sciplex-3.

Model	Validity	Uniqueness	Novelty	Diversity	SA
MolGAN	46.3%	**89%**	**100%**	**1.0**	**4.14**
GxVAE	42.73%	23%	98.7%	0.81	2.85
MolGene-E	**100%**	**89%**	**100%**	0.99	3.20

4.3. MolGene-E Can Be Applied to Single-Cell Data

For single-cell RNA-seq (scRNA-seq) data we used Sciplex-3 dataset which uses "nuclear hashing" to quantify global transcriptional responses to thousands of independent perturbations at single-cell resolution. It was applied to 3 cancer lines exposed to 188 compounds, where we filtered the dataset to take samples only from the MCF7 cell line. The scRNA-seq data as such is extremely sparse, making it hard to use directly with deep learning models. Hence in order to impute missing values we used Deep Count Autoencoder (Eraslan et al., 2019). Further, to integrate the scRNA-seq dataset with the L1000RNAtoseq dataset, we used an MLP based network to map gene expression profiles of Sciplex-3 to L1000toRNAseq to introduce homogeneity in the dataset and remove batch effects (results shown in Figure 4 in Appendix B). Further, since each chemical sample has several replicates, gene expression profiles for each chemical perturbation were randomly sampled, and 200 molecules were generated for each gene expression signature. The one with the highest score was chosen as the candidate. Figure 3 shows the Tanimoto similarity distributions of molecules generated using gene expression profiles from the Sciplex-3 single-cell dataset. Results listed in Table 2 indicate that MolGene-E performs at par when generating molecules using gene expression profiles for single dataset Sciplex-3. For a sample of generated molecules using gene expressions from single-cell data and corresponding reference molecules, we show them in Figure 5 in Appendix C.

5. Conclusions

In this paper, we developed an AI-based generative model that utilizes phenotypic properties from single-cell omics data to generate high-quality lead candidates for drug discovery. MolGene-E consistently outperforms baseline methods in generating high-quality, hit-like molecules from gene expression profiles obtained from single-cell datasets and gene expressions induced by CRISPR-based knockout targets. This superior performance is demonstrated across de novo molecule generation metrics, including novelty, diversity, uniqueness, and synthesizability.

Future work includes incorporating multiple cell lines and conditioning drugs on multi-omics data, leading to a robust framework capable of more accurately reflecting the complex biological environments found in vivo. Additionally, expanding the model to integrate diverse datasets will en-

hance its ability to generalize across different biological contexts, thereby improving its predictive power and utility in identifying effective therapeutic compounds. This approach will pave the way for more personalized and precise drug discovery, ultimately accelerating the development of new treatments and improving patient outcomes.

Impact Statement

This paper presents a novel AI based generative model whose ability to generate molecules based on single-cell phenotype gene expression data in healthcare holds the promise of personalized medicine, faster drug development, and reduced healthcare costs.

References

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):390, 2019.

Gao, W., Fu, T., Sun, J., and Coley, C. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in neural information processing systems*, 35:21342–21357, 2022.

Han, Y., Wang, D., Peng, L., Huang, T., He, X., Wang, J., and Ou, C. Single-cell sequencing: a promising approach for uncovering the mechanisms of tumor metastasis. *Journal of Hematology Oncology*, 15, 05 2022. doi: 10.1186/s13045-022-01280-w.

Jeon, M., Xie, Z., Evangelista, J. E., Wojciechowicz, M. L., Clarke, D. J., and Ma'ayan, A. Transforming l1000 profiles to rna-seq-like profiles with deep learning. *BMC bioinformatics*, 23(1):374, 2022.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Li, C. and Yamanishi, Y. Gxvaes: Two joint vaes generate hit molecules from gene expression profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13455–13463, 2024.

Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D., and Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature communications*, 11(1):10, 2020a.

Méndez-Lucio, O., Zapata, P. A. M., Wichard, J., Rouquié, D., and Clevert, D.-A. Cell morphology-guided de novo hit design by conditioning generative adversarial networks on phenotypic image features. 2020b.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Peidli, S., Green, T., Shen, C., Gross, T., Min, J., Garda, S., Yuan, B., Schumacher, L., Taylor-King, J., Marks, D., Luna, A., Blüthgen, N., and Sander, C. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21: 1–10, 01 2024. doi: 10.1038/s41592-023-02144-y.

Pham, T.-H., Qiu, Y., Zeng, J., Xie, L., and Zhang, P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to covid-19 drug repurposing. *Nature machine intelligence*, 3(3):247–257, 2021.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Salame, N., Fooks, K., El-Hachem, N., Bikorimana, J. P., Mercier, F., and Rafei, M. Recent advances in cancer drug discovery through the use of phenotypic reporter systems, connectivity mapping, and pooled crispr screening. *Frontiers in Pharmacology*, 13:852143, 06 2022. doi: 10.3389/fphar.2022.852143.

Srivatsan, S. R., McFaline-Figueroa, J. L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H. A., Jackson, D. L., Daza, R. M., Christiansen, L., et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Subramanian, A., Narayan, R., Corsello, S., Peck, D., Natoli, T., Lu, X., Gould, J., Davis, J., Tubelli, A., Asiedu, J., Lahr, D., Hirschman, J., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I., Lam, D., and Golub, T. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. 05 2017. doi: 10.1101/136168.

Van de Sande, B., Lee, J. S., Mutasa-Gottgens, E., Naughton, B., Bacon, W., Manning, J., Wang, Y., Pollard, J., Mendez, M., Hill, J., Kumar, N., Cao, X., Chen, X., Khaladkar, M., Wen, J., Leach, A., and Ferran, E. Applications of single-cell rna sequencing in drug discovery and development. *Nature Reviews Drug Discovery*, 22, 04 2023. doi: 10.1038/s41573-023-00688-4.

Vincent, F., Nueda, A., Lee, J., Schenone, M., Prunotto, M., and Mercola, M. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, 21, 05 2022. doi: 10.1038/s41573-022-00472-w.

Zeng, X., Wang, F., Luo, Y., Kang, S.-G., Tang, J., Lightstone, F., Fang, E., Cornell, W., Nussinov, R., and Cheng, F. Deep generative molecular design reshapes drug discovery. *Cell reports. Medicine*, 3:100794, 10 2022. doi: 10.1016/j.xcrm.2022.100794.

A. Algorithms for Training and Inference

In this section, we present the training and inference procedures.

Algorithm 1 Training the CLIP module

Require: $\mathcal{D}$: Dataset of pairs of SELFIES strings and corresponding gene expressions (s, g), S: pre-trained and frozen SELFIES VAE, G_θ: Gene expression VAE
Require: $\mathcal{B}$: Number of mini-batches
for each mini-batch $\mathcal{M}$ in $\mathcal{B}$ **do**
  Sample mini-batch $\mathcal{M} = \{(s_1, g_1), \ldots, (s_k, g_k)\}$
  for each pair $(s, g) \in \mathcal{M}$ **do**
    $z_{\text{selfies}} \leftarrow S.Encoder(s)$ {Encode SELFIES string}
    $z_{\text{gene}} \leftarrow G_\theta.Encoder(g)$ {Encode gene expression}
    $\ell \leftarrow L_{\text{contrastive}}(z_{\text{selfies}}, z_{\text{gene}})$ {Compute contrastive loss, ref Equation (1)}
  end for
  Compute average loss $\bar{\ell}$ over the mini-batch
  Update weights of E_{gene} using gradient descent with $\bar{\ell}$
end for

Algorithm 2 Training the Prior Module

Require: $\mathcal{D}$: Dataset of pairs of chemical embeddings and corresponding gene expression embeddings (S_θ, G_θ) from the pretrained CLIP module
Require: $\mathcal{B}$: Number of mini-batches
Require: Initialize Prior Model θ
for each mini-batch $\mathcal{M}$ in $\mathcal{B}$ **do**
  Sample mini-batch $\mathcal{M} = \{(s_1, g_1), \ldots, (s_k, g_k)\}$
  for each pair $(s, g) \in \mathcal{M}$ **do**
    $z_{\text{recons}} \leftarrow \theta(g)$ {Map gene expression to chemical space}
    $\ell \leftarrow L_{\text{RMSE}}(z_{\text{recons}}, s)$ {Compute RMSE loss}
  end for
  Compute average loss $\bar{\ell}$ over the mini-batch
  Update weights of Prior Model using gradient descent with $\bar{\ell}$
end for

Algorithm 3 Inference

Require: Gene expression g
Require: G_θ: Gene expression VAE, S: SELFIES VAE, P: Prior model,
$z_{\text{gene}} \leftarrow G_\theta(\text{g})$ {Encode gene expression to gene embedding}
$z_{\text{chemical}} \leftarrow P(z_{\text{gene}})$ {Generate chemical embedding using Prior model}
$s_{\text{molecule}} \leftarrow S.Decoder(z_{\text{chemical}})$ {Decode chemical embedding}
return s_{molecule} {Return molecule represented as SELFIES string}

B. Mean Gene Expression Signatures After Harmonizing Single Cell Dataset With L1000

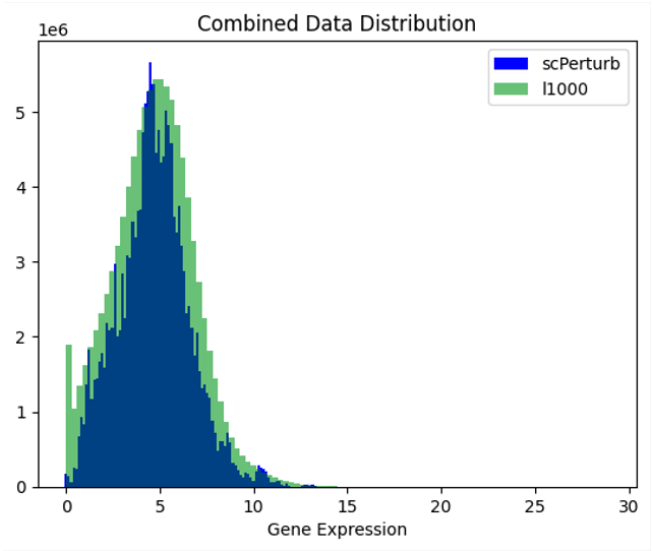

Figure 4. Mean gene expression signatures after harmonizing single cell dataset with L1000.

C. Sample reference and generated molecules

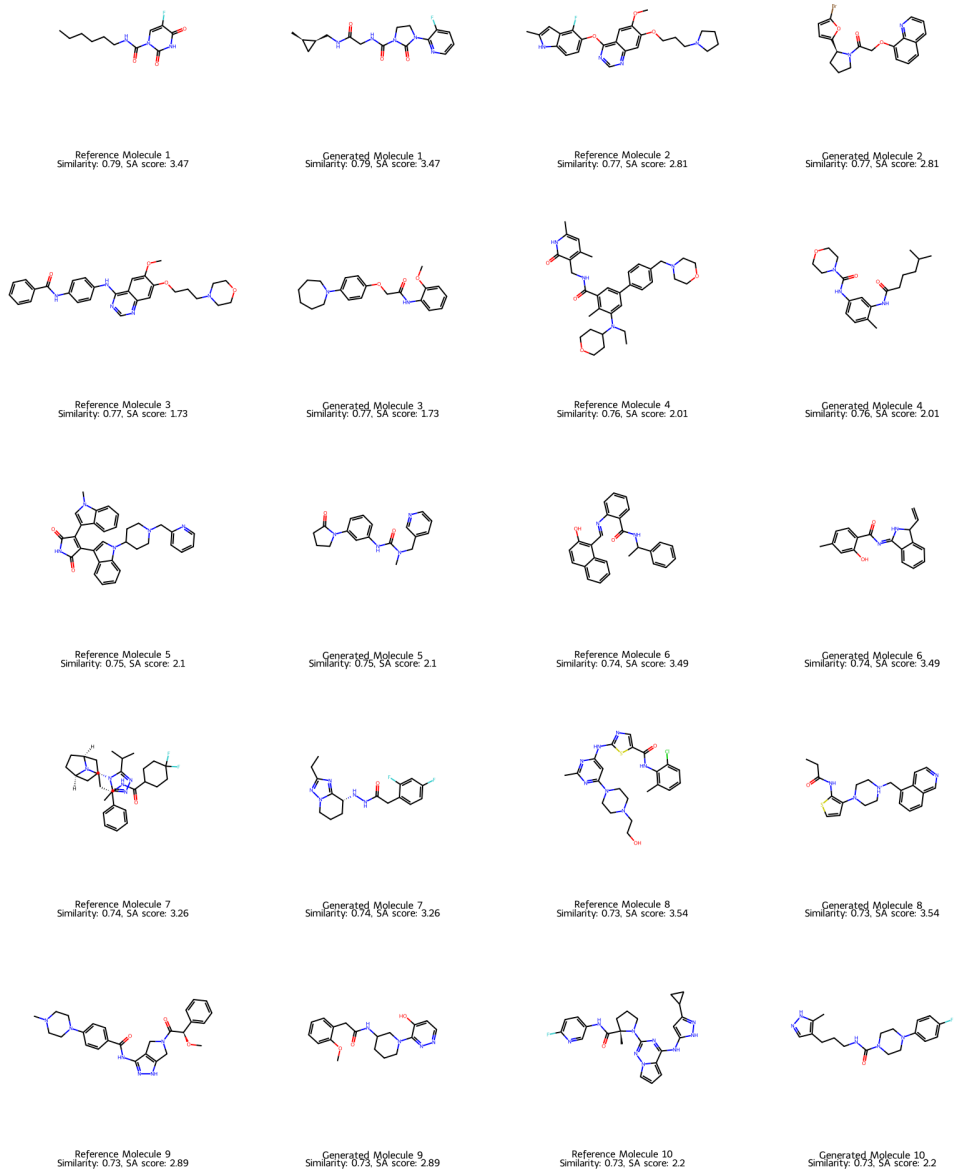

Figure 5. Sample reference and generated molecules from gene expression profiles from single-cell dataset.