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

Rahul Ohlan[1], Raswanth Murugan[3], Li Xie[3], Mohammedsadeq Mottaqi[2], Shuo Zhang[3,4*], Lei Xie[1,2,3,4*]

[1]Ph.D. program in Computer Science, The Graduate Center, The City University of New York, New York, NY, 10016, USA.
[2]Ph.D. program in Biochemistry, The Graduate Center, The City University of New York, New York, NY, 10016, USA.
[3]Department of Computer Science, Hunter College, The City University of New York, New York, NY, 10065, USA.
[4]Helen and Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University, New York, NY, 10065, U.S.A.

*Corresponding author(s). E-mail(s): sz780@hunter.cuny.edu; lei.xie@hunter.cuny.edu;
Contributing authors: rohlan@gradcenter.cuny.edu; raswanth99@gmail.com; lxie9509@gmail.com; mmottaqi@gradcenter.cuny.edu;

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 to 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 contrastive learning-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 as validated by target knock-out experiments using CRISPR. This superior performance is demonstrated across diverse *de novo* molecule generation metrics. Extensive evaluations demonstrate that MolGene-E achieves state-of-the-art performance for zero-shot molecular generations. This makes MolGene-E a potentially powerful new tool for drug discovery.

Keywords: Systems Pharmacology, Drug Discovery, Phenotypic Screening, Generative AI

1 Introduction

Capitalizing on the success of deep learning across various domains such as natural language, images, and videos, deep generative models have been extensively applied to the generation of small organic compounds targeting a specific disease gene for drug discovery [1]. 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 [2]. 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 [3][4]. Several deep learning methods have been proposed to leverage chemical-induced bulk gene expression data for inverse molecule design. Notably, MolGAN [5] generates molecules conditioning a generative adversarial network with bulk transcriptomics data. Although it shows promising results, we observe that generative adversarial networks (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 [6], which employs two joint variational autoencoders (VAEs) to facilitate the extraction of latent gene expression features and use it as a condition to generate molecules using a second VAE. However, GxVAE has not been developed for single-cell transcriptomics data.

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 [7]. 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 [8]. 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 [9] 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 [9]. 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 [10] and Sciplex-3 [11] 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 tool for drug discovery.

2 Results

2.1 Overview of MolGene-E

MolGene-E is a deep generative framework designed to generate novel drug-like molecules from single-cell transcriptomics data. The pipeline draws inspiration from OpenAI's DALL-E 2 model [12], which generates images from textual descriptions through a two-step encoding and generative process. Similarly, MolGene-E employs a cross-modal learning approach, where gene expression profiles are aligned with molecular representations to enable de novo molecule generation.

As shown in Figure 1, the framework involves a five-step process that integrates diverse data representations and deep learning modules to align chemical and gene expression profile information effectively.

First, a Variational Autoencoder (VAE) denoises gene expression profiles corresponding to replicate chemical perturbations by reconstructing their median gene

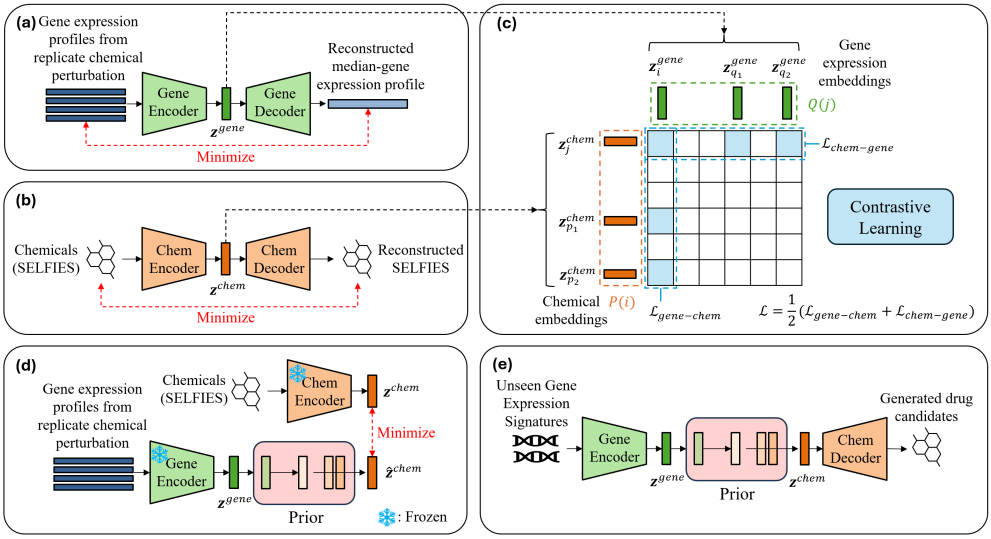


Fig. 1: **(a)** A Variational Autoencoder (VAE) denoises the gene expression profiles corresponding to replicate chemical perturbation in a batch by reconstructing their median gene expression profiles. **(b)** We represent chemical structures via SELFIES and use a pretrained frozen VAE to extract the chemical embeddings. **(c)** The gene expression encoder is fine-tuned to align gene embeddings $\boldsymbol{z}^{gene}$ to the chemical embeddings $\boldsymbol{z}^{chem}$ via a contrastive learning module. A supervised objective $\mathcal{L}$ (Equation 3) is optimized to maximize the agreement between positive pairs while minimizing the similarity between negative pairs. **(d)** A prior model is trained to map the inferred $\boldsymbol{z}^{gene}$ to the inferred $\boldsymbol{z}^{chem}$. **(e)** Given unseen gene expression profiles, the inferred $\boldsymbol{z}^{gene}$ are mapped to $\boldsymbol{z}^{chem}$ via the pretrained prior model, and are further decoded using the SELFIES VAE's decoder to generate drug candidates.

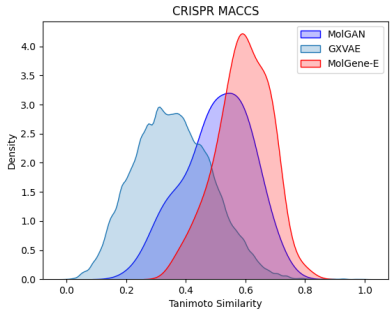
expression profiles. This step ensures robustness against biological noise in transcriptomic data. Next, chemical structures are represented using SELFIES and passed through a pretrained, frozen VAE to obtain chemically meaningful embeddings.

A contrastive learning module is then employed to align the gene expression space with the chemical space. The gene expression encoder is fine-tuned to maximize the agreement between gene embeddings $\boldsymbol{z}^{gene}$ and their corresponding chemical embeddings $\boldsymbol{z}^{chem}$ while minimizing similarity with negative samples. This supervised contrastive loss optimizes the model's ability to infer chemical perturbations from gene expression data.

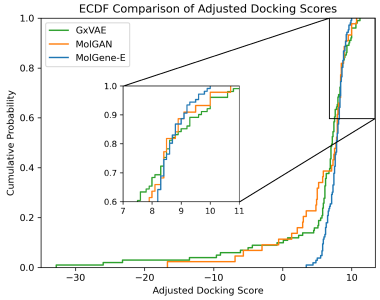
To enable molecule generation, a prior model is trained to learn the mapping from $\boldsymbol{z}^{gene}$ to $\boldsymbol{z}^{chem}$. Given an unseen gene expression profile, the trained prior model predicts its corresponding $\boldsymbol{z}^{chem}$, which is subsequently decoded by the SELFIES VAE to generate novel molecular structures.

By harmonizing multi-modal data and leveraging generative modeling principles, MolGene-E provides a powerful approach for discovering drug candidates based on transcriptomic responses, advancing the field of AI-driven drug discovery.

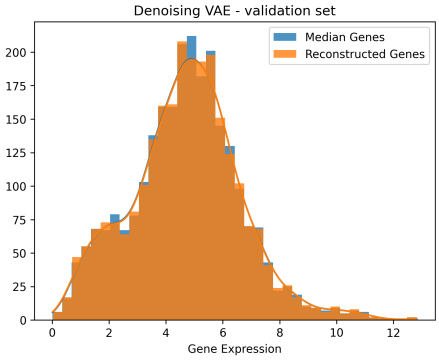
2.2 MolGene-E Improves the Success Rate of Inverse Molecule Design



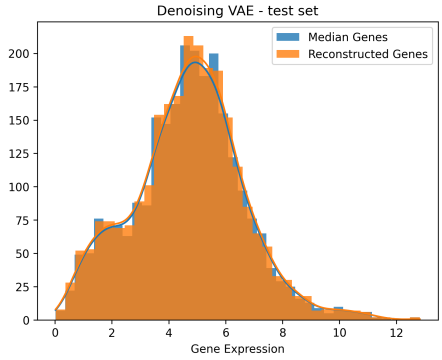
(a)



(b)



(c)



(d)

Fig. 2: **a**: Distributions of MACCS key Tanimoto similarities between molecules generated using gene expression signatures induced by CRISPR target knockouts and reference molecules. **b**: Cumulative distribution of docking scores for generated molecules across corresponding protein targets. **c**: Reconstruction performance of the denoising VAE on the validation set. **d**: Reconstruction performance of the denoising VAE on the test set.

We use a challenge task to evaluate the performance of the molecular generation from gene expressions. If a drug can correctly revert gene expressions from a disease state to

a healthy state, the drug could interact with disease-causing genes, i.e., drug targets. In other words, the gene expression changes that are caused by the target gene knock-out or knock-down should be similar to those that result from the chemical perturbation targeting the knock-out/down gene. In our experiments, reference molecules from the test-split of L1000toRNAseq dataset were considered which had single target knock-outs in the CRISPR gene perturbation dataset [13] for the MCF7 cell line. Gene expression profiles for these targets were used for the inference of drug candidates.

As shown in Figure 2a, 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.

Table 1: Evaluation metrics on CRISPR perturbation dataset. We mark the best results in bold and the second-best results with underline.

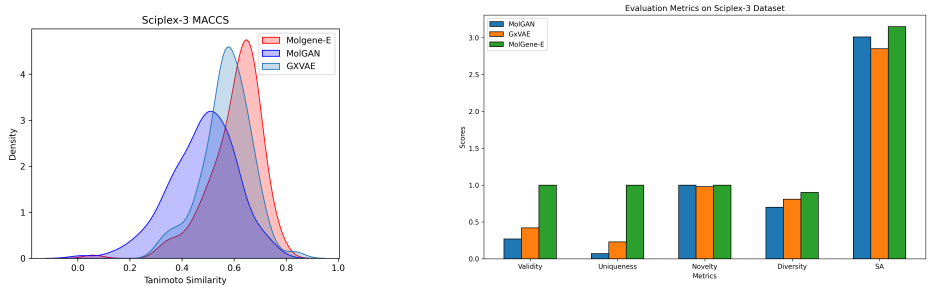
Model	Validity ↑	Uniqueness ↑	Novelty ↑	Diversity ↑	SA ↓
MolGAN	0.31	<u>0.99</u>	**1.00**	**0.89**	4.14
GxVAE	<u>0.93</u>	0.91	0.21	0.73	**2.87**
MolGene-E	**1.00**	**1.00**	**1.00**	**0.89**	<u>3.20</u>

For further evaluation of the quality of generated molecules, other metrics including uniqueness, validity, novelty, diversity, and synthesizability (SA) were used as listed in Table 1. MolGene-E demonstrates superior performance across most metrics compared to the baseline models, MolGAN and GxVAE. It achieves perfect scores for **Validity**, **Uniqueness**, and **Novelty**, highlighting its robust capability to generate valid, unique, and novel molecular candidates. Additionally, MolGene-E matches MolGAN in **Diversity**, achieving the highest score, while slightly trailing behind GxVAE in **SA (Synthetic Accessibility)**. These results affirm MolGene-E's effectiveness and versatility across diverse datasets. MolGene-E achieves the best performance on validity, uniqueness, novelty, and diversity while maintaining a high level of SA. This indicates that MolGene-E not only generates molecules that closely resemble the reference compounds in terms of structural similarity but also proposes novel and diverse chemical scaffolds that are synthetically feasible.

Figures 2c and 2d compares the reconstructed gene expression values from a denoising VAE on a test set to the median gene expression values derived from replicate samples of a perturbation. The histogram shows the density distribution of gene expression values, where the blue bars represent the median gene expression values, and the orange bars represent the reconstructed values. The close overlap of the two distributions indicates that the VAE effectively captures and reconstructs the original distribution of gene expression values. The smooth curve overlay demonstrates the similarity between the reconstructed and true distributions, reflecting the model's ability to denoise and reconstruct gene expression patterns accurately, even in the presence of noise from biological or experimental variability.

For a comprehensive evaluation, we further applied AutoDock[14] to the generated compounds to assess their docking performance, illustrated in Figure 2b. Human ligand-protein complex structures were retrieved from the PDB database to identify binding pockets in the target proteins. Binding pockets associated with solvents and metal ions were excluded, and ligand-protein complexes with the best resolution for each target were selected for the docking process. A total of 113 human protein targets with valid ligand-binding pockets were identified. For each protein target, we docked the predicted Top-1 compound that has the highest Tanimoto similarity with reference molecules (utilizing the MACCS keys) using Autodock. The generated compounds that could not be converted to 3D mol2 format using Open Babel were excluded, which resulted in 101 chemical-protein pairs for GxVAE, 44 pairs for MolGAN, and 106 pairs for MolGene-E. The results of the Kolmogorov-Smirnov (K-S) test between MolGene-E and the baseline generative models, MolGAN and GxVAE, indicate significant differences in the cumulative distributions of their docking scores with corresponding targets. Specifically, the K-S statistic for the comparison between MolGene-E and MolGAN is 0.3032, which suggests a considerable divergence between their cumulative distribution functions. The associated p-value of 0.0049, being below the 0.05 threshold, confirms that this difference is statistically significant. In contrast, the K-S statistic between MolGene-E and GxVAE is 0.2304, indicating a smaller divergence between these two models, with a p-value of 0.0065, which also suggests a statistically significant difference, though less pronounced than in the MolGene-E vs. MolGAN comparison. These findings highlight that MolGene-E generates molecular distributions that are distinct from both baseline models, with a more substantial divergence from MolGAN compared to GxVAE.

2.3 MolGene-E Can Be Applied to Single-Cell Data



(a) (b)

Fig. 3: **(a)** Distributions of MACCS key Tanimoto similarities between reference molecules and molecules generated using gene expression signatures from the Sciplex-3 dataset. **(b)** Comparison of Evaluation metrics across baselines

For single-cell RNA-seq (scRNA-seq) data, we used Sciplex-3 dataset [11]. Furthermore, 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 3a shows the distributions of the Tanimoto similarity scores of molecules generated using gene expression profiles from the Sciplex-3 dataset. Results listed in Figure 3b indicate that MolGene-E consistently outperforms the baseline models, MolGAN and GxVAE, across most evaluation metrics on the Sciplex-3 dataset. For **Validity**, MolGene-E achieves a perfect score of **1.00**, far exceeding MolGAN (0.27) and GxVAE (0.42). This perfect validity score is primarily attributed to the use of **SELFIES** instead of SMILES for molecular representation, ensuring that all generated molecules are chemically valid by design. Similarly, MolGene-E attains a score of **1.00** for **Uniqueness**, significantly outperforming MolGAN (0.07) and GxVAE (0.23), demonstrating its ability to generate entirely unique molecules without duplication. For **Novelty**, MolGene-E again achieves a perfect score of **1.00**, matching MolGAN and slightly surpassing GxVAE (0.98), reflecting its capability to generate novel molecules that diverge from known examples. Furthermore, MolGene-E achieves the highest score for **Diversity** at **0.90**, surpassing MolGAN (0.70) and GxVAE (0.81), indicating its effectiveness in exploring a broad chemical space. While MolGene-E's score for **SA (Synthetic Accessibility)** is **3.15**, slightly higher (worse) than those of MolGAN (3.01) and GxVAE (2.85), this difference is marginal and does not detract from its strengths in other metrics. Overall, MolGene-E demonstrates superior performance, particularly in Validity, Uniqueness, Novelty, and Diversity, showcasing its ability to generate high-quality, chemically valid, and diverse molecular candidates. Although there is a minor trade-off in synthetic accessibility, this is a reasonable compromise given its exceptional performance in other critical areas. A sample of generated molecules using gene expression perturbation profiles from single-cell data and corresponding reference molecules are presented in Figure 4.

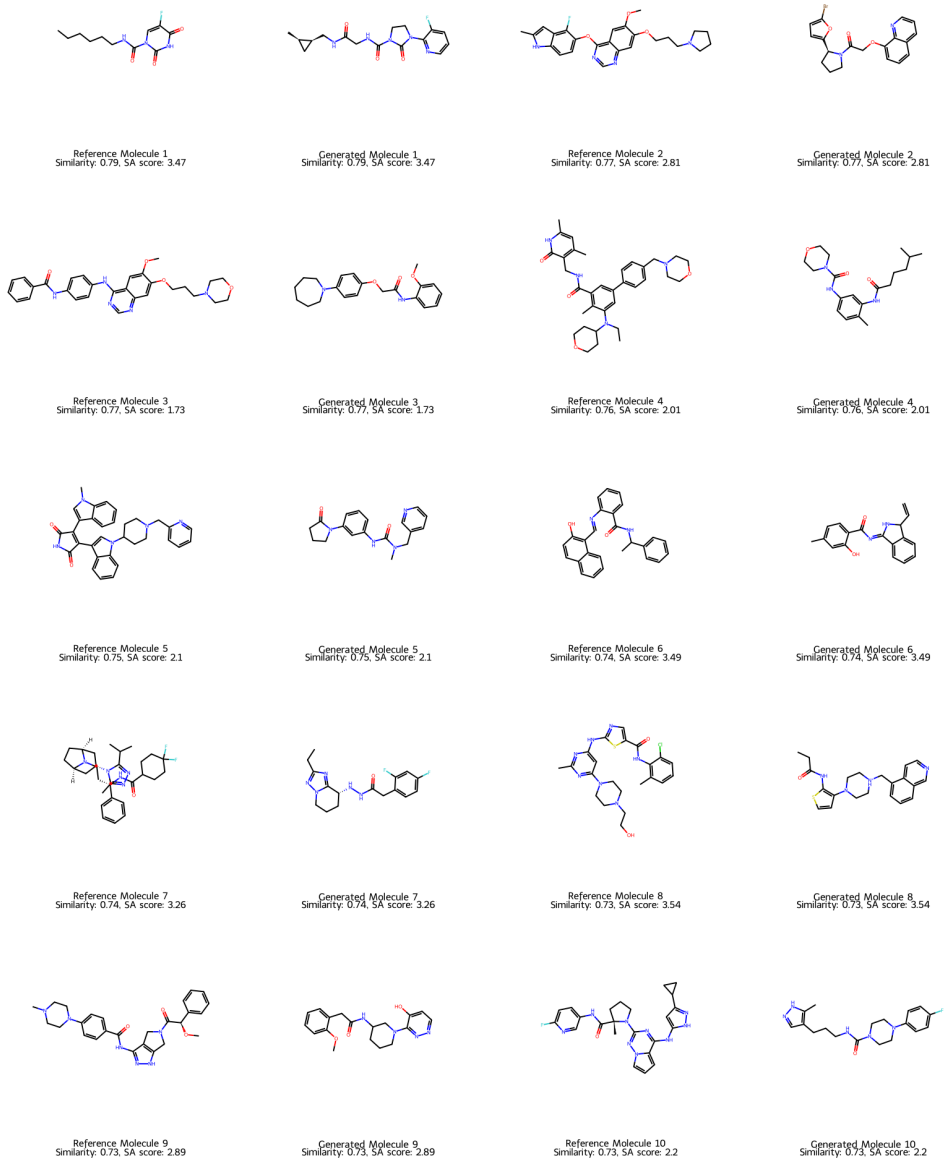
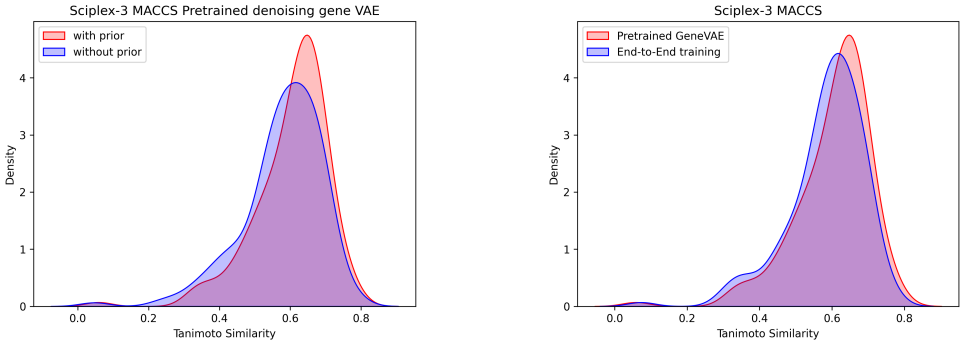


Fig. 4: Reference and generated molecules using gene expression profiles from Sciplex-3.

2.4 Ablation Studies

To evaluate the efficacy of various design choices in our molecular generation pipeline, we conducted comprehensive ablation studies on the Sciplex-3 dataset in a zero-shot testing setting. The results are discussed below.

Effect of prior model: We remove the prior model in MolGene-E to build an ablated model called MolGene-E$_{NP}$. When a prior model was utilized, gene expression embeddings were mapped to latent space embeddings, which were then used for molecule generation. Without the prior model, gene expression embeddings were directly utilized from the CLIP encoder. Figure 5a shows two distributions of Tanimoto similarity comparing the effect of using the prior model: one when the encoder was pretrained for denoising and the other when the pipeline was trained in an end-to-end manner without pretraining the encoder. Using the prior model results in higher density in a high similarity range (>0.6), suggesting better alignment with the reference molecules. The molecule generation metrics for this comparison are shown in Table 2. While both approaches maintain high validity, uniqueness, and novelty, the use of the prior model in MolGene-E reduces the synthetic accessibility (SA) score.



(a) Molecule generation with and without using the Prior Model

(b) Using pretrained VAE vs end to end training.

Fig. 5: Effect of using a prior model on molecule similarity for the ScPerturb dataset. (a) Results with a pretrained gene expression encoder. (b) Results comparing effectiveness of pre-trained VAE over end to end training

Effect of pretraining the gene expression encoder for denoising: We compared the performance of MolGene-E when the gene expression encoder was pretrained as a denoising autoencoder with MolGene-E$_E$ when the whole model was trained in an end-to-end manner without pretraining. Figure 5b shows the distributions of the Tanimoto similarity scores for both setups. The pretrained encoder produces a slightly sharper and narrower peak in the high similarity range, indicating a better alignment with reference molecules. The molecule generation metrics for this comparison are listed in Table 2. MolGene-E achieves a lower SA score than both MolGene-E$_E$ and

Table 2: Results of ablation study on Sciplex-3 dataset.

Model	Prior	Pretrain	SA↓
MolGene-E	✓	✓	**3.15**
MolGene-E$_{NP}$	×	✓	3.49
MolGene-E$_{E}$	✓	×	3.25

MolGene-E$_{NP}$ highlighting the importance of using the pretraining for denoising and using the prior model.

The results in our ablation studies demonstrate the importance of incorporating a prior model and pretraining the gene expression encoder. The prior model enhances the alignment between generated and reference molecules, while the pretrained encoder improves precision without sacrificing diversity. These findings provide valuable insights into optimizing the pipeline for high-quality molecule generation.

3 Discussion

In this paper, we developed a deep 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 enhance 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.

4 Methods

4.1 Denoising VAE for Gene Expression Profiles

In chemical-induced bulk gene expression data, multiple distinct gene expression profiles perturbed by replicate chemicals can exist. In order to manage and interpret the complex data from multiple replicates, MolGene-E employs a VAE for denoising, which is trained with the objective of reconstructing the median gene expression profile from the gene expression profiles corresponding to replicate chemical perturbations in a batch (Figure 1a). The training objective incorporates a standard reconstruction loss

and KL divergence loss [15], with dynamic weighting between these components. During training, the weight of the reconstruction component is gradually reduced while the KL divergence component is increased, ensuring that the latent space becomes well-regularized and captures the underlying structure of the data. This approach ensures that the VAE captures the most representative gene expression profile, reducing noise and focusing on the core response to chemical perturbations. This process enhances the reliability of the gene expression data used in further steps.

4.2 SELFIES VAE for Chemicals

For representing the chemical structures in perturbations, MolGene-E leverages SELFIES (Self-Referencing Embedded Strings) [16] due to its guaranteed 100% validity in contrast to using SMILES strings to represent molecules for molecule design [5]. These SELFIES strings are encoded using a VAE model pretrained on ZINC dataset [17] (Figure 1b). The use of SELFIES allows for a comprehensive and error-resistant encoding of molecular structures, facilitating seamless integration with machine learning models.

4.3 Alignment of Gene Expressions and Chemical Representations

The key innovation in MolGene-E lies in aligning the gene expression profiles with their corresponding chemical perturbations. This is achieved through a contrastive learning module (Figure 1c) trained with a supervised contrastive loss $\mathcal{L}$ (Equation 3) inspired by CLIP [18] and SupCon loss [19]. 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. It is also specifically designed to deal with the existence of multiple distinct gene expression profiles perturbed by replicate chemicals in a batch. By doing so, MolGene-E ensures that the biological effects of chemicals are accurately reflected in their encoded representations.

4.4 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.

4.5 Generation of Drug Candidates

After training the prior model, gene expressions corresponding to chemical perturbations can be used for inference to generate drug candidates that might result in similar perturbation effects (Figure 1e). The gene expression embeddings $\boldsymbol{z}^{gene}$ are extracted using the pretrained gene expression encoder in the contrastive learning module and subsequently input to the prior model to compute chemical embeddings $\boldsymbol{z}^{chem}$. $\boldsymbol{z}^{chem}$

are then decoded via the SELFIES VAE model to obtain potential drug candidates in the form of novel molecular structures.

4.6 Implementation Details

4.6.1 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 [10]. 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 CRISPR Perturbations L1000 dataset, sourced from the sigcom portal [13], consists of 1218 L1000 signatures for 44 different transcription factors (TFs) targeted by CRISPR knockout perturbations. We used this data for zero-shot inference filtering it for the MCF7 cell line and the samples consisting of single targets in order to avoid confounding effects that may arise from multi-target perturbations, allowing for a clearer evaluation of the model's ability to infer the effects of individual transcription factor knockouts.

The Sciplex-3 dataset, sourced from [11] uses "nuclear hashing" strategy that relies on labeling the unmodified single-stranded DNA oligos to quantify global transcriptional responses to thousands of independent perturbations at single-cell resolution, and harmonized by scPerturb [20], includes single-cell transcriptomic profiles of 188 compounds across three cancer cell lines. We focused on the MCF7 cell line to be used for inference since MolGene-E was trained on MCF7 cell line data and filtered the data to improve quality. Additionally, the dataset was harmonized with the L1000 data bulk rna seq using the deep count autoencoder (DCA) method [21] to impute missing values and align with the L1000toRNAseq dataset using an MLP-based network to remove batch effects. Figure B1 shows the distribution of gene expression profiles of the processed dataset compared with the distribution of original gene expression profiles from the L1000toRNAseq dataset. This dataset was only used for inference (molecule generation).

4.6.2 Evaluation Metrics

For performance evaluation, the following measures were used. **Tanimoto similarity scores** [22] between reference and generated molecules are computed on encoding molecules to MACCS keys [23]. Generated molecules with higher Tanimoto similarity scores are considered as more potential drug candidates. **Novelty** is the fraction of generated molecules not observed in the training set. **Uniqueness** is the fraction of distinct molecules generated for each input gene expression profile. The mean of uniqueness for all generated molecules corresponding to their reference molecules was reported. **Diversity** measures the chemical variability among the generated molecules.

It is computed as the average pairwise Tanimoto distance between all generated molecules, with higher diversity indicating a broader exploration of the chemical space. **Validity** and **SA** (synthesizability and accessibility scores) are computed using the RDKit library [24].

4.6.3 Model Settings

For the denoising VAE, we used hidden layers of sizes [1024, 512, 256] with layer normalization [25] 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 200 epochs.

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 pre-trained model and architecture identical to the one implemented in MOSES [26] to model SELFIES strings. In detail, 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 [27] with a learning rate of 3e-4.

For the prior model, it is an MLP 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 contrastive learning objective $\mathcal{L}$ (Equation 3) with an approach similar to [18]. 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. When generating drug candidates, 400 unique chemical candidates are generated by sampling from the latent space for each gene expression profile. For further details on the training (Algorithm 1,2) and inference (Algorithm 3) process, please refer to Appendix A.

To define the contrastive loss, we introduce $\mathcal{L}_{\text{gene-chem}}$ and $\mathcal{L}_{\text{chem-gene}}$. The former aligns each arbitrary anchor gene expression embedding $\boldsymbol{z}_i^{gene}$ with an index i to all corresponding replicate chemical perturbation embeddings $\boldsymbol{z}_p^{chem}$ with indices $p \in P(i)$ in a batch:

$$\mathcal{L}_{\text{gene-chem}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i^{gene} \cdot \boldsymbol{z}_p^{chem} / \tau\right)}{\sum_{k \in I} \exp\left(\boldsymbol{z}_i^{gene} \cdot \boldsymbol{z}_k^{chem} / \tau\right)}, \tag{1}$$

while the latter aligns each arbitrary anchor chemical embedding $\boldsymbol{z}_j^{chem}$ with an index j to all corresponding perturbed gene expression embeddings $\boldsymbol{z}_q^{gene}$ with indices $q \in Q(j)$ in a batch:

$$\mathcal{L}_{\text{chem-gene}} = \sum_{j \in I} \frac{-1}{|Q(j)|} \sum_{q \in Q(j)} \log \frac{\exp\left(\boldsymbol{z}_j^{chem} \cdot \boldsymbol{z}_q^{gene} / \tau\right)}{\sum_{k \in I} \exp\left(\boldsymbol{z}_j^{chem} \cdot \boldsymbol{z}_k^{gene} / \tau\right)}. \tag{2}$$

In the two equations above, τ denotes the temperature parameter controlling the sharpness of the similarity scores, $|P|$ denotes the cardinality of P, and I is the set of all indices in the batch.

The final contrastive loss $\mathcal{L}$ is obtained by combining the two losses above:

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\text{gene-chem}} + \mathcal{L}_{\text{chem-gene}}\right), \tag{3}$$

The SELFIES chemical embeddings are used directly from the pre-trained chemical encoder model underscored in the previous section and its parameters are frozen while training.

Data Availability

The datasets analyzed in the study are publicly available. The L1000RNAtoseq chemical perturbations and CRISPR perturbations can be downloaded from https://lincs-dcic.s3.amazonaws.com/LINCS-data-2020/RNA-seq/cp_predicted_RNAseq_profiles.gctx and https://lincs-dcic.s3.amazonaws.com/LINCS-data-2020/RNA-seq/xpr_predicted_RNAseq_profiles.gctx respectively. The sciplex-3 dataset can be obtained from the scPerturb portal at https://zenodo.org/records/7041849/files/SrivatsanTrapnell2020_sciplex3.h5ad?download=1.

Code Availability

The code used in this study will be made accessible to reviewers via Code Ocean during the review phase. The scripts and necessary data to reproduce the key results presented in this manuscript will be included.

Acknowledgement

This project has been funded with federal funds from the National Institute of General Medical Sciences of the National Institute of Health (R01GM122845), the National Institute on Aging of the National Institute of Health (R01AG057555, R21AG083302), and the National Science Foundation (NSF2230354).

Author Contributions

Lei X. conceived the presented idea and acquired funding. R. O. and R. M. performed the computations, baseline experiments, and verified the analytical methods. M. M. organized the datasets used in the study. Li X. conducted the docking experiments using AutoDock. Lei X. and S. Z. supervised the design of methods and experiments, and the writing of the manuscript.

References

[1] Zeng, X., Wang, F., Luo, Y., Kang, S.-G., Tang, J., Lightstone, F., Fang, E., Cornell, W., Nussinov, R., Cheng, F.: Deep generative molecular design reshapes

drug discovery. Cell reports. Medicine **3**, 100794 (2022) https://doi.org/10.1016/j.xcrm.2022.100794

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

[3] Salame, N., Fooks, K., El-Hachem, N., Bikorimana, J.P., Mercier, F., 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 (2022) https://doi.org/10.3389/fphar.2022.852143

[4] Pham, T.-H., Qiu, Y., Zeng, J., Xie, L., 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)

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

[6] Li, C., Yamanishi, Y.: Gxvaes: Two joint vaes generate hit molecules from gene expression profiles. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 13455–13463 (2024)

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

[8] 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., Ferran, E.: Applications of single-cell rna sequencing in drug discovery and development. Nature Reviews Drug Discovery **22** (2023) https://doi.org/10.1038/s41573-023-00688-4

[9] Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., *et al.*: A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell **171**(6), 1437–1452 (2017)

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

[11] 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)

[12] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)

[13] Evangelista, J.E., Clarke, D.J., Xie, Z., Lachmann, A., Jeon, M., Chen, K., Jagodnik, K.M., Jenkins, S.L., Kuleshov, M.V., Wojciechowicz, M.L., *et al.*: Sigcom lincs: data and metadata search engine for a million gene expression signatures. Nucleic acids research **50**(W1), 697–709 (2022)

[14] Morris, G.M., Huey, R., Olson, A.J.: Using autodock for ligand-receptor docking. Current protocols in bioinformatics **24**(1), 8–14 (2008)

[15] Kingma, D.P.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[16] Krenn, M., Häse, F., Nigam, A., Friederich, P., Aspuru-Guzik, A.: Self-referencing embedded strings (selfies): A 100% robust molecular string representation. Machine Learning: Science and Technology **1**(4), 045024 (2020)

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

[18] 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 (2021). PMLR

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

[20] 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., Sander, C.: scperturb: harmonized single-cell perturbation data. Nature Methods **21**, 1–10 (2024) https://doi.org/10.1038/s41592-023-02144-y

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

[22] Bajusz, D., Rácz, A., Héberger, K.: Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of cheminformatics **7**, 1–13 (2015)

[23] Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.: Reoptimization of mdl keys for use in drug discovery. Journal of chemical information and computer sciences **42**(6), 1273–1280 (2002)

[24] Riniker, S., Landrum, G.: RDKit: Open-source cheminformatics. URL: https://www.rdkit.org (2013)

[25] Lei Ba, J., Kiros, J.R., Hinton, G.E.: Layer normalization. ArXiv e-prints, 1607 (2016)

[26] 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)

[27] Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

Appendix A Algorithms for Training and Inference

Algorithm 1 Training the Contrastive Learning 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: $\mathcal{B}$: Set of mini-batches
for each mini-batch $\mathcal{M}$ in $\mathcal{B}$ **do**
  for each pair $(s, g) \in \mathcal{M}$ **do**
    $\boldsymbol{z}^{chem} \leftarrow S.Encoder(s)$ {Encode SELFIES string}
    $\boldsymbol{z}^{gene} \leftarrow G_\theta.Encoder(g)$ {Encode gene expression}
  end for
  Compute average loss $\bar{\ell} \leftarrow \mathcal{L}(\boldsymbol{z}^{chem}, \boldsymbol{z}^{gene})$ over $\mathcal{M}$ (Equation 3)
  Update weights of $G_\theta.Encoder$ using gradient descent with $\bar{\ell}$
end for

Algorithm 2 Training the Prior 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: pre-trained and frozen gene expression VAE via Algorithm 1
Require: $\mathcal{B}$: Set of mini-batches
Require: Initialize Prior Model θ
for each mini-batch $\mathcal{M}$ (batch size=N) in $\mathcal{B}$ **do**
  for each pair $(s, g) \in \{(s_k, g_k), \forall k \in [1, N]\}$ **do**
    $\boldsymbol{z}^{chem} \leftarrow S.Encoder(s)$ {Encode SELFIES string}
    $\boldsymbol{z}^{gene} \leftarrow G_\theta.Encoder(g)$ {Encode gene expression}
    $\hat{\boldsymbol{z}}^{chem} \leftarrow \theta(\boldsymbol{z}^{gene})$ {Map gene expression to chemical space}
    $\ell \leftarrow \mathcal{L}_{\text{RMSE}}(\boldsymbol{z}^{chem}, \hat{\boldsymbol{z}}^{chem})$ {Compute RMSE loss}
  end for
  Compute average loss $\bar{\ell}$ over the mini-batch
  Update weights of Prior Model θ 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: pre-trained and frozen prior model via Algorithm 2
$\boldsymbol{z}^{gene} \leftarrow G_\theta.Encoder(\text{g})$ {Encode gene expression to gene embedding}
$\boldsymbol{z}^{chem} \leftarrow P(\boldsymbol{z}^{gene})$ {Generate chemical embedding using Prior model}
$s_{\text{molecule}} \leftarrow S.Decoder(\boldsymbol{z}^{chem})$ {Decode chemical embedding}
return s_{molecule} {Return molecule represented as SELFIES string}

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

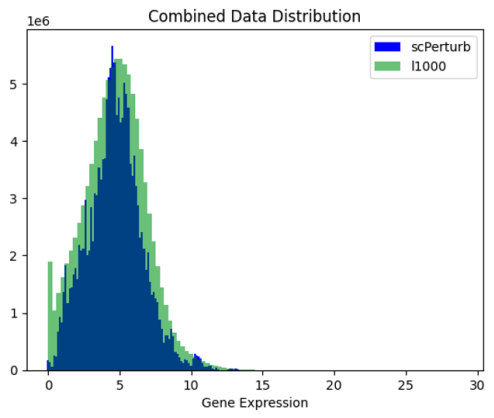


Fig. B1: Mean gene expression signatures after harmonizing single cell dataset with L1000.