arXiv:2511.02769v2 [cs.LG] 22 Sep 2026

STAR-VAE: A Scalable Latent-Variable Transformer for Controllable Molecular Generation

Bum Chul Kwon[1,*], Ben Shapira[3], Moshiko Raboh[3], Shreyans Sethi[2], Shruti Murarka[2], Joseph A. Morrone[2], Leili Zhang[2], Wendy Cornell[2], Pablo Meyer[2], Jianying Hu[2], and Parthasarathy Suryanarayanan[2,*]

[1]IBM Research, Cambridge, MA 02142, USA

[2]IBM Research, Yorktown Heights, NY 10598, USA

[3]IBM Research, Haifa, Israel

*Corresponding Authors:

Bum Chul Kwon bumchul.kwon@us.ibm.com

Parthasarathy Suryanarayanan psuryan@us.ibm.com

Abstract

Many molecular Transformers lack probabilistic latent variables for posterior inference and latent interpolation. We introduce STAR-VAE, a **S**ELFIES-encoded, **T**ransformer-based, **A**uto**R**egressive **V**ariational **A**uto**E**ncoder combining a bidirectional encoder with an autoregressive decoder pretrained on 79 million PubChem molecules. A property signal jointly conditions the prior, posterior, and decoder, while LoRA adapters support fine-tuning on small datasets without modifying the backbone. STAR-VAE achieves 100% validity and near-perfect novelty under unconditional MOSES sampling, the lowest KL divergence on five of ten GuacaMol descriptors, Spearman $\rho = 0.62$ at 98% validity for synthetic-accessibility conditioning, and directional docking-score control for three Tartarus protein targets. Across four ChEMBL targets, seed-based posterior sampling recovers target-associated held-out scaffolds while label-conditioned sampling produces structurally diverse outputs. Code is available at https://github.com/BiomedSciAI/STAR-VAE.

Keywords

STAR-VAE; variational autoencoder; molecular generation; SELFIES; Transformer; conditional generation; LoRA; latent variable model; drug discovery

Introduction

The space of synthesizable small molecules is estimated to exceed 10^{33} compounds [1], far beyond what can be exhaustively enumerated or tested experimentally. Generative modeling offers a principled route through this space, enabling data-driven exploration focused on molecules with desirable properties [2, 3]. Deep generative models have consequently become central to de novo drug design, with a wide range of architectures now proposed for molecular generation [4, 5]. Useful molecular generative models must learn broad chemical distributions, support label-guided generation, and remain computationally tractable.

Variational autoencoders (VAEs) were among the first deep generative models applied to molecules and remain influential for their ability to produce smooth latent spaces amenable to interpolation and optimization. Gómez-Bombarelli et al. [6] showed that a convolutional encoder and gated recurrent unit (GRU) decoder operating on SMILES strings could navigate chemical space via gradient-based latent optimization. Graph-based approaches [7–9] moved to direct graph generation, and conditional VAEs [10–12] extended the paradigm to label-guided synthesis. Despite this breadth, early molecular VAEs relied on recurrent or graph-based architectures and were trained on relatively small datasets, leaving open the question of whether the VAE paradigm can scale to modern data regimes.

Subsequent work shifted toward large Transformer architectures. Models such as MolGPT [13] and MolGen [14] generate SMILES [15] or SELFIES one token at a time. They achieve high validity and diversity but do not use an explicit variational latent space for posterior inference or latent interpolation. Chemformer [16], SELF-BART [17], and SELFIES-TED [18] use encoder–decoder architectures with denoising pretraining. They support several molecular prediction and generation tasks, but they also lack an explicit variational latent distribution. Adversarial, flow, and diffusion methods [19–26] provide other approaches to molecular generation. Target-aware graph models [27] address goal-oriented generation [28]. Large training sets and large decoders have improved validity and diversity, while limited labeled data remain a challenge for conditional generation.

We revisit the latent-variable paradigm and ask whether a modernized VAE, scaled to 79 million pretraining molecules and paired with principled conditioning, can reproduce properties of the MOSES and GuacaMol benchmark distributions and support controlled generation. We present STAR-VAE (**S**elfies-

ENCODED, **T**RANSFORMER-BASED, **A**UTO**R**EGRESSIVE **V**ariational **A**uto**E**ncoder), which integrates a bidirectional Transformer encoder and an autoregressive Transformer decoder with a variational latent space. The model is pretrained on 79 million drug-like PubChem molecules represented as SELFIES [29]. This design supports posterior inference, latent-space sampling, and property-conditioned generation in one model. A single scalar label conditions the latent prior, inference network, and decoder jointly; the current implementation conditions on one property at a time. Label sources differ by experiment: SA score and docking score are computed directly with RDKit and QVina respectively, while the rediscovery experiments use MMELON-derived pseudo-labels. LoRA modules in both encoder and decoder enable label-conditioned fine-tuning on small datasets with modest compute.

This work contributes a Transformer-based variational model pretrained at scale on SELFIES and a conditional formulation in which one property signal shapes the prior, approximate posterior, and decoder. LoRA supports adaptation to smaller labeled datasets. On MOSES [30] and GuacaMol [31], STAR-VAE shows high validity and novelty and obtains the lowest KL divergence for five of ten molecular properties among the four retrained GuacaMol baselines. Quantitative conditioning experiments show control of synthetic accessibility and docking scores for all three Tartarus protein targets [32]. The SA-score comparison with REINVENT 4, MolGPT, and DrugDiff characterizes trade-offs among different conditioning approaches. Compared with MolGPT under the shared continuous-conditioning protocol, STAR-VAE produces a higher fraction of valid molecules and more unique scaffolds, while MolGPT tracks requested SA scores more closely among its valid outputs. A five-factor ablation study with three independent fine-tuning runs per selected configuration shows that measurable control requires a conditioning path to the decoder and that differences among conditioning-capable variants fall within run-to-run variation. Finally, experiments on four ChEMBL targets [33] show complementary generation modes: seed-based VAE sampling recovers more target-associated held-out scaffolds, while label-conditioned CVAE sampling produces more varied structures.

Methods

Data

The objective of the pretraining dataset is to learn a task-agnostic molecular representation through large-scale self-supervision. We use PubChem [34], a public chemical database curated by the National Center for Biotechnology Information (NCBI) that aggregates data on small molecules from over 870 contributing sources, including substances, bioassays, targets, and patents. From this resource, we curate a drug-like

subset of approximately 79 million molecules. The curation process involves extracting the largest molecular fragment, removing duplicates, and applying standard drug-likeness filters: MW $\leq$ 600 Da, $\leq$5 H-bond donors, $\leq$10 acceptors, $\leq$10 rotatable bonds. These filters yield a dataset covering broad chemical diversity while remaining within drug-relevant physicochemical ranges.

Molecular Representation

We represent molecules using the SELFIES (Self-Referencing Embedded Strings) representation [29]. SELFIES substantially reduces syntax-related generation failures because valid SELFIES strings decode to molecular graphs; downstream conversion and RDKit sanitization can nevertheless reject some generated structures, so we report empirical RDKit validity as the operational metric throughout. Compared to SMILES, SELFIES removes most syntactic failure modes and is therefore well-suited for sequence-based generative modeling at scale.

Tokenization. Each molecule is converted into a SELFIES string and then split into discrete tokens drawn from a fixed vocabulary. Padding, start-of-sequence, end-of-sequence, and unknown tokens are added for sequence handling during training and generation. The resulting vocabulary covers atoms, branches, rings, and stereochemical markers encountered in the dataset. For efficient batching and consistent positional encoding, we limit sequence length to a maximum of 71 tokens, which covers over 99% of drug-like molecules in our training dataset. Shorter sequences are padded, while longer ones are truncated.

Positional encoding. Absolute positional encodings are used throughout; the fixed 71-token maximum makes rotary or relative schemes unnecessary.

Model Architecture

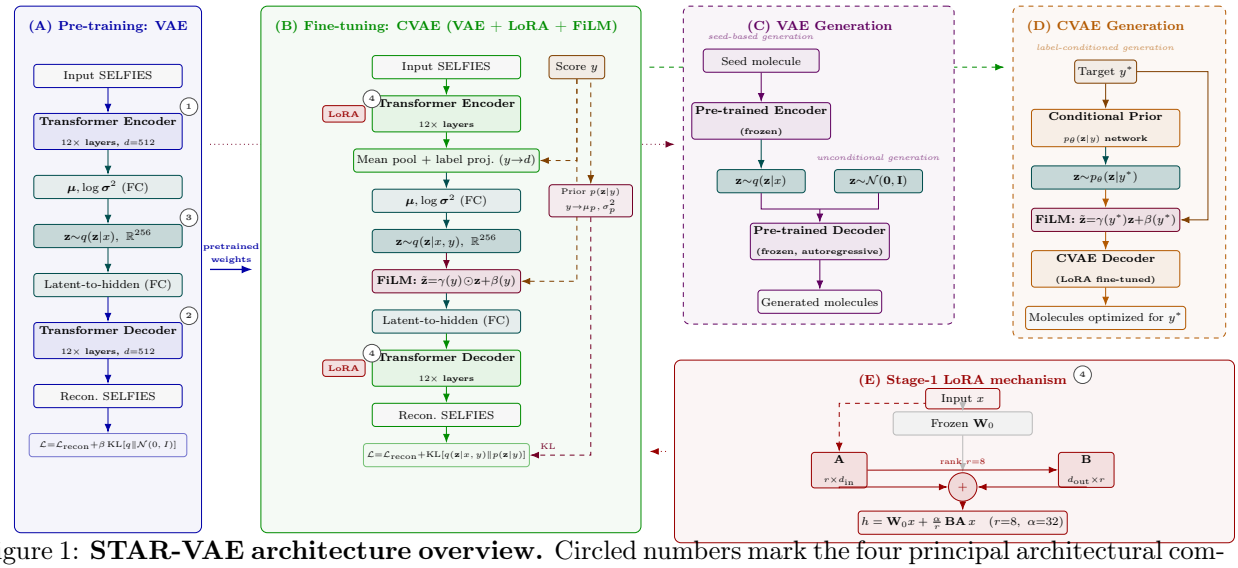


Figure 1: **STAR-VAE architecture overview.** Circled numbers mark the four principal architectural components: (1) bidirectional Transformer encoder, (2) autoregressive Transformer decoder, (3) latent-variable bottleneck, and (4) LoRA-based adaptation. **(A)** Pretraining: a Transformer-based VAE (12-layer encoder/decoder, d=512, latent $\mathbb{R}^{256}$) is trained on 79 M SELFIES molecules with a reconstruction + KL-divergence objective. **(B)** Fine-tuning: CVAE (VAE + LoRA + FiLM): the pretrained backbone is frozen and augmented with LoRA adapters (r=8, α=32) on the attention output projections and feed-forward projections of the encoder and decoder layers. A FiLM module modulates the latent vector with the property score y; a learned conditional prior $p(\mathbf{z}|y)$ replaces the isotropic Gaussian prior in the KL term. Which of these components is trainable differs by experiment (Table 1). **(C)** VAE seed-based generation: a seed molecule is encoded to obtain $\mathbf{z}\sim q(\mathbf{z}|x_{\text{seed}})$, which is then decoded autoregressively to produce structurally similar molecules. **(D)** CVAE label-conditioned generation: the requested label y^* is passed to the conditional-prior network, which parameterizes $p_\theta(\mathbf{z}|y^*)$; a latent sampled from this prior is FiLM-modulated and decoded autoregressively. The encoder is not used in this protocol. **(E)** LoRA adapter mechanism (stage 1): each frozen weight $\mathbf{W}_0$ is augmented by a trainable low-rank perturbation $\frac{\alpha}{r}\mathbf{BA}$ (r=8, scaling =4). The panel illustrates the LoRA-specific update; per-experiment trainable modules differ, and the reported SA-score and docking models use a second stage in which LoRA is frozen and only the conditioning pathway is updated (Table 1).

Alt text: Five-panel diagram of the STAR-VAE architecture. Panel A: pretraining; a bidirectional Transformer encoder compresses SELFIES molecules to a 256-dimensional latent vector decoded autoregressively. Panel B: conditional fine-tuning; Transformer states are mean-pooled, the projected property score is added to the pooled representation before the posterior heads, and decoder-side Feature-wise Linear Modulation conditions the latent. Panels C and D: VAE generation via seed encoding or Gaussian prior sampling; Conditional VAE generation in which the requested label y-star is passed to the conditional-prior network, which supplies the latent vector. Panel E: stage-1 Low-Rank Adaptation rank-8 decomposition. Per-experiment trainable modules are reported in Table 1.

STAR-VAE (**S**ELFIES-ENCODED, **T**RANSFORMER-BASED, **A**UTO**R**EGRESSIVE **VAE**) consists of a bidirectional Transformer encoder, a latent-variable bottleneck, and an autoregressive Transformer decoder (Figure 1). Both encoder and decoder use absolute positional encoding, consistent with the fixed maximum

SELFIES length of 71 tokens. This architecture employs 12 encoder and 12 decoder layers, 8 attention heads, a feed-forward width of 2048, dropout 0.1, and a latent dimensionality of 256. Model size varies with the components used in each experiment. The pretrained unconditional backbone contains 89.18 M parameters. Adding the label projection, conditional-prior networks, and FiLM modulation increases the total to 89.77 M parameters. Adding encoder and decoder LoRA adapters increases it to 91.05 M parameters. Table 1 gives the total and trainable counts for every experiment.

Encoder. The encoder is a stack of Transformer layers operating on SELFIES token embeddings; contextualized hidden states are mean-pooled over sequence positions to produce the mean and log-variance of a Gaussian latent distribution. In the label-conditioned configurations, a learned linear projection of the scalar label is added to the pooled representation before the latent mean and log-variance heads, so that the encoder parameterizes $q_\phi(\mathbf{z}|x, y)$. LoRA adapters may additionally be applied within the encoder layers to adapt the backbone to the task (Table 1).

Latent bottleneck. The latent vector is sampled via the reparameterization trick, linearly projected to the decoder width, and repeated across memory positions to serve as decoder cross-attention memory.

Decoder. The autoregressive Transformer decoder generates SELFIES tokens one at a time, conditioned on previously generated tokens and on the latent-derived cross-attention memory. In label-conditioned configurations, the property value enters through FiLM modulation of the latent vector before it is projected to decoder memory: a small network maps the scalar label to per-dimension scale and shift parameters $(\gamma(y), \beta(y))$, and the modulated latent is $\gamma(y) \odot \mathbf{z} + \beta(y)$. LoRA adapters may be applied in the decoder layers to adapt the backbone while keeping the majority of parameters frozen.

Pretraining Setup

The model is trained with Adam (lr $= 10^{-5}$, batch 256) on 8 GPUs for approximately 72 hours, with no learning-rate schedule. The KL coefficient follows a linear warmup from 0 to $\beta = 1.0$ over the first 10,000 optimizer steps and is held at $\beta = 1.0$ thereafter. The cyclical β annealing and free-bits options described in the Ablation Studies apply to conditional fine-tuning, not to this pretraining run.

Conditional Generative Formalism

After pretraining, we fine-tune the model on a label-guided objective where label y encodes a property of interest (activity, ADMET endpoint, or multi-component objective). We formalize conditional generation

via three distributions:

$$p_\theta(z \mid y) = \text{prior over latents given property } y, \tag{1}$$

$$q_\phi(z \mid x, y) = \text{approximate posterior inferred from molecule } x \text{ and property } y, \tag{2}$$

$$p_\theta(x \mid z, y) = \text{autoregressive decoder generating molecule } x \text{ given } (z, y). \tag{3}$$

The inference network is modeled as a Gaussian:

$$q_\phi(z \mid x, y) = \mathcal{N}\big(\mu_\phi(x, y),\ \mathrm{diag}(\sigma^2_\phi(x, y))\big)\,, \tag{4}$$

where μ_ϕ and σ_ϕ are outputs of the Transformer encoder. The label-conditioned prior is also Gaussian:

$$p_\theta(z \mid y) = \mathcal{N}\big(\mu_\theta(y),\ \mathrm{diag}(\sigma^2_\theta(y))\big)\,, \tag{5}$$

with parameters predicted from the property embedding y.

The decoder defines the conditional likelihood:

$$p_\theta(x \mid z, y) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}, z, y)\,. \tag{6}$$

Training maximizes the conditional evidence lower bound (ELBO):

$$\mathcal{L}(x, y) = \mathbb{E}_{q_\phi(z|x,y)}\big[\log p_\theta(x \mid z, y)\big] - \beta\,\mathrm{KL}[q_\phi(z \mid x, y) \,\|\, p_\theta(z \mid y)]\,, \tag{7}$$

where β controls the strength of the KL regularization.

LoRA-based task adaptation. To adapt the pretrained backbone to each conditional task, we insert LoRA adapters into the Transformer encoder and decoder layers. Adapters are inserted at the attention *output* projections (W_O in encoder self-attention, and in both decoder self-attention and cross-attention) and at the two position-wise feed-forward projections of each layer. The fused query/key/value projection of each attention block is left unmodified. One implementation detail materially qualifies this description. The attention modules dispatch to a fused attention kernel that consumes the output-projection weight directly rather than calling the wrapped module, so the adapters inserted at W_O are never evaluated in the forward pass. Their B matrices remain at their zero initialization and receive no gradient. We verified this by

gradient inspection and by checking the released checkpoints, in which all attention-output B matrices are exactly zero while the feed-forward B matrices are not. In effect, therefore, LoRA adapts the position-wise feed-forward projections only. These inert adapters account for 294,912 of the 1,277,952 LoRA parameters (23%), so the reference ablation configuration trains approximately 1.64 M effective parameters rather than the 1.94 M reported as trainable. Parameter counts elsewhere in this paper are reported as instantiated, since that is what the checkpoints contain. A unit test accompanying the released code pins this behaviour. The attention-output LoRA implementation issue does not alter the reported results because those adapters remained at zero and were inactive. Effective feed-forward LoRA was trained in Stage 1, inherited and frozen in Stage 2, and remains active in the decoder during inference (Table 1). For each adapted linear map $W_\bullet$,

$$\tilde{W}_\bullet = W_\bullet^{(0)} + \tfrac{\alpha}{r}\, B_\bullet A_\bullet, \tag{8}$$

where $W_\bullet^{(0)}$ are the frozen pretrained weights, $A_\bullet \in \mathbb{R}^{r\times d_{\text{in}}}$, and $B_\bullet \in \mathbb{R}^{d_{\text{out}}\times r}$ with $r \ll d$. LoRA adapts the task-specific weights; it does not itself introduce the input-dependent property value. The property value enters through the label-conditioned prior, the label projection in the approximate posterior, and the decoder-side FiLM modulation described above. The scaling factor α/r is fixed and embedded into the trained matrices in accordance with standard LoRA practice [35]. In all experiments, $r = 8$ and $\alpha = 32$, resulting in a fixed scaling factor of 4.

Per-Experiment Configurations

The conditional components described above are not all trainable, or even all present, in every experiment. Table 1 therefore reports, for each reported model, which modules are trainable, whether the label enters the approximate posterior, whether the label-conditioned prior exists and is trained, whether it is sampled from at generation time, and the exact trainable and total parameter counts.

Two aspects of this table warrant emphasis, because they qualify how the reported results should be interpreted. First, the SA-score and docking models were adapted in two stages. LoRA adapters, FiLM, and the latent-to-hidden projection were trained in the first stage. Those weights were then frozen, and only the conditioning pathway, including the conditional prior, was trained in the second stage. The reported models are the product of both stages, and the conditional-prior parameters are updated only in the second. Second, the ablation study samples latents from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ rather than from $p_\theta(\mathbf{z}|y)$ for all variants, so it characterizes the FiLM and LoRA conditioning pathways but does not isolate the contribution of the learned conditional prior.

Table 1: **Per-experiment model configuration, conditioning pathway, and parameter counts.** Trainable and total parameter counts are derived from the freezing logic in the released code and cross-checked against training logs. "Post. label" indicates whether the scalar label is projected into the pooled encoder representation before the latent heads, i.e. whether the encoder parameterizes $q_\phi(\mathbf{z}|x, y)$ rather than $q_\phi(\mathbf{z}|x)$. "Prior" columns distinguish whether the label-conditioned prior $p_\theta(\mathbf{z}|y)$ exists, whether its parameters are updated, and whether it is sampled from at generation time. Where the prior exists but is frozen, it contributes to the KL term with randomly initialized parameters. The SA-score and docking models were adapted in two stages (see footnotes).

		LoRA		Post.	Conditional prior			Parameters	
Model / experiment	FiLM	Enc.	Dec.	label	exists	trained	sampled	Trainable	Total
Pretrained VAE backbone	–	–	–	–	–	–	–	89,175,368	89,175,368
SA-score and docking conditional models (two-stage adaptation)									
Stage 1: LoRA adaptation[a]	✓	✓	✓	✓	✓	–	–	1,936,384	91,047,752
Stage 2: SA score (reported)[b]	✓	frozen	frozen	✓	✓	✓	✓	726,016	91,047,752
Stage 2: 1SYH / 4LDE / 6Y2F[b]	✓	frozen	frozen	✓	✓	✓	✓	726,016	91,047,752
Rediscovery CVAE[c]	✓	–	✓	–	–	–	–	1,346,560	90,391,368
Ablations (SA-score setting; all generate from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, not from $p_\theta(\mathbf{z}|y)$)[d]									
REF (reference)	✓	✓	✓	✓	✓	–	–	1,936,384	91,047,752
A2 deterministic AE	✓	✓	✓	✓	✓	–	–	1,936,384	91,047,752
A3 encoder-only conditioning	–	✓	–	✓	✓	–	–	721,408	89,832,776
A3 decoder-only conditioning	✓	–	✓	–	✓	–	–	1,346,560	90,457,928
A3 no conditioning[e]	–	✓	✓	✓	–	–	–	1,409,536	90,454,344
A4 full fine-tuning	✓	–	–	✓	✓	✓	–	89,769,800	89,769,800
A4 frozen backbone	✓	–	–	✓	✓	✓	–	988,672	89,769,800
A5 fixed β / linear β	✓	✓	✓	✓	✓	–	–	1,936,384	91,047,752

[a] Counts are instantiated counts. Attention-output LoRA adapters are bypassed by the fused attention kernel; 294,912 both-side LoRA parameters are therefore inert. Stage 1 trains the effective feed-forward LoRA, FiLM, and latent-to-hidden parameters, while the conditional prior and posterior label projection remain frozen.

[b] The reported SA-score and docking models add a second stage in which all Transformer and LoRA weights are frozen and only FiLM, the conditional-prior networks, latent-to-hidden projection, and label projection are updated. The posterior label projection is updated during Stage-2 training but is inactive during label-only generation because the encoder is not used.

[c] The rediscovery CVAE uses decoder-only LoRA, no conditional prior, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ followed by FiLM; counts follow the released configuration because the original checkpoints were unavailable.

[d] All ablations generate from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, so the study does not isolate conditional-prior sampling.

[e] The "no conditioning" configuration disables the prior and FiLM but still routes training through a frozen randomly initialized label projection; labels have no effect at generation. In rows marked "–" under FiLM, no modulation module is constructed.

Generation Protocols

STAR-VAE supports three generation protocols, distinguished by what information is available at inference time.

Protocol I: Unconditional generation (VAE). The latent vector is sampled directly from the standard Gaussian prior: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. No seed molecule and no property label are required. The pretrained decoder then generates molecules autoregressively from $\mathbf{z}$. This protocol is used for the MOSES and GuacaMol "Random" unconditional benchmarks.

Protocol II: Seed-based generation (VAE). A seed molecule x_{seed} is first encoded by the pretrained encoder to obtain a posterior sample $\mathbf{z} \sim q_\phi(\mathbf{z}|x_{\text{seed}})$. The pretrained decoder then generates new molecules from $\mathbf{z}$, producing outputs that are structurally similar to the seed. No property label is used. This protocol is used for the GuacaMol "Seed" experiment and for the VAE arm of the rediscovery evaluation.

Protocol III: Label-conditioned generation (CVAE). A target label y^* (such as a target-binding activity level or an ADMET endpoint) is the sole input at inference time; no seed molecule is encoded. The label is passed to the conditional-prior network, which parameterizes $p_\theta(\mathbf{z}|y^*)$; a latent vector is sampled from this prior, FiLM-modulated, and decoded autoregressively. The encoder is not used in this protocol, since it represents $q_\phi(\mathbf{z}|x, y)$ and requires a molecule. This protocol is used for the Tartarus protein-ligand design tasks. For the CVAE arm of the rediscovery evaluation, the conditional prior was not enabled in the fine-tuned model. Latents were instead drawn from the isotropic Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and FiLM-modulated with the requested label before decoding. Table 1 reports this configuration.

Label-reference decoder guidance. At inference, we can strengthen the effect of the requested property label by combining two decoder predictions. We first draw one latent vector from the conditional prior, $\mathbf{z} \sim p_\theta(\mathbf{z}|y^*)$. We then apply FiLM to this same latent vector with the requested label y^* and with the reference label $y_\varnothing = 0$. At each decoding step, both branches use the same token prefix and produce logit vectors $\boldsymbol{\ell}_{y^*}$ and $\boldsymbol{\ell}_\varnothing$. We combine them as

$$\boldsymbol{\ell}_{\text{guided}} = \boldsymbol{\ell}_\varnothing + \alpha \left(\boldsymbol{\ell}_{y^*} - \boldsymbol{\ell}_\varnothing \right), \tag{9}$$

where $\alpha \geq 1$ controls the guidance strength. At $\alpha = 1$, the equation gives the prediction for the requested label. Larger values increase the difference between the requested-label and reference-label predictions. The decoder selects the token with the highest guided logit at each step.

This procedure is inspired by classifier-free guidance [36], but it is not the same method. The models were trained without condition dropout, and $y_{\varnothing} = 0$ is a real endpoint of the property scale rather than a learned null label. Both branches also share the same latent vector and token prefix. The procedure therefore adjusts the decoder logits relative to a reference property value without adding latent sampling noise.

Results

Unconditional Generation

Table 2: Comparison of models on the MOSES benchmark using seven key metrics. Modern baselines (SAFE-GPT, DiGress, FragFM, MolHIT, GVT) do not report all metrics. Columns that are uniformly absent for modern baselines have been omitted here (see Table S1 in the Supporting Information for the full 15-column table). Baseline values for classic models are from the official MOSES repository [30]. Modern baseline values are from the cited publications. STAR-VAE runs are means ± SD over three independent evaluation seeds. **Bold** indicates the best value per column across all models with reported data. Column headers: Validity (Valid), Uniqueness (Unique@10k), Fréchet ChemNet Distance (FCD/Test, lower is better), Nearest Neighbor Similarity (SNN/Test), Scaffold Similarity (Scaf/Test), drug-likeness filter pass rate (Filters), and Novelty.

Models	Valid (↑)	Unique@10k (↑)	FCD/Test (↓)	SNN/Test (↑)	Scaf/Test (↑)	Filters (↑)	Novelty (↑)
Classic baselines							
HMM	0.076	0.5671	24.4661	0.3876	0.2065	0.9024	0.9994
NGram	0.2376	0.9217	5.5069	0.5209	0.5302	0.9582	0.9694
Combinatorial	**1.0**	0.9909	4.2375	0.4514	0.4445	0.9557	0.9878
CharRNN	0.9748	0.9994	**0.0732**	0.6015	0.9242	**0.9943**	0.8419
AAE	0.9368	0.9973	0.5555	0.6081	0.9022	0.9960	0.7931
VAE	0.9767	0.9984	0.0990	0.6257	**0.9386**	0.9970	0.6949
JTN-VAE	**1.0**	0.9996	0.3954	0.5477	0.8964	0.9760	0.9143
LatentGAN	0.8966	0.9968	0.2968	0.5371	0.8867	0.9735	0.9498
Modern baselines							
SAFE-GPT [37]	0.998	0.989	0.72	0.57	0.063	0.977	0.437
DiGress [38]	0.857	**1.000**	1.19	0.52	0.148	0.971	0.950
FragFM [39]	0.998	**1.000**	0.58	0.56	0.109	0.991	0.871
MolHIT [40]	0.991	0.998	1.03	0.55	0.144	0.980	0.914
GVT [41]	0.994	**1.000**	0.16	**0.64**	0.034	0.996	0.614
Ours (STAR-VAE)							
PubChem 80M (Random)	**1.0000**±0.0000	**1.0000**±0.0000	14.1140±0.0607	0.3140±0.0002	0.2831±0.0095	0.4766±0.0014	**0.9999**±0.0000
PubChem 80M (Seed)	**1.0000**±0.0000	0.9998±0.0000	2.0062±0.0084	0.5090±0.0007	0.8987±0.0029	0.8313±0.0013	0.7583±0.0010
MOSES 1.5M (Random)	**1.0000**±0.0000	0.9999±0.0001	9.1489±0.0291	0.3842±0.0003	0.5622±0.0120	0.6398±0.0029	0.9939±0.0002
MOSES 1.5M (Seed)	**1.0000**±0.0000	0.9996±0.0002	6.6121±0.0335	0.4263±0.0003	0.7583±0.0057	0.6931±0.0018	0.9567±0.0004

To establish a task-specific baseline alongside the PubChem-80M model, we trained a smaller variant of STAR-VAE (denoted STAR-VAE-MOSES) directly on the official MOSES training set. The architecture

uses $d_{\text{model}} = 256$, $n_{\text{head}} = 4$, 4 encoder layers, 4 decoder layers, a feedforward dimensionality of 1,024, and a 128-dimensional latent bottleneck, yielding approximately 7.7 million parameters (versus 89.18 million for the PubChem-80M model). Training used SMILES converted to SELFIES, with maximum sequence length 74 tokens, for 15 epochs with the Adam optimizer, learning rate 0.0002, and batch size 256 on a single GPU. For evaluation, we generate exactly 30,000 molecules following the official MOSES benchmark experiments [30]. Protocol I (Random) draws latent vectors from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Protocol II (Seed) encodes 30,000 randomly selected training molecules, adds a small perturbation around each posterior mean, and decodes one molecule per seed. The resulting molecules are evaluated offline against the official MOSES splits.

On the MOSES [30] benchmark, we evaluate STAR-VAE under two pretraining datasets (PubChem 80M and MOSES 1.5M) and two generation protocols (Unconditional Random Prior sampling and Seed-Conditioned perturbation), with means and standard deviations over three independent evaluation seeds (Table 2).

Under the **Unconditional (Random)** setting ($\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$), both models achieve perfect Validity (1.0000 ± 0.0000), near-perfect Uniqueness@10K, and high Novelty, with IntDiv1 $= 0.8990 \pm 0.0001$ for PubChem, above all classic baselines. FCD/Test is higher (14.1140 ± 0.0607 and 9.1489 ± 0.0291) and scaffold similarity is lower, reflecting unconstrained sampling from a prior trained on PubChem rather than on the MOSES corpus.

Under the **Seed-Conditioned** protocol, 30,000 randomly selected training SMILES are encoded and their latent vectors perturbed with $\epsilon \sim \mathcal{N}(0, I)$ before decoding. STAR-VAE PubChem 80M (Seed) reduces FCD/Test from 14.1140 to 2.0062 ± 0.0084, a substantial improvement over the Random prior but still above the classic baselines trained on the MOSES corpus (CharRNN: 0.0732, AAE: 0.5555). Nearest Neighbor Similarity increases from 0.3140 to 0.5090 ± 0.0007. Fragment Similarity and Scaffold Similarity reach 0.9964 ± 0.0001 and 0.8987 ± 0.0029. Generation remains within the topological neighborhood of the seeds without mode collapse (Unique@10K $= 0.9998 \pm 0.0000$).

Under seed conditioning, STAR-VAE PubChem 80M outperforms the MOSES 1.5M model on FCD/Test (2.0062 ± 0.0084 vs. 6.6121 ± 0.0335), SNN/Test (0.5090 ± 0.0007 vs. 0.4263 ± 0.0003), and filter pass rate (0.8313 ± 0.0013 vs. 0.6931 ± 0.0018). This comparison should be read with two asymmetries in view, both of which favour the MOSES-1.5M arm. First, that model was trained on the official MOSES training split and used the official MOSES test split for validation and checkpoint selection, so the reported FCD/Test and SNN/Test are computed against molecules it selected checkpoints on. Second, the seed protocol draws its 30,000 seeds from the MOSES training split, which the MOSES-1.5M model saw directly during training. STAR-VAE PubChem 80M received no parameter updates on any MOSES split, but an overlap audit (Appendix F, Table S5) shows that 71.45% of MOSES training molecules and 71.18% of MOSES test molecules

match the full curated PubChem source corpus, providing upper bounds on exposure to this chemistry. The defensible reading is therefore narrow. Corpus-scale pretraining yields a latent manifold that supports better local sampling around a supplied seed than a model trained on the benchmark distribution itself, even though the smaller model had direct access to both the seeds and the evaluation split. Neither arm demonstrates transfer to chemistry absent from pretraining.

GuacaMol evaluates distributional similarity via KL divergence across ten RDKit physicochemical descriptors between generated and reference molecules; higher scores indicate closer alignment (Table 3). Four baselines (CharRNN [42], VAE [43], AAE [44], MolGPT [45]) were retrained on the official GuacaMol split and each sampled for 10,000 molecules, providing full per-descriptor KL values for Table 3. Four additional baselines (DiGress [38] 0.929, MolHIT [40] 0.967, FragFM [39] 0.974, GVT [41] 0.996) are taken from their respective publications, which report only a 5-column summary (Validity, Uniqueness, Novelty, KL Divergence Score, FCD). Per-descriptor breakdowns are not available, so a row-by-row comparison with Table 3 is not possible for these models. STAR-VAE (Random) scores 0.9157 and STAR-VAE (Seed) scores 0.9979 on the same overall KL Divergence Score scale.

Table 3: Molecular physicochemical distribution scores on the GuacaMol benchmark. The first row (KL Divergence Score) is the overall GuacaMol score $\in [0, 1]$ (*higher is better*↑), computed as the mean of exp(−KL) across ten descriptors. The remaining rows report raw per-descriptor KL divergence values (*lower is better*↓). Bold indicates the best value per row. STAR-VAE (Random) samples from the pretrained PubChem prior; STAR-VAE (Seed) samples from latent encodings of GuacaMol training molecules.

Metric	**CharRNN**	**VAE**	**AAE**	**MolGPT**	**STAR-VAE (ours)**	
					Random	Seed
KL Divergence Score[†] (↑)	.8894	.9173	.9356	.8817	.9157	**.9979**
BertzCT	.3460	.2617	.1660	.3926	**.0274**	**.0012**
MolLogP	.0684	.0723	.0714	.0744	**.0105**	**.0026**
MolWt	.2337	.1986	.1358	.2804	**.0076**	**.0018**
TPSA	.1092	.0725	.0643	.1051	**.0052**	**.0015**
NumHAcceptors	.1414	.0852	.0687	.1549	**.0055**	**.0008**
NumHDonors	.0234	.0203	.0223	.0149	.0199	**.0005**
NumRotatableBonds	.0075	.0141	**.0067**	.0196	.0083	.0053
NumAliphaticRings	.0116	.0069	.0107	.0117	.0414	**.0035**
NumAromaticRings	.2047	.1355	.1027	.1779	.1846	**.0007**
DiffInternalSimilarities	.0797	.0280	.0298	.0943	.7993	**.0031**

[†] The overall KL Divergence Score (↑) is the mean of $\exp(-\mathrm{KL}_i)$ across all ten descriptors, normalized to $[0, 1]$. Per-descriptor raw KL values (↓): lower is better.

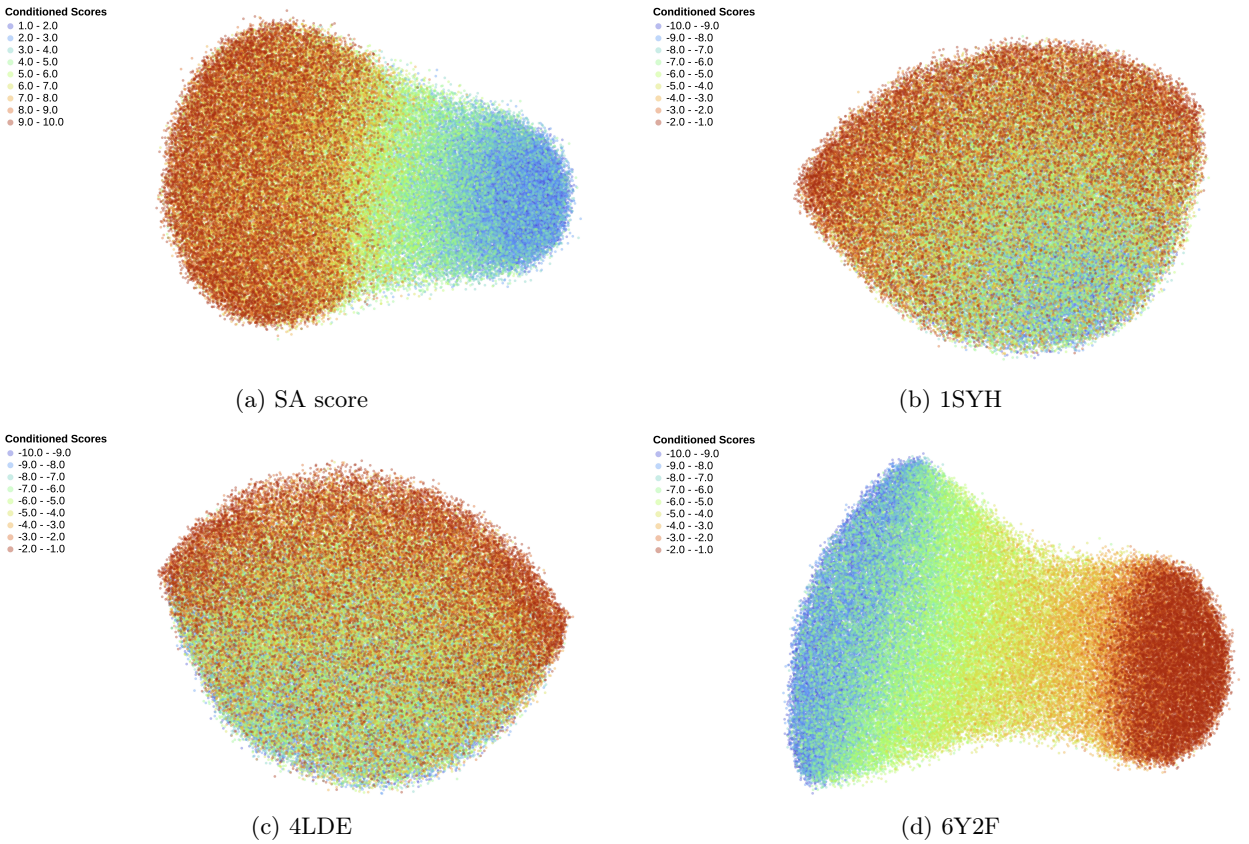
Under the GuacaMol evaluation protocol, STAR-VAE, pretrained on the full PubChem dataset in contrast to the benchmark's ChEMBL-derived training set, attains an overall KL-divergence score of 0.9157

(Table 3, "Random" column). The model achieves the lowest KL divergence among the four retrained baselines on five descriptors: BertzCT (0.0274), MolLogP (0.0105), MolWt (0.0076), TPSA (0.0052), and NumHAcceptors (0.0055). Two descriptors show substantially higher KL divergence than all baselines: NumAromaticRings (0.1846, vs. AAE best of 0.1027) and DiffInternalSimilarities (0.7993, vs. VAE best of 0.0280). DiffInternalSimilarities measures the per-molecule spread of pairwise Tanimoto similarities to the rest of the generated set. The broader PubChem chemical space yields a wider spread than ChEMBL24, raising this KL divergence.

The higher internal dissimilarity under GuacaMol reflects the distributional mismatch between the PubChem pretraining corpus and the ChEMBL24 reference set. To quantify this, we generated 10,000 molecules from latent encodings of randomly selected GuacaMol training-set molecules (the "Seed" column). This setting yields the lowest KL divergences across all descriptors. Because the seeds are drawn from the GuacaMol training set itself, this protocol measures reconstruction within a supplied domain rather than unconditional distribution matching, and its scores are not comparable to unconditional benchmark results. The overlap audit further shows that 87.04% of GuacaMol training molecules match the full curated PubChem source corpus (Appendix F), providing an upper bound on exposure to the reconstructed domain. The result establishes that seed-based posterior sampling can localize generation to a narrower property distribution without fine-tuning. It does not establish generalization to chemistry absent from pretraining.

Conditional Generation

Hereafter, "CVAE" denotes a fine-tuned conditional model and "pretrained VAE" denotes the base model. The sampling protocol (random prior or seed-based posterior) is stated separately. We evaluate conditioning on two directly computable scores: the RDKit SA score for synthetic accessibility and the QVina docking score for protein–ligand binding.



(a) SA score (b) 1SYH

(c) 4LDE (d) 6Y2F

Figure 2: UMAP projections of the CVAE conditional prior colored by property score. **(a)** SA score: molecules colored by SA score bin, from low SA score (easier to synthesize) to high SA score (harder to synthesize); adjacent bins overlap continuously. **(b–d)** Docking score for targets 1SYH, 4LDE, and 6Y2F: molecules colored from dark blue (docking score near 0, weaker predicted binding) to yellow (most negative docking score, strongest predicted binding). In all four panels, property-score levels occupy visibly different latent regions with smooth overlap at boundaries, showing score-associated organization in the two-dimensional projection. UMAP is a nonlinear embedding and is used here for visualization only; the quantitative evidence for conditioning is the rank correlation and calibration analysis in Table 4, not this projection.

Alt text: Four Uniform Manifold Approximation and Projection scatterplots of the latent space. Panel a: synthetic accessibility score conditioning; molecules colored from low SA score (easier to synthesize) to high SA score (harder to synthesize), with smooth overlap between adjacent bins. Panel b: target 1SYH (GluR2) docking score; dark blue indicates weak binding near 0 kcal/mol, yellow indicates strong binding near -10 kcal/mol. Panel c: target 4LDE (beta-2 adrenoceptor) docking score; analogous color-coded clustering. Panel d: target 6Y2F (SARS-CoV-2 main protease) docking score; analogous clustering. All panels show property-level separation with smooth boundary overlap.

SA score. We fine-tuned a CVAE on 9,990 PubChem molecules labeled with RDKit SA scores [46], normalized to $[0, 1]$ via min-max over the observed range $[1.44, 7.51]$ ($y=0$: easiest synthesis, $y=1$: hardest). In this final adaptation stage, we updated only the FiLM module, conditional-prior mean and log-variance

networks, latent-to-hidden projection, and label projection (726,016 of 91,047,752 parameters, 0.80%). The Transformer encoder and decoder, their LoRA adapters, and the posterior heads remained frozen. LoRA adapters were trained in a preceding adaptation stage. Table 1 gives the full configuration. Figure 2(a) shows a UMAP of the conditional prior colored by SA score bin: low-SA (easy) and high-SA (hard) molecules occupy distinct regions with smooth overlap at the boundary.

Docking-score conditioning. We used the protein-ligand design subset of the Tartarus benchmark [32], which provides QVina docking scores against three targets: 1SYH (GluR2/GluA2 ligand-binding core), 4LDE (β_2 adrenoceptor), and 6Y2F (SARS-CoV-2 main protease). Lower docking scores indicate stronger predicted binding.

The Tartarus training corpus contains 105,338 molecules with docking scores for all three targets. For 6Y2F all molecules dock with negative scores, so min-max normalization over the full set spans $[0, 1]$ without compression. For 1SYH and 4LDE, however, a small fraction of molecules yield large positive docking scores (up to +255.6 and +18.6 kcal/mol respectively), compressing the labels of the binding-relevant negative-score molecules into only [0, 0.037] and [0, 0.346] of the unit interval. We corrected this by re-normalizing within the negative-score subset for each target, mapping $[s_{\min}, 0]$ linearly to [0, 1] ($s_{\min}$: most negative docking score in training set) and discarding positive-score molecules from *training-label construction*. At evaluation, by contrast, a generated molecule is treated as a completed docking calculation whenever QVina returns a non-sentinel score. Real non-negative scores are retained as unfavorable outcomes. This distinction avoids labeling an unfavorable but completed docking score as a computational failure. This yielded 89,086 training molecules for 1SYH, 104,988 for 4LDE, and 105,338 for 6Y2F. The label convention is $y=0$ for the strongest predicted binder ($s=s_{\min}$) and $y=1$ for the weakest ($s\to 0$). As for SA score, only the conditioning pathway (FiLM, conditional-prior networks, latent-to-hidden projection, and label projection) was updated in the final adaptation stage, with the Transformer weights, LoRA adapters, and posterior heads frozen (Table 1).

We generated 3,003 molecules per target per guidance scale, sweeping the requested label from 0 to 1 in steps of 0.001 and docking all valid molecules with QVina. The label direction is defined below. Note that the network-side label y_{model} and the reported sweep coordinate y^* run in opposite directions. RDKit SA scores were computed for the valid set at each α: mean SA ranged from 2.9 to 3.6 across all three targets and all guidance scales. SA score is a heuristic estimate rather than a measurement of synthesizability, and a mean can conceal a problematic tail, so we report these as favorable heuristic SA scores rather than as confirmation of synthetic accessibility. They do indicate that docking-score conditioning did not shift the bulk of the distribution toward the high-SA range [46].

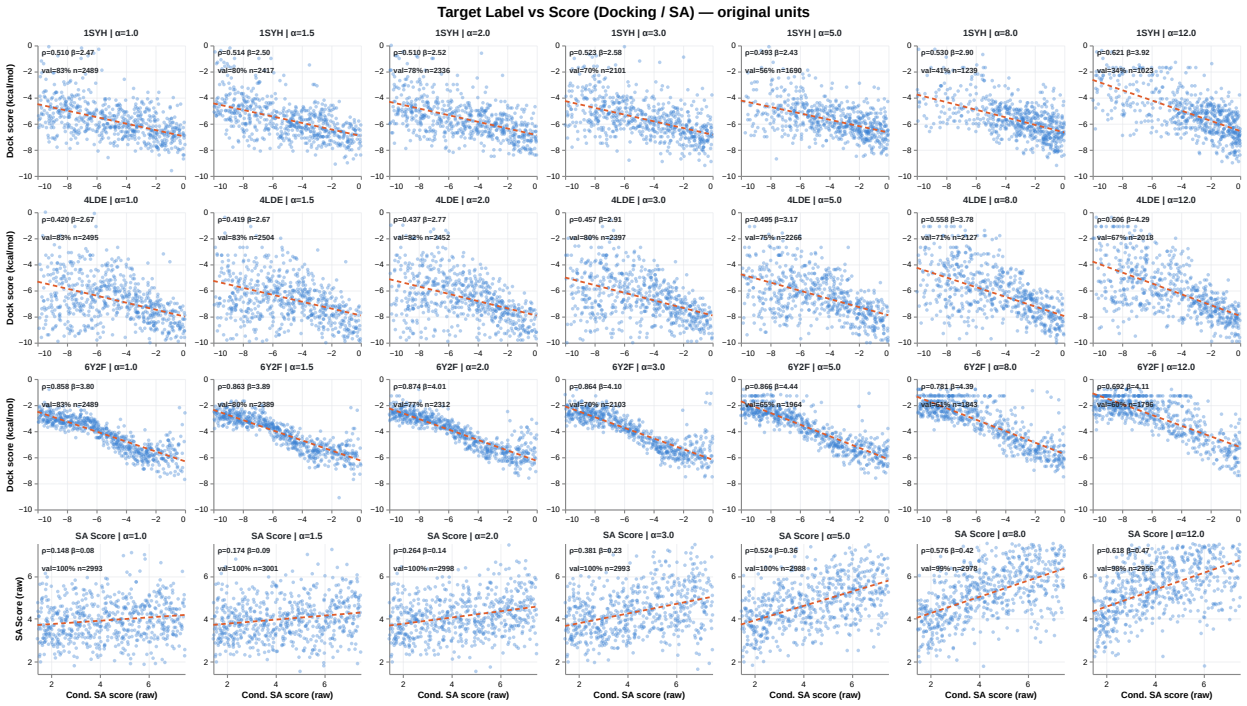


Figure 3: Conditioning label versus measured score for all four property targets across seven label-reference guidance strengths (α). Rows 1–3 show docking score (kcal/mol) against conditioning label for targets 1SYH, 4LDE, and 6Y2F. The label axis is the network-side label y_{model}, for which $y_{\text{model}} = 0$ targets strong binding (most negative). Row 4 shows SA score (raw units, 1–7) against conditioning label. The value $y^* = 0$ targets easy synthesis (low SA score). Each panel shows a random subsample of up to 600 of the scored molecules for legibility (fixed seed), while the least-squares regression line and the annotated Spearman ρ and calibration slope β are computed on all scored molecules for that condition. The dependence on α is target-specific rather than uniformly increasing. SA score and 4LDE strengthen with α, 6Y2F is strongest at moderate α and declines thereafter, and 1SYH is nearly flat until the largest α. Validity also falls as α increases for the docking targets, so panels at high α summarize fewer molecules (Table 4).

Alt text: Grid of 28 scatter plots (4 rows, 7 columns). Rows correspond to docking score targets 1SYH (GluR2), 4LDE (beta-2 adrenoceptor), and 6Y2F (SARS-CoV-2 protease), plus synthetic accessibility score. Columns span label-reference guidance strength alpha from 1.0 to 12.0. Each panel shows conditioning label versus measured score with a regression line. The association is target-dependent: it strengthens with alpha for synthetic accessibility and 4LDE, peaks at moderate alpha for 6Y2F, and is flat then rising for 1SYH.

Target Condition Alignment. Figure 3 shows y^* against the measured score for all four properties across seven guidance scales ($\alpha \in \{1.0,\ 1.5,\ 2.0,\ 3.0,\ 5.0,\ 8.0,\ 12.0\}$), with regression slope and Spearman ρ annotated per panel. Table 4 reports ρ, calibration slope β, validity, SR, and IntDiv at all tested α for each target.

For SA score, ρ increases from 0.15 at $\alpha = 1$ to 0.62 at $\alpha = 12$, with molecular validity remaining above 98% across all tested α (Table 4, rows 1–7). At $\alpha = 12$, molecules conditioned on $y^* \in [0,\ 0.33]$ have a mean

normalized SA score of 0.53 versus 0.84 for $y^* \in [0.67, 1]$, a separation of $\Delta = 0.31$ ($\approx$1.9 raw SA-score units).

The docking labels require care, because two label coordinates appear in the pipeline. The label supplied to the network is y_{model}, for which $y_{\text{model}} = 0$ denotes the strongest (most negative) training score. The saved generation files also contain sweep coordinate $y^* = 1 - y_{\text{model}}$. To remove the prior sign ambiguity, Table 4 and Figure 3 report y_{model} throughout. Rank correlations and calibration slopes are computed between y_{model} and raw docking score. Positive values therefore denote increasing score as the requested condition moves from strong toward weak binding. The strength of this association is target-dependent and not monotone in α. Under the standardized sentinel-only population, 6Y2F reaches $\rho = 0.813$ at $\alpha = 1$, peaks at 0.826 at $\alpha = 2$, and declines to 0.647 at $\alpha = 12$. 1SYH remains near 0.52 from $\alpha = 1$ to $\alpha = 5$ before rising to 0.630 at $\alpha = 12$. 4LDE increases from 0.482 at $\alpha = 1$ to 0.635 at $\alpha = 12$. Increasing α also reduces generation validity and docking completion for every protein target (Table 4), so larger α is not uniformly preferable. UMAP projections show latent-space organization by docking-score range (Figure 2(b–d)).

Success rate (SR) is the fraction of successfully scored molecules whose externally recomputed score falls within ε of the target score implied by the reported conditioning coordinate, at $\varepsilon_{10} = 0.1$ and $\varepsilon_{20} = 0.2$ times the property range. The score is recomputed with the same deterministic procedure that produced the training labels (RDKit SA score, or QVina against the same receptor), so it is independent of the model's parameters but not an orthogonal predictor of the same property. For docking, one operational rule is now used throughout: only the explicit QVina failure sentinel denotes a failed docking job. Real non-negative scores are retained. Gen.Valid is the RDKit-parseable fraction of all 3,003 attempts, Dock is the fraction receiving a non-sentinel QVina score, and ρ, β, SR, and IntDiv are computed over that docked population. Table 5 separately reports attempt-level SR, where every invalid or failed-docking attempt counts as failure. For docking, the target score is $\hat{s}(y_{\text{model}}) = s_{\min}(1 - y_{\text{model}})$, so ρ, β, and SR use the same coordinate and population. For SA score at $\alpha = 12$, $\text{SR}_{10}/\text{SR}_{20}$ are 0.214/0.411. At the rank-correlation-maximizing docking settings, docked-only $\text{SR}_{10}/\text{SR}_{20}$ are 0.449/0.725 (6Y2F, $\alpha = 2$), 0.189/0.463 (1SYH, $\alpha = 12$), and 0.366/0.657 (4LDE, $\alpha = 12$). The corresponding attempt-level SR_{20} values are 0.563, 0.161, and 0.404. Internal diversity (IntDiv) computed as $1 - \bar{\tau}$ over Morgan fingerprints (radius = 2, 1024 bits) is reported for both the full scored set and the SR_{20} subset. Calibration curves (Figure 4) visualize the bin-level trend.

Table 4: Conditioning performance across all label-reference guidance scales α for each property target. Bold rows mark the α with the highest ρ per target. Because ρ is not the only relevant objective, the full sweep is reported and Table 5 gives an explicit trade-off view. ρ: Spearman rank correlation between the reported conditioning coordinate and the recomputed score. Docking rows use the network-side coordinate $y_{\text{model}} = 1 - y^*$. β: calibration slope. Gen.Valid: RDKit-parseable molecules divided by 3,003 attempts. Dock: molecules with a non-sentinel QVina score divided by 3,003. Not applicable to SA score. SR_{10}/SR_{20}: valid-/docked-only success at $\varepsilon = 0.1/0.2\times$ property range. IntDiv: internal diversity of the full scored set and the SR_{20} subset. Docking rows use the sentinel-only failure rule.

Target	α	ρ	β	Gen.Valid	Dock	SR_{10}	SR_{20}	IntDiv	IntDiv SR_{20}
	1.0	0.148	0.077	99.7%	–	0.206	0.423	0.901	0.898
	1.5	0.174	0.095	99.9%	–	0.219	0.435	0.902	0.901
	2.0	0.264	0.145	99.8%	–	0.230	0.458	0.903	0.902
SA score	3.0	0.381	0.230	99.7%	–	0.247	0.498	0.907	0.905
	5.0	0.524	0.360	99.5%	–	0.288	0.552	0.915	0.914
	8.0	0.576	0.423	99.2%	–	0.245	0.482	0.925	0.927
	12.0	**0.618**	**0.473**	**98.4%**	–	**0.214**	**0.411**	**0.934**	**0.939**
	1.0	0.813	2.611	85.8%	83.6%	0.449	0.708	0.919	0.919
	1.5	0.802	1.771	83.6%	80.6%	0.448	0.711	0.919	0.919
	2.0	**0.826**	**1.929**	**81.9%**	**77.7%**	**0.449**	**0.725**	**0.920**	**0.921**
6Y2F	3.0	0.800	2.092	77.7%	71.0%	0.449	0.727	0.921	0.921
	5.0	0.821	2.940	75.2%	66.2%	0.416	0.730	0.926	0.924
	8.0	0.726	2.513	72.9%	62.1%	0.340	0.665	0.931	0.921
	12.0	0.647	2.516	72.9%	60.4%	0.309	0.596	0.931	0.911
	1.0	0.523	2.972	86.2%	84.2%	0.270	0.522	0.902	0.905
	1.5	0.528	3.097	84.0%	81.9%	0.278	0.536	0.901	0.905
	2.0	0.526	3.122	81.2%	79.2%	0.270	0.538	0.901	0.903
1SYH	3.0	0.540	3.398	74.7%	71.4%	0.280	0.562	0.902	0.904
	5.0	0.516	3.501	63.4%	57.5%	0.284	0.574	0.904	0.905
	8.0	0.544	3.962	54.2%	41.9%	0.253	0.553	0.912	0.913
	12.0	**0.630**	**5.203**	**54.8%**	**34.7%**	**0.189**	**0.463**	**0.927**	**0.932**
	1.0	0.482	9.927	89.2%	87.4%	0.322	0.559	0.915	0.917
	1.5	0.482	7.507	89.6%	87.9%	0.321	0.557	0.914	0.917
	2.0	0.498	11.226	88.6%	86.1%	0.331	0.567	0.914	0.917
4LDE	3.0	0.525	14.356	87.7%	85.0%	0.335	0.573	0.914	0.919
	5.0	0.563	12.848	85.0%	81.2%	0.350	0.590	0.914	0.919
	8.0	0.616	11.188	82.0%	75.8%	0.358	0.622	0.917	0.922
	12.0	**0.635**	**8.744**	**80.6%**	**70.2%**	**0.366**	**0.657**	**0.925**	**0.928**

Because rank correlation alone does not capture the cost of stronger guidance, Table 5 reports the same sentinel-only sweep at attempt level, adding the number of molecules that were actually generated and

docked, the attempt-level success rate, and the fraction of attempts reaching a target-specific strong-binding cutoff. Two patterns are visible there that the valid-only view obscures. First, the number of successfully docked molecules falls substantially with α: for 1SYH, from 2,530 of 3,003 attempts at $\alpha = 1$ to 1,041 at $\alpha = 12$. Second, and more consequentially for practical use, the fraction of attempts producing a strong binder *decreases* monotonically with α for all three targets (1SYH: 9.0% to 2.7%, 4LDE: 9.7% to 3.7%, and 6Y2F: 9.7% to 3.0%). Increasing guidance therefore improves the rank ordering of scores while reducing the absolute yield of strong-binding molecules per attempt. We consequently do not designate a single recommended α as uniformly best, and report the full sweep instead.

Table 5: Attempt-level accounting and trade-off analysis for docking-score conditional generation. Every condition begins from the same 3,003 generation attempts (1,001 requested labels × 3 samples). n_{gen}: RDKit-parseable molecules produced. n_{dock}: molecules receiving a usable QVina score. Only the explicit failure sentinel is excluded, and real non-negative scores are retained. Valid.: $n_{\text{gen}}/3003$. Dock: $n_{\text{dock}}/3003$. ρ: Spearman correlation between network-side conditioning coordinate $y_{\text{model}} = 1 - y^*$ and raw docking score over the same docked population used in Table 4. $\text{SR}_{20}^{\text{att}}$: success rate at $\varepsilon = 0.2\times$ property range with all 3,003 attempts in the denominator. Strong: fraction of *attempts* yielding a docking score at or below the target-specific strong-binding cutoff (the 10th percentile of the $\alpha = 1$ docked distribution: −7.4, −8.9, and −6.0 kcal/mol for 1SYH, 4LDE, and 6Y2F). Increasing α raises ρ for 1SYH and 4LDE and reduces strong-binder yield for all three targets. Attempt-level success also declines overall but is not strictly monotone. No single α is optimal across these measures.

Target	α	n_{gen}	n_{dock}	**Valid.**	**Dock**	ρ	$\text{SR}_{20}^{\text{att}}$	**Strong**
1SYH	1.0	2588	2530	0.862	0.842	0.523	0.440	0.0896
	1.5	2522	2460	0.840	0.819	0.528	0.439	0.0886
	2.0	2439	2377	0.812	0.792	0.526	0.426	0.0759
	3.0	2243	2145	0.747	0.714	0.540	0.402	0.0703
	5.0	1904	1727	0.634	0.575	0.516	0.330	0.0430
	8.0	1627	1258	0.542	0.419	0.544	0.232	0.0383
	12.0	1646	1041	0.548	0.347	0.630	0.161	0.0270
4LDE	1.0	2680	2625	0.892	0.874	0.482	0.489	0.0966
	1.5	2692	2639	0.896	0.879	0.482	0.489	0.0876
	2.0	2660	2586	0.886	0.861	0.498	0.488	0.0842
	3.0	2635	2552	0.877	0.850	0.525	0.487	0.0733
	5.0	2554	2437	0.850	0.812	0.563	0.479	0.0586
	8.0	2461	2277	0.820	0.758	0.616	0.472	0.0476
	12.0	2420	2109	0.806	0.702	0.635	0.461	0.0370
6Y2F	1.0	2577	2512	0.858	0.836	0.813	0.592	0.0966
	1.5	2512	2419	0.836	0.806	0.802	0.572	0.0836
	2.0	2460	2334	0.819	0.777	0.826	0.563	0.0806
	3.0	2333	2133	0.777	0.710	0.800	0.516	0.0703
	5.0	2259	1988	0.752	0.662	0.821	0.484	0.0606
	8.0	2189	1866	0.729	0.621	0.726	0.413	0.0423
	12.0	2188	1815	0.729	0.604	0.647	0.360	0.0303



Figure 4: Calibration curves for all four conditioning targets at the α with the highest reported rank correlation. Each panel bins the successfully scored molecules from the 3,003 generation attempts into ten equal-width intervals and plots the mean $\pm$ 1 SD of the externally recomputed normalized property score in each bin (blue line and shaded band) against the target-specific ideal line (dashed grey). The SA panel uses requested label y^*. Docking panels use network-side label $y_{\text{model}} = 1 - y^*$. The shaded band shows the molecule-to-molecule spread within a bin, not the uncertainty of the binned mean, and the number of molecules per bin varies with validity. For docking targets (plotted against the network-side label y_{model}, for which $y_{\text{model}} = 0$ is the strongest binder) the observed score is normalized to $[0, 1]$ via $\hat{s} = (s - s_{\min})/(0 - s_{\min})$ so all four panels share the same axis. Deviation from the ideal line indicates residual miscalibration. Each panel indicates a monotone association between the requested label and the recomputed property, with the largest departures from the diagonal at the extremes of the label range.

Alt text: Four calibration curve panels. Each plots conditioning label (0 to 1) against mean normalized observed property score across 10 bins, with a dashed target-specific ideal line and shaded standard deviation band. Panels cover synthetic accessibility score and docking scores for targets 1SYH (GluR2), 4LDE (beta-2 adrenoceptor), and 6Y2F (SARS-CoV-2 main protease). All panels show monotone trends with the largest miscalibration at the extremes of the label range. Shaded bands show per-molecule spread within each bin, not uncertainty of the mean.

Scope and limitations. Each CVAE model reported above is trained independently for a single fixed protein target using docking scores as the scalar conditioning signal. No cross-target transfer is performed or claimed: adapting to a new target uses the two-stage adaptation procedure, effective LoRA and FiLM adaptation followed by conditioning-pathway refinement with inherited LoRA frozen, on per-target docking data. Both stages together complete in under two hours on a single GPU using the pipeline described in the Model Architecture subsection. This per-target training requirement is a current limitation that distinguishes STAR-VAE from approaches that explicitly represent target identity (such as protein-sequence or pocket-structure embeddings) as a conditioning variable, and is an open direction for future work.

Baseline Comparison for SA score Conditioning.

Baseline selection. We compare STAR-VAE SA score conditioning against three baselines representing distinct conditioning paradigms: reinforcement learning (REINVENT 4 [47]), autoregressive prefix-token

conditioning (MolGPT [13]), and latent diffusion with predictor guidance (DrugDiff [48]). Together with STAR-VAE (latent-variable conditioning), these four models span the principal approaches to property-conditioned molecular generation. MolGuidance [49] and PropMolFlow [50] were excluded because both require 3D SDF input and condition on quantum-chemistry properties not supported by the 2D SA score setting.

Baseline configurations. Each baseline is evaluated in its standard deployment configuration, which is the comparison relevant to practitioners choosing among frameworks. **REINVENT 4** is first transfer-learned for five epochs on the same 9,990 PubChem SMILES to align its generation distribution with the PubChem-10K chemical space. It is then optimized through staged reinforcement learning with the RDKit SA score as the reward signal. This workflow is the standard deployment procedure for goal-directed generation with REINVENT 4. **MolGPT** is trained from scratch on the same 9,990 molecules with SA score as the conditioning property; it encodes the target SA score value as a prefix token prepended to the SMILES sequence and has no pretrained backbone. **DrugDiff** retains its published ZINC-250K pretrained VAE and latent diffusion model; an SA score prediction head (multi-layer perceptron, 3 hidden layers) is trained on ZINC-250K molecules scored with RDKit and used to steer reverse diffusion via predictor gradient guidance.

Performance differences therefore reflect both conditioning-mechanism design and initialization scale. Controlling for initialization by re-pretraining all models on PubChem-80M would require weeks of computation per baseline and produce non-standard configurations outside each method's intended use. The comparisons below are deployment-realistic, not architecturally controlled.

Evaluation protocol. All models generate exactly 3,003 molecules by sweeping the conditioning target y^* uniformly from 0 to 1 in steps of 0.001 ($\times 3$ samples per level), matching the protocol used for STAR-VAE exactly. Generated SMILES are rescored with the same deterministic RDKit SA-score procedure and normalized using the same scaler ($s_{\min} = 1.437$, $s_{\max} = 7.513$). We report Spearman ρ between y^* and realized score, calibration slope β, MAE, success rate (SR_{10}/SR_{20}), molecular validity, and internal diversity (IntDiv). The ρ, β, and validity values are computed as in Table 4. IntDiv, however, was computed by three different pipelines that differ in fingerprint width and in which molecules enter the pairwise average, and we report the values as originally computed rather than silently re-deriving them. For STAR-VAE in this table the estimate uses Morgan radius 2 with 2048 bits over the first 2,000 molecules. For the three external baselines it uses Morgan radius 2 with 2048 bits over a random 1,000-molecule subsample. And in Table 4 it uses Morgan radius 2 with 1024 bits over all valid molecules. The same STAR-VAE molecules at $\alpha = 12$ therefore yield IntDiv = 0.928 here and 0.934 in Table 4, and would yield 0.946 under the subsampling convention used

for the baselines. IntDiv differences of this size ($\lesssim 0.02$) are consequently not meaningful in this table, and we do not draw conclusions from the IntDiv ranking. Because REINVENT 4 and DrugDiff do not natively support a fine-grained 1001-level y^* sweep, we assign bucket midpoints ($y^* \in \{0.167, 0.500, 0.833\}$) to the low, mid, and high generation campaigns and sample 1,001 molecules per bucket; Spearman ρ is computed over the resulting three-level ranking, which is structurally different from the 1,001-point sweep used for STAR-VAE and MolGPT (a distinction we address below).

Table 6: SA score conditional generation benchmark. All models evaluated on 3,003 generation attempts using the same deterministic RDKit SA-score procedure and min–max scaler ($s_{\min} = 1.437$, $s_{\max} = 7.513$). Bold indicates the best value *within each paradigm group*. The two groups use fundamentally different conditioning protocols and are not directly rank-comparable (see text). ρ: Spearman rank correlation between y^* and realized score. β: calibration slope (OLS). MAE: mean $|y^* - \hat{y}|$. $\mathrm{SR}_{10}/\mathrm{SR}_{20}$: success rate at $\varepsilon = 0.10/0.20$. Valid.: $n_{\text{valid}}/3003$, i.e. the denominator is the number of generation attempts for every model. IntDiv: internal diversity. Computed by three different pipelines that differ in fingerprint width and evaluated subset, so it is not bolded and is not comparable across rows or against Table 4 (see text). Scaffolds: unique Murcko frameworks. Scaf/mol = unique scaffolds / n_{valid}. $|\Delta|$: |mean(high bucket) − mean(low bucket)| (larger = better separation). ‡Spearman ρ is a 3-point test. *Metrics use the 50.4% valid survivors only.

Model	ρ	β	MAE	SR_{10}	SR_{20}	Valid.	IntDiv	Scaffolds	Scaf/mol	$\lvert\Delta\rvert$
Paradigm A: oracle-free continuous conditioning (1,001 levels)										
STAR-VAE (ours, $\alpha = 12$)	0.618	0.473	0.261	0.215	0.411	**0.984**	0.928	**1,974**	0.668	0.314
MolGPT*	**0.895**	**0.703**	**0.092**	**0.685**	**0.890**	0.505	0.907	1,088	**0.718**	**0.428**
Paradigm B: oracle-in-the-loop / 3-bucket conditioning										
REINVENT 4‡	**0.941**	**1.482**	**0.145**	**0.228**	**0.910**	0.875	0.918	1,347	**0.513**	**0.971**
DrugDiff‡	−0.026	−0.013	0.273	0.207	0.390	**1.000**	0.922	**1,373**	0.457	0.008

Results. Table 6 summarizes all metrics. Because the two paradigms use different evaluation protocols, Spearman ρ is only meaningful when compared within the same paradigm group.

Paradigm A: oracle-free continuous conditioning (STAR-VAE, MolGPT). Both models are trained on static (label, SELFIES/SMILES) pairs, with no oracle access during training, and are evaluated across 1,001 uniformly-spaced conditioning targets $y^* \in [0, 1]$. In this setting, Spearman ρ measures how well each model tracks the requested property value across the full conditioning range. STAR-VAE generates valid molecules at 98.4% of all attempts versus 50.4% for MolGPT, and produces 1,974 distinct Murcko scaffolds [51] from 2,956 scored molecules (scaf/mol = 0.668). MolGPT's $\rho = 0.895$ is based only on its 1,515 valid outputs and does not account for the 1,488 failed attempts (49.6% of all tries). Validity drops sharply as the conditioning target moves toward harder synthesis: 75.7% at low y^* ($y^* \leq 0.33$, raw SA $\leq$ 3.5, easy-to-synthesize) versus 24.4% at high y^* ($y^* \geq 0.67$, raw SA $\geq$ 5.5, hard-to-synthesize); Table S2 shows the full per-bucket breakdown. This reflects a known failure mode of autoregressive SMILES decoders, which struggle to produce valid outputs for structurally complex molecules. The valid survivors skew toward easy-

to-synthesize structures (mean $SA_{norm} = 0.35$ vs. ≈ 0.60 for the training set), so the effective SR_{10} across all 3,003 attempts falls from the reported 0.685 to approximately 0.346, close to STAR-VAE's 0.215. On the same attempt basis, MolGPT's scaffold coverage rate is $1{,}088/3{,}003 = 0.362$, versus STAR-VAE's 0.668.

Paradigm B: oracle-in-the-loop / 3-bucket conditioning (REINVENT 4, DrugDiff). REINVENT 4 optimizes three separate RL policies, one per conditioning bucket ($y^* \in \{0.167, 0.500, 0.833\}$), using approximately 384,000 SA score oracle calls in total. Its $\rho = 0.941$ is a 3-point rank test: it simply checks whether the three bucket-level means are correctly ordered, which is expected when each policy was directly trained to hit its bucket. $SR_{10} = 0.228$ shows that pointwise accuracy remains modest despite the high rank correlation. Its aggregate SR values are also computed on retained molecules only: REINVENT 4 retained 2,628 unique valid molecules from 3,003 requested (87.5%), and retention is strongly bucket-dependent (63.9% at $y^* = 0.167$ versus 99.1% at $y^* = 0.833$). Because its sampler discards duplicates as well as unparseable strings, the low-y^* shortfall reflects repeated structures rather than invalid chemistry. On a per-requested-molecule basis its SR_{20} in that campaign falls from 1.000 to 0.639 (Table S2). The calibration slope $\beta = 1.482$ indicates that the model over-shoots the requested targets: the easy-synthesis campaign ($y^* = 0.167$) produces molecules with a mean realized SA score of 0.003 (target: 0.167), and the hard-synthesis campaign ($y^* = 0.833$) reaches a mean of 0.974 (target: 0.833); both are farther from their targets than their nominal distance suggests (Table S2). Scaffold diversity is also highly uneven: the easy-synthesis campaign converges to only 61 unique Murcko scaffolds from the 640 unique molecules it retained out of 1,001 requested (scaf/mol = 0.095), while the hard-synthesis campaign yields near-unique scaffolds per molecule (979 unique from 992 valid; scaf/mol = 0.987), because the RL agent saturates the accessible scaffold pool when the reward target is easy (Table S2). DrugDiff generates valid molecules at 100% but produces no conditioning effect on the PubChem distribution ($\rho \approx 0$, $|\Delta| = 0.008$) and the fewest distinct scaffolds (scaf/mol = 0.457). The SA score predictor trained on ZINC-250K molecules does not transfer to the PubChem chemical space without retraining.

Scope and limitations. The four baselines differ in pretraining scale and initialization, so observed performance gaps reflect both the conditioning mechanism and the starting distribution. A controlled comparison would require re-pretraining all models on PubChem-80M, which is outside the scope of this study.

Rediscovery Evaluation

Rediscovery evaluation was performed across four ChEMBL target datasets. Each dataset contains fewer than 11,000 labeled compounds after curation, which is insufficient for direct generative fine-tuning. We therefore adopted a silver-label strategy. MMELON [52] (Multi-view Molecular Encoder with Long-context and Omics Networks), a multi-view molecular foundation model, was fine-tuned on each target dataset and

applied to 100,000 randomly selected PubChem molecules to produce pseudo-labels for STAR-VAE fine-tuning. Pseudo-label quality was validated against held-out experimental ChEMBL values (Table 7). RMSE values of 0.095–0.114 log-units across targets indicate that the pseudo-labels are sufficiently accurate for use as conditioning signals. All independent property predictors used in controllability validation were trained exclusively on experimental labels, without access to the pseudo-labeled PubChem set.

Four targets were selected from ChEMBL for molecular rediscovery evaluation: ABL1 kinase (CHEMBL1862), Bromodomain-containing protein 4 (BRD4, CHEMBL1163125), Glucocorticoid receptor (GRL, CHEMBL2034), and HIV-1 protease (CHEMBL243). Standard curation was applied: IC50-only assays, canonical SMILES, charge removal, duplicate averaging, and Murcko scaffold computation. After refinement, the ABL1 dataset was reduced from 18,830 to 2,064 samples, BRD4 from 58,432 to 10,800, GRL from 15,052 to 2,744, and HIV-1 protease from 9,226 to 2,843 samples.

MMELON was fine-tuned on each target dataset using an 80/10/10 train/validation/test split. Fine-tuning used the AdamW optimizer with a learning rate of 10^{-4}, batch size 32, and gradient clipping at 1.0. Models were trained for up to 200 epochs with early stopping (patience 20) based on validation RMSE, and the best checkpoint was selected by minimum validation RMSE. Predictions on 100,000 randomly selected PubChem molecules provided silver labels for STAR-VAE fine-tuning. All models converged without overfitting (Table 7).

Table 7: Evaluation metrics for ChEMBL molecular property prediction across four target datasets. Training completed successfully for all datasets. The evaluation metrics (MSE, RMSE, MAE) are reported. Sample counts reflect the 80/10/10 train/validation/test split.

Dataset	**Samples**			**MSE**	**RMSE**	**MAE**
	Train	**Val**	**Test**			
ABL1	1650	207	207	0.0090	0.0951	0.0695
BRD4	8640	1080	1080	0.0050	0.0704	0.0442
GRL	2194	275	275	0.0120	0.1093	0.0707
HIV-1 Protease	2273	285	285	0.0130	0.1140	0.0845

Three generation approaches were compared: PubChem sampling (baseline), VAE (temperature $T = 1.1$), and CVAE (LoRA with label conditioning). Each was run on the four ChEMBL targets at five sample sizes (10K to 50K), for 60 conditions total. **PubChem sampling** (baseline): molecules were drawn uniformly at random from the 100,000-molecule PubChem set used for silver-label generation, with no model involvement.

VAE: the pretrained VAE encodes each training-set seed molecule into its posterior mean μ, samples latent vectors in the neighbourhood of μ (perturbation scale σ, decoder temperature $T = 1.1$), and decodes to SMILES. Seeds were drawn from the training set by cluster-proportional sampling to reflect the full scaffold distribution. **CVAE**: the fine-tuned CVAE generates from a label alone, with no seed molecule. In this configuration the label-conditioned prior was disabled, so latent vectors are drawn from the isotropic Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and modulated by FiLM with the requested label before decoding (Table 1). Conditioning therefore acts on the decoder rather than on the latent distribution. Conditioning targets y^* were drawn proportionally from all ten label bins of the training distribution, covering the full range of the min-max-normalized $\log_{10}(\mathrm{IC}_{50}/\mathrm{nM})$ scale (label = 0: strongest binder. Label = 1: weakest binder). This broad sampling was chosen to evaluate generation coverage. Conditioning on a targeted label tier is addressed in the Scope and Limitations paragraph below. Metrics cover Exact Rediscovery (exact SMILES matches), Nearest Neighbor Similarity (Tanimoto, Morgan fingerprints radius = 2, 2048 bits), Scaffold Rediscovery (Bemis–Murcko), and Clustering Rediscovery (Butina, threshold 0.6).

The held-out split was constructed by Butina clustering with Morgan fingerprints (radius = 2, 2048 bits) and a similarity threshold of 0.6. Approximately 20% of clusters (∼20% of molecules) were held out using stratified sampling across 10 score bins, so that entire clusters, rather than individual molecules, were reserved for evaluation. This removes close analogues from the target-specific fine-tuning set. It does not remove the held-out compounds from the pretraining corpus (Appendix F). The clustering proportion metric measures the fraction of held-out clusters reached by at least one generated molecule, evaluating whether models generate molecules in chemically feasible regions that were excluded from training. The number of molecules in the held-out sets is 364, 2028, 615, and 492 for ABL1, BRD4, GRL, and HIV-1 protease, respectively.

Table 8: Comprehensive rediscovery metrics for all sample sizes (10K, 20K, 30K, 40K, 50K) across four datasets and three approaches. Within each 3-row group (CVAE, PUBCHEM, VAE at the same dataset and sample size), the highest value is bolded for Exact P*, Exact R*, NN Sim, Clust 0.6, Scaff Redisc, CrossSim, Div, and Uniq; for InterSim the lowest value is bolded. Exact P* and Exact R* are reported in $\times 10^{-4}$. CrossSim is the mean Tanimoto similarity over all generated–held-out scaffold pairs. It is a raw similarity in $[0, 1]$, not a ratio. Appendix A gives the definitions.

Dataset	Model	N	Exact P*	Exact R*	NN Sim	Clust 0.6	Scaff Redisc	CrossSim	InterSim	Div	Uniq
ABL1	CVAE	10000	0	0	0.180	0.000	0.032	0.110	**0.104**	**1.096**	**1.014**
ABL1	PUBCHEM	10000	0	0	0.210	0.000	0.059	0.129	0.127	1.086	**1.014**
ABL1	VAE	10000	**1.16**	**27.47**	**0.246**	**0.073**	**0.100**	**0.133**	0.120	1.078	0.954
ABL1	CVAE	20000	0	0	0.181	0.000	0.046	0.110	**0.104**	**1.096**	**1.014**
ABL1	PUBCHEM	20000	0	0	0.210	0.018	0.078	0.130	0.127	1.086	**1.014**
ABL1	VAE	20000	**1.77**	**82.42**	**0.246**	**0.101**	**0.110**	**0.133**	0.120	1.078	0.937

Continued on next page

Table 8 continued from previous page

Dataset	Model	N	Exact P*	Exact R*	NN Sim	Clust 0.6	Scaff Redisc	CrossSim	InterSim	Div	Uniq
ABL1	CVAE	30000	0	0	0.181	0.000	0.046	0.110	**0.104**	**1.096**	**1.014**
ABL1	PUBCHEM	30000	0	0	0.210	0.037	0.082	0.129	0.127	1.086	**1.014**
ABL1	VAE	30000	**1.98**	**137.36**	**0.245**	**0.138**	**0.128**	**0.132**	0.120	1.078	0.928
ABL1	CVAE	40000	0	0	0.181	0.000	0.050	0.110	**0.104**	**1.096**	**1.014**
ABL1	PUBCHEM	40000	0	0	0.210	0.037	0.087	0.130	0.127	1.086	**1.014**
ABL1	VAE	40000	**1.8**	**164.84**	**0.246**	**0.128**	**0.160**	**0.132**	0.120	1.078	0.917
ABL1	CVAE	50000	0	0	0.181	0.000	0.055	0.110	**0.104**	**1.096**	**1.014**
ABL1	PUBCHEM	50000	0	0	0.210	0.037	0.087	0.130	0.127	1.086	**1.014**
ABL1	VAE	50000	**1.93**	**219.78**	**0.246**	**0.138**	**0.174**	**0.132**	0.120	1.078	0.912
BRD4	CVAE	10000	0	0	0.207	0.000	0.009	0.105	**0.103**	**1.066**	1.043
BRD4	PUBCHEM	10000	0	0	**0.240**	0.009	0.015	**0.119**	0.127	1.056	**1.043**
BRD4	VAE	10000	**1.09**	**4.93**	0.240	**0.025**	**0.049**	0.114	0.107	1.053	1.034
BRD4	CVAE	20000	0	0	0.207	0.000	0.012	0.106	**0.104**	**1.066**	1.043
BRD4	PUBCHEM	20000	0	0	0.240	0.012	0.020	**0.119**	0.127	1.056	**1.043**
BRD4	VAE	20000	**1.1**	**9.86**	**0.241**	**0.056**	**0.063**	0.114	0.107	1.053	1.024
BRD4	CVAE	30000	0	0	0.207	0.000	0.013	0.106	**0.104**	**1.066**	1.043
BRD4	PUBCHEM	30000	0	0	0.240	0.012	0.027	**0.119**	0.127	1.056	**1.043**
BRD4	VAE	30000	**0.74**	**9.86**	**0.241**	**0.053**	**0.077**	0.114	0.107	1.053	1.018
BRD4	CVAE	40000	0	0	0.207	0.000	0.013	0.106	**0.104**	**1.066**	1.043
BRD4	PUBCHEM	40000	0	0	**0.240**	0.012	0.031	**0.120**	0.127	1.056	**1.043**
BRD4	VAE	40000	**1.67**	**29.59**	0.240	**0.062**	**0.070**	0.114	0.107	1.053	1.014
BRD4	CVAE	50000	0	0	0.207	0.000	0.014	0.106	**0.104**	**1.066**	1.043
BRD4	PUBCHEM	50000	0	0	0.240	0.019	0.033	**0.120**	0.127	1.056	**1.043**
BRD4	VAE	50000	**0.45**	**9.86**	**0.240**	**0.062**	**0.088**	0.114	0.107	1.053	1.011
GRL	CVAE	10000	**0**	**0**	0.203	0.000	0.064	0.106	**0.104**	**1.036**	**1.025**
GRL	PUBCHEM	10000	**0**	**0**	0.231	0.020	0.120	**0.118**	0.127	1.027	**1.025**
GRL	VAE	10000	**0**	**0**	**0.242**	**0.030**	**0.127**	0.110	0.111	1.036	0.906
GRL	CVAE	20000	0	0	0.203	0.000	0.090	0.107	**0.105**	**1.036**	**1.025**
GRL	PUBCHEM	20000	0	0	0.231	**0.035**	**0.142**	**0.118**	0.127	1.028	**1.025**
GRL	VAE	20000	**2.38**	**48.78**	**0.241**	0.025	0.139	0.110	0.110	1.034	0.872
GRL	CVAE	30000	0	0	0.203	0.000	0.105	0.107	**0.105**	**1.036**	**1.025**
GRL	PUBCHEM	30000	0	0	0.231	**0.066**	**0.187**	**0.118**	0.127	1.028	**1.025**
GRL	VAE	30000	**1.63**	**48.78**	**0.242**	0.051	0.157	0.110	0.111	1.035	0.851
GRL	CVAE	40000	0	0	0.203	0.000	0.109	0.107	**0.105**	**1.036**	**1.025**
GRL	PUBCHEM	40000	0	0	0.231	**0.076**	**0.217**	**0.118**	0.127	1.027	**1.025**
GRL	VAE	40000	**1.65**	**65.04**	**0.241**	0.045	0.184	0.110	0.111	1.034	0.840
GRL	CVAE	50000	0	0	0.203	0.000	0.112	0.107	**0.105**	**1.036**	**1.025**
GRL	PUBCHEM	50000	0	0	0.231	**0.086**	**0.228**	**0.118**	0.127	1.027	**1.025**
GRL	VAE	50000	**1.34**	**65.04**	**0.242**	0.051	0.180	0.110	0.110	1.034	0.828
HIV-1 Protease	CVAE	10000	**0**	**0**	0.184	**0.000**	0.017	0.112	**0.104**	**1.173**	**1.070**
HIV-1 Protease	PUBCHEM	10000	**0**	**0**	**0.208**	**0.000**	**0.021**	**0.133**	0.127	1.163	**1.070**
HIV-1 Protease	VAE	10000	**0**	**0**	0.207	**0.000**	**0.021**	0.128	0.117	1.152	1.047

Continued on next page

Table 8 continued from previous page

Dataset	Model	N	Exact P*	Exact R*	NN Sim	Clust 0.6	Scaff Redisc	CrossSim	InterSim	Div	Uniq
HIV-1 Protease	CVAE	20000	**0**	**0**	0.184	0.000	0.017	0.113	**0.104**	**1.173**	**1.070**
HIV-1 Protease	PUBCHEM	20000	**0**	**0**	**0.208**	0.000	0.026	**0.133**	0.127	1.163	**1.070**
HIV-1 Protease	VAE	20000	**0**	**0**	0.206	**0.012**	**0.030**	0.127	0.116	1.153	1.044
HIV-1 Protease	CVAE	30000	**0**	**0**	0.184	0.000	0.017	0.113	**0.105**	**1.173**	**1.070**
HIV-1 Protease	PUBCHEM	30000	**0**	**0**	**0.208**	0.000	**0.030**	**0.133**	0.127	1.163	**1.070**
HIV-1 Protease	VAE	30000	**0**	**0**	0.206	**0.025**	**0.030**	0.127	0.116	1.153	1.039
HIV-1 Protease	CVAE	40000	**0**	**0**	0.184	0.000	0.017	0.113	**0.105**	**1.173**	**1.070**
HIV-1 Protease	PUBCHEM	40000	**0**	**0**	**0.208**	0.000	**0.038**	**0.133**	0.127	1.163	**1.070**
HIV-1 Protease	VAE	40000	**0**	**0**	0.206	**0.012**	**0.038**	0.127	0.117	1.153	1.036
HIV-1 Protease	CVAE	50000	**0**	**0**	0.184	0.000	0.017	0.114	**0.105**	**1.173**	**1.070**
HIV-1 Protease	PUBCHEM	50000	**0**	**0**	**0.208**	0.000	**0.043**	**0.133**	0.127	1.163	**1.070**
HIV-1 Protease	VAE	50000	**0**	**0**	0.206	**0.025**	0.030	0.127	0.116	1.153	1.036

Exact P (Exact Discovery Precision): Proportion of generated molecules that exactly match held-out molecules. **Exact R** (Exact Discovery Recall): Proportion of held-out molecules that are exactly regenerated. **NN Sim** (Nearest Neighbor Mean Similarity Score): Average Tanimoto similarity between each generated molecule and its closest match in the held-out set. **Clust 0.6** (Clustering Held-Out Cluster Recall at similarity ≥ 0.6): Fraction of *held-out clusters* reached by at least one generated molecule whose similarity to that cluster is at least 0.6. The denominator is the number of held-out clusters, not the number of generated molecules. **Scaff Redisc** (Scaffold Rediscovery): Proportion of held-out Murcko scaffolds recovered in the generated set. **CrossSim** (Cross-Similarity between Scaffolds): Mean Tanimoto similarity over *all* generated–held-out scaffold pairs. It is a raw similarity in $[0, 1]$ and an all-pairs average, not a nearest-neighbour distance. Because it aggregates over all pairs, it reflects global proximity and does not establish that generated molecules occupy specific held-out regions. **InterSim** (Internal Similarity of Generated Scaffolds): Average pairwise Tanimoto similarity within the generated scaffold set. **Div** and **Uniq**: ratios of diversity and uniqueness in the generated set to the held-out set. Because these are ratios, values may exceed 1 and are not percentages. Uniqueness is computed on the generated collection, in which repeated molecules are counted, whereas the scaffold-similarity metrics operate on deduplicated scaffold sets.

Across all datasets, seed-based sampling from the pretrained VAE recovered more held-out structures than label-only CVAE sampling on exact rediscovery, scaffold recovery, and held-out cluster recovery (Table 8). For ABL1 and BRD4 the CVAE recovered no held-out compounds exactly, so the ratio to VAE recall is undefined; in absolute terms VAE exact recall reached 2.2% (ABL1) and 0.3% (BRD4). PubChem random sampling remained competitive on GRL. Nearest-neighbor similarity spanned 0.18–0.25 across all methods, and scaffold rediscovery ranged 2.1–18.4% for the VAE versus 0.9–11.2% for the CVAE, depending strongly on target. The CVAE produces zero exact rediscoveries in most conditions. Label conditioning in this arm acts through decoder-side FiLM modulation on latents drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the fine-tuning set excludes the held-out ChEMBL molecules, so conditioning shifts generation toward the requested property range

without recovering specific known compounds.

Upper bound on pretraining exposure of the held-out compounds. Cluster-held-out splitting removes close analogues of the evaluation compounds from the target-specific fine-tuning set, but it does not remove them from the pretraining source corpus. An overlap audit against all 79,311,161 molecules in the full curated PubChem source corpus (Appendix F, Table S5) finds matches for 82.45% (ABL1), 75.98% (BRD4), 66.33% (GRL), and 21.79% (HIV-1 protease) of held-out compounds. Because that corpus is larger than the training dataset, these percentages are upper bounds on actual pretraining exposure. This bounds the interpretation of exact rediscovery: recovering a held-out structure that matched the source corpus may be retrieval from the pretrained distribution rather than proposal of a previously unseen molecule. We therefore report exact rediscovery as a bounded observation and treat scaffold recovery and cluster recovery as the primary evidence, since those metrics credit reaching a Bemis–Murcko framework or a cluster rather than reproducing a specific structure. The task is accordingly described throughout as recovery of task-held-out ChEMBL compounds by a broadly pretrained model. All VAE and CVAE configurations produce zero exact rediscoveries for HIV-1 protease at every sample size (10K–50K). That held-out set (492 molecules) has a median molecular weight of 566 Da with 35.0% of compounds above 600 Da, against medians of 425–473 Da and 4.7–6.3% above 600 Da for the other three targets. It also has the lowest pretraining exposure (21.79%). Scaffold class and pretraining exposure are confounded across these four targets, so the absence of exact rediscoveries for this target cannot be attributed to either factor alone.

Most metrics improved with sample size (10K to 50K): clustering gains were largest (up to +330% for GRL–PubChem at threshold 0.6), and scaffold rediscovery improved 42 to 91% for the VAE. Exact rediscovery occurred only for the VAE, and was already nonzero at 10K samples for ABL1 and BRD4 rather than emerging only at larger sizes; the counts are small throughout (at most 2.2% of held-out compounds). NN similarity remained stable (±0.5%) across all sample sizes and all methods, reflecting the intrinsic distance between model and held-out distributions rather than a coverage effect.

After correcting the table's audited column mapping, the CVAE has diversity ratios of 1.036–1.173 versus 1.034–1.153 for the VAE and lower internal scaffold similarity (InterSim: 0.103–0.105 versus 0.107–0.120). Raw generated–held-out CrossSim values are 0.106–0.114 for the CVAE and 0.111–0.131 for the VAE, indicating similar global scaffold proximity despite different internal diversity. The VAE shows 2–3× higher scaffold rediscovery for ABL1 (10.0–17.4% versus 3.2–5.5%) and better cluster rediscovery. Uniqueness among CVAE outputs is essentially saturated (99.99–100% of generated molecules distinct). The uniqueness *ratio* relative to the held-out set exceeds one (1.014–1.070) and is not a percentage. For GRL, PubChem random sampling matches the VAE on scaffold rediscovery above 20K samples (PubChem: 14.2–22.8% and

VAE: 13.9–18.0%), indicating that GRL-relevant scaffolds are broadly covered by PubChem and that model-based generation adds little over sampling the corpus for this target. The two modes therefore serve different purposes. Seed-based sampling recovers structures related to supplied examples, while label-only sampling under a broad label distribution produces more structurally varied outputs. Because the label distribution spans weak as well as strong binders, we do not interpret the CVAE arm as enriching for active compounds.

Scope and limitations. This evaluation measures scaffold recovery and structural diversity, not property optimization. The CVAE uses MMELON-predicted pseudo-labels on PubChem because experimental ChEMBL data are too sparse for direct fine-tuning. The pseudo-labels are accurate (RMSE 0.095–0.114 log-units on held-out ChEMBL data), but distributional shift between PubChem and ChEMBL means the CVAE steers toward a predicted activity range rather than confirmed active structures. In the rediscovery experiments, the CVAE was conditioned on labels sampled from the full training distribution to produce broad coverage across the label range. Wet-lab validation is required to confirm biological activity.

Ablation Studies

We varied five design factors one at a time against a fixed reference configuration (REF): (A1) tokenization scheme, (A2) stochastic latent bottleneck, (A3) conditioning scope, (A4) backbone fine-tuning strategy, and (A5) KL annealing schedule. All runs use the same 9,990 PubChem molecules and SA score conditioning task as the Conditional Generation subsection.

Reference configuration. REF uses SELFIES tokenization, a stochastic VAE bottleneck, joint encoder-input and decoder FiLM conditioning, LoRA adaptation ($r = 8$, $\alpha = 32$) inserted at the attention output and feed-forward projections of both encoder and decoder (effective at the feed-forward projections only, as described under Model Architecture), and cyclical β-annealing ($n_{\text{cycles}} = 4$, ratio $= 0.5$) with free bits $= 0.5$. REF trains 1,936,384 of 91,047,752 parameters. The conditional-prior networks are present but frozen at initialization in this configuration (Table 1). The backbone is the PubChem-80M pretrained SELFIES TransformerVAE with 12 encoder and decoder layers, $d_{\text{model}} = 512$, a 256-dimensional latent space, and a validation loss of 0.29. Fine-tuning uses Adam, lr $= 10^{-4}$, batch size 64, up to 20 epochs with early stopping (patience 8). All ablation results use $\alpha = 1$ (no label-reference guidance). The REF Spearman $\rho = 0.120$ is computed over 9,000 pooled molecules (3 bucket targets $\times$ 3 seeds $\times$ 1,000 molecules). Table 4 reports $\rho = 0.149$ at $\alpha = 1$ over 3,003 molecules swept across 1,001 uniform y^* levels. The higher value in Table 4 reflects the finer evaluation granularity. This rounded value is numerically similar to, but distinct from, the separate A3 decoder-FiLM-plus-decoder-LoRA result: Exp17-C uses the two-stage model and a 1,001-level

sweep, whereas A3 uses three fixed labels and three generation seeds.

Metrics. We report Spearman ρ between conditioning target y^* and realized SA score (pooled over 9,000 molecules: 3 targets $\times$ 3 seeds $\times$ 1,000 molecules), success rate SR_{10} (fraction of molecules within $|y_{\mathrm{realized}} - y^*| \leq 0.10$), internal diversity (IntDiv), and Bemis-Murcko scaffold diversity (ScafDiv = unique scaffolds / 9,000). Per-bucket SR_{10} is reported at three fixed conditioning targets: SR_{10_h} at $y^* = 0.165$ (easy synthesis, raw SA ≈ 2.4), SR_{10_m} at $y^* = 0.500$ (mid synthesis, raw SA ≈ 4.5), and SR_{10_l} at $y^* = 0.835$ (hard synthesis, raw SA ≈ 6.5). The subscripts $h/m/l$ denote high/mid/low success rate across buckets, which corresponds to easy/mid/hard synthesis difficulty.

Table 9: Configuration-study results. REF (A0): reference configuration (SELFIES, stochastic VAE, encoder+decoder conditioning, LoRA, cyclical β); the study was designed around nominal one-factor changes from REF. The A1 row is an exception because it uses a separately pretrained backbone with a different sequence-length budget. ρ: Spearman rank correlation (pooled 9,000 molecules). SR_{10}: success rate at $\varepsilon = 0.10$. $SR_{10_h}/SR_{10_m}/SR_{10_l}$: per-bucket SR_{10} at $y^* = 0.165/0.500/0.835$. ScafDiv: unique Bemis-Murcko scaffolds / 9,000. IntDiv: mean pairwise $1 - $Tanimoto (Morgan-2, 2048-bit). All runs: validity $= 1.000$. **Bold** marks values that exceed the REF baseline on that metric; values in rows with $\rho \approx 0$ (conditioning failure) that exceed REF are incidental, not indicative of improved conditioning.

Factor	Run	ρ	SR_{10}	SR_{10_h}	SR_{10_m}	SR_{10_l}	ScafDiv	IntDiv
A0: Reference								
REF	SELFIES + VAE + Enc + Dec FiLM + LoRA + Cyclical β	0.120	0.211	0.252	0.365	0.016	0.706	0.906
A1: Tokenization								
A1	SMILES (PubChem-80M backbone)	**0.416**	**0.272**	**0.687**	0.127	0.003	**0.722**	0.893
A2: Latent bottleneck								
A2	Deterministic AE (no KL)	0.034	0.202	0.180	**0.408**	**0.018**	0.607	**0.908**
A3: Conditioning scope								
A3	Encoder only (no FiLM)	0.000	0.195	0.235	0.339	0.011	0.260	0.904
A3	Decoder-side FiLM with decoder-only LoRA	**0.149**	**0.217**	**0.262**	0.364	**0.026**	0.619	**0.907**
A3	No conditioning	0.000	0.194	0.237	0.333	0.010	0.262	0.902
A4: Fine-tuning strategy								
A4	Full fine-tuning (89.8M trainable)	0.000	0.196	**0.283**	0.290	0.015	0.278	0.899
A4	Frozen backbone (heads only)	0.000	0.193	0.225	0.337	**0.018**	0.265	0.905
A5: KL annealing schedule								
A5	Monotone linear annealing	0.066	0.207	0.252	0.353	**0.017**	0.696	0.906
A5	Fixed $\beta = 1.0$	**0.153**	**0.217**	0.252	**0.373**	**0.026**	**0.737**	0.906

A1: Tokenization. The SMILES model achieves $\rho = 0.416$ versus REF $\rho = 0.120$, but this aggregate gain is concentrated at the easy-synthesis target. $SR_{10_h} = 0.687$ at $y^* = 0.165$, while $SR_{10_m} = 0.127$ and $SR_{10_l} = 0.003$ at mid and hard targets. PubChem-80M is skewed toward easily synthesizable molecules (normalized mean ≈ 0.24), so the SMILES prior is concentrated near $y = 0.2$. Conditioning on $y^* = 0.835$ therefore requires molecules that are underrepresented in pretraining, producing near-zero SR_{10} at that target. SELFIES keeps mid-range control higher ($SR_{10_m} = 0.365$ vs. 0.127 for SMILES) at the cost of

lower easy-target SR_{10}. This row is not a clean one-factor comparison, and we do not read it as isolating tokenization. The two arms use separately pretrained backbones that differ in several respects at once: the SELFIES backbone was pretrained for 260,496 optimizer steps on 8 GPUs at global batch size 256, whereas the SMILES backbone was pretrained for 100,980 steps on 4 GPUs at global batch size 1,024. The SMILES arm also uses a maximum sequence length of 128 tokens rather than 71, because character-level SMILES require longer sequences, which changes the positional-encoding extent and the total parameter count. Their validation losses are not comparable either, since the vocabularies and sequence lengths differ. The aggregate ρ difference therefore reflects tokenization confounded with pretraining trajectory, sequence-length budget, and model size.

A2: Stochastic bottleneck. Removing KL regularization (deterministic AE) reduces ρ from 0.120 to 0.034 and ScafDiv from 0.706 to 0.607. A plausible account is that KL regularization keeps the aggregated posterior close to $\mathcal{N}(0, I)$, so that Gaussian sampling at generation time lands in a region the decoder was trained on. Without it, encoded molecules may occupy arbitrary regions and prior samples fall outside the covered space. We did not measure the aggregated posterior directly, so this remains an interpretation consistent with the observed drop rather than a demonstrated mechanism.

A3: Conditioning scope. Encoder-only conditioning (no FiLM on the latent) gives $\rho = 0.000$ and ScafDiv $= 0.260$, matching the no-conditioning baseline ($\rho = 0.000$, ScafDiv $= 0.262$). The label does reach the latent parameters during training: it is projected and added to the pooled encoder representation immediately before the latent mean and log-variance heads, so a pathway exists. What this variant lacks is a pathway at *inference*: label-only generation in this ablation draws $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ rather than from a label-conditioned prior, and with FiLM disabled the requested label has no remaining route into the decoder. The observation is therefore that encoder-side injection alone produced no measurable generative control under the tested setup, not that no pathway to the latent exists. Decoder-side FiLM with decoder-only LoRA recovers $\rho = 0.149$ and ScafDiv $= 0.619$, showing that decoder-side FiLM is sufficient to produce measurable conditioning. Adding encoder-side injection (REF) raises ScafDiv to 0.706 at comparable ρ. Replication across training seeds finds these two configurations indistinguishable (Table 10), so we do not claim that either is the stronger pathway. Because no variant in this study samples from the learned conditional prior at generation time, this ablation does not isolate the contribution of that prior (Table 1).

A4: Fine-tuning strategy. Full fine-tuning and frozen backbone both give $\rho = 0$ and ScafDiv ≤ 0.278. Full fine-tuning substantially reduces scaffold diversity under the tested small-data optimization settings, from 0.706 to 0.278, consistent with loss of pretrained chemical coverage when all 89.8 M parameters are

updated on 9,990 molecules. The frozen backbone also failed to condition (ScafDiv = 0.265). LoRA, which in this implementation effectively adapts the position-wise feed-forward projections via low-rank updates while leaving the backbone weights intact, performed better than both tested alternatives ($\rho = 0.120$, ScafDiv = 0.706). Both alternatives were run at the learning rate and stopping policy inherited from the LoRA setting. We therefore do not draw a general conclusion about full fine-tuning or backbone freezing as adaptation strategies from these rows.

A5: KL annealing schedule. Fixed $\beta = 1.0$ gives $\rho = 0.153$ and ScafDiv 0.737. Cyclical annealing (REF) gives $\rho = 0.120$ and ScafDiv = 0.706. Monotone linear annealing gives $\rho = 0.066$ and ScafDiv = 0.696. This comparison also does not isolate the β schedule alone. In the implementation, the free-bits floor is applied only when cyclical annealing is enabled, so although all three configurations request free bits = 0.5, the constraint was active only in the cyclical (REF) run and inactive in the fixed-β and linear runs. Schedule and free-bits treatment therefore vary together across these rows, and we do not attribute the observed differences to the schedule by itself.

Cross-factor summary. Across the tested configurations, those that failed to condition ($\rho = 0$) all produced ScafDiv ≤ 0.28, while those that conditioned measurably ($\rho > 0$) all produced ScafDiv ≥ 0.61. This association is consistent with the interpretation that configurations retaining pretrained coverage are also the ones able to express the conditioning signal, but with one run per configuration and a single tokenization/backbone pairing per row, it is an observed pattern rather than a demonstrated causal relationship. Table 9 reports one training run per configuration (seed 42) evaluated at three generation seeds (42, 0, 101), which measures sampling variability at fixed weights but not variability from model fitting.

Replication across independent fine-tuning seeds. We trained six configurations with three different random seeds (Table 10). All other settings remained unchanged. We also evaluated each trained model with the same three generation seeds. To compare two configurations, we check whether their mean $\pm$ 1 SD intervals overlap. We do not apply a significance test because three training runs per configuration are insufficient for one.

The SMILES configuration gives $\rho = 0.422 \pm 0.049$, compared with 0.138 ± 0.039 for the reference. Their intervals do not overlap, so the difference is consistent across the three training seeds. However, the confounds described above remain, and this result does not show that tokenization alone causes the difference. Full fine-tuning and the frozen backbone give $\rho = 0.000$ in all three runs.

The intervals for the reference, decoder-side FiLM with decoder-only LoRA, and fixed β overlap. Their means differ by 0.003–0.033, while their standard deviations across training seeds are 0.014–0.039. We

therefore do not rank these three configurations. The results show a clear difference between configurations that achieve measurable control and those that do not, but they do not resolve smaller differences among the configurations that condition successfully.

Table 10: Spearman ρ across three independent fine-tuning runs for six SA-score ablation configurations. Each run differs only in its training seed. The model, data, objective, and generation protocol are identical, and every run is evaluated at the same three generation seeds (9,000 molecules per run). Seed 42 is the run reported in Table 9. "Range" is max − min across training seeds. Under a criterion of non-overlapping mean ± 1 SD, A1 and both A4 variants separate from the reference. The intervals for REF, A3 decoder-side FiLM with decoder-only LoRA, and A5 fixed β overlap. We do not apply a significance test because three runs per configuration cannot support one.

		Per training seed			
Configuration	**ρ (mean ± SD)**	**42**	**7**	**13**	**Range**
A1 SMILES tokenization	0.422 ± 0.049	0.416	0.376	0.474	0.098
A5 fixed $\beta = 1.0$	0.171 ± 0.026	0.153	0.201	0.158	0.048
A3 decoder-side FiLM with decoder-only LoRA	0.140 ± 0.014	0.149	0.147	0.125	0.024
REF (reference)	0.138 ± 0.039	0.120	0.183	0.110	0.072
A4 full fine-tuning	0.000 ± 0.000	0.000	0.000	0.000	0.000
A4 frozen backbone	0.000 ± 0.000	0.000	0.000	0.000	0.000

Discussion

Effective molecular generation for drug discovery requires separating three concerns: what the model knows from pretraining, how it adapts to a specific task, and which generation mode fits the current campaign stage. STAR-VAE addresses each concern with a distinct design decision: large-scale SELFIES-based VAE pretraining, LoRA plus FiLM conditioning for task adaptation, and two complementary generation modes. The experiments characterize how pretraining, adaptation, and inference choices affect controllability and structural diversity under the tested SA-score setting.

The MOSES and GuacaMol experiments show that a corpus-scale latent model can be localized to a narrower domain at inference time without fine-tuning, and they also delimit what that demonstration covers. On MOSES, STAR-VAE pretrained on PubChem-80M achieves FCD/Test = 2.01 in seed mode, versus FCD/Test = 6.61 for the MOSES-1.5M model. An overlap audit against the full curated PubChem source corpus (Appendix F) finds matches for 71.45% of MOSES training molecules and 87.04% of GuacaMol training molecules. These percentages upper-bound actual exposure because the source corpus is larger than the training dataset. Even as upper bounds, they preclude interpreting either benchmark as transfer to unfamiliar chemistry. The comparison also favours the smaller model in two respects: it was trained on the

MOSES training split and validated on the MOSES test split, and the seeds are drawn from that same training split. That corpus-scale pretraining still yields better seeded FCD under those conditions is the result worth reporting, and it is a statement about local sampling quality rather than about generalization. Pretraining on 79 million molecules produces a latent manifold where local sampling near any seed generates chemically coherent, distribution-aligned molecules. On GuacaMol, the two largest deviations from the ChEMBL24 reference distribution are DiffInternalSimilarities KL = 0.80 and NumAromaticRings KL = 0.18. They trace to the broader chemical diversity of PubChem. Seed-based generation on GuacaMol training molecules reduces all ten per-descriptor KL values to near zero with no training on GuacaMol data. Since the seeds are themselves GuacaMol training molecules, that setting measures reconstruction within a supplied domain and is not comparable to unconditional benchmark scores. Consistent with this, removing KL regularization in the ablation reduces Spearman ρ from 0.120 to 0.034 and ScafDiv from 0.706 to 0.607, which is what one would expect if prior sampling depends on the aggregated posterior remaining close to the prior. We did not measure the aggregated posterior directly, so we present this as a supporting observation rather than a demonstrated mechanism.

The ablation and baseline comparison experiments identify the fine-tuning strategy as the most consequential design decision. Full fine-tuning of all 89.8M parameters on 9,990 task molecules substantially reduces scaffold diversity under the tested small-data optimization settings, from 0.706 to 0.278, with Spearman ρ falling to zero, consistent with loss of pretrained chemical coverage. A frozen backbone also failed to condition ($\rho = 0$, ScafDiv = 0.265). Among the three strategies tested, only LoRA achieved non-zero controllability and high scaffold diversity simultaneously ($\rho = 0.120$, ScafDiv = 0.706), using low-rank updates that in this implementation act on the position-wise feed-forward projections while leaving the backbone weights intact. Both alternatives were run at hyperparameters inherited from the LoRA setting, so this is a statement about the tested configurations.

Within the conditioning design, the ablation also localizes where property information must enter the model. Encoder-only conditioning (no FiLM on the latent) gives $\rho = 0.000$, identical to the no-conditioning baseline. The label does reach the latent parameters during training, but with FiLM disabled and label-only generation drawing $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, it has no route into the decoder at inference. Decoder-side FiLM with decoder-only LoRA recovers $\rho = 0.149$, showing that a decoder-side pathway is sufficient for measurable conditioning. Adding encoder-side label injection (REF) raises ScafDiv to 0.706 at comparable ρ. Across three independent fine-tuning runs, the reference, decoder-side FiLM with decoder-only LoRA, and fixed-β variants have overlapping intervals. Their means differ by 0.003–0.033, while their between-seed standard deviations are 0.014–0.039. The ablation therefore supports a boundary rather than a ranking. Configurations with a decoder-side conditioning path show measurable control, while differences among these conditioning-

capable variants remain within run-to-run variation. No tested variant samples from the learned conditional prior at generation time, so these experiments do not measure that prior's incremental contribution.

The SA score conditioning experiment quantifies what this design achieves in practice. Spearman ρ rises from 0.15 at $\alpha = 1$ to 0.62 at $\alpha = 12$, while molecular validity stays above 98% across all guidance scales. The guidance scale therefore trades rank-level controllability for pointwise accuracy ($SR_{10} = 0.215$ at $\alpha = 12$) without degrading chemical validity. For docking targets, stronger guidance comes at a steeper cost: at $\alpha = 12$ for 1SYH, RDKit generation validity is 54.8% and docking completion is 34.7% under the sentinel-only failure rule. The α that maximizes rank correlation differs by target (2 for 6Y2F, 12 for 1SYH and 4LDE), but rank correlation alone is not a sufficient selection criterion: at $\alpha = 12$ for 1SYH, RDKit generation validity is 54.8%, docking completion is 34.7%, and both valid-/docked-only and attempt-level success rates are lower than at $\alpha = 1$. Reporting the full sweep rather than a single recommended value lets readers choose an operating point against their own tolerance for generation and docking failure.

The baseline comparison places these results in context. REINVENT 4 achieves $\rho = 0.941$ but uses ≈384,000 oracle calls and retains only 640 unique molecules of 1,001 requested in the easy-synthesis campaign, spanning just 61 unique scaffolds, consistent with reward saturation in the RL optimizer. MolGPT achieves $\rho = 0.895$ on its valid outputs, but 49.6% of generation attempts produce invalid SMILES, and its attempt-level SR_{10} across all 3,003 attempts falls to ≈0.346. DrugDiff shows negligible conditioning response ($\rho = -0.026$) in the evaluated cross-dataset configuration, in which its ZINC-250K SA score predictor is applied to PubChem molecules without retraining. STAR-VAE achieves $\rho = 0.618$, 98.4% validity, and 1,974 unique Murcko scaffolds without oracle access. These comparisons also clarify what ρ measures in each paradigm: REINVENT 4's 3-point Spearman tests only whether three bucket-level means are correctly ordered, while STAR-VAE's ρ is computed over 1,001 uniformly spaced conditioning targets. The two values are not directly comparable.

The rediscovery experiments address a different question: given that the model can generate molecules, which generation mode is appropriate for which task?

The rediscovery evaluation illustrates the trade-offs between label-conditioned generation using a CVAE and seed-based generation using a VAE. Across four ChEMBL target datasets and sample sizes ranging from 10,000 to 50,000 molecules, the seed-based VAE consistently outperformed the CVAE on exact rediscovery, scaffold recovery, and held-out cluster recovery, while the CVAE achieved higher diversity ratios and lower internal similarity; raw generated–held-out scaffold CrossSim was similar between the two modes. These are two different operating regimes rather than one uniformly better mode. Seed-based posterior sampling remains near the supplied molecular neighbourhoods and therefore recovers more held-out structures. Sampling without a structural seed, under a label distribution drawn proportionally from all ten activity bins,

yields more structurally varied outputs. Because that protocol requests weak as well as strong binders, the present experiment measures structural coverage under label-conditioned sampling and does not establish enrichment for active compounds. The comparison is further bounded by an overlap upper bound: 21.79%–82.45% of the held-out compounds per protein target match the full curated PubChem source corpus, which is larger than the training dataset (Appendix F). Exact rediscovery of such compounds may therefore be retrieval from the pretrained distribution, so we base the comparison on scaffold and cluster recovery, which credit reaching a framework or cluster rather than reproducing a specific structure.

One result warrants explicit attention. For GRL, PubChem random sampling matches the VAE on scaffold rediscovery above 20K samples (PubChem: 14.2–22.8% and VAE: 13.9–18.0%). This indicates that GRL-relevant scaffolds are broadly represented in PubChem, so model-based generation provides no advantage over random sampling from the pretraining corpus. The useful implication for practitioners is that the added value of the VAE over random sampling scales with how underrepresented the target scaffold class is in PubChem. For targets with abundant PubChem coverage, random sampling from the corpus is a reasonable baseline. For targets with sparser coverage, seed-based sampling recovered more of the held-out compounds in these experiments.

The VAE and CVAE serve different but complementary roles. Under seed-based sampling the pretrained VAE recovers target-associated held-out scaffolds, as measured across four ChEMBL targets. The CVAE shifts generation toward a requested scalar value, demonstrated here for SA score and QVina docking score and extensible in principle to other scalar endpoints. The two modes address different needs: recovering structures related to supplied examples, and exploring under a specified scalar objective.

Latent-variable models remain worth pursuing alongside large decoder-only and diffusion generators. Recent molecular foundation models have demonstrated the value of large pretraining corpora and chemically robust tokenizations, but they do not always offer the same direct mechanisms for latent interpolation, posterior inference, and probabilistic conditioning. STAR-VAE shows that these classical strengths of VAEs can be retained while incorporating modern components such as Transformer encoder-decoders, SELFIES tokenization, and parameter-efficient adaptation. In medicinal chemistry, control over how generated molecules change is useful alongside generation quality, and the latent-variable structure studied here supports both.

This study should be interpreted within the scope of its computational validation. The benchmarks assess distribution learning, property alignment, rediscovery behavior, and docking-based target conditioning, showing that the model captures useful chemical regularities and can steer generation meaningfully. However, docking scores, predicted properties, silver-label supervision, and rediscovery metrics are proxies; they do not provide definitive measures of biological efficacy. Wet-lab experiments are required to confirm these findings, which are beyond the scope of this study.

STAR-VAE operates at the sequence level without modeling 3D geometry or quantum-mechanical properties, unlike physics-informed models such as LiTENexus [53] or structure-based diffusion models such as TargetDiff [54] and DiffSBDD [55] that condition generation on protein pocket geometry. Its strength is fast, large-scale distribution learning and data-efficient conditional adaptation. A natural next step is to combine the two approaches: use STAR-VAE as a fast candidate generator, then filter or refine its outputs with physics-based scoring or pocket-conditioned optimization.

Several directions remain open. Conditioning could be extended to multiple simultaneous properties, guiding latent sampling jointly across activity and synthesizability. Integrating seed-based and label-conditioned generation within a single inference step would let users move between scaffold recovery and property-guided exploration without switching modes. Incorporating pocket geometry or molecular dynamics signals would strengthen target-conditioned generation beyond ligand-only docking supervision. Calibration and uncertainty quantification for the conditional prior and auxiliary predictors would improve the reliability of generated candidates for downstream decision-making.

The complementary capabilities of the VAE and CVAE support latent-variable Transformers as one practical architecture for large-scale molecular generation and scalar-conditioned adaptation.

Author Contributions

B.C.K. and **P.S.** led the project from conception to completion. They jointly conceptualized the model architecture and conditional generative framework, sourced and curated the pretraining and evaluation datasets, implemented model training and fine-tuning, designed and executed all benchmark and computational validation experiments, performed the analyses, and wrote the original manuscript draft. **B.S.**, **M.R.**, **S.S.**, **S.M.** and **J.A.M.** contributed to the implementation of model training and fine-tuning experiments and to the interpretation of results. **L.Z.** and **W.C.** contributed to the conceptualization of the computational validation strategy and to the sourcing and preparation of the datasets used in validation experiments. **P.M.** and **J.H.** contributed to scientific discussions and provided critical feedback on the manuscript. All authors contributed to writing and reviewing the manuscript throughout the research project and approved the final version for submission.

Funding

This work was supported by internal research funding from IBM Research. No external funding was received for this work.

Notes

The authors declare no competing financial interest.

Data Availability

All datasets and benchmark resources used in this work are publicly available. The PubChem dataset used for pre-training is available at `https://pubchem.ncbi.nlm.nih.gov/`. The MOSES benchmarking platform and its associated dataset are available at `https://github.com/molecularsets/moses`. The GuacaMol benchmark is available at `https://github.com/BenevolentAI/guacamol`; its reference molecules are derived from ChEMBL, which is available at `https://www.ebi.ac.uk/chembl/`. The Tartarus protein–ligand benchmark dataset is available at `https://github.com/aspuru-guzik-group/tartarus`. The synthetic accessibility (SA) score is computed with RDKit (`https://www.rdkit.org`) following Ertl and Schuffenhauer [46]. ChEMBL target datasets used for rediscovery evaluation (ABL1, BRD4, GRL, HIV-1 protease) are available at `https://www.ebi.ac.uk/chembl/`. The STAR-VAE source code, training scripts, and pretrained model checkpoints are available at `https://github.com/BiomedSciAI/STAR-VAE`.

Acknowledgments

The authors thank colleagues at IBM Research for helpful discussions throughout this project. AI-based tools were used to assist with code writing, code documentation, and text correction (spelling and grammar). The authors reviewed the full analysis pipeline to verify its correctness and completeness, and take full responsibility for the code and its documentation. All scientific content, including research design, experimental methodology, data analysis, interpretation of results, and all figures and tables, was produced by the authors. No AI tool was used to draft manuscript text, generate figures or data, or produce any substantive scientific content.

Supporting Information

The Supporting Information is included after the main manuscript. Appendix A: Formal definitions of all rediscovery evaluation metrics including exact discovery precision/recall, nearest-neighbor similarity, scaffold rediscovery, clustering held-out proportion, and diversity/uniqueness ratios (PDF). Appendix B: Representative samples of generated molecules from the rediscovery evaluation illustrating the complementary behavior of the pretrained VAE under seed-based posterior sampling and the label-conditioned CVAE (PDF). Ap-

pendix C: Full MOSES benchmark comparison table (Table S1) including all 15 metrics and all columns for all models (classic and modern baselines) (PDF). Appendix D: Per-bucket breakdown of SA score conditional generation (Table S2), including attempted and valid counts, retention or validity, success rates, and scaffold diversity for each target range and model (PDF). Appendix E: Reproducibility details covering software and hardware, training and conditional adaptation, validation and checkpoint selection, generation and decoding settings, evaluation procedures, and statistical reporting (PDF). Appendix F: Exact-structure overlap of the curated PubChem source corpus with MOSES, GuacaMol, and held-out ChEMBL sets (Tables S5 and S6), providing upper bounds on pretraining exposure (PDF).

References

[1] Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design* **2013**, *27*, 675–679.

[2] Gao, W.; Coley, C. W. The synthesizability of molecules proposed by generative models. *Journal of chemical information and modeling* **2020**, *60*, 5714–5723.

[3] Di, L.; Kerns, E. H.; Carter, G. T. Drug-like property concepts in pharmaceutical design. *Current pharmaceutical design* **2009**, *15*, 2184–2194.

[4] Cheng, Y.; Gong, Y.; Liu, Y.; Song, B.; Zou, Q. Molecular design in drug discovery: a comprehensive review of deep generative models. *Briefings in Bioinformatics* **2021**, *22*, bbab344.

[5] Tang, X.; Dai, H.; Knight, E.; Wu, F.; Li, Y.; Li, T.; Gerstein, M. A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation. *Briefings in Bioinformatics* **2024**, *25*, bbae338.

[6] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **2018**, *4*, 268–276.

[7] Simonovsky, M.; Komodakis, N. Graphvae: Towards generation of small graphs using variational autoencoders. International conference on artificial neural networks. 2018; pp 412–422.

[8] Ma, C.; Zhang, X. GF-VAE: a flow-based variational autoencoder for molecule generation. Proceedings of the 30th ACM international conference on information & knowledge management. 2021; pp 1181–1190.

[9] Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. International conference on machine learning. 2018; pp 2323–2332.

[10] Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of cheminformatics* **2018**, *10*, 31.

[11] Kang, S.; Cho, K. Conditional molecular design with deep generative models. *Journal of chemical information and modeling* **2018**, *59*, 43–52.

[12] Yang, M.; Feng, S.; Liao, Y.; Liu, R.; Dong, J.; Wang, D.; Deng, L.; Guan, J.; Yuan, Y. CMGN: a conditional molecular generation net to design target-specific molecules with desired properties. *Briefings in Bioinformatics* **2023**, *24*, bbad185.

[13] Bagal, V.; Aggarwal, R.; Vinod, P.; Priyakumar, U. D. MolGPT: molecular generation using a transformer-decoder model. *Journal of chemical information and modeling* **2021**, *62*, 2064–2076.

[14] Fang, Y.; Zhang, N.; Chen, Z.; Guo, L.; Fan, X.; Chen, H. Domain-agnostic molecular generation with chemical feedback. *arXiv preprint arXiv:2301.11259* **2023**,

[15] Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences* **1989**, *29*, 97–101.

[16] Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **2022**, *3*, 015022.

[17] Priyadarsini, I.; Takeda, S.; Hamada, L.; Brazil, E. V.; Soares, E.; Shinohara, H. Self-bart: A transformer-based molecular representation model using selfies. *arXiv preprint arXiv:2410.12348* **2024**,

[18] Priyadarsini, I.; Takeda, S.; Hamada, L.; Brazil, E. V.; Soares, E.; Shinohara, H. SELFIES-TED : A Robust Transformer Model for Molecular Representation using SELFIES. 2025; `https://openreview.net/forum?id=uPj9oBH80V`.

[19] Li, C.; Yamanishi, Y. SpotGAN: A reverse-transformer GAN generates scaffold-constrained molecules with property optimization. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2023; pp 323–338.

[20] Li, C.; Yamanishi, Y. Tengan: Pure transformer encoders make an efficient discrete gan for de novo molecular generation. International Conference on Artificial Intelligence and Statistics. 2024; pp 361–369.

[21] Kuznetsov, M.; Polykovskiy, D. MolGrow: A graph normalizing flow for hierarchical molecular generation. Proceedings of the AAAI Conference on Artificial Intelligence. 2021; pp 8226–8234.

[22] Jo, J.; Lee, S.; Hwang, S. J. Score-based generative modeling of graphs via the system of stochastic differential equations. International conference on machine learning. 2022; pp 10362–10383.

[23] Huang, H.; Sun, L.; Du, B.; Lv, W. Conditional diffusion based on discrete graph structures for molecular graph generation. Proceedings of the AAAI Conference on Artificial Intelligence. 2023; pp 4302–4311.

[24] Lee, S.; Jo, J.; Hwang, S. J. Exploring chemical space with score-based out-of-distribution generation. International Conference on Machine Learning. 2023; pp 18872–18892.

[25] Park, S.; Kim, H.; Lee, S.; Ryu, J.-K.; Kim, S. Global and local integrated gradient-based diffusion model for de novo drug design. *Briefings in Bioinformatics* **2026**, *27*, bbag033.

[26] Norouzi, R.; Karimi, M.; Abadpour, A.; Abbasi, A. MT-ConBiFormer-GPT: multi-target molecular generation for low-data drug discovery via a contrastive BiFormer-GPT architecture and curriculum learning with cross-domain generalization. *Briefings in Bioinformatics* **2026**, *27*, bbag079.

[27] Kang, S.-g.; Weber, J. K.; Morrone, J. A.; Zhang, L.; Huynh, T.; Cornell, W. D. In-Pocket 3D Graphs Enhance Ligand–Target Compatibility in Generative Small-Molecule Creation: A Dopamine D2 Receptor Model System. *The Journal of Physical Chemistry B* **2026**, *130*, 2965–2973.

[28] Du, Y.; Jamasb, A. R.; Guo, J.; Fu, T.; Harris, C.; Wang, Y.; Duan, C.; Liò, P.; Schwaller, P.; Blundell, T. L. Machine learning-aided generative molecular design. *Nature Machine Intelligence* **2024**, *6*, 589–604.

[29] 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* **2020**, *1*, 045024.

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

[31] Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* **2019**, *59*, 1096–1108.

[32] Nigam, A.; Pollice, R.; Tom, G.; Jorner, K.; Willes, J.; Thiede, L.; Kundaje, A.; Aspuru-Guzik, A. Tartarus: A Benchmarking Platform for Realistic And Practical Inverse Molecular Design. Advances in Neural Information Processing Systems. 2023; pp 3263–3306.

[33] Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; others ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research* **2019**, *47*, D930–D940.

[34] Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 update. *Nucleic Acids Research* **2022**, *51*, D1373–D1380, _eprint: https://academic.oup.com/nar/article-pdf/51/D1/D1373/48441598/gkac956.pdf.

[35] Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations* **2022**,

[36] Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* **2022**,

[37] Noutahi, E.; Gabellini, C.; Craig, M.; Lim, J. S. C.; Tossou, P. Gotta be SAFE: A New Framework for Molecular Design. *Digital Discovery* **2024**, *3*, 796–804.

[38] Vignac, C.; Krawczuk, I.; Siraudin, A.; Wang, B.; Cevher, V.; Frossard, P. DiGress: Discrete Denoising Diffusion for Graph Generation. International Conference on Learning Representations. 2023.

[39] Kim, J.; others FragFM: Hierarchical Framework for Efficient Molecule Generation via Fragment-Level Discrete Flow Matching. *arXiv preprint arXiv:2502.15805* **2025**,

[40] Jung, S.; others MolHIT: Advancing Molecular-Graph Generation with Hierarchical Discrete Diffusion. *arXiv preprint arXiv:2602.17602* **2026**,

[41] Zheng, S.; others Graph VQ-Transformer (GVT): Fast and Accurate Molecular Generation via VQ-VAE and Transformer. *arXiv preprint arXiv:2512.02667* **2025**,

[42] Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science* **2018**, *4*, 120–131.

[43] Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. ICLR. 2014.

[44] Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I. J. Adversarial Autoencoders. *CoRR* **2015**, *abs/1511.05644*.

[45] Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J. Chem. Inf. Model.* **2022**, *62*, 2064–2076.

[46] Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics* **2009**, *1*, 8.

[47] Loeffler, H. H.; He, J.; Tibo, A.; Janet, J. P.; Voronov, A.; Mervin, L. H.; Engkvist, O. REINVENT 4: Modern AI–driven generative molecule design. *Journal of Cheminformatics* **2024**, *16*, 20.

[48] Oestreich, M.; Bremer, J.; Scher, E. DrugDiff: a latent diffusion model for small molecule generation guided by molecular properties. *Journal of Cheminformatics* **2025**, *17*, 36.

[49] Jin, J.; Zeng, C.; Ambrose, C.; Liu, M. MolGuidance: Advanced Guidance Strategies for Conditional Molecular Generation with Flow Matching. *Journal of Chemical Information and Modeling* **2025**, *66*, 8860–8872.

[50] Zeng, C.; Jin, J.; Ambrose, C.; Karypis, G.; Transtrum, M.; Tadmor, E. B.; Hennig, R. G.; Roitberg, A.; Martiniani, S.; Liu, M. PropMolFlow: property-guided molecule generation with geometry-complete flow matching. *Nature Computational Science* **2026**, *6*, 1–10.

[51] Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893.

[52] Suryanarayanan, P.; Qiu, Y.; Sethi, S.; Mahajan, D.; Li, H.; Yang, Y.; Eyigoz, E.; Saenz, A. G.; Platt, D. E.; Rumbell, T. H.; others Multi-view biomedical foundation models for molecule-target and property prediction. *arXiv preprint arXiv:2410.19704* **2024**,

[53] Liu, S.; others LiTENexus: A Physics-Informed End-to-End Framework for Molecular Generation and Optimization. *arXiv preprint* **2024**,

[54] Guan, J.; Qian, W. W.; Peng, X.; Su, Y.; Peng, J.; Ma, J. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. *International Conference on Learning Representations* **2023**,

[55] Schneuing, A.; Du, Y.; Harris, C.; Jamasb, A.; Igashov, I.; Du, W.; Blundell, T.; Lió, P.; Gomes, C.; Welling, M.; others Structure-based Drug Design with Equivariant Diffusion Models. *Nature Computational Science* **2024**, *4*, 899–909.

Supporting Information

STAR-VAE: A Scalable Latent-Variable Transformer for Controllable Molecular Generation

Bum Chul Kwon[1,*], Ben Shapira[3], Moshiko Raboh[3], Shreyans Sethi[2], Shruti Murarka[2], Joseph A. Morrone[2], Leili Zhang[2], Wendy Cornell[2], Pablo Meyer[2], Jianying Hu[2], Parthasarathy Suryanarayanan[2,†]

[1]IBM Research, Cambridge, MA 02142, USA

[2]IBM Research, Yorktown Heights, NY 10598, USA

[3]IBM Research, Haifa, Israel

[*]Corresponding author: bumchul.kwon@us.ibm.com [†]Corresponding author: psuryan@us.ibm.com

This Supporting Information provides technical details in six appendices. Appendix A defines the rediscovery metrics, including exact discovery precision and recall, nearest-neighbor similarity, scaffold rediscovery, cluster recovery, and diversity and uniqueness ratios. Appendix B shows representative molecules from the rediscovery evaluation for seed-based VAE sampling and label-conditioned CVAE sampling. Appendix C reports all 15 MOSES metrics for the evaluated models, including retrained and literature baselines. Appendix D reports validity, success rate, and scaffold diversity for each SA-score target range and each conditional-generation model. Appendix E lists the training, generation, evaluation, and statistical-reporting details needed to interpret the experiments. Appendix F measures exact molecular overlap between the curated PubChem source corpus and the MOSES, GuacaMol, and held-out ChEMBL sets. Because the audited corpus is larger than the training dataset, these percentages are upper bounds on pretraining exposure.

Rediscovery Metric Definitions

Rediscovery Metric Definitions

Overview

We evaluate molecular rediscovery with a cluster-based held-out split. For each target, we partition the full dataset by Butina clustering using Morgan fingerprints with radius 2, 2048 bits, and a similarity threshold of 0.6. We reserve about 20% of the clusters for evaluation and stratify the selection across 10 score bins. so that entire clusters, instead of individual molecules, are withheld. This split removes close analogues of the held-out compounds from the target-specific fine-tuning set. It does not remove those compounds from the

79-million-molecule PubChem source corpus. The overlap audit in Appendix finds source-corpus matches for 21.79%–82.45% of the held-out compounds, depending on the protein target. Because the audited corpus contains more molecules than the training dataset, these values are upper bounds on pretraining exposure. We therefore describe this task as recovery of *task-held-out* ChEMBL compounds and treat exact rediscovery as a bounded observation.

Let:

- T: training set (molecules in training clusters)
- H: held-out set (molecules in held-out clusters)
- G: molecules generated by the model. We treat G as a multiset, so the same molecule can appear more than once and $|G|$ counts samples. Scaffold similarity and clustering use deduplicated sets derived from G. Uniqueness uses the full multiset.

We denote the canonical SMILES of a molecule m by $c(m)$, Morgan fingerprints (radius = 2, 2048 bits) by $f(m)$, and Tanimoto similarity by $\mathrm{sim}(a, b) = \mathrm{Tanimoto}(f(a), f(b))$. Murcko scaffolds are computed with the Bemis–Murcko decomposition.

Exact Discovery Precision and Recall (Exact P / Exact R)

Exact rediscovery records whether the model reproduces a held-out molecule up to canonical SMILES equivalence.

Let

$$H_{\mathrm{canon}} = \{c(h) : h \in H\}, \qquad G_{\mathrm{canon}} = \{c(g) : g \in G\}.$$

The set of exactly rediscovered molecules is

$$R_{\mathrm{exact}} = H_{\mathrm{canon}} \cap G_{\mathrm{canon}}.$$

Exact Discovery Precision (Exact P):

$$\mathrm{Exact\,P} = \frac{|R_{\mathrm{exact}}|}{|G_{\mathrm{canon}}|}.$$

Exact Discovery Recall (Exact R):

$$\mathrm{Exact\,R} = \frac{|R_{\mathrm{exact}}|}{|H_{\mathrm{canon}}|}.$$

Because exact matches are rare, both values are reported in units of $\times 10^{-4}$ in the main results table.

Nearest Neighbor Mean Similarity Score (NN Sim)

For each *generated* molecule $g \in G$, let its best-matching held-out molecule be

$$s_{\text{best}}(g) = \max_{h \in H} \text{sim}(g, h).$$

The **Nearest Neighbor Mean Similarity (NN Sim)** is the average of these best-match scores across all generated molecules:

$$\text{NN Sim} = \frac{1}{|G|} \sum_{g \in G} s_{\text{best}}(g).$$

Higher values mean that generated molecules lie closer to held-out molecules in fingerprint space. This direction measures the proximity of generated molecules to the held-out set. Averaging in the reverse direction would measure coverage of the held-out set. These directions are not interchangeable. We report the implemented proximity direction.

Clustering Held-Out Proportion (Clust 0.6)

We apply Butina clustering to the full dataset with the same fingerprint settings. Each cluster belongs to the training or held-out split. For each generated molecule $g \in G$, we find the nearest cluster by Tanimoto similarity and denote the score by $s_{\text{clust}}(g)$. We count the assignment only when $s_{\text{clust}}(g) \geq \tau$.

Here $\tau = 0.6$ is an evaluation-time assignment cutoff. It is separate from the threshold used to construct the Butina clusters. Without this cutoff, every generated molecule would receive a cluster assignment regardless of distance.

Let C_H denote the held-out clusters. This metric is a cluster recall. The numerator counts held-out clusters reached by at least one qualifying generated molecule, and the denominator counts all held-out clusters. It is not the fraction of generated molecules assigned to held-out clusters.

Clustering Held-Out Proportion at threshold τ (Clust 0.6):

$$\text{Clust}\,0.6 = \frac{\left|\{\, c \in C_H \;:\; \exists\, g \in G \text{ with } s_{\text{clust}}(g) \geq \tau \text{ and } g \text{ assigned to } c \,\}\right|}{|C_H|}.$$

Generating multiple molecules in the same cluster does not increase the score. The value is bounded by 1 for any sample size. It measures coverage of held-out cluster space at the stated cutoff.

Scaffold Rediscovery (Scaff Redisc)

Let scaff(m) denote the Bemis–Murcko scaffold of molecule m. Define the scaffold sets

$$S_H = \{\text{scaff}(h) : h \in H\}, \qquad S_G = \{\text{scaff}(g) : g \in G\}.$$

Scaffold Rediscovery (Scaff Redisc):

$$\text{Scaff Redisc} = \frac{|S_H \cap S_G|}{|S_H|}.$$

This metric gives the fraction of held-out scaffolds recovered by the generated set. It evaluates core chemical frameworks instead of exact structures.

Cross-Similarity between Scaffold Sets (CrossSim)

Cross-Similarity (CrossSim) is the mean Tanimoto similarity over all generated–held-out scaffold pairs:

$$\text{CrossSim} = \frac{1}{|S_G| \cdot |S_H|} \sum_{s \in S_G} \sum_{h \in S_H} \text{sim}(s, h).$$

S_G and S_H are deduplicated scaffold sets, so repeated scaffolds do not add repeated pairs. CrossSim lies in $[0, 1]$ and summarizes average proximity between the two scaffold sets. It is not a nearest-neighbor distance. Because it averages all pairs, a high value does not show that generated molecules occupy a specific held-out region. The released code includes a worked example and unit tests.

Internal Similarity of Generated Scaffolds (InterSim)

Internal Similarity (InterSim) measures the average pairwise Tanimoto similarity among scaffolds within the generated set:

$$\text{InterSim} = \frac{1}{|S_G|(|S_G| - 1)} \sum_{\substack{s, s' \in S_G \\ s \neq s'}} \text{sim}(s, s').$$

Lower InterSim indicates greater scaffold diversity in the generated output. Higher values indicate that generated molecules share similar core frameworks.

Diversity Ratio (Div)

Let $D(X)$ denote the internal diversity of a set X, defined as the average pairwise Tanimoto *dissimilarity* $(1 - \text{sim})$ among molecules in X.

Diversity Ratio (Div):

$$\text{Div} = \frac{D(G)}{D(H)}.$$

Values greater than 1 indicate that the generated set is more diverse than the held-out set. Values below 1 indicate less diversity.

Uniqueness Ratio (Uniq)

Let $U(X) = |X_{\text{canon}}|/|X|$ denote the uniqueness of a collection X. It divides the number of distinct canonical SMILES by the number of samples. This definition applies because X is a multiset. For a set, it would always equal 1.

Uniqueness Ratio (Uniq):

$$\text{Uniq} = \frac{U(G)}{U(H)}.$$

Values near 1 mean that the generated and held-out collections have similar uniqueness. The ratio can exceed 1 because both its numerator and denominator are fractions. We therefore do not report it as a percentage.

Generated Molecules

Figure S1 shows representative examples of the two generation modes evaluated in the main manuscript. Column 1 contains five task-held-out molecules reproduced exactly by the VAE. The overlap audit only bounds exposure to the source corpus, so these examples do not show that the structures were absent from pretraining. In Columns 2–3, VAE outputs remain close to the reference scaffold, with Tanimoto similarities above 0.6. This pattern agrees with the scaffold and cluster recovery results in Table 8 of the main manuscript. The selected CVAE outputs in Columns 4–5 are more varied, in line with their higher aggregate diversity ratios and lower internal similarity across the four ChEMBL targets. These examples were selected for illustration. They are not a random sample and do not establish that conditioning caused the difference in diversity.

Full MOSES Benchmark Comparison (All 15 Metrics)

Table S1 reports all 15 MOSES metrics for every model evaluated in the main text. We obtained values for the eight classic baselines (HMM, NGram, Combinatorial, CharRNN, AAE, VAE, JTN-VAE, and LatentGAN) from the official molecularsets repository [30]. We obtained values for the five modern baselines (SAFE-

GPT, DiGress, FragFM, MolHIT, and GVT) from their publications. These papers report only Validity, Unique@10k, FCD/Test, Filters, and Novelty. We mark the other ten columns with —. For STAR-VAE, values are means ± standard deviations over three evaluation seeds.

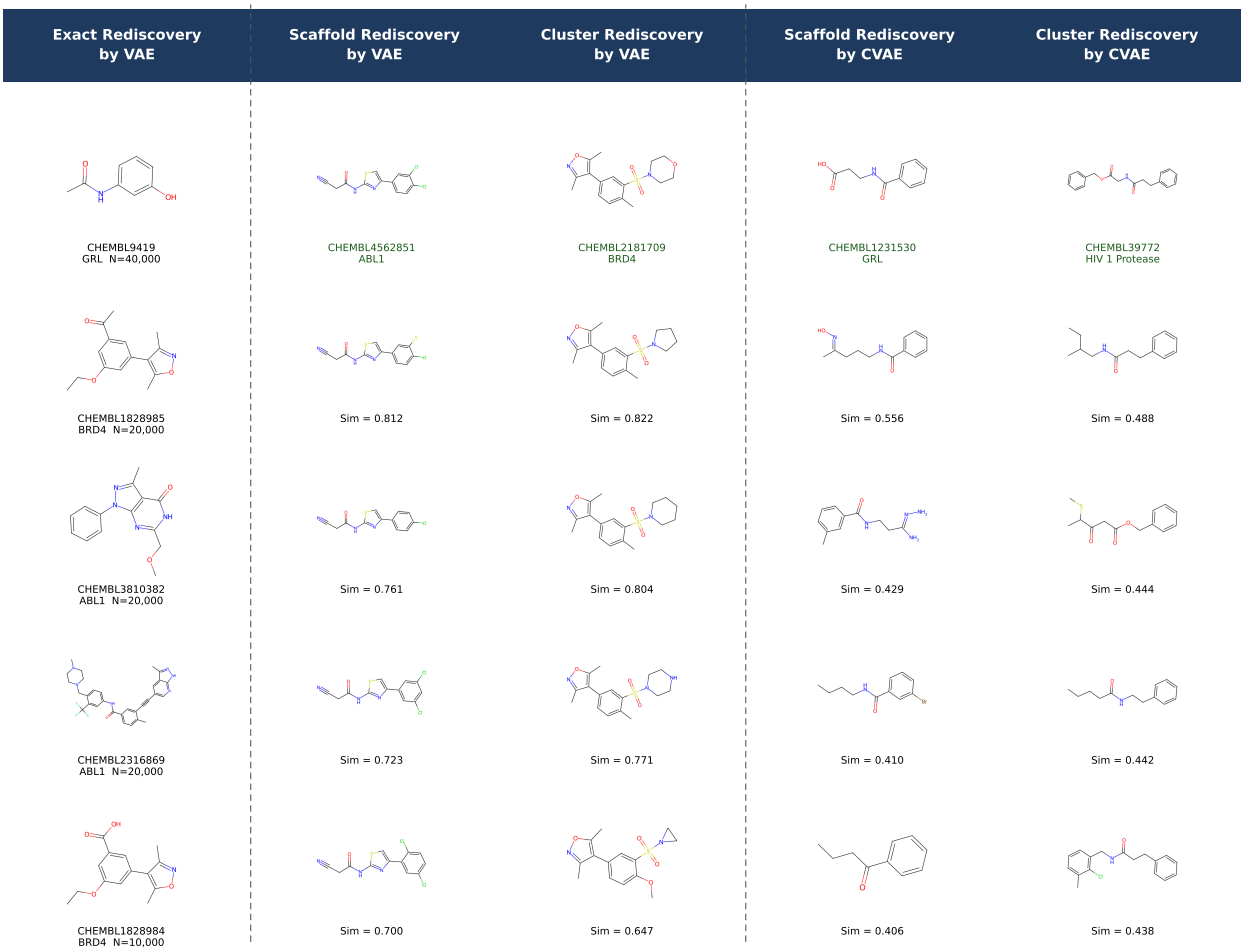


Figure S1: Samples of generated molecules from the rediscovery evaluation. Column 1 lists five held-out molecules exactly reproduced by the VAE (ChEMBL ID, dataset, N). Columns 2–3 show scaffold and cluster rediscovery by the VAE. Columns 4–5 present the same for the CVAE. In Columns 2–3 and 4–5, a reference molecule in the held-out set is shown in the first row, followed by its four most similar generated molecules by Tanimoto similarity. The VAE outputs share nearly identical scaffolds or cluster centroids (Tanimoto > 0.6), whereas the selected CVAE outputs are more varied. These examples are illustrative and do not establish a causal effect of conditioning.

Alt text: A grid of selected chemical structure diagrams across five columns. Column 1 shows five task-held-out molecules exactly reproduced by the VAE and labeled by ChEMBL identifier. Columns 2 and 3 show VAE scaffold and cluster recovery. A reference molecule is followed by four generated molecules with Tanimoto similarity above 0.6. Columns 4 and 5 show analogous CVAE examples selected by the same nearest-neighbor display procedure.

Table S1: Full MOSES benchmark comparison across all 15 metrics. Baseline values (HMM, NGram, Combinatorial, CharRNN [42], AAE [44], VAE [43], JTN-VAE [9], LatentGAN [17]) are retrieved directly from the official molecularsets (MOSES) repository [30]. Our STAR-VAE runs are reported as means ± standard deviations computed over three independent evaluation seeds. Modern baseline rows (SAFE-GPT, DiGress, FragFM, MolHIT, GVT) are literature values (†) taken from the cited publications. Columns not reported in the original papers are indicated by —.

Models	Valid (↑)	Unique@1k (↑)	Unique@10k (↑)	FCD/Test (↓)	FCD/TestSF (↓)	SNN/Test (↑)	SNN/TestSF (↑)	Frag/Test (↑)	Frag/TestSF (↑)	Scaf/Test (↑)	Scaf/TestSF (↑)	IntDiv1 (↑)	IntDiv2 (↑)	Filters (↑)	Novelty (↑)
Reference Split															
Train	*1.0*	*1.0*	*1.0*	*0.008*	*0.4755*	*0.6419*	*0.5859*	*1.0*	*0.9986*	*0.9907*	*0.0*	*0.8567*	*0.8508*	*1.0*	*1.0*
Classic baselines															
HMM	0.076±0.0322	0.623±0.1224	0.5671±0.1424	24.4661±2.5251	25.4312±2.5599	0.3876±0.0107	0.3795±0.0107	0.5754±0.1224	0.5681±0.1218	0.2065±0.0481	0.049±0.018	0.8466±0.0403	0.8104±0.0507	0.9024±0.0489	0.9994±0.001
NGram	0.2376±0.0025	0.974±0.0108	0.9217±0.0019	5.5069±0.1027	6.2306±0.0966	0.5209±0.001	0.4997±0.0005	0.9846±0.0012	0.9815±0.0012	0.5302±0.0163	0.0977±0.0142	0.8738±0.0002	0.8644±0.0002	0.9582±0.001	0.9694±0.001
Combinatorial	1.0±0.0	0.9983±0.0015	0.9909±0.0009	4.2375±0.037	4.5113±0.0274	0.4514±0.0003	0.4388±0.0002	0.9912±0.0004	0.9904±0.0003	0.4445±0.0056	0.0865±0.0027	0.8732±0.0002	0.8666±0.0002	0.9557±0.0018	0.9878±0.0008
CharRNN	0.9748±0.0264	1.0±0.0	0.9994±0.0003	**0.0732**±0.0247	**0.5204**±0.0379	0.6015±0.0206	0.5649±0.0142	**0.9998**±0.0002	**0.9983**±0.0003	0.9242±0.0058	**0.1101**±0.0081	0.8562±0.0005	0.8503±0.0005	**0.9943**±0.0034	0.8419±0.0509
AAE	0.9368±0.0341	1.0±0.0	0.9973±0.002	0.5555±0.2033	1.0572±0.2375	0.6081±0.0043	0.5677±0.0045	0.991±0.0051	0.9905±0.0039	0.9022±0.0375	0.0789±0.009	0.8557±0.0031	0.8499±0.003	0.996±0.0006	0.7931±0.0285
VAE	0.9767±0.0012	1.0±0.0	0.9984±0.0005	0.099±0.0125	0.567±0.0338	0.6257±0.0005	0.5783±0.0008	0.9994±0.0001	0.9984±0.0003	0.9386±0.0021	0.0588±0.0095	0.8558±0.0004	0.8498±0.0004	0.997±0.0002	0.6949±0.0069
JTN-VAE	1.0±0.0	1.0±0.0	0.9996±0.0003	0.3954±0.0234	0.9382±0.0531	0.5477±0.0076	0.5194±0.007	0.9965±0.0003	0.9947±0.0002	0.8964±0.0039	0.1009±0.0105	0.8551±0.0034	0.8493±0.0035	0.976±0.0016	0.9143±0.0058
LatentGAN	0.8966±0.0029	1.0±0.0	0.9968±0.0002	0.2968±0.0087	0.8281±0.0117	0.5371±0.0004	0.5132±0.0002	0.9986±0.0004	0.9972±0.0007	0.8867±0.0009	0.1072±0.0098	0.8565±0.0007	0.8505±0.0006	0.9735±0.0006	0.9498±0.0006
Modern baselines († literature values)															
SAFE-GPT† [37]	0.998	—	0.989	0.72	—	0.57	—	—	—	0.063	—	—	—	0.977	0.437
DiGress† [38]	0.857	—	1.000	1.19	—	0.52	—	—	—	0.148	—	—	—	0.971	0.950
FragFM† [39]	0.998	—	1.000	0.58	—	0.56	—	—	—	0.109	—	—	—	0.991	0.871
MolHIT† [40]	0.991	—	0.998	1.03	—	0.55	—	—	—	0.144	—	—	—	0.980	0.914
GVT† [41]	0.994	—	1.000	0.16	—	**0.64**	—	—	—	0.034	—	—	—	0.996	0.614
Ours (STAR-VAE)															
STAR-VAE PubChem 80M (Random)	**1.0000**±0.0000	**1.0000**±0.0000	**1.0000**±0.0000	14.1140±0.0607	15.0960±0.0637	0.3140±0.0002	0.3059±0.0003	0.9116±0.0012	0.9058±0.0013	0.2831±0.0095	0.1303±0.0041	**0.8990**±0.0001	**0.8939**±0.0001	0.4766±0.0014	**0.9999**±0.0000
STAR-VAE PubChem 80M (Seed)	**1.0000**±0.0000	**1.0000**±0.0000	0.9998±0.0000	2.0062±0.0084	2.7635±0.0122	0.5090±0.0007	0.4798±0.0005	0.9964±0.0001	0.9931±0.0001	**0.8987**±0.0029	0.0708±0.0055	0.8683±0.0002	0.8626±0.0002	0.8313±0.0013	0.7583±0.0010
STAR-VAE MOSES 1.5M (Random)	**1.0000**±0.0000	**1.0000**±0.0000	0.9999±0.0001	9.1489±0.0291	10.0725±0.0265	0.3842±0.0003	0.3722±0.0004	0.9676±0.0011	0.9598±0.0012	0.5622±0.0120	0.1123±0.0112	0.8809±0.0001	0.8755±0.0001	0.6398±0.0029	0.9939±0.0002
STAR-VAE MOSES 1.5M (Seed)	**1.0000**±0.0000	0.9997±0.0005	0.9996±0.0002	6.6121±0.0335	7.5617±0.0315	0.4263±0.0003	0.4080±0.0003	0.9795±0.0001	0.9731±0.0002	0.7583±0.0057	0.0880±0.0124	0.8806±0.0001	0.8745±0.0001	0.6931±0.0018	0.9567±0.0004

Per-Bucket Breakdown: SA score Conditional Generation

Table S2 separates the aggregate SA-score benchmark results into three target regions: low y^* (easy synthesis, $y^* \approx 0.167$), middle y^* ($y^* \approx 0.500$), and high y^* (hard synthesis, $y^* \approx 0.833$). All rows use the same RDKit SA scorer and min–max scaler as the main experiment.

Denominator convention. We assign each attempt to a bucket from its *requested* label y^* before validity filtering. Every attempt therefore belongs to one bucket, including attempts that do not produce a parseable molecule. STAR-VAE and MolGPT sample three molecules at each of 1,001 y^* values. This gives $n_{\text{att}} = 999/1002/1002$ in the low, middle, and high regions, for 3,003 attempts in total. REINVENT 4 ran three campaigns of 1,001 requests (`num_smiles` = 1001, `unique_molecules` = true) and retained 640, 996, and 992 unique valid molecules. Its sampler removes duplicate and unparseable outputs. The low-y^* campaign recorded no sanitization failures, so nearly all 361 removed outputs were duplicates. We therefore describe n_{val} for REINVENT 4 as the number of *unique valid* molecules retained from 1,001 requests and label the ratio "Retain." DrugDiff generates 1,001 molecules per bucket and is valid by construction. We report two success rates because output retention differs across models. SR^{val} divides by n_{val}. SR^{att} divides by n_{att} and counts invalid or unscored attempts as failures. The table reports both n_{att} and n_{val} so that validity and retention are calculated from the appropriate denominators.

Bucket-level observations. **MolGPT.** Validity decreases most strongly in the high-y^* region. Only 244 of 1,002 attempts (24.4%) produce valid SMILES, compared with 756 of 999 attempts (75.7%) in the low-y^* region.

REINVENT 4. The low-y^* campaign produces 61 unique Murcko scaffolds from 640 retained molecules (Scaf/mol = 0.095). The high-y^* campaign produces 979 scaffolds from 992 molecules (Scaf/mol = 0.987). The reward used in each campaign may contribute to this difference. The low-y^* campaign has the lowest retention outside MolGPT (640/1001 = 63.9%). Because REINVENT 4 removes duplicates, this result reflects repeated generation as well as chemical validity. Its $\text{SR}^{\text{att}}_{20} = 0.639$ measures success per requested molecule.

STAR-VAE. Validity remains between 97% and 100% across the three regions (995/999, 985/1002, and 976/1002). The model produces 1,974 distinct scaffolds, the highest total in this comparison. Scaf/mol declines from 0.774 at low y^* to 0.578 at high y^*. Hard-synthesis structures are less common in the training corpus, which may contribute to this trend.

DrugDiff. All generated molecules are valid, but the mean SA scores do not move toward the requested values. The observed means are 0.521, 0.541, and 0.513 for targets of 0.167, 0.500, and 0.833. Under this

test, the predictor trained on ZINC-250K does not transfer to the PubChem distribution.

Table S2: Per-bucket breakdown of SA score conditional generation metrics for all four models. Each row reports results for one model–bucket combination. **Bucket**: low y^* = easy synthesis ($y^* \approx 0.167$, $\mathrm{SA}_{\mathrm{raw}} \approx 2.45$); mid y^* ($y^* \approx 0.500$, $\mathrm{SA}_{\mathrm{raw}} \approx 4.48$); high y^* = hard synthesis ($y^* \approx 0.833$, $\mathrm{SA}_{\mathrm{raw}} \approx 6.49$). n_{att}: generation attempts assigned to the bucket by requested label, *before* validity filtering. n_{val}: valid SMILES retained. Ret.: $n_{\mathrm{val}}/n_{\mathrm{att}}$; this is RDKit validity for STAR-VAE, MolGPT, and DrugDiff, but for REINVENT 4 it is unique-valid retention, because that sampler also discards duplicates (see above). $\hat{y}$: mean realized normalized SA score $\pm\,\sigma$ over valid molecules. MAE: mean $|y^* - \hat{y}|$ over valid molecules. $\mathrm{SR}^{\mathrm{val}}_{10}/\mathrm{SR}^{\mathrm{val}}_{20}$: fraction of *valid* molecules within $\varepsilon = 0.10/0.20$ of y^*. $\mathrm{SR}^{\mathrm{att}}_{10}/\mathrm{SR}^{\mathrm{att}}_{20}$: the same numerator divided by n_{att}, so invalid attempts count as failures. The two coincide only at 100% validity. Scaf: unique Bemis–Murcko scaffolds; Scaf/mol = Scaf/n_{val}. Bold: best value per metric within each paradigm group (Paradigm A: oracle-free continuous conditioning; Paradigm B: oracle-in-the-loop / 3-bucket conditioning). ‡REINVENT 4 and DrugDiff use a single fixed y^* per bucket.

Model	Bucket	y^*	n_{att}	n_{val}	Ret.	$\hat{y}$	MAE	$\mathrm{SR}^{\mathrm{val}}_{10}$	$\mathrm{SR}^{\mathrm{att}}_{10}$	$\mathrm{SR}^{\mathrm{val}}_{20}$	$\mathrm{SR}^{\mathrm{att}}_{20}$	Scaf	Scaf/mol
Paradigm A: oracle-free continuous conditioning (1,001 levels × 3 samples; buckets are y^* bins)													
STAR-VAE (ours, $\alpha=12$)	low y^*	0.167	999	995	0.996	0.531	0.366	0.044	0.044	0.143	0.142	**770**	**0.774**
STAR-VAE (ours, $\alpha=12$)	mid y^*	0.500	1002	985	0.983	0.739	0.265	0.179	0.176	0.374	0.367	**691**	0.702
STAR-VAE (ours, $\alpha=12$)	high y^*	0.833	1002	976	**0.974**	**0.845**	**0.148**	**0.424**	**0.413**	**0.722**	**0.704**	**564**	0.578
MolGPT	low y^*	0.167	999	756	0.757	**0.189**	**0.056**	**0.843**	**0.638**	**0.978**	**0.740**	561	0.742
MolGPT	mid y^*	0.500	1002	515	0.514	**0.469**	**0.090**	**0.656**	0.337	**0.915**	0.470	386	**0.750**
MolGPT	high y^*	0.833	1002	244	0.244	0.616	0.204	0.258	0.063	0.566	0.138	183	**0.750**
Paradigm B: oracle-in-the-loop / 3-bucket conditioning (fixed y^* per campaign‡)													
REINVENT 4‡	low y^*	0.167	1001	640	0.639	**0.003**	0.164	0.000	0.000	**1.000**	0.639	61	0.095
REINVENT 4‡	mid y^*	0.500	1001	996	**0.995**	**0.395**	**0.132**	**0.473**	**0.471**	0.764	0.760	312	0.313
REINVENT 4‡	high y^*	0.833	1001	992	**0.991**	**0.974**	0.144	0.129	0.128	**0.998**	**0.989**	**979**	**0.987**
DrugDiff‡	low y^*	0.167	1001	**1001**	**1.000**	0.521	0.356	0.053	0.053	0.191	0.191	509	**0.508**
DrugDiff‡	mid y^*	0.500	1001	**1001**	**1.000**	0.541	0.136	0.465	0.465	0.756	0.756	**551**	0.550
DrugDiff‡	high y^*	0.833	1001	**1001**	**1.000**	0.513	**0.327**	0.103	0.103	0.223	0.223	440	0.440

Reproducibility Details

This appendix lists the training, generation, and evaluation settings recovered from the released code, configuration files, checkpoints, and run logs. We mark settings that could not be recovered instead of estimating them.

Pretraining

Training used Adam, a constant learning rate of 10^{-5}, and a batch size of 256. No learning-rate schedule was used. The KL coefficient increased linearly from 0 to $\beta = 1.0$ during the first 10,000 optimizer steps and remained at 1.0 thereafter. The checkpoint records no changes to these defaults. We exclude fixed sinusoidal positional encodings from parameter counts. The pretrained checkpoint contains 89,175,368 parameters.

Tokenization allows at most 71 SELFIES tokens, or 72 positions including the beginning-of-sequence token. Longer sequences are *truncated*. Truncation can remove the end-of-sequence token and cause the decoded molecule to differ from the input. We did not measure how many training molecules were affected. Future preprocessing should exclude overlength sequences instead.

Conditional adaptation

We adapted each conditional model in two stages. The reported SA-score and docking results use the second-stage checkpoints.

1. **LoRA adaptation.** We train the LoRA adapters, FiLM network, and latent-to-hidden projection. We freeze all other parameters, including the conditional-prior networks and posterior label projection. The fused attention kernel bypasses adapters at the attention output projections, so those adapters receive no gradient. This stage therefore trains the feed-forward adapters, FiLM, and latent-to-hidden projection. About 1.64 M of the 1.94 M parameters reported as trainable receive updates.

2. **Conditioning-pathway refinement.** We freeze all Transformer weights and LoRA adapters. We update the FiLM network, the conditional-prior mean and log-variance networks, the latent-to-hidden projection, and the label projection. This gives 726,016 trainable parameters out of 91,047,752. The stage uses a fixed KL weight of 1.0 and a per-dimension free-bits floor of 0.3 from its first step.

Per-experiment trainable and total parameter counts, conditioning pathways, and prior status are given in the main text.

Generation and decoding

Table S3: Generation settings by experiment. "Greedy" indicates that the next token is the argmax of the (possibly temperature-scaled) logits, so temperature has no effect on the selected token.

Experiment	**Latent source**	**Token selection**	**Seeds**
MOSES random	$\mathcal{N}(\mathbf{0}, \mathbf{I})$	multinomial, T=1.1, top-k 20	3 evaluation seeds
MOSES seed-based	$\mu + \varepsilon\sigma$, scale 1.0	multinomial, T=1.1, top-k 20	3 evaluation seeds
SA / docking sweeps	$p_\theta(\mathbf{z}\|y^*)$	greedy	42
Ablations	$\mathcal{N}(\mathbf{0}, \mathbf{I})$	greedy	42, 0, 101
Rediscovery (VAE)	posterior-local, scale 1.0	multinomial, T=1.1	not recorded
Rediscovery (CVAE)	$\mathcal{N}(\mathbf{0}, \mathbf{I})$ + FiLM	greedy	not recorded

The MOSES seed protocol encodes 30,000 molecules from the official MOSES training split and draws one latent per seed as $\mathbf{z} = \mu + \varepsilon\,\sigma_{\text{posterior}}$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with a multiplier of 1.0. This samples at the full learned posterior scale, not as a small perturbation around μ. The released artifacts do not record the exact perturbation scale used for the rediscovery runs. The code default is 1.0.

The primary ablation and refinement runs used random seed 42. Six SA-score ablation configurations were also trained with seeds 7 and 13, giving three independent fine-tuning runs per configuration. Deterministic cuDNN kernels were not enabled, so a fixed seed does not guarantee bitwise-identical results.

Docking evaluation populations. Each guidance condition starts with 3,003 generation attempts. We count an output as generation-valid when it decodes from SELFIES and RDKit can parse it. We count docking as complete when QVina returns a value other than the sentinel 10,000. We retain real non-negative QVina scores as unfavorable outcomes. Table 4 calculates correlation, calibration slope, valid- or docked-only SR, and IntDiv over the completed-docking population. It separately reports generation validity and docking completion. Table 5 reports attempt-level SR and counts every invalid or failed-docking attempt as a failure. The analysis uses fixed normalization ranges: $s_{\min} = -10.0$ for 1SYH and 4LDE, and -9.4 for 6Y2F. It does not estimate a new range from each generated sample.

Rediscovery table provenance. The 60-row rediscovery table was generated from the combined results using explicit metric names. This keeps CrossSim, InterSim, Div, and Uniq aligned with their definitions.

Statistical reporting

MOSES results use three evaluation seeds. They repeat sampling from one fixed checkpoint and are not independent training runs.

For the replicated SA-score ablations, we report the mean and standard deviation across three independently fine-tuned models. The remaining results use one trained model per condition. GuacaMol, SA-score guidance, docking, and rediscovery therefore report point estimates, with bootstrap or generation-seed uncertainty where stated. Those intervals do not include variation from model fitting, data splitting, pseudo-label estimation, or checkpoint selection.

Pretraining Overlap Audit

The backbone was pretrained on 79 million PubChem molecules. The downstream reference sets come from public sources that may overlap with PubChem. We therefore measured overlap instead of assuming that the

benchmark and held-out structures were absent from the pretraining source corpus. This appendix reports the results for MOSES, GuacaMol, and the cluster-held-out ChEMBL sets.

Procedure

We parsed every molecule in each reference set with RDKit and converted it to a standard InChIKey. We applied the same procedure to all 79,311,161 molecules in the curated PubChem corpus and intersected the resulting key sets. We matched molecules by the full InChIKey and by its first block. The full key includes stereochemistry and isotopic labeling. The first block represents molecular connectivity only. Reporting both separates exact structural overlap from connectivity matches with different stereochemistry or isotopes. The main text uses the full-InChIKey results. The two measures are identical for all four held-out ChEMBL sets.

The ChEMBL reference sets are the cluster-held-out splits used for the rediscovery evaluation, so the audited molecules are exactly those the rediscovery metrics are computed against.

Two qualifications apply to Table S4. First, the audited corpus contains more molecules than the training dataset. The reported percentages are therefore upper bounds on exposure. Second, overlap describes the data, not the model's behavior. It limits how benchmark and rediscovery results can be interpreted, but it does not show that the model memorized a particular generated molecule.

Table S4: Overlap between the PubChem pretraining corpus and the benchmark and evaluation sets. Each reference molecule was canonicalized to a standard InChIKey. "Unique keys" is the number of distinct full InChIKeys after deduplication. These keys were matched against all 79,311,161 curated PubChem molecules. A total of 252 rows (3×10^{-4}%) failed to parse and were excluded. "Exact" requires full InChIKey identity, including stereochemistry. "Connectivity" matches only the first InChIKey block, ignoring stereochemistry and isotopes, and is therefore an upper bound on structural exposure. The two measures agree exactly for all four held-out ChEMBL sets, which carry the leakage argument, and differ by at most 0.15 percentage points elsewhere (GRL fine-tuning), so the overlap reflects structural identity rather than stereochemical variants. Percentages quoted in the main text are the exact-InChIKey values, which are the more conservative of the two. Because the audited corpus is larger than the training dataset, these values bound exposure from above.

Reference set	**Unique keys**	**Exact matches**	**Exact**	**Connectivity**
MOSES train	1,584,399	1,132,095	71.45%	71.45%
MOSES test	176,072	125,331	71.18%	71.19%
MOSES test-scaffolds	176,222	126,042	71.52%	71.53%
GuacaMol train	1,272,168	1,107,308	87.04%	87.10%
ABL1 held-out	359	296	82.45%	82.45%
ABL1 fine-tuning	944	809	85.70%	85.70%
BRD4 held-out	1,944	1,477	75.98%	75.98%
BRD4 fine-tuning	4,475	3,572	79.82%	79.82%
GRL held-out	600	398	66.33%	66.33%
GRL fine-tuning	1,306	1,048	80.25%	80.40%
HIV-1 protease held-out	459	100	21.79%	21.79%
HIV-1 protease fine-tuning	1,137	317	27.88%	27.97%

Consequences for the benchmark results

Between 71% and 72% of MOSES molecules and 87% of GuacaMol training molecules match the curated PubChem source corpus. These values are upper bounds on pretraining exposure. The MOSES and GuacaMol results therefore do not measure transfer to wholly unfamiliar chemical distributions. The model received no parameter updates on these benchmark splits and no benchmark-specific tuning, but pretraining substantially covered their chemical space. We use this narrower interpretation throughout the manuscript.

The overlap does not affect the within-model comparison of random-prior and seed-based sampling because both protocols use the same weights and pretraining corpus. The improvement from seed-based sampling reflects conditioning on a supplied structure. It does not demonstrate generalization to unfamiliar chemistry.

Consequences for the rediscovery results

The cluster-held-out split removes close analogues of the evaluation compounds from target-specific fine-tuning. It does not remove those compounds from the PubChem source corpus. Table S4 shows source-corpus matches for 21.79%–82.45% of held-out compounds, depending on the protein target. These values are upper bounds on pretraining exposure.

This result directly affects the interpretation of exact rediscovery. A recovered held-out structure that matches the source corpus may come from the pretrained distribution. It is not evidence of a previously unseen molecule. Exposure and exact recall follow the same general ordering across the four targets:

Table S5: Held-out pretraining exposure and the highest exact rediscovery recall observed for the pretrained VAE under seed-based sampling across sample sizes from 10K to 50K. The ordering across four protein targets is descriptive. It informs the interpretation of exact rediscovery but does not estimate a memorization effect.

Protein target	**Held-out exposure**	**Max exact recall**
ABL1	82.45%	2.198%
BRD4	75.98%	0.296%
GRL	66.33%	0.650%
HIV-1 protease	21.79%	0.000%

HIV-1 protease has the lowest exposure and no exact rediscoveries at any sample size. ABL1 has the highest exposure and the highest recall. Spearman correlation across the four targets is 0.80. With only four points, this is a descriptive observation and not an estimate of the relationship.

The PubChem molecular-weight filter explains much of the lower exposure for HIV-1 protease. The corpus excludes molecules above 600 Da, so none of those molecules can match. Across all four targets, exposure is 0.00% above 600 Da and 33.4–87.1% at or below 600 Da. The HIV-1 protease held-out set has a median molecular weight of 566 Da, and 35.0% of its compounds exceed 600 Da. The other targets have median molecular weights of 425–473 Da, with 4.7–6.3% of compounds above 600 Da. The curation filter and molecular-weight distribution therefore account for much of the lower HIV-1 protease overlap. Exposure and scaffold class are also confounded across these four targets, so the reported correlation does not estimate a memorization effect.

Scaffold and cluster recovery are less sensitive to exact pretraining overlap. They record whether a generated molecule reaches a Bemis–Murcko framework or a cluster, not whether it reproduces one structure. We therefore use scaffold and cluster recovery as the main rediscovery evidence and treat exact rediscovery as a bounded observation.