ENERGY-GUIDED DIFFUSION FOR VALID SMILES GENERATION

Ivan Gurev & Nikolay Nikitin
ITMO University
Saint-Petersburg, Russia
ivan.gurev@itmo.ru

ABSTRACT

Diffusion models provide a flexible framework for molecular generation, yet their application to SMILES sequences is fundamentally constrained by chemical validity. In continuous diffusion over token embeddings, denoising trajectories often drift off the discrete manifold of valid SMILES, producing syntactic errors, chemical violations, and corrupted stereochemistry after decoding. We analyze these failure modes and show that they stem from a misalignment between smooth probability paths in embedding space and the rule-governed structure of symbolic molecular representations. We frame SMILES validity as an inference-time alignment problem and interpret valid generation as sampling from a tilted distribution that reweights a base diffusion model toward structurally valid regions. Based on this perspective, we introduce validity-aware diffusion mechanisms that combine auxiliary training objectives with energy-based guidance during sampling, steering diffusion trajectories toward the valid SMILES manifold without changing the underlying representation or retraining the base model. Experiments demonstrate substantial improvements in SMILES validity while preserving diversity and novelty, showing that inference-time aligned diffusion can be competitive with autoregressive and masked language models for molecular string generation and suggesting broader applicability to structured symbolic domains such as code and discrete diffusion language models.

1 INTRODUCTION

Diffusion models generate complex data by transporting a simple reference distribution along a continuous probability path (Ho et al., 2020; Song & Ermon, 2019). While originally developed for continuous domains such as images and audio (Saharia et al., 2022; Chung et al., 2022), they have recently been extended to symbolic data, including text (Nie et al., 2025), code (Singh et al., 2023), and molecules (Cheng et al., 2021; Luo et al., 2023). A core challenge in these settings is alignment: enforcing discrete, rule-based validity within a continuous generative process.

Molecular generation exemplifies this difficulty. Chemical space is vast—exceeding 10^{60} drug-like molecules (Polishchuk et al., 2013)—yet valid symbolic representations are exceedingly rare. SMILES strings (Weininger, 1988) are widely used due to their compactness and tooling support, but must satisfy strict syntactic and chemical constraints, including balanced parentheses, ring closures, and valid valences (Schoenmaker et al., 2023).

Score-based diffusion models (Song et al., 2020) for SMILES operate in continuous embedding spaces, enabling stable training but decoupling the diffusion trajectory from the discrete manifold of valid strings. As a result, intermediate states often decode to syntactically or chemically invalid SMILES, yielding low validity and persistent structural errors.

Existing solutions either modify the representation, e.g., using SELFIES (Krenn et al., 2019; Nguyen et al., 2024), or abandon strings in favor of graph- or geometry-based diffusion models (Tao et al., 2025; Weng et al., 2025). These approaches improve validity but reduce compatibility with SMILES workflows or increase modeling complexity. Recent work on diffusion alignment instead treats validity as a constraint or reward imposed on an unconstrained generative process (Tang et al., 2025; Cardei et al., 2025; Schoenmaker et al., 2023; Gong et al., 2024).

We adopt this perspective and frame valid SMILES generation as sampling from a tilted distribution that reweights a base diffusion model toward valid regions of embedding space. We introduce validity-aware diffusion methods that combine auxiliary training objectives with inference-time guidance from learned energy models over SMILES grammar, steering trajectories toward the valid manifold without modifying the representation or retraining the base model.

Our contributions are threefold: (i) an analysis of grammatical and chemical failure modes in SMILES diffusion models; (ii) energy-guided inference method for validity alignment; and (iii) empirical results showing improved validity while preserving diversity and novelty. More broadly, our results demonstrate how diffusion models can be aligned with discrete, rule-governed symbolic domains.

2 Continuous Diffusion on SMILES Sequences

We formulate molecular generation as a diffusion process over continuous embeddings of SMILES token sequences. Let $S = [s_1, \ldots, s_L]$ be a SMILES string with tokens $s_i \in \mathcal{V}$, mapped to embeddings via a learnable embedding function $\mathrm{Emb} : \mathcal{V} \to \mathbb{R}^d$, yielding $z_{\text{pure}} = \mathrm{Emb}(S) \in \mathbb{R}^{L \times d}$. We follow Gong et al. (2024) and sample final embedding z from Gaussian distribution centered at z_{pure}: $z \sim \mathcal{N}(z_{\text{pure}}, \sigma_0 I)$

Following standard score-based diffusion, training is performed by gradually corrupting embeddings with Gaussian noise along a predefined noise schedule. Specifically, samples at time $t \in [0, 1]$ are generated as

$$x_t = \alpha_t z + \beta_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where α_t and β_t control signal and noise magnitudes. A neural network is trained to predict the injected noise, enabling reverse-time sampling from a simple Gaussian prior.

After sampling, the final continuous embeddings $\hat{z}$ are projected back to discrete SMILES tokens using nearest-neighbor decoding in embedding space:

$$\hat{s}_i = \arg\min_{v \in \mathcal{V}} \|\hat{z}_i - \mathrm{Emb}(v)\|_2.$$

This discrete projection creates a fundamental mismatch between smooth diffusion trajectories and the discrete, rule-based SMILES grammar. Understanding and mitigating the resulting validity errors is the focus of the following sections.

3 Validity-Aware Diffusion via Energy-Guided Sampling

We address the SMILES validity problem by treating it as an inference-time alignment task for diffusion models. Rather than modifying the diffusion architecture or replacing the SMILES representation, we steer sampling toward structurally valid regions of embedding space using energy-based guidance. Let $p_\theta(x)$ denote the distribution implicitly defined by a trained diffusion model over continuous SMILES embeddings $x \in \mathbb{R}^{L \times d}$. Only a small subset of these embeddings decode to valid SMILES strings. To bias generation toward this subset, we define a validity-aware target distribution

$$p_{\text{valid}}(x) \sim p_\theta(x) \exp(-E_\phi(x))$$

where $E_\phi(x)$ is a learned energy function that assigns low energy to embeddings corresponding to valid SMILES and high energy to invalid ones.

3.1 Energy Model for SMILES Validity

The energy function $E_\phi(x)$ is implemented using a bidirectional transformer initialized from a SMILES-pretrained BERT model (UniKei, 2025). Continuous diffusion embeddings are projected into the BERT hidden space using a lightweight multi-layer perceptron net (MLP). This projection network is trained jointly with the energy model and maps diffusion embeddings to the input dimensionality expected by BERT, enabling compatibility between continuous representations and discrete-token pretraining. The model outputs a scalar energy representing grammatical and structural consistency.

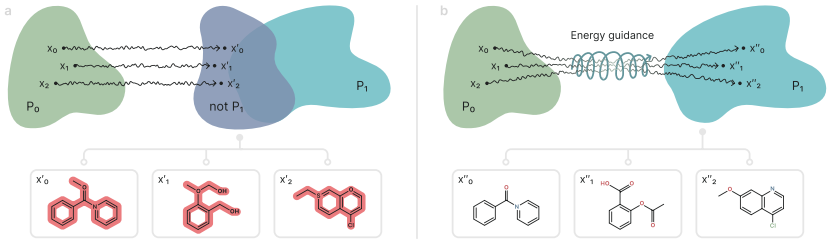

Figure 1: Inference-time alignment of diffusion trajectories for SMILES generation. Continuous diffusion over token embeddings drifts away from the discrete manifold of valid SMILES, producing invalid strings after decoding (left). Energy-guided sampling tilts the diffusion distribution toward low-energy regions corresponding to valid SMILES, improving validity without retraining (right).

To train the energy model, we construct negative examples by stochastically corrupting valid SMILES sequences using grammar-breaking operations, including unbalanced parentheses, invalid ring indices, and token substitutions (see Appendix B, C for details). A contrastive objective encourages lower energy for valid embeddings and higher energy for corrupted ones, enabling the model to capture global syntactic constraints without explicit token-level supervision.

3.2 Energy-Guided Diffusion Sampling

At inference time, validity constraints are incorporated via energy-guided sampling. The standard diffusion score is augmented with the gradient of a learned energy function:

$$\nabla_x \log p_{\text{valid}}(x_t) = \nabla_x \log p_\theta(x_t) - \lambda \nabla_x E_\phi(x_t)$$

Guidance is applied during the later stages (see Appendix E) of the reverse diffusion process, steering samples toward regions corresponding to valid SMILES.

Importantly, this guidance operates entirely at sampling time and does not require retraining the diffusion model, making it a plug-and-play mechanism that can be combined with other forms of conditioning.

4 Experiments

4.1 Dataset

We use the ChEBI-20 dataset (Edwards et al., 2022) despite its molecule–text pairing structure, and restrict our evaluation to unconditional molecular generation. This allows us to isolate the effect of validity guidance independently of conditional signals. Extending energy-guided diffusion to text-conditional generation is a promising direction for future work. Dataset contains 33,010 entries, split into 80/10/10% train/validation/test sets. All experiments adhere to this split. Both models are trained on train set. For training details please refer to Appendix D.

4.2 Results

Without validity guidance, the base diffusion model achieves 0.61 SMILES validity (Table 1), consistent with known limitations of continuous diffusion over symbolic representations. Applying energy-guided sampling improves validity to 0.68, corresponding to a relative gain of over 10% without retraining the model.

Table 1: Comparison of raw and guided variants

Type	Validity ↑	Novelty, ↑, %	Diversity ↑
Unguided	0.61	100	0.89
Guided	0.68	100	0.86

For context, Table 2 reports results from prior methods on ChEBI-20. These approaches are designed for text-conditioned generation, a more complex setting that introduces additional constraints and is not directly comparable to our unconditional setup. We include them as a reference for SMILES generation quality, while our method isolates validity as the primary objective.

Table 2: Text-guided molecule generation results on ChEBI-20 test split.

Model	BLEU ↑	Levenshtein ↓	MACCS FTS ↑	RDK FTS↑	Morgan FTS ↑	Validity ↑
Transformer	0.499	57.660	0.480	0.320	0.217	**0.906**
T5-Base	0.762	24.950	0.731	0.605	0.545	0.660
MolT5-Base	0.769	24.458	0.721	0.588	0.529	0.772
TGM-DLM	**0.826**	**17.003**	**0.854**	**0.739**	**0.688**	0.871

Importantly, the improvement in validity does not come at the cost of novelty or diversity. Generated molecules remain structurally diverse, and common failure modes - such as unbalanced parentheses and incorrect ring closures -are substantially reduced. This indicates that energy guidance primarily corrects global syntactic inconsistencies rather than encouraging trivial or degenerate outputs.

5 Discussion

Our results show that a substantial portion of SMILES invalidity in diffusion models can be mitigated through inference-time alignment alone. The improvement from 61% to 68% validity highlights both the potential and the limits of energy-guided sampling: validity constraints can be imposed without retraining or architectural changes, yet the discrete structure of SMILES still poses challenges for continuous models.

Beyond molecular generation, similar alignment issues arise in domains with strict syntactic or semantic constraints, such as code, mathematical expressions, and discrete diffusion language models. Energy-guided diffusion offers a general inference-time mechanism for enforcing global structural constraints, complementing representation choices and training-time objectives.

6 Conclusion

We examined SMILES validity in continuous diffusion models and framed it as an inference-time alignment problem caused by the mismatch between smooth embedding trajectories and discrete molecular syntax. To address this, we proposed a validity-aware diffusion approach based on energy-guided sampling, interpreting valid generation as sampling from a tilted distribution over a pretrained model.

Experiments show that energy guidance increases validity from 61% to 68% without retraining or loss of diversity. These findings demonstrate that energy-based inference-time alignment is an effective and flexible strategy for controlled generation in structured symbolic domains, with implications for grammar-constrained and discrete diffusion language models.

Acknowledgments

This research is financially supported by the Foundation for National Technology Initiative's Projects Support as a part of the roadmap implementation for the development of the high-tech field of Artificial Intelligence for the period up to 2030 (agreement 70-2021-00187).

The authors express the gratitude to Nina Gubina for her assistance in preparing the illustrations for this paper.

REFERENCES

Michael Cardei, Jacob K Christopher, Bhavya Kailkhura, Thomas Hartvigsen, and Ferdinando Fioretto. Constrained molecular generation with discrete diffusion for drug discovery. In *NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development*, 2025.

Yu Cheng, Yongshun Gong, Yuansheng Liu, Bosheng Song, and Quan Zou. Molecular design in drug discovery: a comprehensive review of deep generative models. *Briefings in bioinformatics*, 22(6):bbab344, 2021.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. Text-guided molecule generation with diffusion language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 109–117, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Mario Krenn, Florian Häse, A Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Selfies: a robust representation of semantically constrained graphs with an example application in chemistry. *arXiv preprint arXiv:1905.13741*, 1(3), 2019.

Tianze Luo, Zhanfeng Mo, and Sinno Jialin Pan. Fast graph generation via spectral diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3496–3508, 2023.

Nguyen Nguyen, Nhat Truong Pham, Duong Tran, and Balachandran Manavalan. Lang2mol-diff: A diffusion-based generative model for language-to-molecule translation leveraging selfies representation. In *Proceedings of the 1st Workshop on Language+ Molecules (L+ M 2024)*, pp. 128–134, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of computer-aided molecular design*, 27(8): 675–679, 2013.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Linde Schoenmaker, Olivier JM Béquignon, Willem Jespers, and Gerard JP van Westen. Uncorrupt smiles: a novel approach to de novo design. *Journal of Cheminformatics*, 15(1):22, 2023.

Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, Carina Negreanu, and Gust Verbruggen. Codefusion: A pre-trained diffusion model for code generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11697–11708, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. Peptune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. *ArXiv*, pp. arXiv–2412, 2025.

Wen Tao, Jing Tang, Alvin Chan, Bryan Hooi, Baolong Bi, Nanyun Peng, Yuansheng Liu, and Yiwei Wang. How to make large language models generate 100% valid molecules? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26576–26591, 2025.

UniKei. BERT-base-smiles: Bidirectional transformer pretrained on smiles, 2025. URL `https://huggingface.co/unikei/bert-base-smiles`.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Wenchao Weng, Hanyu Jiang, Xiangjie Kong, and Giovanni Pau. Text-guided diverse-expression diffusion model for molecule generation. *Chinese Physics B*, 34(5):050701, 2025.

A Computational overhead

Energy-guided sampling introduces additional computational cost, as the energy model must be evaluated at each guided diffusion step. In our implementation, this results in an approximate 1.5–2$\times$ increase in inference time, depending on sequence length and guidance frequency. However, since guidance is only applied in the later stages of sampling, the overhead remains moderate and can be reduced further with smaller energy models or less frequent evaluations.

B Validity Challenges in SMILES Diffusion

To analyze generation failures, we categorize common decoding errors observed in continuous latent models, summarized in Table 3. A large fraction of invalid samples arises from syntactic violations, such as incorrect ring closure digits or unbalanced parentheses. These errors are often minor in embedding space but catastrophic after discrete decoding. Chemical constraint violations, including impossible valencies or incorrect bond orders, are less frequent but result in fundamentally invalid molecules.

Table 3: Common decoding errors in continuous SMILES generation.

Type	Description	Example
Atom substitution	Replacement of one atom by a chemically similar one.	`...cc1Br` → `...cc1Cl`
Ring closure error	Incorrect or missing ring index, breaking ring structure.	`c1nccc2n1` → `c1ncccn1`
Valency violation	Atom assigned an impossible number of bonds.	`C(C)(C)(C)C`
Bond order error	Incorrect single/double/triple bond assignment.	`C=O` → `CO`
Fragment loss	Omission of a substituent or side chain.	`CCCCCO` → `CCCO`
Fragment addition	Insertion of spurious atoms or groups.	`C1CC1` → `C1CCCC1`
Syntax error	Unbalanced parentheses or malformed branches.	`C(C(C)C` → `C(C)C)C`
Chirality corruption	Loss or inversion of stereochemical markers.	`C[C@@H]` → `C[C@H]`

In addition, diffusion models frequently exhibit semantic mutations, such as atom substitutions or fragment loss, which preserve syntactic validity while significantly altering molecular properties. Finally, stereochemical information is particularly fragile: chirality markers are often corrupted or dropped entirely, leading to loss of enantiomeric specificity (Fig. 2).

Overall, these observations highlight a key limitation of continuous diffusion on SMILES: local smoothness in embedding space does not align with the discrete, rule-based structure of molecular strings.

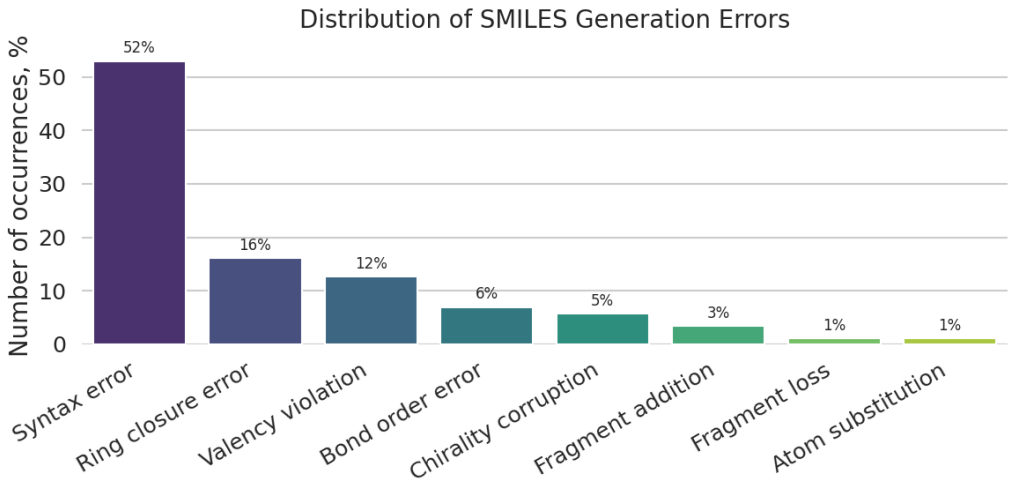

Figure 2: Distribution of error types among generated SMILES.

C Corruption Operators for Invalid Sequence Generation

To train the model to distinguish valid from invalid molecular strings, we apply a set of structured corruption operators that intentionally violate syntactic or chemical constraints of SMILES-like representations. Each operator introduces a specific type of error while preserving most of the original sequence, producing hard negative examples.

Below we describe each corruption mechanism and provide illustrative examples.

C.1 Bond Order Corruption

This corruption introduces an invalid or inconsistent bond specification by replacing two adjacent tokens with bond symbols. Such modifications typically violate local grammar rules, since bond symbols are expected to connect valid atom symbols.

Effect: Breaks local bond–atom consistency.

Example:

$$\text{CCO} \rightarrow \text{C==O}$$

C.2 Parenthesis Flip

All closing parentheses are replaced by opening parentheses, resulting in unmatched branches.

Effect: Violates structural balance of branches.

Example:

$$\text{C(C)O} \rightarrow \text{C((O}$$

C.3 Parenthesis Insertion

An opening or closing parenthesis is inserted at an arbitrary valid position without a corresponding closing or opening parenthesis.

Effect: Creates unbalanced branching.

Example:

$$\text{CCO} \rightarrow \text{C(CO}$$

C.4 Ring Index Corruption

A ring index digit is inserted at two unrelated positions, creating an unmatched or incorrectly paired ring closure.

Effect: Breaks ring consistency rules.

Example:

$$\text{C1CC1} \rightarrow \text{C1CC2}$$

C.5 Transition Corruption

A token is replaced by an arbitrary vocabulary symbol in a position where it violates the expected transition rules of the grammar.

Effect: Introduces illegal token transitions.

Example:

$$\text{CCO} \rightarrow \text{C\#O}$$

C.6 Unbalanced Closing Parenthesis

The last valid token in the sequence is replaced with a closing parenthesis, ensuring that the sequence ends with an unmatched closure.

Effect: Guarantees global syntactic invalidity.

Example:

$$\text{CCO} \rightarrow \text{CC)}$$

C.7 Composite Corruptions

For each sequence, at least one corruption is always applied. With a fixed probability, multiple corruption operators may be combined sequentially, yielding more challenging invalid examples that contain several interacting errors.

Example:

$$\text{C(C)O} \rightarrow \text{C((O2}$$

This mixture of syntactic and chemical violations encourages the model to learn both local and global validity constraints.

D Model Configurations

D.1 Diffusion Transformer

We train a 170M-parameter DiT operating on continuous SMILES token embeddings. Tokens are embedded into a 128-dimensional space and processed with 8 transformer blocks, each with a model dimension of 768, feedforward dimension 3072, and 12 attention heads. Training uses a noise-prediction objective combined with auxiliary reconstruction and cross-entropy losses at low-noise timesteps as in algorithm 1. The optimizer is Adam with a learning rate of 1×10^{-4}, batch size 256, and 1000 epochs. Distributed training with DDP ensures efficiency across GPUs. SMILES strings are decoded via nearest-neighbor projection in embedding space.

The training objective combines three loss terms:

Algorithm 1 Diffusion Training

Require: Batch $\mathcal{B} = \{\mathbf{z}\}$ of SMILES tokens
Require: Model $\mathcal{M}$ with embedding ϕ, denoiser ϵ_θ, and logits head g
Require: Diffusion path with $\alpha(t), \beta(t)$

procedure TRAINSTEP($\mathcal{B}$)
  $\mathbf{z}_{\text{emb}} \leftarrow \phi(\mathbf{z})$ ▷ Embed tokens
  $\mathbf{x}_0 \leftarrow \mathbf{z}_{\text{emb}} + \eta \cdot \mathcal{N}(0, I)$ ▷ Add initial noise
  $t \sim \mathcal{U}[0, 1]$ ▷ Sample diffusion timestep
  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ ▷ Sample Gaussian noise
  $\mathbf{x}_t \leftarrow \alpha(t)\mathbf{x}_0 + \beta(t)\boldsymbol{\epsilon}$ ▷ Forward diffusion
  $\boldsymbol{\epsilon}_\theta \leftarrow \mathcal{M}(\mathbf{x}_t, t)$ ▷ Predict noise
  $\mathcal{L}_{\text{mse}} \leftarrow \|\boldsymbol{\epsilon}_\theta - \boldsymbol{\epsilon}\|^2$ ▷ MSE loss
  $\hat{\mathbf{x}}_0 \leftarrow (\mathbf{x}_t - \beta(t)\boldsymbol{\epsilon}_\theta)/\alpha(t)$ ▷ Denoise
  $\mathcal{L}_{\text{mse0}} \leftarrow \|\hat{\mathbf{x}}_0 - \mathbf{z}_{\text{emb}}\|^2$ ▷ MSE0 loss
  $\mathcal{L}_{\text{ce}} \leftarrow \text{CE}(g(\hat{\mathbf{x}}_0), \mathbf{z})$ ▷ Cross-entropy loss
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{mse}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}}$
  return $\mathcal{L}$ ▷ Backpropagate and update model parameters
end procedure

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{mse0}} \tag{1}$$

$$\mathcal{L}_{\text{mse}} = \mathbb{E}_{t\sim\mathcal{U}[\epsilon,1-\epsilon],\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}\|_2^2\right] \tag{2}$$

$$\mathcal{L}_{\text{mse0}} = \mathbb{E}_{t>0.33}\left[\|\hat{\mathbf{x}}_0(\mathbf{x}_t, t) - \mathbf{z}_{\text{emb}}\|_2^2\right] \tag{3}$$

$$\mathcal{L}_{\text{ce}} = \mathbb{E}_{t>0.45}\left[\text{CE}\left(g(\hat{\mathbf{x}}_0(\mathbf{x}_t, t)), \mathbf{z}\right)\right] \tag{4}$$

D.2 ENERGY MODEL

To guide diffusion toward valid SMILES, we train a bidirectional transformer initialized from UniKei (2025). Continuous embeddings from DiT are projected into BERT's hidden space. Training uses contrastive energy-based objectives (see algorithm 2): positive examples are valid SMILES embeddings, and negatives are corrupted sequences with invalid rings, parentheses, or token substitutions. Loss is the sum of softplus activations over positive and negative examples. Optimization is performed with Adam and a learning rate of 1×10^{-4}.

Algorithm 2 Energy Model Training for SMILES Validity

Require: Batch $\mathcal{B} = \{\mathbf{z}\}$ of SMILES tokens
Require: Energy model $\mathcal{E}$, embedding layer ϕ
Require: Function make_invalid$(\cdot)$ for generating negative examples

procedure TRAINSTEP($\mathcal{B}$)
  $\mathbf{z}_{\text{pos}} \leftarrow \mathbf{z}$ ▷ Positive SMILES
  $\mathbf{z}_{\text{neg}} \leftarrow \text{make\_invalid}(\mathbf{z})$ ▷ Generate corrupted SMILES
  $\mathbf{x}_{\text{pos}} \leftarrow \phi(\mathbf{z}_{\text{pos}})$
  $\mathbf{x}_{\text{neg}} \leftarrow \phi(\mathbf{z}_{\text{neg}})$ ▷ Embed sequences
  $E_{\text{pos}} \leftarrow \mathcal{E}(\mathbf{x}_{\text{pos}})$
  $E_{\text{neg}} \leftarrow \mathcal{E}(\mathbf{x}_{\text{neg}})$ ▷ Compute energy scores
  $\mathcal{L}_{\text{pos}} \leftarrow \text{softplus}(E_{\text{pos}})$
  $\mathcal{L}_{\text{neg}} \leftarrow \text{softplus}(-E_{\text{neg}})$
  $\mathcal{L} \leftarrow \text{mean}(\mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}})$
  return $\mathcal{L}$ ▷ Backpropagate and update model parameters
end procedure

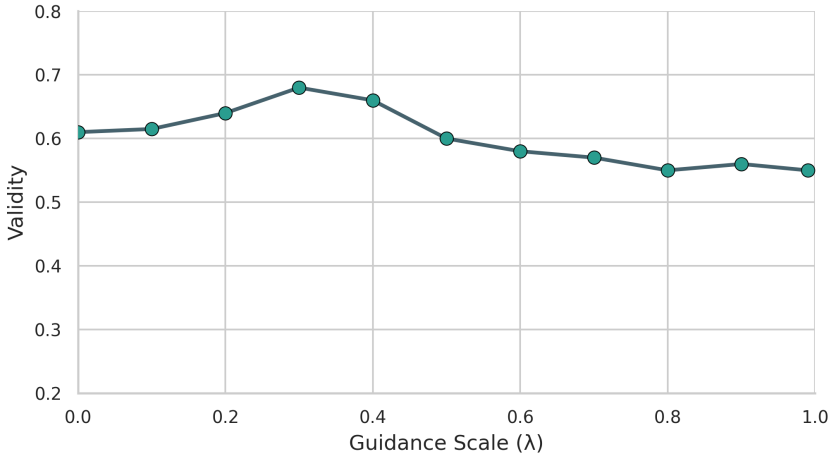

Figure 3: Validity of generated molecules versus guidance scale λ. Validity peaks when λ reaches 0.3-0.4 region.

E Guidance Strength and Scheduling

The guidance strength λ controls the impact of energy model on diffusion trajectory. In practice, we select λ from a small range based on validation performance. Across all experiments, we observe that moderate values of λ_t improve validity, while overly strong guidance can lead to diminishing returns (Fig. 3).

In addition to guidance strength, the point at which guidance is introduced during the reverse diffusion process plays a critical role. We apply guidance only during the later stages of sampling. Empirically, applying guidance too early degrades sample quality, likely because high-noise states lack meaningful structure for the energy model to evaluate. Delaying guidance allows the diffusion process to first form coarse structure before enforcing validity constraints.

To study this effect, we vary the fraction of the trajectory during which guidance is applied (Fig. 4). Specifically, a value of 1 corresponds to no guidance, while 0 indicates guidance applied throughout the entire trajectory.

We observe that introducing guidance too early in the trajectory leads to lower validity, while applying it later yields more stable improvements. This supports our design choice to delay guidance until the later stages of diffusion, where samples have sufficient structure for the energy model to provide meaningful corrections.

F Metrics

Validity. Validity is assessed using RDKit, counting the fraction of generated SMILES strings that are syntactically and chemically valid.

Novelty. Novelty measures the fraction of generated molecules that do not appear in the training dataset. It quantifies the ability of the model to produce new chemical structures rather than memorizing or reproducing known examples. High novelty indicates that the generative model can explore previously unseen regions of chemical space.

Diversity. Diversity measures the structural variety among generated molecules. For each molecule, we compute a Morgan fingerprint (radius 2, 1024 bits) and calculate the pairwise Tanimoto similarity

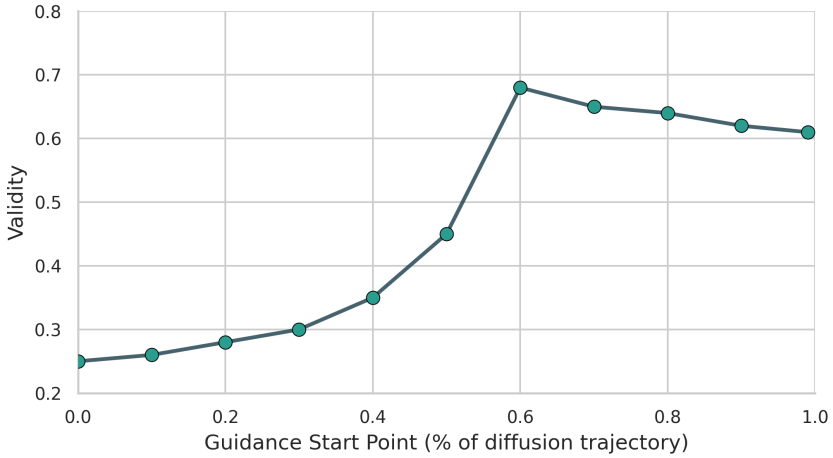

Figure 4: Validity of generated molecules versus guidance start point. Validity peaks when guidance begins after 60% of the diffusion trajectory.

across all generated molecules. Diversity is defined as one minus the average pairwise similarity:

$$\text{Diversity} = 1 - \frac{2}{N(N-1)} \sum_{i<j} T(m_i, m_j)$$

where N is the number of molecules and $T(m_i, m_j)$ is the Tanimoto similarity between molecules i and j. A higher score indicates a more structurally varied set of molecules, reflecting the model's ability to explore diverse regions of chemical space.