International Journal on Advanced Science Engineering Information Technology

Vol.15 (2025) No. 3
ISSN: 2088-5334

Comparative Analysis of Autoregressive and Diffusion-Based Language Models for Complex Molecular Structure Generation

Yihyun Kim [a], Hyeri Yun [b], Jaechoon Jo [c,*]

[a] *Department of Biomedical Informatics, Korea University College of Medicine, Seoul, Republic of Korea*
[b] *Department of Biomedical Informatics, Jeju National University, Jeju, Republic of Korea*
[c] *Department of Computer Education, Jeju National University, Jeju, Republic of Korea*
*Corresponding author: *jjo@jejunu.ac.kr*

Abstract*—Recent advancements in biomedical informatics have opened new avenues for integrating chemical structure data with natural language, enabling innovative approaches in de novo molecular design. In this study, we compare two paradigms for text-guided molecule generation: an autoregressive model, MolT5, built on a T5 framework employing self-supervised pre-training with corrupted span replacement followed by fine-tuning for both molecule captioning and generation, and a diffusion-based model, TGM-DLM, which maps textual descriptions into latent embeddings and iteratively refines SMILES sequences via a denoising process. Evaluated on the ChEBI-20 dataset—partitioned into simple and complex molecular structures—our analysis using metrics such as BLEU, exact match, Levenshtein distance, validity, MACCS, RDK, and Morgan fingerprint similarity reveals that while TGM-DLM exhibits superior performance in capturing the overall architecture of complex molecules, MolT5 achieves higher rates of chemical validity. By leveraging these complementary approaches, our work provides a nuanced assessment of the trade-offs between structural fidelity and chemical correctness in molecular generation. The diffusion-based TGM-DLM model shows particular promise in addressing the complex challenges of intricate molecular configurations, as substantiated by quantitative improvements across multiple evaluation criteria. Conversely, the autoregressive MolT5 model's robustness in preserving chemical integrity underscores its potential for applications where molecular reliability is paramount. These comparative insights not only enhance our understanding of model architectures in multi-modal molecular design but also pave the way for future innovations in computational chemistry and drug discovery.

***Keywords*— Molecular generation; SMILES; autoregressive model; diffusion-based language model.**

Manuscript received 11 Aug. 2024; revised 8 Nov. 2024; accepted 14 Jan. 2025. Date of publication 30 Jun. 2025.

I. INTRODUCTION

The design and generation of novel molecular entities remain pivotal challenges in drug discovery and materials science. Traditional computational chemistry methods, such as structure-based and quantum chemical approaches, have long been employed to predict molecular properties; however, these methods are often limited by the vastness of chemical space and the intricacies of complex molecular architectures [1]. To overcome these challenges, recent years have witnessed a surge of deep learning–based strategies aimed at navigating and exploiting chemical space by leveraging large-scale data.

The evolution of Transformer-based models and diffusion-based generative models has significantly shaped recent advancements in deep learning. Transformer-based architectures were first introduced in the landmark work *Attention Is All You Need* by Vaswani et al. [1]. By employing a self-attention mechanism, Transformers overcame the sequential processing limitations inherent in recurrent neural networks (RNNs), thereby enabling highly parallelized training and improved efficiency in modeling long-range dependencies. This innovation catalyzed the development of large-scale pretrained language models such as BERT and GPT, which have revolutionized natural language processing (NLP) [2]. Moreover, the Transformer framework has been successfully extended beyond text processing into domains such as computer vision (e.g., Vision Transformer) [3], demonstrating its versatility across data modalities.

In parallel, diffusion-based generative models have emerged as a powerful alternative to traditional generative paradigms such as GANs and VAEs. Diffusion models operate by gradually introducing noise into the data and learning to reverse this process, thereby reconstructing high-quality samples from pure noise [4]. Although early diffusion models suffered from computational inefficiencies and slow

sampling speeds, subsequent advancements, particularly the development of Denoising Diffusion Probabilistic Models (DDPMs) [5], have markedly improved training stability and output fidelity. The incorporation of classifier guidance and text-conditional generation techniques has further enabled diffusion models to surpass previous generative approaches, especially in image synthesis and conditional text-to-image tasks [6].

Beyond these recent developments, a rich body of work has explored alternative deep generative approaches for molecular design. Early studies employed Variational Autoencoders (VAEs) to learn continuous latent representations from discrete molecular inputs such as SMILES strings or molecular graphs. VAEs facilitate latent-space interpolation and property optimization, as demonstrated by Gómez-Bombarelli et al [7]. However, they often require extensive data and careful tuning to ensure that decoded molecules adhere to chemical validity. Extensions such as Junction-Tree VAEs have been proposed to capture the underlying chemical structure better. Meanwhile, Generative Adversarial Networks (GANs) have also been applied to molecule generation. For instance, models like MolGAN generate molecular graphs by pitting a generator against a discriminator, thereby learning to produce structures that mimic the distribution of real molecules. Despite their potential, GAN-based methods often face issues such as mode collapse and training instability. In addition, autoregressive language models that generate SMILES sequences token by token have demonstrated robust performance in preserving local syntactic correctness, albeit sometimes at the expense of global structural fidelity [8]. Other approaches—including reinforcement learning–driven models and evolutionary algorithms—further contribute to this diverse landscape, each balancing trade-offs between novelty, validity, and property optimization.

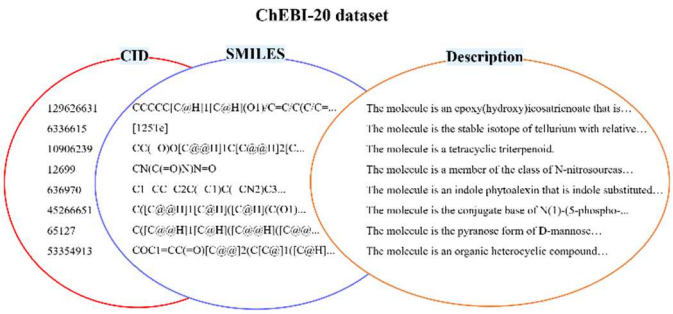


Fig. 1 Example of ChEBI Dataset

Within this context, our study presents a comprehensive comparison of two state-of-the-art paradigms for text-guided molecular generation: the autoregressive MolT5 and the diffusion-based TGM-DLM. Utilizing the ChEBI-20 dataset (e.g., Fig. 1), which is stratified into "Global" (complex) and "Local" (simple) subsets based on chemical criteria such as molecular weight (>500), ring count (>3), and hydrogen bonding capacity (>5) [9], we rigorously evaluate model performance via metrics including BLEU scores, exact match ratios, Levenshtein distance, chemical validity, and fingerprint similarity measures. By integrating our experimental findings with established literature [10]–[13], this work elucidates the trade-offs between preserving local syntactic accuracy and achieving global structural fidelity [14], and it lays the groundwork for future hybrid models that can combine the strengths of diverse generative approaches.

II. Material and Method

A. Dataset and Preprocessing

We employ the ChEBI-20 dataset, which comprises 33,010 paired instances of SMILES strings and their corresponding natural language descriptions. The dataset is divided into training, validation, and test sets in an 80:10:10 ratio. Following established protocols [15], [16], molecules are classified into "simple" and "complex" subsets. Specifically, molecules with a molecular weight greater than 500, more than three ring structures, or more than five hydrogen bond donors are designated as "complex" (hereafter "Global"), while the remaining molecules are considered "simple" (hereafter "Local"). This stratification is performed using RDKit to compute molecular descriptors, ensuring reproducibility and methodological rigor [17]–[19].

SMILES strings are first standardized to mitigate syntactic variability and subsequently tokenized at the character level using a specialized SMILES tokenizer. The tokenizer treats each atom, bond symbol, ring indicator, bracket, and other special characters as distinct tokens. Domain-specific vocabulary enhancements (e.g., distinguishing "Sc" from an "S" followed by "c") are incorporated to reduce tokenization errors. Additional preprocessing steps include noise filtering and alignment between the chemical and text modalities to facilitate joint representation learning.

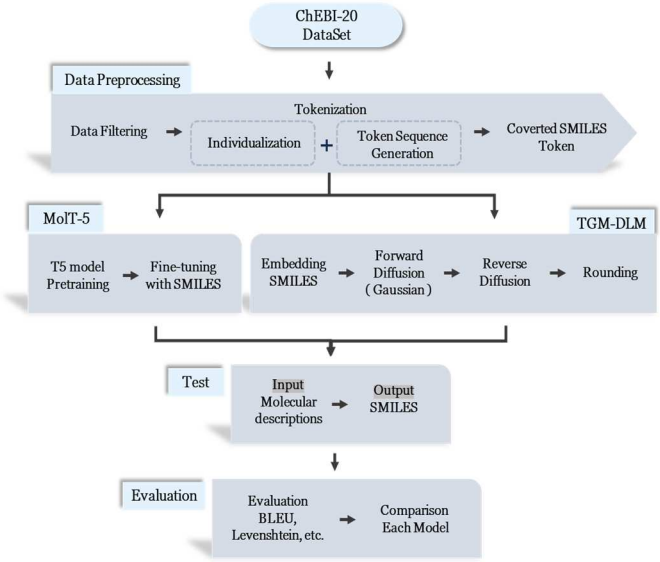


Fig. 2 Study Design Framework

B. Autoregressive Model: MolT5

MolT5 is implemented upon the T5 (Text-to-Text Transfer Transformer) framework and adapted to handle both natural language and SMILES representations. Its training pipeline is divided into two main stages—pre-training and fine-tuning—and each stage is supported by specialized tokenization methods and a decoding strategy designed to capture the nuances of chemical notation (e.g., Fig. 3).

1) Text Tokenizer: For the natural language component, MolT5 leverages a Hugging Face pretrained tokenizer

initialized with SciBERT (i.e., allenai/scibert_scivocab_uncased) to better capture domain-specific terminology [20]. SciBERT's tokenizer is based on the WordPiece (or SentencePiece) subword approach, which splits words into smaller subword units to handle out-of-vocabulary terms and morphological variations effectively. This strategy is particularly advantageous for biomedical text, as it allows the model to capture partial matches of rare or compound terms often found in chemical and biomedical descriptions. During pre-training, the tokenizer is used in conjunction with a "replace corrupted spans" objective: contiguous segments of both natural language text and SMILES are masked and replaced by sentinel tokens. By predicting these masked spans, MolT5 learns a joint embedding space encompassing both linguistic semantics and chemical syntax. This cross-modal representation is critical for subsequent tasks such as molecule captioning and molecule generation [21], [22].

2) SMILES Tokenizer: While the text tokenizer handles domain-specific biomedical vocabulary, SMILES tokenization in MolT5 typically involves a specialized scheme at the character or sub-character level. For example:

- Atoms and Atom Groups: Each element symbol (e.g., "C," "Cl," "Br") is treated as a token. Special care is taken for cases like "Sc" (scandium) to avoid confusion with "S" followed by "c."
- Bonds and Ring Closures: Characters for double bonds ("="), ring indices ("1," "2," etc.), and branching parentheses are each considered distinct tokens.
- Bracketed Notations: Square brackets (e.g., [NH3+]) or stereochemical indicators are also split into separate tokens or sub-tokens to ensure precise reconstruction.

This level of granularity helps the autoregressive decoder predict each symbol in a strictly ordered manner, reducing syntactic errors in the final SMILES output. During fine-tuning, a dedicated tokenizer script resolves these SMILES-specific ambiguities [23], ensuring that tokens representing similar chemical entities are handled consistently.

3) Training Process: During pre-training, MolT5 learns cross-modal embeddings by applying a "replace corrupted spans" objective to both text and SMILES sequences. Random contiguous segments are masked and replaced with sentinel tokens, and the model must predict these missing spans. This approach is similar to multilingual pre-training methods like mBERT [14], allowing MolT5 to jointly encode linguistic semantics and chemical syntax [21], [22]. After this unsupervised phase, fine-tuning is conducted on a curated subset of the ChEBI-20 dataset [24]. Fine-tuning focuses on two key tasks: (1) molecule generation, where the input is a textual description and the output is the corresponding SMILES, and (2) molecule captioning, which reverses the process by translating SMILES into descriptive text. By training on both directions, MolT5 strengthens its ability to bridge language and chemical notation.

4) Decoding Method: MolT5 uses a beam search algorithm at inference time to generate SMILES from textual prompts. In beam search, multiple candidate sequences (beams) are maintained at each decoding step. Each beam is expanded by possible next tokens, and only the highest-scoring beams—according to the model's probability estimates—are retained. This process continues until an end-of-sequence token is reached or a maximum length is met. Although more computationally demanding than simple greedy decoding, beam search generally yields more chemically valid and semantically coherent SMILES because it explores a broader set of candidate sequences. As a result, MolT5 achieves a balance between preserving local syntactic correctness (through its autoregressive architecture) and leveraging domain-specific knowledge acquired during pre-training with SciBERT-initialized parameters.

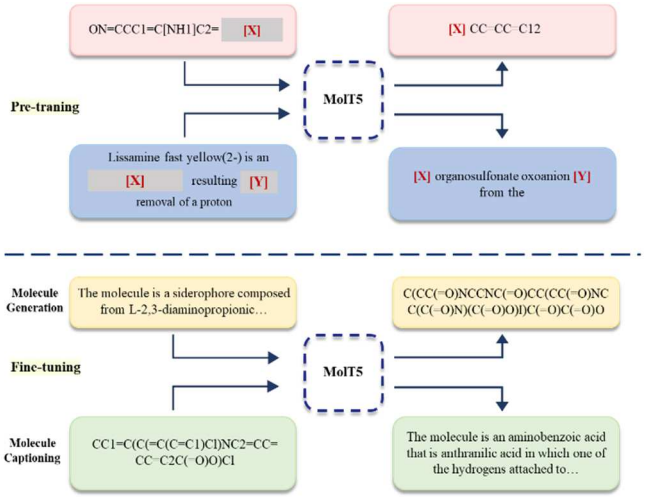


Fig. 3 Design flow chart for applying MolT5.

C. Diffusion-Based Model: TGM-DLM

TGM-DLM employs a diffusion-based strategy for text-guided molecular generation, and its approach can be broadly divided into two sequential phases: text-guided generation and iterative denoising correction. This structure is designed to capture both the global molecular architecture and the local syntactic details essential for valid SMILES representations (e.g., Fig. 4).

1) Text-Guided Generation: In the first phase, a pretrained language model transforms the input textual description into a continuous latent embedding space. This latent space is enriched by cross-attention mechanisms that inject the textual context into every layer, ensuring that critical semantic cues are well represented. From this continuous space, the model generates an initial SMILES sequence. The tokenizer used here is tightly integrated with the diffusion framework, with a detailed vocabulary specifically designed to handle chemical nuances such as ring closures, branching parentheses, bracketed notations, and charge states. This ensures that when the continuous latent vectors are mapped back to discrete tokens, key chemical features are preserved.

2) Iterative Denoising Correction: Recognizing that the initial SMILES output may exhibit minor syntactic discrepancies—such as unbalanced brackets or misaligned ring closures—a second, iterative denoising process is applied. In this phase, Gaussian noise is incrementally introduced into the latent embedding, and then systematically removed over multiple diffusion steps. At each step, the model refines the latent representation to better adhere to global molecular

constraints. A final rounding step maps the continuous latent vectors back to discrete SMILES tokens by finding the closest match in the pre-defined embedding vocabulary (using L_2 distance). This two-stage reverse diffusion process, conceptually analogous to methods in Diffusion-LM and Diffuseq [3], ensures that the final output is both globally coherent and chemically valid. Further details on the denoising framework, including parameter settings for the noise schedule (β_t) and the cumulative product formulation for α_t, are provided in the Supplementary Information.

In summary, while MolT5's autoregressive beam search excels at preserving chemical validity step by step, TGM-DLM's diffusion-based iterative refinement is particularly adept at capturing the global molecular structure, especially for more complex molecules. Together, these approaches represent complementary paradigms in text-guided molecular generation, each offering unique advantages depending on the desired balance between local syntactic accuracy and global structural fidelity.

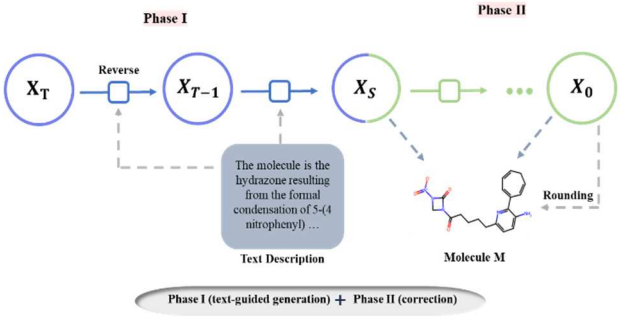


Fig. 4 Design flow chart for applying TGM-DLM.

D. Evaluation Metrics

The performance of both models is rigorously quantified using a suite of evaluation metrics as shown in Table 1 and 2:

TABLE I
COMPLEX STRUCTURE DATASET (1454) RESULT

Model	MolT-5	TGM-DLM
BLEU	0.440	0.821
Exact	0.0	0.162
Levenshtein	69.802	28.765
Validity	0.935	0.714
MACC	0.658	0.904
RDK	0.488	0.826
Morgan	0.317	0.758

TABLE II
SIMPLE STRUCTURE DATASET (1846) RESULT

Model	MolT-5	TGM-DLM
BLEU	0.395	0.802
Exact	0.001	0.211
Levenshtein	30.095	9.148
Validity	0.928	0.882
MACC	0.409	0.799
RDK	0.203	0.646
Morgan	0.164	0.611

1) BLEU Score: This metric assesses n-gram overlap between the generated and reference SMILES strings. SMILES sequences are tokenized to extract n-grams, and precision scores are computed with an additional brevity penalty to account for overly short sequences [25].

2) Exact Match Ratio: The proportion of generated SMILES strings that exactly match the reference molecules is computed, providing a direct measure of generation accuracy.

3) Levenshtein Distance: The minimum number of edit operations (insertions, deletions, substitutions) required to transform the generated SMILES string into the reference string is calculated. This metric indicates the overall sequence similarity.

4) Chemical Validity: The fraction of generated SMILES strings that correspond to chemically plausible molecules is determined using standard cheminformatics toolkits (e.g., RDKit) [26].

5) Fingerprint Similarity Metrics: Structural similarity between generated and reference molecules is quantified using multiple fingerprint representations, including MACCS keys (166-bit vectors), RDKit fingerprints (2048-bit vectors), and Morgan (ECFP) fingerprints. These metrics assess the overlap of molecular substructures, providing a robust measure of chemical similarity [27], [28].

III. RESULTS AND DISCUSSION

A. Experimental Results

Our evaluation on the ChEBI-20 dataset—divided into "Global" (complex) and "Local" (simple) subsets based on molecular weight (>500), ring count (>3), and hydrogen bonding capacity (>5)—reveals distinct performance profiles for the two models. For the complex (Global) subset, the diffusion-based TGM-DLM model achieves substantially higher BLEU scores and superior fingerprint similarity (evaluated using MACCS, RDKit, and Morgan fingerprints) compared to the autoregressive MolT5. This result indicates that TGM-DLM's iterative denoising framework is especially effective at capturing long-range dependencies and reconstructing the intricate global architecture of complex molecules [28], [29] (See Table 3). However, MolT5 consistently attains higher chemical validity scores, suggesting that its sequential, token-by-token generation method is better at preserving local syntactic correctness and ensuring that the resulting SMILES strings correspond to chemically plausible structures [30].

For simple (Local) molecules, the performance gap between the two models narrows, though the observed trends persist (See Table 4). Quantitatively, TGM-DLM's ability to reconstruct overall structural features is reflected in improved global similarity metrics, even as its iterative refinement process can introduce minor syntactic discrepancies. In contrast, MolT5's autoregressive approach robustly preserves chemical validity, albeit at the potential cost of slight deviations in global structural fidelity. Representative qualitative examples and a detailed quantitative summary substantiate these observations, highlighting the complementary strengths and limitations of each modeling approach.

TABLE III
EXAMPLES OF GENERATED MOLECULES FROM BOTH MODELS COMPARED WITH GROUND-TRUTH STRUCTURES (COMPLEX)

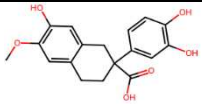
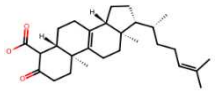
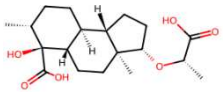
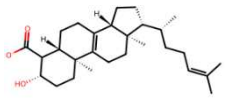
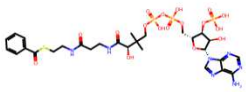
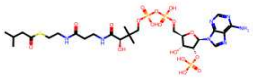
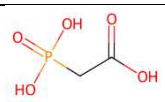
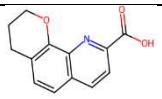
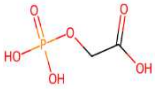
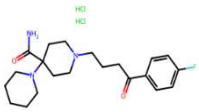
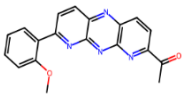
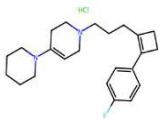
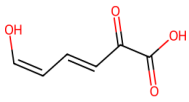
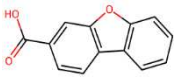
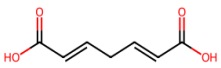
No.	Input	Ground Truth	MolT5	TGM-DLM
1	The molecule is a kanamycin obtained by dehydrogenation at position 2' of kanamycin A. It derives from a kanamycin A. It is a conjugate base of a 2'-oxokanamycin(4+).			
2	The molecule is a steroid acid anion that is the conjugate base of 3-dehydro-4-carboxyzymosterol, obtained by deprotonation of the carboxy group; major species at pH 7.3. It is a conjugate base of a 3-dehydro-4-carboxyzymosterol.			
3	The molecule is a methylbutanoyl-CoA, is the S-isovaleryl derivative of coenzyme A. It has a role as a mouse metabolite. It derives from isovaleric acid and butyryl-CoA. It is a conjugate acid of an isovaleryl-CoA(4-).			

TABLE IV
EXAMPLES OF GENERATED MOLECULES FROM BOTH MODELS COMPARED WITH GROUND-TRUTH STRUCTURES (SIMPLE)

No	Input	Ground Truth	MolT5	TGM-DLM
1	The molecule is a member of the class of phosphonic acids, which is phosphonic acid in which the hydrogen attached to the phosphorus is replaced by a carboxymethyl group. It has a role as an antiviral agent and an EC 2.7.7.7 …			
2	The molecule is a hydrochloride resulting from the reaction of pipamperone with 2 mol eq. of hydrogen chloride. It is used as an antipsychotic. It has a role as a dopaminergic antagonist, a first-generation antipsychotic, and a serotonergic antagonist			
3	The molecule is a muconic semialdehyde having a hydroxy substituent at the 2-position. It is a muconic semialdehyde and an alpha, beta-unsaturated monocarboxylic acid. It is a conjugate acid of a 2-hydroxy-6-oxohexa-2,4-dienoate.			

B. Discussion

Our results underscore a critical trade-off inherent in current text-guided molecular generation methods. The superior global performance of the diffusion-based TGM-DLM model is consistent with recent studies that highlight the efficacy of diffusion processes in capturing long-range dependencies in complex, high-dimensional data [9], [25], [26]. Similar benefits have been reported in other domains, such as image and audio synthesis, where diffusion-based strategies enable the reconstruction of intricate global patterns [10], [31].

Conversely, the autoregressive MolT5 model demonstrates higher chemical validity, a finding that corroborates previous research on sequential generation methods where the stepwise prediction mechanism minimizes local errors and reinforces syntactic integrity [8], [22]. This local accuracy is particularly crucial in molecular design, where minor deviations in token sequences can result in invalid chemical structures.

Bridging these findings with the existing literature, our study reinforces the notion that model architecture is crucial in striking a balance between global structural fidelity and local syntactic accuracy. Recent reviews have advocated for hybrid approaches that integrate the strengths of both diffusion-based and autoregressive models [3], [21], [24]. Our comparative analysis provides empirical support for such strategies: while TGM-DLM excels in reconstructing the overall molecular framework, its iterative denoising process may sometimes compromise chemical validity. This shortcoming could be mitigated by incorporating autoregressive elements.

Furthermore, the comprehensive suite of evaluation metrics employed—ranging from sequence-level BLEU scores and Levenshtein distances to chemical validity and multi-fingerprint similarity analyses—affords a nuanced understanding of model performance. These metrics, refined in recent studies [27], [28], [29], are indispensable for capturing the multifaceted nature of molecular generation tasks. They ensure that assessments reflect both the holistic structural coherence and the granular chemical plausibility of generated molecules.

In summary, our results highlight that while diffusion-based methods, such as TGM-DLM, are adept at capturing global structural features, autoregressive models, like MolT5, excel in maintaining chemical correctness. This complementary performance suggests that future research should explore hybrid architectures that combine the iterative refinement capabilities of diffusion models with the local precision of autoregressive approaches, thereby advancing the frontier of multi-modal molecular design.

IV. CONCLUSION

In this study, we conducted a detailed comparative analysis of two state-of-the-art language models for text-guided molecular structure generation—the autoregressive MolT5

and the diffusion-based TGM-DLM—using the ChEBI-20 dataset. Our experiments demonstrate that TGM-DLM excels at capturing global structural nuances of complex molecules, as evidenced by its superior BLEU and fingerprint similarity scores, while MolT5 consistently achieves higher chemical validity through its sequential token generation approach. These complementary strengths underscore a fundamental trade-off between global structural coherence and local syntactic accuracy.

Our findings suggest that neither model, when used in isolation, fully satisfies the multifaceted requirements of de novo molecular design. Instead, the integration of diffusion-based global modeling with the robust local error-correction capabilities of autoregressive models may represent a promising direction for future research. Further development of such hybrid architectures, along with the refinement of evaluation metrics that capture both molecular validity and novelty, is essential for advancing multi-modal molecular design.

The convergence of NLP techniques and cheminformatics, as evidenced by our study, is poised to accelerate innovations in drug discovery and materials science. As these interdisciplinary fields continue to merge, we anticipate that increasingly sophisticated models will emerge, capable of generating molecules that not only meet stringent chemical criteria but also push the boundaries of design in complex chemical spaces.

ACKNOWLEDGMENT

This research is supported by the 2025 scientific promotion program funded by Jeju National University

REFERENCES

[1] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 6000-6010.

[2] H. Gong et al., "Text-guided molecule generation with diffusion language model," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 1, 2024, pp. 1234-1242, doi: 10.1609/aaai.v38i1.27761.

[3] X. Li et al., "Diffusion-LM improves controllable text generation," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 4328-4343.

[4] C. Edwards et al., "Translation between molecules and natural language," in *Proc. EMNLP*, 2022, pp. 375-413, doi:10.18653/v1/2022.emnlp-main.26.

[5] C. Edwards et al., "Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries," in *Proc. EMNLP*, 2021, pp. 595-607, doi:10.18653/v1/2021.emnlp-main.47.

[6] S. Liu et al., "Multi-modal molecule structure-text model for text-based retrieval and editing," *Nat. Mach. Intell.*, vol. 5, pp. 1447-1457, 2023, doi: 10.1038/s42256-023-00759-6.

[7] R. Gómez-Bombarelli et al., "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Cent. Sci.*, vol. 4, no. 2, pp. 268-276, Feb. 2018, doi: 10.1021/acscentsci.7b00572.

[8] S. Gao et al., "G-MATT: Single-step Retrosynthesis Prediction using Molecular Grammar Tree Transformer," *AIChE J.*, vol. 70, no. 2, 2024, doi: 10.1002/aic.18244.

[9] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Adv. Drug Deliv. Rev.*, vol. 46, no. 1-3, pp. 3-26, 2001, doi: 10.1016/S0169-409X(96)00423-1.

[10] A. Alakhdar, B. Poczos, and N. Washburn, "Diffusion Models in De Novo Drug Design," *J. Chem. Inf. Model.*, vol. 64, no. 19, pp. 7238-7256, Oct. 2024, doi: 10.1021/acs.jcim.4c01107.

[11] Y. C. Lo et al., "Machine learning in chemoinformatics and drug discovery," *Drug Discov. Today*, vol. 23, no. 8, pp. 1538-1546, Aug. 2018, doi: 10.1016/j.drudis.2018.05.010.

[12] Z. Wu et al., "Exploring the trade-offs: Unified large language models vs. local fine-tuned models for highly-specific radiology NLI task," *IEEE Trans. Big Data*, vol. 11, no. 3, pp. 1027-1041, Jun. 2025, doi: 10.1109/tbdata.2025.3536928.

[13] R. Goyal, P. Kumar, and V. P. Singh, "A systematic survey on automated text generation tools and techniques: Application, evaluation, and challenges," *Multimedia Tools Appl.*, vol. 82, no. 28, pp. 43089-43144, Nov. 2023, doi: 10.1007/s11042-023-15224-0.

[14] M. Sako, N. Yasuo, and M. Sekijima, "DiffInt: A Diffusion Model for Structure-Based Drug Design with Explicit Hydrogen Bond Interaction Guidance," *J. Chem. Inf. Model.*, vol. 65, no. 1, pp. 71-82, Jan. 2025, doi: 10.1021/acs.jcim.4c01385.

[15] B. Zagidullin et al., "Comparative analysis of molecular fingerprints in prediction of drug combination effects," *Brief. Bioinform.*, vol. 22, no. 6, 2021, doi: 10.1093/bib/bbab291.

[16] N. Brown et al., "GuacaMol: Benchmarking Models for de Novo Molecular Design," *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1096-1108, Mar. 2019, doi: 10.1021/acs.jcim.8b00839.

[17] G. Sliwoski et al., "Computational methods in drug discovery," *Pharmacol. Rev.*, vol. 66, no. 1, pp. 334-395, Jan. 2014, doi: 10.1124/pr.112.007336.

[18] M. Vogt, "Exploring chemical space - Generative models and their evaluation," *Artif. Intell. Life Sci.*, vol. 3, Dec. 2023, doi:10.1016/j.ailsci.2023.100064.

[19] J. Sieg et al., "MolPipeline: A Python package for processing molecules with RDKit in Scikit-learn," *J. Chem. Inf. Model.*, vol. 64, no. 24, pp. 9027-9033, Dec. 2024, doi: 10.1021/acs.jcim.4c00863.

[20] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, 2019, doi: 10.18653/v1/D19-1371.

[21] Y. Li et al., "Generative Models for Molecular Design," *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 5635-5636, Dec. 2020, doi:10.1021/acs.jcim.0c01388.

[22] S. Ishida et al., "Large language models open new way of AI-assisted molecule design for chemists," *J. Cheminform.*, vol. 17, no. 1, Mar. 2025, doi: 10.1186/s13321-025-00984-8.

[23] H. H. Loeffler et al., "Reinvent 4: Modern AI-driven generative molecule design," *J. Cheminform.*, vol. 16, no. 1, Feb. 2024, doi:10.1186/s13321-024-00812-5.

[24] N. Lee et al., "Vision language model is NOT all you need: Augmentation strategies for molecule language models," in *Proc. CIKM*, 2024, pp. 1153-1162, doi: 10.1145/3627673.3679607.

[25] U. V. Ucak, I. Ashyrmamatov, and J. Lee, "Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization," *J. Cheminform.*, vol. 15, no. 1, May 2023, doi:10.1186/s13321-023-00725-9.

[26] M. A. Skinnider, "Invalid SMILES are beneficial rather than detrimental to chemical language models," *Nat. Mach. Intell.*, vol. 6, no. 4, pp. 437-448, Apr. 2024, doi: 10.1038/s42256-024-00821-x.

[27] H. Safizadeh et al., "Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical-Genetic Interactions," *J. Chem. Inf. Model.*, vol. 61, no. 9, pp. 4156-4172, Sep. 2021, doi: 10.1021/acs.jcim.0c00993.

[28] D. Boldini et al., "Effectiveness of molecular fingerprints for exploring the chemical space of natural products," *J. Cheminform.*, vol. 16, no. 1, Mar. 2024, doi: 10.1186/s13321-024-00830-3.

[29] C. Bilodeau et al., "Generative models for molecular discovery: Recent advances and challenges," *WIREs Comput. Mol. Sci.*, vol. 12, no. 5, 2022, doi: 10.1002/wcms.1608.

[30] J. Deng et al., "Artificial intelligence in drug discovery: Applications and techniques," *Brief. Bioinform.*, vol. 23, no. 1, 2021, doi:10.1093/bib/bbab430.

[31] H. Chen et al., "Comprehensive exploration of diffusion models in image generation: A survey," *Artif. Intell. Rev.*, vol. 58, no. 4, Jan. 2025, doi: 10.1007/s10462-025-11110-3.