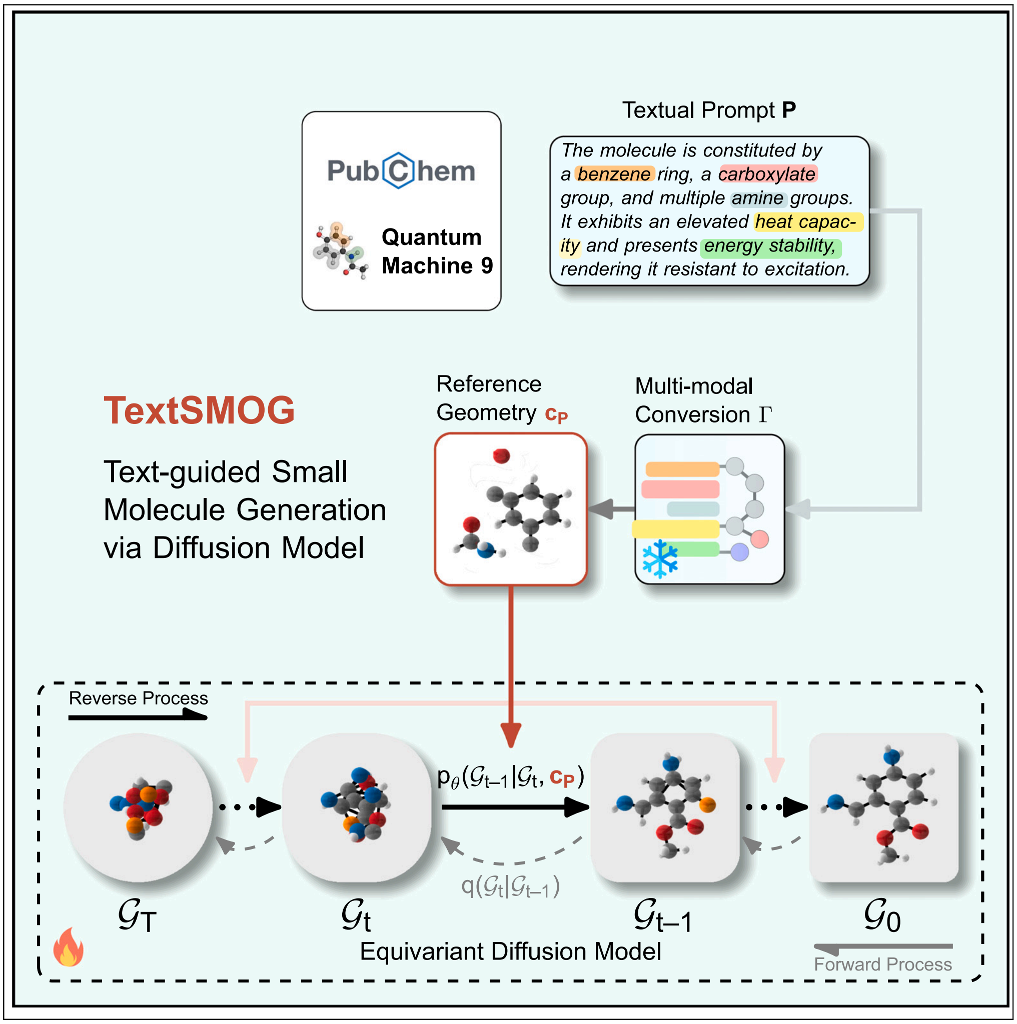
iScience

CellPress
OPEN ACCESS

Article

Text-guided small molecule generation via diffusion model


Textual Prompt P
PubChem
Quantum Machine 9
TextSMOG
Text-guided Small Molecule Generation via Diffusion Model
Reference Geometry c_P
Multi-modal Conversion Γ
Reverse Process
$p_\theta(\mathcal{G}_{t-1}|\mathcal{G}_t, c_P)$
$q(\mathcal{G}_t|\mathcal{G}_{t-1})$
$\mathcal{G}_T$
$\mathcal{G}_t$
$\mathcal{G}_{t-1}$
$\mathcal{G}_0$
Equivariant Diffusion Model
Forward Process

The molecule is constituted by a benzene ring, a carboxylate group, and multiple amine groups. It exhibits an elevated heat capacity and presents energy stability, rendering it resistant to excitation.

Yanchen Luo, Junfeng Fang, Sihang Li, ..., An Zhang, Wenjie Du, Xiang Wang

duwenjie@mail.ustc.edu.cn (W.D.)
xiangwang1223@gmail.com (X.W.)

Highlights

Combines language models and diffusion models into text-guided 3D molecule generation

Multi-modal conversion module translates text conditions into reference geometries

Outperforms existing methods, yielding more stable and diverse subject molecules

Initializes understanding general text descriptions to explore molecular space

Luo et al., iScience 27, 110992
November 15, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110992

iScience

CellPress
OPEN ACCESS

Article

Text-guided small molecule generation via diffusion model

Yanchen Luo,[1] Junfeng Fang,[1,3] Sihang Li,[1] Zhiyuan Liu,[2] Jiancan Wu,[1] An Zhang,[2] Wenjie Du,[1,4,*] and Xiang Wang[1,*]

SUMMARY

The *de novo* generation of molecules with targeted properties is crucial in biology, chemistry, and drug discovery. Current generative models are limited to using single property values as conditions, struggling with complex customizations described in detailed human language. To address this, we propose the text guidance instead, and introduce TextSMOG, a new *text-guided small molecule generation approach* via *diffusion model*, which integrates language and diffusion models for text-guided small molecule generation. This method uses textual conditions to guide molecule generation, enhancing both stability and diversity. Experimental results show TextSMOG's proficiency in capturing and utilizing information from textual descriptions, making it a powerful tool for generating 3D molecular structures in response to complex textual customizations.

INTRODUCTION

De novo molecule design, the process of generating molecules with specific, chemically viable structures for target properties, is a cornerstone in the fields of biology, chemistry, and drug discovery.[1–4] It not only allows for the creation of subject molecules but also provides insights into the relationship between molecular structure and function, enabling the prediction and manipulation of biological activity. Constrained by the immense diversity of chemical space, manually generating property-specific molecules remains a daunting challenge.[5] However, the generation of molecules that precisely meet specific requirements, including the creation of tailor-made molecules, is a complex task due to the vastness of the chemical space and the intricate relationship between molecular structure and function. Overcoming this challenge is crucial for advancing our understanding of biological systems and for the development of new therapeutic agents. In recent years, machine and deep learning methods have initiated a paradigm shift in the molecule generation,[6–11] which enable the direct design of 3D molecular geometric structures with the desired properties.[8,12,13] Notably, diffusion models,[14,15] specifically equivariant diffusion models,[16,17] have gradually entered the center of the stage with its outstanding performance. The core of this method is to introduce diffusion noise on molecular data, and then learn a reverse process in either *unconditional* or *conditional* manners to denoise this corruption, thereby crafting desired 3D molecular geometries. Meanwhile, some conditional inputs (e.g., polarizability $\alpha = 100\text{Bohr}^3$) could be applied for constraining the model to generate more specific molecules types.

However, despite the promise of these methods, a significant proportion of molecules generated by diffusion models do not meet the practical needs of researchers. For instance, they may lack the desired biological activity, exhibit poor pharmacokinetic properties, or be synthetically infeasible. This would be due to the fact that, on one hand, searching for suitable molecules in drug design typically requires consideration of multiple properties of interest (e.g., simultaneously characterized by specific polarizability, orbital energy, properties like aromaticity, and distinct functional groups).[18–20] On the other hand, humans seem to struggle with conveying their needs precisely to the model. While a text segment such as "this molecule is an aromatic compound, with small HOMO-LUMO gaps and possessing at least one carboxyl group" can accurately describe human requirements and facilitate communication among humans, it is still challenging to directly convey this "thoughts" to the model. Therefore, we aspire to develop a method that allows for the interactive inverse design of 3D molecular structures through natural language. In other words, we aim to create a system where researchers can describe the properties they want in a molecule using natural language, and the system will generate a molecule that meets these requirements. This aspiration prompts us to explore text guidance in diffusion models, emphasizing the necessity for models adept at precise language understanding and molecule generation.

Toward this end, we propose TextSMOG, a new text-guided small molecule generation approach. The basic idea is to combine the capabilities of the advanced language models[21–26] with high-fidelity diffusion models, enabling a sophisticated understanding of textual prompts and accurate translation into 3D molecular structures. TextSMOG accomplishes this through integrating textual information with a conversion module that conditions a pre-trained equivariant diffusion model (EDM)[16] following the multi-modal fusion fashion.[9,27–29]

[1]University of Science and Technology of China, Hefei, Anhui, China
[2]National University of Singapore, Singapore, Singapore
[3]These authors contributed equally
[4]Lead contact
*Correspondence: duwenjie@mail.ustc.edu.cn (W.D.), xiangwang1223@gmail.com (X.W.)
https://doi.org/10.1016/j.isci.2024.110992

Specifically, at each denoising step, TextSMOG first generates reference geometry, an intermediate conformation that encapsulates the textual condition signal, through a multi-modal conversion module. Equipped with language and molecular encoder-decoder, corresponding to the textual condition. Then the reference geometry guides the denoising of each atom within the pre-trained unconditional EDM, gradually modifying the molecular geometry to match the condition while maintaining chemical validity. By incorporating valuable language knowledge into the pre-trained diffusion model, TextSMOG enhances the generation of valid and stable 3D molecular conformations that align with a spectrum of diverse directives. This is achieved without the need for exhaustive training on each specific condition, demonstrating the model's ability to generalize from the language input. This integration allows for the incorporation of valuable language knowledge in the high-fidelity pre-trained diffusion model, thereby enabling the conditional generation contingent upon a spectrum of diverse directives, while enhancing the generation of valid and stable 3D molecular conformations, without specific exhaustive training of the condition.

We applied TextSMOG to the standard quantum chemistry dataset QM9[30] and a real-world text-molecule dataset from PubChem.[31] The experimental results show that TextSMOG accurately captures single or multiple desired properties from textual descriptions, thereby aligning the generated molecules with the desired structures. Notably, TextSMOG outperforms leading diffusion-based molecule generation baselines (e.g., EDM,[16] EEGSDE[17]) in terms of both the stability and diversity of the generated molecules. This is evidenced by higher scores on metrics such as the Tanimoto similarity to the target structure, the synthetic accessibility of the generated molecules, and the diversity of the generated molecule set. Furthermore, when applied to real-world textual excerpts, TextSMOG demonstrates its generative capability under general textual conditions. These findings suggest that TextSMOG constitutes a versatile and efficient text-guided molecular diffusion framework. As an advanced intelligent agent, it can effectively comprehend the meaning of textual commands and accomplish generation tasks, thereby paving the way for a more in-depth exploration of the molecular space.

RESULTS

In this section, we present the architecture and the experimental results of our proposed TextSMOG model, showcasing its ability to generate molecules with desired properties.

Architecture

To evaluate our model, we employ the QM9 dataset,[30] which is a standard benchmark containing quantum properties and atom coordinates of over 130K molecules, each with up to 9 heavy atoms (C, N, O, F). For the purpose of training our model under the condition of textual descriptions, we have curated a subset of molecules from QM9 and associated them with real-world descriptions. These descriptions are sourced from PubChem,[31] one of the most comprehensive databases for molecular descriptions, and are linked to the molecules in QM9 based on their unique SMILES.

PubChem aggregates extensive annotations from a diverse array of sources, such as ChEBI,[32] LOTUS,[33] and T3DB.[34] Each of these sources offers an emphasis on the physical, chemical, or structural attributes of molecules. Additionally, we have employed a set of textual templates to generate corresponding descriptions based on the quantum properties of the molecules, thereby enriching the content of the dataset and supplementing textual context for those molecules lacking real-world descriptions. This process has enriched QM9 into a dataset of chemical molecule-textual description pairs. Our proposed TextSMOG model, illustrated in Figure 1, is built upon the pre-trained unconditional diffusion model EDM.[16] It integrates the textual information into the conditional signal of diffusion models by employing a reference geometry that is updated at each step based on the textual prompt. The final molecular geometry is generated by gradually denoising an initial geometry, while noise is added at each step during the forward process until the molecular geometry is fully noise-corrupted.

Experiment on single quantum properties conditioning

Following EDM,[16] we first evaluate our TextSMOG on the task of generating molecule conditioning on a single desired quantum property in QM9. Then we compare our TextSMOG with several baselines to demonstrate the effectiveness of our model on single quantum properties conditioning molecule generation.

Setup

We follow the same data preprocessing and partitions as in EDM,[16] which results in 100K/18K/13K molecule samples for training/validation/test, respectively. In order to assess the quality of the conditional generated molecules *w.r.t.* to the desired properties, we use the property classifier network ϕ_p introduced by.[35] Then for the impartiality, the training partition is further split into two non-overlapping halves $\mathbb{D}_a$ and $\mathbb{D}_b$ of 50K molecule samples each. The property classifier network ϕ_p is trained on the first half $\mathbb{D}_a$, while our TextSMOG is trained on the second half $\mathbb{D}_b$. This ensures that there is no information leak and the property classifier network ϕ_p is not biased toward the generated molecules from TextSMOG. Then ϕ_p is evaluated on the generated molecule samples from TextSMOG as we introduce in the following.

Metrics

Following,[16] we use the mean absolute error (MAE) between the properties of generated molecules and the ground truth as a metric to evaluate how the generated molecules align with the condition (see the supplemental information for details). We generate 10K molecule samples for the evaluation of ϕ_p, following the same protocol as in EDM. Additionally, we then measure novelty,[36] atom stability,[16] and molecule stability[16] to demonstrate the fundamental molecule generation capacity of the model (also see the supplemental information for details).

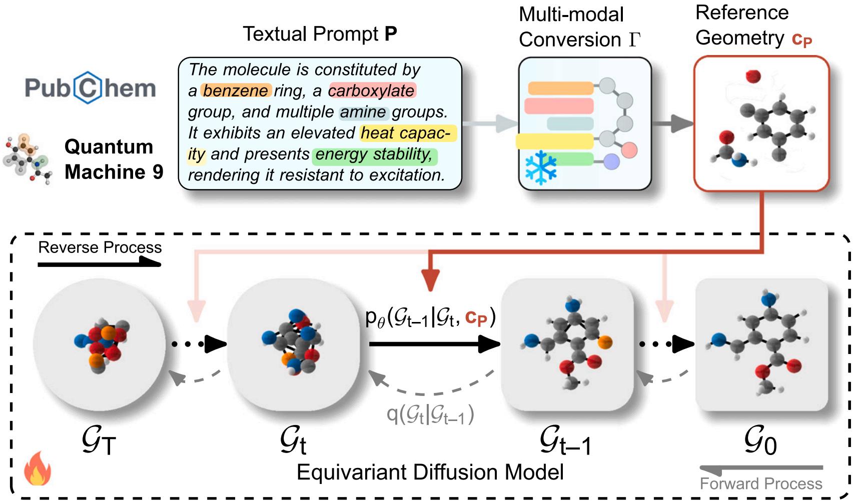

Textual Prompt P
PubChem
Quantum Machine 9
The molecule is constituted by a benzene ring, a carboxylate group, and multiple amine groups. It exhibits an elevated heat capacity and presents energy stability, rendering it resistant to excitation.
Multi-modal Conversion Γ
Reference Geometry $\mathbf{c_P}$
Reverse Process
$p_\theta(\mathcal{G}_{t-1}|\mathcal{G}_t, \mathbf{c_P})$
$q(\mathcal{G}_t|\mathcal{G}_{t-1})$
$\mathcal{G}_T$
$\mathcal{G}_t$
$\mathcal{G}_{t-1}$
$\mathcal{G}_0$
Equivariant Diffusion Model
Forward Process

Figure 1. Architecture of our text-guided small molecule generation via diffusion model (TextSMOG)

The model starts with an initial geometry ($\mathcal{G}_T$) and gradually denoises it to generate the final molecular geometry. The reference geometry ($\mathbf{c_P}$), updated at each step based on the textual prompt (**P**), is employed to integrate the textual information into the conditional signal of diffusion models. Flame 🔥 denotes tunable modules, while snowflake ❄ indicates frozen modules.

Baseline

We compare our TextSMOG with a direct baseline conditional EDM[16] and a recent work EEGSDE which takes energy as guidance.[17] We also compare two additional baselines "U-bound" and "#Atoms" introduced by.[16] In the "U-bound" baseline, any relation between molecule and property is ignored, and the property classifier network ϕ_p is evaluated on $\mathbb{D}_b$ with shuffled property labels. In the "#Atoms" baseline, the properties are predicted solely based on the number of atoms in the molecule. Furthermore, we report the error of ϕ_p on $\mathbb{D}_b$ as a lower bound baseline "L-Bound".

Results

We generate molecules with textual descriptions targeted to each one of the six properties in QM9, which are detailed in the supplemental information. As presented in Figure 2, our TextSMOG has a lower MAE than other baselines on five out of the six properties, suggesting that the molecules generated by TextSMOG align more closely with the desired properties than other baselines. The result underscores the proficiency of TextSMOG in exploiting textual data to guide the conditional *de novo* generation of molecules. Moreover, it highlights the superior congruence of the text-guided molecule generation via the diffusion model with the desired property, thus showing significant potential. Furthermore, as indicated in Figure 3, our proposed TextSMOG exhibits commendable performance in terms of novelty and stability. The text guidance we introduced has transformed the exploration of the model in the molecule generation space, generally enhancing the novelty of the generated molecules while maintaining their stability.

Experiment on multiple quantum properties conditioning

The capacity to generate molecules, guided by multiple conditions, is a crucial aspect of the molecule generation model. When guided by textual descriptions, characterizing the condition with multiple desired properties is highly intuitive and flexible. Following the same setup and metrics in the previous section, we evaluate our TextSMOG on the task of generating molecules with multiple desired quantum properties in QM9. Then we compare TextSMOG with two baselines to showcase the effectiveness of our model in generating molecules conditioned on multiple quantum properties.

As shown in Table 1, our TextSMOG has a remarkably lower MAE than the other two baselines, thereby demonstrating the superiority of our model in generating molecules with multiple desired properties. This also further substantiates that, without necessitating additional targeted interventions, textual conditions can be utilized in our model to guide molecule generation that conforms to multiple desired properties.

Additionally, as highlighted in Table 2, our proposed TextSMOG maintains superior performance in terms of novelty and stability, when generating molecules targeted at multiple desired properties. The results indicate that the flexible integration of multiple conditions through textual description does not compromise the stability of the generated molecules. Furthermore, this approach enhances novelty when compared to the baseline.

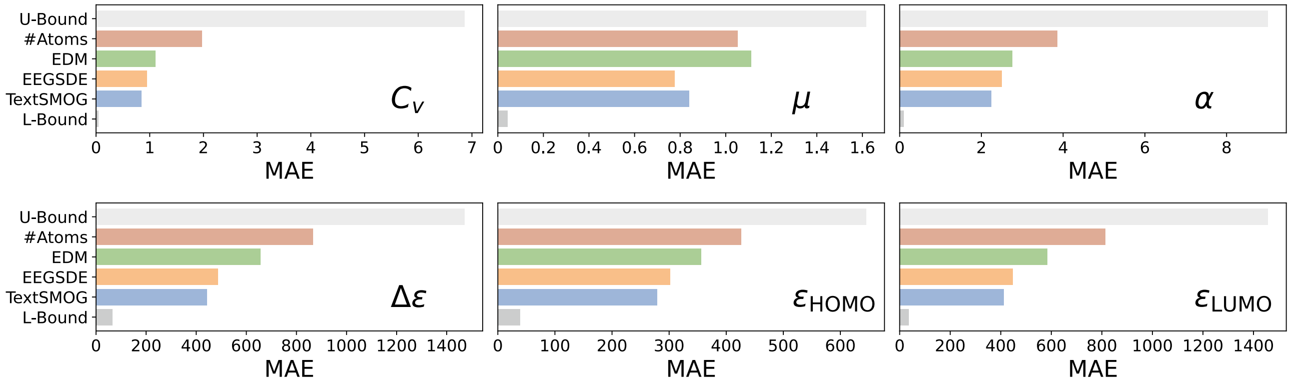


Figure 2. Comparison of MAE for the generated molecules targeted to desired property

Statistics of baselines are from their original papers. The performance of EEGSDE varies depending on the scaling factor, and we report its best results. The numerical values are provided in the supplemental information.

Generation on general textual descriptions

To further assess our model, we undertake additional training on a vast dataset of over 330K text-molecule pairs we gleaned from PubChem.[31] Then, we generate molecules based on general textual descriptions to observe the capacity of our model to generate from generalized textual conditions.

Visual observations, as depicted in Figure 4, illuminate the impressive aptitude of our TextSMOG in aligning molecule structures with the desired property within the textual descriptions. For instance, when the textual description includes affirmatively mentioned terms such as "simple chain structure", "at least one carboxyl group", and "soluble in water", the generated molecules consistently exhibit chain structures with at least one carboxyl group, and characteristics indicative of water solubility.

Moreover, when the textual description includes "polycyclic heteroarene" and specifies the solubility and heat capacity of the molecule, TextSMOG generates a variety of polycyclic aromatic hydrocarbon molecules. The ubiquitously present amino and nitro groups attest to a certain degree of solubility of the molecules. Referring to structurally similar molecules, their expected specific heat capacity is also relatively low.

Lastly, when the text description explicitly demands multiple nitrogen atoms and a low energy gap, the molecules generated by TextSMOG not only possess the required polycyclic structure and multiple nitrogen atoms, but also the rings on the same plane denote the low-energy structures of these molecules that are difficult to excite.

The remarkable alignment between the conditions and the generated molecule stands as a testament to the exceptional generative capabilities of TextSMOG. The result demonstrates that TextSMOG is equipped to deeply explore the chemical molecular space in a text-guided manner, thereby generating prospective molecules for subsequent applications. This capability could potentially expedite drug design and the discovery of materials.

The results highlight TextSMOG's versatility in generating a wide variety of molecular structures, from simple chain structures to complex polycyclic compounds, under the guidance of general text descriptions. This underscores the model's potential to perform well even when the conditions deviate significantly from the distribution of the training set.

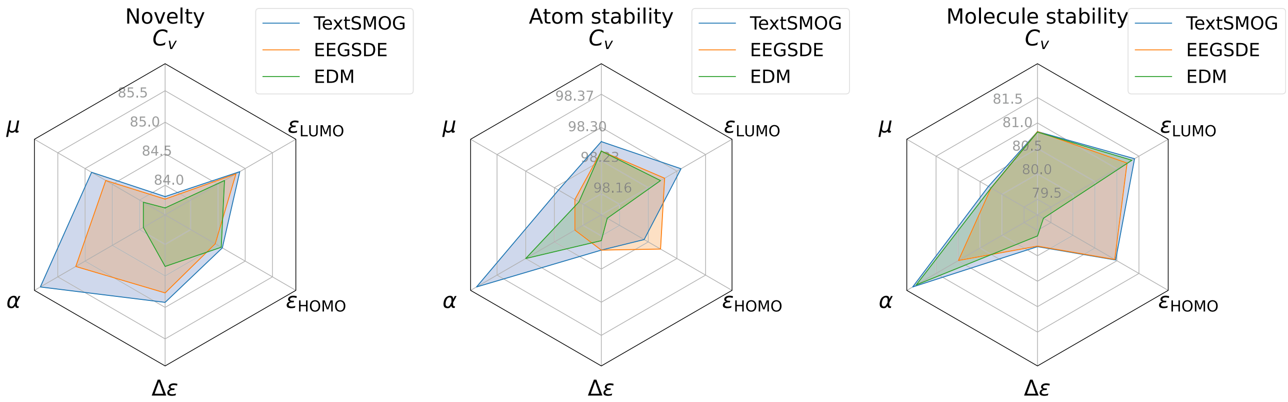


Figure 3. Comparison of novelty (Novel, %), atom stability (A. Stable,%), and molecule stability (M. Stable,%) on generated molecules targeted to the desired property

Statistics of baselines are from EEGSDE. The performance of EEGSDE varies depending on the scaling factor, and we report its best results.

Table 1. Comparison of MAE on the generated molecules targeted to the multiple desired properties

Method	MAE1 ↓	MAE2 ↓
Condition	C_v ($\frac{\text{cal}}{\text{mol}}$K) and μ (D)	
EDM	1.079 ±0.007	1.156 ±0.011
EEGSDE	0.981 ±0.008	0.912 ±0.006
TextSMOG	**0.645** ±0.014	**0.836** ±0.017
Condition	α (Bohr^3) and μ (D)	
EDM	2.76 ±0.01	1.158 ±0.002
EEGSDE	2.61 ±0.01	0.855 ±0.007
TextSMOG	**2.27** ±0.01	**0.809** ±0.010
Condition	$\Delta\varepsilon$ (meV) and μ (D)	
EDM	683 ±1	1.130 ±0.007
EEGSDE	563 ±3	0.866 ±0.003
TextSMOG	**489** ±4	**0.843** ±0.009

Statistics of baselines are from EEGSDE. Boldface indicates the best performance.

DISCUSSION

The translational impacts of TextSMOG are particularly significant for the field of drug discovery and materials science. By enabling the generation of molecular structures directly from textual descriptions, TextSMOG can streamline the early stages of drug design where rapid prototyping and iterative testing are crucial. This approach can facilitate the discovery of subject drug candidates by allowing researchers to quickly generate and evaluate provided molecules based on specific desired properties mentioned in literature or derived from expert knowledge.

Furthermore, TextSMOG can aid in the development of materials with tailored properties by generating molecules that meet specific criteria. This capability is valuable in industries such as polymers, nanomaterials, and catalysts, where precise molecular structures can significantly influence material performance.

In drug discovery, TextSMOG's ability to generate molecules that align with complex textual prompts can accelerate the identification of compounds with potential therapeutic effects. This is especially relevant for targeting diseases with well-characterized biochemical pathways, where detailed descriptions of molecular interactions and desired properties are available. By generating candidate molecules that meet these criteria, TextSMOG can help narrow down the pool of potential drugs, reducing the time and cost associated with experimental validation.

Additionally, TextSMOG's flexibility in handling diverse textual inputs can facilitate interdisciplinary research, where insights from different fields can be integrated into the molecule generation process. For instance, combining insights from biology, chemistry, and pharmacology can lead to more informed and effective drug design strategies.

Table 2. Comparison of novelty (Novel, %), atom stability (A. Stable, %), and molecule stability (M. Stable,%) on the generated molecules targeted to the multiple desired properties

Method	Novel ↑	A. Stable ↑	M. Stable ↑
Condition	C_v($\frac{\text{cal}}{\text{mol}}$K) and μ (D)		
EDM	85.31 ±0.43	**98.00** ±0.07	**77.42** ±0.80
EEGSDE	85.62 ±0.86	97.67 ±0.08	74.56 ±0.54
TextSMOG	**85.79** ±0.66	97.89 ±0.10	77.33 ±0.72
Condition	α (Bohr^3) and μ (D)		
EDM	85.06 ±0.27	97.96 ±0.00	75.95 ±0.30
EEGSDE	85.56 ±0.56	97.61 ±0.04	72.72 ±0.27
TextSMOG	**85.64** ±0.64	**98.01** ±0.07	**75.97** ±0.44
Condition	$\Delta\varepsilon$ (meV) and μ (D)		
EDM	85.18 ±0.35	98.00 ±0.06	77.96 ±0.33
EEGSDE	85.36 ±0.03	97.99 ±0.06	77.77 ±0.26
TextSMOG	**85.44** ±0.41	**98.06** ±0.04	**78.03** ±0.29

Statistics of baselines are from EEGSDE. Boldface indicates the best performance.

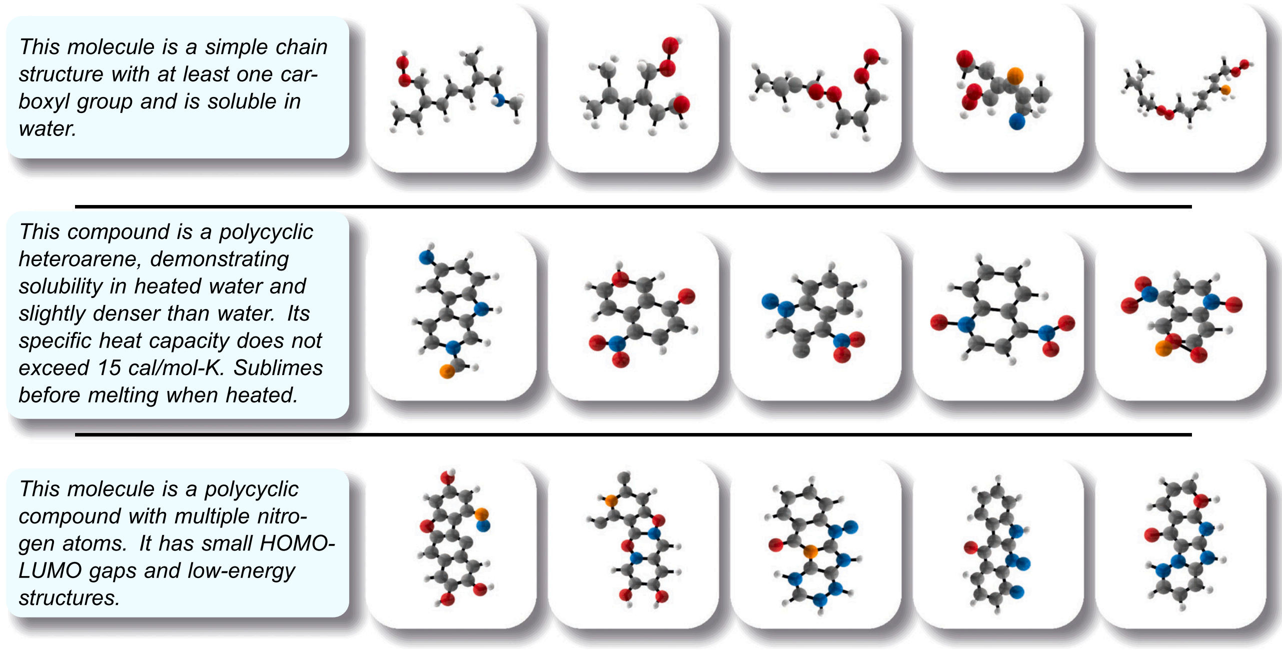

This molecule is a simple chain structure with at least one carboxyl group and is soluble in water.
This compound is a polycyclic heteroarene, demonstrating solubility in heated water and slightly denser than water. Its specific heat capacity does not exceed 15 cal/mol-K. Sublimes before melting when heated.
This molecule is a polycyclic compound with multiple nitrogen atoms. It has small HOMO-LUMO gaps and low-energy structures.

Figure 4. Generated molecules targeted to text description excerpts

Despite the complexity of translating textual prompts into accurate molecular structures, we have successfully integrated advanced language models with high-fidelity diffusion models in TextSMOG, a text-guided diffusion approach for 3D molecule generation. Our experiments on the QM9 and PubChem datasets demonstrate the superior performance of TextSMOG over leading baselines, affirming its efficacy in capturing desired properties from textual descriptions and generating corresponding valid molecules.

Limitations of the study

The integration of textual information with the denoising process of a pre-trained EDM allows TextSMOG to generate valid and stable molecular conformations that closely align with diverse textual directives. This initial success paves the way for significant advancements in the exploration of chemical space and the development of compounds. Nevertheless, our findings are not without limitations.

Our work was constrained by the scarcity of high-quality data linking real-world 3D molecules to their corresponding textual descriptions. This limitation impacted our ability to fully train the model on a diverse set of text-3D molecule pairs, potentially affecting the accuracy of the generated molecules in generating molecules that accurately align with complex textual descriptions. Moreover, the relative slowness of the sampling process due to the iterative nature of the total diffusion steps can pose a challenge in scenarios requiring rapid molecule generation, such as high-throughput drug discovery or material design.

In addition to these limitations, the current design of TextSMOG necessitates that the properties to condition on must be known upfront during the training phase. This might not always be feasible in practical settings, where specific properties linked to a particular drug discovery target may only become available later on, and often with very limited sample data. The generalization of TextSMOG to more complex and real-world scenarios also needs further exploration.

Looking ahead, we are optimistic about the potential of text-guided 3D molecule generation to revolutionize drug discovery and related fields. Future work will focus on overcoming these challenges by expanding and enhancing the quality of datasets linking textual descriptions to molecular structures, improving the efficiency of the sampling process, and making TextSMOG more adaptable to real-world applications. Addressing these limitations will not only enhance the performance of TextSMOG but also contribute significantly to the advancement of text-guided molecule generation technology.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Wenjie Du (duwenjie@mail.ustc.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The datasets generated during this study are available at HuggingFace and are publicly available as of the date of publication. The url is listed in the key resources table.
- All original code has been deposited at GitHub and is publicly available as of the date of publication. The url is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This research is supported by the National Natural Science Foundation of China (92270114). This research was also supported by the advanced computing resources provided by the Supercomputing Center of the USTC.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.L. and J.F.; methodology, Y.L. and S.L.; investigation, Z.L. and S.L.; writing—original draft, Y.L. and J.F.; writing—review and editing, J.W., A.Z., and W.D.; funding acquisition, X.W.; resources, J.W., A.Z., and W.D.; supervision, X.W.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - ○ Notation and background
  - ○ Equivariant diffusion model for molecule generation
  - ○ Integrating textual prompts into 3D molecular reference geometry
  - ○ Conditioning with the reference of text guidance
  - ○ Training objective
  - ○ Evaluation metrics
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ The quantum properties in QM9 dataset

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110992.

SUPPORTING CITATIONS

The following references appear in the Supplemental Information:[37–58].

Received: March 8, 2024
Revised: June 23, 2024
Accepted: September 16, 2024
Published: September 19, 2024

REFERENCES

1. Hajduk, P.J., and Greer, J. (2007). A decade of fragment-based drug design: strategic advances and lessons learned. Nat. Rev. Drug Discov. *6*, 211–219. https://doi.org/10.1038/nrd2220.
2. Mandal, S., Moudgil, M., and Mandal, S.K. (2009). Rational drug design. Eur. J. Pharmacol. *625*, 90–100. https://doi.org/10.1016/j.ejphar.2009.06.065.
3. Pyzer-Knapp, E.O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., and Aspuru-Guzik, A. (2015). What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. Annu. Rev. Mater. Res. *45*, 195–216. https://doi.org/10.1146/annurev-matsci-070214-020823.
4. Barakat, K.H., Houghton, M., Tyrrel, D.L., and Tuszynski, J.A. (2014). Rational Drug Design: One Target, Many Paths to It. Int. J. Comput. Model Algorithm. Med. *4*, 59–85. https://doi.org/10.4018/ijcmam.2014010104.
5. Gaudelet, T., Day, B., Jamasb, A.R., Soman, J., Regep, C., Liu, G., Hayter, J.B.R., Vickers, R., Roberts, C., Tang, J., et al. (2021). Utilizing graph machine learning within drug discovery and development. Briefings Bioinf. *22*, bbab159. https://doi.org/10.1093/bib/bbab159.
6. Alcalde, M., Ferrer, M., Plou, F.J., and Ballesteros, A. (2006). Environmental biocatalysis: from remediation with enzymes to novel green processes. Trends Biotechnol. *24*, 281–287. https://doi.org/10.1016/j.tibtech.2006.04.002.
7. Anand, N., Eguchi, R., Mathews, I.I., Perez, C.P., Derry, A., Altman, R.B., and Huang, P.-S. (2022). Protein sequence design with a learned potential. Nat. Commun. *13*, 746. https://doi.org/10.1038/s41467-022-28313-9.
8. Mansimov, E., Mahmood, O., Kang, S., and Cho, K. (2019). Molecular geometry prediction using a deep generative graph neural network. Sci. Rep. *9*, 20381. https://doi.org/10.1038/s41598-019-56773-5.
9. Zang, C., and Wang, F. (2020). MoFlow: An Invertible Flow Model for Generating Molecular Graphs. In KDD (ACM), pp. 617–626. https://doi.org/10.1145/3394486.3403104.
10. Satorras, V.G., Hoogeboom, E., Fuchs, F., Posner, I., and Welling, M. (2021). E(n) Equivariant Normalizing Flows. In NeurIPS, pp. 4181–4192.
11. Gebauer, N.W.A., Gastegger, M., and Schütt, K. (2019). Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In NeurIPS, pp. 7564–7576. https://doi.org/10.48550/arXiv.1906.00957.
12. Huang, L., Zhang, H., Xu, T., and Wong, K. (2023). MDM: Molecular Diffusion Model for 3D Molecule Generation. In AAAI (AAAI Press), pp. 5105–5112. https://doi.org/10.48550/arXiv.2209.05710.

13. Luo, S., Shi, C., Xu, M., and Tang, J. (2021). Predicting Molecular Conformation via Dynamic Graph Score Matching. In NeurIPS, pp. 19784–19795. https://doi.org/10.5555/3540261.3541774.

14. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., and Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. ICML. In JMLR Workshop and Conference Proceedings, *37*JMLR Workshop and Conference Proceedings (JMLR), pp. 2256–2265. https://doi.org/10.48550/arXiv.1503.03585.

15. Ho, J., Jain, A., and Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In NeurIPS. https://doi.org/10.48550/arXiv.2006.11239.

16. Hoogeboom, E., Satorras, V.G., Vignac, C., and Welling, M. (2022). Equivariant Diffusion for Molecule Generation in 3D. ICML. In Proceedings of Machine Learning Research, *162*Proceedings of Machine Learning Research (PMLR), pp. 8867–8887. https://doi.org/10.48550/arXiv.2203.17003.

17. Bao, F., Zhao, M., Hao, Z., Li, P., Li, C., and Zhu, J. (2023). Equivariant Energy-Guided SDE for Inverse Molecular Design. In ICLR (OpenReview). https://doi.org/10.48550/arXiv.2209.15408.

18. Honório, K., Moda, T., and Andricopulo, A. (2013). Pharmacokinetic properties and in silico ADME modeling in drug discovery. Med. Chem. *9*, 163–176. https://doi.org/10.2174/1573406411309020002.

19. Gebauer, N.W.A., Gastegger, M., Hessmann, S.S.P., Müller, K.R., and Schütt, K.T. (2022). Inverse design of 3d molecular structures with conditional generative neural networks. Nat. Commun. *13*, 973. https://doi.org/10.1038/s41467-022-28526-y.

20. Lee, M., and Min, K. (2022). MGCVAE: Multi-Objective Inverse Design via Molecular Graph Conditional Variational Autoencoder. J. Chem. Inf. Model. *62*, 2943–2950. https://doi.org/10.1021/acs.jcim.2c00487.

21. Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1) (Association for Computational Linguistics), pp. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805.

22. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR. https://doi.org/10.48550/arXiv.1907.11692.

23. Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In EMNLP/IJCNLP (1) (Association for Computational Linguistics), pp. 3613–3618. https://doi.org/10.48550/arXiv.1903.10676.

24. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P.J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Res. *21*, 1–67. https://doi.org/10.5555/3455716.3455856.

25. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. In NeurIPS. https://doi.org/10.48550/arXiv.2005.14165.

26. OpenAI (2023). GPT-4 Technical Report. CoRR. https://doi.org/10.48550/arXiv.2303.08774.

27. Su, B., Du, D., Yang, Z., Zhou, Y., Li, J., Rao, A., Sun, H., Lu, Z., and Wen, J. (2022). A Molecular Multimodal Foundation Model Associating Molecule Graphs with Natural Language. CoRR. https://doi.org/10.48550/ARXIV.2209.05481.

28. Edwards, C., Zhai, C., and Ji, H. (2021). Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. In EMNLP (1) (Association for Computational Linguistics), pp. 595–607. https://doi.org/10.18653/v1/2021.emnlp-main.47.

29. Edwards, C., Lai, T.M., Ros, K., Honke, G., Cho, K., and Ji, H. (2022). Translation between Molecules and Natural Language. In EMNLP (Association for Computational Linguistics), pp. 375–413. https://doi.org/10.48550/arXiv.2204.11817.

30. Ramakrishnan, R., Dral, P.O., Rupp, M., and von Lilienfeld, O.A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. Sci. Data *1*, 140022. https://doi.org/10.1038/sdata.2014.22.

31. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2021). PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. *49*, D1388–D1395. https://doi.org/10.1093/NAR/GKAA971.

32. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. *36*, 344–350. https://doi.org/10.1093/NAR/GKM791.

33. Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Gaudry, A., Graham, J.G., Stephan, R., Page, R., Vondrášek, J., et al. (2022). The LOTUS initiative for open knowledge management in natural products research. Elife *11*, e70780. https://doi.org/10.7554/eLife.70780.

34. Wishart, D., Arndt, D., Pon, A., Sajed, T., Guo, A.C., Djoumbou, Y., Knox, C., Wilson, M., Liang, Y., Grant, J., et al. (2015). T3DB: The toxic exposome database. Nucleic Acids Res. *43*, D928–D934. https://doi.org/10.1093/nar/gku1004.

35. Satorras, V.G., Hoogeboom, E., and Welling, M. (2021). E(n) Equivariant Graph Neural Networks. ICML. In Proceedings of Machine Learning Research, *139*Proceedings of Machine Learning Research (PMLR). https://doi.org/10.48550/arXiv.2102.09844.

36. Simonovsky, M., and Komodakis, N. (2018). GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. Lecture Notes in Computer Science. In ICANN (1), *11139*ICANN (1) (Springer), pp. 412–422. https://doi.org/10.48550/arXiv.1802.03480.

37. Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., and Chan, W. (2021). WaveGrad: Estimating Gradients for Waveform Generation. In ICLR (OpenReview). https://doi.org/10.48550/arXiv.2009.00713.

38. Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2021). DiffWave: A Versatile Diffusion Model for Audio Synthesis. In ICLR (OpenReview). https://doi.org/10.48550/arXiv.2009.09761.

39. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. *28*, 31–36. https://doi.org/10.1021/CI00057A005.

40. Kotsias, P.C., Arús-Pous, J., Chen, H., Engkvist, O., Tyrchan, C., and Bjerrum, E.J. (2020). Direct steering of *de novo* molecular generation with descriptor conditional recurrent neural networks. Nat. Mach. Intell. *2*, 254–265. https://doi.org/10.1038/s42256-020-0174-5.

41. Jin, W., Barzilay, R., and Jaakkola, T.S. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. ICML. In Proceedings of Machine Learning Research, *80*Proceedings of Machine Learning Research (PMLR), pp. 2328–2337. https://doi.org/10.48550/arXiv.1802.04364.

42. Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T.S. (2022). Torsional Diffusion for Molecular Conformer Generation. In NeurIPS. https://doi.org/10.48550/arXiv.2206.01729.

43. Nesterov, V., Wieser, M., and Roth, V. (2020). 3DMolNet: A Generative Network for Molecular Structures. CoRR. https://doi.org/10.48550/arXiv.2010.06477.

44. Hoffmann, M., and Noé, F. (2019). Generating valid Euclidean distance matrices. CoRR. https://doi.org/10.48550/arXiv.1910.03131.eprint:1910.03131.

45. Kusner, M.J., Paige, B., and Hernández-Lobato, J.M. (2017). Grammar Variational Autoencoder. ICML. In Proceedings of Machine Learning Research, *70*Proceedings of Machine Learning Research (PMLR), pp. 1945–1954. https://doi.org/10.48550/arXiv.1703.01925.

46. Dai, H., Tian, Y., Dai, B., Skiena, S., and Song, L. (2018). Syntax-Directed Variational Autoencoder for Structured Data. In ICLR (Poster) (OpenReview). https://doi.org/10.48550/arXiv.1802.08786.

47. Liu, Q., Allamanis, M., Brockschmidt, M., and Gaunt, A.L. (2018). Constrained Graph Variational Autoencoders for Molecule Design. In NeurIPS, pp. 7806–7815. https://doi.org/10.48550/arXiv.1805.09076.

48. Madhawa, K., Ishiguro, K., Nakago, K., and Abe, M. (2019). GraphNVP: An Invertible Flow Model for Generating Molecular Graphs. CoRR. https://doi.org/10.48550/arXiv.1905.11600.

49. Luo, Y., Yan, K., and Ji, S. (2021). GraphDF: A Discrete Flow Model for Molecular Graph Generation. ICML. In Proceedings of Machine Learning Research, *139*Proceedings of Machine Learning Research (PMLR), pp. 7192–7203. https://doi.org/10.48550/arXiv.2102.01189.

50. Bian, Y., Wang, J., Jun, J.J., and Xie, X.-Q. (2019). Deep Convolutional Generative Adversarial Network (dcGAN) Models for Screening and Design of Small Molecules Targeting Cannabinoid Receptors. Mol. Pharm. *16*, 4451–4460. https://doi.org/10.1021/acs.molpharmaceut.9b00500.

51. Assouel, R., Ahmed, M., Segler, M.H.S., Saffari, A., and Bengio, Y. (2018). DEFactor: Differentiable Edge Factorization-based Probabilistic Graph Generation. CoRR. https://doi.org/10.48550/arXiv.1811.09766.

52. Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., and Tang, J. (2020). GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation. In ICLR (OpenReview). https://doi.org/10.48550/arXiv.2001.09382.

53. Popova, M., Shvets, M., Oliva, J., and Isayev, O. (2019). MolecularRNN: Generating realistic molecular graphs with optimized

properties. CoRR. https://doi.org/10.48550/arXiv.1905.13372.

54. Flam-Shepherd, D., Zhu, K., and Aspuru-Guzik, A. (2022). Language Models can learn Complex Molecular Distributions. Nat. Commun. *13*, 3293. https://doi.org/10.1038/s41467-022-30839-x.

55. Wu, L., Gong, C., Liu, X., Ye, M., and Liu, Q. (2022). Diffusion-based Molecule Generation with Informative Prior Bridges. In NeurIPS. https://doi.org/10.48550/arXiv.2209.00865.

56. Kang, S., and Cho, K. (2019). Conditional Molecular Design with Deep Generative Models. J. Chem. Inf. Model. *59*, 43–52. https://doi.org/10.1021/acs.jcim.8b00263.

57. Yang, M., Sun, H., Liu, X., Xue, X., Deng, Y., and Wang, X. (2023). CMGN: a conditional molecular generation net to design target-specific molecules with desired properties. Briefings Bioinf. *24*, bbad185. https://doi.org/10.1093/bib/bbad185.

58. Sanchez-Lengeling, B., and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. Science *361*, 360–365. https://doi.org/10.1126/science.aat2663.

59. Dhariwal, P., and Nichol, A.Q. (2021). Diffusion Models Beat GANs on Image Synthesis. In NeurIPS, pp. 8780–8794. https://doi.org/10.48550/arXiv.2105.05233.

60. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. In CVPR (IEEE), pp. 10674–10685. https://doi.org/10.48550/arXiv.2112.10752.

61. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In CVPR (IEEE), pp. 22500–22510. https://doi.org/10.48550/arXiv.2208.12242.

62. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. In ICLR (OpenReview). https://doi.org/10.48550/arXiv.2011.13456.

63. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., and Norouzi, M. (2023). Image Super-Resolution via Iterative Refinement. IEEE Trans. Pattern Anal. Mach. Intell. *45*, 4713–4726. https://doi.org/10.1109/TPAMI.2022.3204461.

64. Schneider, F. (2023). ArchiSound: Audio Generation with Diffusion. CoRR. https://doi.org/10.48550/ARXIV.2301.13267.

65. Thomas, N., Smidt, T.E., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., and Riley, P. (2018). Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds. CoRR. https://doi.org/10.48550/arXiv.1802.08219.

66. Fuchs, F., Worrall, D.E., Fischer, V., and Welling, M. (2020). SE3-Transformers: 3D Roto-Translation Equivariant Attention Networks. In NeurIPS. https://doi.org/10.48550/arXiv.2006.10503.

67. Finzi, M., Stanton, S., Izmailov, P., and Wilson, A.G. (2020). Generalizing Convolutional Neural Networks for Equivariance to Lie Groups on Arbitrary Continuous Data. ICML. In Proceedings of Machine Learning Research, *119*Proceedings of Machine Learning Research (PMLR), pp. 3165–3176. https://doi.org/10.48550/arXiv.2002.12880.

68. Köhler, J., Klein, L., and Noé, F. (2020). Equivariant Flows: Exact Likelihood Generative Learning for Symmetric Densities. ICML. In Proceedings of Machine Learning Research, *119*Proceedings of Machine Learning Research (PMLR), pp. 5361–5370. https://doi.org/10.48550/arXiv.2006.02425.

69. Xu, M., Yu, L., Song, Y., Shi, C., Ermon, S., and Tang, J. (2022). GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In ICLR (OpenReview). https://doi.org/10.48550/arXiv.2203.02923.

70. Hamilton, W.L., Ying, Z., and Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. In NIPS, pp. 1024–1034. https://doi.org/10.48550/arXiv.1706.02216.

71. Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). How Powerful are Graph Neural Networks? In ICLR. https://doi.org/10.48550/arXiv.1810.00826.

72. Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. (2022). Pre-training Molecular Graph Representation with 3D Geometry. In ICLR (OpenReview). https://doi.org/10.48550/arXiv.2110.07728.

73. Zeng, Z., Yao, Y., Liu, Z., and Sun, M. (2022). A Deep-learning System Bridging Molecule Structure and Biomedical Text with Comprehension Comparable to Human Professionals. Nat. Commun. *13*, 862. https://doi.org/10.1038/s41467-022-28494-3.

74. Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. (2021). ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In ICCV (IEEE), pp. 14347–14356. https://doi.org/10.48550/arXiv.2108.02938.

75. James, A.T., and Wilkinson, G.N. (1971). Factorization of the residual operator and canonical decomposition of nonorthogonal factors in the analysis of variance. Biometrika *58*, 279–294. https://doi.org/10.2307/2334516.

76. Willmott, C.J., and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. *30*, 79–82. https://doi.org/10.3354/cr030079.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Processed PubChem Dataset	This paper, PubChem	https://pubchem.ncbi.nlm.nih.gov/
QM9 Dataset	Quantum Machine	http://quantum-machine.org/datasets/
Software and algorithms		
TextSMOG Model	This paper	https://github.com/lyc0930/TEDMol

METHOD DETAILS

In this section, we elaborate on the proposed text-guided small molecule generation approach via diffusion model (TextSMOG), as illustrated in Figure 1. It integrates the textual information (*i.e.*, text guidance) into the conditional signal of diffusion models by employing the reference geometry that is described in the first subsection following. Subsequently, we introduce an efficient learning approach that incorporates both the encoded conditional signal and pre-trained unconditional signal in the reverse process, to generate molecules that are not only structurally stable and chemically valid but also align well with the specified conditions, as presented in the second subsection.

Notation and background

We begin with a background of diffusion-based 3D molecule generation, introducing the fundamental concepts of the diffusion model and delving into equivariant diffusion models. See the comprehensive literature review on these topics in the Section Related works in supplemental information. In accordance with prior studies,[12,16,17] we use the variable $\mathcal{G} = (\mathbf{x}, \mathbf{h})$ to represent the 3D molecular geometry. Here $\mathbf{x} = (x_1, \ldots, x_M) \in \mathbb{R}^{M\times 3}$ signifies the atom coordinates, while $\mathbf{h} = (h_1, \ldots, h_M) \in \mathbb{R}^{M\times k}$ denotes the atom features. These features encompass atom types and atom charges, characterizing the atomic properties within the molecular structure.

Diffusion model

The diffusion model[14,15] emerges as a leading generative model, having achieved great success in various domains.[59–64] Typically, it is formulated as two Markov chains: a forward process (*aka.* noising process) that gradually injects noise into the data, and a reverse process (*aka.* denoising process) that learns to recover the original data. Such a reverse process endows the diffusion model with enhanced capabilities for effective data generation and recovery.

Forward Process. Given the real 3D molecular geometry $\mathcal{G}_0$, the forward process yields a sequence of intermediate variables $\mathcal{G}_1, \ldots, \mathcal{G}_T$ using the transition kernel $q(\mathcal{G}_t|\mathcal{G}_{t-1})$ in alignment with a variance schedule $\beta_1, \beta_2, \ldots, \beta_T \in (0,1)$. Formally, it is expressed as:

$$q(\mathcal{G}_t|\mathcal{G}_{t-1}) = \mathcal{N}\left(\mathcal{G}_t\middle|\sqrt{1-\beta_t}\,\mathcal{G}_{t-1}, \beta_t\mathbf{I}_n\right), \qquad \text{(Equation 1)}$$

where $\mathcal{N}(\cdot|\cdot,\cdot)$ is a Gaussian distribution and $\mathbf{I}_n$ is the identity matrix. This defines the joint distribution of $\mathcal{G}_1, \ldots, \mathcal{G}_T$ conditioned on $\mathcal{G}_0$ using the chain rule of the Markov process:

$$q(\mathcal{G}_1, \ldots, \mathcal{G}_T|\mathcal{G}_0) = \prod_{t=1}^{T} q(\mathcal{G}_t|\mathcal{G}_{t-1}). \qquad \text{(Equation 2)}$$

Let $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t : = \prod_{s=1}^{t} \alpha_s$. The sampling of $\mathcal{G}_t$ at time step t is in a closed form:

$$q(\mathcal{G}_t|\mathcal{G}_0) = \mathcal{N}(\mathcal{G}_t|\sqrt{\overline{\alpha}_t}\,\mathcal{G}_0, (1-\overline{\alpha}_t)\mathbf{I}_n). \qquad \text{(Equation 3)}$$

Accordingly, the forward process posteriors, when conditioned on $\mathcal{G}_0$, are tractable as:

$$q(\mathcal{G}_{t-1}|\mathcal{G}_t, \mathcal{G}_0) = \mathcal{N}(\mathcal{G}_{t-1}|\tilde{\mu}(\mathcal{G}_t, \mathcal{G}_0), \tilde{\beta}_t\mathbf{I}_n), \qquad \text{(Equation 4)}$$

where

$$\tilde{\mu}(\mathcal{G}_t, \mathcal{G}_0) = \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\mathcal{G}_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}{1-\overline{\alpha}_t}\mathcal{G}_t, \tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t. \qquad \text{(Equation 5)}$$

Reverse Process. To recover the original molecular geometry $\mathcal{G}_0$, the diffusion model starts by generating a standard Gaussian noise $\mathcal{G}_T \sim \mathcal{N}(\mathbf{O}, \mathbf{I}_n)$, then progressively eliminates noise through a reverse Markov chain. This is characterized by a learnable transition kernel $p_\theta(\mathcal{G}_{t-1}|\mathcal{G}_t)$ at each reverse step t, defined as:

$$p_\theta(\mathcal{G}_{t-1}|\mathcal{G}_t) = \mathcal{N}(\mathcal{G}_{t-1}|\mu_\theta(\mathcal{G}_t, t), \Sigma_\theta(\mathcal{G}_t, t)), \tag{Equation 6}$$

where the variance $\Sigma_\theta(\mathcal{G}_t, t) = \tilde{\beta}_t \mathbf{I}_n$ and the mean $\mu_\theta(\mathcal{G}_t, t)$ is parameterized by deep neural networks with parameters θ:

$$\mu_\theta(\mathcal{G}_t, t) = \tilde{\mu}_t\left(\mathcal{G}_t, \frac{1}{\sqrt{\overline{\alpha}_t}}\left(\mathcal{G}_t - \sqrt{1-\overline{\alpha}_t}\,\epsilon_\theta(\mathcal{G}_t, t)\right)\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathcal{G}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\mathcal{G}_t, t)\right), \tag{Equation 7}$$

where ϵ_θ is a noise prediction function to approximate the noise ϵ from $\mathcal{G}_t$.

With the reverse Markov chain, we can iteratively sample from the learnable transition kernel $p_\theta(\mathcal{G}_{t-1}|\mathcal{G}_t)$ until $t = 1$ to estimate the molecular geometry $\mathcal{G}_0$.

Equivariant diffusion models

The molecular geometry $\mathcal{G} = (\mathbf{x}, \mathbf{h})$ is inherently symmetric in 3D space — that is, translating or rotating a molecule does not change its underlying structure or features. Previous studies[65–67] underscore the significance of leveraging these invariances in molecular representation learning for enhanced generalization. However, the transformation of these higher-order representations usually requires computationally expensive approximations or coefficients.[16,35] In contrast, equivariant diffusion models[16,68,69] provide a more efficient approach to ensure both rotational and translational invariance. The approach rests on the assumption that, with the model distribution $p(\mathcal{G}) = p(\mathbf{x}, \mathbf{h})$ remaining invariant to the Euclidean group E(3), identical molecules, despite being in different orientations, will correspond to the same distribution. Based on this assumption, translational invariance is achieved by predicting only the deviations in coordinate with a zero center of mass, *i.e.*, $\sum_{i=1}^{M} x_i = 0$. On the other hand, rotational invariance is accomplished by making the noise prediction network $\epsilon_\theta(\cdot)$ equivariant to orthogonal transformations.[16,35] Specifically, given an orthogonal matrix $\mathbf{R}$ representing a coordinate rotation or reflection, the conformation output $a^{\mathbf{x}}$ from the network $\epsilon_\theta(\mathcal{G}) = \epsilon_\theta(\mathbf{x}, \mathbf{h}) = (a^{\mathbf{x}}, a^{\mathbf{h}})$ is equivariant to $\mathbf{R}$, if the following condition holds for all orthogonal matrices $\mathbf{R}$:

$$\epsilon_\theta(\mathbf{R}\mathbf{x}, \mathbf{h}) = \left(\mathbf{R}a^{\mathbf{x}}, a^{\mathbf{h}}\right). \tag{Equation 8}$$

A model exhibiting rotational and translational equivariance means a neural network $p_\theta(\mathcal{G})$ can avoid learning orientations and translations of molecules from scratch.[16,35] In this paper, we parameterize the noise prediction network ϵ_θ using an E(n) equivariant graph neural network as introduced by,[35] which is a type of Graph Neural Network[70] that satisfies the above equivariance constraint to E(3).

Equivariant diffusion model for molecule generation

Diffusion models, formulated as two Markov chains—a forward process that gradually injects noise into the data and a reverse process that learns to recover the original data—have been successfully applied to various domains, including molecule generation. This process is particularly effective in the context of molecule generation, where the forward process adds noise to the molecular geometry at each step until it is fully noise-corrupted. The reverse process then gradually denoises the initial geometry $\mathcal{G}_T$ to generate the final molecular geometry $\mathcal{G}_0$.

However, molecular geometries are inherently symmetric in 3D space—translations or rotations do not change their underlying structure or features. To take advantage of these invariances for improved generalization, we employ an equivariant diffusion model (EDM). The EDM ensures both rotational and translational invariance by predicting only the deviations in coordinate with a zero center of mass and making the noise prediction network $\epsilon_\theta(\cdot)$ equivariant to orthogonal transformations. This allows the model distribution $p(\mathcal{G})$ to remain invariant to the Euclidean group E(3), meaning identical molecules in different orientations correspond to the same distribution.

In this work, the integration of textual information into the conditional signal of the equivariant diffusion model is achieved by employing a reference geometry $\mathbf{c_P}$ that is updated at each step based on the textual prompt $\mathbf{P}$.

Integrating textual prompts into 3D molecular reference geometry

To ensure high-fidelity 3D molecule generation, the reverse process of the diffusion model is typically guided by tailored conditional information representing desired properties like unique polarizability. We represent this conditional information as c, which allows us to formulate the conditional reverse process as:

$$p_\theta(\mathcal{G}_{t-1}|\mathcal{G}_t, c) = \mathcal{N}(\mathcal{G}_{t-1}|\mu_\theta(\mathcal{G}_t, c, t), \tilde{\beta}_t \mathbf{I}_n) \tag{Equation 9}$$

Unlike previous approaches relying on limited value guidance (*i.e.*, property values), in this work, we aim to steer the reverse process with text guidance (*i.e.*, informative textual descriptions), which can convey a broader range of conditional requirements. Intuitively, utilizing textual descriptions to specify conditional generation criteria not only provides greater expressivity but also better aligns the resulting 3D molecules with diverse and complex expectations.

Practically, we first introduce a textual prompt $\mathbf{P}$ describing desired 3D molecule properties. A multi-modal conversion module Γ, pre-trained on 300K text-molecule pairs from PubChem, is then employed. This module is comprised of a GIN molecular graph encoder[71,72] and a language encoder-decoder extended from BERT.[21,73] It converts $\mathbf{P}$ into a reference geometry $\mathbf{c_P}$, extracting specific information from the target conditions and refining the textual condition signal:

$$\mathbf{c_P} = \Gamma(\mathbf{P}). \quad \text{(Equation 10)}$$

Nevertheless, we should emphasize that valid and stable 3D molecules can hardly be obtained directly from $\mathbf{c_P}$. The chemical fidelity in 3D molecular space may not be guaranteed. In what follows, we describe how to utilize $\mathbf{c_P}$ for conditioning a pre-trained diffusion model to generate molecules that align with the desired properties, meanwhile alleviating the exhaustive training from scratch.

Conditioning with the reference of text guidance

To leverage $\mathbf{c_P}$ for text-guided conditional generation while preserving the validity and stability of the synthesized molecule, TextSMOG employs the iterative latent variable refinement (ILVR)[74] to condition a pre-trained unconditional diffusion model meanwhile maintaining inherent domain knowledge in the unconditional model.

With the pre-trained unconditional diffusion model EDM,[16] we could perform a step-by-step reverse process. Formally, at step t, we can sample an unconditional proposal molecular geometry:

$$\tilde{\mathcal{G}}_{t-1} \sim \tilde{p}_{\tilde{\theta}}(\tilde{\mathcal{G}}_{t-1}|\mathcal{G}_t). \quad \text{(Equation 11)}$$

where $\tilde{\theta}$ is the fixed parameters of the pre-trained unconditional diffusion model.[16] Then, to incorporate the condition signal $\mathbf{c_P}$ in the reverse process, we introduce a linear operation $\phi_\theta(\cdot)$. Therefore the conditional denoising for one step at step t can be formulated as:

$$\mathcal{G}_{t-1} = \phi_\theta(\mathbf{c_P}) + (\mathcal{I} - \phi_\theta)(\tilde{\mathcal{G}}_{t-1}), \quad \text{(Equation 12)}$$

where $\mathcal{I}(\cdot)$ is the identity operation and $(\mathcal{I} - \phi_\theta)(\cdot)$ is the residual operation *w.r.t.* $\phi_\theta(\cdot)$[75]. Accordingly, the condition signal $\mathbf{c_P}$ is projected into the reverse denoising process by $\phi_\theta(\cdot)$, thus $\mathcal{G}_{t-1}$ is obtained as the generated 3D molecular geometry conditioned on $\mathbf{c_P}$. Conceptually, the proposal geometry from unconditional generation $\tilde{\mathcal{G}}_{t-1}$ tries to push the atoms into a chemically valid position, while the reference geometry $\mathbf{c_P}$ pulls the atoms toward the structure targeted to the condition.

By matching latent variables following Equation 12, we enable text-guided conditional generation with the unconditional diffusion model. Accordingly, the one-step denoising distribution conditioned on textual guidance at each step t can be reformulated as:

$$\mathcal{G}_{t-1} \sim p_\theta(\mathcal{G}_{t-1}|\mathcal{G}_t, \mathbf{c_P}). \quad \text{(Equation 13)}$$

Training objective

To guarantee the quality of the generated molecules, the key lies in optimizing the variational lower bound (ELBO) of negative log likelihood, which equals minimizing the Kullback-Leibler divergence between the joint distribution of the reverse Markov chain $p_\theta(\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_T)$ and the forward process $q(\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_T)$:

$$\mathbb{E}\left[-\log p_\theta(\mathcal{G}_0|\mathbf{c_P})\right] \le -\log \sum_{t \ge 1} \underbrace{D_{\mathrm{KL}}\left(q(\mathcal{G}_{t-1}|\mathcal{G}_t, \mathcal{G}_0) \middle\| p_\theta(\mathcal{G}_{t-1}|\mathcal{G}_t, \mathbf{c_P})\right)}_{:=\mathcal{L}_{t-1}} + C, \quad \text{(Equation 14)}$$

where C is a constant independent of θ.

Note that we set $\mathcal{L}_0 = -\log p_\theta(\mathcal{G}_0|\mathcal{G}_1)$ as a discrete decoder following.[15] Further adopting the reparameterization from,[15] $\mathcal{L}_{t-1}$ can be simplified to:

$$\mathcal{L}_{t-1} = \mathbb{E}_{\mathrm{P},\mathcal{G}_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\overline{\alpha}_t}\,\mathcal{G}_0 + \sqrt{1 - \overline{\alpha}_t}\,\epsilon, t, \mathbf{c_P}\right)\right\|^2\right]. \quad \text{(Equation 15)}$$

Evaluation metrics

Mean absolute error (MAE).[76] is a measure of errors between paired observations. Given the property classifier network ϕ_p, and the set of generated molecules $\mathbb{G}$, the MAE is defined as:

$$\mathrm{MAE} = \frac{1}{|\mathbb{G}|} \sum_{\mathcal{G} \in \mathbb{G}} \left|\phi_p(\mathcal{G}) - c_{\mathcal{G}}\right|, \quad \text{(Equation 16)}$$

where $\mathcal{G}$ is the generated molecule, and of which $c_{\mathcal{G}}$ is the desired property.

Novelty.[36] is the proportion of generated molecules that do not appear in the training set. Specifically, let $\mathbb{G}$ be the set of generated molecules, the novelty in our experiment is calculated as:

$$\mathrm{Novelty} = \frac{|\mathbb{G} \cap \mathbb{D}_b|}{|\mathbb{G}|}. \quad \text{(Equation 17)}$$

Atom stability.[16] is the proportion of the atoms in the generated molecules that have the right valency. Specifically, the atom stability in our experiment is calculated as:

$$\text{AtomStability} = \frac{\sum_{\mathcal{G}\in\mathbb{G}}\left|\mathbb{A}_{\mathcal{G},\text{stable}}\right|}{\sum_{\mathcal{G}\in\mathbb{G}}\left|\mathbb{A}_{\mathcal{G}}\right|}, \tag{Equation 18}$$

where $\mathbb{A}_{\mathcal{G}}$ is the set of atoms in the generated molecule $\mathcal{G}$, and $\mathbb{A}_{\mathcal{G},\text{stable}}$ is the set of atoms in $\mathbb{A}_{\mathcal{G}}$ that have the right valency.

Molecule stability.[16] is the proportion of the generated molecules where all atoms are stable. Specifically, the molecule stability in our experiment is calculated as:

$$\text{MoleculeStability} = \frac{\left|\mathbb{G}_{\text{stable}}\right|}{\left|\mathbb{G}\right|}, \tag{Equation 19}$$

where $\mathbb{G}_{\text{stable}}$ is the set of generated molecules where all atoms have the right valency.

QUANTIFICATION AND STATISTICAL ANALYSIS

The quantum properties in QM9 dataset

We consider 6 main quantum properties in QM9:

- C_v: Heat capacity at 298.15K.
- μ: Dipole moment.
- α: Polarizability, which represents the tendency of a molecule to acquire an electric dipole moment when subjected to an external electric field.
- $\varepsilon_{\text{HOMO}}$: Highest occupied molecular orbital energy.
- $\varepsilon_{\text{LUMO}}$: Lowest unoccupied molecular orbital energy.
- Δ_{ε}: The energy gap between HOMO and LUMO.