BMC Biology

RESEARCH

Open Access

Instruction multi-constraint molecular generation using a teacher-student large language model

Peng Zhou[1,9], Jianmin Wang[2], Chunyan Li[3], Zixu Wang[4], Yiping Liu[1], Siqi Sun[5,6], Jianxin Lin[1], Leyi Wei[7,8], Xibao Cai[1], Houtim Lai[9], Wei Liu[9], Longyue Wang[10*], Yuansheng Liu[1*] and Xiangxiang Zeng[1*]

Abstract

Background While various models and computational tools have been proposed for structure and property analysis of molecules, generating molecules that conform to all desired structures and properties remains a challenge.

Results We introduce a multi-constraint molecular generation large language model, TSMMG, which, akin to a student, incorporates knowledge from various small models and tools, namely, the "teachers." To train TSMMG, we construct a large set of text-molecule pairs by extracting molecular knowledge from these "teachers," enabling it to generate novel molecules that conform to the descriptions through various text prompts. We experimentally show that TSMMG remarkably performs in generating molecules that meet complex property requirements described in natural language across two-, three-, and four-constraint tasks, with an average molecular validity of over 99% and success ratio of 82.58%, 68.03%, and 67.48%, respectively. The model also exhibits adaptability through zero-shot testing, creating molecules that satisfy combinations of properties that have not been encountered. It can comprehend text inputs with various language styles, extending beyond the confines of outlined prompts.

Conclusions TSMMG presents an effective model for multi-constraint molecular generation using natural language. This framework is not only applicable to drug discovery but also serves as a reference for other related fields.

Keywords Molecular generation, Large language model, Multi-constraint

*Correspondence:
Longyue Wang
vincentwang0229@gmail.com
Yuansheng Liu
yuanshengliu@hnu.edu.cn
Xiangxiang Zeng
xzeng@hnu.edu.cn
[1] College of Information Science and Engineering, Hunan University, Changsha 410082, Hunan, China
[2] The Interdisciplinary Graduate Program in Integrative Biotechnology, Yonsei University, Incheon 21983, Seoul, Korea
[3] School of Informatics, Yunnan Normal University, Kunming 650500, Yunnan, China
[4] Department of Computer Science, University of Tsukuba, Tsukuba 3058577, Japan
[5] Research Institute of Intelligent Complex Systems, Fudan University, Shanghai 200433, China
[6] Shanghai AI Laboratory, Shanghai 200232, China
[7] Centre for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Science, Macao Polytechnic University, Macao SAR, China
[8] School of Informatics, Xiamen University, Xiamen, China
[9] AI for Life Sciences Lab, Tencent, Shenzhen, China
[10] Alibaba International Digital Commerce, Hangzhou, China

Background

The development and application of molecular generation models play an essential role in the field of artificial intelligence for drug discovery (AIDD). Molecular generation models are instrumental in addressing the challenges and complexities associated with the identification and design of novel therapeutic compounds. In contrast to traditional virtual screening approaches, involving the sift of desired molecules from existing libraries, these innovative models are engineered to directly generate novel molecules. Their ability to navigate vast chemical spaces, optimize lead compounds, and facilitate de novo design positions them as indispensable tools in the pursuit of novel and effective therapeutic interventions [1–7]. These models not only exhibit the ability to generate chemically valid molecules that precisely adhere to the requirements of molecular analysis tools [8, 9], but they also excel in the generation of molecules that meet specific constraints, like quantitative estimate of drug-likeness (QED) and molecular hydrophobicity (LogP) [10, 11].

However, a primary challenge in the realm of drug discovery lies in identifying molecules that conform to a multitude of constraints, including binding affinity, LogP, QED, synthetic accessibility (SA), and toxicity, rather than merely generating compounds that are chemically valid or solely meeting specific criteria [12, 13]. Several works have been introduced to address this challenge, presenting methodologies capable of generating molecules that adhere to a spectrum of concurrent condition constraints. For instance, Li et al. introduced a conditional generative model proficient in generating molecules that meet both SA and QED criteria, even yielding dual-target inhibitors for JNK3 and GSK3 [14]. Jin et al. achieved this feat by extracting diverse substructures with varying properties and reassembling them to produce molecules satisfying QED, SA, and the inhibition of both JNK3 and GSK3 [15]. Bagal et al. employed a transformer decoder architecture, treating constraint conditions as conditional codes, to explore the generation of molecules under various combinations of multiple constraints, including LogP, TPSA (total polar surface area), and SA [16]. Wang et al. utilized a combination of a conditional transformer, knowledge distillation, and reinforcement learning to generate molecules with activity against DRD2, while also ensuring adherence to QED and SA criteria [12].

Although significant progress has been made in prior endeavors, it is important to acknowledge that multi-constraint molecular generation methods still suffer from several noteworthy limitations, which hinder their practical applicability. These limitations undermine the overall effectiveness and efficiency of these methods in generating molecules that simultaneously meet diverse sets of constraints in drug discovery. These limitations fall into the following points: (1) Current multi-constraint molecular generation methods heavily rely on a narrow set of constraints. These methods predominantly focus on specific molecular properties, such as LogP, QED, SA, DRD2, JNK3, and GSK3. As a result, they may overlook other crucial aspects like substructures, bioavailability, and toxicity. The restricted range of constraints limits the comprehensive exploration of diverse chemical properties, potentially hindering the discovery and optimization of molecules with broader applicability in drug discovery and related domains. (2) These methods often require extensive fine-tuning when applied to different tasks. They tend to generate molecules that closely adhere to the feature distribution of the training dataset. As a consequence, adapting these models to changes in the target space or applying them to diverse tasks necessitates significant retraining. This inflexibility makes the models less adaptable, introducing a substantial burden in terms of computational resources and time when confronted with variations in the application context. (3) They often involve intricate designs. The complexity of the models and algorithms used can be a significant obstacle in their practical application. Users may find it challenging to understand and navigate the complexities of the methods, impacting their usability. Improving the simplicity of these models is essential to make them more accessible and applicable in real-world scenarios, especially in drug discovery and related domains.

To address the challenges, we introduce the teacher-student-based multi-constraint molecular generation (TSMMG) model, a natural language-based multi-constraint molecular generation approach. TSMMG offers several pivotal advantages: (1) Broader properties and high scalability: In addition to the constraints often focused on by existing methods, we also consider molecular substructures and ADMET properties. Based on the concept of knowledge distillation, our approach presents a versatile data generation framework that leverages a range of molecular tools and advanced models to selectively extract molecules with diverse properties from publicly available molecular libraries. This paradigm provides a highly scalable approach, facilitating the seamless absorption of molecular knowledge beyond the scope of this paper. Moreover, this approach can be easily extended to other domains, such as materials science. (2) Multi-task capability: Harnessing the capabilities of large language models, we train TSMMG across multiple tasks. By formulating distinct prompts, we delineate unique molecular spaces without the need for repetitive fine-tuning. (3) Simple architecture: TSMMG adopts a transformer-based decoder architecture. This design,

characterized by its simplicity, eliminates the need for intricate preprocessing of molecular data.

To showcase the expressive capabilities of our proposed model, we meticulously designed 16 sets of experiments for multi-constraint molecular generation. These experiments covered a spectrum of tasks, including molecular substructures, physicochemical properties, affinity with targets, and ADMET properties. Our findings from these experiments are compelling: TSMMG not only yields over 99% of legally valid molecules based on natural language instructions but also, notably, a substantial proportion of these molecules impeccably aligns with the specified properties in the textual descriptions. Furthermore, we conducted a noteworthy case study involving a zero-shot 5-constraint task. In this scenario, TSMMG successfully produced molecules capable of simultaneous binding to EP2 and EP4, showcasing favorable drug-likeness and synthetic accessibility, along with the ability to penetrate the blood-brain barrier. This case study serves as an additional testament to the vast potential embedded in TSMMG. Additionally, our model demonstrated its prowess in understanding natural language beyond the prompts outlined in this paper, as empirically validated. This expanded capability further solidifies the model's practical applicability. Moreover, we observed that integrating molecules generated by our model enhances the teacher model's performance. This collaborative synergy fosters continuous improvement between the teacher and student models, underscoring the model's adaptability and potential for iterative refinement.

Results

TSMMG approach

As shown in Fig. 1, TSMMG process involves the following steps: (1) First, a substantial dataset of molecules is collected from publicly available molecular libraries. This dataset undergoes analysis by advanced molecular parsing tools and models, which referred to as "teachers." These teachers extract extensive information, encompassing structural details, physicochemical properties, binding affinities to various targets, and other pertinent attributes for each molecule. The resulting knowledge is then organized into text descriptions, which are paired with the corresponding molecules. (2) Second, the "student" model, TSMMG, is trained using the knowledge obtained in the previous step. TSMMG is designed to create a direct mapping from natural language to molecular language. By absorbing a diverse range of knowledge expressed in natural language, the model acquires the capability to generate molecules that possess the specified properties outlined in the text. It is worth noting that TSMMG undergoes pre-training on a vast corpus of pure text, enabling it to effectively understand and interpret natural language. (3) When presented with a text description that includes multiple constraints, TSMMG can generate entirely novel molecules that fulfill these textual descriptions. In doing so, it effectively bridges the gap between natural language and molecular language for the purpose of multi-constraint molecular generation.

Multi-constraint task

Task setting

To comprehensively demonstrate the efficacy of the TSMMG model, we categorized multi-constraint tasks into three types, each based on different levels of complexity: two-constraint molecular generation, three-constraint molecular generation, and four-constraint molecular generation. Each of these three task categories comprises eight one-constraint tasks. These one-constraint tasks encompass:

- Task 1. Specifying a functional group (FG).
- Task 2. Specifying the level of hydrophilicity and hydrophobicity ($LogP = 1$).
- Task 3. Specifying the level of quantitative estimate of drug-likeness ($QED > 0.6$).
- Task 4. Specifying the level of synthetic accessibility score ($SAs < 4$).
- Task 5. Generate molecules with high affinity for the dopamine type 2 receptor ($DRD2 > 0.5$).
- Task 6. Generate molecules with high affinity for the glycogen synthase kinase-3 beta ($GSK3 > 0.5$).
- Task 7. Generate molecules capable of crossing the blood-brain barrier ($BBB > 0.5$).
- Task 8. Generate molecules that can be absorbed by the human small intestine ($HIA > 0.5$).

Among these, task 1 is classified as a **structure task**, while tasks 2, 3, and 4 are **physicochemical property tasks**. Tasks 5 and 6 fall under **activity tasks**, and tasks 7 and 8 are **ADMET property tasks**. We employ the "+" symbol to concatenate multiple one-constraint tasks, thereby representing multi-constraint tasks. Within the two-constraint tasks, we considered eight subtasks, combining structure tasks with individual physicochemical property tasks, activity tasks, and ADMET property tasks. These include (1) FG+FG, 2FG for short; (2) FG+LogP; (3) FG+QED; (4) FG+SAs; (5) FG+DRD2; (6) FG+GSK3; (7) FG+BBB; and (8) FG+HIA. In the three-constraint tasks, we explored subtasks such as (1) FG+DRD2+QED; (2) FG+GSK3+QED; (3) FG+BBB+QED; and (4) FG+HIA+QED. The four-constraint tasks include (1) FG+DRD2+QED+SAs; (2) FG+GSK3+QED+SAs; (3) FG+BBB+QED+SAs; and (4) FG+HIA+QED+SAs. It is essential to emphasize that all these tasks were completed within a single model, employing different natural

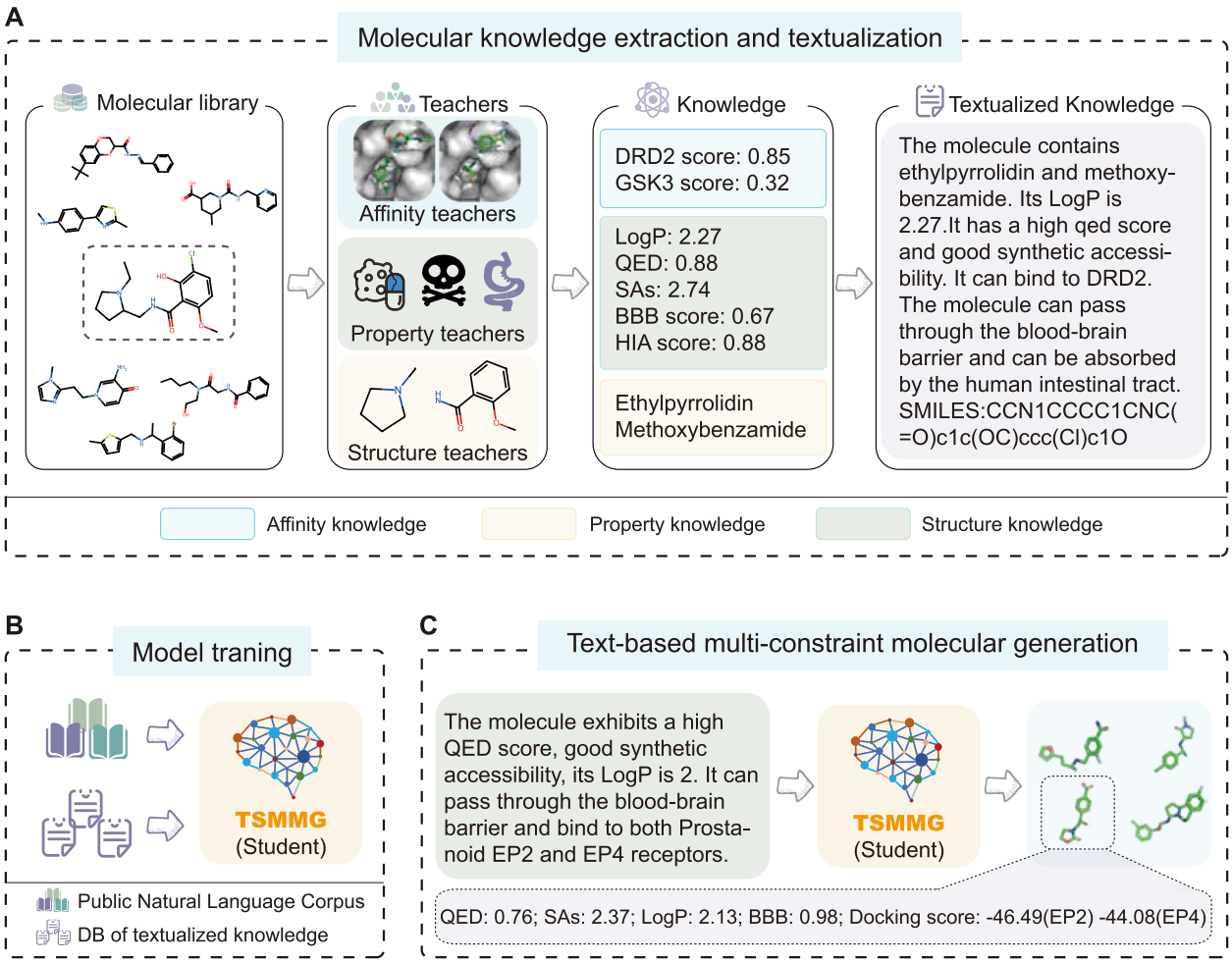

Fig. 1 The process of TSMMG is illustrated as follows: **A** We use "teacher" models to analyze molecules and obtain their properties, such as structural information, physicochemical properties, and binding affinities. These properties are then converted into natural language. The natural language descriptions and molecules form a text-molecule paired dataset. **B** The "student" model, TSMMG, is trained on this text-molecule paired dataset to map descriptions to molecular structures. Pre-training on a large text corpus helps it understand natural language. **C** TSMMG can generate new molecules based on text descriptions with multiple constraints, linking natural language to molecular generation

language prompts. The model underwent comprehensive training in a unified process, eliminating the need for repetitive fine-tuning. The specific prompts used in these experiments are detailed in Table 1.

Performance analysis

The experimental results, depicted in Fig. 2A and B, unveil several noteworthy findings: Validity: The model demonstrates a remarkable ability to generate molecules that adhere to the syntax rules of SMILES (Simplified Molecular Input Line Entry System) [17], with an impressive average validity rate of 99.87%, 99.89%, and 99.87% for two-constraint tasks, three-constraint tasks, and four-constraint tasks, respectively. This underscores the model's proficiency in consistently producing grammatically correct molecules. Uniqueness: Most generated molecules are unique, with an outstanding average uniqueness rate of 90.27%, 81.2%, and 81.89% for two-constraint tasks, three-constraint tasks, and four-constraint tasks, respectively. From a specific task perspective, the uniqueness of tasks related to DRD2 and GSK3 is relatively low, averaging less than 70%, while other tasks score above 90%. In the next section, we will analyze the reasons behind this situation. Overall, the model consistently generates largely distinct molecules across various tasks. Novelty: The average novelty of the generated molecules stands at 92.79%, 87.6%, and 87.87%. Similar to uniqueness, tasks related to DRD2 and GSK3, such as FG+DRD2+QED (82.76%),

Table 1 The prompts we use in this work. [FG], [FG1], and [FG2] refer to any functional group, and [VALUE] refers to a real number

Task	Prompt
2FG	The molecule contains [FG1],[FG2]
FG+LogP	The molecule contains [FG]. Its LogP is [VALUE]
FG+QED	The molecule contains [FG]. It has a high QED score
FG+SA	The molecule contains [FG]. It has good synthetic accessibility
FG+DRD2	The molecule contains [FG]. It is active to DRD2
FG+GSK3	The molecule contains [FG]. It is active to GSK3
FG+BBB	The molecule contains [FG]. It can pass through the blood-brain barrier
FG+HIA	The molecule contains [FG]. It can be absorbed by human intestinal
FG+DRD2+QED	The molecule contains [FG]. It is active to DRD2. It has a high QED score
FG+GSK3+QED	The molecule contains [FG]. It is active to GSK3. It has a high QED score
FG+BBB+QED	The molecule contains [FG]. It can pass through the blood-brain barrier. It has a high QED score
FG+HIA+QED	The molecule contains [FG]. It can be absorbed by human intestinal. It has a high QED score
FG+DRD2+QED+SAs	The molecule contains [FG]. It is active to DRD2. It has a high QED score. It has good synthetic accessibility
FG+GSK3+QED+SAs	The molecule contains [FG]. It is active to GSK3. It has a high QED score. It has good synthetic accessibility
FG+BBB+QED+SAs	The molecule contains [FG]. It can pass through the blood-brain barrier. It has a high QED score. It has good synthetic accessibility
FG+HIA+QED+SAs	The molecule contains [FG]. It can be absorbed by human intestinal. It has a high QED score. It has good synthetic accessibility
BTK	The molecule can bind to BTK
FGFR4	The molecule can bind to FGFR4
KPCD3	The molecule can bind to KPCD3
3CL	The molecule can bind to 3CL

FG+GSK3+QED (82.44%), FG+DRD2+QED+SA (83.9%), and FG+GSK3+QED+SA (83.14%), have relatively lower novelty scores, while other tasks have novelty scores exceeding 90%. In general, the model demonstrates a capacity to generate innovative molecules for most of the tasks at hand. Diversity: Most generated molecules exhibit notable structural differences, as reflected in the outstanding average diversity score of 90.47, 89.3, and 89.37. Similarly, tasks related to DRD2 and GSK3 exhibit lower diversity compared to other tasks. Success ratio: The average success ratio stands at 82.58%, 68.03%, and 67.48% for two-constraint tasks, three-constraint tasks, and four-constraint tasks, respectively, which demonstrates the model's efficacy in generating novel molecules that effectively meet all requirements specified by natural language.

Impact of FGs on performance

An important distinction from previous methods is that we consider functional groups (FG) as an additional constraint, allowing for more precise control over the direction of generation. Given the relatively low uniqueness in tasks related to DRD2 and GSK3, we use the FG+DRD2 task as an example to further discuss the impact of FG on generation results.

Firstly, we analyze the reasons for the low uniqueness (68.48) of the FG+DRD2 task. In this task, our prompt template is "The molecule contains [FG]. It is active to DRD2." We randomly selected 1000 FGs to form 1000 prompts, each generating 5 molecules, totaling 5000 molecules. The only difference between each prompt is the FG, so we group according to the number of unique molecules generated by each prompt and then extract the FG from these prompts for analysis. As shown in Fig. 2C(1), we see that the number of FGs that led to the generation of 1, 2, 3, 4, and 5 unique molecules were 195, 287, 252, 184, and 82, respectively. Over 80% of the FG-associated prompts can generate two or more unique molecules, with 82 FG-associated prompts each generating 5 completely different molecules. A significant portion of FG-associated prompts tend to generate identical molecules. We speculate that the main reason for this is the inconsistent frequency of these FGs in the training set, causing the model to be unable to effectively learn the larger molecular space corresponding to the FG. In light of this speculation, we grouped these 1000 FGs according to the number of unique molecules generated and calculated the average frequency of the FGs in the training set within each group. As shown in Fig. 2C(2), this is basically consistent with our speculation. Except for the group generating one unique number of molecules as group 1, as the number of unique molecules generated increases, the frequency of the corresponding group's FG in the training set also increases, indicating that these

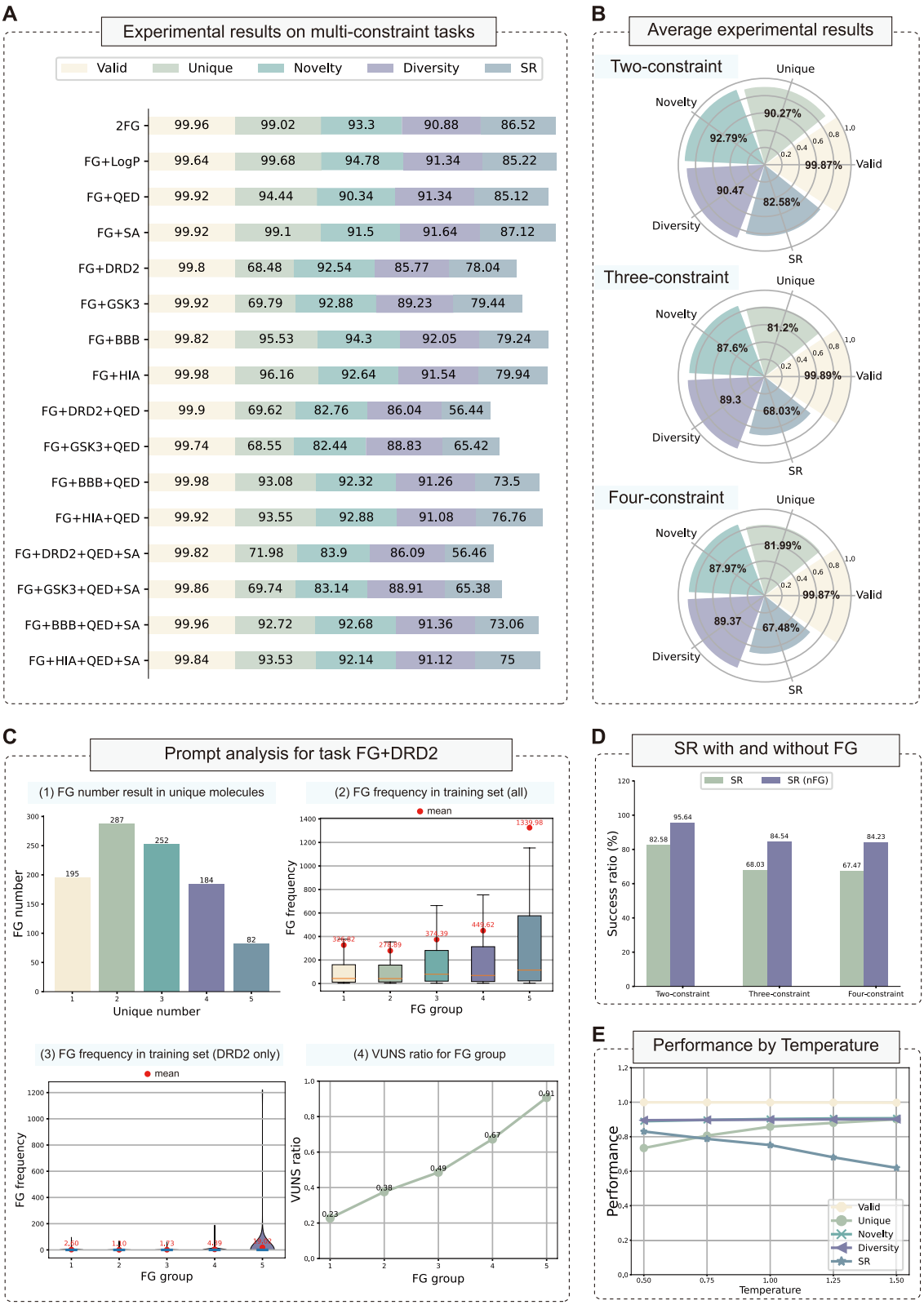

Fig. 2 **A** Experimental results for TSMMG across various tasks, encompassing 8 two-constraint tasks, 4 three-constraint tasks, and 4 four-constraint tasks. **B** Average experimental results on two-constraint, three-constraint, and four-constraint tasks. **C** FG analysis for task FG+DRD2. **D** Shows the comparison of the success ratio SR (nFG) without considering whether FG matches and the success ratio considering whether FG matches under different constraint tasks. **E** Shows the impact of different temperatures on the model

FGs can be better trained. The average frequency of the FGs in the group with one unique number is slightly higher than that of the group with a 2 unique number, and lower than the other groups, which we assume is an acceptable bias. We then checked the frequency of the FGs in the DRD2 related training set. As shown in Fig. 2C(3), a considerable portion of functional groups (FGs) did not appear in the training set related to DRD2. Despite this, our model still demonstrates the capability to generate correct molecules.

Further, we consider the ratio of molecules that simultaneously satisfy valid, unique, novelty, and success (VUNS) criteria generated by different groups. As shown in Fig. 2C(4), combined with Fig. 2C(2) and C(3), as the frequency of FGs in the training set increases, the model is more capable of generating more novel molecules that meet the constraints. In group 5, the average frequency of this group's FG in all training set is 1339, and the ratio of VUNS molecules generated by these FG-associated prompts is as high as 91%.

Given the significant impact of FG on the success ratio, we calculated the success ratio without considering FG matching, abbreviated as SR (nFG). For example, for the FG+DRD2+QED+SA task, SR (nFG) only considers whether DRD2, QED, and SA meet the constraints. The results are shown in Fig. 2D. It can be seen that in the two-constraint task, three-constraint task, and four-constraint task, the success ratio without considering FG is 13.06%, 16.51%, and 16.76% higher than the success ratio considering FG, respectively.

The above observations suggest that as more molecules and FGs are added to the training set, our model should be able to achieve more significant performance.

Effect of temperature on performance

During the inference process of large language models, temperature is a parameter of great interest. A lower temperature implies lower randomness, while a higher temperature means the model has greater freedom. We conducted tests on all tasks by setting different temperatures. Figure 2E shows the average performance of all tasks when the temperature is set to 0.5, 0.75, 1.0, 1.25, and 1.5, respectively. It can be observed that as the temperature increases, the ability to generate valid molecules remains virtually unchanged, still maintaining above 99%. Novelty and diversity also remain almost unchanged. However, unique and SR show a noticeable increase or decrease. Unique increases from 73.42 to 90.04%, an improvement of approximately 17%, indicating that as the temperature increases, the model can generate more unique molecules. At the same time, SR decreases from 83.03 to 61.94%, a reduction of about 21%. The decrease in SR is roughly consistent with the increase in unique, which means that although increasing the temperature from 0.5 to 1.5 generates 17% more unique molecules, most of them do not satisfy all constraint conditions.

Case study of a five-constraint molecular generation

Given the availability of corresponding predictors and a sufficiently large molecular library, it is theoretically feasible to construct training sets for any combination of desired molecular properties. This would enable the model to generate molecules that encompass a wide range of attributes. However, the challenge arises as the number of properties and their combinations increases, resulting in an exponential growth in the total number of possible property combinations. The exhaustive coverage of all these combinations becomes impractical. To address this challenge, we embarked on an investigation to determine if a model could effectively generate molecules when trained using individual properties but tested on arbitrary combinations. This research aimed to assess the model's adaptability to novel scenarios. To this end, we designed a task that entailed the generation of molecules exhibiting high drug-likeness, good synthetic accessibility, blood-brain barrier permeability, and the ability to bind to both the prostaglandin E2 receptor EP2 subtype [18] and prostaglandin E receptor EP4 [19]. The input prompt constructed for this task was: "The molecule exhibits a high QED score, good synthetic accessibility. It can pass through the blood-brain barrier and binds to both Prostanoid EP2 and EP4 receptors." During the training phase, each molecule was associated with only one property, meaning the model was exposed to molecules corresponding to each of the five properties within this prompt. However, the model had not encountered molecules that simultaneously met all five of these properties, and indeed, it had not seen molecules that met even two properties explicitly simultaneously.

This task presents a formidable challenge from multiple perspectives. From the model's input perspective, the model encounters significantly longer input text, a departure from its prior training data. In terms of molecular properties, the model must not only comprehend the mapping of individual properties to molecular spaces but also grasp the complex mapping of multiple properties from a single property mapping. Surprisingly, the model proves to be up to the task, successfully generating molecules that simultaneously satisfy all the condition constraints. As illustrated in Fig. 3, we showcase four molecules that meet all the property requirements specified in the textual description. To validate their compatibility with the receptors of EP2 (PDB ID: 7CX2) and EP4 (PDB ID: 5YWY), we employed UCSF Chimera [20] for molecular docking preparation and UCSF Dock6 [21] to conduct molecular docking. Finally, we used PLIP [22]

and PyMOL [23] for visualizing the docking results. The docking results reveal that these molecules effectively fit into the ligands and establish hydrogen bonds with different residues, demonstrating their potential for fulfilling the specified molecular properties.

This experimental outcome holds profound significance, as it demonstrates the model's robust capability to generate molecules that satisfy complex multi-constraints during zero-shot testing, even when initially trained with one-constraint data. This versatility underscores the model's adaptability and its potential to address intricate challenges in molecular generation.

Diversity of input text

Given that TSMMG is trained based on GPT-2 [24], which has undergone extensive pre-training on natural language datasets, we have a reasonable basis to hypothesize that TSMMG can comprehend the similarities in natural language. Specifically, when provided with prompts that share the same semantics but exhibit subtle differences in their expressions, TSMMG is likely to generate accurate molecules. This hypothesis stems from the fact that GPT-2 has acquired the ability to capture various linguistic patterns and semantic relationships during its training process. Consequently, it may possess the capability to generalize and transfer its knowledge to related but slightly different prompts. In essence, TSMMG's potential to generate correct molecules may persist even with prompt variations due to its underlying understanding of linguistic similarities.

To test this hypothesis, we explored the use of diverse templates that encompass different language habits and variations. By making slight modifications to the original training templates, we aimed to assess TSMMG's ability to generate correct molecules when input prompts were slightly altered. For example, during the training phase, we utilized the template "The molecule contains [FG], it can be absorbed by the human intestine." for the FG+HIA task. We introduced minor adjustments to create two new prompts: "I want a molecule that contains [FG] and can be absorbed by the human intestine." and "Give me a molecule which contains [FG] and can be absorbed by the human intestine." We conducted experiments using these diverse prompts across four different tasks, as presented in Table 2, and the results are summarized in Table 3.

The experiments demonstrated that TSMMG consistently generated molecules that met the specified requirements to a large extent, even with modified prompts. As shown in Table 2, the validity of the generated molecules can still reach over 99% after using prompts of different styles. For the FG+BBB and FG+HIA tasks, using the T1 and T2 templates both resulted in approximately a 9% decrease in SR compared to using the T0 template, while uniqueness, novelty, and diversity showed almost no significant changes. For the FG+DRD2 task, when using the T1 template, SR decreased by 33.36%, novelty decreased by 12.12%, while uniqueness increased by 3.36%; when using the T2 template, SR decreased by 30%, novelty decreased by 11.22%, while uniqueness increased by 1.58%. The FG+GSK3 task and the FG+DRD2 task show the same trend, that is, when using the T1 and T2 templates, SR and novelty show a significant decrease and uniqueness shows a certain degree of increase, while other indicators show relatively small differences.

These results suggest that TSMMG exhibits a certain degree of tolerance to diverse prompts and can continue to generate molecules that meet the specified requirements, even when the prompts are modified. It is important to note that while TSMMG may demonstrate tolerance to prompt variations, the extent of its ability to generalize and generate accurate molecules may vary depending on the specific prompt and task.

Discussion

TSMMG as producer

The development of TSMMG can be viewed as a form of knowledge distillation [25], as depicted in Fig. 4A. Initially, diverse molecular properties are obtained using teacher models. These properties are then encapsulated into natural language descriptions and combined with molecular sequences to create text-molecule pairs. TSMMG is trained using these text-molecule pairs as training data, enabling it to acquire the knowledge inherent in the properties through natural language. By leveraging this process, TSMMG becomes proficient in generating molecules that exhibit the desired properties. Notably, TSMMG has the ability to generate novel molecules possessing specific properties based on the

(See figure on next page.)

Fig. 3 **A** Docking reference for EP2 and EP4. **B** Molecules generated by TSMMG that can simultaneously bind to both EP2 and EP4 receptors. The input prompt is: "The molecule exhibits a high QED score, good synthetic accessibility. It can pass through the blood-brain barrier and binds to both Prostanoid EP2 and EP4 receptors." During training, TSMMG has seen molecules that can individually bind to both EP2 and EP4 receptors, but it has not explicitly received molecules that simultaneously satisfy all the constraints in this prompt. Nevertheless, it still successfully generates the desired molecules

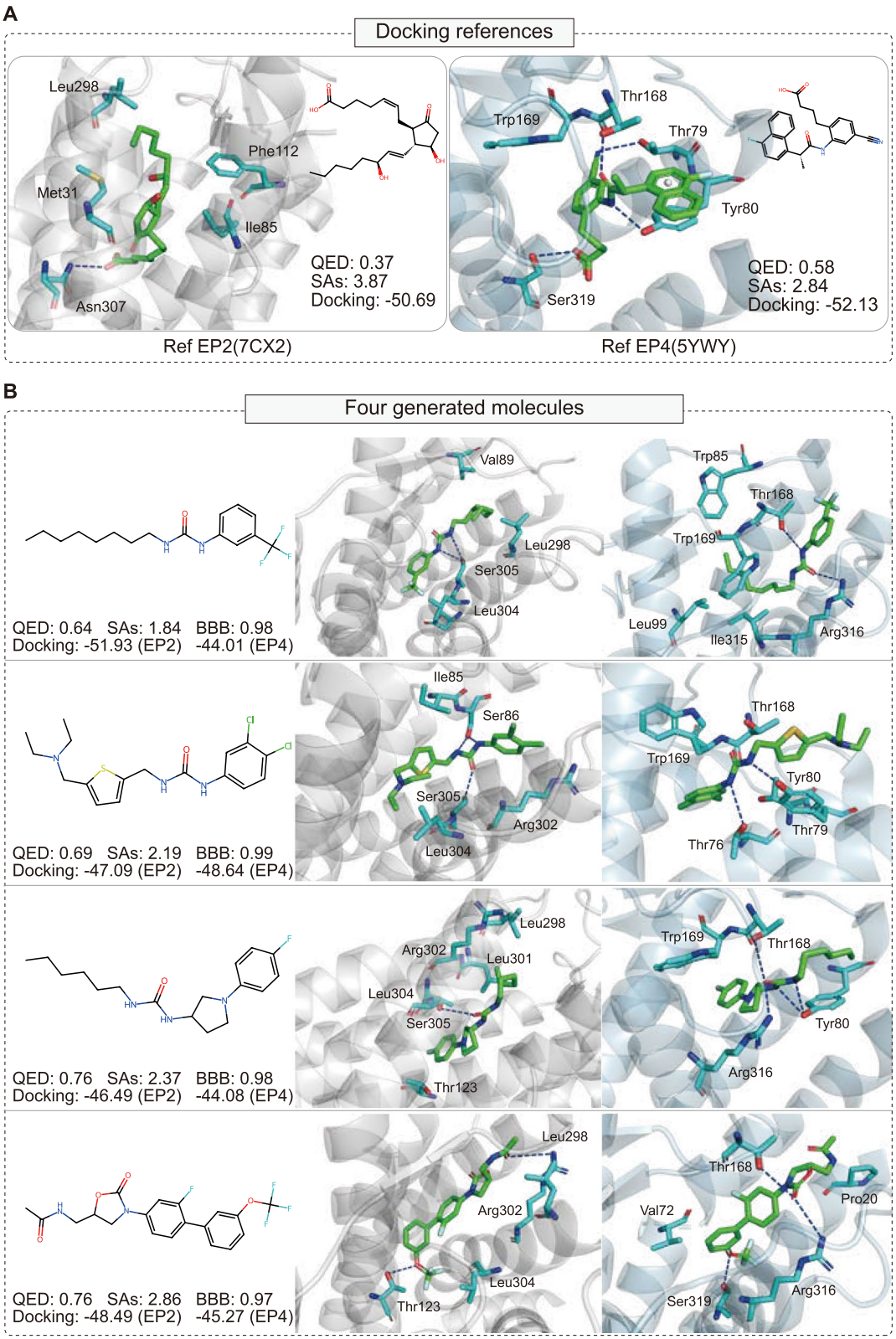

Fig. 3 (See legend on previous page.)

Table 2 Prompts we used in order to test the tolerance of TSMMG to diverse inputs, [FG] refers to any functional group

DRD2/GSK3	T0	The molecule contains [FG]. It is active to DRD2. [D2D2/GSK3]
	T1	I want a molecule that contains [FG] and can bind to [D2D2/GSK3]
	T2	Give me a molecule which contains [FG] and can bind to [D2D2/GSK3]
BBB	T0	The molecule contains [FG]. It can pass through the blood-brain barrier
	T1	I want a molecule that contains [FG] and can pass through the blood-brain barrier
	T2	Give me a molecule which contains [FG] and can pass through the blood-brain barrier
HIA	T0	The molecule contains [FG]. It can be absorbed by human intestine
	T1	I want a molecule that contains [FG] and can be absorbed by human intestine
	T2	Give me a molecule which contains [FG] and can be absorbed by human intestine

Table 3 Experimental results with different template styles

Template	Task	Valid	Unique	Novelty	Diversity	SR	SR (nFG)
T0	FG+DRD2	99.80%	68.48%	92.54%	85.77%	78.04%	93.18%
	FG+GSK3	99.92%	69.79%	92.88%	89.23%	79.44%	94.40%
	FG+BBB	99.82%	95.53%	94.30%	92.05%	79.24%	96.14%
	FG+HIA	99.98%	96.16%	92.64%	91.54%	79.94%	95.80%
T1	FG+DRD2	99.70%	71.84%	80.42%	86.26%	44.68%	82.12%
	FG+GSK3	99.68%	73.49%	83.90%	89.56%	47.36%	84.88%
	FG+BBB	99.68%	94.30%	95.48%	92.28%	70.64%	96.84%
	FG+HIA	99.78%	94.11%	94.82%	91.99%	69.34%	93.86%
T2	FG+DRD2	99.80%	70.06%	81.32%	86.18%	48.04%	83.52%
	FG+GSK3	99.68%	71.89%	84.68%	89.52%	50.22%	85.60%
	FG+BBB	99.74%	95.67%	95.40%	92.17%	70.72%	96.52%
	FG+HIA	99.90%	95.24%	95.48%	91.86%	71.10%	94.12%

acquired knowledge. This offers a feedback loop to the teacher models, allowing them to refine and update their knowledge.

To illustrate this, we conducted further experiments involving serine/threonine-protein kinase D3 (KPCD3), Bruton's tyrosine kinase (BTK), fibroblast growth factor receptor 4 (FGFR4), and papain-like protease 3CL. Initially, each target dataset is randomly divided into training and test sets. Subsequently, a random subset is partitioned from the training dataset to serve as the validation set, and an SVM predictor is trained using the training set and validated using the validation set. This process is repeated 100 times to select the best predictor. Finally, the chosen predictor is applied to the test set to obtain the F1 score. For comparison, different numbers of molecules generated by TSMMG that can bind to the corresponding target are randomly added to the training set to train new SVM predictors. This process is also repeated 100 times to yield consistent statistical data. These added molecules are referred to as pseudo-samples. The experimental results presented in Fig. 4B demonstrate that the addition of pseudo-samples significantly enhances the performance of the predictors. Notable improvements are observed in KPCD3, BTK, FGFR4, and 3CL by approximately 13%, 4%, 17%, and 7%, respectively. Furthermore, as the number of pseudo-samples increases, the performance of each predictor tends to converge. These results indicate that TSMMG can discern the commonalities shared by molecules with specific properties and generate novel molecules that embody these commonalities. Moreover, TSMMG's unique ability to learn from teacher models and provide feedback for updating the knowledge of these models initiates a symbiotic relationship that promotes continuous improvement in their respective capabilities.

Comparison with other methods

Although existing commercial or open-source LLMs such as GPT-4 [26] and Llama [27] perform well on various natural language tasks and exhibit some molecular generation capabilities, they struggle to consistently generate high-quality, novel molecules, especially under multiple constraints. This is primarily because they have not been fine-tuned on specialized molecule datasets. Consequently, we did not consider vanilla LLMs as baselines. Additionally, similar models like ChemLLM

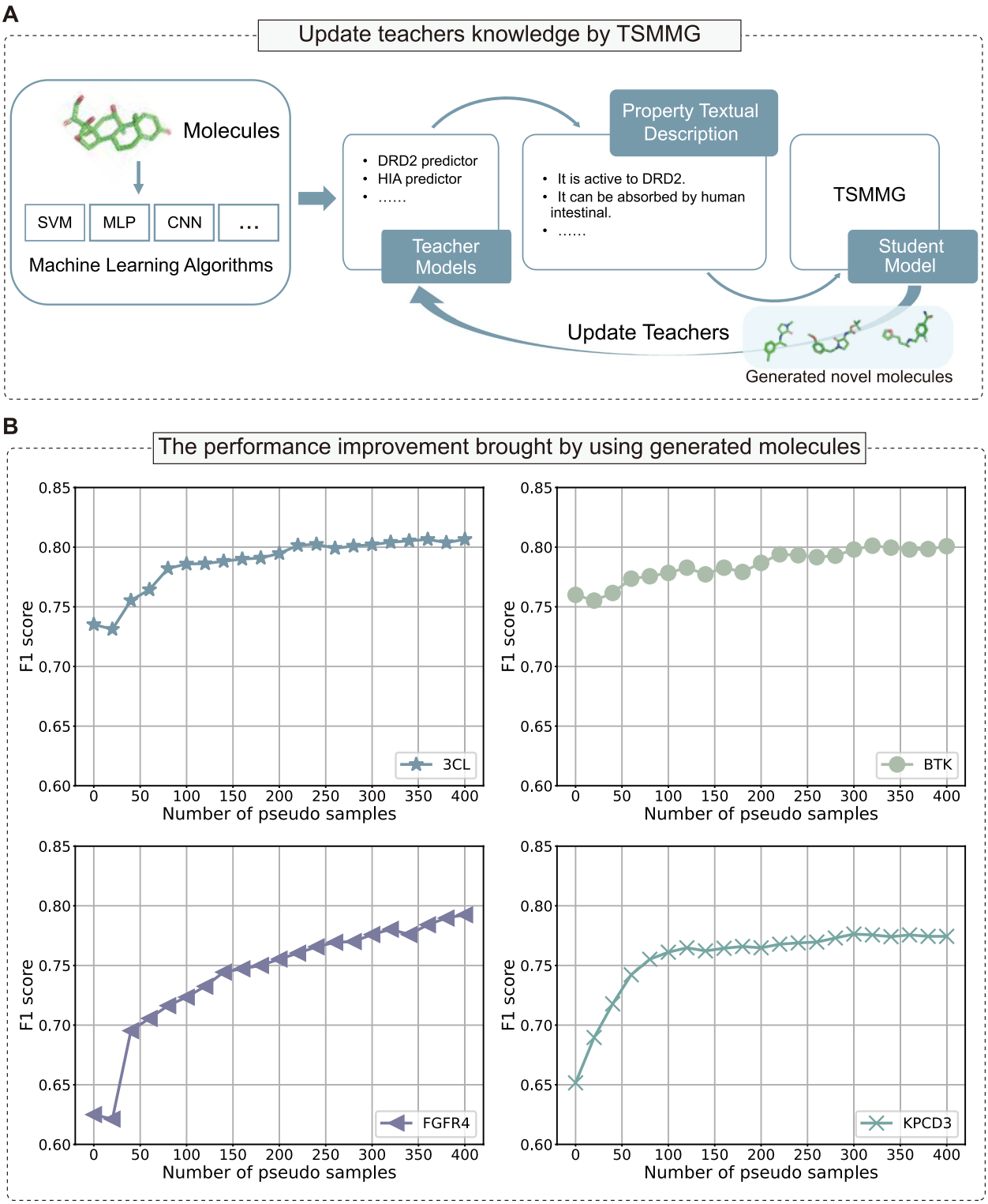

Fig. 4 TSMMG as a producer. Molecules generated by TSMMG can be used to improve the accuracy of the predictor. **A** We leverage a large number of property predictors, which can be regarded as teacher models, to obtain molecular properties, and then use these properties to construct textual descriptions to train TSMMG. The molecules generated by TSMMG can also be used to update the corresponding property predictors. This has two benefits: firstly, it allows us to verify whether TSMMG has effectively extracted the latent representation of the property, and secondly, it can improve the accuracy of the property predictors. **B** The experimental results on four property predictors are shown in the figure. The horizontal axis represents the number of generated molecules added to the training data, which we refer to as pseudo-samples. As can be observed, the accuracy of the property predictors increases and tends to converge as the number of pseudo-samples increases

[28] and ChemCrow [29] are also unsuitable for multi-constraint molecular generation tasks. To better demonstrate the capabilities of TSMMG, we present two sets of comparisons in Additional file 1: (1) using a larger open-source LLM, Llama2, fine-tuned with LoRA [30], as the backbone; and (2) comparing traditional multi-constraint molecular generation methods, including Reinvent [31], Reinvent2 [31], and MCMG [12]. We utilized the reinforcement learning method DPO [32] to further fine-tune TSMMG in order to achieve a relatively fairer comparison. The results of Reinvent, Reinvent2, MCMGL, and MCMGM are quoted from [33].

Conclusions

TSMMG presents an effective method for utilizing natural language to generate molecules with multiple constraints. By employing knowledge distillation, the property prediction capabilities of specialized molecular models are transferred to enable LLMs to generate molecules with specific properties. This transforms the molecule screening process into a molecule generation process, allowing exploration of a larger molecular space. We demonstrate TSMMG's exceptional molecular generation capabilities through tasks with two, three, and four constraints, including structure, physicochemical properties, affinity, and ADMET. TSMMG can generate novel molecules that meet specified requirements based on natural language descriptions. Additionally, TSMMG exhibits zero-shot generation capabilities. TSMMG shows potential not only in drug discovery but also in other molecule-related fields, such as material discovery.

On one hand, the capabilities of TSMMG can be enhanced as the capabilities of the teacher model improve. On the other hand, the novel molecules generated by TSMMG have the potential to further enhance the teacher model's capabilities. Our future work will focus on leveraging more advanced teacher models to improve TSMMG's capabilities and enabling TSMMG to generate molecules with a wider range of properties.

Methods

Problem setting

Natural language serves as a user-friendly means for human-machine interaction, making it an ideal solution for generating molecules from natural language descriptions. Recent successes in the development of large language models (LLMs) [26, 34] inspire the vision that we may achieve the generation of molecules from diverse molecular spaces by simply modifying the input prompt. This approach offers a promising solution to address the challenge of generality in molecular generation. Despite both natural language and SMILES being sequence data formats, SMILES can be viewed as a specialized molecular language that can be challenging for humans to interpret. From this perspective, generating molecular sequences from natural language descriptions can be regarded as a translation task, an area where LLMs excel.

Given a natural language sequence $W = \{w_1, w_2, ..., w_n\}$, the objective is to generate a corresponding molecule represented by a SMILES sequence, $S = \{s_1, s_2, ..., s_m\}$, which can be formulated as conditional probability $P(S|W)$.

In order to ensure the quality of the generated molecules, it is imperative to adhere to the following prerequisites: (1) **Validity**: The generated SMILES representation, S, should strictly adhere to the syntax rules of the SMILES format, guaranteeing that it forms a valid and well-structured molecule. (2) **Relevance**: The molecule represented by S should accurately reflect the physical and chemical properties described by the natural language sequence W. This entails that if there exists a subsequence $W_{i,j} = \{w_i, w_{i+1}, \ldots, w_j\}$ in W that specifies a particular property, there should be a corresponding subsequence $S_{k,l} = \{s_k, s_{k+1}, \ldots, s_l\}$ in S that satisfies the desired property. (3) **Diversity**: While satisfying the validity and relevance criteria, the generated S should exhibit diversity. In other words, the generated molecules should not be identical or overly similar, providing a range of molecular structures that fulfill the given properties. (4) **Novelty**: The model should possess the ability to generate S that are not present in the training set. This capability ensures that the generated molecules introduce new and previously unseen chemical structures, thereby expanding the exploration space beyond the confines of the training data.

The quandary of translating natural language into molecular language, albeit bearing resemblances to conventional machine translation, poses distinctive challenges. In this context, three fundamental patterns of correspondence between natural language and molecular sequences can be discerned: (1) **One-to-one mapping**: In this pattern, a specific text description corresponds to a single, specific molecular sequence. Models like MolT5 [35], MolXPT [36], and MoleculeSTM [37] have tackled this problem as a query task, aiming to establish a direct mapping relationship between text and molecular sequences. However, this approach may not be ideal for generating novel molecules with diverse properties, as it relies on a fixed ground truth and does not explore beyond the known data. (2) **One-to-many mapping**: Here, a text description can correspond to multiple different molecular sequences. This pattern allows the model to learn the feature distribution of the target space, enabling sampling from the distribution to generate new molecules. Models like those proposed by Kotsias et al. [38] and Wang et al. [12] leverage this pattern effectively

by training on specific datasets containing molecules with shared properties which implicitly embracing the one-to-many mapping pattern. (3) **Many-to-one mapping**: In this pattern, a specific molecular sequence can be described in various ways. By understanding the inherent relationship between different attributes, it is possible to discover new properties of a molecule. This pattern offers opportunities for exploring diverse attributes of molecules beyond their known properties. In order to develop a universal molecular generative model capable of generating molecules with various desired properties without the need for retraining, it is essential to accumulate a substantial amount of data that explicitly adheres to the one-to-many and many-to-one mapping patterns. The primary challenge lies in acquiring a sufficient number of text-molecule pairs in a rapid, convenient, and cost-effective manner.

Data generation framework

Several studies have explored the integration of natural language and molecular language. MolT5 [35] aimed to achieve bidirectional translation between natural language and molecular language. The model underwent initial pre-training on an extensive collection of unpaired natural language corpora and molecular sequences, followed by fine-tuning on the text-molecule paired dataset ChEBI-20. However, ChEBI-20 presents two notable limitations. Firstly, it contains a relatively small set of 33,010 text-molecule pairs, making it challenging to establish the correspondence between natural language and molecular language. Secondly, the text descriptions in this dataset, sourced from the comment field in ChEBI [39], often contain information unrelated to molecular properties. Additionally, these descriptions exhibit a strong one-to-one relationship with the molecules, posing challenges for the model to explore the specific molecular space associated with desired properties. MolXPT [36] proposed a method that involves incorporating molecular sequences within the input text for large language models (LLMs). CLAMP [40] introduced a fusion approach, combining a molecule encoder and a text encoder for property prediction tasks. Christofidellis et al. [41] presented a unified model capable of handling various text-to-text, text-to-molecule, molecule-to-text, and molecule-to-molecule tasks. MolReGPT [42] implemented tasks such as molecule captioning and text-based molecule generation by assigning ChatGPT a role as a biochemist, facilitating in-context learning.

However, a common limitation in all of the above-mentioned studies is their reliance on the ChEBI dataset, which constrains their performance due to data scarcity and quality issues. As of now, limited research efforts have been directed at addressing these issues in natural language-based molecular generation. Therefore, we propose a knowledge distillation-based approach to construct an extensive and high-quality dataset of natural language-molecule pairs.

Figure 1A provides an overview of the framework employed for the creation of our dataset. The underlying concept revolves around the utilization of advanced molecular parsing tools and models to extract knowledge related to molecules. Subsequently, this acquired knowledge is transformed into natural language text, resulting in paired data comprising molecules and their corresponding textual descriptions. Within this framework, the tools and models responsible for extracting molecular knowledge are collectively referred to as "teachers," while TSMMG assumes the role of the "student." TSMMG undertakes the task of learning various properties associated with molecules from these "teachers." It also comprehends the mapping relationship between these properties and the molecular structures themselves. This knowledge empowers TSMMG to generate new molecules based on specified properties using natural language descriptions.

Within this framework, multiple "teachers" are employed, each with distinct capabilities related to molecular properties and structures. These teachers encompass a range of tools and models, including:

- Physicochemical property teacher: RDKit, a tool capable of parsing molecules to extract physicochemical properties such as molecular weight (MW), the number of aromatic rings (AROM), LogP, SA, QED, the number of hydrogen bond acceptors (HBA), the number of hydrogen bond donors (HBD), and topological molecular polar surface area (PSA).
- ADMET property prediction models: admetSAR [43], based on support vector machine (SVM), predicts ADMET properties, such as blood-brain barrier permeability and absorptivity.
- Affinity prediction models: Olivecrona et al.'s SVM-based models [44] and Jin et al.'s models [45] can predict the binding probabilities of molecules to specific targets, including DRD2, GSK3, and JNK3. Newer models such as MolTrans [46], DrugBAN [47], and TransformerCPI [48] are designed to predict the affinity of small molecules to receptor proteins and more.
- Structural information extraction: In addition to these property-related teachers, the IUPAC name of a molecule, which bears structural information, is considered. The IUPAC name exhibits a grammar resembling natural language and provides standardized descriptions of molecules. By breaking down IUPAC names, it is possible to extract

structural components of a molecule. For instance, deconstructing the molecule "(2-methyl-5-methylsulfonylphenyl)methanamine" yields the functional groups "methyl," "methylsulfonylphenyl," and "methanamine." Therefore, an IUPAC parser is proposed, along with a set of rules for dissecting IUPAC names, serving as an additional "teacher" for extracting the internal structure of molecules.

Through these "teachers," we acquire extensive knowledge about molecules, including their structural information, physicochemical properties, and binding affinities to specific receptors. This information is then transcribed into natural language descriptions and combined with the corresponding molecules to create text-molecule pairs. For an example as shown in Fig. 1B, let us consider a molecule represented as "CCN1CCCC1CNC(=O)c1c(OC)ccc(Cl)c1O." We can break down its IUPAC name to extract the functional group "methoxybenzamide." By utilizing RDKit, we determine its LogP, QED, and SAs. We further predict its affinity with DRD2 through a classifier proposed by Olivecrona et al. [44] and evaluate its likelihood of passing through the blood-brain barrier using admetSAR. These various properties are then associated with the molecule using natural language templates. The data generation method offers several notable advantages: (1) With numerous publicly accessible molecular databases like PubChem [49] and ZINC [50], our approach allows for the rapid acquisition of a large number of text-molecule pairs. This effectively overcomes the data limitations often encountered in natural language-based molecular generation models; (2) The text molecule pairs generated through this method exhibit a high degree of relevance. Each segment of text contains certain properties of the molecules, enabling the model to learn the mapping relationship between text descriptions and molecular properties more effectively; (3) There is a wealth of advanced tools and models available for molecular structure analysis and property prediction. Our framework simplifies the process of transferring knowledge from these advanced tools and models into a student model in natural language form. This empowers the student model to generate molecules that incorporate this knowledge. (4) The method is highly scalable, allowing for the seamless transfer of knowledge for various molecular properties. It can be applied to an array of properties, making it versatile for different research needs. (5) Our method supports continuous knowledge updates. This means that the student model can benefit from the latest and more robust models, ensuring that it remains up-to-date and well-informed.

Training model

We began by collecting 2 million molecules from PubChem. Subsequently, we harnessed the tools and models mentioned earlier to extract comprehensive knowledge regarding these molecules. This knowledge was then translated into natural language text using predefined templates and combined with the corresponding SMILES representations. To maximize the model's capabilities, we thoughtfully organized the data to encompass both one-to-many and many-to-one patterns. This approach ensures that the model learns the underlying distribution of specific inputs, promoting adaptability and preventing the mere memorization of fixed responses. For instance, let us take molecule M, which possesses ten pieces of extracted knowledge. If we were to compile all ten pieces into a single text, denoted as T, the resulting molecule space associated with T would likely be highly restricted, potentially corresponding to just one specific molecule, let us say, molecule A. This would essentially create a one-to-one data pattern. To overcome this limitation, we adopt a strategy where, for each molecule, we select a subset of its knowledge to compose the text. The goal here is to craft this text in a way that it corresponds to as many molecules as possible. This strategy empowers the model to gain insights into the broader distribution of molecules linked to the provided text, rather than locking it into a specific, isolated instance. Then, the training of TSMMG involves two key steps: pre-training on a large natural language corpus and fine-tuning on text-molecule paired data that we have constructed. In the first stage, TSMMG undergoes pre-training on a large natural language corpus. This enables TSMMG to learn and understand natural language by capturing the statistical patterns and linguistic structures present in the data. The pre-training stage helps TSMMG acquire a general understanding of language and forms the initial foundation for subsequent training stages. The second stage involves a fine-tuning on the text-molecule paired data that contains descriptions of various properties as shown in Table 3. This fine-tuning stage focuses on teaching TSMMG the mapping between text descriptions and molecular sequences as well as the syntax of SMILES. By fine-tuning TSMMG on this specific dataset, it becomes proficient in generating molecules based on specific text-described-property such as functional groups, LogP, physicochemical properties, drug-like properties, and affinity scores to certain targets. The architecture of TSMMG is the same as GPT [24]. TSMMG follows the settings of GPT2small, which consists of 12 layers and has a total of 117 million parameters. We downloaded the weights of GPT2small from Huggingface model repository [51] to initialize TSMMG. This helps with cost and computational considerations

by leveraging pre-trained weights for an efficient starting point. And since the weights are trained by a large number of language corpus, we can directly fine-tune the model using the text-molecule paired data we construct. We fine-tune TSMMG on 8 A100 40G GPUs for around 6 days. We use the subsequent hyperparameters: a batch size of 32, a learning rate set to 5e−4, a warmup steps of 100. We use AdamW [52] as the optimizer.

Metrics

To evaluate the performance of the TSMMG model, we employed four common metrics in molecular generation: validity, uniqueness, novelty, and diversity. Each of these metrics was essential for a comprehensive evaluation: Validity assesses whether the generated molecules conform to the syntax rules of SMILES. We utilized RDKit [53] to parse the generated molecules, considering them valid if the parsing process was successful. Uniqueness measures the proportion of non-repetitive molecules among the generated set. It ensures that the model produces diverse molecules. Novelty signifies whether the generated molecules are previously unseen in the training dataset, preventing the model from regenerating known molecules. Diversity describes the structural differences between generated molecules, it is calculated as:

$$Diversity = 1 - \frac{2}{n(n-1)\sum_{X,Y} sim(X,Y)},$$

where sim(X,Y) is calculated based on the Tanimoto distance with respect to the Morgan fingerprints of generated molecules X and Y. In addition to these standard metrics, we introduced the concept of success ratio (SR) to measure whether the generated molecules meet predefined conditions. We establish different criteria to define the success of generated molecules based on the specific task. These criteria are outlined as follows:

- FG: We leveraged IUPAC nomenclature to identify functional groups within the molecules. By parsing IUPAC names and matching them to generated SMILES-encoded molecules, we checked if the generated molecules contained the specified functional groups.
- LogP: Using RDKit, we calculated the LogP values of the generated molecules and compared them to predefined values. The generation was considered successful if the LogP value fell within a margin of 1 from the specified value.
- QED and SAs: For these tasks, we adopted criteria similar to prior work [12], considering QED as high if its value exceeded 0.6 and SAs as good if the score was less than 4.
- DRD2 and GSK3: We employed the models proposed by Jin et al. [45] to predict the affinity scores of the generated molecules. A molecule was considered successful if its corresponding affinity score exceeded 0.5 for either target.
- BBB and HIA: We used models developed by Cheng et al. [43] to predict scores, determining if a molecule could pass through the blood-brain barrier (BBB) if its BBB score exceeded 0.5 or if it could be absorbed by the human small intestine (HIA) if its HIA score was above 0.5.

Moreover, for each multi-constraint task, we only considered molecules successful if they met all constraints contained in this task simultaneously. We generated 5000 molecules to evaluate the model's performance for each multi-constraint task. Note that we uniformly express all metric results in percentage format. While converting diversity to a percentage may lack intrinsic meaning, for the sake of ease of comparison with other metrics, we multiply it by 100. However, we refrain from appending the "%" symbol to distinguish it from other metrics.

Translating SMILES to IUPAC

There are several open works that provided their solutions for translating SMILES to IUPAC name, such as STOUT [54] and IUPAC2Struct [55], but the interfaces they released are not for high throughput experiments. Considering experimental efficiency, we trained our own SMILES2IUPAC model based on GPT2small. We formulate this problem also as conditional probability $P(I|S)$ where generating a corresponding IUPAC name $I = \{i_1, i_2, \ldots, i_n\}$ by a given SMILES sequence $S = \{s_1, s_2, \ldots, s_m\}$. We collect 2 million SMILES-IUPAC paired data from PubChem to train this model. The model size and settings of SMILES2IUPAC are the same as TSMMG. For evaluating the trained SMILES2IUPAC model, we pass 1000 unseen molecules to it and generate 1000 corresponding predicted IUPAC names. We then break down the predicted IUPAC names to identify the functional groups, and check if all these functional groups exist in the corresponding ground truth IUPAC names. The experimental results show an accuracy rate of 94%.

Abbreviations

TSMMG	Teacher-student-based multi-constraint molecular generation model
AIDD	Artificial intelligence for drug discovery
QED	Quantitative estimate of drug-likeness
LogP	Molecular hydrophobicity
SA	Synthetic accessibility
FG	Functional group
DRD2	Dopamine type 2 receptor
GSK3	Glycogen synthase kinase-3 beta
BBB	Blood-brain barrier

HIA	Human intestinal absorption
KPCD3	Serine/threonine-protein kinase D3
BTK	Bruton's tyrosine kinase
FGFR4	Fibroblast growth factor receptor 4
3CL	Papain-like protease 3CL
SMILES	Simplified Molecular Input Line Entry System

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-025-02200-3.

Additional file 1: Figure S1-A compares the use of GPT-2 and Llama-7b as backbone LLMs; Figure S1-B compares TSMMG with common multi-constraint molecular generation models, including Reinvent, Reinvent2, and MCMG. Figure S2 showcases some molecules generated by TSMMG. Figure S3 illustrates the QED distribution of molecules generated under different conditions. Figure S4 displays the LogP distribution of molecules generated with specified LogP values. Figure S5 shows the weight of each token during the molecular generation process. Figure S6 demonstrates the scaffold similarity among the generated molecules. Figure S7 presents the convergence of models using different learning rates. Figure S8 highlights cases of generation failures. Figures S9, S10, and S11 respectively show docking examples of molecules generated for targets 3CL, BTK, and FGFR4. Figure S12 further illustrates the impact of FG frequency on the uniqueness of generated molecules. Table S1 explains the relationships between the metrics used. Tables S2, S3, and S4 provide additional experimental parameter information.

Additional file 2. Contains the source data corresponding to the figures and tables.

Authors' contributions

L.W., YS.L., and X.Z. conceived the study of TSMMG and the experimental assays. P.Z. developed TSMMG. P.Z., J.W., C.L., and Z.W. performed all experiments. X.Z., YS.L., L.W., and P.Z. drafted the manuscript and charts. Y.L., C.L., S.S., J.L., L.W., X.C., H.L., and W.L. critically revised the manuscript. All authors critically revised and gave final approval of the manuscript.

Funding

This work was supported by the National Science and Technology Major Project (2023ZD0120902), the National Natural Science Foundation of China (U22A2037, 62425204, 62122025, 62450002, 62432011, 62372159), and the Beijing Natural Science Foundation (L248013).

Data availability

Source data and codes are provided with this paper. The data and source code used in this project are freely available at GitHub (https://github.com/HHW-zhou/TSMMG) and Zenodo (https://doi.org/10.5281/zenodo.15093636).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 10 October 2024 Accepted: 27 March 2025
Published online: 23 April 2025

References

1. Schwalbe-Koda D, Gómez-Bombarelli R. Generative models for automatic chemical design. Machine Learning Meets Quantum Physics. 2020. p. 445–467.
2. Gainza P, Sverrisson F, Monti F, Rodola E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods. 2020;17(2):184–92.
3. Wójcikowski M, Kukiełka M, Stepniewska-Dziubinska MM, Siedlecki P. Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. Bioinformatics. 2019;35(8):1334–41.
4. Mahmoud AH, Masters MR, Yang Y, Lill MA. Elucidating the multiple roles of hydration for accurate protein-ligand binding prediction via deep learning. Commun Chem. 2020;3(1):19.
5. Jones D, Kim H, Zhang X, Zemla A, Stevenson G, Bennett WD, et al. Improved protein-ligand binding affinity prediction with structure-based deep fusion inference. J Chem Inf Model. 2021;61(4):1583–92.
6. Cheng F, Wang F, Tang J, Zhou Y, Fu Z, Zhang P, et al. Artificial intelligence and open science in discovery of disease-modifying medicines for Alzheimer's disease. Cell Rep Med. 2024;5(2):101379.
7. Qiu Y, Cheng F. Artificial intelligence for drug discovery and development in Alzheimer's disease. Curr Opin Struct Biol. 2024;85:102776.
8. Zang C, Wang F. Moflow: an invertible flow model for generating molecular graphs. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. Virtual Event, CA, USA: Association for Computing Machinery; 2020. p. 617–26.
9. Kuznetsov M, Polykovskiy D. MolGrow: a graph normalizing flow for hierarchical molecular generation. In: Proceedings of the AAAI conference on artificial intelligence. Virtual Event: Association for the Advancement of Artificial Intelligence; 2021. p. 8226–34.
10. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038–40.
11. Gottipati SK, Sattarov B, Niu S, Pathak Y, Wei H, Liu S, et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In: International conference on machine learning. Online: PMLR; 2020. p. 3668–79.
12. Wang J, Hsieh CY, Wang M, Wang X, Wu Z, Jiang D, et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. Nat Mach Intell. 2021;3(10):914–22.
13. Xie Y, Shi C, Zhou H, Yang Y, Zhang W, Yu Y, et al. MARS: Markov molecular sampling for multi-objective drug discovery. In: International conference on learning representations. Virtual Event: International Conference on Learning Representations; 2021.
14. Li Y, Zhang L, Liu Z. Multi-objective de novo drug design with conditional graph generative model. J Cheminformatics. 2018;10:1–24.
15. Jin W, Barzilay R, Jaakkola T. Multi-objective molecule generation using interpretable substructures. In: International conference on machine learning. Online: PMLR; 2020. p. 4849–59.
16. Bagal V, Aggarwal R, Vinod P, Priyakumar UD. MolGPT: molecular generation using a transformer-decoder model. J Chem Inf Model. 2021;62(9):2064–76.
17. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–6.
18. Qu C, Mao C, Xiao P, Shen Q, Zhong YN, Yang F, et al. Ligand recognition, unconventional activation, and G protein coupling of the prostaglandin E2 receptor EP2 subtype. Sci Adv. 2021;7(14):eabf1268.
19. Toyoda Y, Morimoto K, Suno R, Horita S, Yamashita K, Hirata K, et al. Ligand binding to human prostaglandin E receptor EP4 at the lipid-bilayer interface. Nat Chem Biol. 2019;15(1):18–26.
20. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera-a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605–12.
21. Allen WJ, Balius TE, Mukherjee S, Brozell SR, Moustakas DT, Lang PT, et al. DOCK 6: impact of new features and current docking performance. J Comput Chem. 2015;36(15):1132–56.

22. Adasme MF, Linnemann KL, Bolz SN, Kaiser F, Salentin S, Haupt VJ, et al. PLIP 2021: expanding the scope of the protein-ligand interaction profiler to DNA and RNA. Nucleic Acids Res. 2021;49(W1):W530–4.
23. Schrödinger. Pymol. https://github.com/schrodinger/pymol-open-source. Accessed 19 Aug 2024.
24. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019;1(8):9.
25. Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: a survey. Int J Comput Vis. 2021;129(6):1789–819.
26. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. 2023. Preprint at https://arxiv.org/abs/2303.08774.
27. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: open and efficient foundation language models. 2023. Preprint at https://arxiv.org/abs/2302.13971.
28. Di Z, Wei L, Qian T, Jingdan C, Hang Y, Yuliang Y, et al. ChemLLM: a chemical large language model. 2024. Preprint at https://arxiv.org/abs/2302.13971.
29. M Bran A, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6:525–35.
30. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Lora: low-rank adaptation of large language models. ICLR. 2022;1(2):3.
31. Blaschke T, Arús-Pous J, Chen H, Margreitter C, Tyrchan C, Engkvist O, et al. REINVENT 2.0: an AI tool for de novo drug design. J Chem Inf Model. 2020;60(12):5918–22.
32. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. Adv Neural Inf Process Syst. 2023;36:53728–41.
33. Wang J, Hsieh CY, Wang M, Wang X, Wu Z, Jiang D, et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning, Table 3. 2021. https://www.nature.com/articles/s42256-021-00403-1/tables/3. Accessed 19 Aug 2024.
34. OpenAI. Introducing ChatGPT. 2022. https://openai.com/index/chatgpt/. Accessed 19 Aug 2024.
35. Edwards C, Lai T, Ros K, Honke G, Cho K, Ji H. Translation between molecules and natural language. In: Proceedings of the 2022 conference on empirical methods in natural language processing. Abu Dhabi: Association for Computational Linguistics; 2022. p. 375–413.
36. Liu Z, Zhang W, Xia Y, Wu L, Xie S, Qin T, et al. MolXPT: wrapping molecules with text for generative pre-training. In: Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 2: short papers). Toronto: Association for Computational Linguistics; 2023. p. 1606–16.
37. Liu S, Nie W, Wang C, Lu J, Qiao Z, Liu L, et al. Multi-modal molecule structure-text model for text-based retrieval and editing. Nat Mach Intell. 2023;5(12):1447–57.
38. Kotsias PC, Arús-Pous J, Chen H, Engkvist O, Tyrchan C, Bjerrum EJ. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. Nat Mach Intell. 2020;2(5):254–65.
39. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2007;36(suppl_1):D344–D350.
40. Seidl P, Vall A, Hochreiter S, Klambauer G. Enhancing activity prediction models in drug discovery with the ability to understand human language. In: International conference on machine learning. Hawaii: PMLR; 2023. pp. 30458–90.
41. Christofidellis D, Giannone G, Born J, Winther O, Laino T, Manica M. Unifying molecular and textual representations via multi-task language modelling. In: International conference on machine learning. Hawaii: PMLR; 2023. p. 6140–57.
42. Li J, Liu Y, Fan W, Wei XY, Liu H, Tang J, et al. Empowering molecule discovery for molecule-caption translation with large language models: a chatgpt perspective. IEEE Trans Knowl Data Eng. 2024;36(11):6071–83.
43. Cheng F, Li W, Zhou Y, Shen J, Wu Z, Liu G, et al. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. J Chem Inf Model. 2012;52(11):3099–3105.
44. Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Cheminformatics. 2017;9:1–14.
45. Jin W, Barzilay R, Jaakkola T. Multi-objective molecule generation using interpretable substructures. Int Conf Mach Learn. 2020:4849–59.
46. Huang K, Xiao C, Glass LM, Sun J. MolTrans: molecular interaction transformer for drug-target interaction prediction. Bioinformatics. 2021;37(6):830–6.
47. Bai P, Miljković F, John B, Lu H. Interpretable bilinear attention network with domain adaptation improves drug-target prediction. Nat Mach Intell. 2023;5(2):126–36.
48. Chen L, Tan X, Wang D, Zhong F, Liu X, Yang T, et al. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics. 2020;36(16):4406–14.
49. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373–80.
50. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, et al. ZINC20-a free ultralarge-scale chemical database for ligand discovery. J Chem Inf Model. 2020;60(12):6065–73.
51. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Online: Association for Computational Linguistics; 2020. p. 38–45.
52. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: International conference on learning representations. New Orleans: International Conference on Learning Representations; 2019.
53. Landrum G. Rdkit documentation. Release. 2013;1(1–79):4.
54. Rajan K, Zielesny A, Steinbeck C. STOUT: SMILES to IUPAC names using neural machine translation. J Cheminformatics. 2021;13(1):34.
55. Krasnov L, Khokhlov I, Fedorov MV, Sosnin S. Transformer-based artificial neural networks for the conversion between chemical notations. Sci Rep. 2021;11(1):14798.

Publisher's Note