Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2023, **14**, 1557

All publication charges for this article have been paid for by the Royal Society of Chemistry

SDEGen: learning to evolve molecular conformations from thermodynamic noise for conformation generation†

Haotian Zhang,‡[a] Shengming Li,‡[b] Jintu Zhang,[ac] Zhe Wang,[a] Jike Wang,[ad] Dejun Jiang,[a] Zhiwen Bian,[a] Yixue Zhang,[a] Yafeng Deng,[e] Jianfei Song,[e] Yu Kang *[a] and Tingjun Hou *[ac]

Generation of representative conformations for small molecules is a fundamental task in cheminformatics and computer-aided drug discovery, but capturing the complex distribution of conformations that contains multiple low energy minima is still a great challenge. Deep generative modeling, aiming to learn complex data distributions, is a promising approach to tackle the conformation generation problem. Here, inspired by stochastic dynamics and recent advances in generative modeling, we developed SDEGen, a novel conformation generation model based on stochastic differential equations. Compared with existing conformation generation methods, it enjoys the following advantages: (1) high model capacity to capture multimodal conformation distribution, thereby searching for multiple low-energy conformations of a molecule quickly, (2) higher conformation generation efficiency, almost ten times faster than the state-of-the-art score-based model, ConfGF, and (3) a clear physical interpretation to learn how a molecule evolves in a stochastic dynamics system starting from noise and eventually relaxing to the conformation that falls in low energy minima. Extensive experiments demonstrate that SDEGen has surpassed existing methods in different tasks for conformation generation, interatomic distance distribution prediction, and thermodynamic property estimation, showing great potential for real-world applications.

Received 9th August 2022
Accepted 11th January 2023

DOI: 10.1039/d2sc04429c

rsc.li/chemical-science

Introduction

The conformation of a molecule represents the 3D coordinates of all the atoms in a molecule. It is well acknowledged that in statistical dynamics, we know everything about a macroscopic system if we can search for all the corresponding microstates, *i.e.*, geometries. Therefore, it is quite essential to obtain all the possible conformations for studied systems to solve complicated biomolecule-involved problems such as structure-based drug design. For example, the quality and diversity of the 3D conformations of a molecule are crucial for various tasks in drug discovery, such as three-dimensional quantitative structure–activity relationships (3D-QSAR),[1] pharmacophore searching,[2] molecular docking,[3] and thermodynamic calculations.[4]

Experimental techniques for 3D structure determination, including X-ray crystallography, cryo-electron microscopy (Cryo-EM) and nuclear magnetic resonance (NMR) spectroscopy, have made continuing progress, but typically they can only provide a single or several static snapshots for the studied system.[5] Moreover, all these experiments are time-consuming and costly. Therefore, economical computational methods are needed to generate a series of conformations of molecules to study their dynamical evolution. Existing computational approaches for molecular conformation generation mainly rely on molecular dynamics (MD)[6] and distance geometry (DG).[7,8] In MD, the conformational state of a molecule is sequentially updated based on the forces acting on each atom, starting from an initial state and a chosen approach for force computation. There are

[a]*Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, China. E-mail: tingjunhou@zju.edu.cn; yukang@zju.edu.cn*
[b]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, Zhejiang, China*
[c]*State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang, China*
[d]*School of Computer Science, Wuhan University, Wuhan 430072, Hubei, China*
[e]*Hangzhou Carbonsilicon AI Technology Co., Ltd, Hangzhou 310018, Zhejiang, China*

† Electronic supplementary information (ESI) available: Part 0. Further explanation for COV, MAT, and MMD; Part 1. The relation between conformation generation and protein folding; Part 2. The exponential averaging algorithms on SDEGen; Part 3. The algorithm of the predictor-corrector solver; Part 4. Different calculation settings; Part 5. Examples of generated conformations and the additional 10-rotor examples; Algorithm S1. predictor-corrector solver; Fig. S0. An illustrative example of COV and MAT; Fig. S1. The examples of the conformations generated by SDEGen; Fig. S2. Comparison of the conformations generated by different methods for several examples; Fig. S3. The additional ten two-rotors energy surface and the SDEGen generated samples. The darker the color of the potential energy surface, the lower the energy. See DOI: https://doi.org/10.1039/d2sc04429c

‡ Equivalent authors.

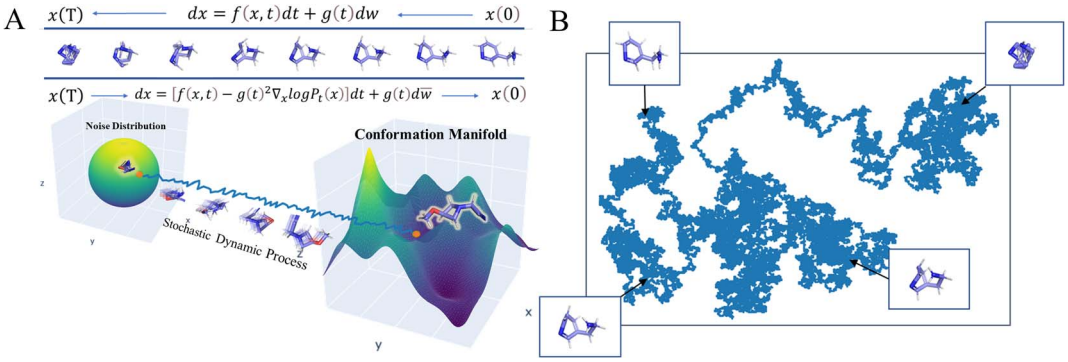

Fig. 1 Physical illustration of (A) conformation generation based on Stochastic Differential Equations (SDEGen) that evolves through the dynamical system, and (B) the evolvement of the random noise to the thermodynamically stable arrangement of atoms.

three classes of approaches to calculating the forces on atoms according to their theoretical principles: *ab initio* methods,[9] density functional theory (DFT),[10] and molecular mechanics (MM) based on empirical force fields.[11–13] *Ab initio* methods and DFT are challenging to be applied to large systems due to extensive computational cost. MM based on empirical force field is much faster than the former two, but it has been considered to give a crude approximation of molecular potential energy.[14] In DG, a randomly sampled set of atomic coordinates is refined against distance constraints to generate rough 3D conformations. However, the estimation of the distance matrix obtained based on traditional triangular constraints is still too coarse, resulting in low-quality conformations.[15] Therefore, generating more natural and diverse low-energy conformations is still a long-standing challenge.

With the advances of artificial intelligence (AI) technologies in recent years, 3D deep generative models have been utilized for conformation generation. In 2019, Mansimov and coworkers reported the first attempt to generate 3D conformations in Cartesian coordinates using the Variational AutoEncoder (VAE) architecture.[17] Subsequently, researchers adopted the idea of DG into conformation generation by changing the learning objective from the distribution in Cartesian coordinates to the distribution in the distance matrix representation, followed by the reconstruction of the 3D conformations of molecules with improved performance.[18,19] Two contemporary works reported by Ganea *et al.*[20] and Xu *et al.*[19] can generate conformations in an end-to-end fashion *via* geometry elements assembly and bi-level programming, respectively. The state-of-the-art (SOTA) score-based method ConfGF[21] reported by Shi *et al.* learns the pseudo-force on each atom and obtains new conformations *via* Langevin Markov chain Monte Carlo (MCMC) sampling. Its performance on the GEOM-Drugs[22] dataset is comparable to that of the rule-based method called experimental-torsion-knowledge distance geometry (ETKDG), which is the default conformation generation model implemented in RDKit.[23] Two other methods that are developed more recently exhibit promising performances: one is DMCG,[24] which directly manipulates Cartesian coordinates, and the other is torsional diffusion,[25] which searches the conformations in the torsional space. Except the generative model framework, reinforcement learning has been used to generate conformations by scanning all accessible torsion angles.[26,27] Furthermore, Luo and coworkers extended the method of ConfGF named DGSM to enhance its performance on the GEOM-Drugs dataset[28] by randomly adding non-bonded edges to graph structures. In conclusion, most of these methods have better performance than the ETKDG approach on the small molecule dataset GEOM-QM9 with the number of atoms less than nine, but there is still a lot of room for improvement in GEOM-Drugs, which is more closely related to the application scenarios of drug design.

Here, inspired by recent advances in generative modeling[29] and stochastic dynamics,[30] we developed a conformation generation model, SDEGen, based on stochastic differential equations (SDE) using a deep generative model (Fig. 1). Different from the regression scheme, our model can generate not only one energetically favorable conformation but also a series of locally optimal conformations, in consistent with the real thermodynamic environment. Three benchmarks for conformation generation, interatomic distance distribution analysis and thermodynamic property prediction were designed to evaluate the accuracy and diversity of the generated conformations by using the metrics including Coverage (COV) and Matching (MAT) of molecular geometry, Max Mean Discrepancy (MMD) of interatomic distance distribution and Mean Absolute Error (MAE) of energy. The results show that SDEGen beats most competitive models on the GEOM-Drugs dataset across almost every tested metric used for benchmarking. In particular, SDEGen outperforms all the competitors on both the COV and MAT metrics to the GEOM-Drugs dataset after force-field refinement. For the interatomic distances, SDEGen beats the other models on 4 out of 6 metrics and achieves comparable results on the other one. Furthermore, with regards to the prediction of thermodynamic properties, the ensemble properties of conformations generated by SDEGen are closest to the results of the DFT calculation (~2 kJ mol^{-1}), which brings benefits to structure-based drug design with more accuracy and effectiveness. By testing on the generated conformations of a randomly selected molecule, we observed an excellent coverage of almost all local minima on the energy landscape at the DFT level, including but not limited to where the crystal structure conformation falls in. Further tests on 12 more

molecules with diverse sizes (up to 12 rotatable bonds) are conducted at the semi-empirical level, which also demonstrate good coverage of most energetically favorable regions. Finally, our model can sample conformations ten times faster than the score-based SOTA method ConfGF, showing powerful application prospects in the real world.

Results and discussion

As shown in Fig. 2, in SDEGen, after randomly selecting a time step, the thermodynamic noise at that time is added to the initial interatomic distances, and then an embedding of the high-dimensional space together with the edge information is added to form the distance embedding conditional on the topological features, *i.e.*, $(\tilde{d}|E)$. At the same time, the attributes of atoms are also embedded and are sent to the Graph Isomorphism Networks(GIN)[31] combined with $(\tilde{d}|E)$ for feature extraction. After three iterations of graph message passing, the final distance features conditional on molecule Graph $(\tilde{d}|G)$ are formed. Finally, we map $(\tilde{d}|G)$ into the vector of dimension one and compute the L_2 loss with the original noise. This process is repeated many times at various molecules and time between [0,1] until convergence. The empirical force field-based energy optimization is embedded in the last step to fine-tune the conformations obtained by the stochastic dynamics system. Through the well-trained SDEGen network, the random samples can be evolved to thermodynamically stable conformations. The Euler–Maruyama[32] solver, predictor-corrector scheme,[33] and ODE-Solver can be used to generate conformations. The details of SDEGen are described in the Method Section.

Physical illustration and model comparison

We schematically construct the physical intuition for the SDEGen model (Fig. 1). The phase space of a molecule is approximately 3N dimensions, where N is the number of the atoms in the molecule, and each point in the phase space represents a conformation. Due to the energy constraints, molecular conformations are not discrete and uniformly distributed in the phase space. The possible conformations of a molecule are distributed over the low-dimensional manifold in the high-dimensional phase space. Our initial sampling is randomly sampled in the high-dimensional phase space. It then evolves through the dynamical system, represented by the stochastic differential equation $dx = f(x,t)dt + g(t)dw$, to the low-dimensional manifold of the original data distribution, forming our final molecular conformation (Fig. 1A). This map in the phase space can be understood as a motion guided by the given

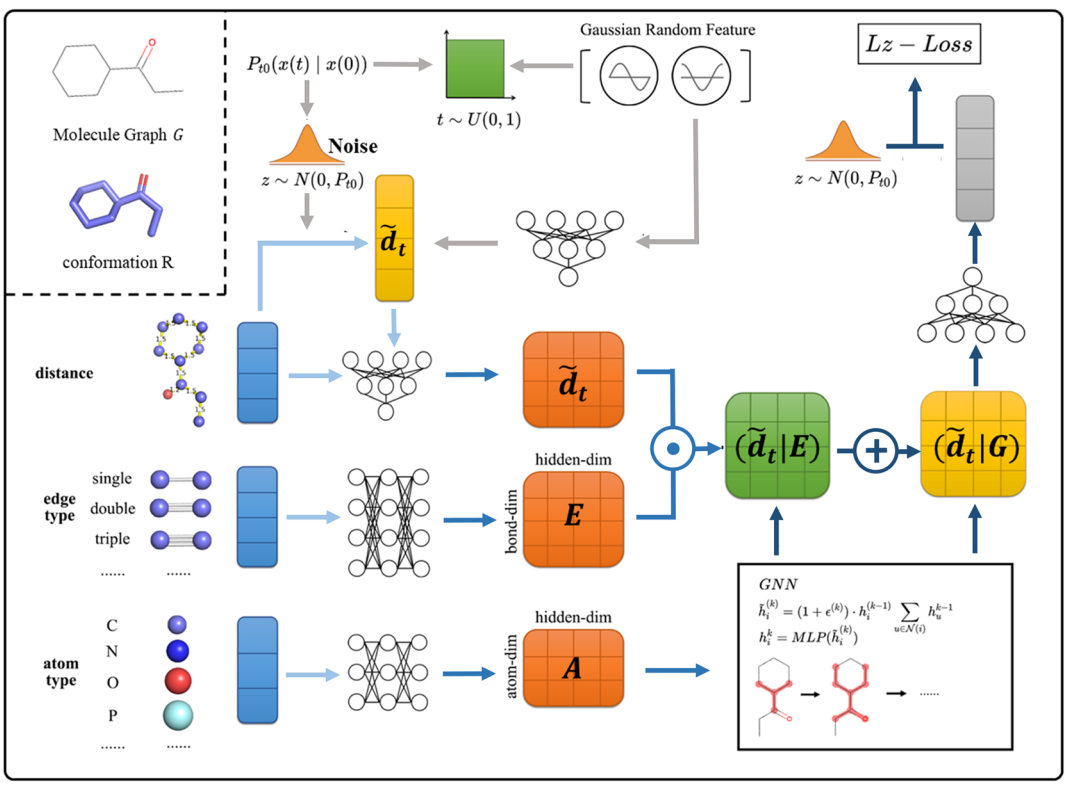

Fig. 2 The framework of SDEGen. At training time, given the graph G and conformation R, we: (1) sample the time from [0,1] uniform distribution and utilize the Gaussian random feature to encode the time information to the model, then this temporal feature is mapped together with the perturbed distance to form the $\tilde{d}$ (2) Map the edge(E) and atom(A) features from molecules to form corresponding embeddings (light blue and slightly darker blue arrows) (3) utilize the GNN model to encode the graph structure to the model $(\tilde{d}|G)$ and train the SDEGen network with denoising score matching (dark blue arrows). The procedure amounts to learning the evolutionary state of the molecule in the stochastic dynamics system at the given time.

force field, and our stochastic differential equation can be viewed as a dynamically driven approach. As illustrated in Fig. 1B, one can imagine scattering a handful of particles into the water, and the positions of the particles obey a random distribution at first. As the particles continue to collide with the water molecules, their final position distribution will tend to be thermodynamically stable. In the language of mapping, this dynamics process can be understood as a diffeomorphism from a D-dimensional hypersphere in 3N-dimensional space to another D-dimensional complex manifold in 3-dimensional space. Fig. S1 and S2† show some examples visually generated by SDEGen and other methods.

Quality of generated conformations

The mean and median of COV and MAT scores were evaluated on both the GEOM-QM9 and GEOM-Drugs datasets for SDEGen and other competitive methods, implying that the comparison was made between the conformations generated by the model and the ensemble of the quantum-computed conformers. The COV represents how much the set of quantum-computed conformations can be covered by the set of generated conformations for a given RMSD threshold: the higher, the better; while the MAT measures how similar the generated and the training QM-level conformations are: the lower, the better. The specific definition and an illustrative example of the metrics COV and MAT, standing for the diversity and accuracy of the generated conformation cluster under a given RMSD threshold, could be found in the ESI† (Part 0). The comparison of different conformation generation models is hardly straightforward because of the different training and evaluation settings used in different work. Here we have done exhaustive experiments to compare five competitive models (*i.e.*, ConfGF, RDKit, DMCG, CGCF, and ConfGFDist) with the same dataset and settings, as shown in Table 1. SDEGen shows excellent performance on all four metrics on both datasets with force-field refinement, and achieves SOTA results on the GEOM-Drugs dataset. All the parameters were untuned. Interestingly, we observe that the performance of DMCG decreases after the force-field refinement, especially on GEOM-Drugs (dropping from 95.36 and 100.0 to 87.02 and 97.73 of mean and median COV). The subsequent experiments on the bond distribution and thermodynamic properties also support this observation. That may be because the DMCG model is designed to directly predict Cartesian coordinates trained through data enhancement, which does not meet the requirements of SE (3)-equivariance originally and causes the overfitting. At the same time, SDEGen embeds the geometry constraints by modeling the three-hop distance, contributing to more robust performance after the force-field refinement. We would like to suggest such refinement as a standard step for conformation generation tasks, just like how it serves in solving crystal or cryo-EM structures, to fine-tune the generated conformations into the nearest stationary point on the specified potential energy surface, and to check the quality of the generated structures, which also fits for the meet of the downstream tasks in real-world applications. In contrast to ConfGF, SDEGen is a multiple-stage approach, which is generally believed to be inferior to the end-to-end model,[34] but it still achieved an overall victory on the two datasets with the force-field refinement settings. Additionally, compared with DGSM that concerns long-range interactions by adding non-bonded edges stochastically, our model achieved better COV and MAT scores with consideration of three-hop distances, which correspond to the truncation value for the calculations of the direct interactions in MM methods.[35]

Table 1 COV and MAT scores of the different methods on the GEOM-QM9 and GEOM-Drugs datasets with Merck molecular force field refinement. The threshold δ was set to 0.5 Å for QM9 and 1.25 Å for Drugs

		COV(%) (↑)		MAT(Å) (↓)	
Dataset	Method	Mean	Median	Mean	Median
QM9	RDKit	81.82	85.98	0.3027	0.2564
	CGCF	83.48	86.70	0.2984	0.2694
	ConfGF	90.99	95.76	0.2648	0.2691
	ConfFDist	83.80	86.72	0.2658	0.2618
	DMCG	**96.14**	**99.55**	0.2035	0.2002
	SDEGen	92.40	96.51	**0.2034**	**0.1918**
Drugs	RDKit	70.47	77.08	1.2069	1.1080
	CGCF	72.41	74.09	1.1198	1.1017
	ConfGF	86.39	89.86	0.8554	0.8347
	ConfFDist	81.08	88.37	0.9624	0.9368
	DMCG	87.02	97.73	0.8794	0.8693
	SDEGen	**92.00**	**98.51**	**0.7892**	**0.7665**

Distribution of interatomic distances

The interatomic distances contain not only the bond lengths between atoms with covalent bonds but also auxiliary bonds, *i.e.*, two-hop and three-hop distances (1–3 bond and 1–4 torsion interactions). As shown in Fig. 3B, the distribution of the interatomic distances roughly shows three peaks corresponding to the three hop bonds, which presents more structural information (including bond angels and bond dihedrals) than the trivial bond length distribution, without being too redundant like the distance matrix evaluated in RDKit. The metric MMD employed here is a kernel-base statistical test to determine whether the given two distributions are the same. The low MMD value indicates similar interatomic distributions. As shown in Table 2, although the 3D reconstruction process has compromised

Table 2 The mean and median MMD of the interatomic distances distribution of different methods compared with the test set. Single: individual distances $p(d_{ij}|G)$, Pair: pairwise distances $p(d_{ij},d_{uv}|G)$, All: all distances $p(d|G)$

	Single		Pair		All	
Method	Mean	Median	Mean	Median	Mean	Median
RDKit	3.4513	3.1602	3.8452	3.6827	4.0866	3.7519
DMCG	4.5088	5.0245	5.2494	5.8464	5.8464	6.3546
CVGAE	4.1789	4.1762	4.9184	5.1856	5.9747	5.9928
GraphDG	0.7645	0.2346	0.8920	0.3287	1.1949	0.5485
CGCF	0.4490	0.1786	0.5509	0.2734	0.8703	0.4447
ConfGF	0.3684	0.2358	0.4582	0.3206	0.6091	0.4240
SDEGen	0.3943	0.1037	0.4518	0.1762	0.6249	0.2742

the estimation of distance distribution, SDEGen still yielded impressive results on all the metrics. In particular, SDEGen outperformed ConfGF in Single-median, Pair-median, all-median, and Pair-mean metrics, and still reached comparable results in Single-mean (0.3943 *vs.* 0.3684) and all-mean (0.6249 *vs.* 0.6091) metrics. It is noted that the performance of DMCG on this task falls short of the SDEGen and ConfGF models despite its superior performance on the above-mentioned conformation generation evaluation task. The comparison of SDEGen with other DG-inspired methods (*i.e.*,GraphDG and CGCF) indicates that SDEGen could learn smooth distance distributions conditional on different types of atoms and chemical bonds, demonstrating the plausibility of the conformations generated by SDEGen from another perspective.

Prediction of thermodynamic properties

Thermodynamic property prediction needs a comprehensive understanding of the macroscopic states of a system. In this task, each conformation corresponds to a microscopic thermodynamic state, and these conformations are aggregated as an ensemble for a specific molecule to represent a thermodynamic system. The more comprehensive microstates considered in thermodynamic calculations, the more accurate the prediction can be. Following this principle, we evaluated the thermodynamic properties of the conformation ensemble generated by SDEGen and its representative opponents (*i.e.*, RDKit, CGCF, ConfGF, and DMCG). The results in Table 3 show that the ensemble properties of the conformations generated by SDEGen are closest ($\sim$2 kJ mol^{-1}) to those obtained by DFT calculations. Among all the results, SDEGen performed considerably better than the classical method RDKit ($\sim$50 kJ mol^{-1}) on this task, implying that the stochastic dynamics method we developed does learn the molecular thermodynamic evolution process with quantum accuracy. In contrast, CGCF performs poorly (50–2500 kJ mol^{-1}) on this task due to its insufficient capability to learn the multi-model conformational manifolds. ConfGF and DMCG perform much better than the above two, but SDEGen still beats all the baselines in all the metrics and achieves the accuracy of quantum chemistry for conformation generation.

Searching for crystal and other thermodynamically stable conformations

The conformations of small-molecule ligands in the bound states suggested by their experimentally determined structures deposited in Protein Data Bank (PDB[36]), are usually regarded as the gold standard in structure-based drug design. In most cases, these near-native conformations fall in the vicinity of the local minima, judging by the free-energy landscape of the molecule in its free state. One of our expectations for the model is that the multiple local minima could be captured so that the natural crystal conformations determined experimentally would be included in the generated conformation ensemble. To test this capability, we treated the Platinum dataset[37] as another external test set. The platinum dataset contains 4626 structures extracted from a total of over 347 k co-crystallized ligand structures stored in PDB by filtering out low-quality co-crystallized ligand structures (resolution > 2.0 Å) according to a set of well-designed criteria. We used the SDEGen model trained on GEOM-Drugs to generate the molecular conformations in the Platinum dataset and compared them with their original crystal conformations. The superimposed structures as well as their RMSDs of representative molecules including macrocyclic and chiral ones are shown in Fig. 3A.

Table 3 The MAE of ensemble thermodynamic properties among different methods (units: kJ mol^{-1})

Method	$\bar{E}$	E_{min}	$\Delta\bar{\varepsilon}$	$\Delta\varepsilon_{min}$	$\Delta\varepsilon_{max}$
RDKit	70.928	46.473	27.531	52.887	18.387
CGCF	2456.7	195.94	197.38	902.02	56.323
DMCG	8.0016	8.1336	13.355	16.262	17.188
ConfGF	3.6643	3.3657	5.5305	13.355	4.8476
SDEGen	2.6406	2.9219	4.4440	12.036	4.5742

As shown in Fig. 3B, among over 85% cases, the conformations generated by our SDEGen can cover the crystal structures in the Platinum dataset (with an RMSD threshold of 1.5 Å). Moreover, the crystal conformation coverage did not increase with the increase of the number of the generated conformations in the ensemble, highlighting the model's good robustness. In general, one can reach more than 75% probability of covering the crystal conformations by generating 50 conformations for a molecule, and $\sim$80% probability by generating 100 conformations.

To further illustrate the quality of the conformations generated by SDEGen, we performed several case studies to test the coverage of all low-energy conformations, probably including but not limited to the crystal structures, for both QM9-level and drugs-level molecules. Firstly, we selected a molecule with two rotatable bonds from the ligand library of PDB[36] and scanned the potential energy values of each conformation as the function of two rotation angles at the DFT level. The orange dots on the potential energy surface in Fig. 3C and D represented the conformations generated by SDEGen, and the yellow dot represented the crystal conformations. It is found that all the 50 conformations generated by SDEGen fall into multiple wells on the potential energy landscape, and cover almost all the captured local minima. It means that the generated 50 conformations can depict most of the potential wells for the tested molecule; meanwhile, the crystal conformation also fell into one of the potential wells covered by our generated conformations, not surprisingly. More examples through semi-empirical x-TB calculations can be found in the ESI† for additional illustration (Fig. S3†). Moreover, we also explored the system of druglike molecules containing 12 rotatable bonds. Fig. 4 shows that the distribution sampled by SDEGen (red dots) allows adequate exploration of the conformational space of the molecule, and most of the sampled points are concentrated near the dominant conformation obtained by quantum chemistry computation (orange dots). Compared with the points sampled by RDKit (pink dots), the sample points from SDEGen are more uniformly distributed on the energy surface. Combined with the more accurate prediction of thermodynamic properties achieved by the previous experiment, we

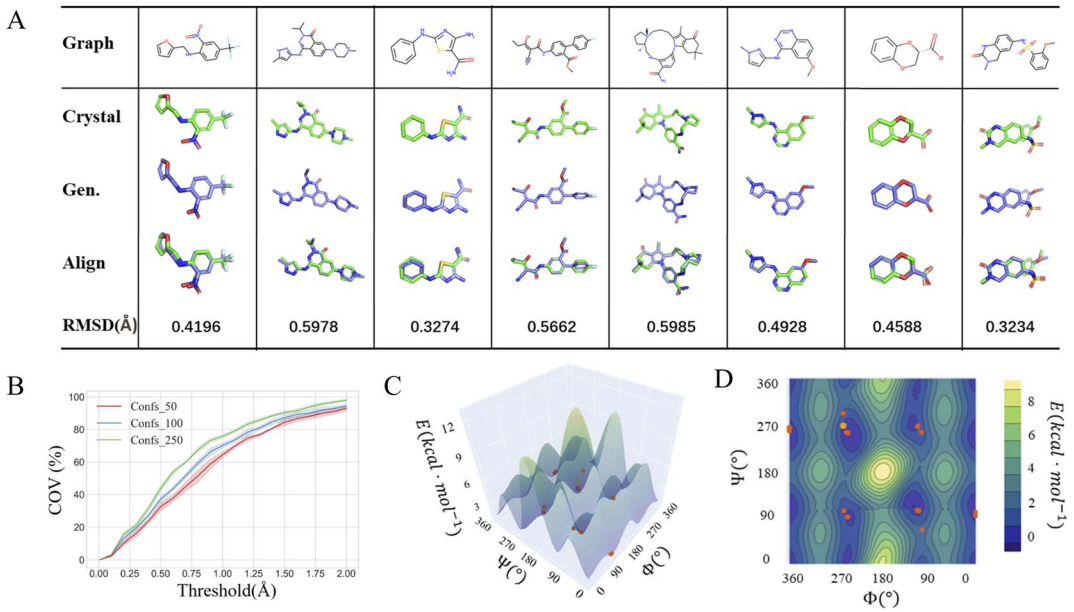

Fig. 3 (A) Comparison of generated conformations with the crystal conformations of representative molecules including macrocyclic and chiral ones; (B) the relationship between the crystal conformation coverage and the RMSD threshold within different numbers of the generated conformation of all molecules. The potential energy surface with (C) 3D and (D) 2D presentations of a randomly selected molecule scanned at DFT level. The yellow dot represents the crystal conformation and the orange dots represent the generated conformations.

believe that the SDEGen could generate a representative and uniformly sampled ensemble. Treating these generated conformations as the inputs to downstream tasks can broaden the representative of energetically favorable conformations, thus fully expressing the molecule's druggability and reducing the likelihood of missing potential active compounds in lead discovery.

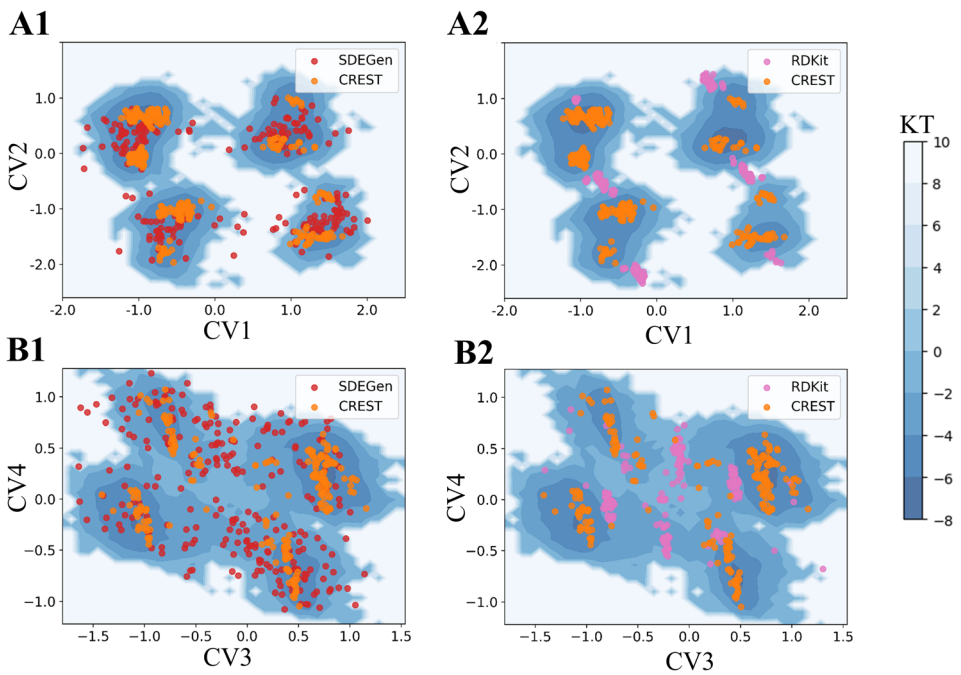

Fig. 4 Free energy surfaces of the two Drugs-level molecules (A) and (B) with 12 rotatable bonds estimated by x-TB based metadynamics. (A1) and (B1) are the comparisons between SDEGen and CREST(GEOM-Drugs) samples; (A2) and (B2) are the comparisons between RDKit and CREST samples. CV is the abbreviation of collective variable, which is the reduced coordinate used for visualization. The specific definition is in the appendix.†

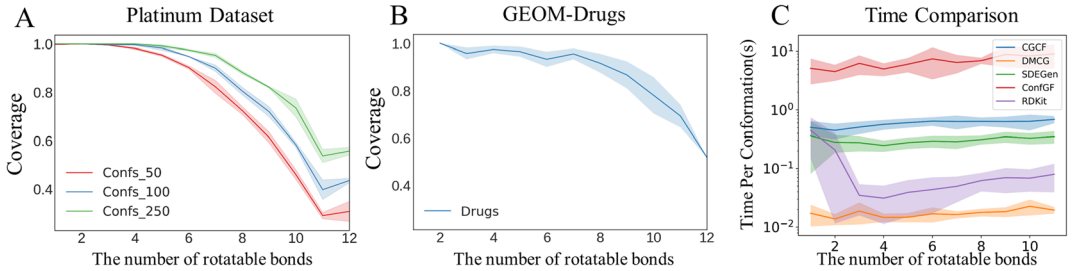

Fig. 5 Coverage *vs.* number of rotatable bonds on the (A) Platinum and (B) GEOM-Drugs datasets. Notation Confs in (A) denotes the number of the generated conformations of a molecule. (C) Comparison of conformation generation times for competitive models.

Discussion about limitations

Although SDEGen achieved good results in the Platinum dataset, the generated conformations for several molecules (about 16.0%) still did not cover the crystal conformations (RMSD threshold 1.5 Å). We attribute the failure cases to the following two reasons. Firstly, it is not easy for the model to handle large systems due to the existence of high degrees of freedom. According to Fig. 5A, the model's performance on Platinum decreases with the increase of the number of rotatable bonds. The same happens with GEOM-Drugs (Fig. 5B), implying that the long-range interactions in these relatively large systems (molecules with 8 or more rotatable bonds) need special consideration. Secondly, the conformations in Platinum are influenced by other biomolecules, whereas SDEGen only counts the internal interactions in molecules. To be specific, the Platinum dataset comprises ligands with protein-bound ligand conformations from the PDB. At the same time, the training set is generated from the DFT calculations without any consideration of the protein pocket environment. Given that protein pockets would exert some kinds of non-bonded interactions on the ligands, *i.e.*, the probability distribution of the small molecule conformation is changed by its binding with the protein (or other effector molecules). Hence the more flexible the molecules, the greater the perturbation by the protein pocket environment. Consequently, it is not surprising to observe that the COV-threshold curve of Platinum falls a little faster than that of GEOM-Drugs, owing to external interaction exerted on Platinum's molecules. We summarize the possible direction for boosting the limitations into two folds: (1) embedding physical/chemical/biological constraints to help the model learn the intrinsic physics behind a large amount of data. (2) Considering multi-scale modeling to capture higher-level interactions, as a famous saying goes, 'More is Different',[38] which is applied to the phenomenon we met here.

Conformation generation speed

Sampling speed is another perspective we should focus on beyond the quality of conformation ensembles. In real-world applications, downstream tasks such as pharmacophore mapping and conformational search require a large number of conformers, *i.e.*, 50 conformers per molecule. To prove the potential application value of SDEGen, we conducted the time cost experiment over Intel(R) Xeon(R) Gold 5218 CPU with 30 CPU cores. We divided the molecules in GEOM-Drugs based on different rotatable bonds and recorded the time used for generating a single conformation by different methods. Fig. 5C shows that SDEGen achieved comparable generation speed, about ten times faster than the score-based SOTA model, ConfGF. RDKit generates conformations at a rate of about 0.1 s a piece; meanwhile, another VAE-based SOTA method, DMCG, generates conformations at a rate of about 0.01 s a piece, which is the fastest model in the baselines. However, considering the quality of the generated conformations, the thermodynamic properties of the conformational system and other factors, we still believe that SDEGen is a competitive model.

Model comparison

The underlying math of SDEGen is $\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w$. In fact, if we made $g(t) = 0$, this stochastic dynamical system would degenerate to an ordinary differential system $\mathrm{d}x = f(x,t)\mathrm{d}t$, which is utilized for constructing the CGCF model. One possible reason why our model works well is that the presence of the stochastic term gives the model a better chance of jumping out of the local optimum. To be specific, the score-based SOTA method ConfGF performed annealing Langevin dynamics to learn the gradient field of molecular conformations, generating samples through given different temperature scales. Nevertheless, since SDEGen learns the $\nabla \log p$ (gradient of the probability distribution of the evolution of the particle over time) for the given interval, implying that one can use any classical integer method or any given step size to evolve this dynamical procedure from the beginning to the end. That is one of the reasons why SDEGen generates samples faster. Compared with another SOTA model, DMCG, which utilizes VAE as its backend, the SDEGen enjoys a lower number of model's parameters (~8 M *vs.* 1283 M) and the ability to compute likelihoods through the ODE solver, implying an additional application to the enhanced importance sampling.

Conclusions

In this study, we exploit the physical intuition and the latest generative model architecture to learn the stochastic dynamics evolution of atoms starting from a random atomic distribution and eventually relaxing to conformations near the energy optimum. This model surpasses most AI-based conformation

generation models in terms of generated conformation quality under real-world application settings, interatomic distance distribution and thermodynamic property prediction. For example, as to the conformation generation quality for drug-like molecules, our model scores best on the COV and MAT metrics with force-field refinement for the drug-like molecules. Besides, our SDEGen model is about ten times faster than the closely related model, ConfGF, which is crucial to generate a large number of conformations for large-scale virtual screening in real scenarios. In the application section, we found that SDEGen can quickly search for the conformations in the crystal structures of small molecules in the Platinum dataset with 80 percent probability. Furthermore, the energy surfaces for both small and large molecules were explored to illustrate that SDEGen could search for the local region which contains the crystal structure and locate other energetically favorable potential wells uniformly.

Method

Data representation

In this study, a molecular graph is represented as an undirected $\mathcal{G} = \langle V, E\rangle$, where $V = \{V_1, V_2, \ldots, V_{|V|}\}$ is the set of atoms of the molecule, and $\mathscr{E} = \{e_{ij} | (i,j) \subseteq V \times V\}$ is the set of bonds in the molecule. Each atom $v_i \in V$ is associated with some atom's attributes, such as element type and atomic coordinates. Each bond $e_{ij} \in \mathscr{E}$ is associated with a chemical bond type and a scalar $d_{ij} = ||v_i - v_j||_2$ denoting the Euclidean distance between the atomic positions of v_i and v_j. As the chemical bonds in a molecule would not suffice to characterize a molecule conformation and cannot express the local interactions within a molecule, we expand our molecule graph to an extended graph by adding auxiliary bonds. The two-hop edges and three-hop edges can be viewed as incorporating bond and dihedral angles information into a 2D graph, *i.e.*, the 1–3 angle interaction and the 1–4 dihedral angle interaction. This technique helps the model capture neighboring features in a molecule and conveys the chemical knowledge that covalent bonds can transmit atomic interaction, where the cutoff setting is always 3. Hereafter, we assume all molecular graphs are extended unless stated.

Generative model based on stochastic differential equation

SDEGen is based on the generative model,[29,39] which aims at learning the process of perturbating a given data distribution to random noise. We can smoothly mold random noise into data for sample generation by reversing this process. This process of perturbating data can be modeled as the solution to an Ito SDE:

$$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w \tag{1}$$

where $f(\cdot, t) : \mathbb{R}^{\mathrm{d}} \mapsto \mathbb{R}^{\mathrm{d}}$ is a vector valued function called the drift coefficient of $x(t)$, and $g(\cdot) : \mathbb{R}^{\mathrm{d}} \mapsto \mathbb{R}^{\mathrm{d}}$ is a scalar function known as the diffusion coefficient of $x(t)$. w is the brownian motion. This formula represents the process of adding noise to the data distribution to another complex distribution that contains no information on data distribution, such as a Gaussian distribution. The reverse process has been proved to satisfy a reverse-time SDE:[32]

$$\mathrm{d}x = [f(x,t) - g(t)^2 \nabla_x \log Pt(x)]\mathrm{d}t + g(t)\mathrm{d}\bar{w} \tag{2}$$

where $\bar{w}$ is a standard Wiener process when time flows back from T to 0. Once the gradient of each marginal distribution, $\nabla_x \log P_t(x)$, is known for all t, then we can derive the reverse stochastic process and simulate it to sample from the data distribution. So our goal is to train a network $s_\theta(\cdot, t) : \mathscr{R}^{\mathrm{d}} \mapsto \mathscr{R}^{\mathrm{d}}$ to approximate $\nabla_x \log P_t(x)$.

To estimate $\nabla_x \log P_t(x)$, we can train a time-dependent model $s_\theta(x,t)$ by:

$$\theta^* = \arg\min_\theta \mathbb{E}_{t\sim\mathscr{U}(0,T)}\left[\lambda(t)\mathbb{E}_{x(0)\sim p_0(x)}\mathbb{E}_{x(t)\sim p_{0t}(x(t)|x(0))}\left[\|s_\theta(x(t),t) - \nabla_{x(t)}\log p_{0t}(x(t)|x(0))\|\right]\right] \tag{3}$$

where $\mathscr{U}(0,T)$ is a uniform distribution over $[0,T]$, $p_t(x)$ is the probability density of x(t), $p_{0t}(x(t)|x(0))$ denotes the transition kernel from $x(0)$ to $x(t)$, and $\lambda(t) \in \mathbb{R} > 0$ denotes a positive weighting function. In the objective, the expectation over $x(t)$ can be estimated with empirical means over data samples from p_0. The expectation over $x(t)$ can be estimated by sampling from $p_{0t}(x(t)|x(0))$, which is efficient when the drift coefficient $f(x,t)$ is affine. The weight function $\lambda(t)$ is typically chosen to be inversely proportional to $E[\|\nabla x \log p_{0t}(x(t)|x(0))\|_2^2]$. After the network $s_\theta(x,t)$ is trained, samples could be generated by solving the reverse-time SDE equation with Euler–Maruyama sampler or predictor-corrector sampler.

Symmetry

Symmetry is ubiquitous in physics systems. Formally, a function $\mathscr{F} : \mathscr{X} \mapsto \mathscr{Y}$ being equivariant can be represented as follows:

$$\mathscr{F} \cdot \rho = \rho \cdot \mathscr{F} \tag{4}$$

where ρ is a transformation function, *e.g.*, rotation. Eqn.(7) says that applying the ρ on the input has the same effect as applying it to the output. In our problem, we find that molecular conformations under Cartesian coordinate are not roto-translational invariance. One approach to tackle this issue is to do normalization;[40] another approach is to redesign this task based on physical intuition. Inspired by a traditional conformation generation method, Distance Geometry(DG), the target can be transformed from learning $P_t(R_i|G)$ to learning $P_t(d_i|G)$, we except:

$$s_\theta(d,t) \approx \nabla_d P_t(d_i|G) \tag{5}$$

where s_θ is the network, d is the distance between atoms in a molecule, G is a molecule graph. For generalization and elegant reasons, we explicitly embed such equivariance into the model architecture.

Based on the above discussion, the framework of SDEGen could be summarized in two stages. Firstly, it learns a conditional probability distribution $P(D|G,t)$ utilizing a generative model scheme based on the SDE. Secondly, it reconstructs the 3D Cartesian conformations from the $P(D|G,t)$ obtained in the first stage.

Specifically, we chose the form of the stochastic differential equation:

$$\mathrm{d}x = \sigma^t \,\mathrm{d}\bar{w},\ t \in [0,1] \tag{6}$$

In this case, the transition kernel is

$$p_{0,t}(x(t)|x(0)) = \mathcal{N}\left(x(t); x(0), \frac{1}{2\log\sigma}\left(\sigma_t^2 - 1\right)\right) \tag{7}$$

and the weighting function is

$$\lambda(t) = \frac{1}{2\log\sigma}\left(\sigma^{2t} - 1\right) \tag{8}$$

when s is large, the prior distribution $p_{i=1}$ can be approximated in the following form:

$$\int p_0(y)\mathcal{N}\left(x; y, \frac{1}{2\log\sigma}\left(\sigma^2 t - 1\right)I\right)\mathrm{d}y \approx \mathcal{N}\left(x; 0, \frac{1}{2\log\sigma}\left(\sigma^{2t} - 1\right)I\right) \tag{9}$$

The eqn (9) indicates that the prior distribution to be chosen is approximately independent of the data distribution and is easy to sample from. Solving this SDE numerically, we can smoothly transform the data $x(0)$ to a simple white noise $x(1)$.

$$\mathrm{d}x = -\sigma^{2t}\nabla_x \log P_t(x)\mathrm{d}t + \sigma t\,\mathrm{d}\bar{w} \tag{10}$$

In this setting, we aim to learn a conditional network to jointly estimate the gradient of perturbed data on all-time steps, which means $s_\theta(\tilde{d},\mathrm{t}) \approx \nabla\tilde{d}\log \mathrm{Pt}(\tilde{d}|G)$. Since $s_\theta(\cdot,t): \mathscr{R}^d \mapsto \mathscr{R}^d$, we can formulate the first stage of conformation generation as an edge regression problem.

Given a molecule graph G and its corresponding set of atomic distances $d \in \mathbb{R}^{|E|}$, we embed the atomic attributes and the corresponding auxiliary bond attributes into a low-dimensional space using a Multilayer Perceptron(MLP)

$$h_i{}^0 = \mathrm{MLP}(V_i), \forall v_i \in V \tag{11}$$

At the same time, in order to embed the time information so that the network can condition on t, the Gaussian random feature[29] is used as an encoding for time step t. Specifically, for a given time step t, the corresponding Gaussian random features are:

$$\mathrm{GRF}(t) = [\sin(2\pi wt)||\cos(2\pi wt)], w \sim \mathcal{N}(0, s^2 I) \tag{12}$$

where $||$ denotes the vector concatenation operation and s is a fixed number. Using the method of adding Gaussian random features to the embedding layer, we can encode time information into our network.

$$h_{e_{ij}} = \mathrm{MLP}(e_{ij}) \times \left(\mathrm{MLP}(d_{ij}) + \mathrm{MLP}(\mathrm{GRF}(t))\right), \forall e_{ij} \in E, \forall d_{ij} \in D \tag{13}$$

We then use a graph neural network to update atom embeddings. We choose Graph Isomorphism Network (GIN)[31] as the GNN module. Since GIN is a provably maximally powerful GNN under the neighborhood aggregating framework. At each layer of GIN, atom embeddings are updated by aggregating messages from neighboring atoms and bonds:

$$h_i{}^l = \mathrm{MLP}\left(h_i{}^{l-1} + \sum_{j \in N(i)} \mathrm{ReLU}\left(h_j{}^{l-1} + h_{e_{ij}}\right)\right) \tag{14}$$

where $N(i)$ denotes ith atom's neighbors. After 3 rounds of message passing, we derive the final bond embedding by concatenating the corresponding atom embeddings for each bond as follows:

$$h_{e_{ij}}^0 = h_i{}^N||h_j{}^N||h_{e_{ij}} \tag{15}$$

where $h_{e_{ij}}^0$ denotes the final embeddings of bond $e_{ij} \in E$. Finally, we use an MLP function to parameterize the SDE network, *i.e.*

$$s_\theta\left(\tilde{d}, t\right) = \mathrm{MLP}\left(h_{e_{ij}}^0\right) \tag{16}$$

We can rescale the output of the SDE network by $1/\sqrt{(E[||\nabla_x \log p_{0t}(x(t)|x(0))||_2^2])}$ to help capture the norm of the actual gradient. Based on the above discussion, the whole loss function takes the form as follows

$$\frac{1}{2L}\sum_{i=1}^{L}\mathbb{E}_{x_0\sim p(0)}\mathbb{E}_{x(t)\sim p_{0t}(x(t)\times(0))}\lambda(t)\left[\left\|s_\theta\left(\tilde{d}, t\right) \times \left(\frac{1}{2\log\sigma}(\sigma^t - 1)\right) - \left(d - \tilde{d}\right)\right\|_2^2\right] \tag{17}$$

At this point, all expectations can be computed by Monte Carlo estimation.

By the way, we also added the Exponential Moving Average (EMA)[41] algorithm to SDEGen and trained a better robust model.

$$v_t = \beta \cdot v_{t-1} + (1-\beta)\cdot\theta_t \tag{18}$$

where θ_t is the model parameters at time t, v_t is the average of the model parameters, and β is the weighted weight value, which is set to 0.999 in our model. The performance of the SDEGen model with EMA algorithms was proved to be greatly improved, especially on the small molecule dataset GEOM-QM9. The probable reason for this improvement is that the randomness introduced by the procedure of adding noise to molecular conformations is averaged out by EMA, a temporal ensembling method that allowed our final model to incorporate more historical states in the learning process. So more molecular structures will be attached attention in the final trained model.

Conformation generation

SDE solver & Langevin dynamics. Given a molecule graph and a well-trained SDE network, the generation process of molecular conformations is performed by numerically solving the stochastic differential eqn (2), *i.e.*, reconstructing the distribution of each atom's position from a noise distribution.

For the numerical solver, we use the predictor-corrector scheme,[29,42] which leverages the additional information, an estimate of the gradient of $p_x(x(t))$ *via* the network, to reduce the error of the numerical SDE solver and improve sample quality. This solver is in two steps. Based on a simple discretization to the SDE, the first step is solving the Euler–Maruyama equation, replacing dt with Δt and dw with $z \sim \mathcal{N}(0, g^2(t)\Delta tI)$. When applied to our reverse-time SDE, we can obtain the following equation:

$$x_{t-\Delta t} = x_t + \sigma^{2t} s_\theta(x_t, t)\Delta t + \sigma^t \sqrt{\Delta t} z_t, \quad z_t \sim \mathcal{N}(0,1) \tag{19}$$

Then, to improve the accuracy of the solution, n steps Langevin MCMC would be implemented as the second step.

$$x_{i+1} = x_i + \epsilon \nabla_x \log p_{t(x_i)} + \sqrt{2\varepsilon} z_i \tag{20}$$

Finally, the conformation would be sampled from Gaussian distribution, and then the PC solver would integrate the SDE in the reverse time direction to obtain the reconstructed conformation. This modified sampling scheme ensures that the sample fully converges to the probability distribution under the given time at each step of solving the SDE equation, reducing the risk of a spatial clash of conformations.

ODE solver. For eqn (2), there exists an ordinary differential equation

$$\mathrm{d}x = \left[f(x,t) - \frac{1}{2} g(t)^2 \nabla_x \log p_t(x) \right] \mathrm{d}t \tag{21}$$

which shares the same marginal probability density p_t(x) with eqn (2). Therefore, we can solve this differential equation through classical integrated algorithms to sample the new energetically favorable conformations and track how the probability evolves after the sampling procedure.

We obtained the likelihood of conformation with the following equation:

$$\log p_0(x(0)) = \log p_1(x(1)) - \frac{1}{2} \int_0^1 \frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t} \mathrm{div}\, s_\theta(x,t) \mathrm{d}t \tag{22}$$

Through this by-product of the ODE sampler, the weights of the samples generated by SDEGen are known. Furthermore, combined with the energy function we define, we can use the SDEGen as an importance sampler to overcome the so-called rare event problem in molecular simulation. But this section is beyond conformation generation. We only provide a demo version code to the interested readers and leave it for a future adventure. The molecule conformations, after evolving through the stochastic system, will fall near the local optimal point, and then we use deterministic optimization to make it converge further.

Experiments

To thoroughly evaluate the performance of SDEGen, we compare it with multiple competitive methods on multiple benchmark datasets with the various tasks.

Tasks and metrics

Quality of generated conformations. In this task, we generated twice the number of conformations as its benchmark conformations for each molecular graph in the test set following conventions.[19,21,43] We then computed the COV and MAT between the generated and benchmark conformations. As the fundamental metric for our conformation evaluation, we used root mean square deviation (RMSD), a standard measure of the difference between two conformations in MD simulations analysis.

$$\mathrm{RMSD}\left(\tilde{R}, R\right) = \min_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \Phi\left(\tilde{R}_i\right) - R_i \right\|^2 \right)^{\frac{1}{2}} \tag{23}$$

where n is the number of heavy (non-hydrogen) atoms and Φ is an alignment function that aligns two conformations by rotation and translation. Following,[19] the COV and MAT used to quantify the quality of conformations are defined as follows:

$$\mathrm{COV}(S_\mathrm{g}, S_\mathrm{r}) = \frac{R \in S_\mathrm{r} \left| \mathrm{RMSD}\left(R, \hat{\mathrm{R}}\right) < \delta, \hat{R} \in S_\mathrm{g} \right.}{|S_\mathrm{r}|} \tag{24}$$

$$\mathrm{MAT}(S_\mathrm{g}, S_\mathrm{r}) = \frac{1}{|S_\mathrm{r}|} \sum_{R \in S_\mathrm{r}, \hat{R} \in S_\mathrm{g}} \min \mathrm{RMSD}\left(R, \hat{R}\right) \tag{25}$$

where S_g and S_r are generated and reference molecular conformation ensembles, respectively. δ is a given RMSD threshold. While COV is effective to assess the diversity and detect the model-collapse phenomenon, MAT is a complement to measure how close the generated conformations and the reference conformations. In general, a higher COV score represents greater diversity performance, while a lower MAT score represents better accuracy of the generated conformations. An illustrative example is prepared in the appendix† for better understanding.

Distribution of interatomic distances

Since the covalent bond lengths are insufficient to represent the information of three-dimensional geometry, we consider the interatomic distances measured in the second task, including the bond lengths (1–2 connection) and 1–3 and 1–4 connections. This consideration amounts to measuring the direct local interactions between atoms that drive the atoms to relax to the real-world thermodynamic distribution from a random distribution. In this task, we sampled 1000 conformations for each test molecule as pseudo-trajectories, and then calculated the MMD between the two distributions using a Gaussian kernel. In specific, for each molecule in the test set, we evaluated distributions of all distances $p(d|G)$ (All), pairwise distances $p(d_{ij}, d_{uv}|G)$ (Pair), and individual distances $p(d_{ij}|G)$ (Single).

Prediction of Thermodynamic Properties

As mentioned earlier, a macroscopic thermodynamic property of an ensemble is obtained by weighting all accessible microscopic states. For each molecule in the test set, we utilized PySCF[44] with DFT(M06-2X/def2-TZVPP) to calculate electron

energy and HOMO–LUMO for each generated and benchmark conformation. We then computed the MAE metric for macroscopic thermodynamic properties by statistical averaging. The ensemble properties considered here include average energy $\bar{E}$, lowest energy $E_{\min}$, highest energy $E_{\max}$, average HOMO–LUMO gap $\Delta\bar{\varepsilon}$, minimum gap $\Delta\varepsilon_{\min}$, maximum gap $\Delta\varepsilon_{\max}$. The mean absolute error (MAE) was used for measuring the accuracy of property prediction.

Datasets

Three well-known datasets, GEOM-QM9, GEOM-Drugs[22] and ISO17,[45–47] were used. GEOM-QM9 is a small molecule dataset containing neutral molecules with up to nine atoms, not counting hydrogen. GEOM-Drugs is a drug-like molecule dataset whose molecule species are accessed as part of AICures.[48] These conformers were generated with the CREST[49] program, which adopts semi-empirical DFT to generate reliable and accurate structures. Following the[21] sampling scheme, the resulting split is 40 000 molecules in the training set with 200 000 conformations and 200 molecules in the test set with 22 408 and 14 324 conformations for GEOM-QM9 and GEOM-Dugs, respectively. The molecules in the ISO17 dataset were randomly drawn from the largest set of isomers in the QM9 dataset, which consists of molecules with a fixed composition of atoms($C_7O_2H_{10}$). These conformers were generated with the Fritz–Haber Institute *ab initio* simulation package (FHI-aims),[50] reaching a higher level of accuracy than the DFT method. So this dataset was assigned to evaluate the interatomic distance distribution task. The default split results in the training set with 357 621 conformations of 167 molecules and the test set with 73 071 conformations of 30 molecules.

Baselines

We tested our model compared with a classical rule-based method and other ML-based methods. The rule-based method is ETKDG,[8] the default program in RDKit for molecular conformation generation. The other AI-based method go as follows: CVGAE,[16] GraphDG,[18] CGCF,[19] ConfGF,[21] DMCG[24] and DGSM.[18] Among these methods, we focus on the ConfGF since it is the SOTA method built upon the score-based generative model, achieving impressive results on both GEOM-QM9 and GEOM-Drugs datasets. Although DGSM is an improved version of ConfGF, we still have no access to its source code until now, so the reported performance[28] in its original paper was used in our study. It is noted that the official codes of GraphDG and CVGAE are utilized old versions of Tensorflow,[51] which does not match the version of our machine. However, these two methods are not our main competitors, so we just extract these results from the ConfGF[21] paper.

Data and code availability

The data and source code of this study is freely available at GitHub (**https://github.com/HaotianZhangAI4Science/SDEGen**) to allow replication of the results.

Author contributions

H. Zhang and S. Li contributed to the main code and wrote the manuscript. J. Zhang and Z. Wang performed the experiment. J. Wang and D. Jiang provided partial codes of this work. Z. Bian and Y. Zhang helped perform the analysis with constructive discussions. Y. Deng and J. Song contributed to the visualization and technique support. T. Hou and Y. Kang provided essential financial support and conception, and were responsible for the overall quality.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was financially supported by National Key Research and Development Program of China (2021YFF1201400), and National Natural Science Foundation of China (22220102001, 81973281).

References

1 J. Verma, V. M. Khedkar and E. C. Coutinho, *Curr. Top. Med. Chem.*, 2010, **10**, 95–115.
2 C. H. Schwab, *Drug Discovery Today: Technol.*, 2010, **7**, e245–e253.
3 M. McGann, *J. Chem. Inf. Model.*, 2011, **51**, 578–596.
4 S. Wlodek, A. Skillman and A. Nicholls, *J. Chem. Theory Comput.*, 2010, **6**, 2140–2152.
5 J.-P. Renaud, A. Chari, C. Ciferri, W.-t. Liu, H.-W. Remigy, H. Stark and C. Wiesmann, *Nat. Rev. Drug Discovery*, 2018, **17**, 471–492.
6 B. J. Alder and T. E. Wainwright, *J. Chem. Phys.*, 1959, **31**, 459–466.
7 G. M. Crippen and T. F. Havel, *Distance geometry and molecular conformation*, Research Studies Press Taunton, 1988.
8 S. Riniker and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2562–2574.
9 M. Yin and M. L. Cohen, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1982, **25**, 7403.
10 R. G. Parr, in *Horizons of quantum chemistry*, Springer, 1980, pp. 5–15.
11 B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels and S. Boresch, *J. Comput. Chem.*, 2009, **30**, 1545–1614.
12 D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker and W. Zhang, *Amber 10*, University of California, 2008.
13 A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.
14 I. Y. Kanal, J. A. Keith and G. R. Hutchison, *Int. J. Quantum Chem.*, 2018, **118**, e25512.
15 J. Dunitz and J. Waser, *J. Am. Chem. Soc.*, 1972, **94**, 5645–5650.

16 E. Mansimov, O. Mahmood, S. Kang and K. Cho, *Sci. Rep.*, 2019, **9**, 1–13.
17 D. P. Kingma and M. Welling, arXiv, 2013, preprint, arXiv:1312.6114.
18 G. N. Simm and J. M. Hernández-Lobato, arXiv, 2019, preprint arXiv:1909.11459.
19 M. Xu, S. Luo, Y. Bengio, J. Peng and J. Tang, arXiv, 2021, preprint, arXiv:2102.10240.
20 O. Ganea, L. Pattanaik, C. Coley, R. Barzilay, K. Jensen, W. Green and T. Jaakkola, *Adv. Neural Inf. Process. Syst.*, 2021, **34**, 13757–13769.
21 C. Shi, S. Luo, M. Xu and J. Tang, *presented in part at the Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2021.
22 S. Axelrod and R. Gomez-Bombarelli, arXiv, 2020, preprint, arXiv:2006.05531.
23 G. Landrum, Rdkit documentation, *Release 1.1-79*, 2013, p. 4.
24 J. Zhu, Y. Xia, C. Liu, L. Wu, S. Xie, T. Wang, Y. Wang, W. Zhou, T. Qin and H. Li, arXiv, 2022, preprint, arXiv:2202.01356.
25 B. Jing, G. Corso, J. Chang, R. Barzilay and T. Jaakkola, arXiv, 2022, preprint, arXiv:2206.01729.
26 T. Gogineni, Z. Xu, E. Punzalan, R. Jiang, J. Kammeraad, A. Tewari and P. Zimmerman, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 20142–20153.
27 B. Rai, V. Sresht, Q. Yang, R. J. Unwalla, M. Tu, A. M. Mathiowetz and G. A. Bakken, *J. Chem. Inf. Model.*, 2022, **62**(4), 785–800.
28 S. Luo, C. Shi, M. Xu and J. Tang, *Adv. Neural Inf. Process. Syst.*, 2021, **34**, 19784–19795.
29 Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon and B. Poole, arXiv, 2020, preprint, arXiv:2011.13456.
30 T. Lelievre and G. Stoltz, *Acta Numer.*, 2016, **25**, 681–880.
31 K. Xu, W. Hu, J. Leskovec and S. Jegelka, arXiv, 2018, preprint arXiv:1810.00826.
32 B. D. Anderson, *Stochastic Processes and their Applications*, 1982, vol. 12, pp. 313–326.
33 S. Bhalekar and V. Daftardar-Gejji, *J. Fractional Calc. Appl.*, 2011, **1**, 1–9.
34 X. Z. Ma and E. Hovy, *arXiv*, 2016, preprint, arXiv:1603.01354, Berlin, Germany.
35 K. Vanommeslaeghe and O. Guvench, *Curr. Pharm. Des.*, 2014, **20**, 3281–3292.
36 S. K. Burley, H. M. Berman, G. J. Kleywegt, J. L. Markley, H. Nakamura and S. Velankar, in *Protein Crystallography: Methods and Protocols*, ed. A. Wlodawer, Z. Dauter and M. Jaskolski, 2017, vol. 1606, pp. 627–641.
37 N.-O. Friedrich, C. de Bruyn Kops, F. Flachsenberg, K. Sommer, M. Rarey and J. Kirchmair, *J. Chem. Inf. Model.*, 2017, **57**, 2719–2728.
38 P. W. Anderson, *Science*, 1972, **177**, 393–396.
39 Y. Song and S. Ermon, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 12438–12448.
40 F. Noé, S. Olsson, J. Köhler and H. Wu, *Science*, 2019, **365**, eaaw1147.
41 A. Tarvainen, H. Valpola, *Advances in Neural Information Processing Systems*, ed. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Curran Associates, Inc., 2017, vol. 30.
42 S. Bhalekar and V. Daftardar-Gejji, *J. Fractional Calc. Appl.*, 2011, **1**, 1–9.
43 M. Xu, W. Wang, S. Luo, C. Shi, Y. Bengio, R. Gomez-Bombarelli, J. Tang, *International Conference on Machine Learning*, 2021, pp. 11537–11547.
44 Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z. H. Cui, J. J. Eriksen, Y. Gao, S. Guo, J. Hermann, M. R. Hermes, K. Koh, P. Koval, S. Lehtola, Z. Li, J. Liu, N. Mardirossian, J. D. McClain, M. Motta, B. Mussard, H. Q. Pham, A. Pulkin, W. Purwanto, P. J. Robinson, E. Ronca, E. R. Sayfutyarova, M. Scheurer, H. F. Schurkus, J. E. T. Smith, C. Sun, S. N. Sun, S. Upadhyay, L. K. Wagner, X. Wang, A. White, J. D. Whitfield, M. J. Williamson, S. Wouters, J. Yang, J. M. Yu, T. Zhu, T. C. Berkelbach, S. Sharma, A. Y. Sokolov and G. K. Chan, *J. Chem. Phys.*, 2020, **153**, 024109.
45 K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, K.-R. Müller, *Advances in Neural Information Processing Systems*, ed. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, 2017, vol. 30.
46 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, *Nat. Commun.*, 2017, **8**, 1–8.
47 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, *Sci. Data*, 2014, **1**, 1–7.
48 AICures, **https://www.aicures.mit.edu/data**, accessed 22-05-2020.
49 S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 2847–2862.
50 V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Comput. Phys. Commun.*, 2009, **180**, 2175–2196.
51 M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving and M. Isard, *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.