Supporting Information

SDEGen: Learning to Evolve Molecular Conformations from Thermodynamic Noise for Conformation Generation

Haotian Zhang[1,#], Jintu Zhang[1,3#],Shengming Li[2], Zhe Wang[1], Jike Wang[1,4], Dejun Jiang[1], Zhiwen Bian[1], Yixue Zhang[1], Yafeng Deng[5], Jianfei Song[5], Yu Kang[1,*], Tingjun Hou[1,3,*]

[1]Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, China

[2]College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, Zhejiang, China

[3]State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang, China

[4]School of Computer Science, Wuhan University, Wuhan 430072, Hubei, China

[5]Hangzhou Carbonsilicon AI Technology Co., Ltd, Hangzhou 310018, Zhejiang, China

[#]Equivalent authors

Corresponding authors

Tingjun Hou

E-mail: tingjunhou@zju.edu.cn

Yu Kang

Email: yukang@zju.edu.cn

Part 0. Further Explanation for COV, MAT, and MMD

The RMSD defination goes as follows:

$$RMSD(\tilde{R},R) = \min_{\Phi}(\frac{1}{n}\sum_{i=1}^{n}||\Phi(\tilde{R}_i) - R_i||^2)^{\frac{1}{2}} \#(1)$$

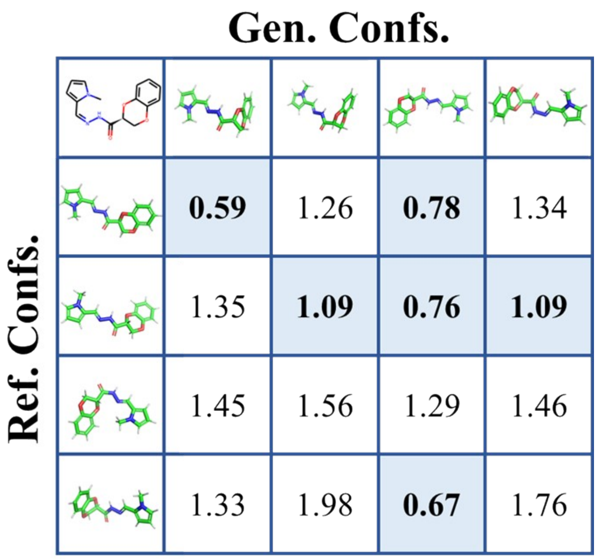
where n is the number of heavy atoms and Φ is an alignment function that aligns two conformations by rotation and translation.

Following (M. Xu, Luo, et al., 2021), we adopted Coverage (COV) and Matching (MAT) to quantify the results of the SDEGen model.

$$COV(S_g,S_r) = \frac{\left|R \in S_r \middle| RMSD(R,\hat{R}) < \delta, \hat{R} \in S_g\right|}{|S_r|} \#(2)$$

$$MAT(S_g,S_r) = \frac{1}{|S_r|}\sum_{R \in S_r, \hat{R} \in S_g} \min RMSD(R,\hat{R}) \#(3)$$

where S_g and S_r are generated and reference molecular conformation ensembles, respectively. δ is a given RMSD threshold. In general, a high COV metric represents higher diversity performance, while a low MAT metric represents better accuracy of generated conformations.



	0.59	1.26	**0.78**	1.34
	1.35	**1.09**	**0.76**	**1.09**
	1.45	1.56	1.29	1.46
	1.33	1.98	**0.67**	1.76

Figure S0. An illustrative example of COV and MAT

To visually explain the COV and MAT metrics, let's take an example to illustrate. The horizontal line is the generated conformation(Gen. Confs.), and the vertical line is the reference conformation(GT Confs.) calculated by quantum mechanics. The number of the (i,j) square grid is the RMSD value between the i-th GT Conf and the j-th Gen. Conf., the unit is Ångstrom and the RMSD threshold is 1.25 Å. To compute the COV metric, we count whether the Gen. Confs. could lie in the threshold of GT confs, which is, to count the rows containing colored squares. In this example is 3/4, so the COV is 75%. And the MAT is the mean of RMSD values, in this example is 1.2575 Å.

$$MMD^2(P,Q) = \parallel \mu_P - \mu_Q \parallel_F^2$$

$$MMD^2(P,Q) = E_P[k(X,X)] - 2E_{P,Q}[k(X,Y)] + E_Q[k(Y,Y)]$$

Maximum mean discrepancy (MMD) is a kernel-based statistical test used to determine whether given two distributions are the same which is proposed in[1].

Since the original design of the AI-based model(except the DMCG is to model the interatomic distributions, so we claim that the lower MMD is, the better estimation of distances the model makes.

Part 1. The Relation between Conformation Generation and Protein Folding Problem

Protein folding is a problem of broad interest in academia and industry, and it is a subproblem of the generalized conformation generation problem, where the goal is to search for the lowest energy static conformation. It is worth noting that most of the current popular protein conformation prediction models do not utilize the ab initio approach, but an approach aided by additional prior knowledge. There is a lot of prior knowledge about proteins, and the initial conformation of a protein can be estimated to a reasonable point on the complex potential energy surface by methods such as homology modeling and Multiple Sequence Alignment (MSA)[2], and then optimized by physical methods so that the protein conformation can be further searched to the optimal point. Most of the protein motifs are known, which also greatly simplifies the complexity of the protein folding problem.

Generalized conformation generation is believed to be a more challenging problem, where the goal is not to generate a static structure but to generate a representative set of conformations. It is generally believed that any conformation within 3-8 kcal/mol of the lowest energy conformation is likely to be the active conformation of a drug molecule[3]. Therefore, our goal is not only to find the lowest energy conformation, but also to identify a series of conformations that are at the trough of the molecular potential energy surface. So conformation prediction methods on proteins such as AlphaFold[4] cannot be directly applied to our small molecule conformation generation problem.

For the dynamic conformation generation of macromolecules, researchers have proposed several deep learning algorithms, which are broadly classified into three categories. For example, the Boltzmann Generator[5] based on the Flow model, which aims at an unbiased, once-generated sample from a thermal equilibrium system, belongs to an enhanced sampling approach. Lemke[6] proposed a dimensionality reduction algorithm based on VAE and nonlinear distance gauge; Zhang invented the TALOS[7] based on the architecture of GAN[8]. These models tend to input sampling data for a particular protein target and then perform targeted learning, so their learned coordinate transformations are not directly applicable to other molecules. In contrast, we condition the conformation generation problem to the graph structure, allowing us to generalize to molecules that the model has not seen once we have trained the models on a portion of the molecular conformations.

Part 2. The exponential averaging algorithms on SDEGen

To improve the model's performance on small molecule datasets, we used an exponential averaging algorithm to update the parameters of the SDEGen. If the number of iteration steps reaches 100, the parameters could be written as follows:

$$v_{100} = (1-\beta)\theta_{100} + \beta(1-\beta)\theta_{99} + \beta^2(1-\beta)\theta_{98} + \ldots + \beta^{99}(1-\beta)\theta_1$$

where β is a hyperparameter that controls the degree of smoothing while the θ is the optimization parameter of the model. It can be found that the improvement of the SDEGen effect on the QM9 dataset is obvious.

Part 3. The Algorithm of the Predictor-Corrector Solver

Algorithm S1. Predictor-Corrector solver

Input molecular graph G, the standard deviation of the perturbation kernel std, diffusion coefficient g(t), the number of Euler-Maruyama sampling steps N, the Langevin MCMC sampling steps M, the smallest time step for numerical stability eps,

```
Initialize interatomic distances $d_T^0$ from a prior distribution $N(0,std * I)$
$\Delta t = (1 - eps)/(N - 1)$        ▻ discretized time step
for i = N-1 to 0 do:
    $g \leftarrow g(t)$        ▻ Euler-Maruyama
    $\overline{d}_i^0 = \overline{d}_{i+1}^0 + g^2 * s_\theta(d_{i+1}^0, i) * \Delta t$
    $z \sim N(0,1)$
    $d_i \leftarrow \overline{d}_i + \sqrt{\Delta t} * g * z$
    for j = 1 to M do:        ▻ $Langevin\ MCMC$
        $z \sim N(0,1)$
        $grad \leftarrow s_\theta(d_i^{j-1}, i)$
        $\epsilon = 2 * (r * ||z||_2/||grad||_2)^2$        ▻ $\epsilon\ is\ the\ Langevin\ time\ step$
        $d_i^j \leftarrow d_i^{j-1} + \epsilon * grad + \sqrt{2} * \epsilon * z$
    end for
    $d_{i-1}^0 = d_i^M$
end for
Reconstruct $d_0^0$ to $R_0$
Output generated conformation $R_0$
```

Part 4. Multiscale Calculation Settings

During the conformation generation the MMFF94[9]-based refinement is performed with the max number of iterations 200, and the energy tolerance 1.0e-6 kcal mol^{-1} Å^{-1}. In the thermodynamic property prediction and the two-rotor with crystal ligand torsional scanning, the generated samples for the two-rotor are 50 per molecule, while for the twelve-rotor are 250 per molecule. The quantum chemistry calculations (single-point calculations) were carried out using the PySCF[10] code. Restricted density functional theory (DFT) calculations were performed under the M06-2X/def2-TZVPP level of theory [11, 12] to determine the ground state electron energies and the HOMO-LUMO gaps of different conformations of the tested molecules.

In the additional ten two-rotors and two twelve-rotors sampling experiments, the rigid scanning of the molecular potential energy surfaces that spanned by two torsion angles was performed under the GFN2-xTB level[13] of theory using the xTB[14] code. The step size of the scanning was set as 10 degrees. As a result, the obtained potential energy surfaces were all sampled by 1296 points. However, for molecules with twelve torsion angles, direct rigid scanning would require 36e10 single-point energy calculations, which is unaffordable even under the semiempirical level of theory. Thus, we invoked the semiempirical ab-initio molecular dynamics (AIMD) method to sample the distributions of their torsion angles. These AIMD calculations were carried out under the GFN2-xTB level of theory using the xTB code as well. For each molecule with twelve torsion angles, a MD run of the length of 2.5 ns was performed with a time step of 2 fs under the NVT ensemble of 300 K. The temperature control was achieved by the Berendsen thermostat[15]. To further accelerate the conformational sampling, the metadynamics[16] methodology was adopted during the simulations. The scaling factor for RMSD criteria was chosen as 0.02 and the width of the gaussian potentials was chosen as 2.0. To maintain the stabilities of the simulations, masses of the hydrogen atoms were set as 2 a.m.u, while the SHAKE method[17] was applied to constrain the length of all the chemical bonds. The simulation trajectories were saved every 50 fs. Since the dimension of the torsion angle space is rather high for an intuitive study, we performed dimensional reduction with the TICA[18] method with the pyEMMA code[19]. The lag time of the TICA method was chosen as 5 ps and the input features were selected as the twelve torsion angles of each molecule. After the dimensional reductions, the first two most significant patterns of linear combinations of the input features were adopted as the collective variables to present the overall distributions of the torsion angles.

Part5. Examples of Generated Conformations and the Additional 10-rotor Examples

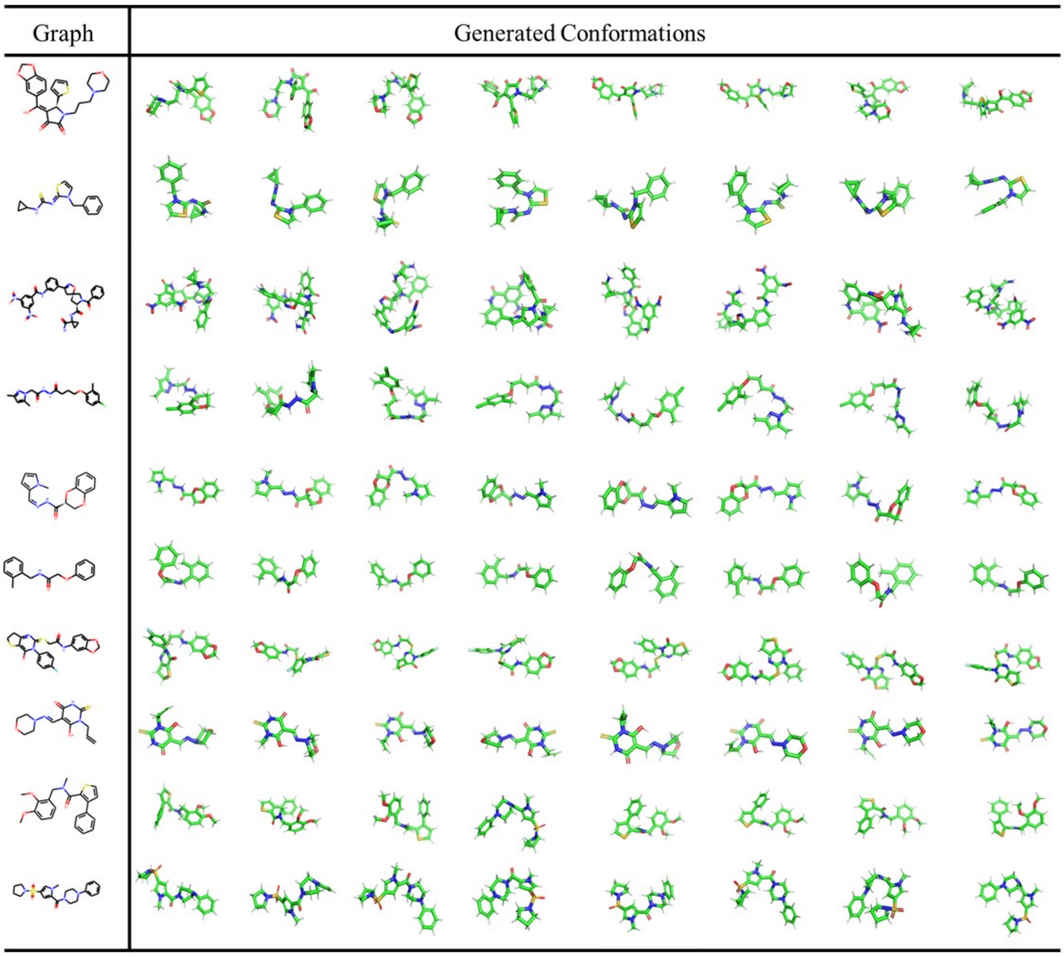


Figure S1.. The examples of the conformations generated by SDEGen.

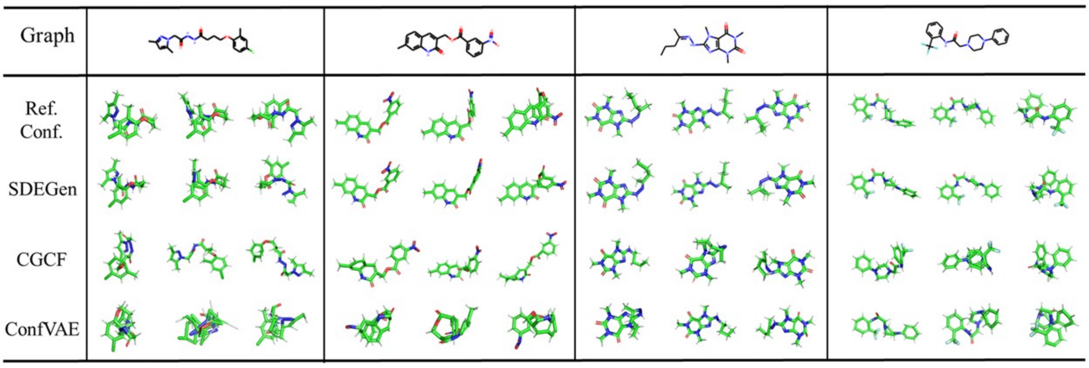
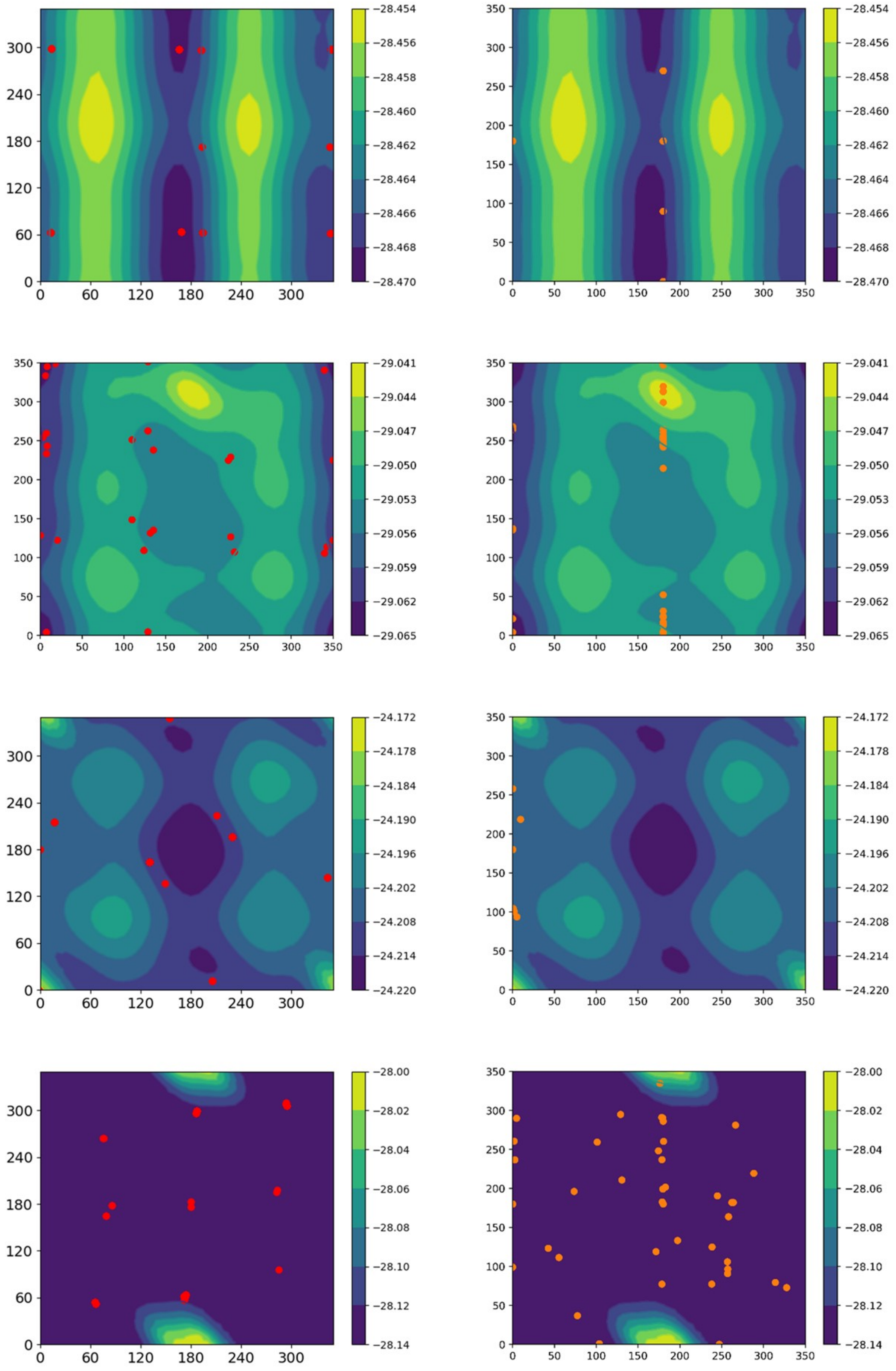
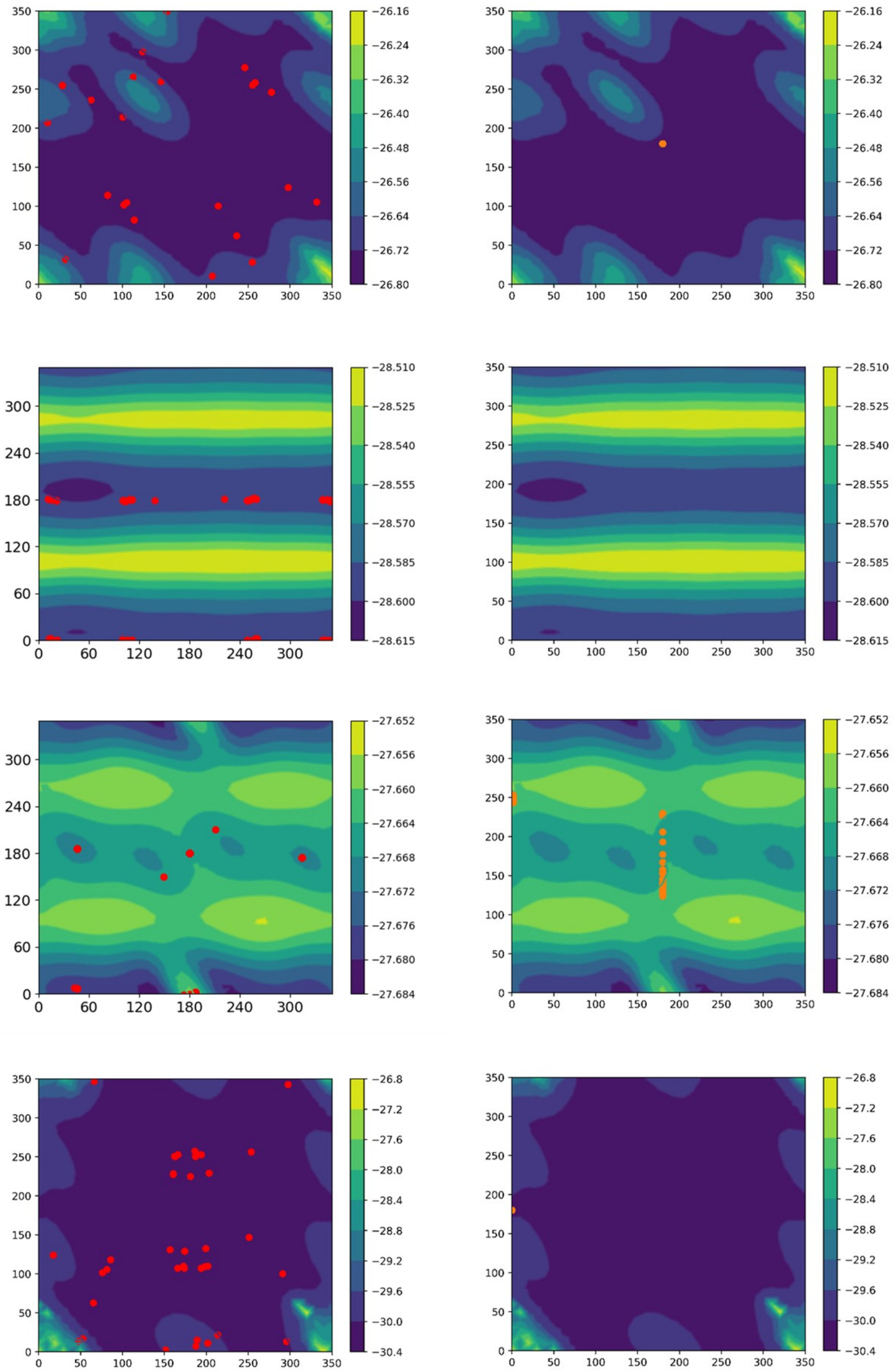


Figure S2.. Comparison of the conformations generated by different methods for several examples.



0
60
120
180
240
300
−28.454
−28.456
−28.458
−28.460
−28.462
−28.464
−28.466
−28.468
−28.470
0
50
100
150
200
250
300
350
−29.041
−29.044
−29.047
−29.050
−29.053
−29.056
−29.059
−29.062
−29.065
−24.172
−24.178
−24.184
−24.190
−24.196
−24.202
−24.208
−24.214
−24.220
−28.00
−28.02
−28.04
−28.06
−28.08
−28.10
−28.12
−28.14



0
50
100
150
200
250
300
350
−26.16
−26.24
−26.32
−26.40
−26.48
−26.56
−26.64
−26.72
−26.80
0
60
120
180
240
300
−28.510
−28.525
−28.540
−28.555
−28.570
−28.585
−28.600
−28.615
−27.652
−27.656
−27.660
−27.664
−27.668
−27.672
−27.676
−27.680
−27.684
−26.8
−27.2
−27.6
−28.0
−28.4
−28.8
−29.2
−29.6
−30.0
−30.4

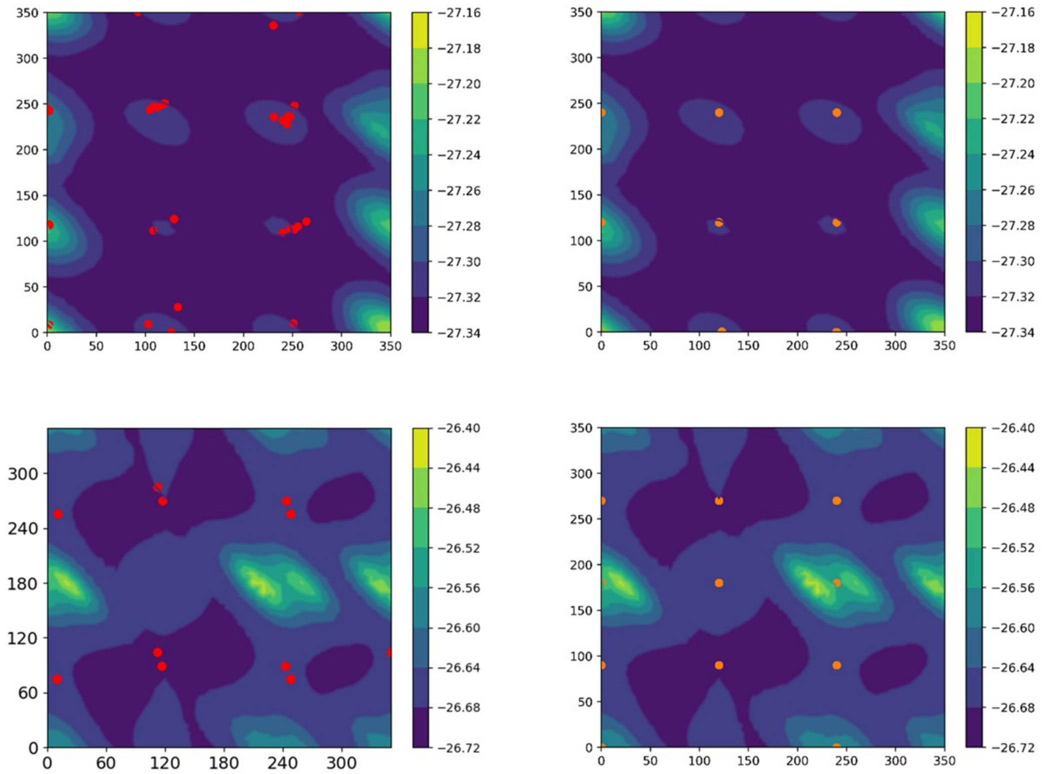


Figure S3. The additional 10 two-rotors energy surface and the SDEGen generated samples. The darker the color of the potential energy surface, the lower the energy. Red color points represent samples of SDEGen, while orange points represent samples of RDKit. Two axes represent two rotatable angles. There are some potential energy surfaces that are empty because RDKit did not succeed in generating the conformations of these molecules.

Reference

1. A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf and A. Smola, *The Journal of Machine Learning Research*, 2012, **13**, 723-773.
2. R. C. Edgar and S. Batzoglou, *Current opinion in structural biology*, 2006, **16**, 368-373.
3. J. Boström, J. R. Greenwood and J. Gottfries, *Journal of Molecular Graphics and Modelling*, 2003, **21**, 449-462.
4. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek and A. Potapenko, *Nature*, 2021, **596**, 583-589.
5. F. Noé, S. Olsson, J. Köhler and H. Wu, *Science*, 2019, **365**, eaaw1147.
6. T. Lemke and C. Peter, *Journal of chemical theory and computation*, 2019, **15**, 1209-1215.
7. J. Zhang, Y. I. Yang and F. Noé, *The journal of physical chemistry letters*, 2019, **10**, 5791-5797.
8. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, *Communications of the ACM*, 2020, **63**, 139-144.
9. P. Tosco, N. Stiefl and G. Landrum, *Journal of cheminformatics*, 2014, **6**, 1-4.
10. Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova and S. Sharma, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2018, **8**, e1340.
11. Y. Zhao and D. G. Truhlar, *Theoretical chemistry accounts*, 2008, **120**, 215-241.
12. F. Weigend and R. Ahlrichs, *Physical Chemistry Chemical Physics*, 2005, **7**, 3297-3305.
13. C. Bannwarth, S. Ehlert and S. Grimme, *Journal of chemical theory and computation*, 2019, **15**, 1652-1671.
14. C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2021, **11**, e1493.
15. H. J. Berendsen, J. v. Postma, W. F. Van Gunsteren, A. DiNola and J. R. Haak, *The Journal of chemical physics*, 1984, **81**, 3684-3690.
16. S. Grimme, *Journal of chemical theory and computation*, 2019, **15**, 2847-2862.
17. J.-P. Ryckaert, G. Ciccotti and H. J. Berendsen, *Journal of computational physics*, 1977, **23**, 327-341.
18. G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis and F. Noé, *The Journal of chemical physics*, 2013, **139**, 07B604_601.
19. M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz and F. Noé, *Journal of chemical theory and computation*, 2015, **11**, 5525-5542.