거리 정보를 통합한 분자 그래프 생성을 위한 Diffusion 모델

박형준[01], 최승연[1], 김환희[1], 이승용[2], 김윤주[3], 박상현[1†]
연세대학교 컴퓨터과학과[1], 홍익대학교 컴퓨터공학과[2], 숙명여자대학교 소프트웨어학부[3]
{ktori1361, tmddus1553, hwanhee, sanghyun}@yonsei.ac.kr,
bluearth4587@g.hongik.ac.kr, angelkim0409@sookmyung.ac.kr

Extended Diffusion Model for Molecular Graph Generation Incorporating Distance Matrix

Hyoungjoon Park[01], Seungyeon Choi[1], Hwanhee Kim[1], Seungyong Lee[2], Yoonju Kim[3], Sanghyun Park[1†]
Dept. of Computer Science, Yonsei University[1]
Dept. of Computer Engineering, Hongik University[2]
Division of Computer Science, Sookmyung Women's University[3]

요 약

신약 개발에서 새로운 분자 구조를 생성하여 약물 후보를 설계하는 과정은 많은 시간과 자원을 요구한다. 이에 따라, 딥러닝 기반의 Diffusion 모델을 활용한 분자 생성 모델이 기존 전통적인 방법의 비효율성을 해결할 대안으로 주목받고 있다. 그러나 기존 모델은 분자의 2D 그래프 정보만을 활용하여, 실제 약물-표적 상호작용에서 중요한 3D 공간적 구조와 상호작용을 충분히 반영하지 못하는 한계가 있다. 본 연구에서는 기존의 한계를 극복하기 위해 원자 간 3D 거리 행렬(Distance Matrix)을 추가로 활용한 확률적 미분 방정식(SDEs) 기반의 Diffusion 모델을 제안한다. 그 결과, 제안된 모델은 기존 베이스라인 모델들보다 생성된 분자의 Novelty와 Uniqueness에서 우수한 결과를 보였다.

1. 서 론

스마트시티는 질병의 신속한 진단과 치료가 핵심 요소이며, 본 연구는 이러한 환경을 마련하는 데 있어 중요한 역할을 한다. 신약 개발은 새로운 분자 구조를 설계하는 과정을 포함하지만, 화학 공간의 복잡성으로 인해 전통적인 실험 방식은 많은 시간과 자원이 소모된다. 이를 해결하기 위해 최근 딥러닝 기반의 Diffusion 모델이 소분자, 항체, mRNA 백신 등 치료제 설계에 활용되며 주목받고 있다. Diffusion 모델은 데이터에 단계적으로 노이즈를 추가하고 제거하는 과정을 통해, 고차원 화학 공간에서 분자 데이터 분포를 학습하고, 실험적으로 탐색하기 어려운 새로운 분자 그래프 구조를 효율적으로 생성할 수 있다.

그러나 현재의 Diffusion 기반 그래프 생성 모델들은 분자 내 원자를 나타내는 노드 특징(Node Features, X)과 원자 간 결합을 나타내는 인접 행렬(Adjacency Matrix, A)의 복원에만 초점을 맞추고 있어, 분자의 3차원 공간적 구조를 반영하는 거리 행렬(Distance Matrix, D)을 생성하지 못한다. 이러한 3차원 공간적 정보의 부재는 약물의 결합 친화력, 효능, 부작용 예측에 한계를 초래한다.

이에 본 연구에서는 분자의 3차원 공간적 구조를 반영하는 3D 거리 행렬(Distance Matrix, D)를 Message Passing 과정에 통합하고, 확률적 미분 방정식(Stochastic Differential Equations, SDE) 시스템[1]을 활용하여 노드 특징, 인접 행렬, 그리고 거리 행렬까지 동시에 생성하는 분자 그래프 생성 방법을 제안한다. 본 모델의 주요 기여는 다음과 같다 :

- 3차원 분자 구조의 통합 모델링 : SDE 기반 그래프 Diffusion 기법을 통해 노드 특징, 인접 행렬, 거리 행렬을 동시에 생성함으로써 분자의 복잡한 3차원 구조를 효과적으로 표현한다.
- 생성 분자의 품질 향상 : 기존 분자 그래프 생성 모델과의 비교실험에서 본 모델은 (1) 중복되지 않은 분자를 생성하고, (2) 훈련 데이터셋에 포함되지 않은 새로운 분자를 생성할 수 있음을 입증하였다.

† 교신저자 : sanghyun@yonsei.ac.kr

* 이 논문은 2025년도 정부(과기정통부)의 재원으로 한국연구재단의 지원(No. RS-2023-00229822)과 국토교통부의 스마트시티 혁신인재육성사업으로 지원을 받아 수행된 연구임.

2. 방 법

2.1 입력 데이터셋 구축

분자 데이터를 그래프 구조로 표현하기 위해, 각 분자는 다음 3가지 요소로 전처리된다 :

1. 노드 특징 행렬 $X \in \{0,1\}^{N \times F}$: 각 원자의 유형(예 : C, N, O)을 원-핫 인코딩으로 나타낸다. 여기서 N은 데이터셋 내 분자의 최대 원자 수이고, F는 가능한 원자 유형의 수이다.
2. 인접 행렬 $A \in \{0,1,2,3\}^{N \times N}$: 원자 간 결합 유형을 나타내며, 0은 결합 없음, 1은 단일 결합, 2는 이중 결합, 3은 삼중 결합을 의미한다.
3. 거리 행렬 $D \in R^{N \times N}$: 원자 간 3D 공간적 유클리드 거리를 실수 값으로 표현한 행렬이다.

2.2 제안 모델

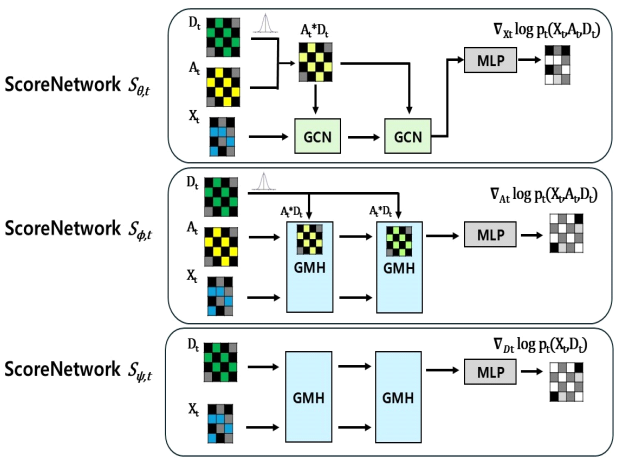


그림 1 . 제안하는 Score 모델 Framework

제안하는 Score 모델 $s_{\theta,t}$, $s_{\phi,t}$, $s_{\psi,t}$은 학습을 통해, 각 시간 단계 t에서의 데이터 분포의 확률밀도함수에 대한 기울기, 즉 Score 함수 $\nabla_{X_t}\log p_t(X_t,A_t,D_t)$, $\nabla_{A_t}\log p_t(X_t,A_t,D_t)$, $\nabla_{D_t}\log p_t(X_t,D_t)$를 추정한다. 각 Score 모델의 아키텍처는 다층 GCN 및 Multi-head Attention을 활용하여 시간에 따라 변화하는 X_t,A_t,D_t의 의존성을 포착할 수 있도록 수식 1과 같이 설계하였다. [2]

$$\begin{aligned} s_{\theta,t}(G_t) &= MLP([GCN(X_t,A_t,D_t)]) \\ s_{\phi,t}(G_t) &= MLP([GMH(GCN(X_t,A_t,D_t),A_t)]) \\ s_{\psi,t}(G_t) &= MLP([GMH(GCN(X_t,D_t),D_t)]) \end{aligned} \quad (1)$$

2.2.1 가우시안 커널을 통한 입력 처리

모델에서 입력받는 거리 행렬 D를 인접 행렬 A에 효과적으로 반영하기 위해, Score 모델 $s_{\theta,t}$, $s_{\phi,t}$에 가우시안 커널(수식 2)을 도입하였다.

$$k(d_{ij}) = \exp(-\frac{d_{ij}^2}{2\sigma^2}) \quad (2)$$

이 가우시안 커널의 값 $k(d_{ij})$은 d_{ij}(원자 i와 j 사이의 3D 공간 거리) = 0일 때 1의 최댓값을 가지며, d_{ij}가 커질수록 지수적으로 감소하여 0에 수렴한다. 이 $k(d_{ij})$는 인접 행렬 A에 가중치로서 곱해지며, 모델이 원자 간 결합 정보와 분자의 3D 공간적 거리를 동시에 학습할 수 있도록 한다.

2.2.2 손실 함수

모델의 학습을 위한 손실 함수는 모든 가능한 시간 t, 모든 원본 그래프 G_0, 그리고 그에 따른 노이즈가 섞인 그래프 $G_t|G_0$에 대해, 모델이 예측한 Score 함수과 실제 Score 함수 간 차이를 L2 norm 제곱으로 정의하였다. (수식 3)

$$\begin{aligned} &\min_\theta E_t[E_{G_0}E_{G_t|G_0}\|s_{\theta,t}(G_t)-\nabla_{X_t}\log p_{0t}(X_t|X_0)\|_2^2] \\ &\min_\phi E_t[E_{G_0}E_{G_t|G_0}\|s_{\phi,t}(G_t)-\nabla_{A_t}\log p_{0t}(A_t|A_0)\|_2^2] \\ &\min_\psi E_t[E_{G_0}E_{G_t|G_0}\|s_{\psi,t}(G_t)-\nabla_{D_t}\log p_{0t}(D_t|D_0)\|_2^2] \end{aligned} \quad (3)$$

2.3 역시간 SDE 해결을 통한 그래프 생성

Reverse Diffusion Process는 학습된 Score 모델 $s_{\theta,t}$, $s_{\phi,t}$, $s_{\psi,t}$을 사용하여, 노이즈에서 시간에 따라 데이터를 복원하며 그래프를 생성하는 과정이다. (그림2) 이 과정은 수식 4에서 역시간 SDE의 해 dX_t, dA_t, dD_t를 구하는 방식으로 진행된다.

$$\begin{aligned} dX_t &= [f_{1,t}(X_t)-g_{1,t}^2\nabla_{X_t}\log p_t(Xt,At,Dt)]d\bar{t}+g_{1,t}d\overline{w}_1 \\ dA_t &= [f_{2,t}(A_t)-g_{2,t}^2\nabla_{A_t}\log p_t(Xt,At,Dt)]d\bar{t}+g_{2,t}d\overline{w}_2 \\ dD_t &= [f_{3,t}(D_t)-g_{3,t}^2\nabla_{D_t}\log p_t(Xt,Dt)]d\bar{t}+g_{3,t}d\overline{w}_3 \end{aligned} \quad (4)$$

여기서 $[f_{i,t}(\bullet)-g_{i,t}^2\nabla\log p_t(\bullet)]d\bar{t}$ 항은 각 그래프 구성 요소 X_t, A_t,D_t에 대해 노이즈가 제거되는 방향을 나타내며, Foward Diffusion에서 노이즈가 추가되는 방향을 나타내는 함수 $f_{i,t}$ 와 Score 함수 $\nabla\log p_t(\bullet)$로 구성된다. $g_{i,t}d\overline{w_i}$항은 확산계수 $g_{i,t}$와 랜덤성을 나타내는 Wiener Process $d\overline{W_i}$가 결합되어, 역방향 과정에서 다양한 데이터를 생성하는 역할을 한다.

이 역시간 SDE를 해결하는 과정은 Symmetric Splitting 기법을 사용하여 다음과 같은 단계를 반복한다 :

1. 시간 $t=T$에서 노드 특징 X_T, 인접행렬 A_T, 거리행렬 D_T를 가우시안 분포에서 샘플링한다.
2. 각 시간 단계에서 학습된 Score 모델을 통해 Score 함수, 즉 log-likelihood의 기울기 $\nabla\log p_t(\bullet)$를 추정한다.
3. 추정된 Score 함수를 바탕으로 현재 상태에서 노이즈가 줄어든 다음 상태 X_{T-1},A_{T-1},D_{T-1}를 생성한다.
4. 위 과정을 $t=T$에서 $t=0$까지 반복하여, 최종적으로 새로운 그래프 X_0,A_0,D_0를 생성한다.

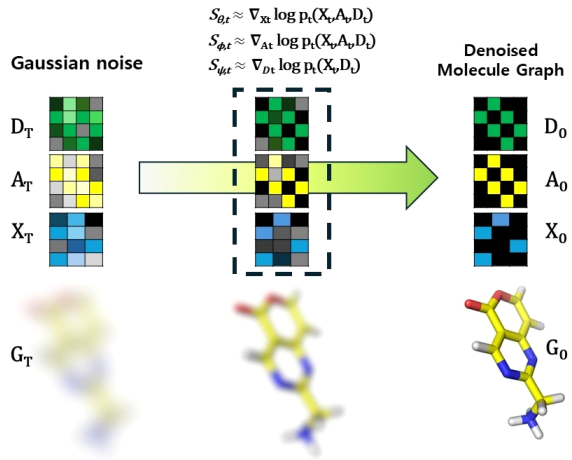
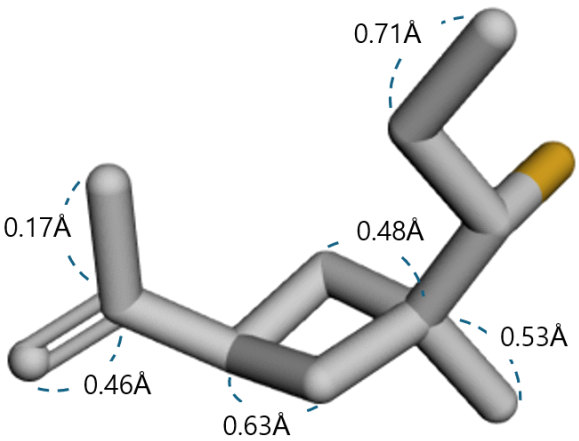


그림 2 . Reverse Diffusion 과정 시각화



SMILES : C=C(C)C1CC(C)(C(F)CC)C1

그림 3 . 모델이 생성한 분자 그래프 시각화

3. 실험 및 결과

3.1 실험 설정

제안된 모델의 성능을 평가하기 위해 두 개의 널리 사용되는 분자 데이터셋, QM9[3]와 ZINC250k[4]를 사용하였다. QM9 데이터셋은 4가지 유형의 최대 9개의 원자를 가진 133,885개의 소분자로 구성된다. ZINC250k 데이터셋은 9가지 유형의 최대 38개의 원자를 가진 249,455개의 약물 유사 분자로 구성된다. 비교분석을 위해 베이스라인 모델로 MoFlow[5], EDP-GNN[6], GraphEGM[7], GDSS[1]를 포함시켰다.

표 1. QM9과 ZINC250K 분자 데이터셋 생성 결과

Data	QM9					ZINC250k				
Method	%Valid ↑	%Unique↑	%Novel ↑	NSPDK ↓	FCD ↓	%Valid ↑	%Unique↑	%Novel ↑	NSPDK ↓	FCD ↓
MoFlow[3]	**100**	98.65	94.72	0.017	4.467	100	99.99	100	0.046	20.93
EDP-GNN[4]	**100**	99.25	86.58	0.005	**2.680**	100	99.79	100	0.049	16.73
GraphEBM[5]	**100**	97.60	**97.01**	0.03	6.143	100	98.79	100	0.212	35.47
GDSS[6]	**100**	97.80	82.12	**0.003**	2.552	100	99.64	100	**0.019**	**14.65**
Ours	**100**	**99.73**	95.47	0.015	6.032	**100**	**100**	**100**	0.216	35.70

3.2 평가 지표

모델의 성능을 평가하기 위해 생성된 10,000개의 분자 샘플에 대해 다음 5가지 지표를 사용하였다. **Validity**는 생성된 분자 중 화학 원자가 규칙을 위반하지 않는, 유효한 분자의 비율이다. **Uniqueness는** 유효한 분자 중 중복되지 않고 생성된 분자의 비율이다. **Novelty**는 훈련 세트에 포함되지 않은 유효한 분자의 비율이다. **FCD**는 분자와 훈련 데이터 간 화학 공간에서의 분포 유사성을 측정하고, **NSPDK**는 그래프 구조 관점에서 분자 그래프의 분포 학습 능력을 평가한다.

3.3 실험 결과 및 분석

표 1은 제안한 모델과 베이스라인 모델들의 성능 비교를 보여준다. 1등은 굵은 글씨, 2등은 밑줄로 표시하였다. 제안한 모델은 QM9과 ZINC250k 데이터셋 모두 **Validity**에서 100%를 달성하며 화학적 원자 규칙을 효과적으로 학습했음을 입증했다. 또한, **Uniqueness**에서 모든 베이스라인 모델을 능가하고, **Novelty**에서도 대부분의 모델보다 우수한 성능을 보였다. 반면, **NSPDK**와 **FCD**에서는 제안한 모델이 타 모델 대비 조금 낮은 성능을 보이는데, 이는 높은 Novelty로 인해 생성된 분자가 훈련 분자 데이터 분포와 다소 차이가 나는 결과로 분석된다.

3.4 Ablation Study

거리 정보를 통합한 생성 모델에서, 거리 행렬의 요소별 역수($\frac{1}{D}$)를 가중치로 사용한 모델과 가우시안 커널을 가중치로 사용한 모델을 QM9 데이터셋에서 비교 분석하였다. (표 2) 실험 결과, 가우시안 커널을 사용한 모델($\sigma=1$)이 Validity, Uniqueness, Novelty 성능에서 더 우수한 성능을 나타냈다. 이는 거리 행렬의 요소별 역수는 거리가 작을 때 지나치게 큰 가중치를 부여하여 모델의 안정성을 저해하는 반면, 가우시안 커널은 거리가 0에 가까워져도 가중치가 1에 수렴하며 거리가 가까운 원자 간의 상호작용을 적절히 반영하기 때문으로 분석된다.

표 2. 거리 정보의 역수($\frac{1}{D}$)와 가우시안 커널의 성능 비교

	%Valid ↑	%Unique↑	%Novel ↑	NSPDK ↓	FCD↓
$\frac{1}{D}$	**100**	70.65	94.50	0.046	8.689
$\sigma=0.1$	**100**	84.59	95.05	0.026	6.812
$\sigma=1$	**100**	**99.73**	**95.47**	0.015	6.032
$\sigma=2$	**100**	98.75	88.32	0.03	3.55
$\sigma=5$	**100**	97.32	85.73	**0.005**	**2.911**

4. 결론 및 향후 연구

본 연구는 원자 간 거리 행렬을 활용하여 분자의 3D 구조 정보를 반영한 Diffusion 기반 분자 그래프 생성 방법을 제안하였다. 실험 결과, 제안한 모델은 베이스라인 모델 대비 생성한 분자의 Uniqueness과 Novelty에서 우수한 성능을 입증하였다. 또한, Ablation study를 통해 가우시안 커널의 중요성을 분석하였고, σ값을 조절하며 하이퍼파라메터 튜닝을 진행하였다.

그러나 분자 그래프의 복원 지표(FCD, NSPDK MMD) 측면에서 일부 베이스라인 모델에 비해 낮은 성능을 보이는 한계가 있다. 이는 높은 참신성과 분포 유사성 간의 Trade-off로 해석된다. 또한, 생성한 3D 거리 행렬이 RDkit을 기준으로 한 거리 행렬과 일부 차이가 나는 것으로 확인되었다. 향후 연구에서는 분자 그래프를 잠재 공간에 인코딩하고 Diffusion 과정을 적용한 후 복원하는 접근을 통해 생성된 분자의 품질과 구조적 정확성을 개선하고자 한다.

5. 참고문헌

[1] Jo, Jaehyeong et al. "Score-based Generative Modeling of Graphs via the System of Stochastic Differential Equations." International Conference on Machine Learning (2022).

[2] Baek, J. et al. "Accurate learning of graph representations with graph multiset pooling" International Conference on Learning Representations (2021).

[3]Ramakrishnan, R., Dral et al. Quantum chemistry structures and properties of 134 kilo molecules. Scientific data, 1(1):1-7, 2014.

[4]Irwin, J. J. et al Zinc: a free tool to discover chemistry for biology. Journal of chemical information and modeling, 52(7):1757-1768, 2012.

[5] Zang, Chengxi et al. "MoFlow: An Invertible Flow Model for Generating Molecular Graphs." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020): n. pag.

[6] Niu, Chenhao et al. "Permutation Invariant Graph Generation via Score-Based Generative Modeling." International Conference on Artificial Intelligence and Statistics (2020).

[7] Liu, Meng et al. "GraphEBM: Molecular Graph Generation with Energy-Based Models." ArXiv abs/2102.00546 (2021): n. pag.