MolPaint: DDPM 및 RePaint 기법을 활용한 효과적인 분자구조 이미지 생성

이기정[1], 최종환[2], 최승연[2], 박상현[2†]
[1] 국민대학교 전자공학부, [2] 연세대학교 컴퓨터과학과
rlwjd4177@kookmin.ac.kr, {mathcombio, tmddus1553, sanghyun}@yonsei.ac.kr

MolPaint: Effective Molecular Structure Image Generation using DDPM and RePaint methods

Kijung Lee[1], Jonghwan Choi[2], Seungyeon Choi[2], and Sanghyun Park[2†]
[1] School of Electronic and Electrical Engineering, Kookmin University
[2] Department of Computer Science, Yonsei University

요 약

문헌에서의 화학구조 정보들은 이미지로 표현되어 있기 때문에, 신약개발을 포함한 화학정보학 연구에서 분자구조 이미지를 인식하고 이를 디지털 정보로 변환할 수 있는 기술의 개발은 중요한 과제이다. 효과적인 분자구조 이미지 인식을 위해 심층신경망이 사용될 수 있으나, 이러한 모델을 훈련하기 위해서는 많은 양의 분자구조 이미지 데이터가 필요하며, 다양한 문헌으로부터 사람이 일일이 학습용 데이터를 수집하는 것은 비용과 시간이 많이 든다. 본 논문에서는 효과적인 분자구조 이미지 데이터 수집을 위해 diffusion 생성모델 및 inpainting 기법을 활용한 분자구조 이미지 생성 기법을 제안한다. 벤치마크 데이터를 활용하여 성능 평가를 수행하였으며, 제안하는 방법이 최신 이미지 생성모델들 대비 1.8-12.1배 더 우수한 Fréchet inception distance 점수를 보여주는 것을 확인할 수 있었다.

1. 서 론

COVID-19 팬데믹으로 인해 전세계적으로 신약개발 연구에 관한 관심이 증대하였으며, 이를 위해 논문, 특허 등의 문헌으로부터 분자구조 데이터를 효과적으로 수집하는 기술의 개발은 중요한 연구과제 중 하나이다[1,2]. 대다수 문헌에서 화학구조 정보들은 Kekule 구조식과 같은 이미지로 표현되어있기 때문에 분자구조 이미지를 효과적으로 인식할 수 있는 기술의 개발은 새로운 저분자 물질 설계 및 신약 개발과 같은 연구를 위해 필수적이다. 문서 상의 화학구조 이미지를 인식하고, 이를 simplified molecular-input line-entry system (SMILES)와 같은 기계가 읽을 수 있는 표현으로 변환하는 연구분야로 optical chemistry structure recognition (OCSR)이 있다. 향상된 OCSR을 위해 다양한 심층신경망 모델이 제안되고 있으나, 이미지로 표현된 화학구조를 심층신경망이 정확하게 인식하기 위해서는 많은 양의 학습용 분자구조 이미지 데이터가 필요하다. DECIMER[3]에서 학습용 분자구조 그림 데이터를 공개하고 있으나, 이용할 수 있는 이미지가 5088개로 충분하다고 보기는 어렵다. 다양한 문헌으로부터 사람이 직접 데이터를 수집하는 것은 비용과 시간이 많이 필요로 하기 때문에, 효율적으로 분자구조 그림 데이터를 증강시킬 수 있는 기술의 개발이 필요한 실정이다.

효율적인 분자구조 그림 데이터 증강을 위해 분자구조 이미지 생성 모델이 활용될 수 있다. 문헌에서 발췌된 것과 같은 분자구조 그림들을 생성할 수 있는 기술의 확보는 학습용 데이터 준비를 용이하게 하며, OCSR을 포함한 다양한 분자구조 그림 처리성능을 향상시킬 수 있는 이점을 얻을 수 있게 한다. 따라서 본 연구에서는 분자구조 이미지를 효과적으로 생성하기 위해 RePaint 기법[4]을 활용한 MolPaint를 제안한다. MolPaint는 denoising diffusion probabilistic model (DDPM)[5]에 기반하며, 일반적인 DDPM[5] 생성과정과 다르게, 랜덤 마스크를 이용하여 다양한 화학구조 이미지를 생성하는 방법을 보여준다.

2. 본 론

MolPaint는 그림1과 같이 inpainting기법을 이용하여 분자구조 이미지를 생성하는 DDPM[5] 계열의 모델이다. MolPaint의 아키텍처 및 학습과정은 그림 1(A)와 같이 RePaint[4]의 DDPM[5]과 동일한 반면에, 새로운 분자구조 이미지를 생성하는 과정이 그림 1(B)와 같이 inpainting을 두 번 활용하는 방식으로 설계되어 있어서 보다 더 사실적인 분자구조 그림을 생성할 수 있다.

2.1. DDPM

DDPM[5]의 훈련 및 추론과정은 원본 이미지 $\boldsymbol{x_0}$ 로부터 $\boldsymbol{T}$ 시

* 본 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(IITP-2017-0-00477, (SW 스타랩) IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발)과 국토교통부의 스마트시티 혁신인재육성사업으로 지원을 받아 수행된 연구임.

† 교신저자: sanghyun@yonsei.ac.kr

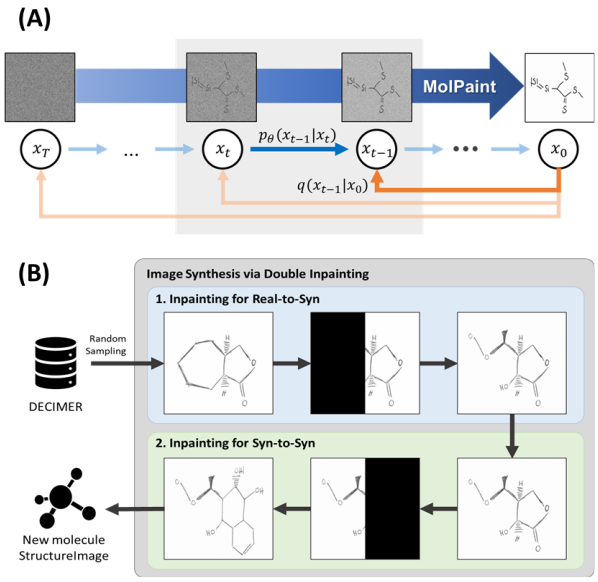

그림 1. Overview of MolPaint. **(A)** Model architecture and **(B)** Molecular structure image generation of MolPaint

간 동안 매시점마다 노이즈를 주입하여 노이즈 벡터 $\boldsymbol{x_T} \sim \boldsymbol{\mathcal{N}(0,I)}$를 만드는 forward 절차 및 $\boldsymbol{x_T}$로부터 매시점마다 노이즈 제거(denoising)를 수행하여 이미지를 점진적으로 복원하는 reverse 절차로 이루어져 있다. Forward 절차에서는 이전 시점 벡터 $\boldsymbol{x_{t-1}}$에 대한 가우시안 분포(Gaussian distribution)를 이용하여 노이즈가 추가된 다음 시점 벡터 $\boldsymbol{x_t}$를 무작위 추출한다.

$$q(x_t|x_{t-1}) \sim \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

여기서 β_t는 사용자가 지정 가능한 하이퍼파라미터(hyper-parameter)이다. 첫 시점부터 $\boldsymbol{T}$ 시점까지 재귀적으로 진행되는 과정을 정리하면, 원본 이미지 $\boldsymbol{x_0}$로부터 임의의 시점 벡터 $\boldsymbol{x_t}$를 다음과 같이 계산할 수 있다.

$$q(x_t|x_0) \sim \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$$

위 식에서 $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{t=1}^{T}\alpha_t$이다.

Forward 절차와 달리 reverse 절차는 학습이 필요한 파라미터를 가지고 있으며, 이들로 정의되는 가우시안 분포를 이용해 이전 시점 벡터를 다음과 같이 복원한다.

$$p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$$

여기서 μ_θ, Σ_θ가 학습을 통해 정의되는 평균벡터 및 공분산행렬 함수이며, DDPM[5]은 이들을 U-Net[6]을 이용하여 학습한다. μ_θ, Σ_θ를 학습하기 위한 목적함수 L은 다음과 같다.

$$L = E_q[\sum_{t>1} D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1},x_t))]$$

본 목적함수를 통해 DDPM[5]은 정의에 의해 지정된 가우시안 분포 q와 학습해야 하는 가우시안 분포 p_θ의 Kullback-Leibler divergence를 줄이는 것을 목표하며, 이는 아래와 같이 계산된다.

$$KL(p||q) = \ln\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

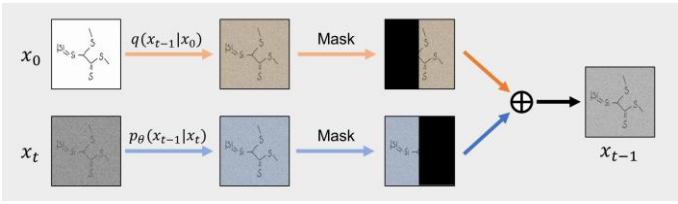

그림 2. RePaint procedure for molecular image synthesis

Ho et al. [5]은 위의 목적함수를 효율적으로 간소화하는 방법을 고안했으며, μ_θ, Σ_θ을 ϵ_θ로 함축하고, 다음과 같이 각 시점에서의 노이즈 ϵ를 예측하는 형태로 목적함수를 변형하여 DDPM [5]을 학습시킨다.

$$L_{simple} = E_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(x_t,t)||^2]$$

2.2. MolPaint

제안하는 모델은 RePaint[4]에서 제안하는 inpainting 기법을 활용하여 분자구조 그림을 효과적으로 생성하는 방법을 제안한다. 구체적으로, 그림 1(B)와 같이, 참조를 목적으로 주어진 분자구조 이미지의 일부를 가린 뒤 가려진 부분을 RePaint 기법[4]으로 복원하고, 복원된 이미지를 참조하여 이전 단계에서 가려지지 않았던 부분을 가리고 복원하는 것으로 새로운 분자구조 이미지를 합성한다. 그림 2는 RePaint[4]의 inpainting 절차를 보여준다. DDPM [5]및 두 번의 inpainting 절차를 활용하여 MolPaint는 고품질의 다양한 분자구조 이미지를 생성할 수 있다.

3. 실험 및 결과

3.1. 실험 환경

본 논문에서는 MolPaint 성능 평가를 위해 최신의 이미지 생성 모델들 DDPM[5], iDDPM[7], StyleGAN2[8]을 베이스라인 모델로 사용하였다. 벤치마크 데이터로 DECIMER[3]의 5088개의 이미지 데이터를 사용하였으며, 각각 이미지 크기를 256x256으로 조정한 후에 학습에 활용하였다. 모든 생성모델의 훈련, 이미지 생성 및 결과분석은 NVIDIA GeForce RTX 3090 이 장착된 Ubuntu 20.04 서버에서 수행되었으며, MolPaint는 Python 3.9.16 및 PyTorch 1.10.0으로 구현되었다.

3.2. 실험 결과

각 모델의 분자구조 이미지 생성 성능을 정량적으로 나타내기 위해 Fréchet inception distance (FID)를 사용했다. FID 는 Inception[10]과 같이 사전 훈련된 신경망 모델을 이용하여 각 이미지의 특징벡터를 추출하고, 원본 이미지들에 대한 특징벡터 분포와 합성 이미지들에 대한 특징벡터 분포 간의 차이를 Fréchet distance로 계산한 값이다. FID 값이 작을수록 원본에 가까운 이미지를 잘 만들어 냈다고 평가한다. 본 연구에서는 분자구조 이미지에 적합한 특징벡터 추출을 위해 최신 OCSR 모델인 MolScribe[9]를 이용하였으며, 해당 인코더가 산출한 특징벡터를 사용하여 FID 점수를 계산하였다. 각 모델로부터 1000 개의 이미지를 무작위 생성하였으며,

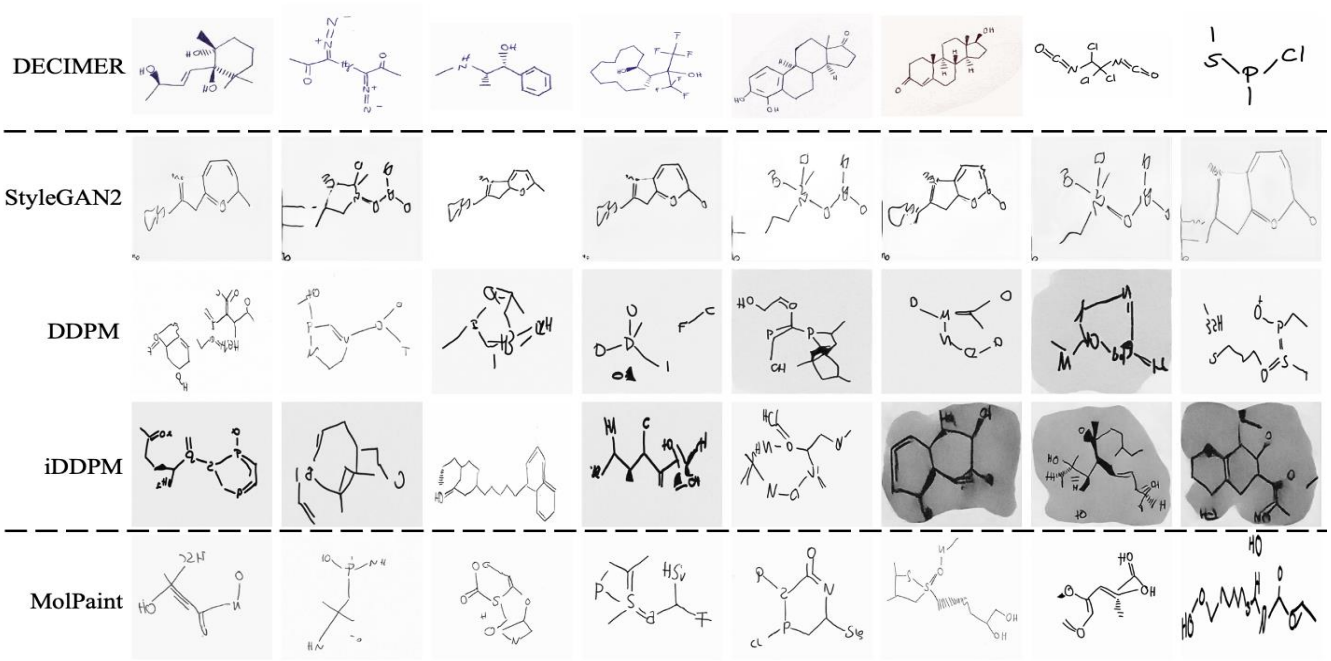

그림 3. Visualization of original molecular images of DECIMER and synthesized images by MolPaint and baseline models

표 1. Comparison of molecular image generation performance between MolPaint and baseline models

Model	StyleGAN2	DDPM	iDDPM	MolPaint
FID	1.23e-03	2.51e-04	1.87e-04	**1.02e-04**
FID/FID_{our}	12.10	2.47	1.84	1.00

DECIMER[3]와의 FID를 계산한 결과는 MolPaint가 가장 높은 성능을 가짐을 보여주었다(표 1). 그림 3은 정성적인 분석을 위해 DECIMER[3] 및 합성된 이미지 집합으로부터 무작위로 8개씩 추출한 결과를 보여준다. StyleGAN2[8]는 그럴듯한 화학구조 이미지를 만들어내지만, 다양성이 떨어지는 mode collapse 현상이 나타나는 것을 관찰할 수 있었다. DDPM[5]및 iDDPM[7]은 다양한 화학구조를 만들어내지만, 물리화학적으로 불가능한 분자구조를 보여주는 생성하는 경우를 볼 수 있었다. FID 성능이 가장 높았던 MolPaint는 다양하고 유의미한 분자구조 이미지를 생성하고 있음을 보여주어 베이스라인 모델들보다 우수한 것을 재확인할 수 있었다.

4. 결 론

본 연구에서는 DDPM[5] 및 inpainting 기법을 활용하여 분자구조 이미지를 효과적으로 생성하는 방법을 제안하였다. 제안하는 방법은 임의의 분자구조 이미지에 마스크를 씌우고, 가려진 부분을 복원하는 방식을 두 번 적용하여 완전한 분자구조 이미지를 합성할 수 있으며, DECIMER[3] 벤치마크를 통해 우수성을 확인하였다. 제안하는 방법은 OCSR 모델을 위한 학습데이터 증강 기법으로 활용될 수 있으며, 나아가 분자구조와 같이 이미지 내 구조 정보가 중요한 CT와 같은 의료영상 데이터 증강 목적으로도 활용되어 스마트병원[11]의 스마트 진단 기술 발전에 기여할 수 있을 것으로 기대한다.

참 고 문 헌

[1] El Bakri, Youness, et al. "One-pot synthesis, X-ray crystal structure, and identification of potential molecules against COVID-19 main protease through structure-guided modeling and simulation approach." Arabian Journal of Chemistry Vol. 15, No. 11, pp. 104230, 2022.

[2] Hormazabal, Rodrigo, et al. "CEDe: A collection of expert-curated datasets with atom-level entity annotations for Optical Chemical Structure Recognition." Advances in Neural Information Processing Systems Vol. 35, pp. 27114-27126, 2022.

[3] Brinkhaus, Henning Otto, et al. "DECIMER—hand-drawn molecule images dataset." Journal of Cheminformatics Vol. 14, No. 1, pp. 1-4, 2022.

[4] Lugmayr, Andreas, et al. "Repaint: Inpainting using denoising diffusion probabilistic models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[5] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems Vol. 33, pp. 6840-6851, 2020.

[6] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI Vol. 18, pp. 234-241, 2015.

[7] Nichol, Alexander Quinn, and Prafulla Dhariwal. "Improved denoising diffusion probabilistic models." International Conference on Machine Learning. PMLR, 2021.

[8] Viazovetskyi, Yuri, Vladimir Ivashkin, and Evgeny Kashin. "Stylegan2 distillation for feed-forward image manipulation." Computer Vision–ECCV pp. 170-186, 2020.

[9] Qian, Yujie, et al. "MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation." Journal of Chemical Information and Modeling. 2023.

[10] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[11] Hu, He-Xuan, et al. "Multimodal brain tumor segmentation based on an intelligent UNET-LSTM algorithm in smart hospitals." ACM Transactions on Internet Technology, Vol. 5, pp. 1-14, 2021.