# 광학 분자구조 인식 성능 향상을 위한 DDPM 기반의 분자구조 생성 및 준지도학습 연구

김진혁, 송태웅, 최종환‡
한림대학교 소프트웨어학부

rlawlsgurjh@naver.com, bbq9088@gmail.com, jonghwanc@hallym.ac.kr‡

# A Study on DDPM-based Molecular Generation and Semi-Supervised Learning for Improving the Performance of Optical Chemical Structure Recognition

Jin-Hyeok Kim, Tae-Woong Song, Jonghwan Choi‡
Division of Software, Hallym University

**요 약**

문헌자료에 나타나는 분자구조 정보를 인식하고, 분석에 용이한 형태로의 데이터 변환하는 기술은 화학정보학 데이터 수집을 용이하게 만드는 중요 정보처리 기술 중 하나이다. 딥러닝 기반의 분자구조 인식 기술이 여럿 개발되었으나, 소규모 분자구조 이미지 데이터집합에 대해서는 학습이 충분하기 어려워 인식 정확도를 향상시키기 위한 학습 전략이 필요하다. 본 연구에서는 데이터 부족으로 인한 학습 효율 저하 문제를 극복하기 위해 이미지 생성 모델을 활용한 준지도학습 알고리즘을 연구하였다. 제안하는 학습 알고리즘은 대조군 대비 5.4%p 성능 향상을 보여주었다.

## 1. 서론

광학 분자구조 인식(optical chemical structure recognition; OCSR)은 사진 속에 등장하는 분자구조를 인식하고 판독하는 기술을 말한다[1-4]. OCSR은 문헌자료에 삽입되어 있는 분자구조 정보를 효과적으로 수집할 수 있는 도구이기 때문에, 인식 정확도가 높은 OCSR 기술을 개발하는 것은 화학정보학에 대한 인공지능 응용 분야에서 중요한 과제 중 하나이다.

인식 정확도가 높은 인공지능 기반의 OCSR 모델 개발하기 위해서는 다량의 학습 데이터 확보가 필수적이다. OCSR 모델 훈련에 필요한 데이터는 1) 분자구조 이미지와 2) 대응되는 문자열 형태의 분자구조 정보이다. 이러한 데이터를 다량 확보하는 것은 수작업으로 만들어 내기에는 상당한 시간과 비용이 요구되는 문제가 있다.

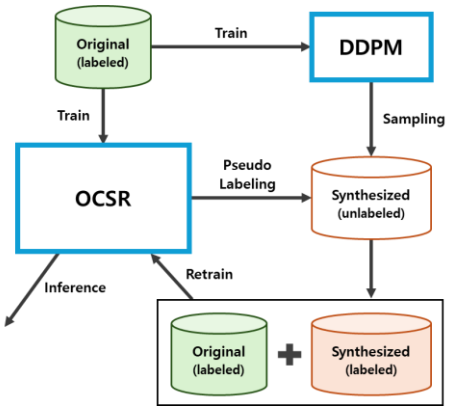
소규모의 데이터로 효과적인 OCSR 모델을 개발하기 위해 준지도학습(semi-supervised learning) 방법이 활용될 수 있다. 준지도학습은 레이블이 있는 데이터와 레이블이 없는 데이터를 모두 활용하는 학습 방법이다[5-7]. 대표적으로 의사 레이블(pseudo label)[6]이 있으며 다음과 같이 세 단계로 진행된다.

① 레이블이 있는 데이터를 이용하여 지도학습(supervised learning)을 수행
② 학습된 모델을 이용하여 레이블이 없는 데이터의 의사 레이블을 예측
③ 의사 레이블을 가진 데이터를 레이블이 있는 데이터와 통합 후 재학습(retrain) 수행

본 논문에서는 소규모 분자구조 이미지 데이터를 효과적으로 학습할 수 있는 준지도학습 기반의 OCSR 학습 전략을 제안한다. 제안하는 알고리즘은 생성형 모델(generative model)을 이용하여 분자구조 이미지 데이터를 생성하고 의사 레이블을 이용하여 OCSR의 학습 데이터를 증강한다. 벤치마크를 통해 대조군 대비 제안 알고리즘의 OCSR 성능 향상도를 평가하였다.

## 2. 방법

그림 1은 제안하는 OCSR 학습 알고리즘을 보여준다. 소규모 분자구조 이미지 데이터를 이용하여 의사 레이블 부여를 위한 OCSR 모델 및 합성 데이터 생성을 위한 DDPM(denoising diffusion probabilistic model)[8]을 구축하고, DDPM으로 합성된 데이터를 OCSR을 이용하여 의사 레이블을 부여한 뒤, 데이터




(그림 1) OCSR 모델 학습을 위한 준지도학습 전략

를 통합하여 OCSR을 재훈련하는 것을 보여준다. 만약 소규모 데이터와 유사한 정보를 가진 대규모 데이터를 이용할 수 있다면 해당 데이터로 OCSR 모델을 사전학습(pretrain)하고, 재훈련대신 파인튜닝(fine-tuning) 전략을 취할 수도 있다.

## 3. 실험 결과

### 3-1. 벤치마크 데이터

제안하는 알고리즘의 성능 평가를 위해 원본데이터(original data)로 ChemDraw 데이터[4]를 이용하였고, OCSR 모델로 MolScribe[4]를 활용하였다. ChemDraw 데이터집합은 총 5,704개의 분자구조에 대한 2D 이미지와 SMILES(simplified molecular input line entry system) 문자열 데이터 쌍으로 구성되어 있으며, 교차검증(cross validation)로 평가하기 위해 500개의 분자구조를 시험 데이터집합(test dataset)으로 설정하고, 나머지 데이터를 훈련(training) 및 검증(validation) 데이터로 사용하였다. MolScribe 저자들이 ChemDraw와는 다른 1.6백만여개의 분자구조 데이터로 사전학습된 모델을 제공하고 있기 때문에, 본 연구에서는 이를 활용하여 제안 알고리즘의 파인튜닝 성능만 평가하였다.

### 3-2. OCSR 성능 향상도 평가

OCSR 모델의 성능을 정량적으로 평가하기 위해 정답 레이블과 OCSR 모델이 예측한 SMILES 간의 유사도를 측정하는 타니모토 유사도(Tanimoto similarity; TS)[10]를 사용하였다. TS의 범위는 0~1이며, 1에 가까울수록 OCSR이 정확하게 인식하였다고 해석한다.

표 1은 ChemDraw 시험 데이터집합에 대한 3가지 버전의 OCSR 모델의 평균 타니모토 유사도 점수를 보여준다. 사전학습에 사용된 데이터는 ChemDraw와는 상이한 분자구조 이미지들이기 때문에, 사전학습만 이루어진 OCSR 모델보다는 파인튜닝을 수행한 모델이 더 나은 TS 점수를 보여주었다. 나아가, ChemDraw 데이터만으로 훈련된 경우보다 제안하는 준지도학습으로 파인튜닝하는 경우의 인식 정확도가 5.4%p 더 우수한 것을 확인하였다.

(표 1) 준지도학습에 따른 OCSR 성능 평가

| Model | Synthesized data | Tanimoto similarity |
|---|---|---|
| Pretrained | - | 0.714 |
| ChemDraw | 4,163 | 0.847 |
| ChemDraw+DDPM (ours) | 10,565 | **0.901** |

## 4. 결론

본 연구에서는 OCSR 모델의 성능 향상을 위한 DDPM 기반의 준지도학습 알고리즘을 제안하였다. 벤치마크 시험을 통해 제안하는 학습 알고리즘의 우수성을 확인하였다. 추후 연구에서는 다양한 OCSR 모델에 제안 알고리즘을 적용해볼 것이며 의사 레이블 기능을 고도화하기 위한 기술을 연구할 계획이다.

## 감사의 글

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2024-00345226)

## 참고문헌

[1] Clevert, Djork-Arné, et al. "Img2Mol–accurate SMILES recognition from molecular graphical depictions." *Chemical science* 12.42 (2021): 14174-14181.

[2] Xu, Zhanpeng, et al. "SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer." *Journal of Cheminformatics* 14.1 (2022): 41.

[3] Rajan, Kohulan, et al. "DECIMER 1.0: deep learning for chemical image recognition using transformers." *Journal of Cheminformatics* 13 (2021): 1-16.

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

[5] Lee, Dong-Hyun. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks." *Workshop on challenges in representation learning, ICML*. Vol. 3. No. 2. 2013.

[6] He, Junxian, et al. "Revisiting self-training for neural sequence generation." *arXiv preprint arXiv:1909.13788* (2019).

[7] Amini, Massih-Reza, et al. "Self-training: A survey." *arXiv preprint* arXiv:2202.12040 (2022).

[8] Nichol, Alexander Quinn, and Prafulla Dhariwal. “Improved Denoising Diffusion Probabilistic Models” *International Conference on Machine Learning*. PMLR, 2021

[9] Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." *Journal of chemical information and computer sciences* 28.1 (1988): 31-36.

[10] Dávid Bajusz, et al. “Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?” *Journal of Chemical Information and Modeling* 2010. Vol. 50, No. 5, pp.742-754