논문 요약 보고서

논문 제목: Analyzing Learned Molecular Representations for Property Prediction
저자명: Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay
출판 정보: J. Chem. Inf. Model. 2019, 59, 8, 3370−3388

- 소개 배경

분자의 물성(property) 예측을 위한 알고리즘에 대한 연구가 활발하게 진행되고 있지만, 가장 대표적인 두 방법인 fixed molecular fingerprints을 적용한 neural networks와 graph convolutional neural networks 중 어떤 것이 더 우수한지 명확하게 결정짓지 못했다. 이에 현재 산업에서 사용되고 있는 기존의 모델보다 상당히 개선된 새로운 GCN model 소개한다.

- 결과

새로운 모델인 D-MPNN은 GCN에 속하며, 기존 논문에서 소개된 바 있는 Message Passing Neural Network (MPNN)를 기반으로 한다. 성능 향상을 위해 Bayesian Optimization을 이용해 Hyperparameter 최적화를 실시하고, noise를 줄이기 위해 앙상블(Ensemble)기법과 Cross Validation을 이용한다. 19개의 공개dataset와 16개의 독점dataset에 대해 850개 이상의 실험을 수행하여 fixed descriptor 또는 학습된 분자 표현에 기초한 분자 속성 예측 모델 비교를 수행하였다.

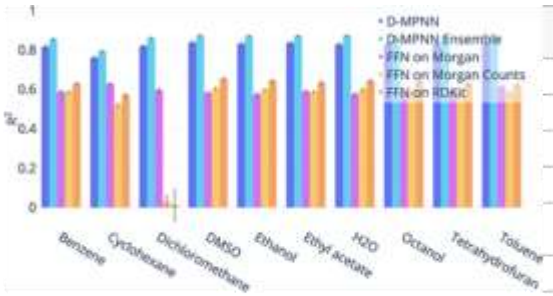

Figure 7. D-MPNN 모델과 baseline models의 비교 (regression data sets)

baseline	D-MPNN is *better*	D-MPNN is the *same*	D-MPNN is *worse*	no. of data sets
Molecule Net	5	3	2	10
Mayr et al.	8	10	1	19
RF on Morgan	14	0	1	15
FFN on Morgan	14	5	0	19
FFN on Morgan Counts	15	4	0	19
FFN on RDKit	8	5	4	19

Table 8. D-MPNN이 baseline 모델보다 성능이 통계적으로 높거나 낮은 public dataset의 수

결과적으로, 이 논문에서 구축한 새로운 모델인 D-MPNN은 모든 데이터 집합에 걸쳐 일관되게 강력하며, 모든 경우의 실험에 대해 baseline(convolution 또는 fingerprint 기반 모델)의 성능과 일치하거나 이를 능가한다.

- 시사점

D-MPNN 모델은 분자의 물성 예측이 필요한 다양한 분야의 연구에서 활용 및 확장될 수 있다. 특히 약물 발견에 종사하는 화학자들을 위한 강력한 도구로 사용이 가능하다
분자의 물성(property) 예측에 있어서 현재 최고의 model 성능을 능가하며 안정적인 모델인 D-MPNN을 제안한다. D-MPNN이 성능이 떨어지는 두 가지 경우는 Dataset의 불균형이 심할 때와 3D 정보가 사용되는 경우인데, 앞으로 이를 해결하기 위한 방향의 개선 연구가 필요하다.