인공지능을 이용한 End-to-End 방식의 오픈소스 서비스 성능 최적화 및 화학 분야 응용

: A Data-driven Approach for Efficient Physical Chemistry

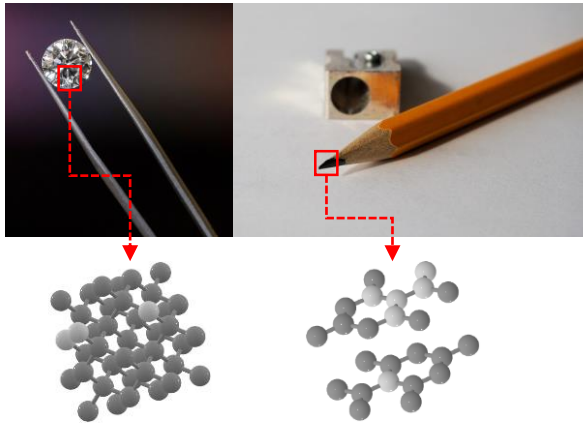
나경석
한국화학연구원 (KRICT)

Contents

01 신물질 개발과 계산 과학

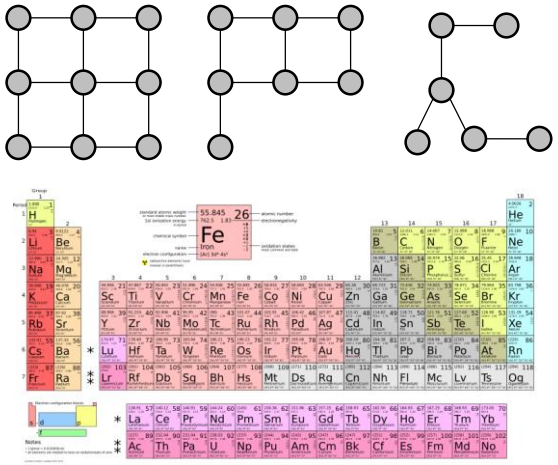
- 식품, 소재, 의약품, 태양 전지 등 세상에 존재하는 것은 기본적으로 분자, 결정 구조, 단백질 등 원자의 배열로 구성된 화합물로 이루어져 있다.
- 신물질 개발의 목적은 원자의 종류와 배열로 만들어지는 조합들을 고려하여 우리가 원하는 특성의 새로운 화합물을 만드는 것이다.
- 그러나 원자의 종류와 배열로 만들어지는 조합의 수는 무한에 가깝기 때문에 모든 후보물질을 검증하는 것은 불가능하며, 효율적인 검증을 위한 방법론이 필요하다.
- 계산 과학은 물리 및 화학 이론을 바탕으로 주어진 원자 배열에 대한 구조 최적화 및 특성 계산을 수행하며, 효율적인 신물질 개발을 위해 다양한 계산 과학 방법론이 활용되고 있다.

우리 생활에 존재하는 다양한 화합물
(음식, 첨단소재, 의약품, 컴퓨터 부품 등)



다이아몬드 구조

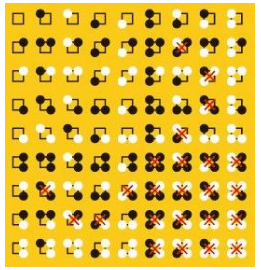
그라파이트 구조



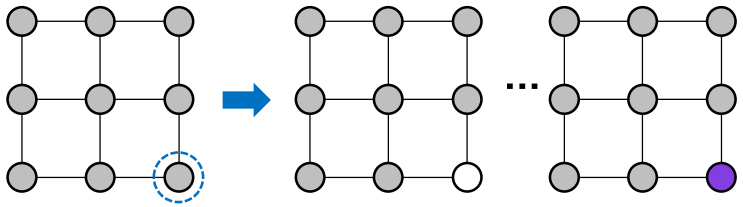
01 신물질 개발과 계산 과학

- 일반적으로 19×19 공간에서 두는 바둑은 우주에 있는 원자의 수보다 큰 경우의 수를 갖는다고 말하며, 한동안 인공지능의 성능을 평가하기 위한 기준으로써 많이 활용되었다.
- 그러나 신물질 개발을 위한 원자 배열 문제에서는 바둑보다 더 다양한 경우의 수를 고려해야하는 어려움이 있다.
- 원자의 배열은 원자의 종류, 원자의 연결 종류, 원자의 수 등을 모두 고려해야 하기 때문에 원자 배열에 대한 모든 경우의 수를 실험적으로 검증하는 것은 불가능하다.

2016 서울에서 있었던 알파고와 이세돌의 구글 딥마인드 챌린지

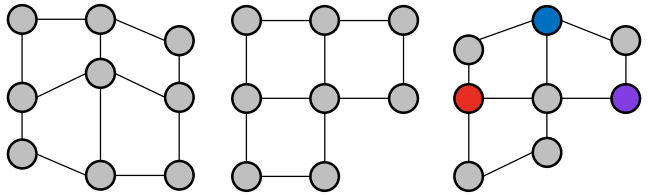


2×2 바둑판의 경우의 수



수소, 탄소, …

약 180개의 경우의 수



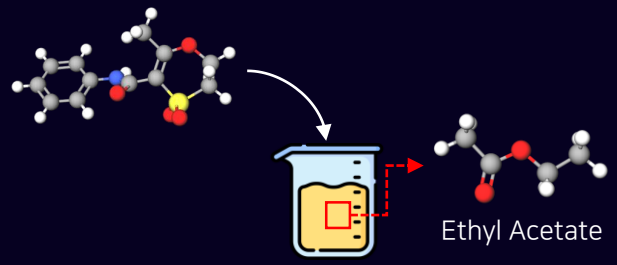
무한에 가까운 경우의 수

계산 과학 방법론은 우리가 원하는 물성을 가질 것으로 예상되는 화합물을 효율적으로 탐색하기 위해 사용된다.

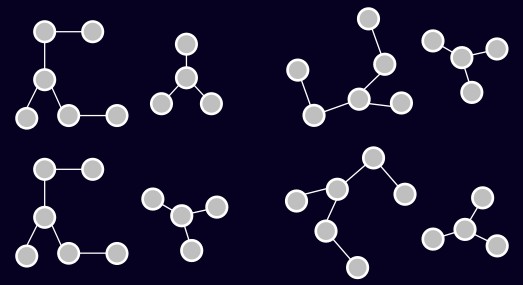
02 전통적인 계산 과학 방법론의 한계점

- 일반적으로 계산 과학 방법론은 분자의 크기에 대해 다항 또는 지수 시간 복잡도를 갖고 있기 때문에 실세계에 존재하는 분자에 대한 계산 과학 방법론의 적용은 매우 어렵다.
- 실제 공학 및 산업에서 사용되는 화합물은 대부분 두 개 이상의 화합물의 상호작용 또는 혼합으로 만들어지기 때문에 화합물의 특성을 파악하기 위한 계산량은 더욱 크게 증가한다.

단일 분자의 구성과 분자 특성



분자-분자 상호작용에 의한 화학 반응



분자-분자 상호작용에 대한 경우의 수

분자-분자 상호작용에 대한 경우의 수 = 분자1의 구성에 대한 경우의 수 × 분자2의 구성에 대한 경우의 수 × 기하학적 배열에 대한 경우의 수

- 분자와 분자의 상호작용에서는 각 분자의 구성뿐만 아니라, 분자와 분자가 만나거나 결합할 때의 기하학적 배열에 의해서도 결과가 달라진다.
- 분자-분자 상호작용을 효율적으로 계산하기 위한 다양한 방법론이 제안되었지만, 화학 분야 전문가가 직접 초기값을 설정해야하는 것과 여전히 높은 계산량에 대한 한계점이 있다.
- 기존 계산 소프트웨어는 오픈소스가 아니기 때문에 사용을 위한 많은 비용이 발생하며, 사용자 필요에 의한 확장이 불가능하다.

03 오픈소스 기반 계산 방법론 효율화

ORCA (https://orcaforum.kofo.mpg.de/app.php/portal)

- Semi-empirical methods, many-body perturbation 등을 구현한 다목적 양자화학 방법론
- 공개 소프트웨어이며, 기존 코드와의 결합이 용이함
- 효율적인 양자화학 계산을 위한 다양한 방법론을 제공
- 상용 소프트웨어보다는 낮은 계산 정확도

연구 목표: 활용이 편리하지만, 상용 소프트웨어에 비해 정확도가 낮은 공개 소프트웨어를 이용하여 서비스를 고도화

1 → 2 → 3 → 4

난수 기반 초기 구조 설계

분자-분자 상호작용 데이터베이스의 각 항목에 대해 물리 및 화학 기반지식을 바탕으로 주어진 두 분자에 대한 초기 기하학적 구조를 생성하기 위한 방법론을 설계

파이썬 및 RDKit 결합

설계한 초기 구조 생성 알고리즘에 따라 머신러닝 모델이 읽을 수 있는 형태의 분자-분자 상호작용 구조를 생성하는 파이썬 스크립트 작성하고, 이를 통해 분자-분자 상호작용의 초기 구조 생성을 자동화

ORCA를 이용한 대량 계산

난수를 기반으로 생성한 분자-분자 상호작용 초기 구조에 대해 ORCA를 이용하여 양자화학적 구조 최적화를 수행하고, 계산된 결과를 정형화된 형태로 저장하여 프로그램을 통해 읽을 수 있도록 구현

학습 데이터셋 구축

파이썬 스크립트와 ORCA를 이용하여 생성한 분자-분자 상호작용에 대한 구조 최적화 결과를 구조 파일과 메타데이터의 형태로 저장하고, 이를 학습 데이터셋으로 활용할 수 있도록 PyTorch 기반 데이터로 변환

04 Representation Learning을 위한 확률 모델

- 본 연구의 핵심이 되는 내용은 "공개 소프트웨어를 활용한 것 " 이 아니라, 공개 소프트웨어를 활용하여 "연구에 필요한 데이터를 생성한 것"이다.
- 연구에서 생성한 분자-분자 상호작용 화학 데이터는 기존 화학데이터와 다른 특성 및 목적을 가지며, 두 데이터의 특징과 비교는 아래의 표와 같다.

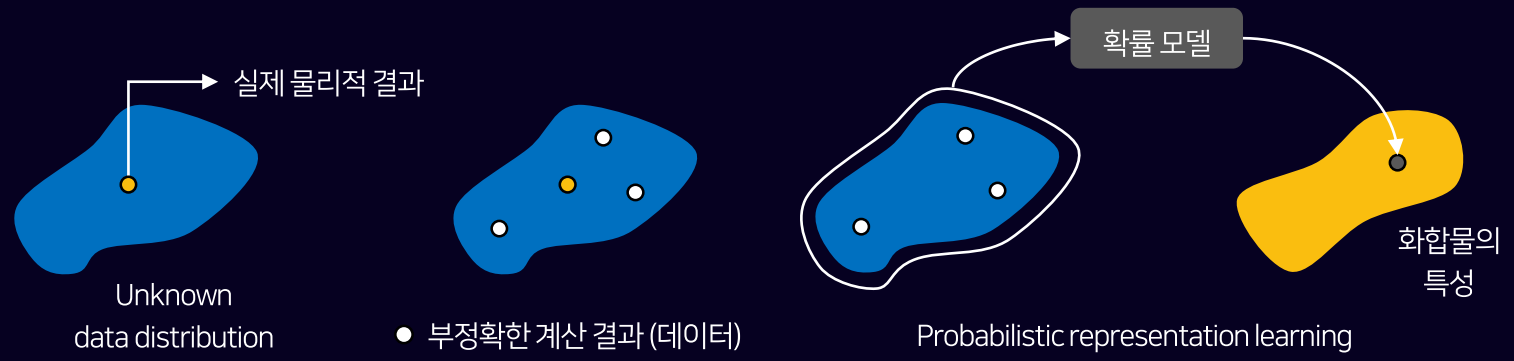
	기존 화학 데이터	본 연구의 화학 데이터
계산 정확도	매우 정확하지만 물리 및 화학 분야 전문가에의한 실행을 요구	상용 소프트웨어에 비해 부정확하지만 자동화가 편리함
계산량	정확한 계산 결과를 얻기 위해 많은 양자화학 계산이 필요	다소 부정확하지만 빠르게 계산을 수행하도록 소프트웨어를 설정
데이터의 양	하나의 관측에 대해 하나의 화학 데이터를 생성	하나의 관측에 대해 난수를 이용하여 여러 화학 데이터를 생성
계산 설정값	하나의 관측에 대해 매우 정확한 계산 결과를 얻기 위해 각 관측에 대해 특화된 계산 설정값을 이용	다양한 관측에 대한 계산 수행을 자동화하기 위해 동일한 계산 설정값을 이용
목적	관측된 현상에 대한 물리 및 화학 분야 전문가의 해석을 보조	관측된 현상들에 대한 머신러닝 모델 구축을 위해 학습 데이터셋으로 활용

다소 부정확한 대량의 데이터로부터 정확한 예측 및 생성 모델을 구축하기 위한 머신러닝 방법론이 필요

04 Representation Learning을 위한 확률 모델

연구에서의 기본 가정

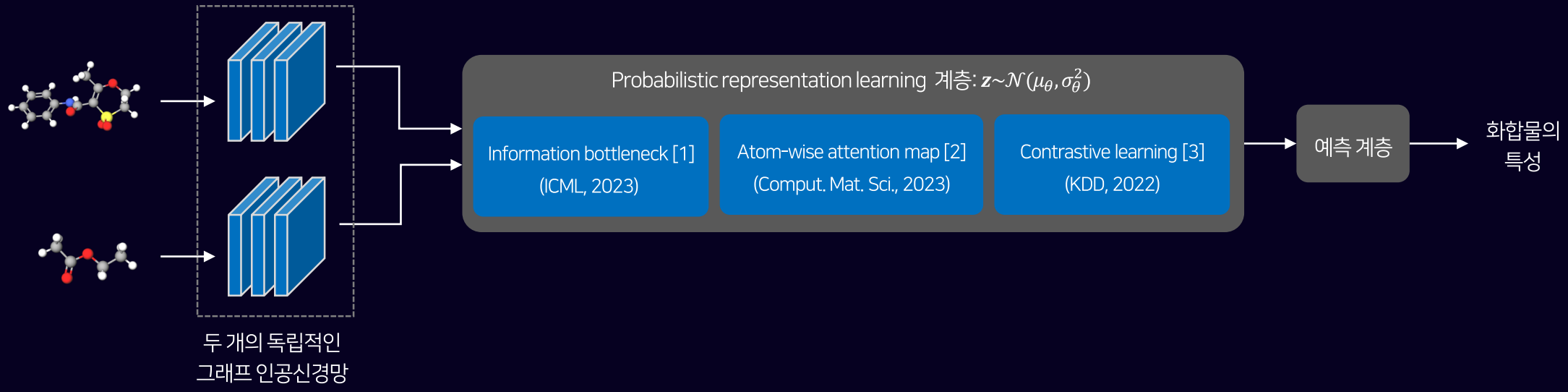
하나의 관측에 대한 여러 부정확한 계산 결과는 관측에 대한 데이터 분포를 형성한다.



- 실제 물리적 결과와 가장 일치하는 정확한 데이터가 없더라도 다수의 부정확한 데이터에 대한 probabilistic representation learning을 수행하여 정확한 예측 모델을 구축한다.
- 실제 물리적 결과가 나타내는 최적 구조의 분포는 모르지만, 연구 가정을 기반으로 부정확한 계산 결과를 분포에서 생성된 샘플로 생각하여 확률 모델을 설계한다.
- 확률 모델은 probabilistic representation learning을 통해 샘플이 나타내는 분포를 특성 도메인의 화합물 특성으로 사상한다.

04 Representation Learning을 위한 확률 모델

- 머신러닝 모델은 두 개의 독립적인 그래프 인공신경망과 probabilistic representation learning 계층, 예측 계층으로 구성된다.
- Probabilistic representation learning 계층은 information bottleneck, atom-wise attention map, contrastive learning을 조합하여 개발하였다.
- 정확도가 낮은 여러 계산 결과를 하나의 분포에서 생성된 샘플로 간주하여 probabilistic representation learning을 수행함으로써 분자-분자 상호작용에 대한 예측 정확도를 향상시킨다.



[1] Lee, N. et al. (2023). Conditional graph information bottleneck for molecular relational learning. ICML, (pp. 18852-18871). PMLR.
[2] Na, G. S. (2023). Substructure interaction graph network with node augmentation for hybrid chemical systems of heterogeneous substructures. *Comput. Mat. Sci.*, *216*, 111835.
[3] Na, G. S., & Park, C. (2022). Nonlinearity encoding for extrapolation of neural networks. KDD (pp. 1284-1294).

05 확률 모델 기반의 서비스 성능 최적화

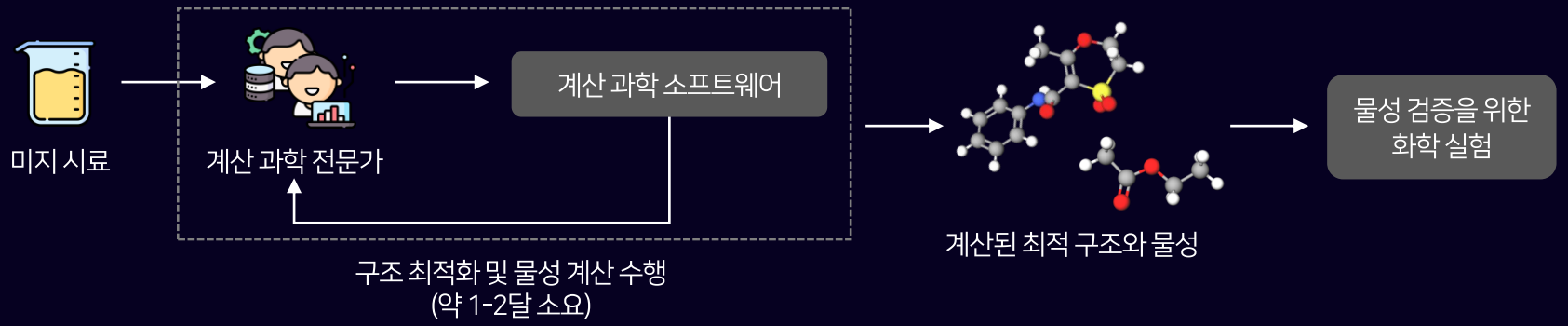
- 개발된 확률 모델의 성능 평가를 위해 각각 약 18,000, 550, 2200 개의 분자-분자 상호작용 데이터를 포함하는 Chromophore, FreeSolv, MNSol 데이터셋을 사용하였다.
- 성능 평가에 이용된 3개의 데이터셋은 상호작용이 발생하는 분자쌍 (입력 데이터)과 상호작용에 의한 결과로 나타나는 광학 및 물리화학적 성질 (출력 데이터)로 구성된다.
- 비교 분석을 위해 기존 AttentiveFP를 변형한 모델과 분자-분자 상호작용 예측을 위해 제안된 두 모델 (CIGIN, CGIB)의 성능을 같이 평가했다.
- 아래의 표와 같이 제안하는 방법론은 모든 예측 작업에서 기존의 다른 방법론보다 더 낮은 예측 오차 (RMSE)를 보여주었다.

머신러닝 방법론	Chromophore			FreeSolv	MNSol
	Absorption max	Emission max	Lifetime		
AttentiveFP [1]	21.25 (0.33)	28.84 (0.31)	0.85 (0.02)	1.05 (0.02)	0.69 (0.01)
CIGIN [2]	19.32 (0.35)	25.09 (0.32)	0.80 (0.01)	0.91 (0.01)	0.61 (0.02)
CGIB [3]	18.11 (0.38)	23.90 (0.35)	0.77 (0.01)	0.85 (0.02)	0.54 (0.01)
Our method	**16.05 (0.28)**	**20.14 (0.28)**	**0.69 (0.01)**	**0.80 (0.01)**	**0.47 (0.02)**

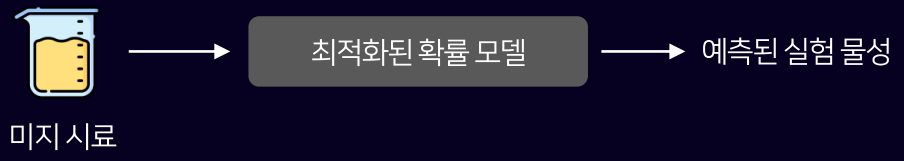
[1] Xiong, Z. et al. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J. Med. Chem., *63*(16), 8749-8760.
[2] Pathak, Y. et al. (2020). Chemically interpretable graph interaction network for prediction of pharmacokinetic properties of drug-like molecules. AAAI (Vol. 34, No. 01, pp. 873-880).
[3] Lee, N. et al. (2023). Conditional graph information bottleneck for molecular relational learning. ICML, (pp. 18852-18871). PMLR.

05 확률 모델 기반의 서비스 성능 최적화

- 화학연에서는 신물질 개발 과정을 효율화하기 위해 다양한 데이터 기반 서비스를 개발하고 있으며, 연구된 확률 모델을 신물질 개발 서비스에 이식하는 작업을 수행하고 있다.
- 기존의 전통적인 방식은 계산 과학 전문가가 직접 최적 구조를 계산해야하기 때문에 약 1-2달 정도 소요되는 계산 과학 작업이 필요했다.



- 개발된 확률 모델을 이용하여 계산 과학을 통한 구조 최적화 과정을 대체할 수 있으며, 이를 통해 서비스의 구조를 단순화할 수 있다.
- 서비스 구조의 단순화뿐만 아니라, 확률 모델을 이용하여 분자-분자 상호작용에 의한 실험 물성을 매우 빠르게 예측할 수 있기 때문에 서비스의 효율성을 크게 향상시킬 수 있다.



06 연구 결과 및 결론

- 본 연구에서는 계산 과학 상용 소프트웨어 사용 시에 발생하는 비용 및 효율성의 한계점을 극복하기 위해 공개 소프트웨어와 인공지능 기술을 결합했다.
- 개발된 확률 모델은 비교적 낮은 정확도 수준에서 생성된 데이터를 기반으로 probabilistic representation learning을 수행하여 높은 예측 정확도를 달성했다.
- 본 연구를 신물질 개발 서비스에 적용함으로써 서비스 과정을 단순화하고, 예측 과정의 효율성을 향상시켰다.
- 향후 작업에서는 개발된 probabilistic representation learning 모델에 대한 고도화 및 더욱 다양한 화학 데이터셋에서의 검증을 수행할 계획이다.

감사합니다.