

AI로 가속된 계산화학 연구

Universal ML-potential model의 전이학습과 응용 연구

학생명 **강민기**
지도교수 **신재윤**
팀명 **(ML | QM)**

Korea University, Sejong

Department of Advanced Material Chemistry

kangmg@korea.ac.kr

프로젝트 학기 발표회

2024.12.10

Background

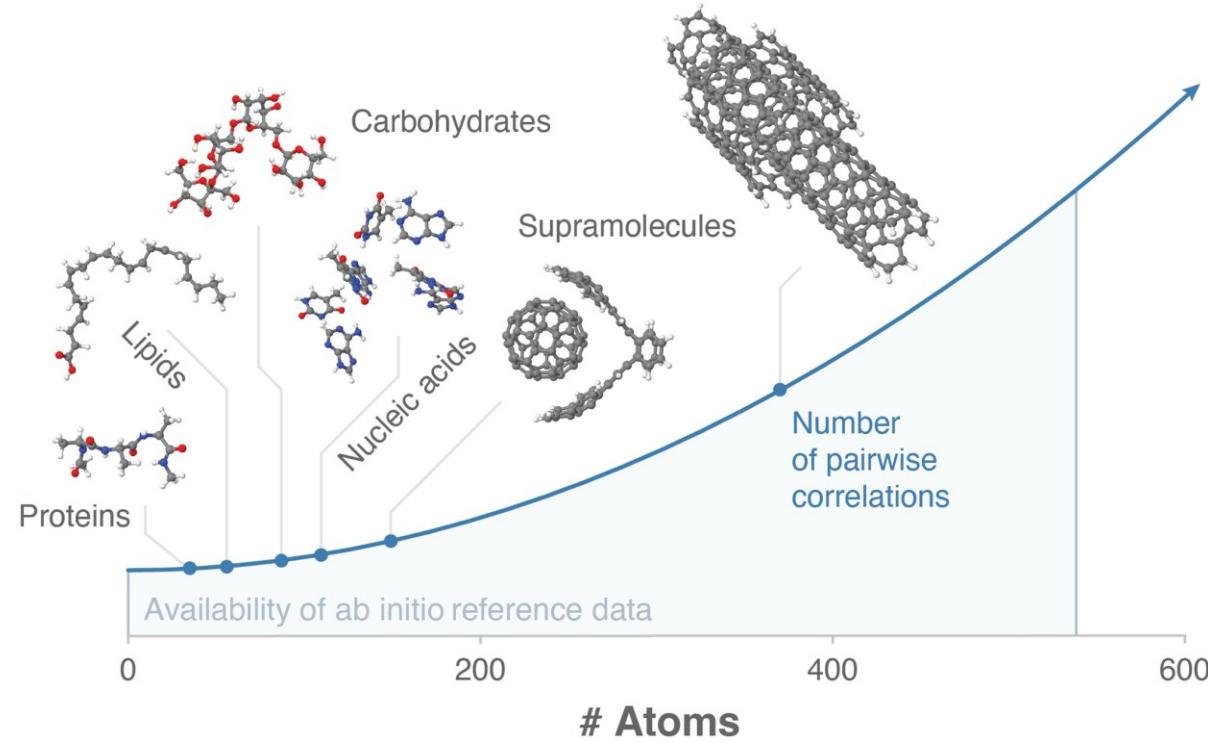
방법론의 선택

Accuracy vs. Cost



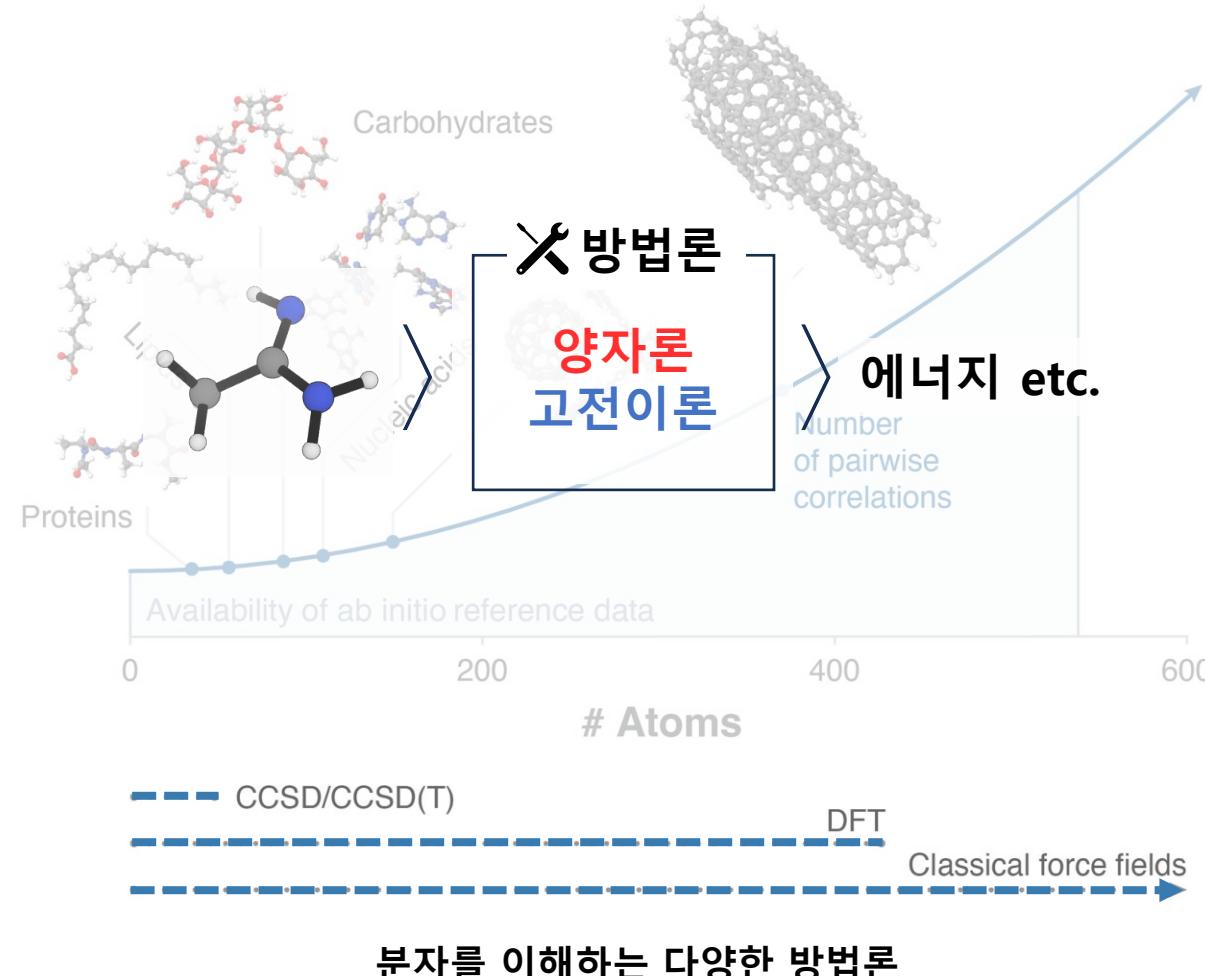
Background

✓ Quantum & Classical Methods



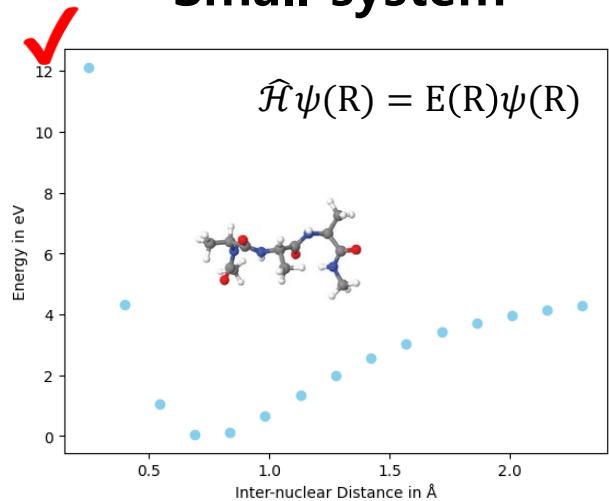
Background

✓ Quantum & Classical Methods



Background

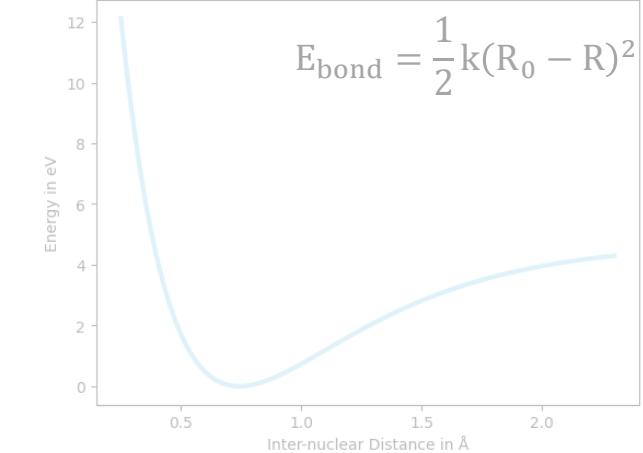
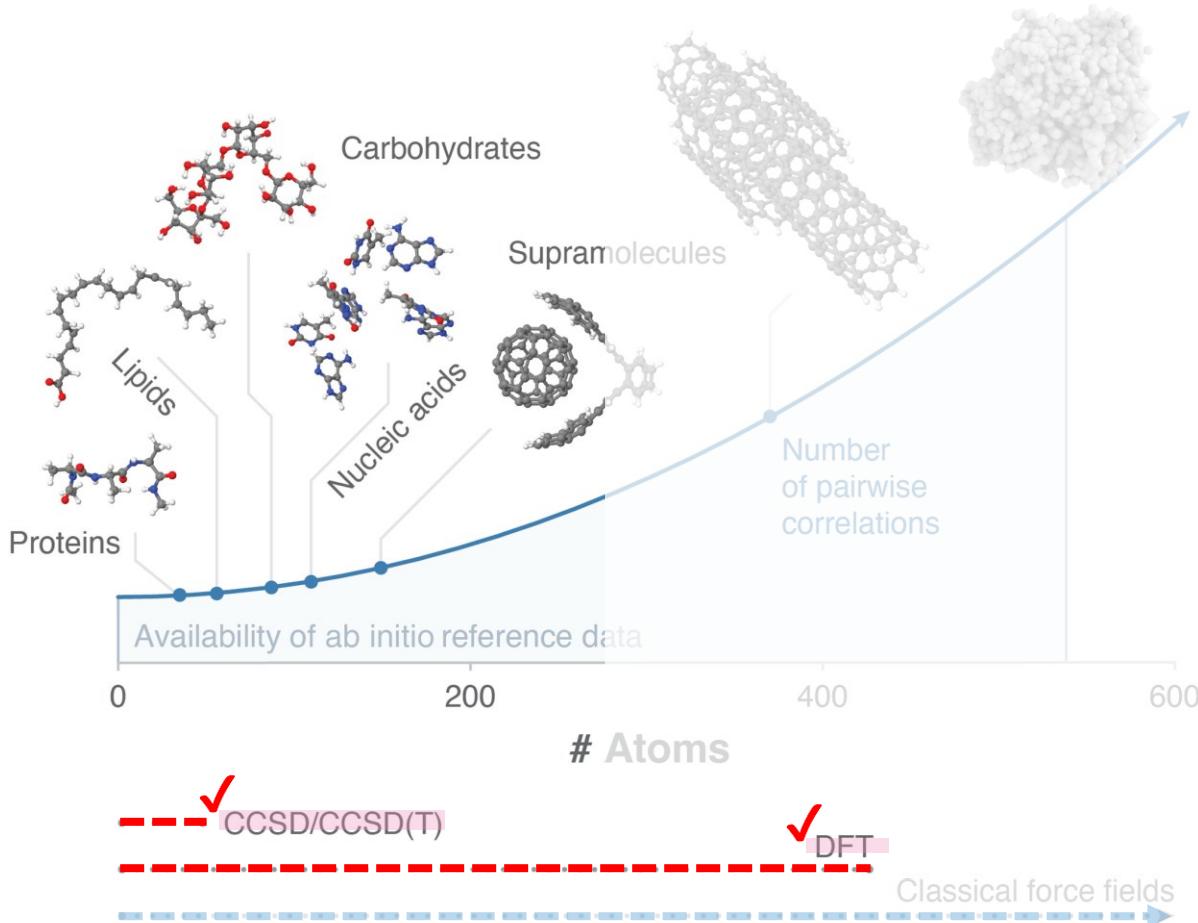
Small system



양자론(ab initio)



Accuracy

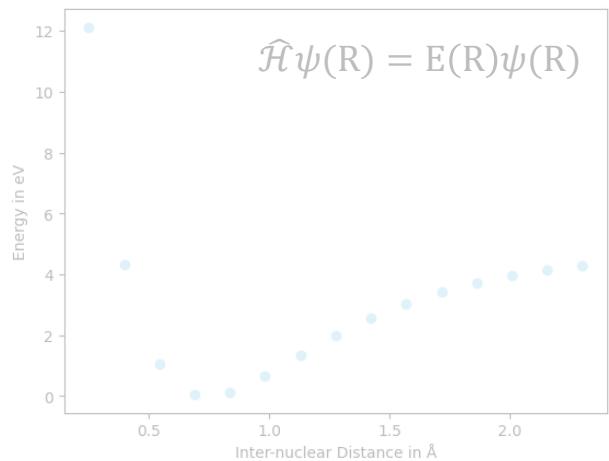


고전이론



Cost

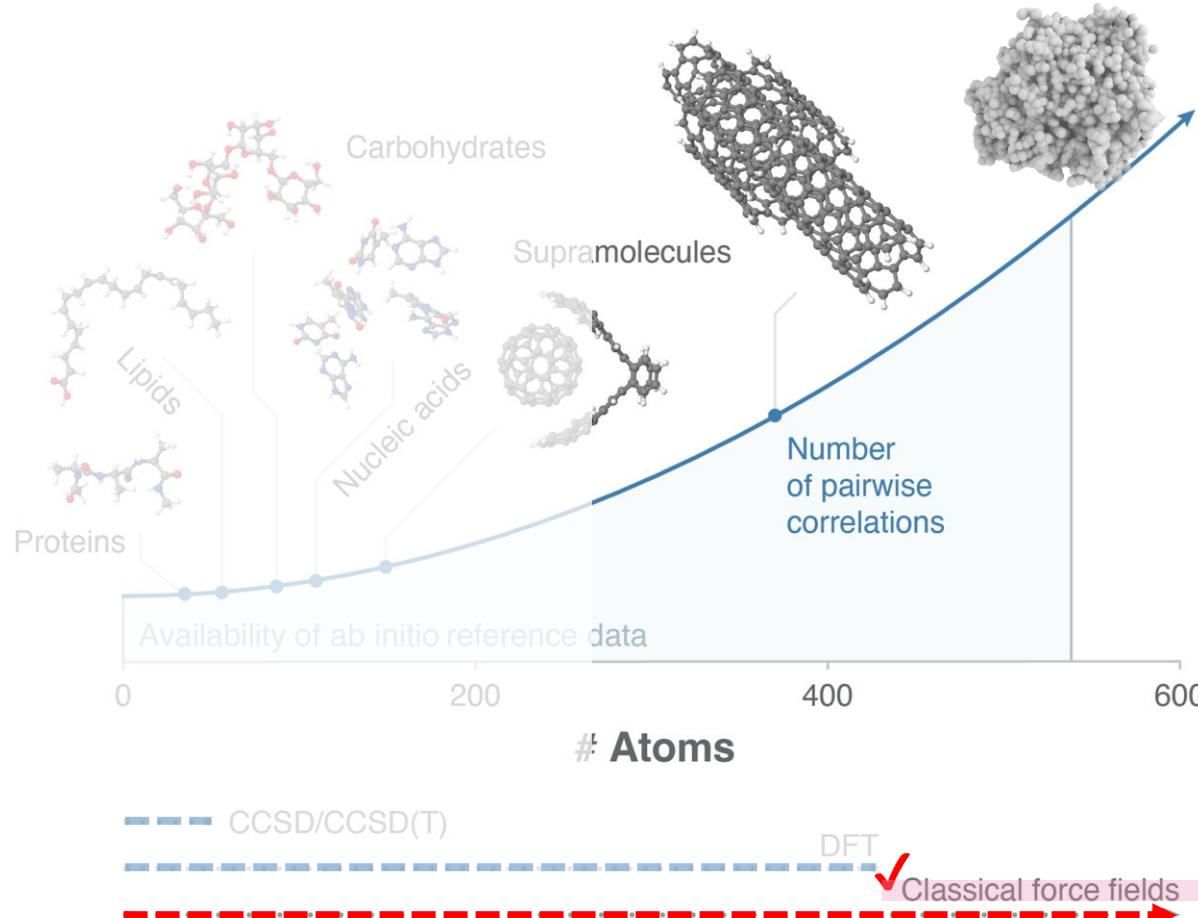
Background



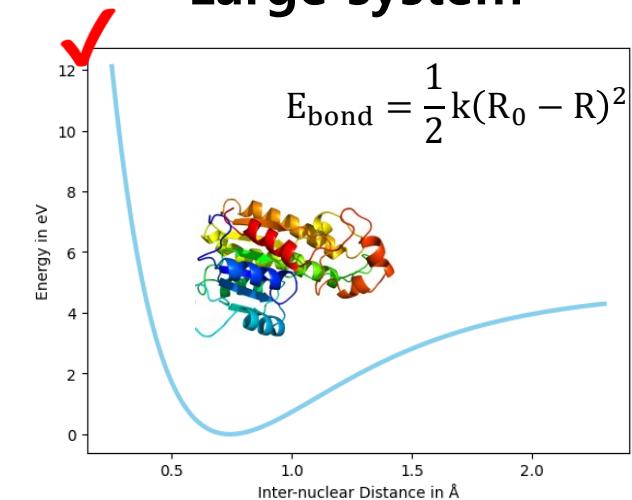
양자론(ab initio)



Accuracy



Large system



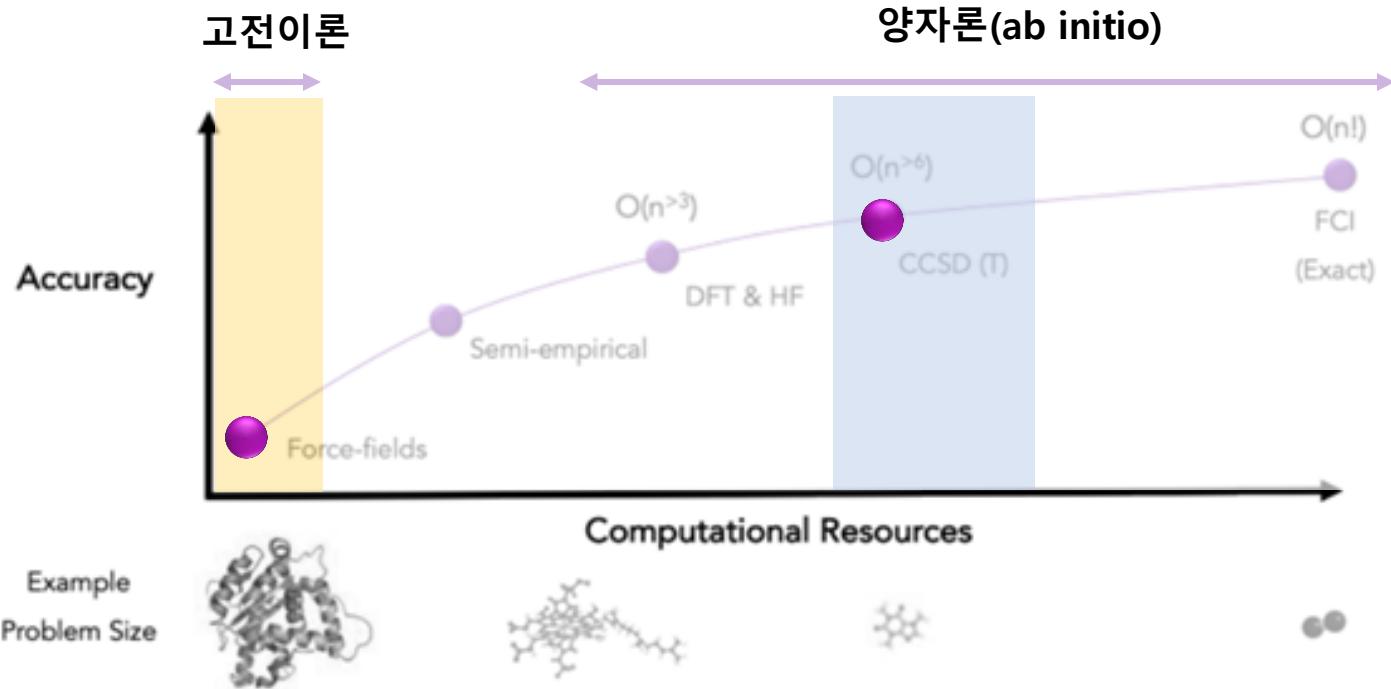
고전이론



Cost

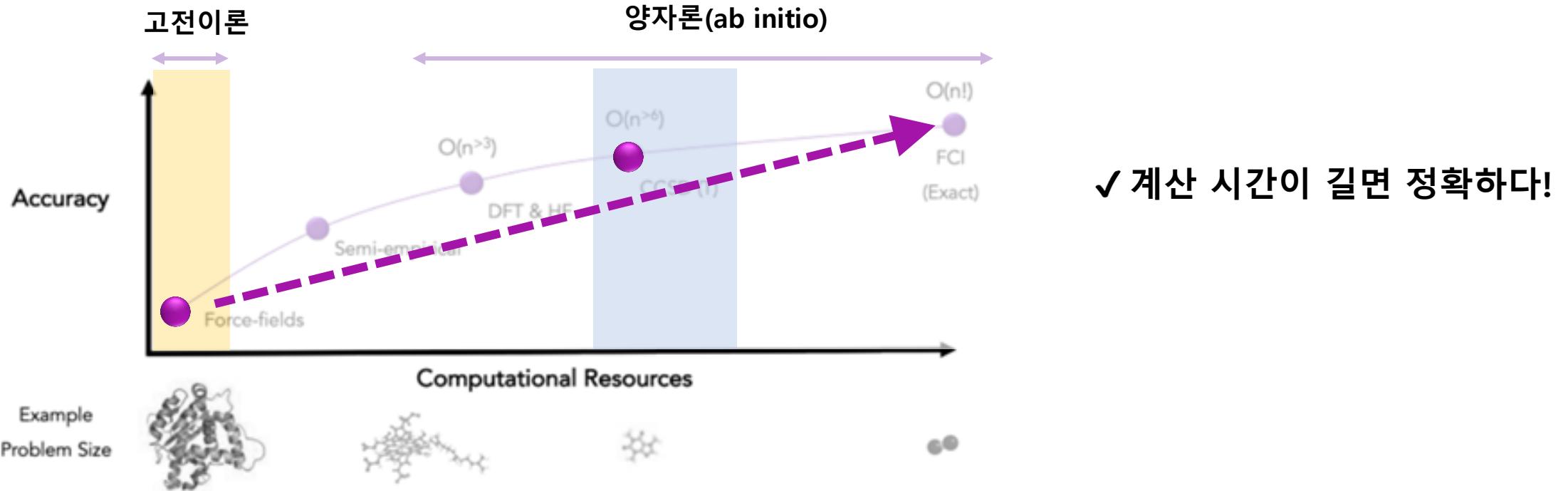
Background

✓ Trade-off btw. **Cost** vs. **Accuracy**



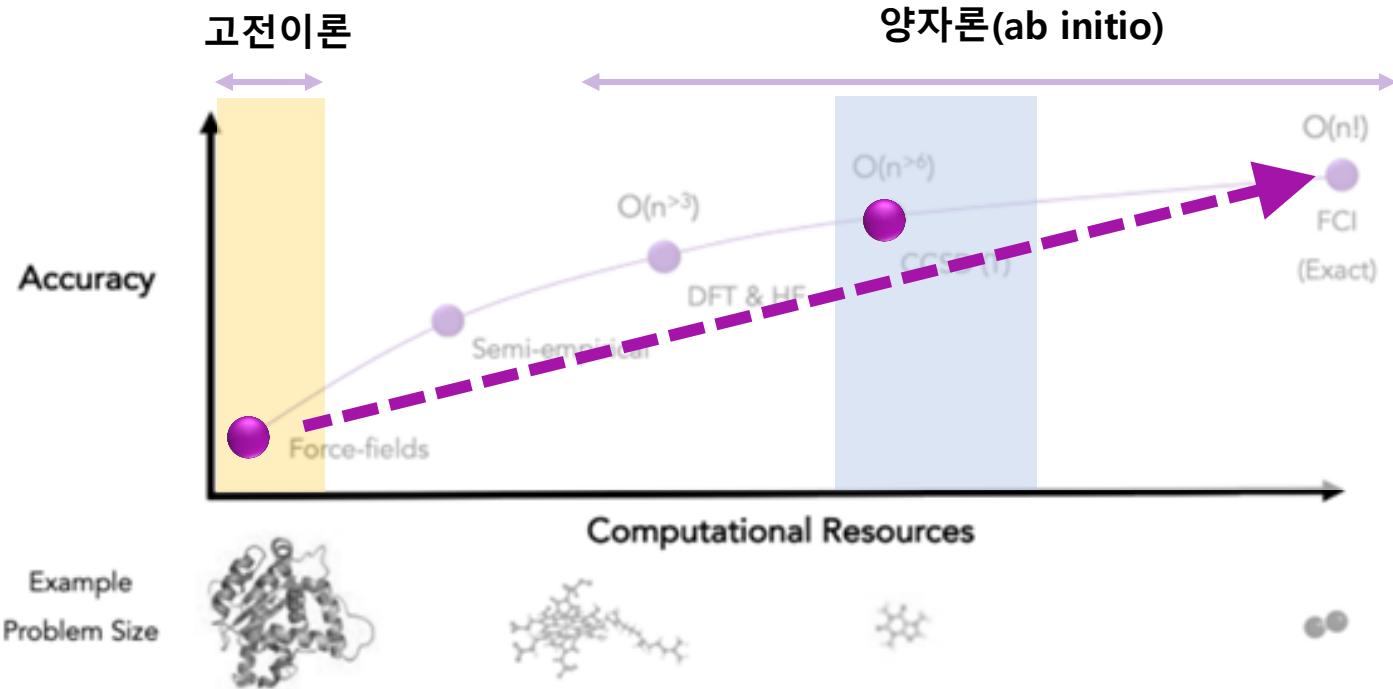
Background

✓ Trade-off btw. **Cost** vs. **Accuracy**



Background

✓ Trade-off btw. **Cost** vs. **Accuracy**



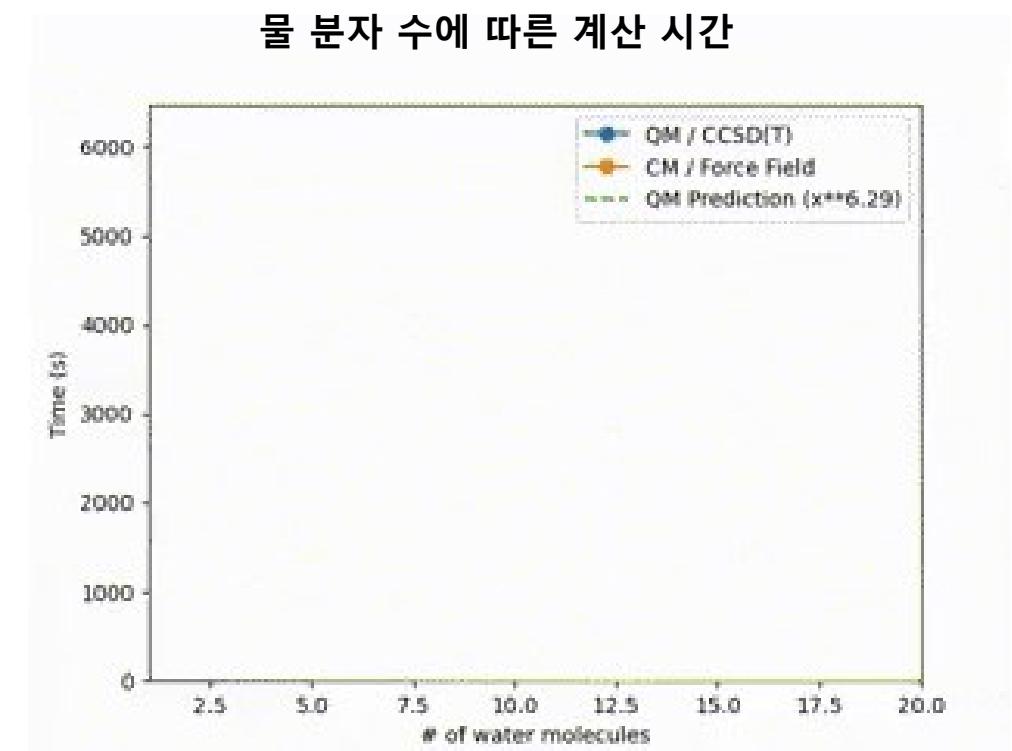
✓ 계산 시간이 길면 정확하다!
얼마나 길면?

Background

✓ Trade-off btw. **Cost** vs. **Accuracy**

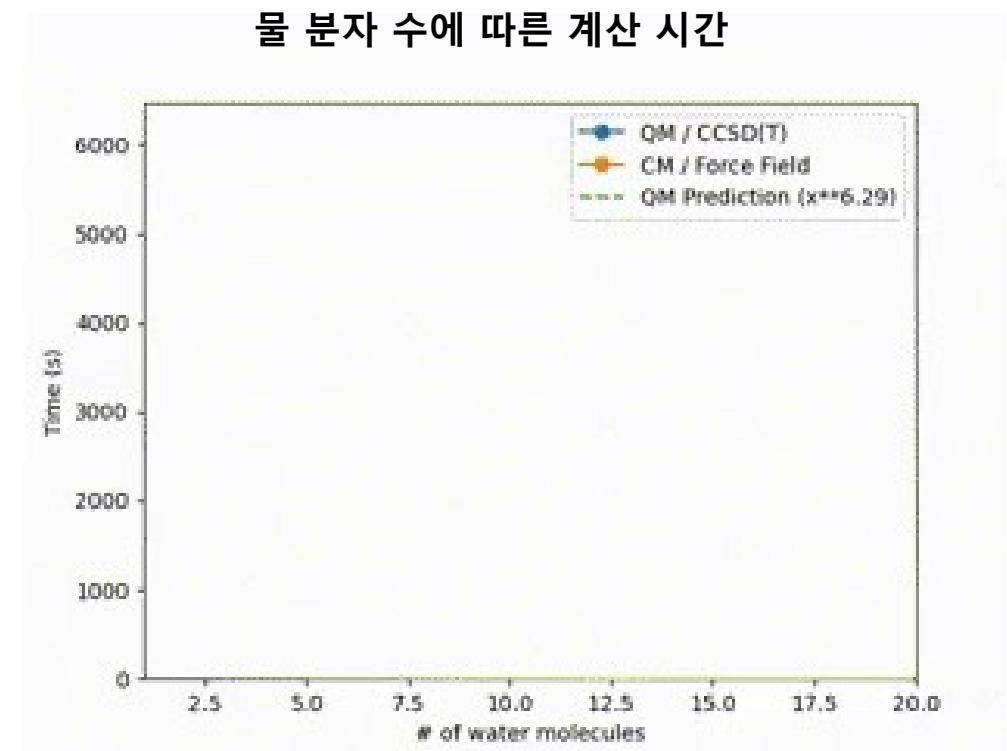
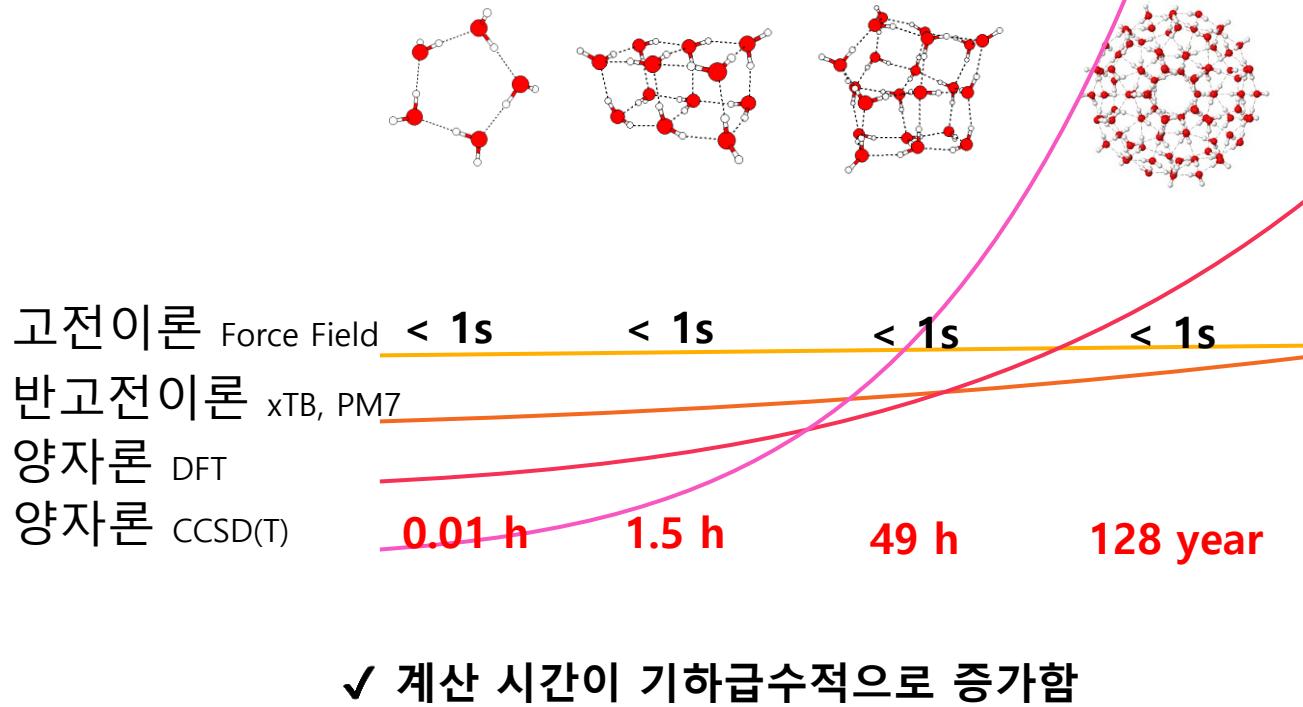
고전이론 Force Field	< 1s	< 1s	< 1s	< 1s
반고전이론 xTB, PM7				
양자론 DFT				
양자론 CCSD(T)	0.01 h	1.5 h	49 h	128 year

✓ 계산 시간이 기하급수적으로 증가함



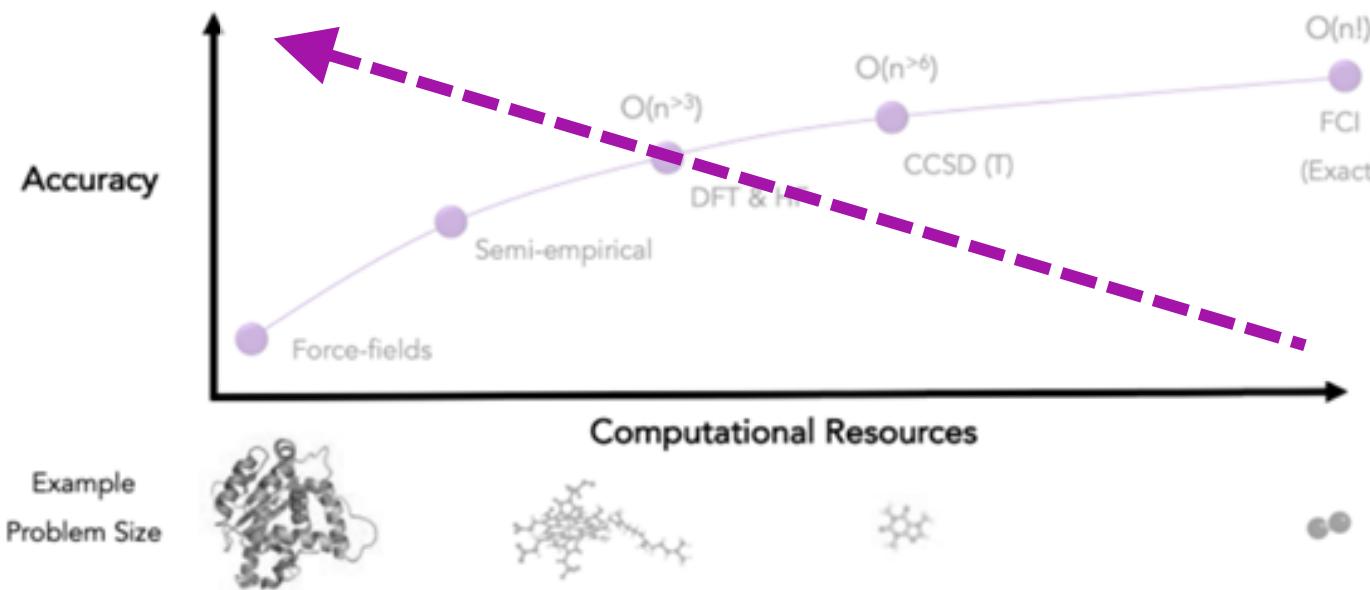
Background

✓ Trade-off btw. **Cost** vs. **Accuracy**



Background

✓ *Novel & Breakthrough Paradigm Shift*



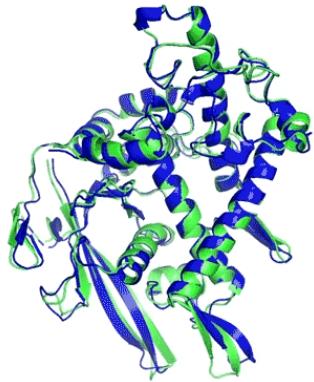
✓ 새로운 방법론이 필요



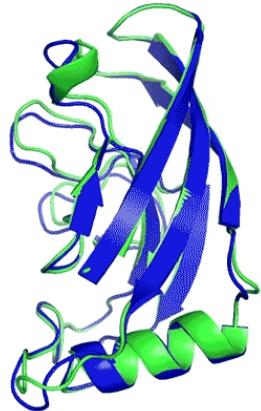
Accurate & Cost-effective

Background

AlphaFold



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction
2024 Nobel Prize in Chemistry



기상 예보



계산**과학**에서의 AI 성공 사례

Background

✓ 새로운 방법론의 등장



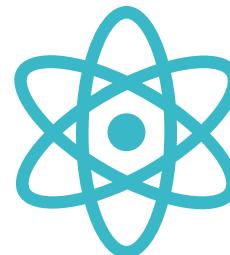
분자 에너지 계산에 AI(Data-Driven model)를 이용하자!

Background

✓ 새로운 방법 중 하나



ML-potential
Machine Learning + Potential Energy



Background

✓ 많은 기업들이 **ML-Potential** 모델과 데이터셋 생성에 투자



✓ 최근 3개월 주요 변화



10월 소재 SOTA 발표 (Orb-V2)



10월 소재 SOTA 발표 (eqV2)



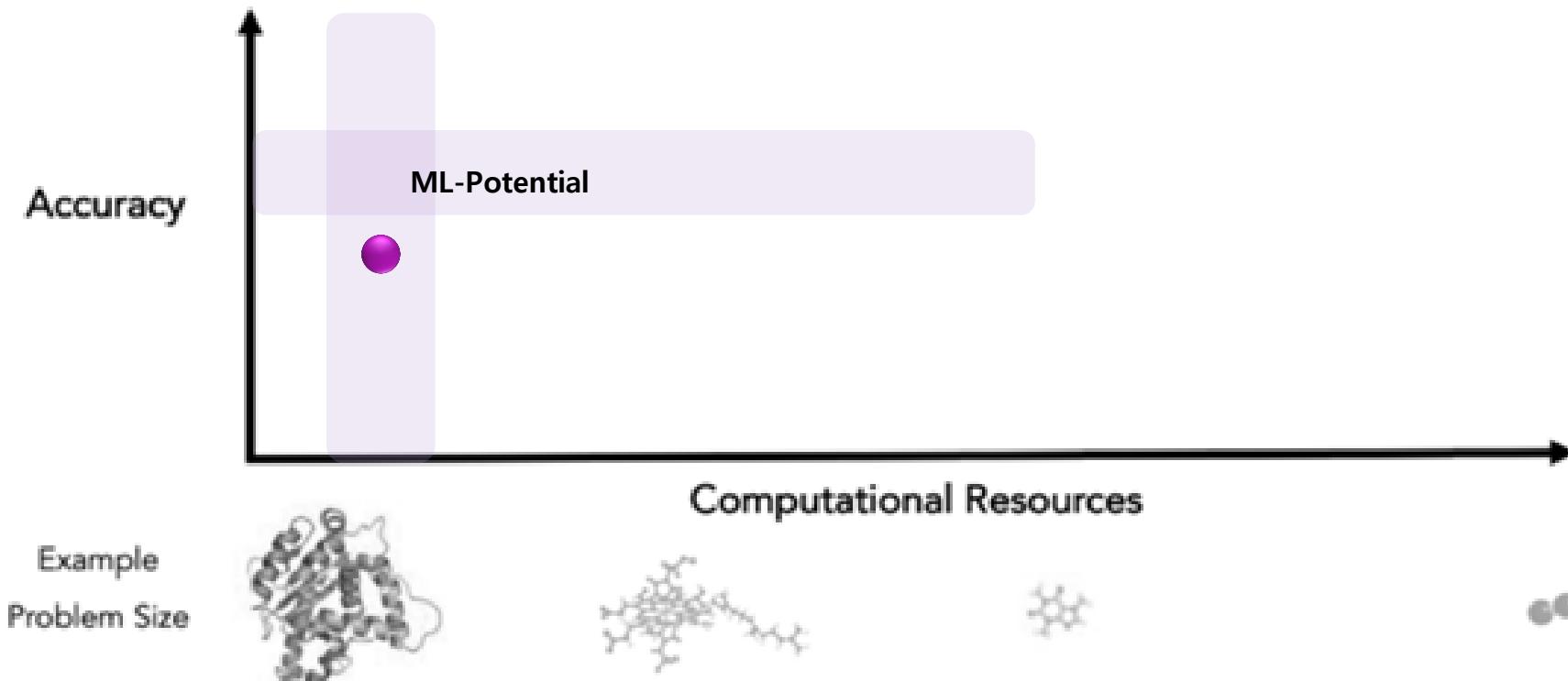
11월 단백질 SOTA 발표 (AI2BMD)



12월 소재 모델 발표 (MatterSim)

Background

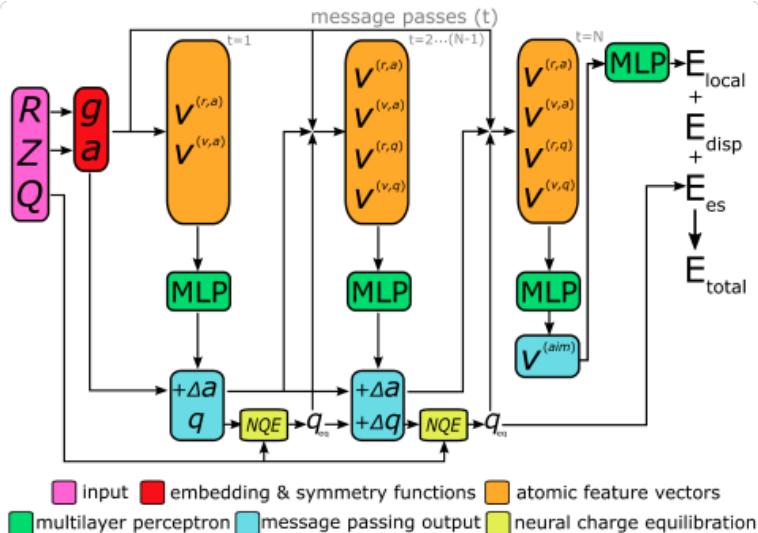
✓ ML-Potential



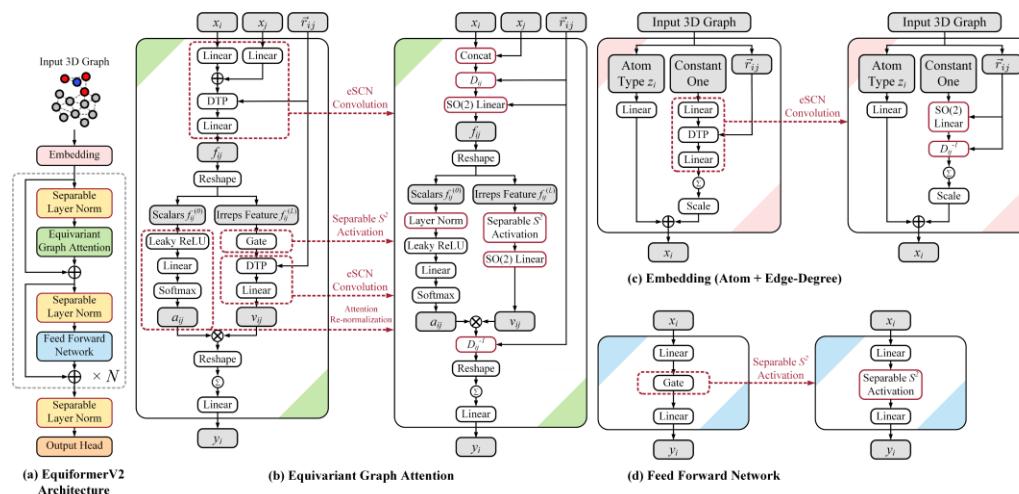
프로젝트 주제

Universal ^①**ML-potential** model의 전이학습과
ML-Potential을 활용한 ^②**응용 연구**

작동 원리

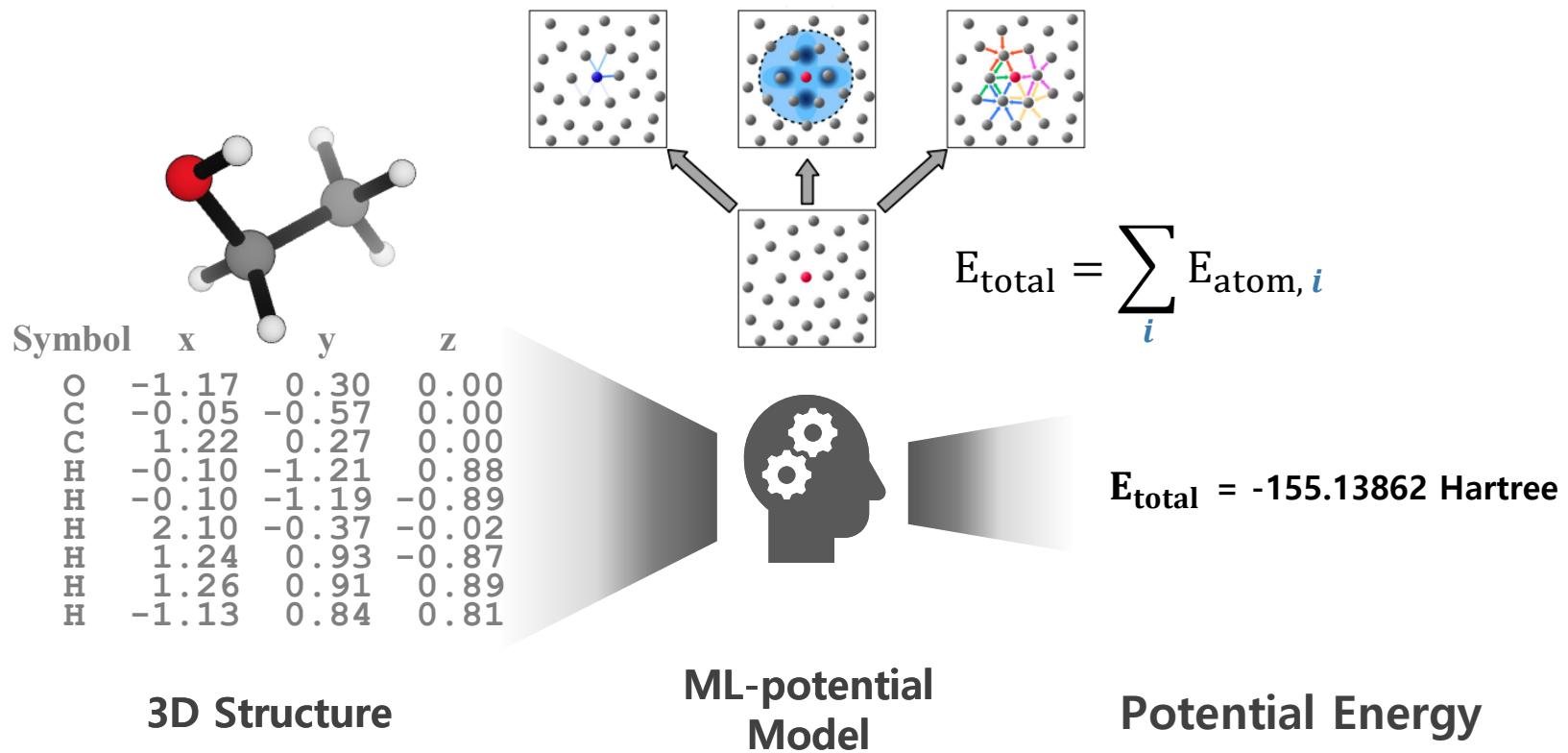


█ input █ embedding & symmetry functions █ atomic feature vectors
█ multilayer perceptron █ message passing output █ neural charge equilibration



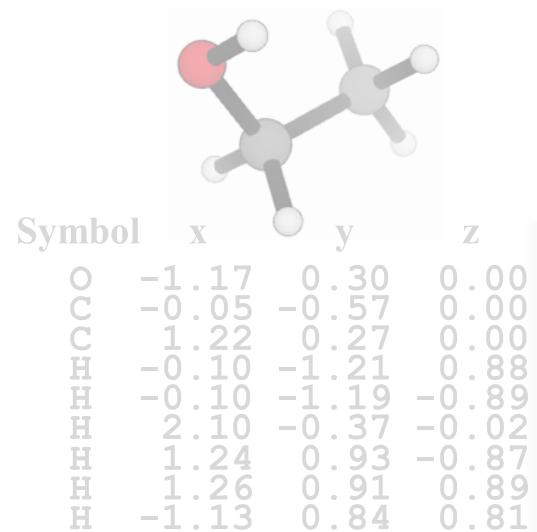
ML-Potential 작동 원리

✓ ML-Potential



ML-Potential 작동 원리

✓ ML-Potential



3D Structure



ML-potential
Model

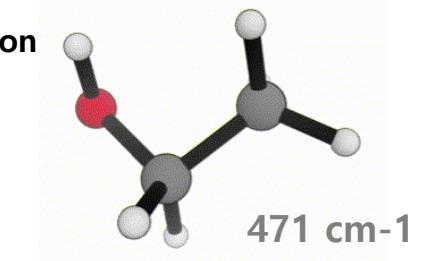
$$E_{\text{total}} = -155.13862 \text{ Hartree}$$

Potential Energy

$$H^f = \begin{pmatrix} \frac{\partial E}{\partial O^2} & \cdots & \frac{\partial E}{\partial O \partial H} \\ \vdots & \ddots & \vdots \\ \frac{\partial E}{\partial H \partial O} & \cdots & \frac{\partial E}{\partial O^2} \end{pmatrix}$$

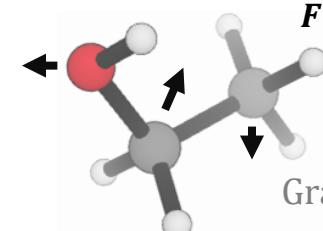
$$H^f X = \varepsilon X$$

Eigenvalue Equation



2nd Derivative

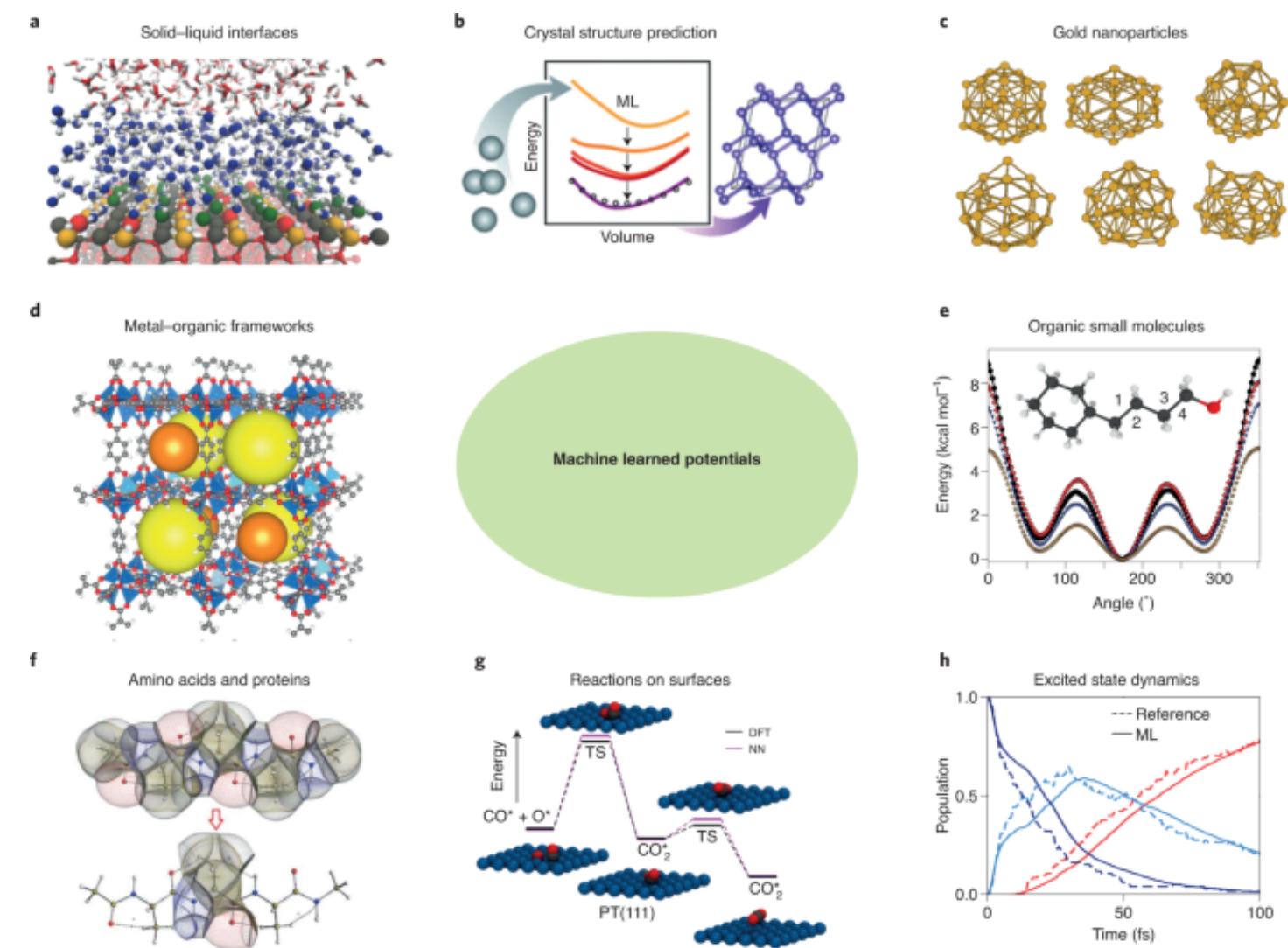
1st Derivative



$$\text{Grad}_E = \begin{pmatrix} \frac{\partial E}{\partial O} \\ \vdots \\ \frac{\partial E}{\partial H} \end{pmatrix}$$

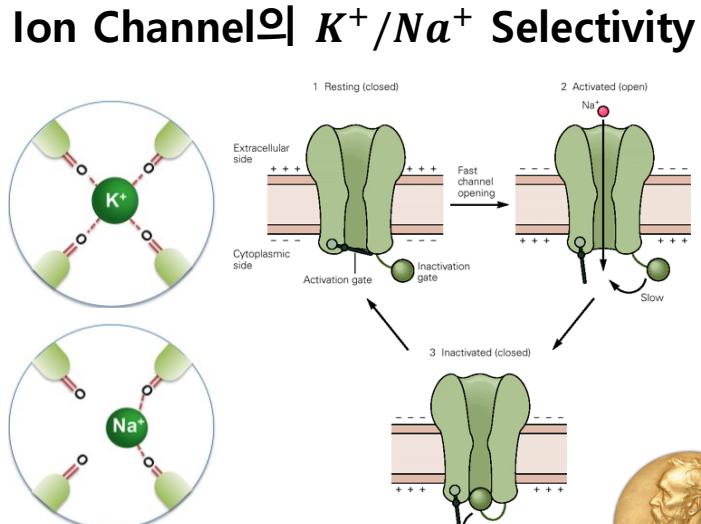
Forces

00용 연구 분야

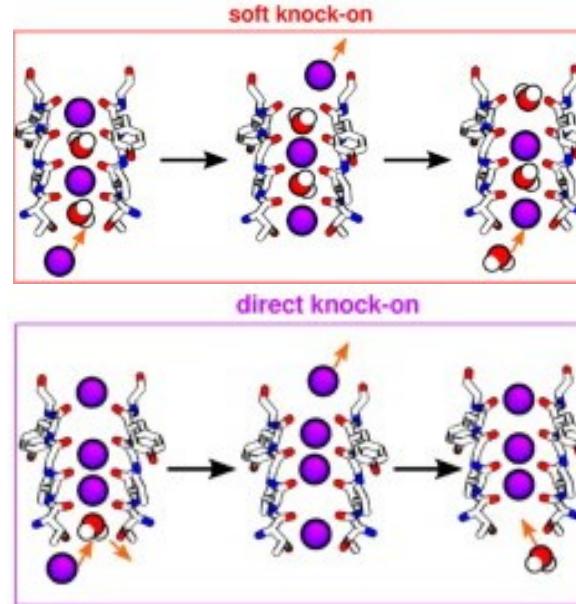
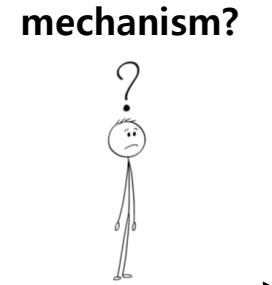


Biomolecular Science

✓ ML-Potential in Biomolecular System



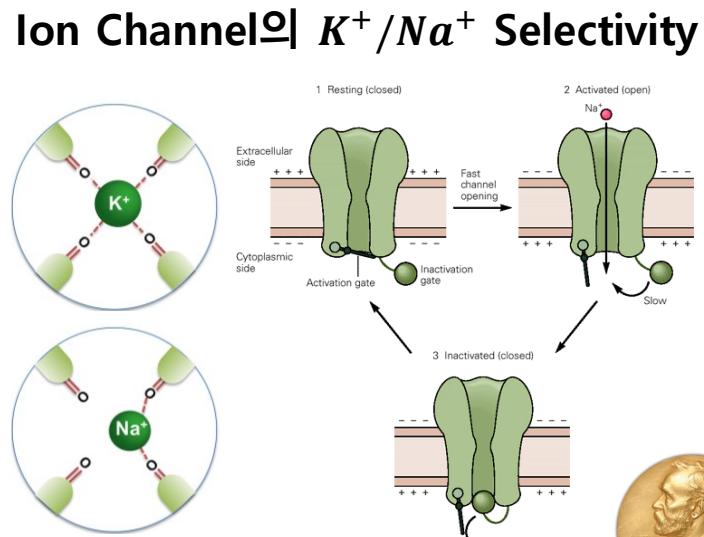
2003 Nobel Prize in Chemistry



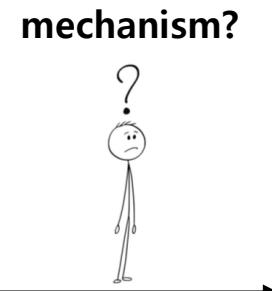
Mechanism은 여전히 논쟁 대상

Biomolecular Science

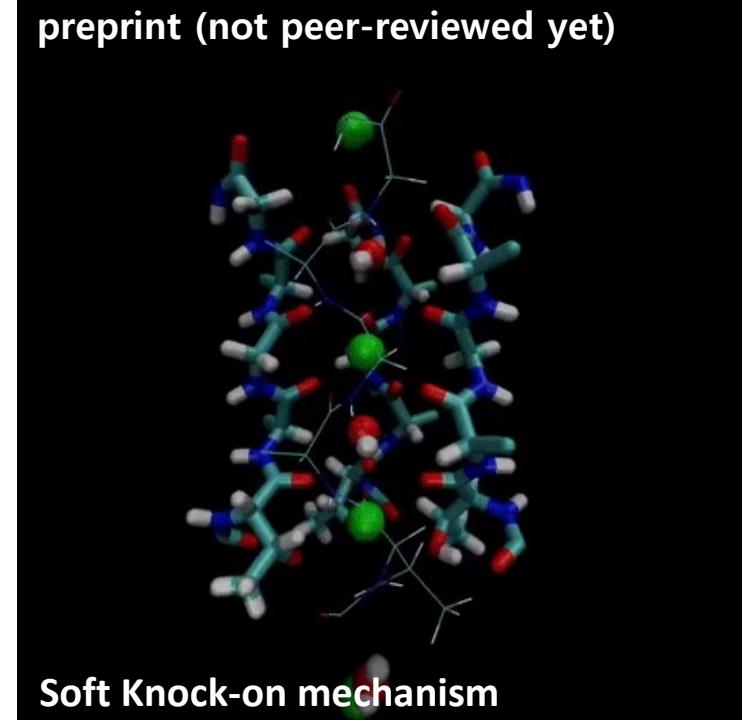
✓ ML-Potential in Biomolecular System



2003 Nobel Prize in Chemistry



preprint (not peer-reviewed yet)



Soft Knock-on mechanism

Ion Channel Transport Mechanism 연구
ML-Potential을 이용한 시뮬레이션 *

Material Science

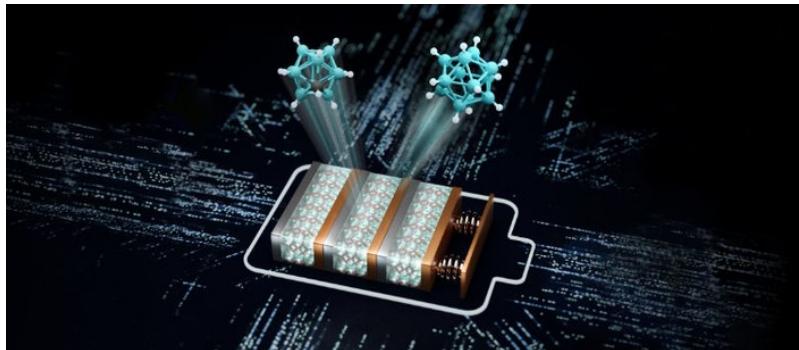
✓ ML-Potential in Material Science

NEWS RELEASE 10-FEB-2022

A new electrolyte for greener and safer batteries

A UNIGE team has developed a new material that improves the performance of solid-state sodium batteries, a less dangerous and more durable alternative to lithium.

Peer-Reviewed Publication
UNIVERSITÉ DE GENÈVE



Next Generation Battery

dynamic properties?



Static



Material Science

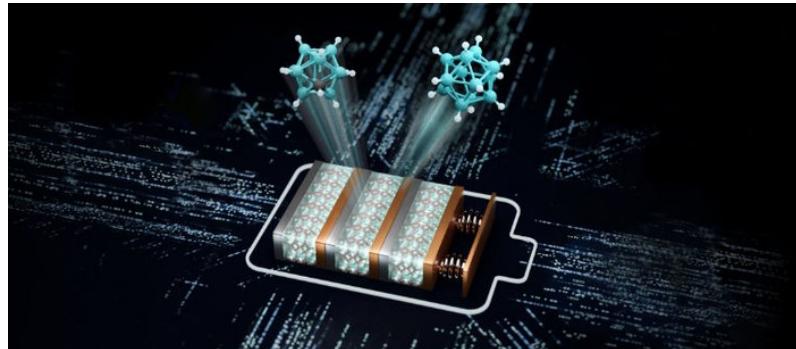
✓ ML-Potential in Material Science

NEWS RELEASE 10-FEB-2022

A new electrolyte for greener and safer batteries

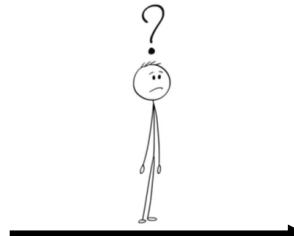
A UNIGE team has developed a new material that improves the performance of solid-state sodium batteries, a less dangerous and more durable alternative to lithium.

Peer-Reviewed Publication
UNIVERSITÉ DE GENÈVE

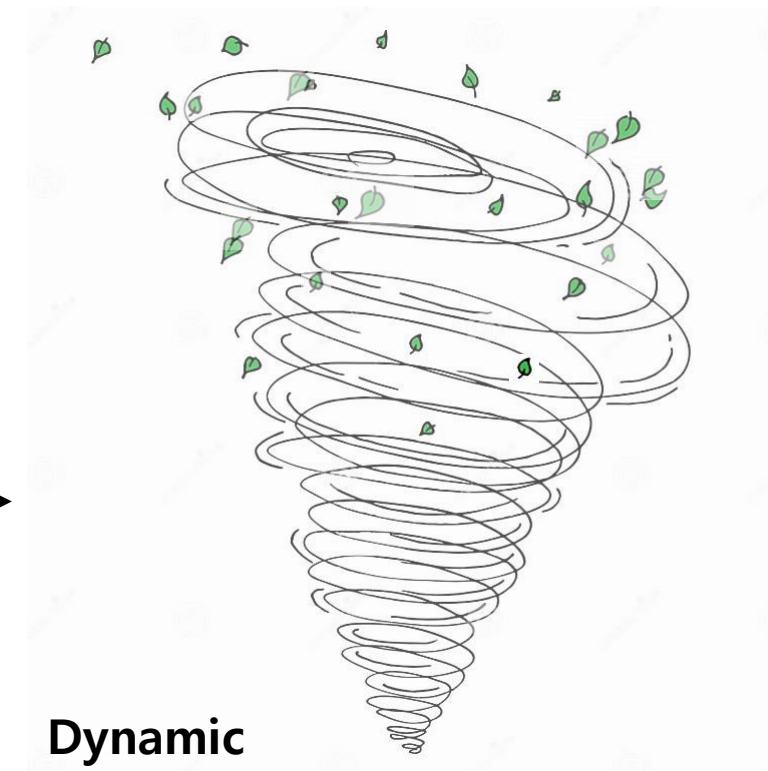


Next Generation Battery

dynamic properties?



Dynamic



Material Science

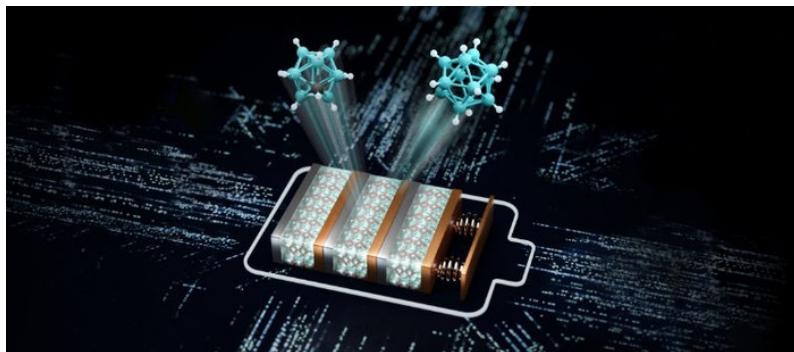
✓ ML-Potential in Material Science

NEWS RELEASE 10-FEB-2022

A new electrolyte for greener and safer batteries

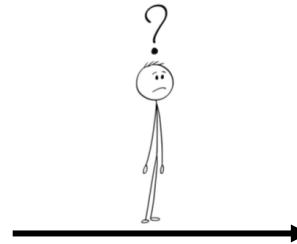
A UNIGE team has developed a new material that improves the performance of solid-state sodium batteries, a less dangerous and more durable alternative to lithium.

Peer-Reviewed Publication
UNIVERSITÉ DE GENÈVE



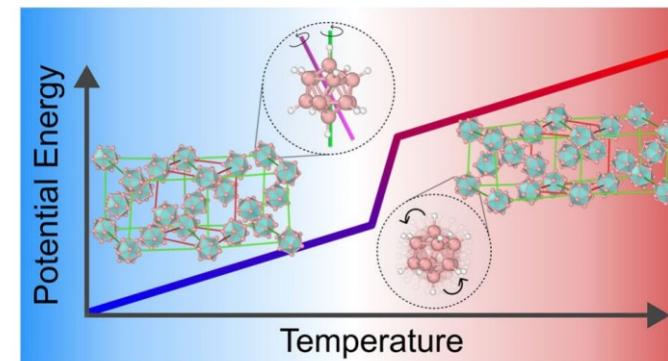
Next Generation Battery

dynamic properties?



Unraveling Superionic Conductivity in Na₂B₁₂H₁₂: Molecular Dynamics Study of Phase Transition, Anion Reorientation, and Ionic Conductivity via Machine Learning Force Field

By Rui Zhou, Kun Luo, Ling Fei and Qi An



ACS
Electrochemistry

ACS Publications
Most Trusted. Most Cited. Most Read.

고체의 동역학적 성질을
ML-Potential로 시뮬레이션

Material Science

✓ ML-Potential in Material Science

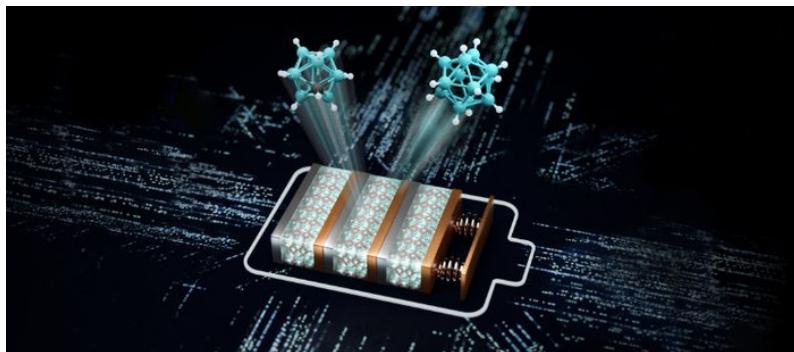
NEWS RELEASE 10-FEB-2022

A new electrolyte for greener and safer batteries

A UNIGE team has developed a new material that improves the performance of solid-state sodium batteries, a less dangerous and more durable alternative to lithium.

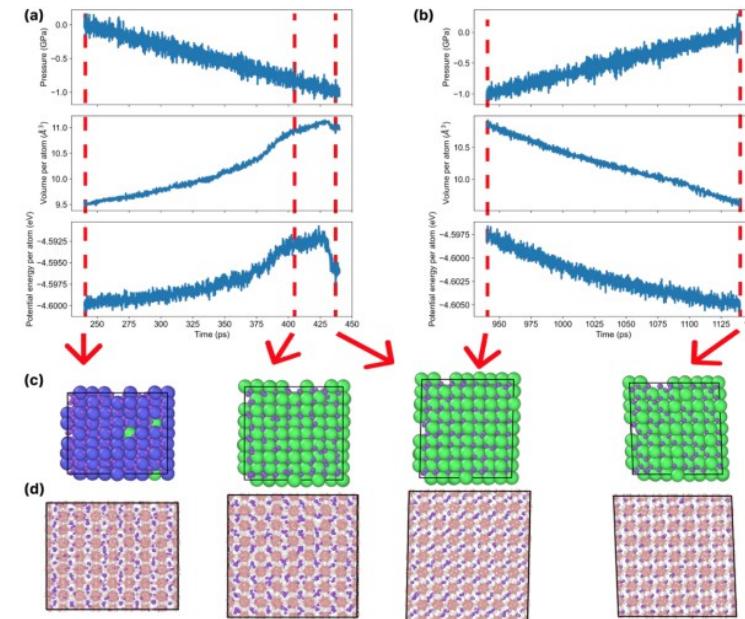
Peer-Reviewed Publication

UNIVERSITÉ DE GENÈVE



Next Generation Battery

dynamic properties?



고체의 동역학적 물성을
ML-Potential로 시뮬레이션

Spectroscopy

✓ ML-Potential in Spectroscopy



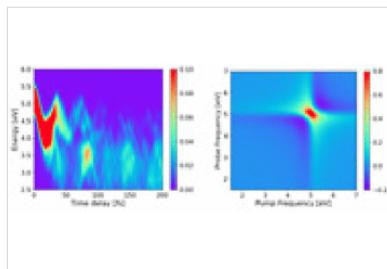
PHYSICAL INSIGHTS INTO LIGHT INTERACTING WITH MATTER | February 22, 2024

Artificial-Intelligence-Enhanced On-the-Fly Simulation of Nonlinear Time-Resolved Spectra

Sebastian V. Pios, Maxim F. Gelin, Arif Ullah, Pavlo O. Dral*, and Lipeng Chen*

Abstract

Time-resolved spectroscopy is an important tool for unraveling the minute details of structural changes in molecules of biological and technological significance. The nonlinear femtosecond signals detected for such systems must be interpreted, but it is a challenging task for which theoretical simulations are often indispensable. Accurate simulations of transient absorption or two-dimensional electronic spectra are, however, computationally very expensive, prohibiting the wider adoption of existing first-principles methods. Here, we report an artificial-intelligence-enhanced protocol to drastically reduce the computational cost of simulating nonlinear time-resolved electronic spectra, which makes such simulations affordable for polyatomic molecules of increasing size. The protocol is based on the doorway–window approach for the on-the-fly surface-hopping simulations. We show its applicability for the prototypical molecule of pyrazine for which it produces spectra with high precision with respect to *ab initio* reference while cutting the computational cost by at least 95% compared to pure first-principles simulations.

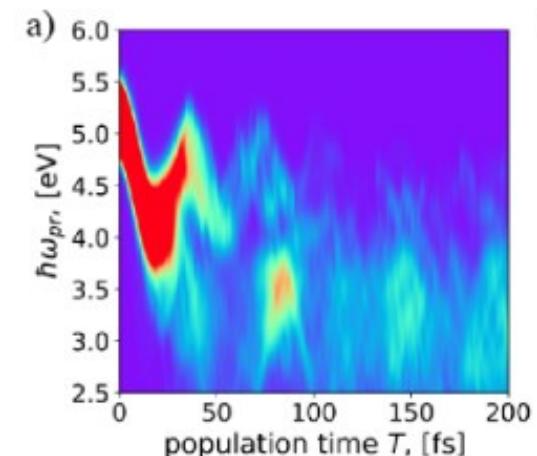


Copyright © 2024 American Chemical Society

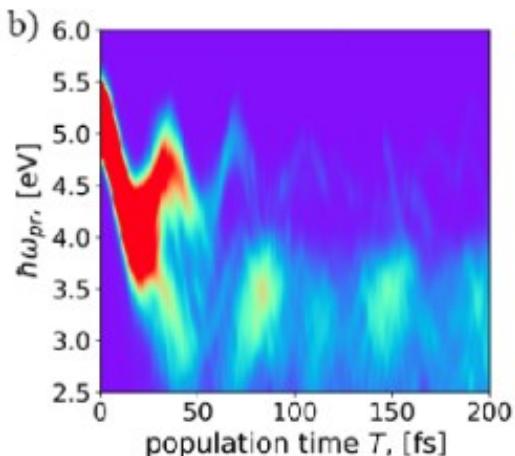
Subjects ⓘ

Ab Initio Calculations Computational Chemistry Pyrazine Quantum Mechanics Spectroscopy

ML-Accelerated



Ref. (pure ab initio)



Pure ab initio 대비 최대 95% 계산 비용 단축

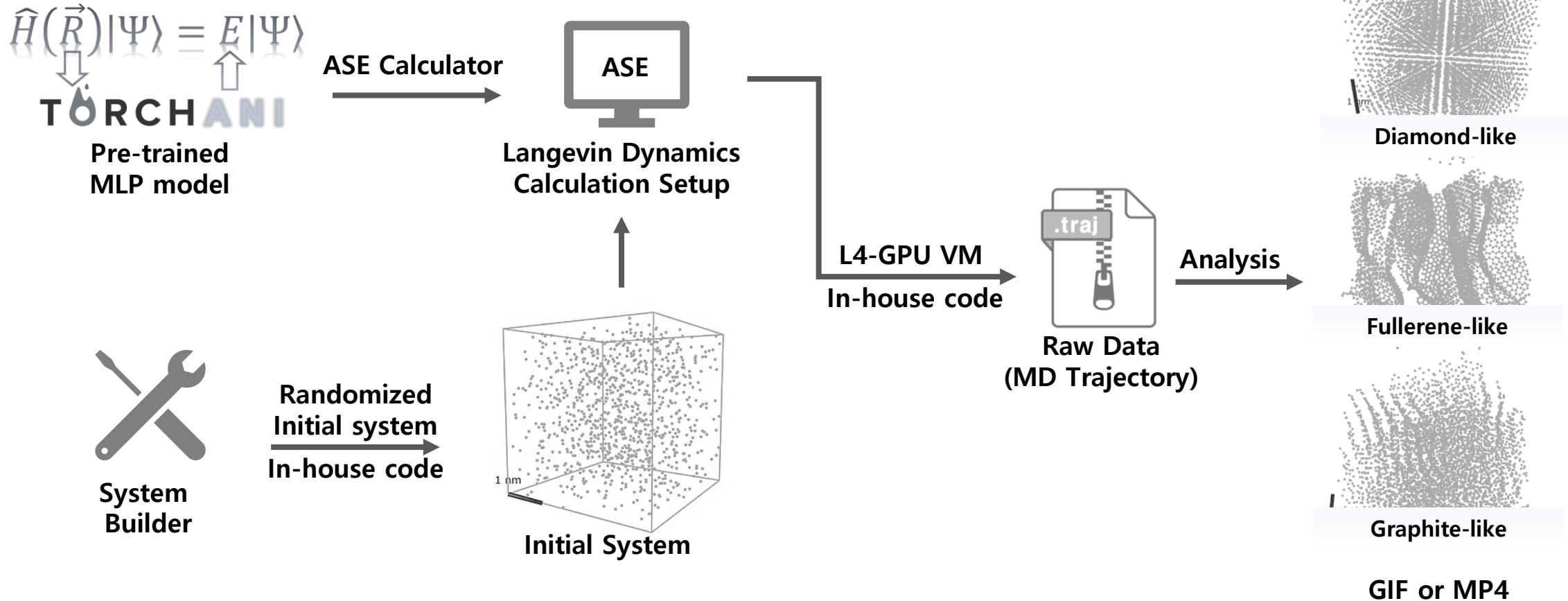
프로젝트 주제

ML-Potential 응용 연구

ML-Potential 모델 연구

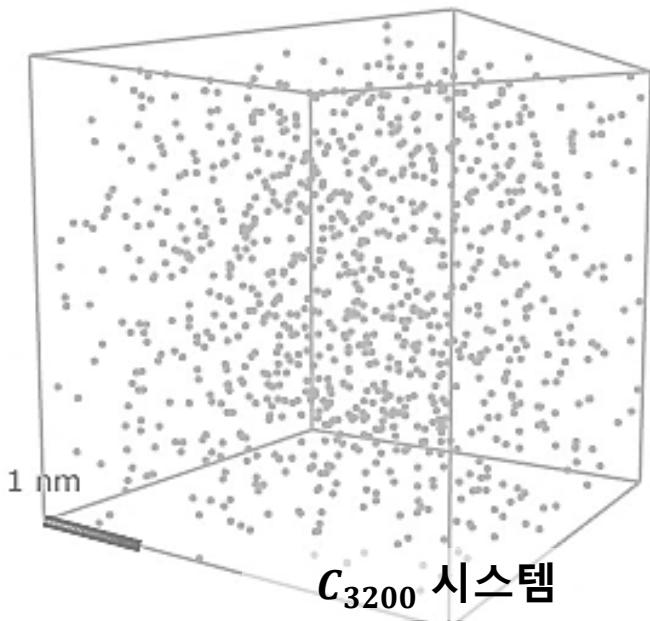
ML-Potential 응용 연구 (1)

✓ ML-Potential 기반 Molecular Dynamics (MD) 시뮬레이션



ML-Potential 응용 연구 (1)

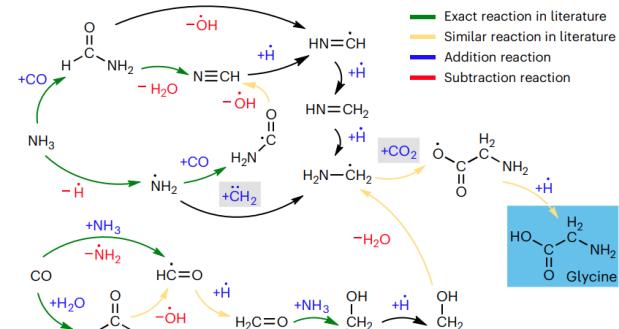
✓ NanoReactor (컴퓨터 속 작은 화학반응기)



가상의 화학반응기

나노 사이즈의 박스에 반응물을 넣고 반응을 시켜보자!

컴퓨터 속에서 온도와 압력 등을 조절하며 화학 반응 구현



nature chemistry

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature chemistry > articles > article

Article | Open access | Published: 07 March 2024

Exploring the frontiers of condensed-phase chemistry with a general reactive machine learning potential

Shuhao Zhang, Małgorzata Z. Makoś, Ryan B. Jadrich, Elfi Kraka, Kipton Barros, Benjamin T. Nebgen, Sergei Tretiak, Alexandr Isayev, Nicholas Lubbers, Richard A. Messerly & Justin S. Smith

Nature Chemistry 16, 727–734 (2024) | Cite this article

17k Accesses | 21 Citations | 155 Altmetric | Metrics

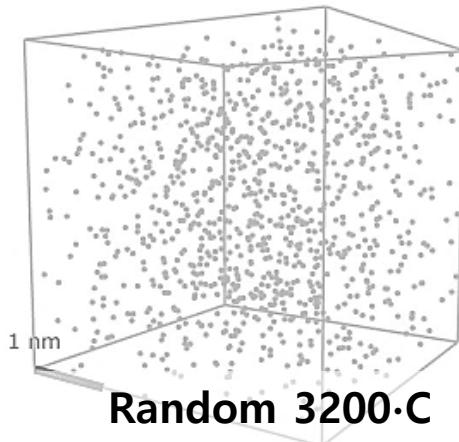
ML-Potential 응용 연구 (1)

✓ 탄소 고분자 합성 시뮬레이션

```
● ● ●  
from nanoreactor import NanoReactor  
  
params = {  
    'concentration': 0.5, # g/cc  
    'system_filepath': 'box_0.5[gcc-1].points_3200[atoms].xyz',  
    'temperature_K': 2500,  
    'time_step_fs': 0.5,  
    'total_steps': 3 * 10**6,  
    'modeltype': 'ANI1xnr',  
    'friction': 0.01,  
    'logfile': 'auto', # '-' for stdout  
    'loginterval': 100,  
    'trajfile': 'auto',  
    'trajinterval': 100,  
    'logger': 'default',  
    'optimize_geometry':False,  
}  
  
# configure system params  
simulation_box = NanoReactor(**params)  
  
# run simulation  
simulation_box.run_simulation()
```

nanoreactor.py (module)

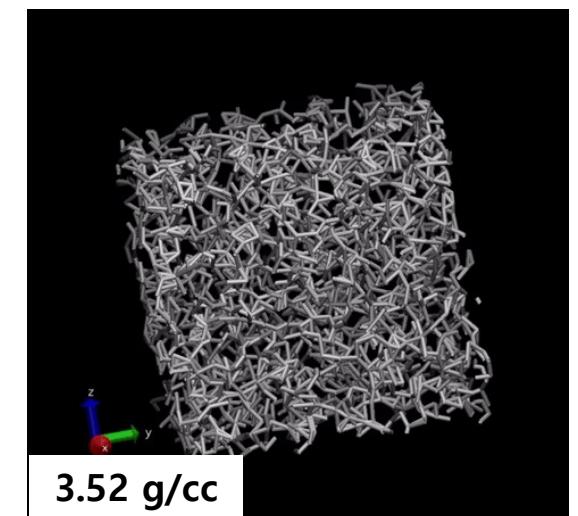
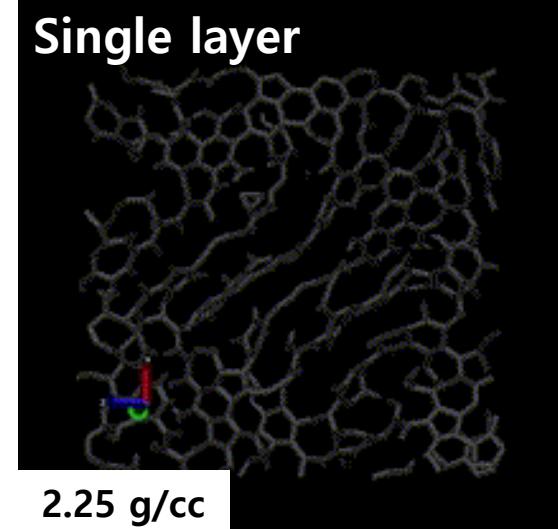
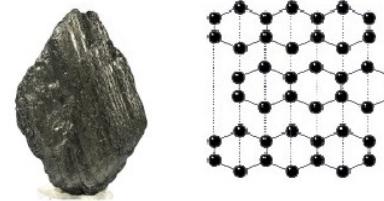
✓ 논문 구현



초기 농도 의존
~2 ns

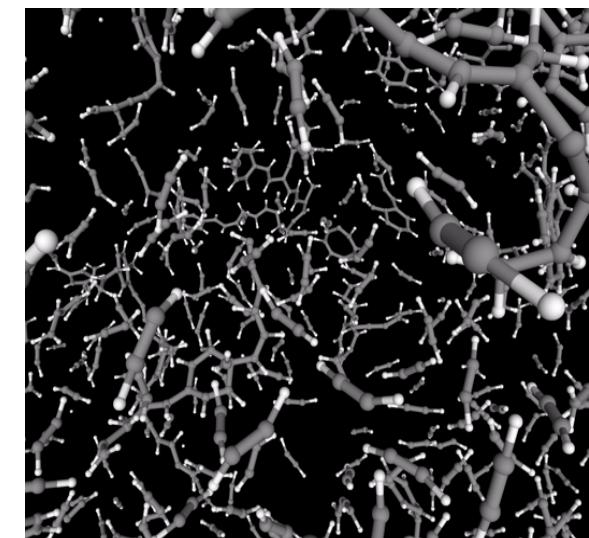
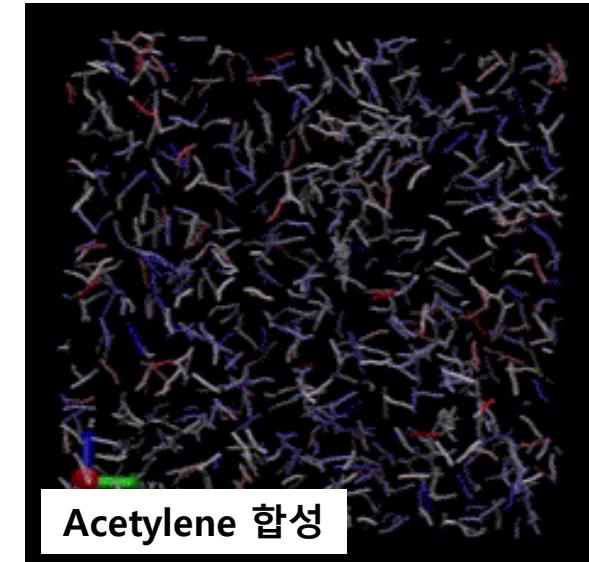
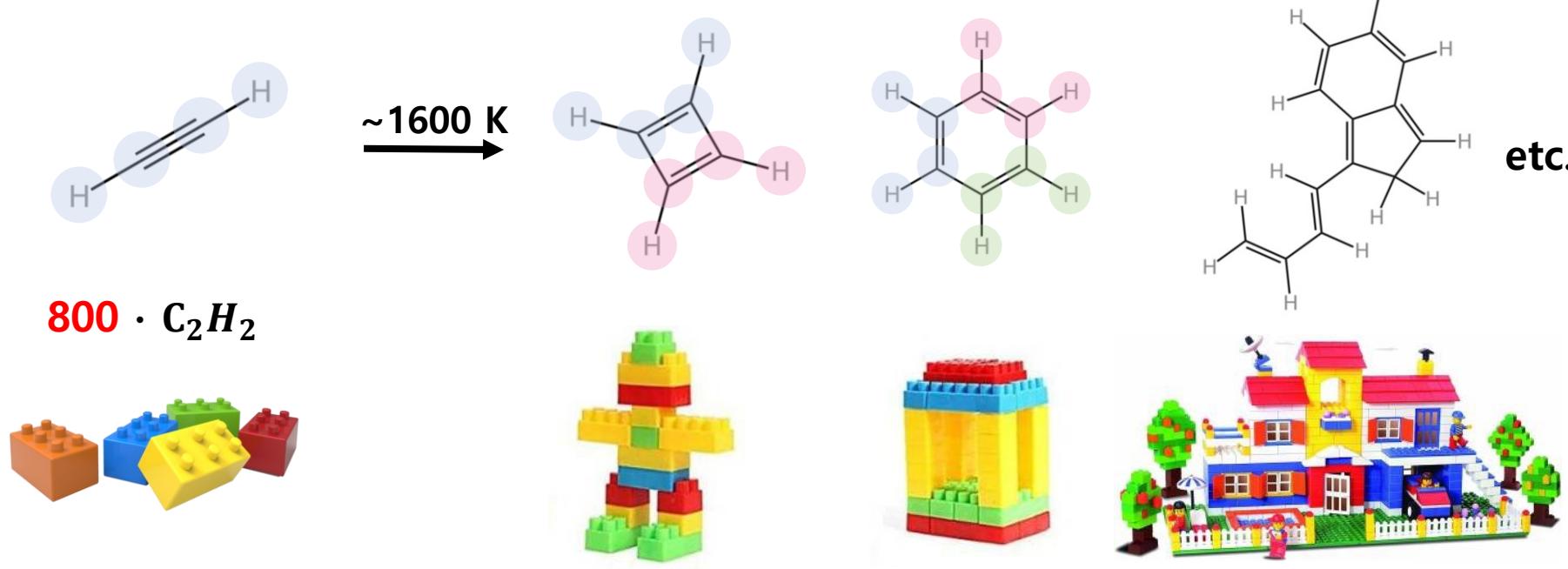


~2 ns simulation (4×10^6 steps, 0.5 fs/step)



ML-Potential 응용 연구 (1)

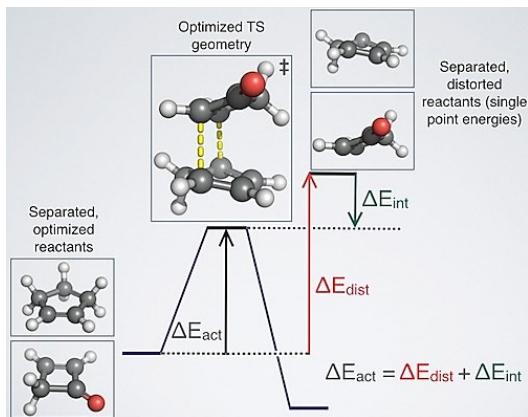
✓ Acetylene Oligomerization



$\sim 1 \text{ ns simulation } (2 \times 10^6 \text{ steps}, 0.5 \text{ fs/step})$

ML-Potential 응용 연구 (2)

✓ Single Point Distortion Interaction Analysis (DIAS)



$$\Delta E_{act} = \Delta E_{dist} + \Delta E_{int}$$

$$\Delta E_{dist, frag i} = E_{preopt, frag i} - E_{opt, frag i}$$

$$\Delta E_{dist} = \sum_i \Delta E_{dist, frag i}$$

✓ Key Features

- ✓ 쉬운 사용법, Open Source
 - * 기존 프로그램 : autoDIAS, pyfrag, D2AF(Oct 2024)
- ✓ 쉬운 설치, (pip 배포)
- ✓ ML-Potential 기반 pre-optimizer도입으로 ab initio 계산도 >2배 빠르게 수행
- ✓ ASE라는 통합 interface를 지원하여 ML-Potential을 포함한 수많은 양자화학 프로그램과 호환

Official calculators



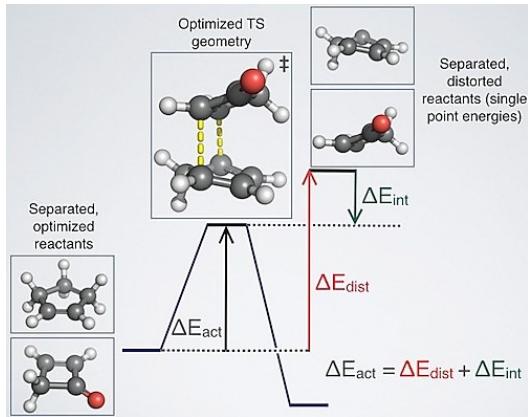
Unofficial calculators

TorchANI, XPaiNN, MACE, AIMNet2, PySCF, TBLite, AI2BMD etc.



ML-Potential 응용 연구 (2)

✓ Single Point Distortion Interaction Analysis (DIAS)



$$\Delta E_{act} = \Delta E_{dist} + \Delta E_{int}$$

$$\Delta E_{dist, frag i} = E_{preopt, frag i} - E_{opt, frag i}$$

$$\Delta E_{dist} = \sum_i \Delta E_{dist, frag i}$$

Simple Python API

```
from ase_dias import dias_run, AIMNet2Calculator, load_aimnet2

# fragments_params
fp_n3 = {
    "Br-": (-1, [2]),
    "CH3+": (+1, [1,3,4,5]),
    "Cl-": (-1, [6])
}

aimnet2_calc = lambda **kwargs : AIMNet2Calculator(model=load_aimnet2(), **kwargs)

dias_run(
    calc=aimnet2_calc,
    trajFile="sn2.xyz",
    fragments_params=fp_n3,
)
```

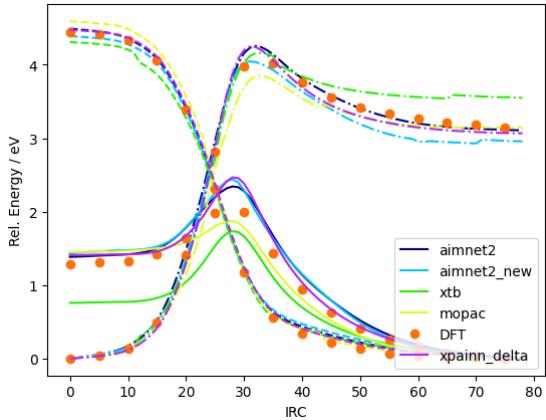
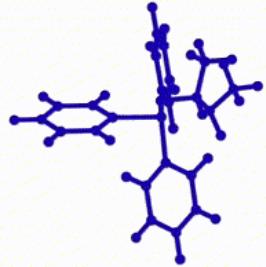
Processing . . .
89.47% |-----| 17 / 19
Optimizer : optimizing Br-



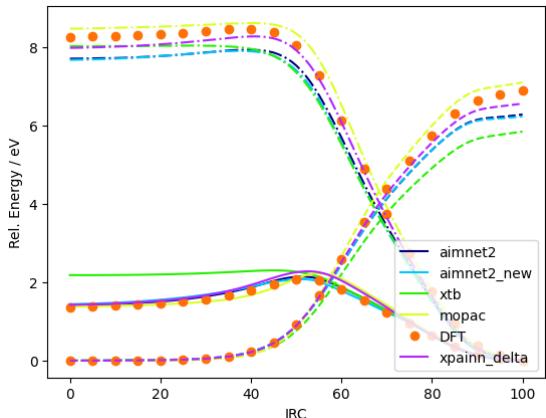
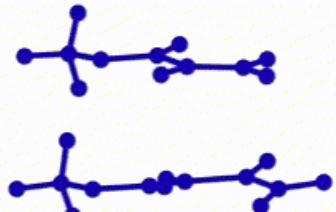
ML-Potential 응용 연구 (2)

✓ Distortion Interaction Analysis

Wittig Reaction



Diels Alder Reaction



MOPAC, xTB
AIMNet2, XPAiNN
DFT (wb97x-d3)

: semi-empirical method
: ML-Potential
: Reference level

✓ Cost

ab initio(DFT) : $\sim 10^2$ hr (\$3000 L4 GPU)

semi-empirical : $\sim 10^3$ sec (\$300 Laptop)

ML-Potential : $\sim 10^2$ sec (\$300 Laptop)

✓ Accuracy

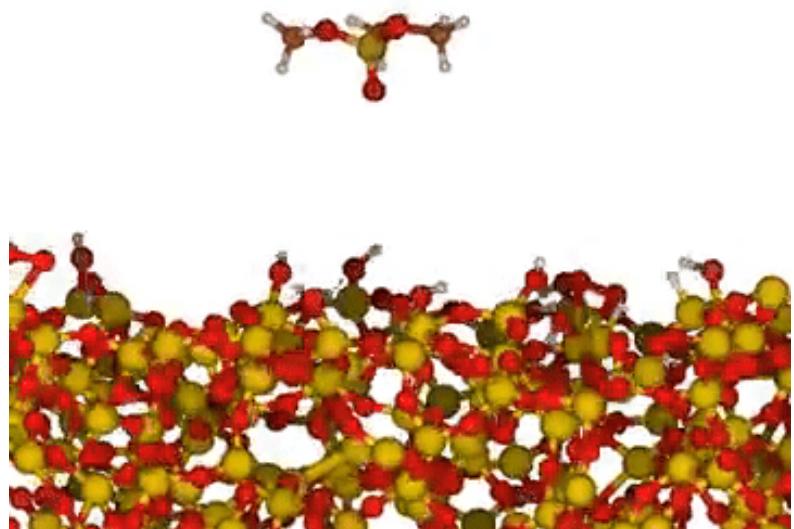
ab initio \approx ML-Potential

~ hr

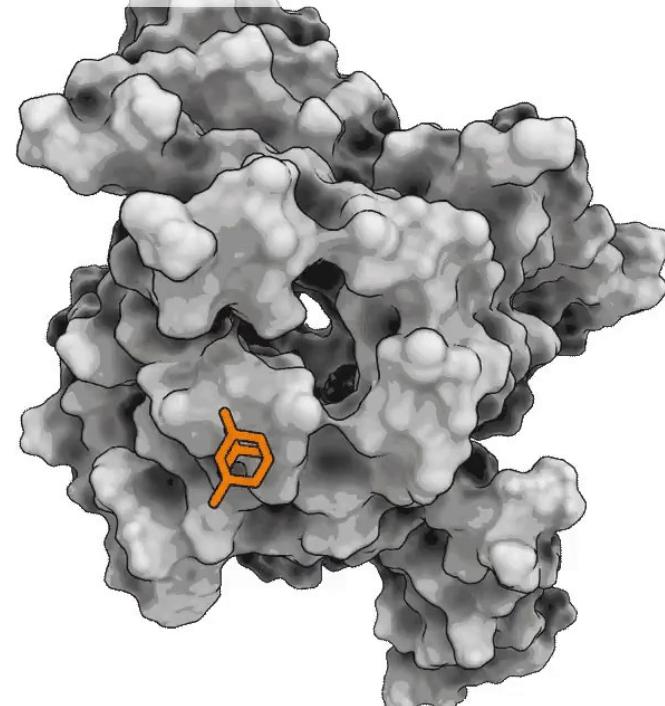
~ sec

Future Work

Surface Dynamics



Protein Ligand Dynamics



거대 시스템에서 일어나는 반응 메커니즘을 이해

프로젝트 주제

ML-Potential 응용 연구

ML-Potential 모델 연구

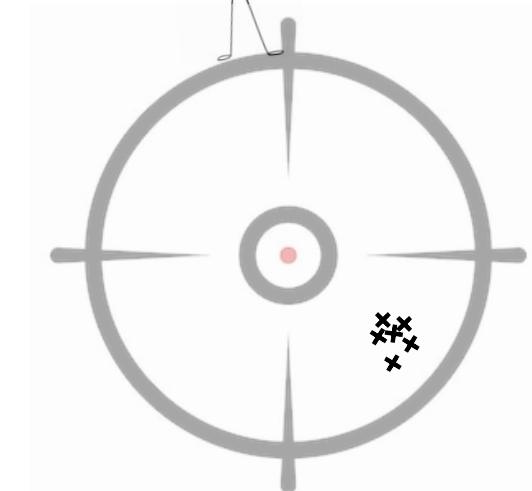
기준 모델의 한계

Seen it!



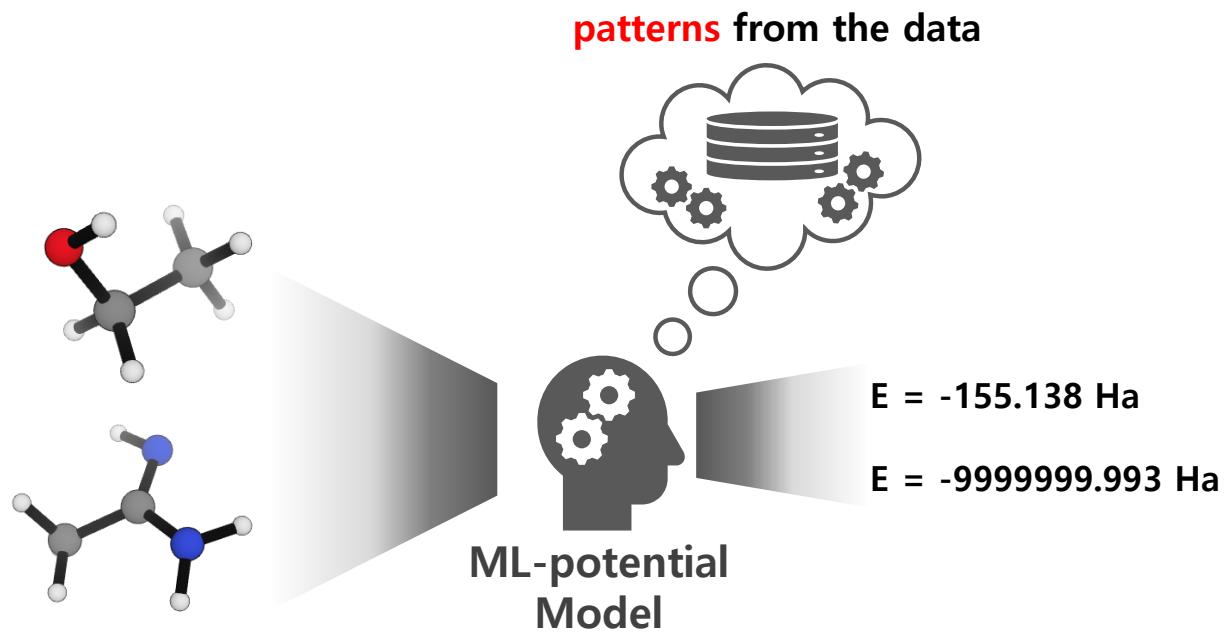
High Accuracy
High Precision

Never seen it...

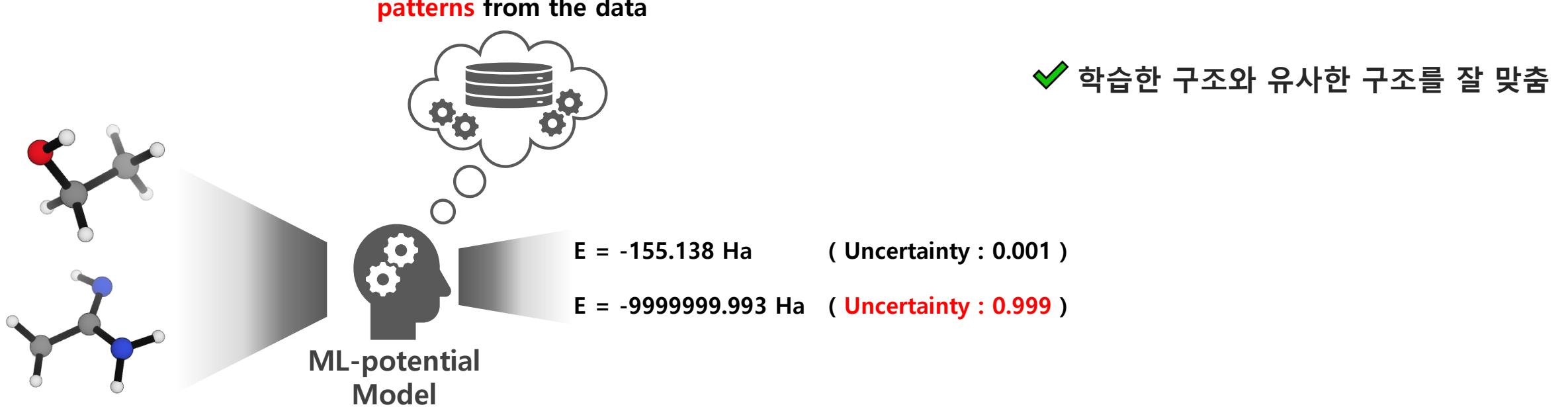


Low Accuracy
High Precision

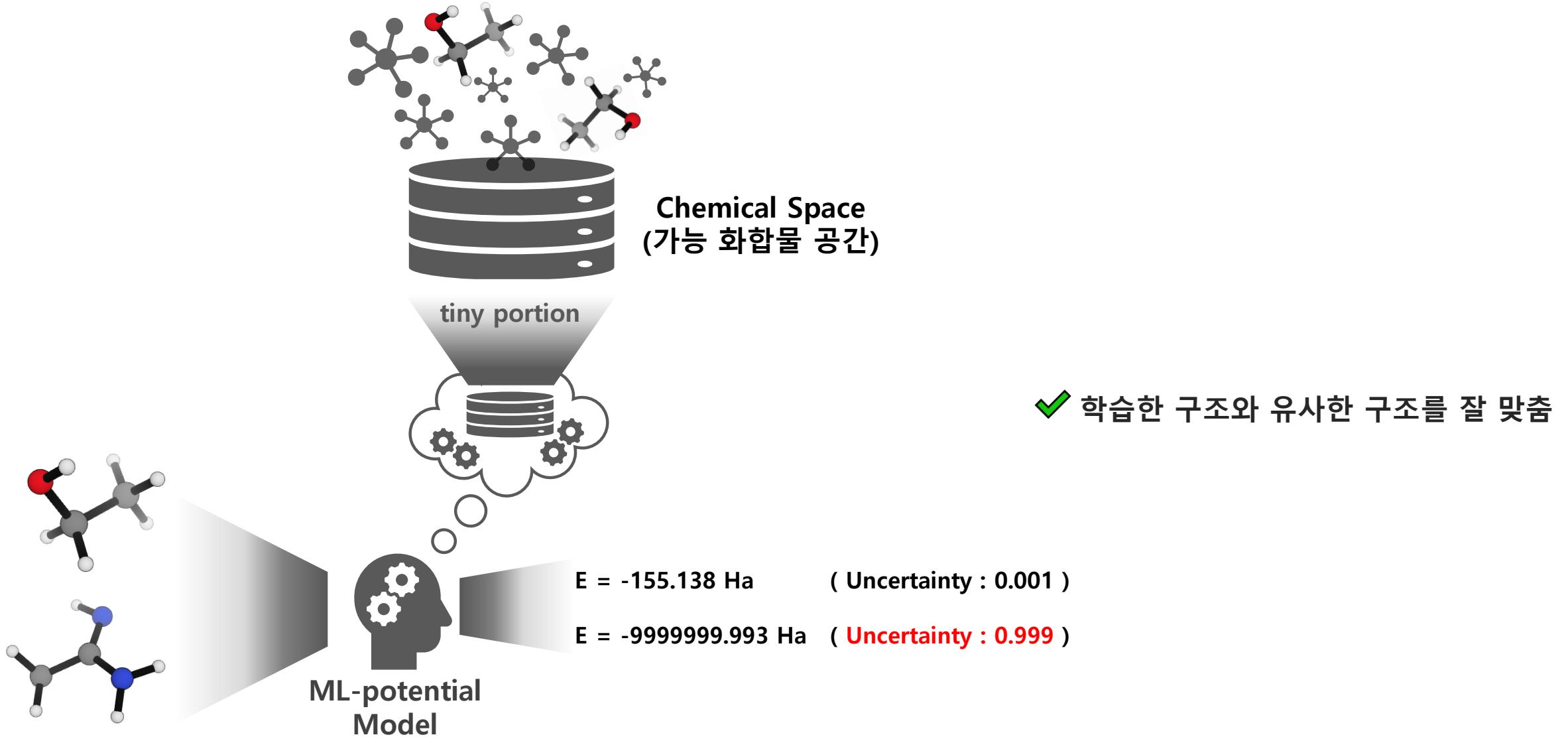
기존 모델의 한계



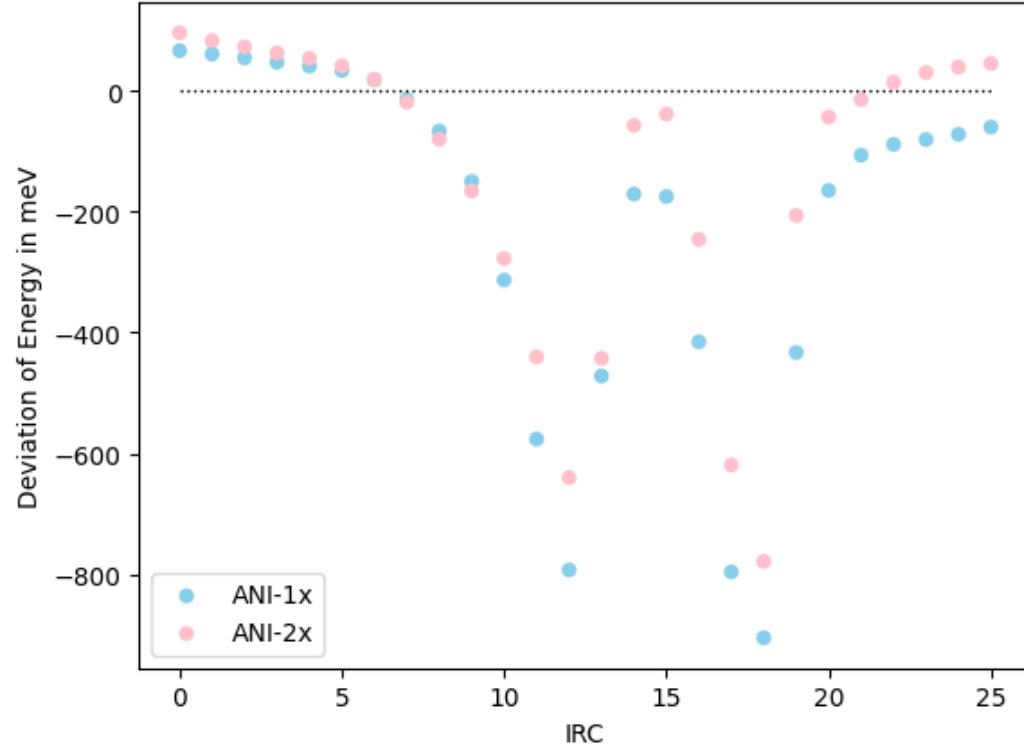
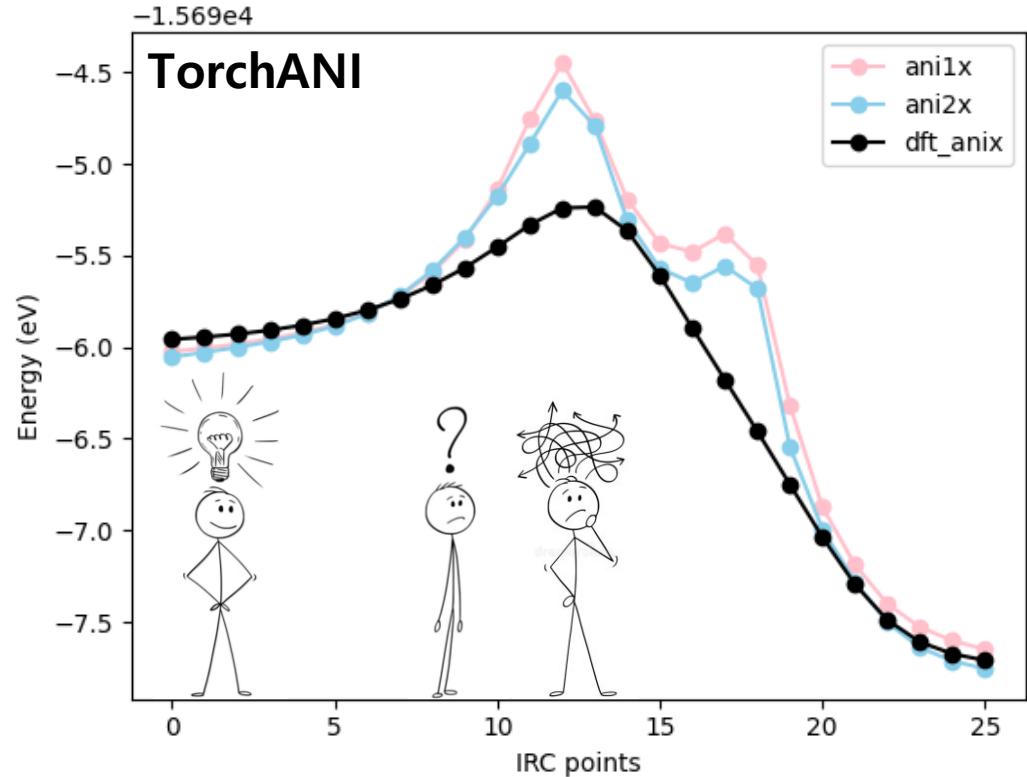
기존 모델의 한계



기존 모델의 한계

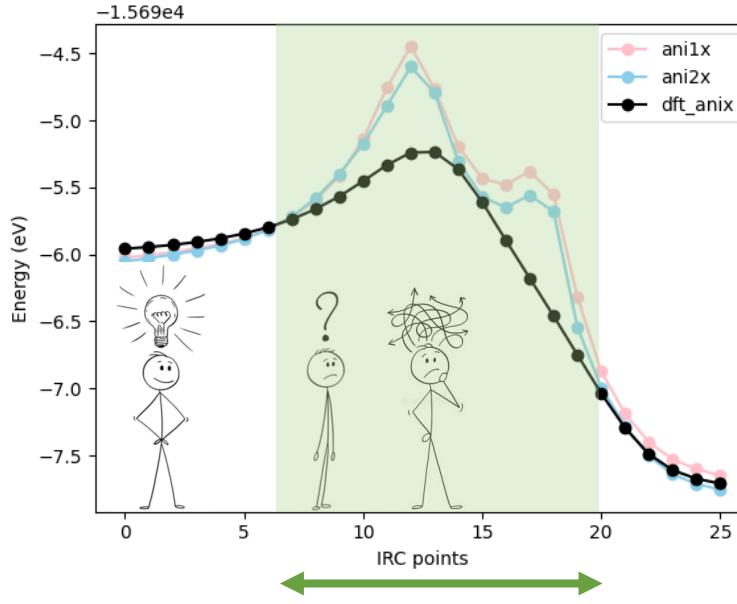


기존 모델의 한계



전이 구조 근처의 훈련 데이터 부재

ML-Potential 모델 연구

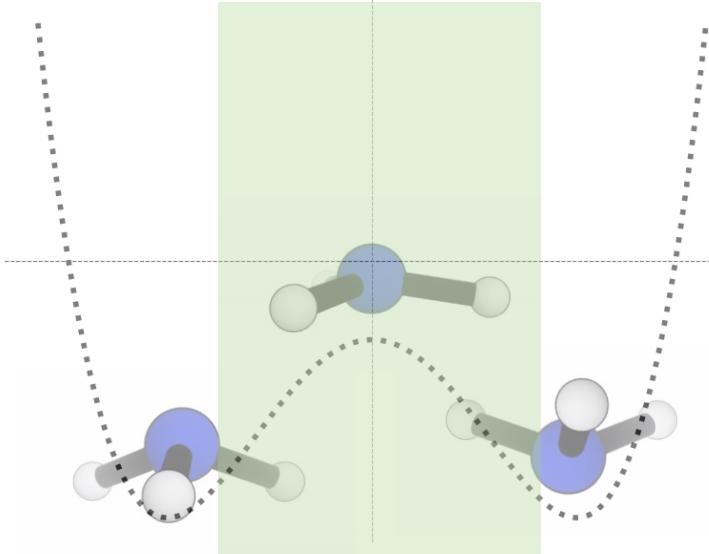


전이 상태 근처 영역



Mission : 전이 상태와 유사한 구조 데이터셋을 만들어보자
(기존 ML-Potential 모델까지 개선하면 더 좋고)

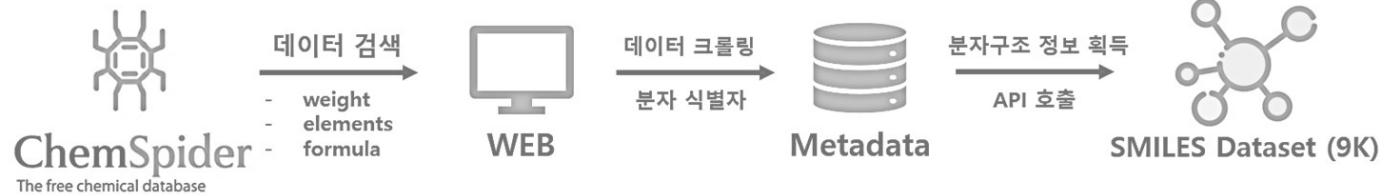
ML-Potential 모델 연구



Idea : 분자는 진동하고 반응은 진동과 관련된다

ML-Potential 모델 연구

<분자 메타 데이터 수집>

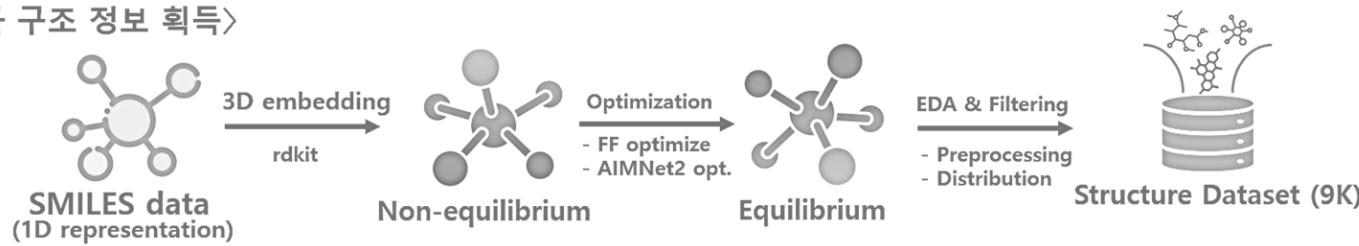


ML-Potential 모델 연구

<분자 메타 데이터 수집>

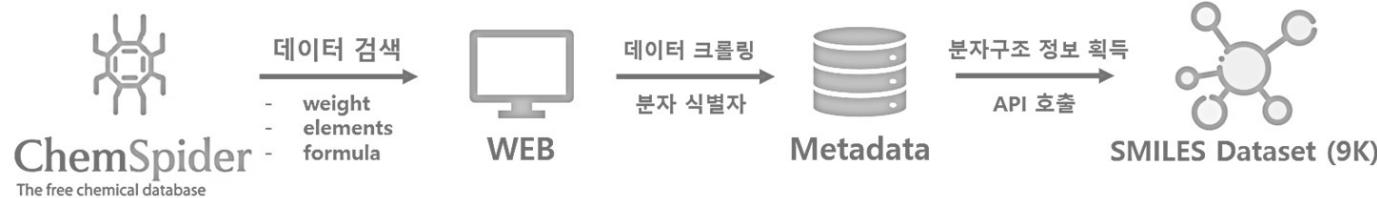


<분자 구조 정보 획득>

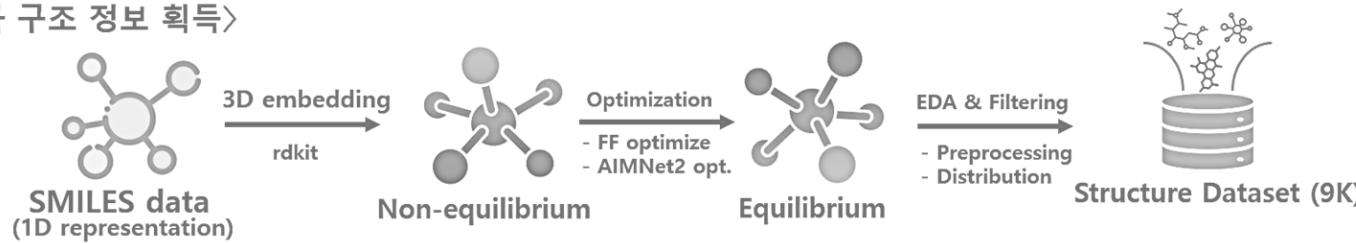


ML-Potential 모델 연구

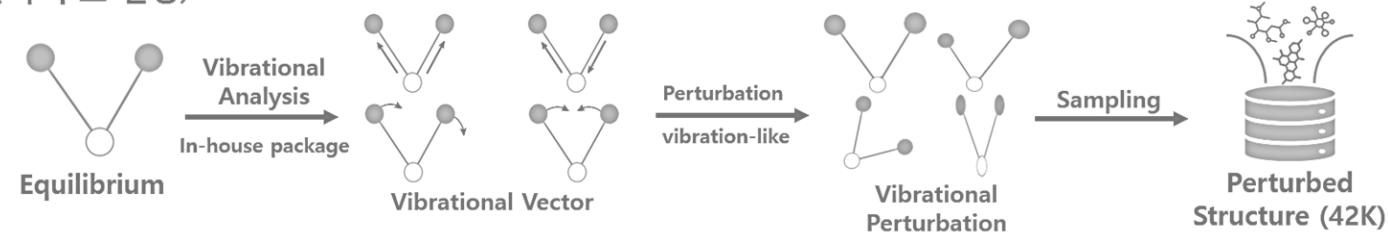
<분자 메타 데이터 수집>



<분자 구조 정보 획득>

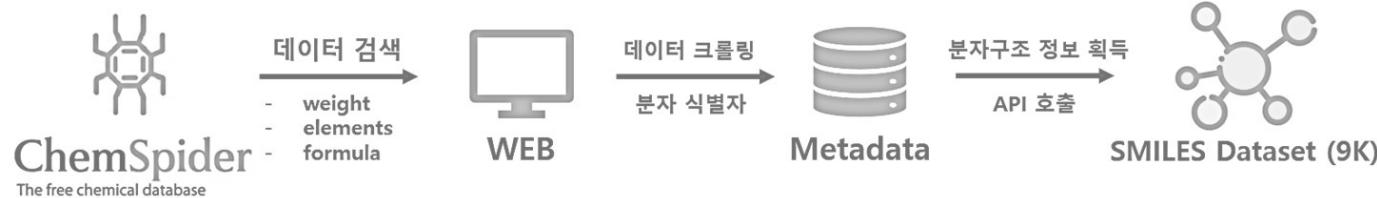


<분자 구조 변형>

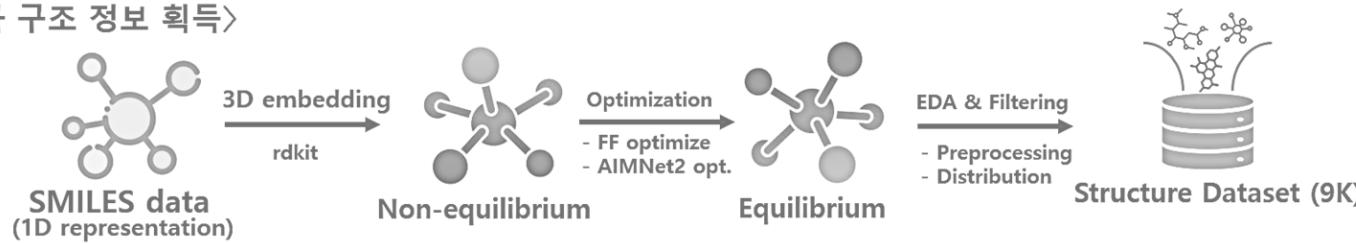


ML-Potential 모델 연구

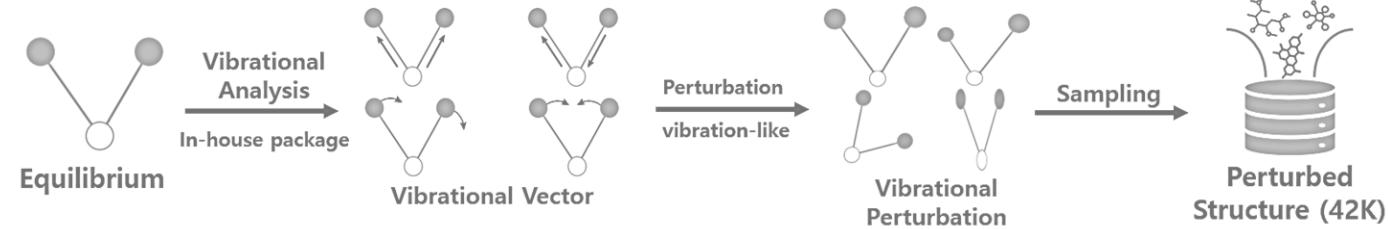
<분자 메타 데이터 수집>



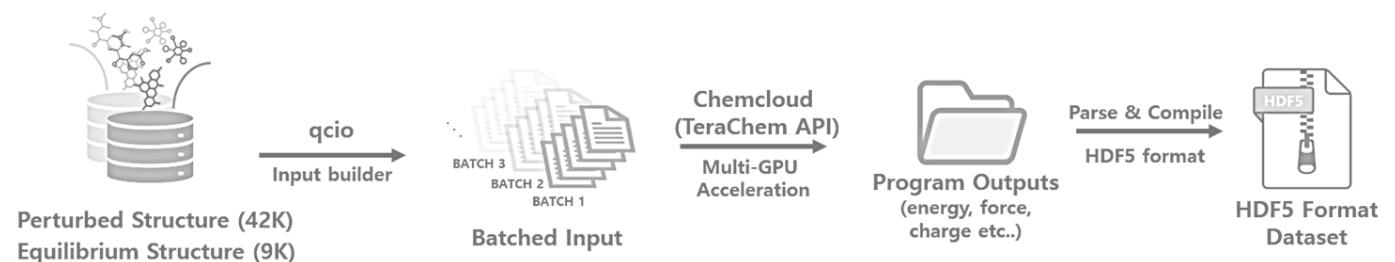
<분자 구조 정보 획득>



<분자 구조 변형>



<유기 분자 양자화학 계산>

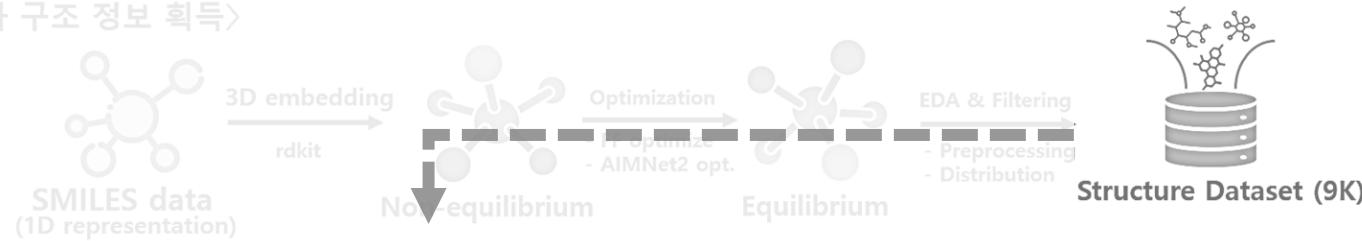


ML-Potential 모델 연구

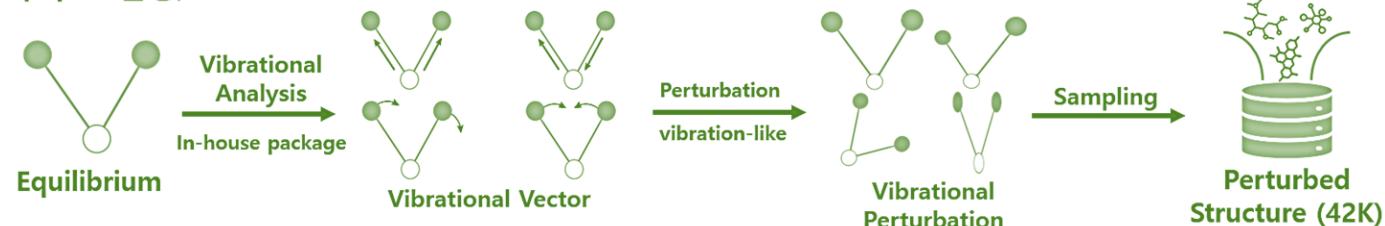
<분자 메타 데이터 수집>



<분자 구조 정보 획득>



<분자 구조 변형>



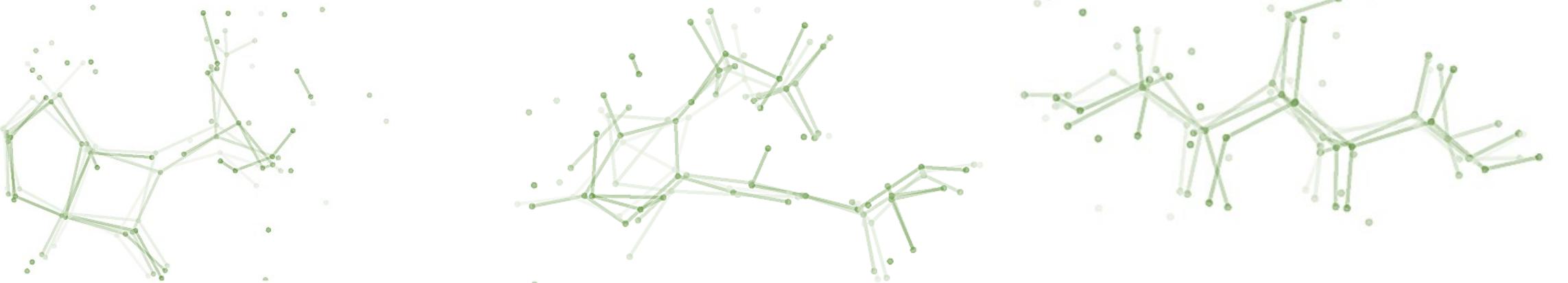
<유기 분자 양자화학 계산>

분자의 진동 방향으로 구조를 변형

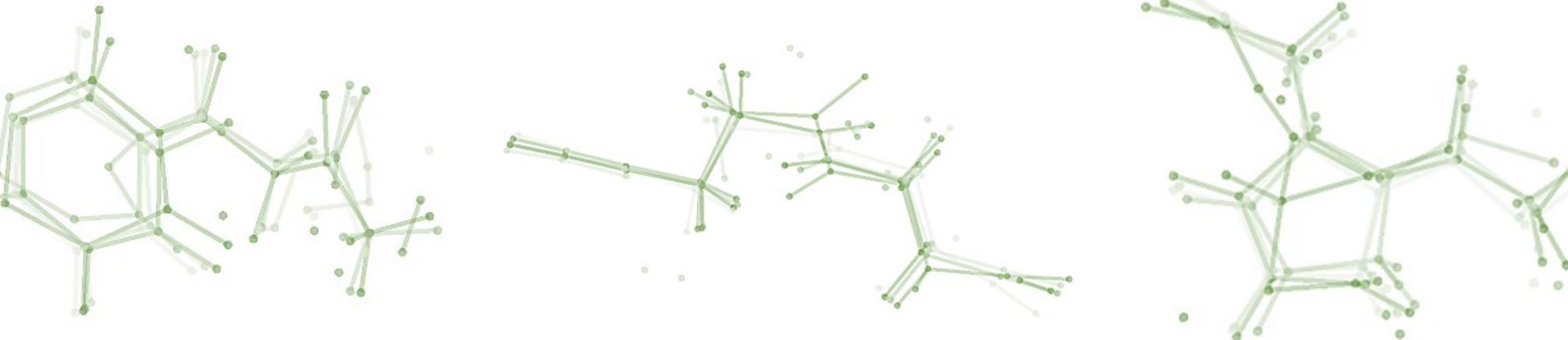
Perturbed Structure (42K)
Equilibrium Structure (9K)



ML-Potential 모델 연구



다양한 결합 각도, 길이 데이터를 포함한 데이터셋 (~42K)



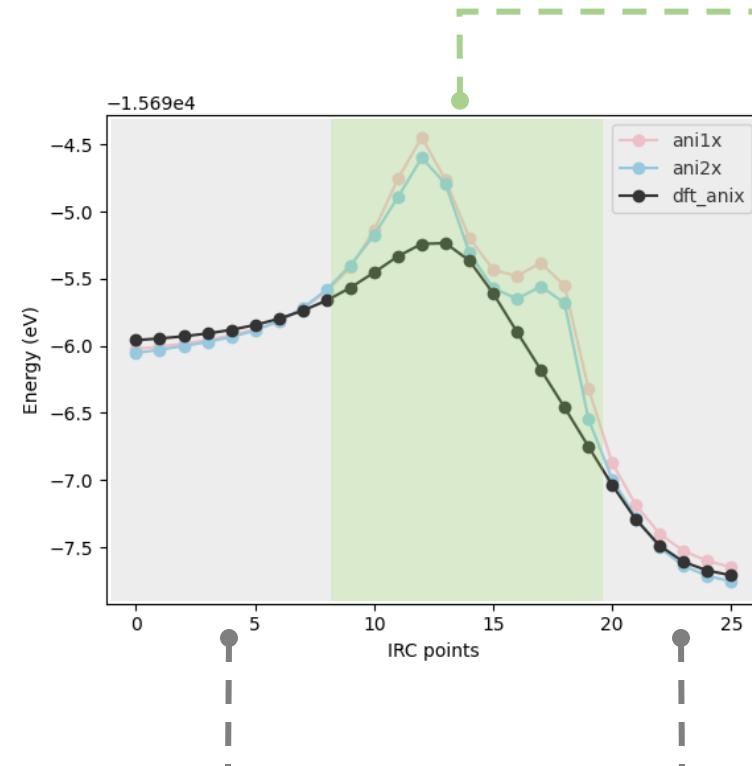
Benchmark



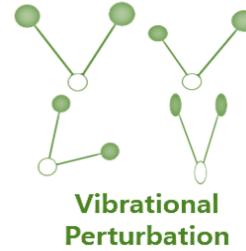
안정한 구조



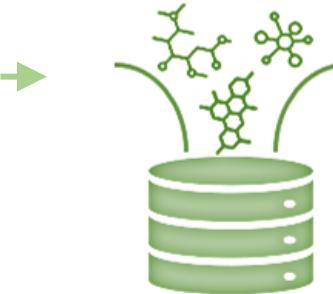
Structure Dataset (9K)



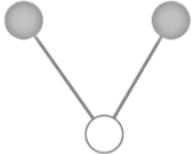
TS와 유사한
진동-변형 구조



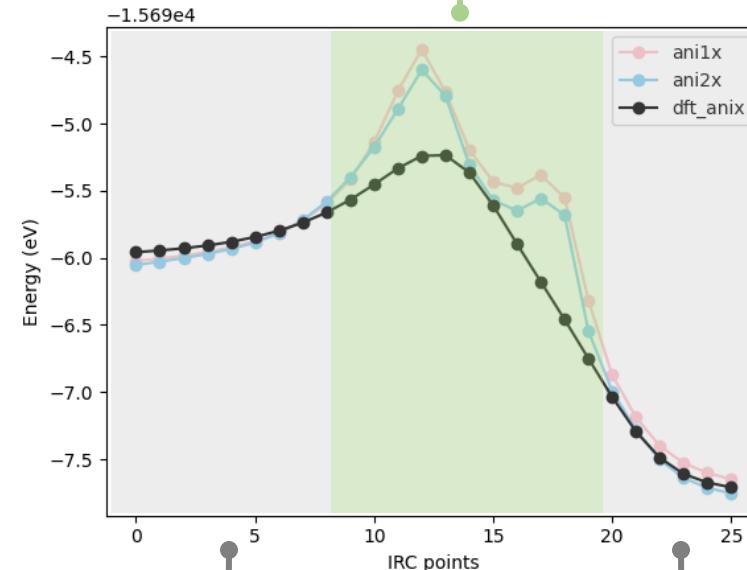
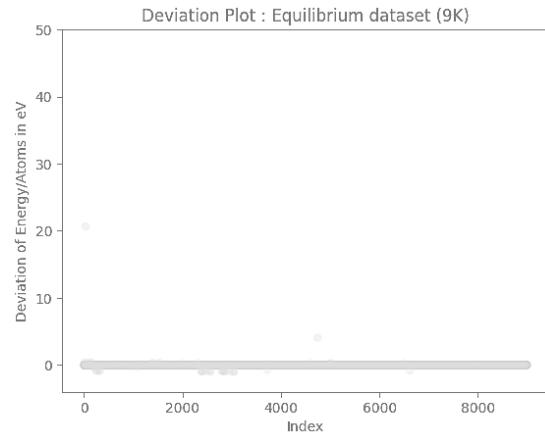
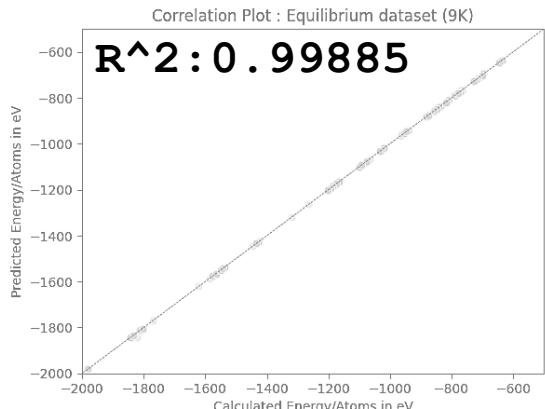
Perturbed
Structure (42K)



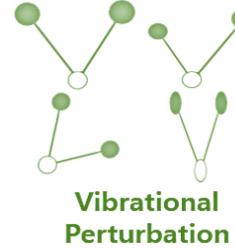
Benchmark



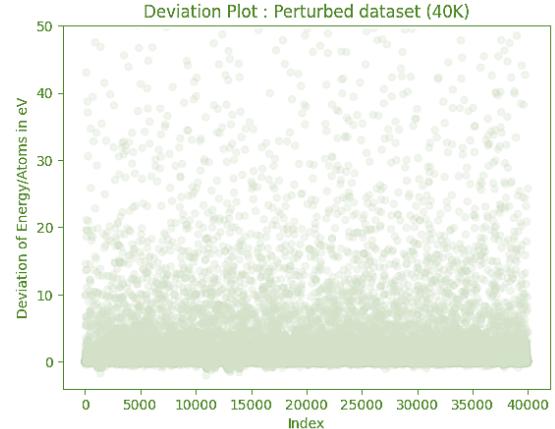
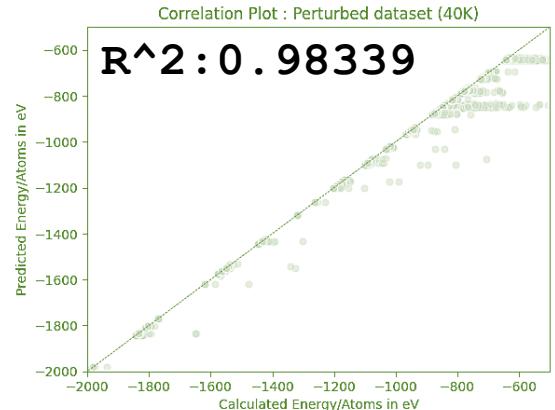
안정한 구조
Equilibrium



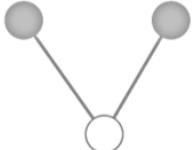
TS와 유사한
진동-변형 구조



Vibrational
Perturbation

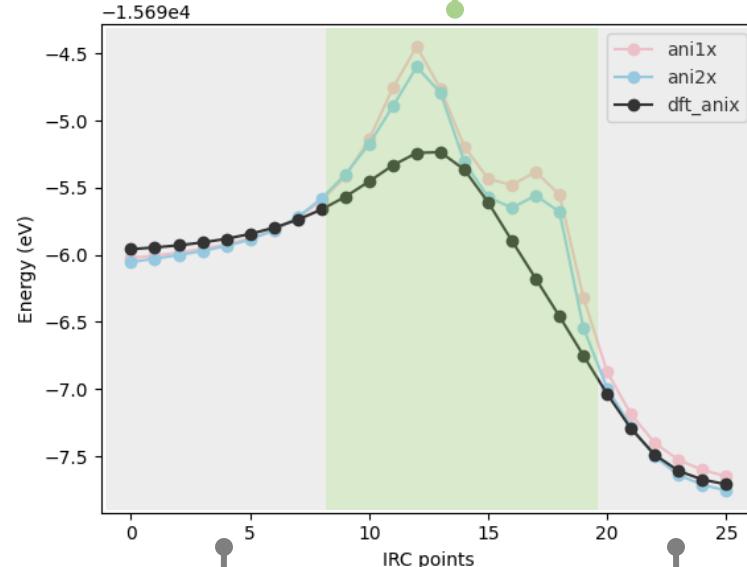
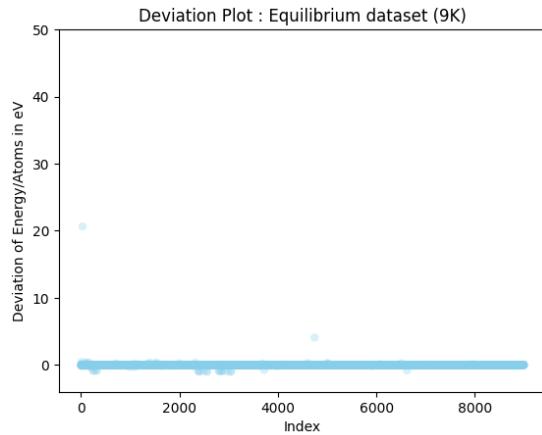
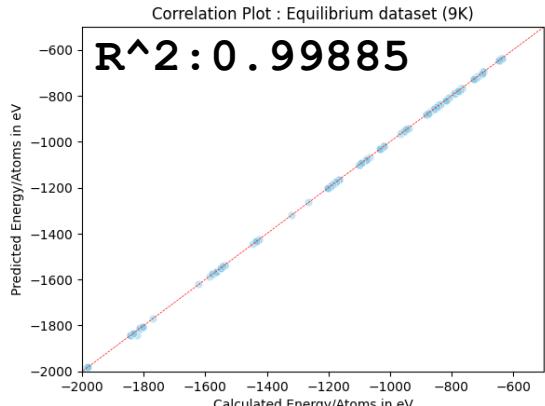


Benchmark

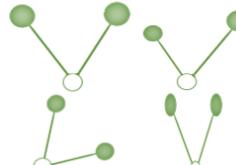


안정한 구조

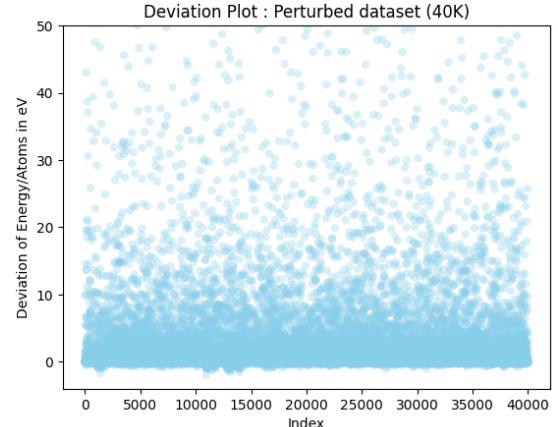
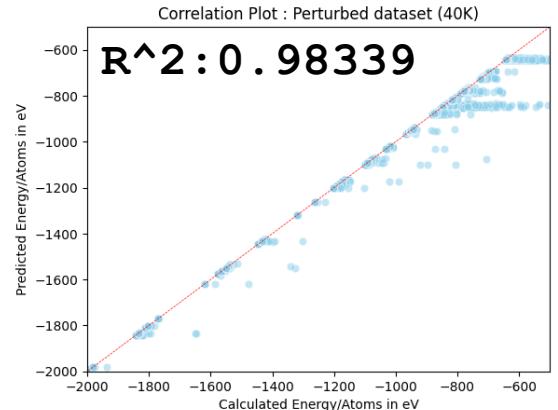
Equilibrium



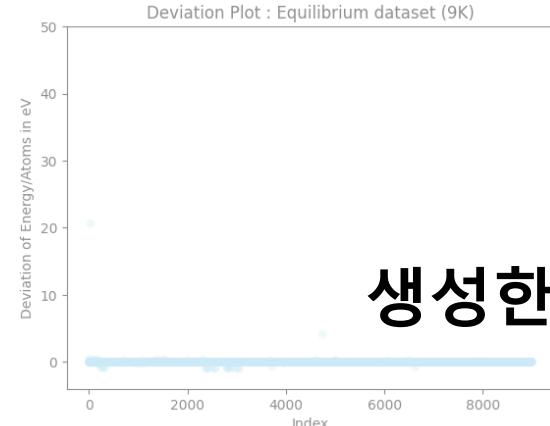
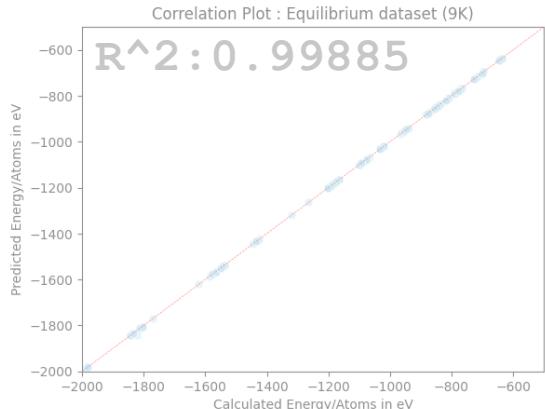
TS와 유사한
진동-변형 구조



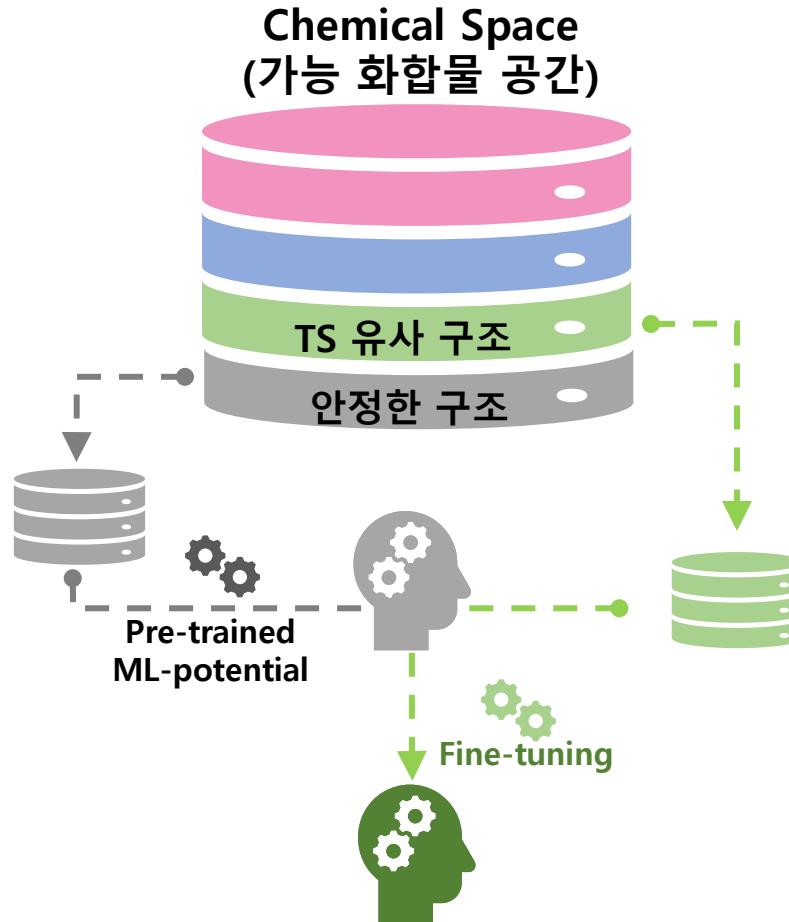
Vibrational
Perturbation



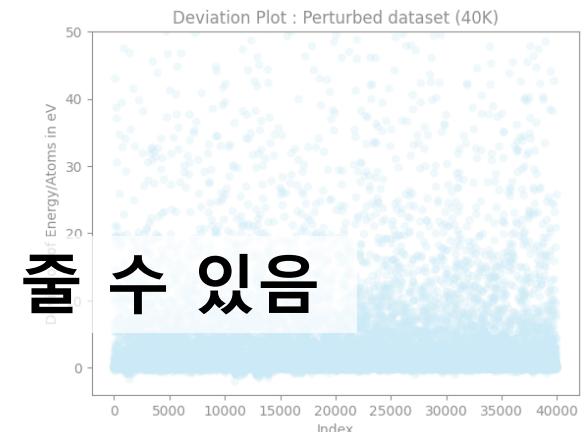
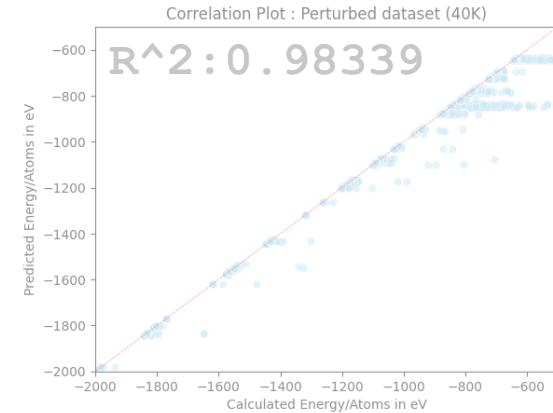
Benchmark



생성한 데이터셋이 기존 모델 성능 개선에 도움을 줄 수 있음



TS와 유사한
진동-변형 구조



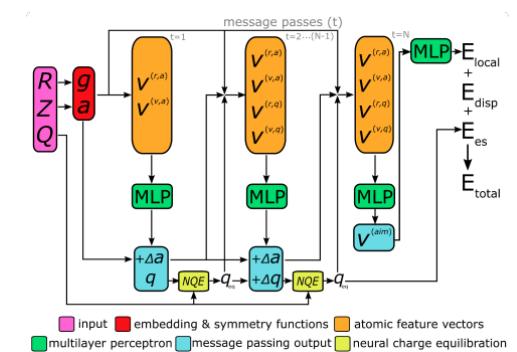
ML-Potential Modeling

1. TorchANI 모델 파인 투닝

모델링에 많은 시간을 투자하지 못함. 100 epoch 학습으로 **모델 성능 개선만 확인**

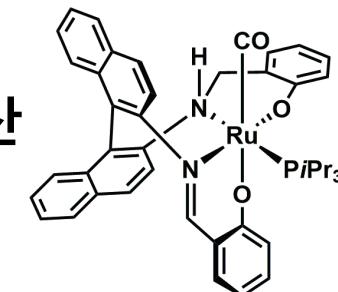
2. Organic AIMNet2 모델 훈련

- Molecular ML-Potential 분야의 **SOTA 모델**
- pre-trained model의 ckpt(pt)가 공개 X
- SPICE와 합친 후 **1M 규모 데이터로 훈련**

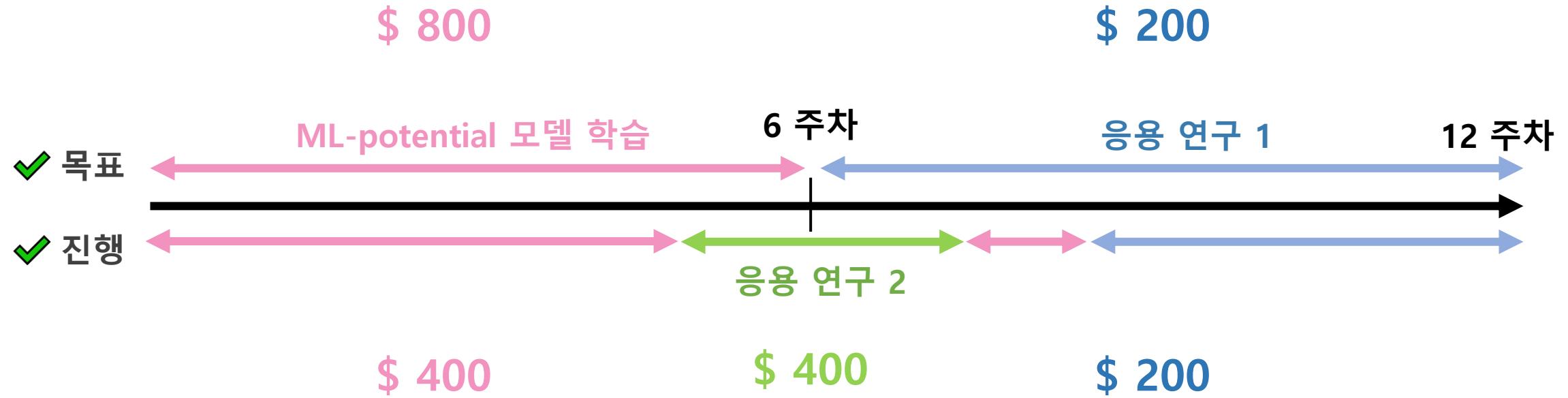


3. Ru-Catalyst AIMNet2 모델 훈련

- tmQM의 **Ru 촉매** 데이터셋에 대한 2.5K 규모 양자화학 계산
- **촉매 타겟** 모델 훈련



Fine Tuned Models



- ✓ 초기 계획에 없던 추가 프로젝트 진행 (더욱 유의미한 MD 관련 연구 수행)
- ✓ 프로젝트 기간, 지원금 문제 → 테스트 데이터셋 부족

ML-Potential Modeling

- ✓ 유의미한 데이터셋 생성 및 배포에도 큰 의미,
데이터셋 생성 방법론 제안 및 코드 공개
- ✓ 아직 ongoing project로 추후 테스트 후 모델 배포 예정

Acknowledgement

Questions?
(Feel free to email me!)

신재윤 교수님
프로젝트 지도 및 방향 조언

Yicheng Chen @SHU
ase-distortion-interaction 벤치마크를 위한 XPaiNN 모델 ckpt 공유

Roman Zubatyuk @CMU
AIMNet2 모델 오류 수정 및 훈련 관련 조언

고려대학교
프로젝트 연구비 \$1000를 지원받아 연구를 수행함

kangmg@korea.ac.kr

The code and model files can also be shared upon request.