

ADMET 예측을 위한 초다중모달 (약물, 단백질, 유전체, 세포 이미징, 문헌) 파운데이션 모델 개발

연구책임자: 안성수(KAIST 김재철 AI 대학원)

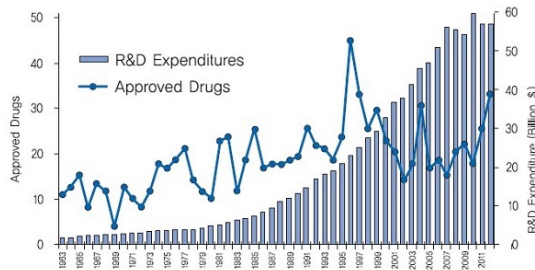
연구개발과제의 개요

목표

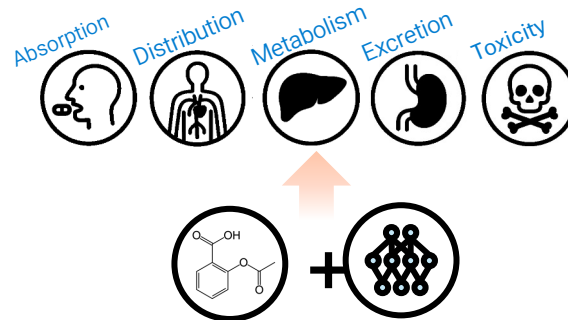
Federated ADMET 예측 모델(FAM) 개발

과제명

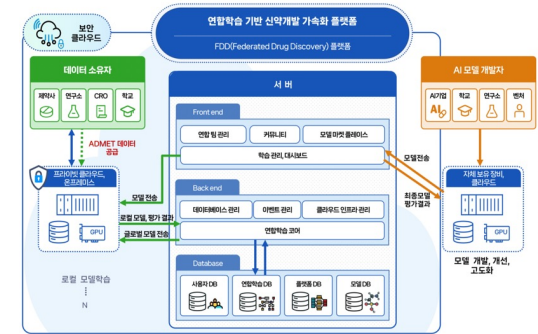
ADMET 예측을 위한 **초다중모달(약물, 단백질, 유전체, 세포 이미징, 문헌)** 파운데이션 모델 개발



해마다 하락하는 신약 생산성



AI 기반 ADMET 예측 모델



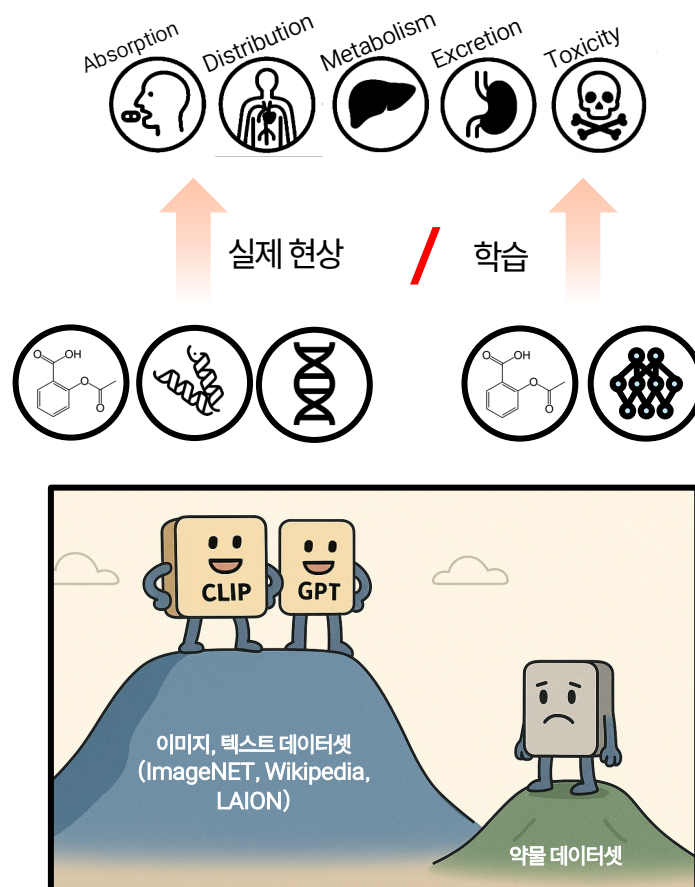
연합학습기반 신약개발 가속화 플랫폼
개요

연구개발과제의 필요성

단일 모달리티 모델의 한계

- 약물 화학 구조에만 의존하여 복합적인 생체 상호작용을 직접적으로 학습하지 못함
- 약물 데이터셋의 크기는 다른 AI 분야(이미지, 자연어 처리)에 비해 현격하게 크기가 작음

→딥러닝 모델의 강한 표현력을 활용 못함



연구개발과제의 필요성

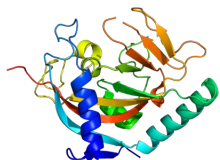
다중 모달리티 파운데이션 모델의 중요성

- 다중 모달리티 파운데이션 모델의 성능이 증가한 다수의 사례(CLIP, Flamingo) 존재
- 단백질, 유전체, 문헌, 세포 이미징 등의 데이터가 추가적인 정보로 존재
- 기존에 학습된 파운데이션 모델(ESM-2, Evo2, ChatGPT 등)이 존재

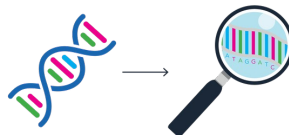
→ 약물 파운데이션 모델에 대한 지식 전이 가능



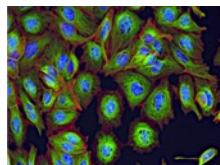
문헌: 방대한 생물의학 논문에는 약물의 표적 독성에 대한, 인과지식이 축적돼 있음



단백질: 많은 약물 관련 물성은 약물-단백질 상호작용으로 표현됨



유전체: 약물의 작용 기전, 부작용 예측, 바이오마커 정의 등에 큰 영향

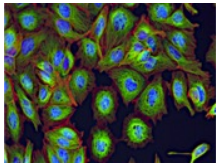


세포 이미지: 약물이 세포에 끼치는 영향을 이미지로 정보화하여 제공함.

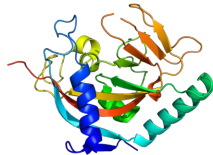
연구개발과제의 목표

정성적 목표

- ADMET 예측을 위한 초다중모달(약물, 단백질, 유전체, 세포 이미징, 문헌) 파운데이션 모델 개발
- 약물, 단백질, 유전체, 세포 이미징, 문헌을 통합한 데이터베이스 구축
- 대조학습을 기반 학습 알고리즘 개발



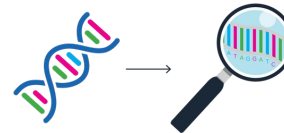
셀 이미지



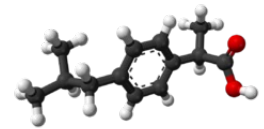
단백질 서열



과학적 문헌



유전체 서열



ADMET
파운데이션 모델

연구개발과제의 목표

정량적 목표

- 데이터: 분자구조, 문헌, 단백질, 유전체, 세포 이미지로 이루어진 대형 데이터수집(130만종 이상)
- 성능: TDC 리더보드 기준 최고성능 방법론(ADMET-AI) 추월(22개의 지표중 13개 이상 최고 성능 기록)

Leaderboard

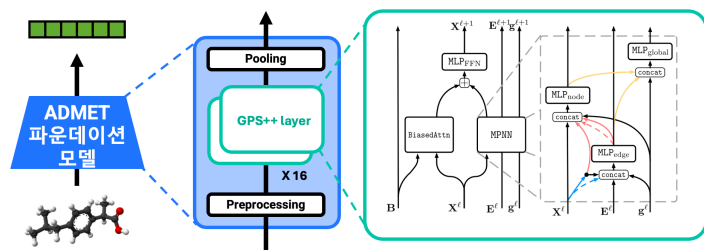
Rank	Model	Contact	Link	#Params	AUPRC ↓
1	MapLight + GNN	Jim Notwell	GitHub, Paper	N/A	0.916 ± 0.000
2	ContextPred	Kexin Huang	GitHub, Paper	2,067,053	0.904 ± 0.002
3	AttrMasking	Kexin Huang	GitHub, Paper	2,067,053	0.902 ± 0.002
4	MapLight	Jim Notwell	GitHub, Paper	N/A	0.881 ± 0.001
5	Chemprop-RDKit	Kyle Swanson	GitHub, Paper	N/A	0.876 ± 0.003
6	ZairaChem	Gemma Turon	GitHub, Paper	N/A	0.875 ± 0.002
7	Chemprop	Kyle Swanson	GitHub, Paper	N/A	0.862 ± 0.003
8	AttentiveFP	Kexin Huang	GitHub, Paper	300,806	0.851 ± 0.006
9	NeuralFP	Kexin Huang	GitHub, Paper	480,193	0.849 ± 0.004
10	GCN	Kexin Huang	GitHub, Paper	191,810	0.840 ± 0.010
11	RDKit2D + MLP (DeepPurpose)	Kexin Huang	GitHub, Paper	633,409	0.829 ± 0.007
12	Morgan + MLP (DeepPurpose)	Kexin Huang	GitHub, Paper	1,477,185	0.827 ± 0.009
13	CNN (DeepPurpose)	Kexin Huang	GitHub, Paper	226,625	0.821 ± 0.003
14	Euclia ML model	Euclia	GitHub, Paper	50	0.696 ± 0.004
15	Basic ML	Nilavo Boral	GitHub, Paper	N/A	0.654 ± 0.000

TDC 벤치마크 CYP3A4 _Veith 데이터셋
지표 예시

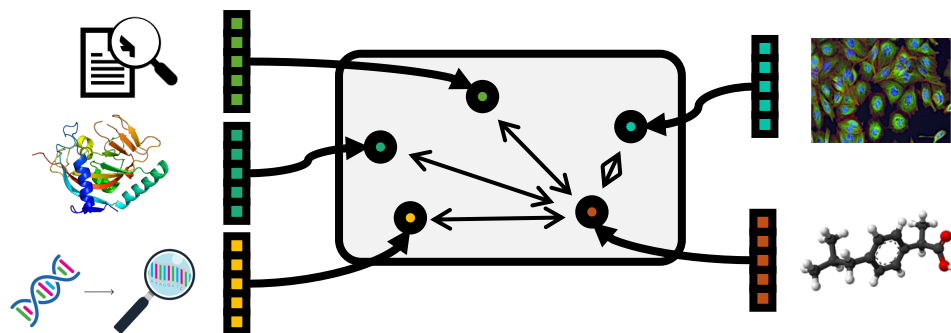
연구개발과제의 내용

신경망 구조 및 학습 방법

- 분자 구조 대상 최고성능을 보이는 **MolGPS++** 구조 사용
- 모달리티 간 관계성에 대한 **대조학습** 진행
- 각 모달리티별로 기존에 학습된 **초대형 파운데이션 모델 임베딩을 사용**
(알파폴드, Evo2, Galactica, OpenPhenom)



MolGPS++ 구조

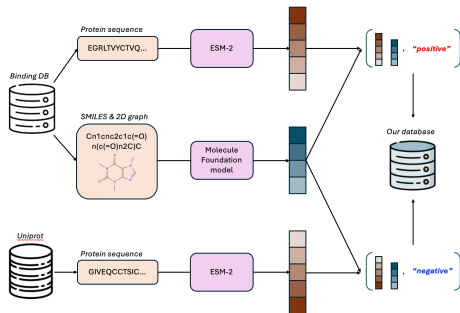


대조학습을 사용한 모달리티 간 관계성 학습

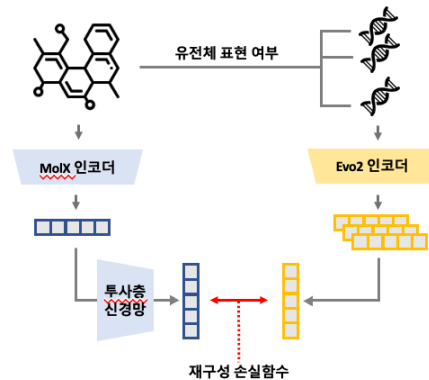
연구개발과제의 내용

문헌, 단백질, 유전체, 세포 이미지 모달리티의 활용

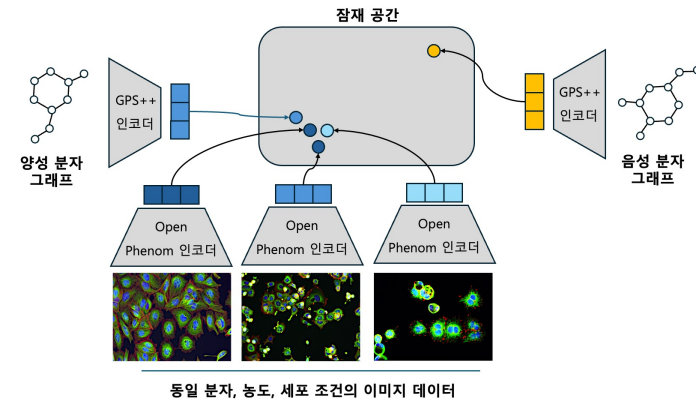
- 약물에 대한 **문헌**, 약물과 바인딩 된 **단백질**, 약물에 의해 억제/활성화된 **유전체(mRNA)**, 약물에 반응한 **세포 이미지** 등에 해당하는 임베딩 도출
- 약물과 해당 데이터에 대한 관계성을 고정된 임베딩과 약물 파운데이션 모델 간의 **대조학습**을 통해 학습



단백질 대조학습 방법론 개요



유전체 대조학습 방법론 개요

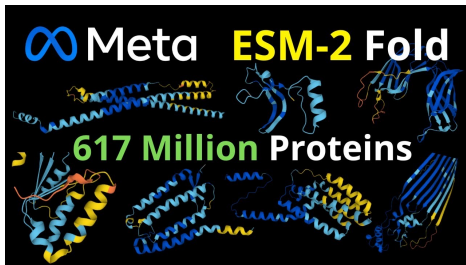


세포 이미지 대조학습 방법론 개요

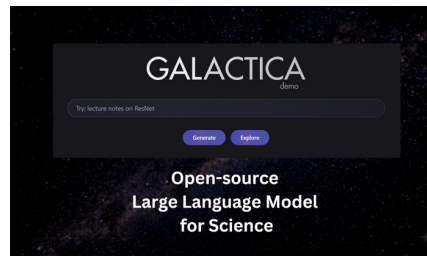
연구개발과제의 내용

파운데이션 모델의 적극적인 활용

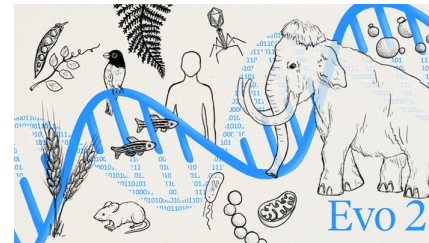
- ESM-2, Evo2, Galactica, OpenPhenom 등의 기존에 학습된 **대형 파운데이션 모델** 존재
- 해당 임베딩과 대조학습을 통해 약물 파운데이션 모델에 대한 **지식 전이**



Meta ESM-2
모델(단백질)



Meta Galactica 모델(문헌)



Meta Evo2
모델(유전체)



Recursion OpenPhenom
모델(세포 이미지)

연구개발과제의 내용

학습용 데이터셋 구성

- (단백질, 유전체, 세포 이미지)
다양한 **공개 데이터**를 활용하여 구축
- (문헌) 강력한 **대형 언어 모델 API를 활용**하여
분자 설명 문헌 데이터 베이스를 신규 구축

```
System: You are an expert medicinal chemist.  
User: Given the SMILES string, return a JSON with:  
  • "caption" – a 1-sentence plain-English description of the scaffold and key FGs  
  • "physchem" – pKa, logP and polar surface area (estimate)  
  • "druglikeness" – one sentence  
  • "use_case" – likely therapeutic area  
Answer only with valid JSON.  
  
Example SMILES: CC(=O)Oc1ccccc1C(=O)O
```

LLM에게 분자구조 설명문을 요구하는 프롬프트의 예시

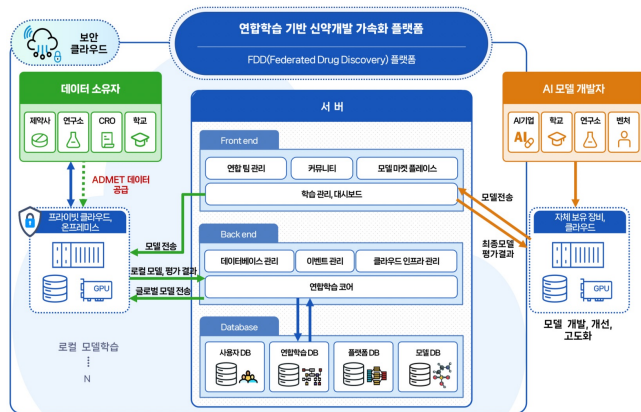
모달리티	데이터셋 이름	크기
약물	TDC (Absorption)	17K
	TDC (Distribution)	5K
	TDC (Metabolism)	50K
	TDC (Excretion)	2K
	TDC (Toxicity)	15K
	MoleculeNet (Absorption)	3K
	MoleculeNet (Distribution)	2K
	MoleculeNet (Toxicity)	19K
단백질	BindingDB	9K
	Uniprot	250M
유전체	Hetionet	2M
	ChemBL2K	2K
	Broad6k	6K
세포	RxRx3	2.2M
	BBBC021	39K

데이터 셋 세부 설명과 크기

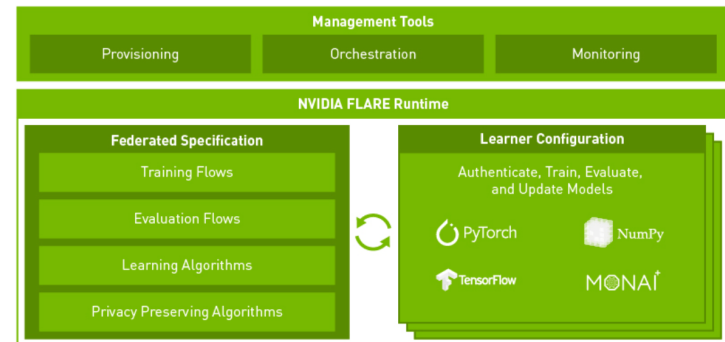
연구개발과제의 내용

연합학습 방법론

- NVIDIA FLARE 프레임워크를 사용하여 기존 학습 방법론 확장.
- 클라이언트 장비 스펙(H100 96GB GPU)에 맞춰 모델 학습 방법론 준비 예정.
- Dataset, model, training, evaluation에 대해 모듈화된 코드를 명시한 뒤 표준적인 학습 절차를 따름.



플랫폼 개요

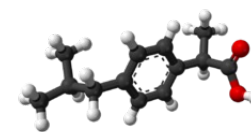
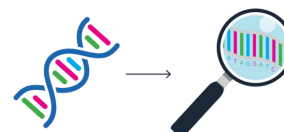
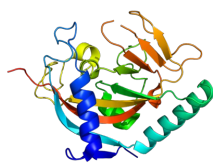
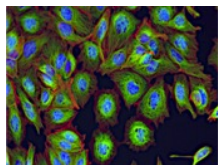


NVIDIA FLARE 개요

결론

ADMET 예측을 위한 초다중모달 파운데이션 모델 개발

- 문헌, 단백질, 유전체, 세포 이미징 모달리티에 대한 약물 모델 데이터의 확장
- 기존 파운데이션 모델(ESM-2, AlphaFold, Evo2, ChatGPT, OpenPhenom)을 적극 활용하는 학습 기법



셀 이미지

단백질 서열

과학적 문헌

유전체 서열

ADMET
파운데이션 모델