Boost up AI 2025: 신약 개발 경진대회

Seongsu Moon (dalcw@jnu.ac.kr)

School of Artificial Intelligence

Chonnam National University

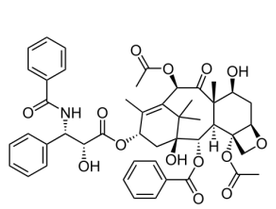
배경

- **CYP3A4 효소 저해 예측 모델 개발**
  - CYP3A4는 많은 약물 대사에 관여함. 즉, CYP3A4는 약물을 분해하는 효소임
  - 만약, 특정 약물이 **CYP3A4를 억제**하면,

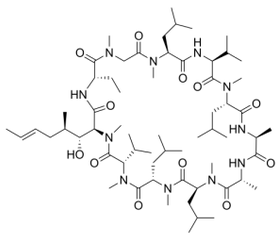
    동시에 투여되는 다른 약물의 대사가 느려져 **독성이 증가**하거나, 반대로 **효능이 소실**될 수 있음

→ **따라서, CYP3A4의 저해율을 예측하는 것은, 신약 개발 과정에서 필수적인 요소임**

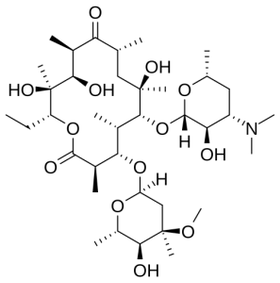
CYP3A4 효소



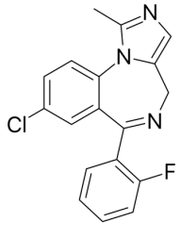
항암제



면역 억제제

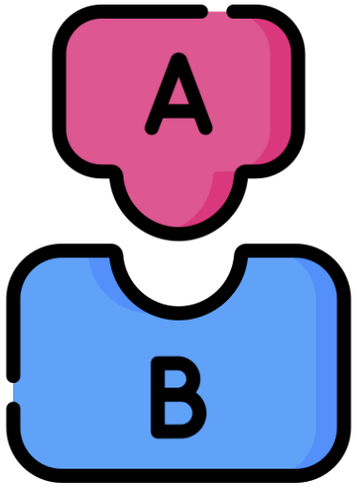


항생제



항우울제

기질



분해산물

데이터 셋

- **데이터 셋에는 두 가지 ‘어려움’이 존재함**

1. 데이터 셋의 수가 매우 적음

- 학습에 사용 가능한 데이터의 수가 1,681개 정도 밖에 안됨
- 이는, 딥러닝 방법론은 커녕, 머신러닝 방법론을 적용하는 것까지도 한계가 있음 (차원의 저주)

→ 정확한 예측 모델을 개발하기 위해서는 추가적인 데이터가 필요함

2. 수동적 특징 추출

- 제공되는 데이터 셋은, 화합물의 구조를 담고 있는 smiles 정보 (e.g., Brc1ccc2OCCc3ccnc1c23)와, 타깃 정보에 해당하는 inhibition 수치 (e.g., 4.45)만이 제공됨

→ 따라서, 특징은 smiles 정보에서 직접 추출해야함

본 태스크는 단순한 회귀 문제가 아님!
→ 화합물의 구조에서 유의미한 저해 특성을 추출해야하며,
→ 소량의 데이터에서 신뢰서 있는 예측을 수행해야하는
고난이도 예측 문제임

추가 데이터 셋

- **정확한 예측을 위해서는 추가 데이터 셋이 필요함**
  - CYP 저해율을 예측하는 다른 연구들에서 사용하는 **PubChem AID 1851, 884** 데이터 셋 사용
  - 해당 데이터 셋은 공개 라이선스로 지정되어 있음

Among the most important resources of measured data on CYP inhibition are the CYP assay data sets from the PubChem Bioassay database[13] (AID 1851,[14], 410,[15] 883,[16] 884,[17] 899[18] and 891[19] data sets), as well as measured CYP data from the ChEMBL database[20] and the ADME Database.[21] PubChem Bioassay 1851, deposited in the year 2009, provides enzyme inhibition data for 17143 compounds measured in an enzyme panel assay covering CYP1A2, 2C9, 2C19, 2D6 and 3A4.[22] The earlier deposited data records, namely PubChem Bioassay 410, 883, 884, 899 and 891, have a significant overlap with assay 1851. The ChEMBL database and ADME Database store manually curated bioactivity data collected from a variety of resources, including individual research papers. They include thousands of records relevant to CYP inhibition.

CYPlebrity: Machine learning models for the prediction of inhibitors of cytochrome P450 enzymes 논문 일부[2]

PubChem

PubChem

Content	
Description	Chemicals and their bioassays
Organisms	Humans and other animals
Contact	
Research center	NCBI
Primary citation	PMID 15879180
Access	
Website	pubchem.ncbi.nlm.nih.gov
Download URL	FTP
Web service URL	PUG-View[1]
Miscellaneous	
License	Public domain

License

추가 데이터 셋

- 그러나, 데이터 셋이 대회에서 주어진 것과 완전히 동일하지는 않음

	대회 제공	PubChem
실험 환경	10um	11.43um
표현형태	Inhibition(%)	Activation(%)

- 이 데이터를 사용하는 방법으로는 2가지가 있음

A. Inhibition(%) = 100 - Activation(%)로 가정하는 방법

B. Hill equation을 이용하여 10uM 환경에서의 Inhibition(%)를 추정하는 방법

$$\text{Activity(L)} = A_0 + \frac{A_\infty - A_0}{1 + (\frac{L}{AC_{50}})^{hill}} \qquad \text{Inhibition(L)} = \frac{A_0 - \text{Activity(L)}}{A_0 - A_\infty} \times 100$$

Hill equation

추가 데이터 셋

- 그러나, 데이터 셋이 대회에서 주어진 것과 완전히 동일하지는 않음

	대회 제공	PubChem
실험 환경	10um	11.43um
표현형태	Inhibition(%)	Activation(%)

- 이 데이터를 사용하는 방법으로는 2가지가 있음

  A. Inhibition(%) = 100 - Activation(%)로 가정하는 방법

  B. Hill equation을 이용하여 10uM 환경에서의 Inhibition(%)를 추정하는 방법

$$\text{Activity(L)} = A_0 + \frac{A_\infty - A_0}{1 + (\frac{L}{AC_{50}})^{hill}} \qquad \text{Inhibition(L)} = \frac{A_0 - \text{Activity(L)}}{A_0 - A_\infty} \times 100$$

Hill equation

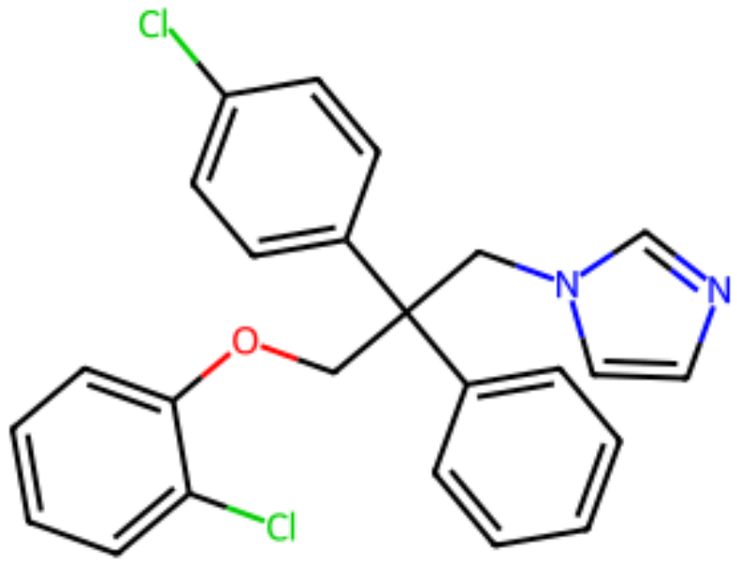
→ **결론적으로는 "A. Inhibition(%) = 100 - Activation(%)"의 방법을 사용함 (성능적으로 우위에 있음)**

→ **학습 데이터가 17,996개로 증강됨**

도메인 분석

Physical features[5,6]
: **분자의 크기, 유연성, 극성 등의 물리학적 특징**을 의미하며, 이는 리간드가 CYP3A4와 상호작용 가능한지 예측하는 데 필수적임

Fingerprint (Morgan, MACCS)[5, 6]
: **분자의 구조를 벡터화**하여 머신러닝 모델이 학습 가능하게 함. 이로써, 모델에게 리간드의 구조적 힌트를 알려줄 수 있음



SMARTS patterns[2,5]
: **CYP3A4에서 중요하다고 알려진 구조적 패턴**을 반영함. 이를 플래그 형식 (0 또는 1)로 표현함

Graph representation[7]
: **원자 간 연결과 주변 환경 정보를 그대로 반영**하여, 구조적 맥락까지 학습할 수 있도록 함. 주로 GNN 모델을 사용함

특징 추출

- 선행 연구에 근거하여 다음과 같은 특징들을 추출함 (논문 근거)

	features
Atom features[7] (22개)	원자번호, 방향족 여부, 혼성화 여부, 형식전하, 고리 구조 여부, 총 수소 수, 연결된 원자 수, 암시적 원자가 수, 명시적 수소 수, 암시적 수소 수, 원자 질량, 동위 원소 번호, 입체 중심 태그, 암시적 수소 금지 여부, 입체 중심 여부, Gasteiger 전하, 라디칼 전자 수, 총 원자가수
Physical features[5,6] (15개)	**분자량**(MolWt), **지용성**(MolLogP), **극성 표면적**(TPSA), **회전 가능한 결합 수** (NumRotatableBonds), **수소 결합 공여체/수용체 수**(NumHDonors/ NumHAcceptors), **방향족 고리 수**(CalcNumAromaticRings), **전체 고리 수** (CalcNumRings), **sp3 탄소 비율**(CalcFractionCSP3), **수소 제외 원자 수** (HeavyAtomCount), **접근 가능한 표면적**(CalcLabuteASA), **물 굴절률** (MolMR), **정밀 분자량**(CalcExactMolWt), **원자가 전자 수** (NumValenceElectrons), **인(P) 원자수**
Fingerprint[5,6] (2215개)	• **Morgan**: radius=2, fpSize=2048 • **MACCS**
SMARTS[2,5] (6개)	**다음 구조들이 존재하는지 플래그 형태로 기록** Imidazole, Tertiary amine, Furan, Acetylene, Pyridine, Thiophene

특징 추출

- 선행 연구에 근거하여 다음과 같은 특징들을 추출함 (논문 근거)

	features
Atom features[7] (22개)	원자번호, 방향족 여부, 혼성화 여부, 형식전하, 고리 구조 여부, 총 수소 수, 연결된 원자 수, 암시적 원자가 수, 명시적 수소 수, 암시적 수소 수, 원자 질량, 동위 원소 번호, 입체 중심 태그, 암시적 수소 금지 여부, 입체 중심 여부, **Gasteiger** 전하, 라디칼 전자 수, 총 원자가수
Physical features[5,6] (15개)	**분자량**(MolWt), **지용성**(MolLogP), **극성 표면적**(TPSA), **회전 가능한 결합 수** (NumRotatableBonds), **수소 결합 공여체/수용체 수**(NumHDonors/ NumHAcceptors), **방향족 고리 수**(CalcNumAromaticRings), **전체 고리 수** (CalcNumRings), **sp3 탄소 비율**(CalcFractionCSP3), **수소 제외 원자 수** (HeavyAtomCount), **접근 가능한 표면적**(CalcLabuteASA), **물 굴절률** (MolMR), **정밀 분자량**(CalcExactMolWt), **원자가 전자 수** (NumValenceElectrons), **인(P) 원자수**
Fingerprint[5,6] (2215개)	• **Morgan**: radius=2, fpSize=2048 • **MACCS**
SMARTS[2,5] (6개)	**다음 구조들이 존재하는지 플래그 형태로 기록** Imidazole, Tertiary amine, Furan, Acetylene, Pyridine, Thiophene

Deep Learning! for GNN

Machine Learning! for h2o

모델링
머신러닝 방법론

- **H2O(AutoML)을 이용하여 모델링 함**
  - **고차원 희소 특징에 강한 트리 모델 기반**
    - 학습에 사용되는 데이터는 2,236 차원의 희소 벡터임
    - 트리 기반 모델은 노드를 분할하면서 중요한 특징을 자동으로 탐색하기에, 해당 데이터를 분석하기에 유리함

  → H2O AutoML은 이러한 트리 모델을 기반을 다수 포함하고 있어, 구조적 특성에 최적화된 학습 가능

  - **다양한 트리 모델 앙상블로 일반화 성능 극대화**
    - 다양한 트리 모델을 학습하고, 이러한 모델들을 통합한 stacked ensemble을 자동으로 생성함

  → 단일 모델보다 잡음에 강하고 일반화 능력이 뛰어난 모델을 생성할 수 있음

  - **자동화된 파이프라인**
    - H2O AutoML은 하이퍼파라미터 튜닝, 모델 탐색, 교차 검증을 자동으로 처리함

  → 개발 시간이 제한된 환경에서 높은 효율을 보임

“수동적인 모델 탐색에 비해, H2O AutoML은 고성능과 고효율을 동시에 달성함”

모델링
머신러닝 방법론

H2O.ai

- **학습 및 결과**
  - **학습**
    - **max_runtime_sces**: 20000 (약, 5시간 30분 탐색)
    - **sort_metrics**: RMSE
    - **data**: 대회 제공 데이터 + 추가 수집 데이터 (2,236개의 데이터)
  - **결과**
    - 평가는, cross-valiation을 이용함
    - cross-validation은 학습 데이터 손실 없이 모델의 적합성을 판단할 수 있음

Train data에 대한 평가 결과

```
ModelMetricsRegressionGLM: stackedensemble
** Reported on train data. **

MSE: 99.17267330113596
RMSE: 9.958547750607813
MAE: 7.4673272275430955
RMSLE: NaN
Mean Residual Deviance: 99.17267330113596
R^2: 0.872202271597651
Null degrees of freedom: 10046
Residual degrees of freedom: 10018
Null deviance: 7798359.301059321
Residual deviance: 996387.848656513
AIC: 74756.82847984841
```

약간의 과적합이 보임

Validation data에 대한 평가 결과

```
ModelMetricsRegressionGLM: stackedensemble
** Reported on validation data. **

MSE: 356.24624876324873
RMSE: 18.87448671522616
MAE: 14.08883633266921
RMSLE: 0.8772002958062967
Mean Residual Deviance: 356.24624876324873
R^2: 0.5596155084900729
Null degrees of freedom: 1905
Residual degrees of freedom: 1877
Null deviance: 1545053.6405246744
Residual deviance: 679005.350142752
AIC: 16667.929605092748
```

모델링
딥러닝 방법론

- **리간드의 그래프 구조를 이용하여 모델링함**
  - **분자는 본질적으로 그래프 구조를 지님**
    - 분자는 원자(노드)와 결합(엣지)로 이루어진 구조로, 이는 자연스럽게 그래프로 표현됨

    → 구조 기반 특성이 중요한 CYP3A4 저해제 예측 같은 문제에서는 분자의 위치/연결성/구조 자체가 중요한 특징이 됨

  - **CYP3A4 효소는 약물의 “구조적인 모양”에 따라 달라짐**
    - CYP3A4는 결합 포켓이 넓고 유연하여, 리간드의 접근 방향/위치/형태가 저행 여부에 큰 영향을 줌

    → 이와 같은 통합 구조적 맥락은 Morgan이나 MACCS를 사용하는 방식으로는 한계가 있음

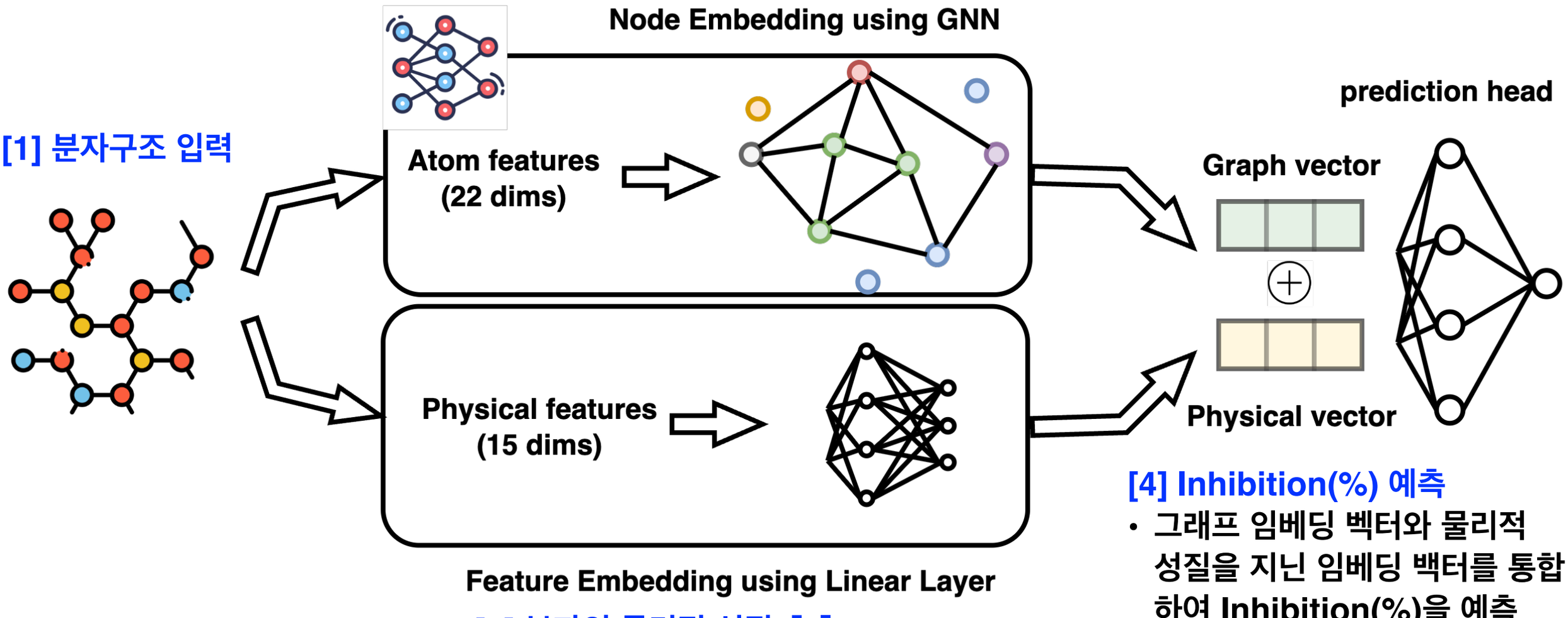
  - **그러나, 딥러닝(e.g., GNN) 모델을 학습하기에는 데이터 셋의 크기가 매우 적음**
    - 적은 수 의 데이터에서는, 학습의 정도가 불안정하며, 효과적인 예측이 불가능할 수 있음

    → 데이터 증강 기법(e.g., mixup)을 이용하고, 이에 더해 앙상블 시 weight를 조절함

모델링
딥러닝 방법론

[2] 분자 → 그래프 변환
- 분자의 화학 구조를 G = (V, E) 로 변환
- 각 원자는 노드 특성 벡터를 지님 (원자 속성)

[1] 분자구조 입력



[3] 분자의 물리적 성질 추출
- 15개의 물리학적 특징 (e.g., 분자량, 극성 표면성)을 추출함

[4] Inhibition(%) 예측
- 그래프 임베딩 벡터와 물리적 성질을 지닌 임베딩 백터를 통합하여 Inhibition(%)을 예측

모델링
딥러닝 방법론

- **학습 및 결과**
  - **학습**
    - **optimizer**: AdamW(lr = 0.0001)
    - **epochs**: 100
    - **data**: 대회 제공 데이터 (1,681개)

      학습 데이터가 적은 한계를 극복하기 위해 Mix-up 방법을 학습 과정에 도입하여 데이터 증강

c.f., mix-up

- 두 샘플 $(x_i, y_i), (x_j, y_j)$에 대해 새로운 샘플 $(\tilde{x}, \tilde{y})$를 생성하는 방식은 다음과 같음

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$
$$(\lambda \sim \text{Beta}(\alpha, \alpha) : \alpha > 0)$$

  - **결과**
    - 검증 데이터로는 전체 데이터에서 랜덤으로 100개의 데이터를 샘플링함
    - 평가 지표는 RMSE와 대회에서 주어진 평가 산식을 이용함

```
RMSE: 21.5
Contest metrics: 0.608
```

모델링
앙상블

- **H2O 모델과 GNN 모델의 결과를 통합함**[3]
  - 다수의 논문에서 딥러닝 기반의 모델보다, 머신러닝 모델의 성능이 우세하다고 알려짐
  - 따라서, ensemble weight를 **{h2o: 0.9, gnn: 0.1}**로 설정함

formula: $0.9 \times \{\text{H2O model}\} + 0.1 \times \{\text{GNN model}\}$

모델링
정리

- **본 모델은 다음과 같은 전략으로 높은 성능을 달성함**
  1. 구조적 특성 + 수치적 특성 융합
  2. 희소고차원 입력에 강한 트리 기반 AutoML 적용
  3. GNN + Feature Ensemble을 통한 구조 보완
  4. 데이터 부족 문제를 고려한, Mix-up 및 외부 데이터 증강

결과 및 '그 외의 다양한 시도'

- 다양한 시도 및 그 결과(Public score 기반)

	Description	Public score
사전학습 모델 이용	self-supervised GNN 활용	≅0.65
	Semisupervised learning	≅0.65
	Pretrained Chemberta 사용	≅0.65
	Pretrained PYTDC 사용	≅0.65
	Pretrained IBM 모델 사용	≅0.65
머신러닝 및 AutoML 사용	다수의 XGBoost 모델을 앙상블	0.69
	Optuna를 이용하여 XGBoost 튜닝	0.68
	Autogluon AutoML을 이용한 학습	0.68
	TPOT AutoML을 이용한 학습	0.68
	H2O AutoML을 이용한 학습	0.72
GNN 모델 사용	GAT 모델을 이용하여 학습	-
	GIN 모델을 이용하여 학습	0.71
고차원 특징 추출	docking score 계산 및 추가[1]	0.69
	H2O + GIN 앙상블	0.77

최종 모델

참고

- 학습 데이터가 매우 적기 때문에, 학습 환경이 동일하다고 하더라도, 완벽히 동일한 재현은 불가능함
- 특히 AutoML과 딥러닝 모델은 학습 과정상 다수의 확률적 요소들이 있기에 매 실행마다 결과가 달라질 수 있음

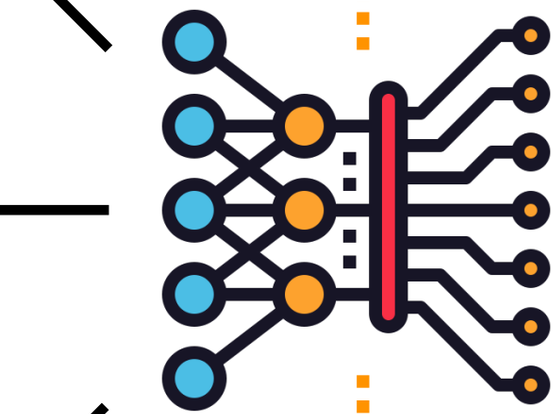
범용성

CYP 계열 저해 예측

: CYP2D6, CYP2C9 등 다른 대사 효소의 저해 예측에 확장 가능

재료 과학

: 분자 또는 결정 구조를 기반으로 전기 전도도, 열전도율, band gap 등 물성 예측



약물 간 상호작용 예측

: 한 약물이 CYP3A4를 억제할 경우 다른 약물의 대사가 억제됨
→ 병용 위험 사전 예측

소프트웨어 보안

: 그래프를 기반으로 소프트웨어에서의 취약점 존재 여부 또는 그 경로를 탐지 가능

약물 독성 예측

: hERG 독성, 간 독성, 흡수/분포/대사 특성 예측으로 확장 가능

화학 공정

: 두 개 이상의 화합물이 반응할 때, 반응 조건에 따라 반응 수율(%) 또는 반응 시간 예측

바이오 분야

비 바이오 분야

감사합니다

Appendix

Appendix A. 개발 환경

- **OS**: Ubuntu 24.04
- **Processor**: Intel Xeon Silver
- **RAM**: 256GB
- **GPU**: RTX 4090 (24GB)

Appendix B. 외부 데이터 출처

- **PubChem AID 1851 dataset**
  - https://pubchem.ncbi.nlm.nih.gov/bioassay/1851

- **PubChem AID 884 dataset**
  - https://pubchem.ncbi.nlm.nih.gov/bioassay/884

Appendix C. 참고자료

1. **Beck, Tyler C., et al. "Descriptors of cytochrome inhibitors and useful machine learning based methods for the design of safer drugs."** ***Pharmaceuticals*** **14.5 (2021): 472.**

2. **Plonka, Wojciech, et al. "CYPlebrity: Machine learning models for the prediction of inhibitors of cytochrome P450 enzymes."** ***Bioorganic & medicinal chemistry*** **46 (2021): 116388.**

3. **Permadi, Elpri Eka, Reiko Watanabe, and Kenji Mizuguchi. "Improving the accuracy of prediction models for small datasets of Cytochrome P450 inhibition with deep learning."** ***Journal of Cheminformatics*** **17.1 (2025): 66.**

4. **Fluetsch, Andrin, et al. "Deep learning models compared to experimental variability for the prediction of CYP3A4 time-dependent inhibition."** ***Chemical research in toxicology*** **37.4 (2024): 549-560.**

5. **Gong, Changda, et al. "Evaluation of machine learning models for cytochrome P450 3A4, 2D6, and 2C9 inhibition."** ***Journal of Applied Toxicology*** **44.7 (2024): 1050-1066.**

6. **Ai, Daiqiao, et al. "DEEPCYPs: A deep learning platform for enhanced cytochrome P450 activity prediction."** ***Frontiers in Pharmacology*** **14 (2023): 1099093.**

7. **Weiser, Benjamin, et al. "Machine learning-augmented docking. 1. CYP inhibition prediction."** ***Digital Discovery*** **2.6 (2023): 1841-1849.**