

·
·
·

**시뮬레이션을 활용한
멀티 오믹스 데이터 분석
및 인공지능 신약개발 활용**

·
·
·

목차

Index

⋮

01

제안이유

02

제안내용

03

활용 방안
및 기대효과

04

참고문헌

⋮

01 제안이유

기존 모델의 한계점

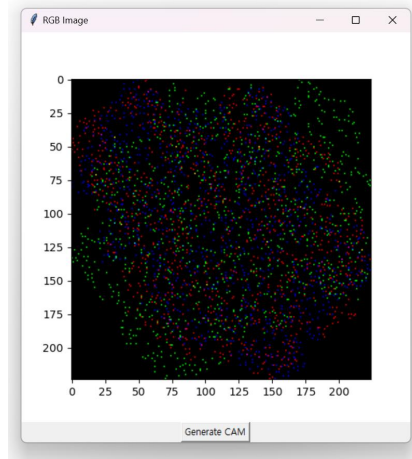
- 1 잠재적인 관계를 파악하기 어려운 표 형식 데이터 기반
- 2 개별 오믹스 데이터 분석을 통한 결과 도출
- 3 복잡한 자연 과정을 설명하는 해석 가능성이 부족한 '블랙박스' 모델
- 4 데이터 수집 및 모델 선별 비용

01 제안이유

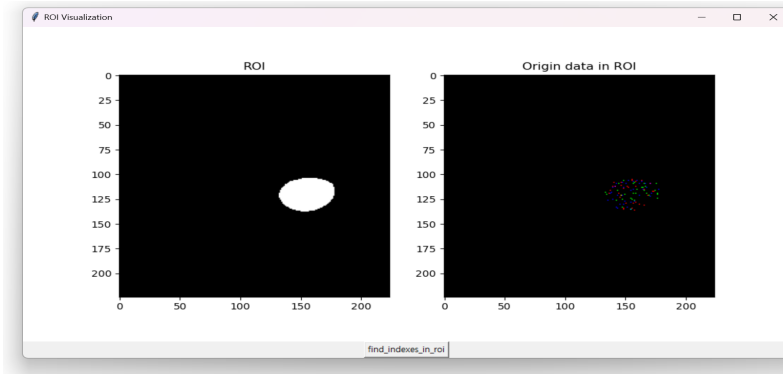
시뮬레이션을 활용한
멀티 오믹스 데이터 분석
프로그램

- 1 테이블 형식 오믹스 데이터를 이미지와 같은 표현으로 변환
=> CNN(:Convolutional Neural Network) 등 딥러닝 모델을
활용하여 잠재적 관계 후보 판별
- 2 전이 학습 활용으로 예측력 향상 뿐만 아니라 계산 시간
절약과 분석 정확도 향상 기대
- 3 CNN을 활용한 **멀티 오믹스** 데이터의 예측 모델링은
기존 기계 학습 기술의 한계 극복과 혁신적 성과 달성 기대
- 4 데이터 시뮬레이션을 통한 데이터 수집 및
모델 선별을 위한 비용 절감

시뮬레이션을 활용한 멀티 오믹스 데이터 분석 프로그램

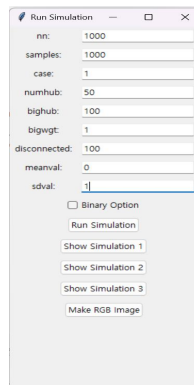


이미지 데이터 변환

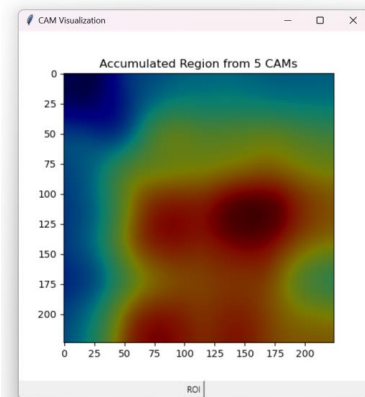


관심 변수 도출

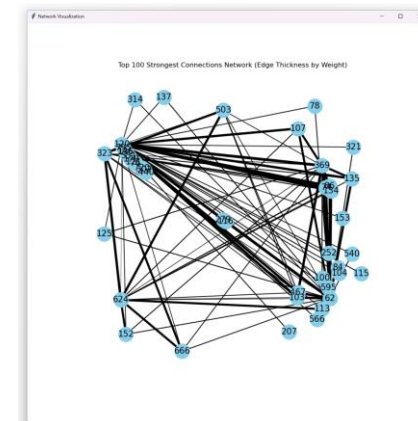
멀티-오믹스 데이터 시뮬레이션



전이학습

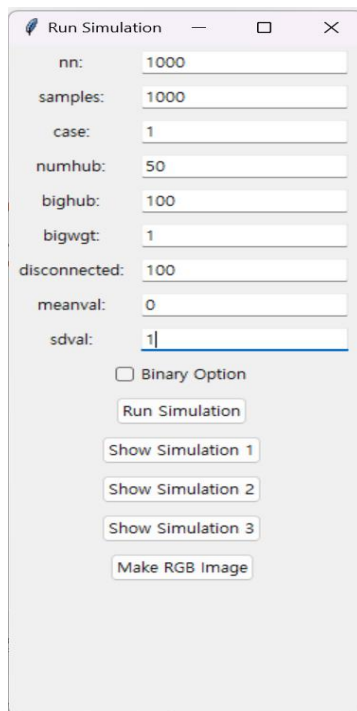


네트워크 분석



02 제안내용

데이터 시뮬레이션

A screenshot of a 'Run Simulation' dialog box. It contains several input fields with the following values: 'nn:' 1000, 'samples:' 1000, 'case:' 1, 'numhub:' 50, 'bighub:' 100, 'bigwgt:' 1, 'disconnected:' 100, 'meanval:' 0, and 'sdval:' 1. Below these fields is a checkbox labeled 'Binary Option' which is unchecked. At the bottom, there are four buttons: 'Run Simulation', 'Show Simulation 1', 'Show Simulation 2', and 'Show Simulation 3'. A 'Make RGB Image' button is also present at the very bottom.

[데이터 시뮬레이션 옵션 윈도우]

- 다양한 분포와 상황을 가정한 데이터를 통한 모델의 일반화 성능 향상
- 실제 데이터 수집의 한계를 극복하여 충분한 양의 데이터 확보
- 특정 시나리오에 대한 모델 성능 사전 테스트로 모델 개발 과정 효율화

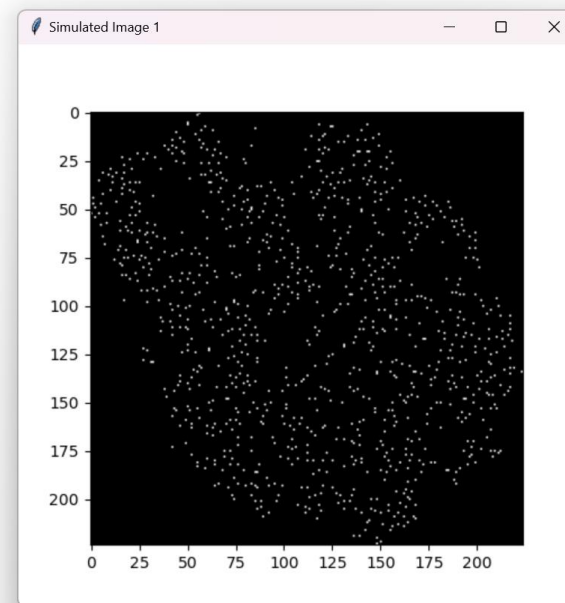
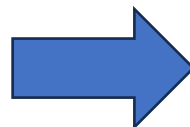
02 제안내용

이미지 데이터로의 표현

```
a["simudata"][1]
```

```
array([ -0.         ,  0.         ,  0.         ,  0.         ,  
        0.45036076, -0.         , -0.         , -0.         ,  
        0.         , -0.         , -0.         ,  0.         ,  
        0.         ,  0.         ,  0.         ,  0.1749668 ,  
       -0.         ,  0.         , -0.         ,  0.         ,  
        0.         , -0.59856086,  0.         ,  0.20804128,  
        0.         , -0.         ,  0.         , -0.87392217,  
       0.35861055, -0.39910944, -0.         ,  0.         ,  
        0.         , -0.51800457,  0.         , -0.         ,  
       -0.         ,  0.16057246, -0.         , -0.79830721,  
      -0.07304266, -0.         ,  0.         , -0.         ,  
       0.67910423, -0.         ,  0.         , -0.         ,  
       -0.         ,  0.         ,  0.         , -0.         ,  
      69.6313113 ,  0.         ,  0.         ,  0.         ,  
        0.         ,  0.         , -0.         ,  0.         ,  
      20.3081136 , -0.         , -0.         ,  0.         ,
```

[시뮬레이션으로 생성된 표 형식 데이터]

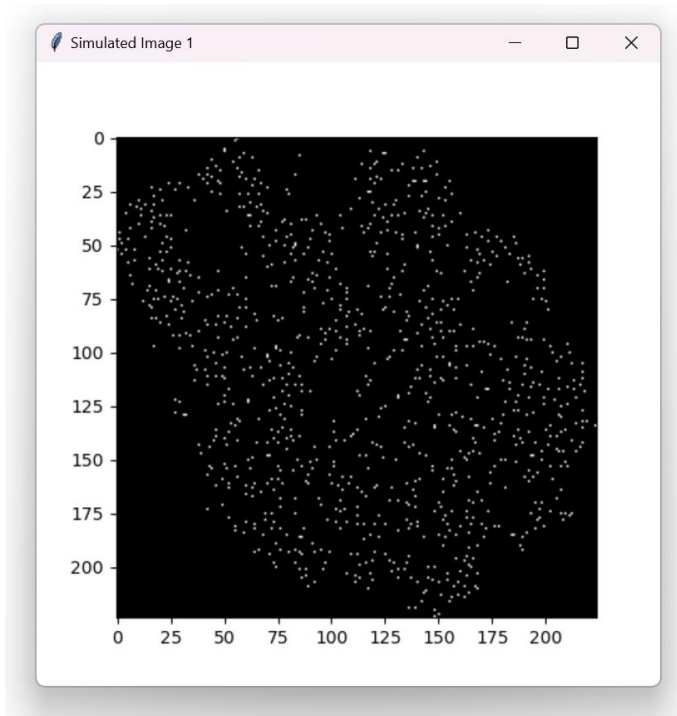


[이미지로 표현된 시뮬레이션 데이터]

대부분 변수가 서로 독립적으로 표시되는 " 표 형식 " 데이터를 이미지 데이터로 변환 후
이미지 분석 기법을 적합하게 적용하여 오믹스 데이터 간의 잠재적인 관계를 보다 잘 파악할 수 있음

02 제안내용

이미지 데이터로의 표현



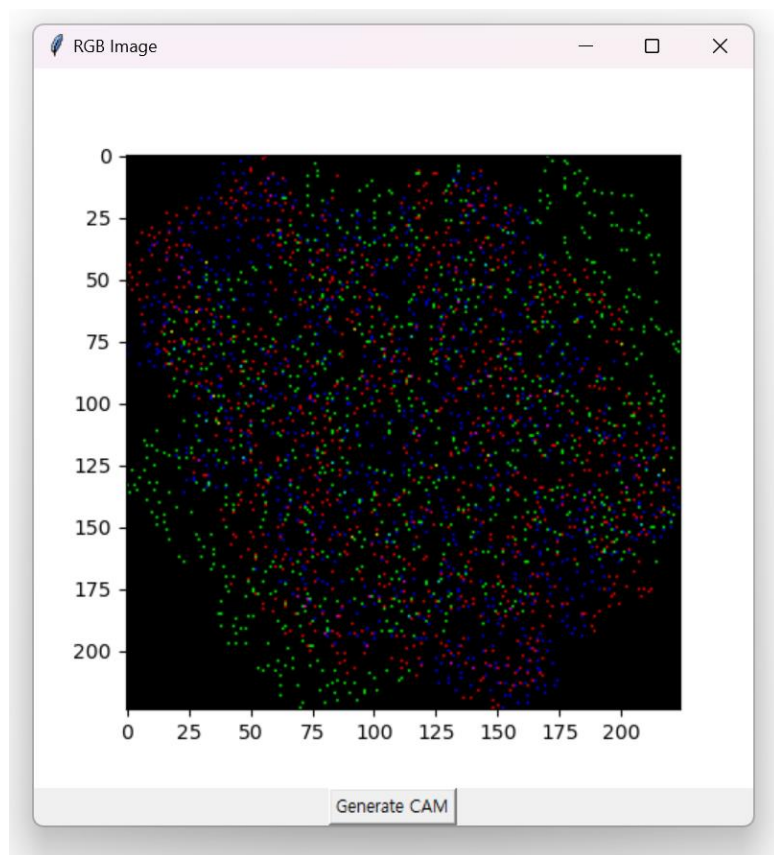
[이미지로 표현된 시뮬레이션 데이터]

이미지 데이터로 변환하여 유사한 특성을 공유하는 요소를 근접한 이웃 특성으로 배열하고, 근접하지 않은 요소는 기존 상태를 유지하면서, CNN이 다각적이고 효율적으로 적용될 수 있는 풍부한 환경을 생성

=> 유사한 변수를 가까이 모아 하나의 그룹으로 처리하여 오믹스 데이터의 복잡한 구조와 패턴을 반영

02 제안내용

이미지 데이터로의 표현




[이미지로 표현된 멀티오믹스 데이터]

오믹스 데이터는 유전체, 전사체, 단백질체, 대사체 등 개별 생물학적 층위를 다루지만, 생물학적 시스템은 이러한 층위들이 복잡하게 상호작용하며 작동

시뮬레이션으로 얻은 오믹스 데이터들의 정보를 통합하여 표현한 하나의 이미지를 딥러닝을 통해, 다양한 오믹스 데이터들 간의 상호 작용을 포착하여 생물 구조적 이해를 촉진하고 분석을 위한 보다 풍부한 내용을 제공

02 제안내용

이미지데이터 분석을 위한 딥러닝



딥러닝을 이용한
멀티오믹스데이터의
관계추론

- 1 수많은 상호 작용을 식별하고 비선형 효과를 모델링
=> 과적합 위험을 완화할 수 있는 효과적인 정규화 기능 제공
- 2 오믹스 데이터의 다양한 유형의 구조화된 정보 수용
- 3 이질적인 데이터를 통합적으로 분석

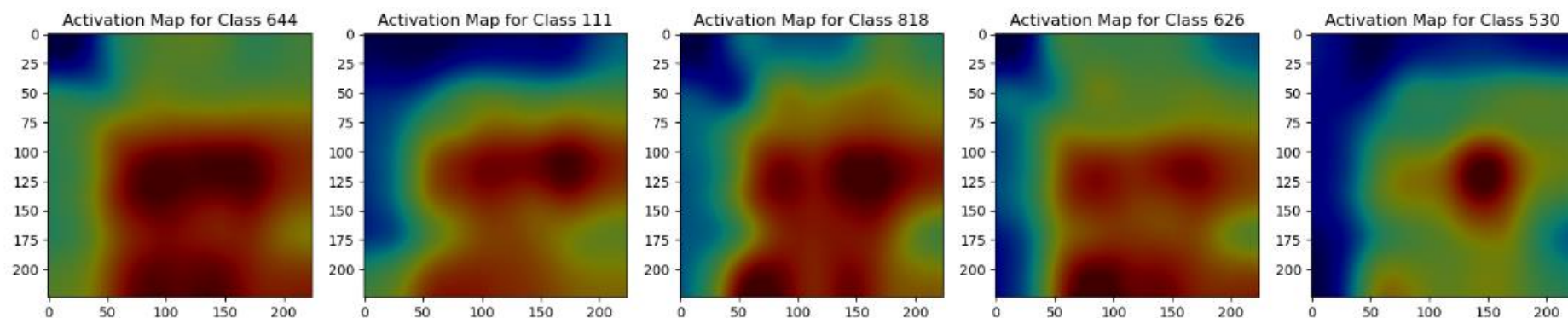
전이학습

- 딥러닝 모델 구축에는 많은 자원이 필요
- 전이학습은 대규모 데이터 세트의 지식을 재사용하여 소규모 데이터에서도 성능 향상 가능
- 사전 훈련된 모델을 새로운 작업에 맞게 특성화하면서 데이터 수집 필요성을 줄이면서도 성능 향상

02 제안내용

이미지데이터 분석을 위한 딥러닝

관심영역을 보여주는 CAM



[각 클래스 별로 예측에 영향을 미치는 이미지 내 영역 강조 표시]

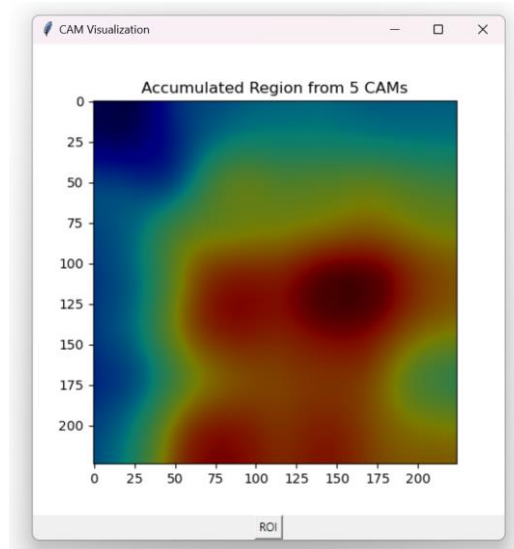
CAM(:Class Activation Map)을 통해 각 클래스(예: 질병 유무, 치료 반응 등)에 대해 모델이 예측할 때 큰 영향을 미치는 이미지의 특정 영역을 강조 표시

=> 중요한 생물학적 층위나 변수들을 식별할 수 있어 모델의 예측 과정을 해석할 수 있게 도와주기 때문에 모델의 성능을 개선하거나 모델의 신뢰성을 높일 수 있음

02 제안내용

이미지데이터 분석을 위한 딥러닝

관심영역을 보여주는 CAM



[모든 클래스에 공통적으로 영향을 미치는 영역 강조하여 표시]

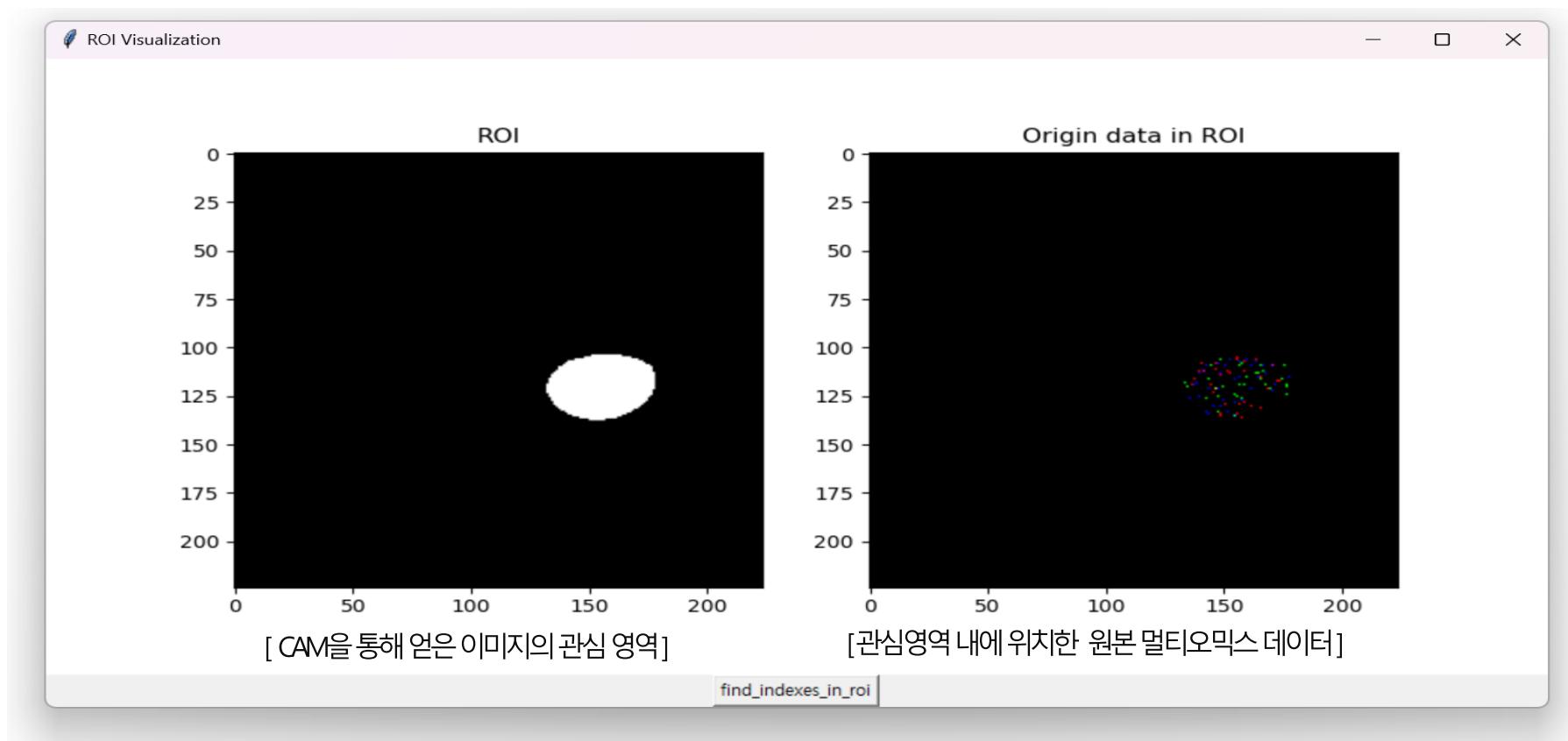
모든 클래스에서 공통적으로 활성화되는 멀티오믹스 데이터의 영역을 살펴보는 평균 CAM을 통해 복잡한 생물학적 네트워크와 상호작용을 통합적으로 이해할 수 있음.

클래스별 CAM과 평균 CAM결과를 원본 멀티오믹스 데이터와 연계하여 분석하면, 질병 발생, 치료 반응 등의 생물학적 메커니즘을 보다 심도 있게 이해할 수 있음.

02 제안내용

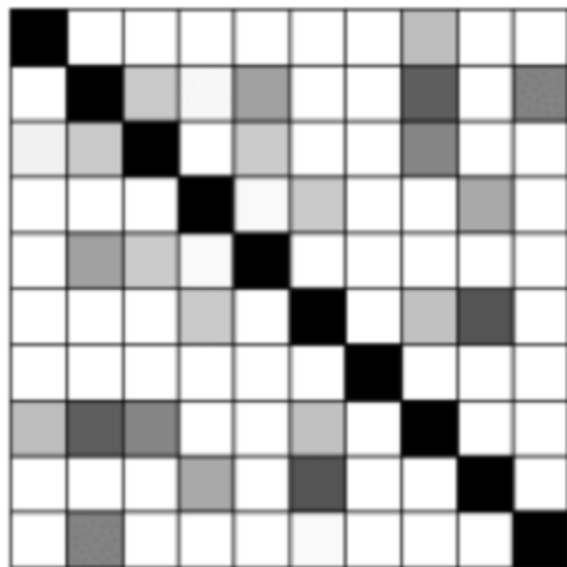
이미지데이터 분석을 위한 딥러닝

관심영역을 보여주는 CAM

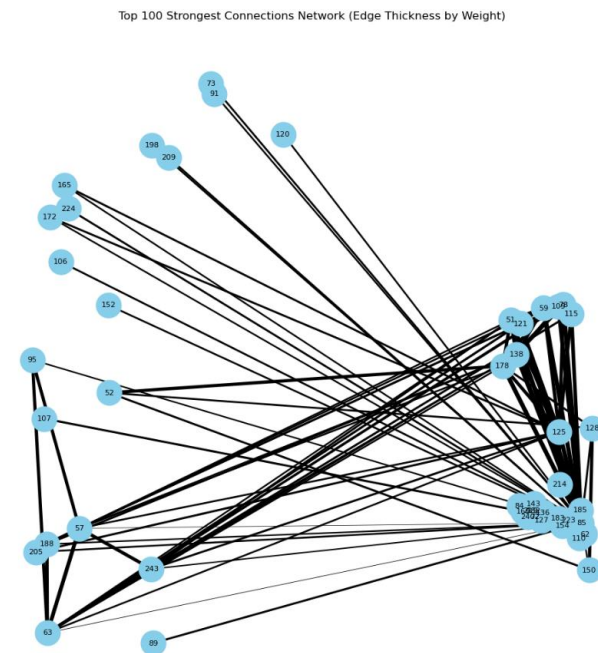


02 제안내용

네트워크 분석을 통한 인사이트 도출



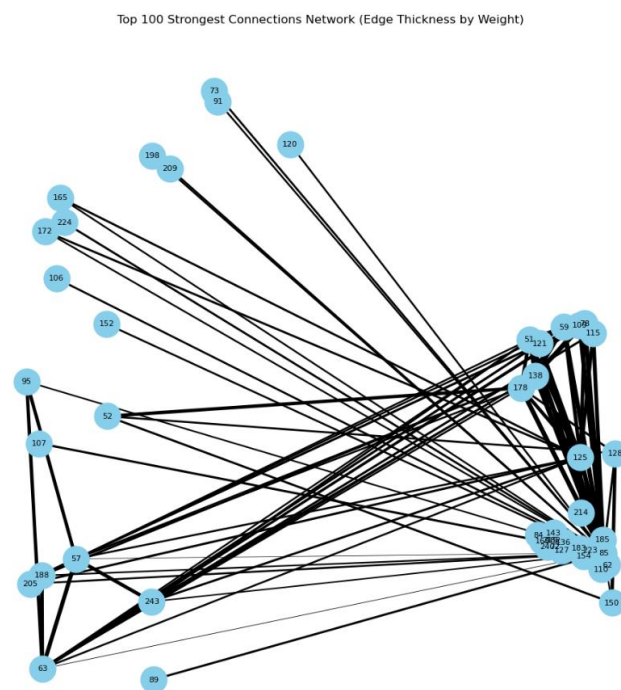
[관심영역 변수들의 공분산 행렬]



관심영역 변수들 간 주요 네트워크

02 제안내용

네트워크 분석을 통한 인사이트 도출



[관심영역 변수들 간 주요 네트워크]

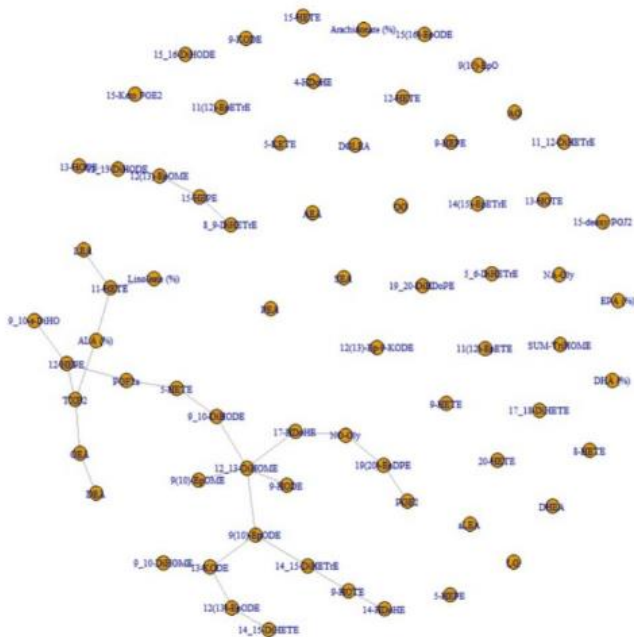
관심 영역의 변수들 간 상관관계, 네트워크 구조 등을 분석하여 생물학적 시스템 내 변수 간 상호작용 이해

=> 질병 발생, 치료 반응 등의 복잡한 생물학적 과정을 보다 구체적으로 설명가능

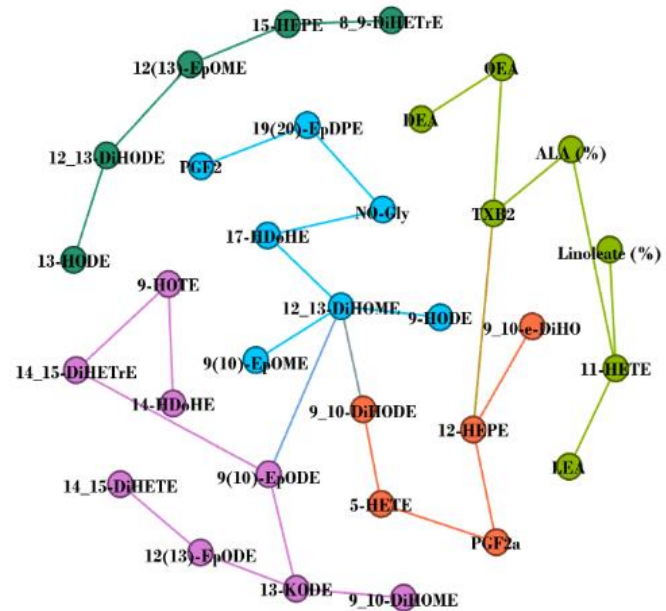
딥러닝을 활용한 모델의 생물학적 해석력 향상

02 제안내용

네트워크 분석을 통한 인사이트 도출



[네트워크 내 주요 관계 파악]



[주요 관계 내 클러스터링]

02 제안내용

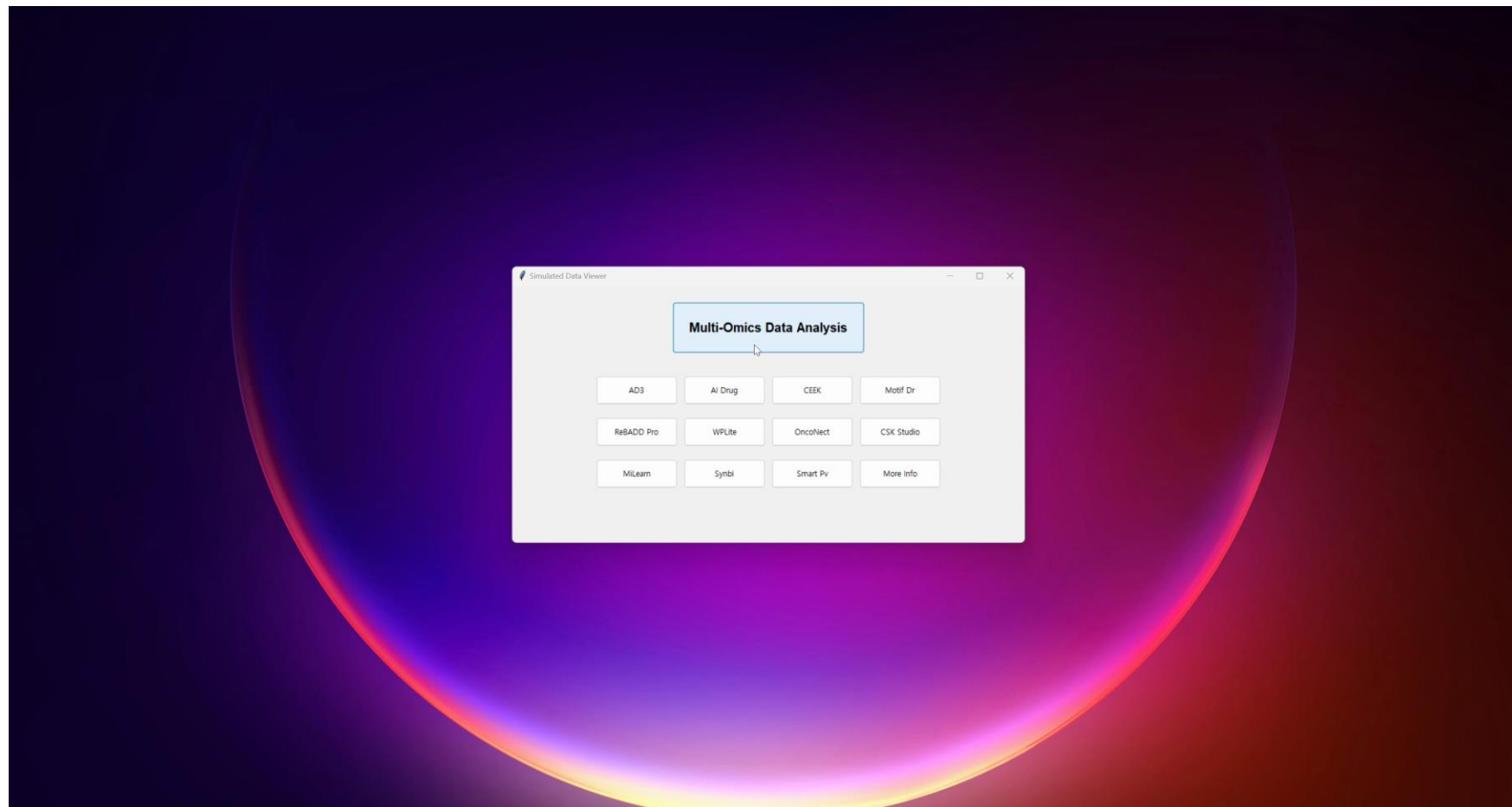
네트워크 분석을 통한 인사이트 도출

Metabolites	Averaged Coefficients
TXB2	0.445
15-deoxy PGJ2	18.12
13-HOTE	0.336
9-HOTE	3.016
15-HEPE	283.357
12-HEPE	4.409
17-HDoHE	9.801
11(12)-EpETE	10.343
13-KODE	0.063
5-KETE	1.505
EPA	430.641
SEA	0.004
OEA	0.477
aLEA	78.785
DHEA	249.065

Metabolites	Averaged Coefficients
PGE2	-4.13
12_13-DiHOME	-12.459
9-HODE	-0.025
12(13)-EpOME	-22.397
9(10)-EpOME	-1.527
11(12)-EpETrE	-43.01
9-KODE	-0.035
Linoleate	-76.875
Arachidonate	-246.647
PEA	-0.015
DGLEA	-2065.24
AEA	-52.77
DEA	-121.534
NA-Gly	-27.862





[발병 가능성과 양적 관계를 가지는 변수들] [발병 가능성과 음적 관계를 가지는 변수들]

02 제안내용



03 활용 방안 및 기대 효과

기존 모델의 한계점

-  1 잠재적인 관계를 파악하기 어려운 표 형식 데이터 기반
-  2 개별 오믹스 데이터 분석을 통한 결과 도출
-  3 복잡한 자연 과정을 설명하는 해석 가능성이 부족한 '블랙박스' 모델
-  4 데이터 수집 및 모델 선별 비용

03 활용 방안 및 기대 효과

생물학적 시스템 이해 증진

- 유전체, 전사체, 단백질체, 대사체 등 다양한 오믹스 데이터를 통합적으로 분석하여 생물학적 시스템의 복잡한 상호작용 이해

- 시뮬레이션을 통해 실험적으로 검증하기 어려운 가설에 대한 테스트 및 검증 가능

Motif Dr 및 WPLite과 같은 신약개발 플랫폼 사용 전, 특정 유전자나 단백질의 구조와 기능을 이해하는 데 도움을 줄 수 있음

생물학적 시스템 최적화

- 시뮬레이션을 통해 생물학적 시스템의 특정 매개변수를 조절하여 사용자에게 맞춤형 경험 제공

- 다양한 시나리오에 대한 가정과 실험은 유전자 조작, 대사 경로 설계, 약물 개발 등에 활용 가능

03 활용 방안 및 기대 효과

개인 맞춤형 의료 및 건강 관리

- 개인의 오믹스 데이터를 분석하고 시뮬레이션하여 질병 예방, 치료, 관리 등에 활용할 수 있음
- 개인의 유전적 특성과 생활 습관을 고려할 수 있기 때문에 다양한 플랫폼의 분석 결과를 개인에 맞추어 살펴볼 수 있도록 도움을 줄 수 있음
- AD3 및 AI Drug 플랫폼 등에서 신약 후보 물질을 더욱 정밀하게 발굴하도록 도움을 줄 수 있음.

교육 프로그램의 역할

- 의약학 교육에서 복잡한 생물학적 시스템을 시각화하고 이해를 돕는 도구로 활용 가능
- 의약학 연구 분야에서 새로운 가설 검증, 실험 설계, 데이터 분석 등에 활용 가능
- 쉬운 동작 방법과 다양한 시나리오를 가정하는 경험을 제공하기 때문에 KAIDD 내 새로운 교육 프로그램의 역할 기대

04 참고문헌

Alok Sharma, Artem Lysenko, Shangru Jia, Keith A. Boroevich & Tatsuhiko Tsunod, Advances in AI and machine learning for predictive medicine, Journal of Human Genetics, 2024

Alok Sharma, Edwin Vans, Daichi Shigemizu, Keith A. Boroevich & Tatsuhiko Tsunoda, DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture, Scientific Reports , 2019

Anastasia Zompola, Aigli Korfiati, Konstantinos Theofilatos, and Seferina Mavroudid, Omics-CNN: A comprehensive pipeline for predictive analytics in quantitative omics using one-dimensional convolutional neural networks, Heliyon, 2023