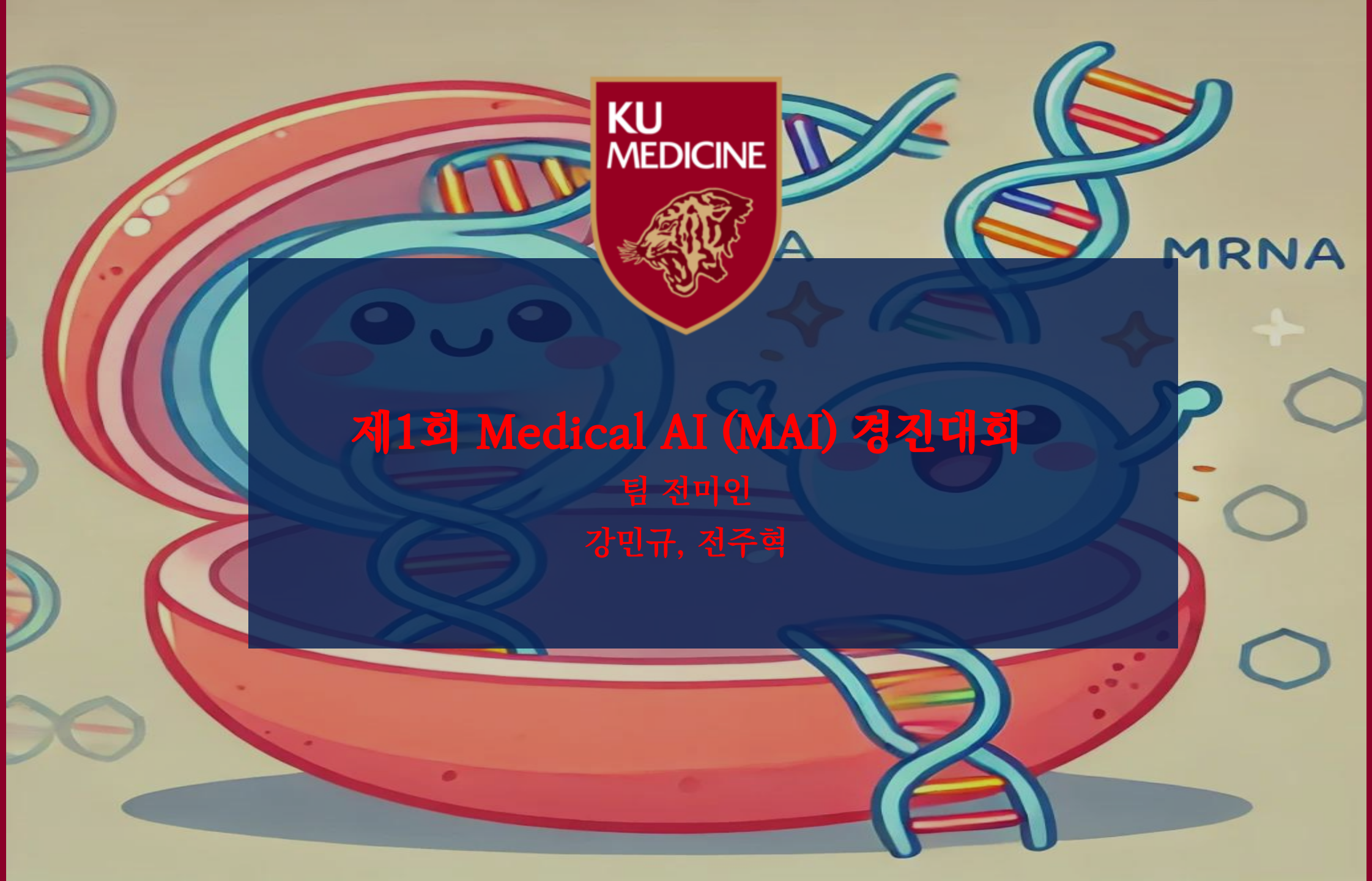


제1회 Medical AI (MAI) 경진대회

팀 전미인

강민규, 전주혁

목차

01
INTRO

02
데이터 분석
1. 이미지 데이터
2. 유전체 데이터

03
상관관계 고려 모델 학습 전략
1. 모델 접근
2. 유전자(데이터) 접근
3. 유사도 기반 접근

04
학습 과정 최적화
1. 이미지 증강
2. 학습 프로세스
3. 학습 결과

05
발전 가능성

INTRO

INTRO

대회 주제

H&E 염색된 조직 이미지로부터 유전자 발현 예측

제공 데이터

H&E 염색된 조직 이미지 샘플 및 각 유전자의 발현 정보

대회 목적

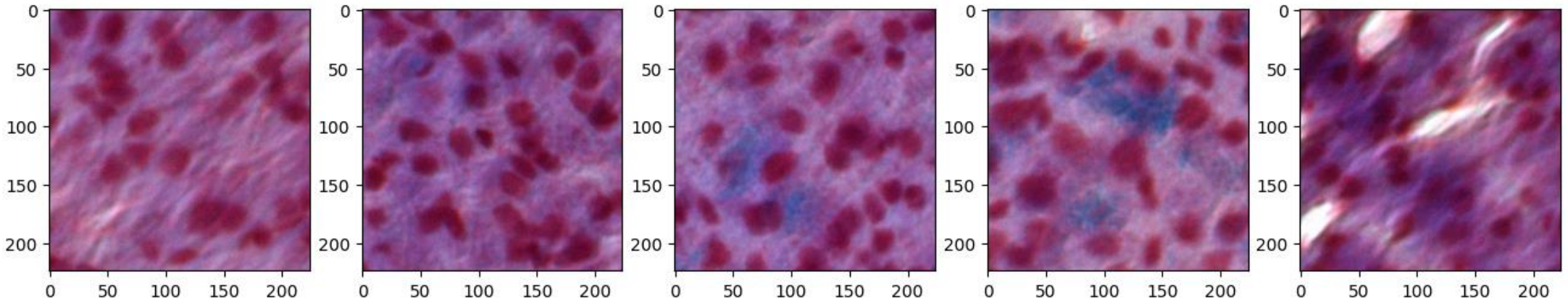
의료 데이터를 활용한 AI 기술로 실제 문제 해결에 어떻게 기여할 수 있는지 탐구하는 것이 목적

평가

의료 모델 성능 개선 및 문제 해결 관련 솔루션 평가

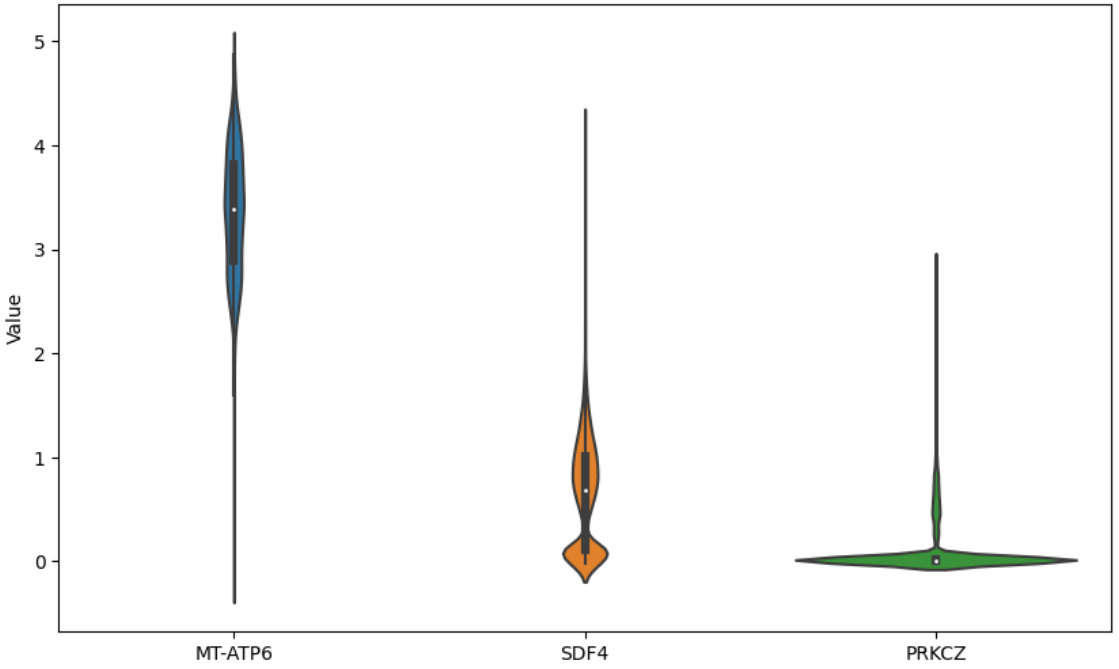
데이터 분석

데이터 분석 : 이미지 데이터



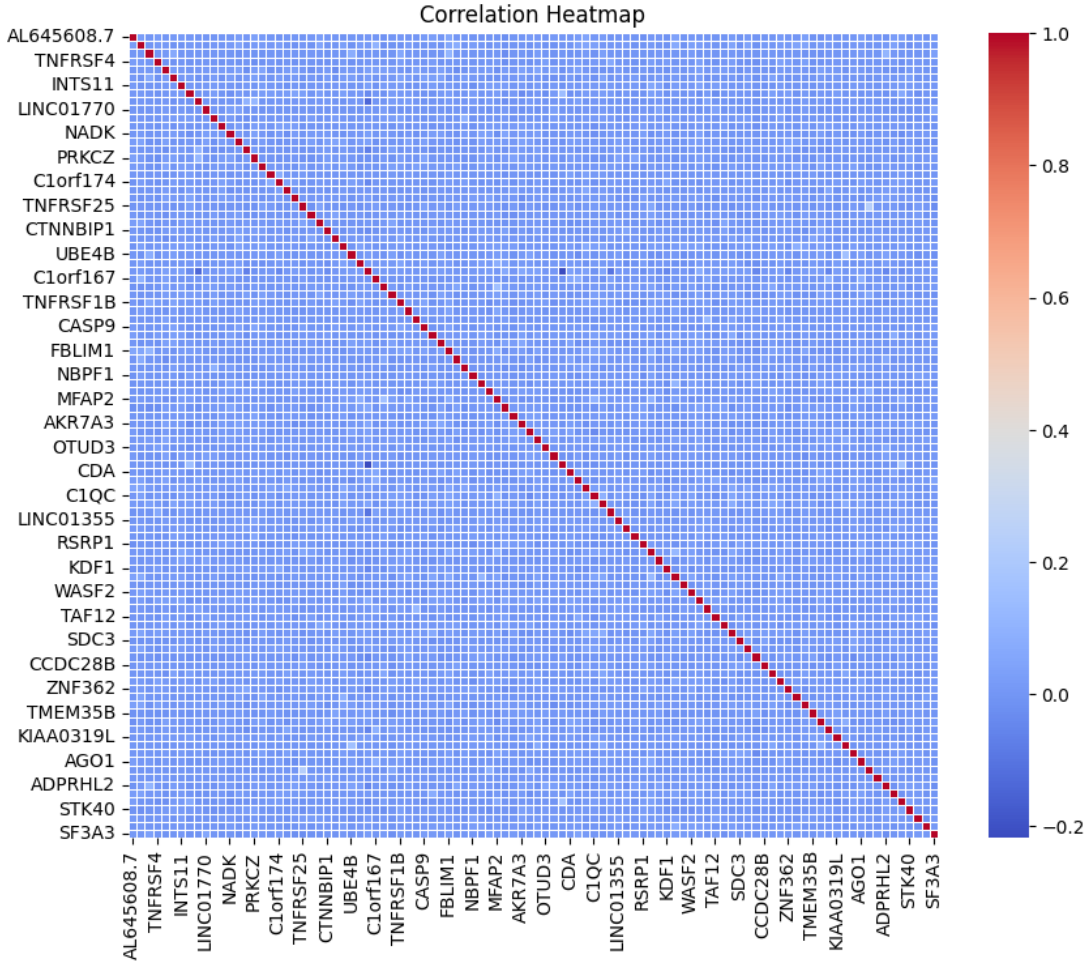
- 모든 이미지는 224x224의 해상도를 가짐
- H&E 염색 이미지는 세포핵은 진한 보라색, 세포질과 세포외 기질을 분홍색/붉은색으로 염색한 이미지로 병리적 변화 확인에 용이
  - 잘못된 색상, 대비 조정은 중요한 병리적 특성을 왜곡할 수 있음
  - **염색 특성을 변형하지 않는 방식으로 증강** 기법을 신중히 선택해야 함

데이터 분석 : 유전자 데이터



- 예측할 타겟은 3467개의 유전자의 발현 정보
- 각각의 유전자들이 **다양한 스케일과 분포**를 가지고 있음
- 모든 유전자들을 동시에 예측하는 것은 어려움
  - 모델이 특정 타겟에 맞춰 과적합되거나
    일부 타겟의 예측 성능이 떨어지게 됨
  - **유사한 스케일과 상관관계를 가지는 유전자들의 학습 및 예측이 필요**

데이터 분석 : 유전자 데이터



- 각각의 유전자들은 높은 상관관계를 가지지 않는 것들이 다수 존재
- 발현 패턴이 유전자마다 매우 상이하게 나타나는 것을 의미
- 모든 유전자를 일괄 예측 or 무조건 분할 예측은 모델이 잡음을 학습할 가능성 증가

➢ **도메인 or 데이터를 고려한 전략이 필수**

상관관계 고려 모델 학습 전략

상관관계 고려한 학습 전략

Hypothesis

- 유전자 발현에는 **특정 유전자들끼리의 상관관계가 존재할 것**이라는 가설 설정

A General Framework for Weighted Gene Co-Expression Network Analysis

Bin Zhang and Steve Horvath

Abstract

Gene co-expression networks are increasingly used to explore the system-level functionality of genes. The network construction is conceptually straightforward: nodes represent genes and nodes are connected if the corresponding genes are significantly co-expressed across appropriately chosen tissue samples. In reality, it is tricky to define the connections between the nodes in such networks. An important question is whether it is biologically meaningful to encode gene co-expression using binary information (connected=1, unconnected=0). We describe a general framework for `soft' thresholding that assigns a connection weight to each gene pair. This leads us to define the notion of a weighted gene co-expression network. For soft thresholding we propose several adjacency functions that convert the co-expression measure to a connection weight. For determining the parameters of the adjacency function, we propose a biologically motivated criterion (referred to as the scale-free topology criterion).

We generalize the following important network concepts to the case of weighted networks. First, we introduce several node connectivity measures and provide empirical evidence that they can be important for predicting the biological significance of a gene. Second, we provide theoretical and empirical evidence that the `weighted' topological overlap measure (used to define gene modules) leads to more cohesive modules than its `unweighted' counterpart. Third, we generalize the clustering coefficient to weighted networks. Unlike the unweighted clustering coefficient, the weighted clustering coefficient is not inversely related to the connectivity. We provide a model that shows how an inverse relationship between clustering coefficient and connectivity arises from hard thresholding.

We apply our methods to simulated data, a cancer microarray data set, and a yeast microarray data set.

Macroscopic Biclustering of Gene Expression Data

Jaegyoon Ahn† · Youngmi Yoon†† · Sanghyun Park†††

ABSTRACT

A microarray dataset is 2-dimensional dataset with a set of genes and a set of conditions. A bicluster is a subset of genes that show similar behavior within a subset of conditions. Genes that show similar behavior can be considered to have same cellular functions. Thus, biclustering algorithm is a useful tool to uncover groups of genes involved in the same cellular process and groups of conditions which take place in this process. We are proposing a polynomial time algorithm to identify functionally highly correlated biclusters. Our algorithm identifies 1) the gene set that has hidden patterns even if the level of noise is high, 2) the multiple, possibly overlapped, and diverse gene sets, 3) gene sets whose functional association is strongly high, and 4) deterministic biclustering results. We validated the level of functional association of our method, and compared with current methods using GO.

위 논문들에서 가설에 기반이 되는 유전자 발현 간 상관관계를 언급

→ 예측한 유전자 발현값들이 실제 유전자 발현값들의 상관관계를 잘 반영하는 모델 개발 필요

상관관계 고려한 학습 전략

예측한 유전자 발현값들이 실제 유전체 발현값들의

상관관계를 잘 반영하는 모델 개발 전략

Strategy 1

모델 접근

Strategy 2

유전자(데이터) 접근

Strategy 3

데이터 유사도 기반 접근

상관관계 고려 모델 학습 전략
- 모델 접근 -

상관관계 고려 : 모델 접근

모델의 학습 과정에서 실제 유전자 발현값들의 상관관계를 잘 반영하도록 손실 함수 개발 전략

Problem **MSE Loss 함수의 한계**

개별 유전자 발현값의 예측 정확도를 높이는 데 초점을 맞추는 손실 함수

한계점

- **유전자 간 상관관계 미반영**

  : 유전자들 간의 상호 관계나 패턴을 고려하지 않기 때문에,

  예측된 발현값들이 실제 발현값들의 상관관계를 반영하지 못함

- **복잡한 생물학적 정보 손실**

  : 유전자 데이터에서 중요한 유전자 간의 상호 작용이나 공동 발현 패턴을 학습하지 못하여,

  모델의 생물학적 타당성이 떨어짐

상관관계 고려 : 모델 접근

모델의 학습 과정에서 실제 유전자 발현값들의 상관관계를 잘 반영하도록 손실 함수 개발 전략

Upgrade V1 **피어슨 상관관계 기반 손실 함수 도입**

MSE Loss의 한계를 극복하기 위해 예측값과 실제값 사이의 상관관계를 고려한 손실 함수

도입 효과

➢ **상관관계 최대화**

: 모델이 유전자 간의 변동 패턴을 학습하도록 유도

➢ **스케일 독립적**

: 상관계수는 데이터의 스케일(크기)에 영향을 받지 않으므로,
유전자 발현값의 상대적인 변화를 학습하는 데 효과적

상관관계 고려 : 모델 접근

모델의 학습 과정에서 실제 유전자 발현값들의 상관관계를 잘 반영하도록 손실 함수 개발 전략

Upgrade V1 피어슨 상관관계 기반 손실 함수의 한계

한계점

- **절대적 크기 반영 불가(스케일 정보 손실)**

  → 상관계수는 두 변수 간의 관계 방향과 정도를 나타내지만, 값의 크기나 분산에 대한 정보를 제공X

  (유전자 발현값의 절대적인 크기나 변동성을 반영하지 못함)

- **복잡한 상호 관계 학습의 어려움**

  → 유전체 데이터에서의 유전자 간의 복잡한 공변성 구조를 완전히 반영하기 어려움

상관관계 고려 : 모델 접근

모델의 학습 과정에서 실제 유전자 발현값들의 상관관계를 잘 반영하도록 손실 함수 개발 전략

Upgrade V2 **공분산 행렬 손실 함수 도입**

도입 효과

상관관계 기반 손실 함수의 한계를 보완하기 위해 공분산 행렬을 활용한 손실 함수를 도입

- **스케일과 변동성 반영**

  → 공분산은 데이터의 스케일과 분산 정보를 포함하므로,
  모델이 유전자 발현값의 절대적인 크기와 변동성을 학습할 수 있음

- **복잡한 상호 관계 학습**

  → 유전자들 간의 공변성 구조 전체를 고려하여, 복잡한 상호 작용 패턴을 모델링 가능

상관관계 고려 : 모델 접근

모델의 학습 과정에서 실제 유전자 발현값들의 상관관계를 잘 반영하도록 손실 함수 개발 전략

총 손실 함수 구성

$$\text{Total Loss} = \alpha \times \text{MSELoss} + \beta \times \text{Correlation Loss} + (1-\alpha-\beta) \times \text{Covariance Loss}$$

1. 피어슨 상관관계 기반 손실 함수를 도입하여 상관관계를 반영

2. 공분산 행렬 손실 함수를 추가하여 모델이 유전자 발현값의 절대적인 크기와 복잡한 공변성 구조를 학습

3. 각 손실 함수의 장점을 조합함으로써,
   모델이 예측 정확도와 유전자 간의 관계성을 모두 잘 반영

상관관계 고려 모델 학습 전략

- 유전자(데이터) 접근 -

상관관계 고려 : 유전자(데이터) 접근

유전자 기능적 유사성 또는 이름을 기반으로 유전자를 군집화하여 모델의 예측 정확도 향상

Hypothesis 1

특정 유전자는 그 기능에 따라 발현 패턴과 상호작용 네트워크에서 유사성을 나타낼 것

→ **기능적 군집화** (Functional Clustering) / Gene Ontology(GO)

Hypothesis 2

특정 접두사를 공유하는 유전자들은 기능과 상관없이,

공통된 유전체적 특성 반영 가능성 존재

→ **접두사 연관성 기반 군집화** (Prefix Correlation-based Clustering)

상관관계 고려 : 유전자(데이터) 접근

유전자 기능적 유사성 또는 이름을 기반으로 유전자를 군집화하여 모델의 예측 정확도 향상

1. 기능적 군집화(Functional Clustering)

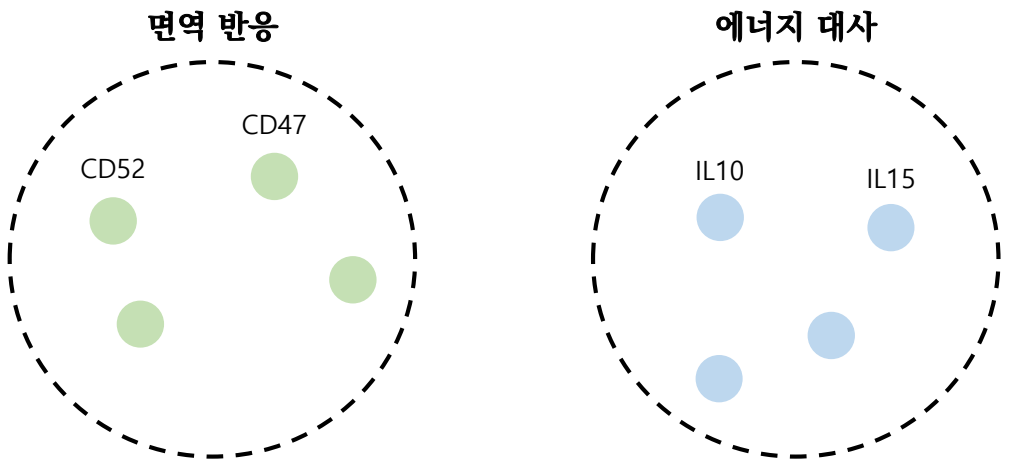
목적: 기능적 유사성을 기반으로 유전자를 군집화

방식:

- 특정 기능적 역할(예: 면역 반응, 대사 과정 등)을 수행하는 유전자들을 묶음
- **실제 기능적 유사성을 기준**으로 군집화

기대효과:

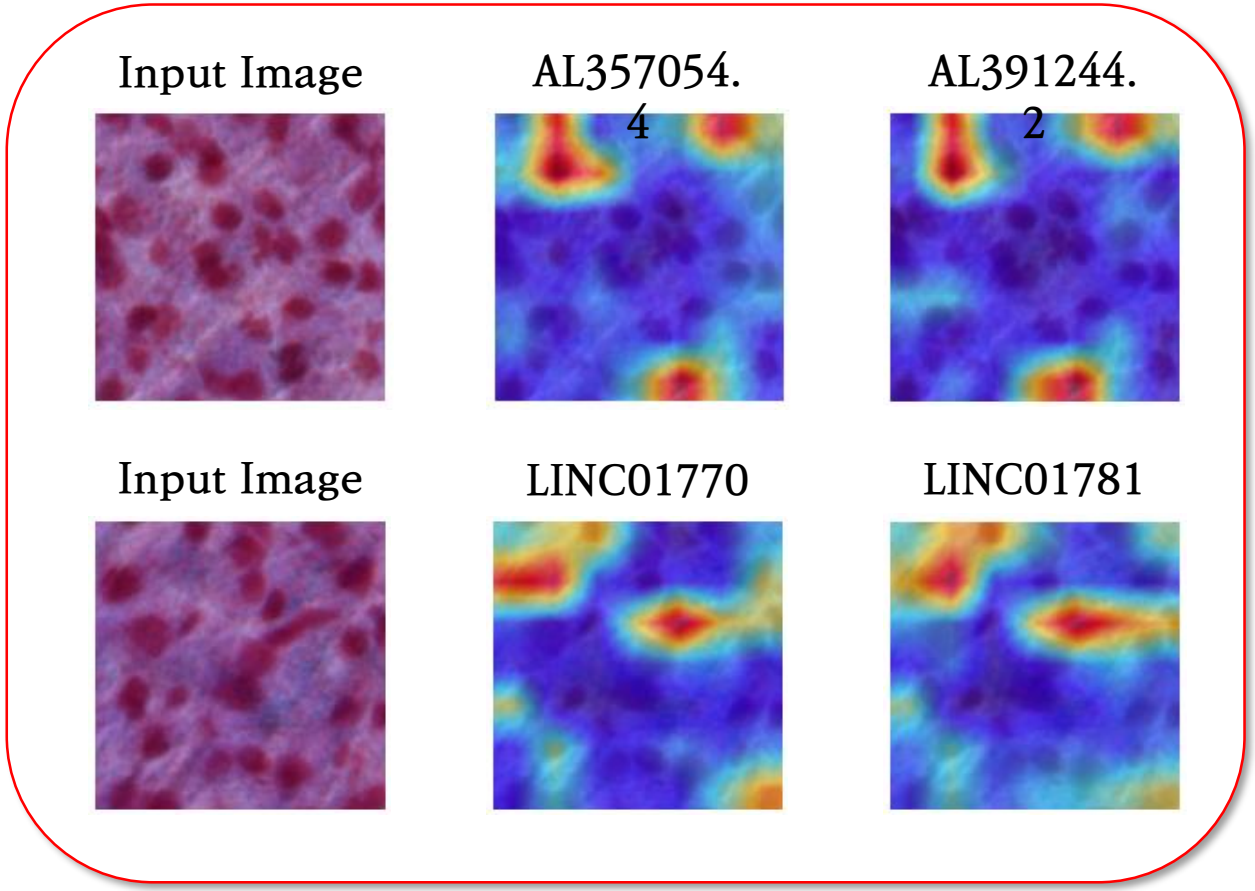
- 공동 발현 패턴을 효과적으로 학습 가능



상관관계 고려 : 유전자(데이터) 접근

유전자 기능적 유사성 또는 이름을 기반으로 유전자를 군집화하여 모델의 예측 정확도 향상

2. 접두사 연관성 기반 군집화 (Prefix Correlation-based Clustering)



분석: Grad-CAM 시각화 분석

결과:

- **동일 접두사를 가진 유전자들이 유사한 활성화 영역**이 나타남
- **해당 유전자들 간의 유전체적 또는 기능적 연관성이 있음을 의미**

상관관계 고려 : 유전자(데이터) 접근

유전자 기능적 유사성 또는 이름을 기반으로 유전자를 군집화하여 모델의 예측 정확도 향상

2. 접두사 연관성 기반 군집화 (Prefix Correlation-based Clustering)

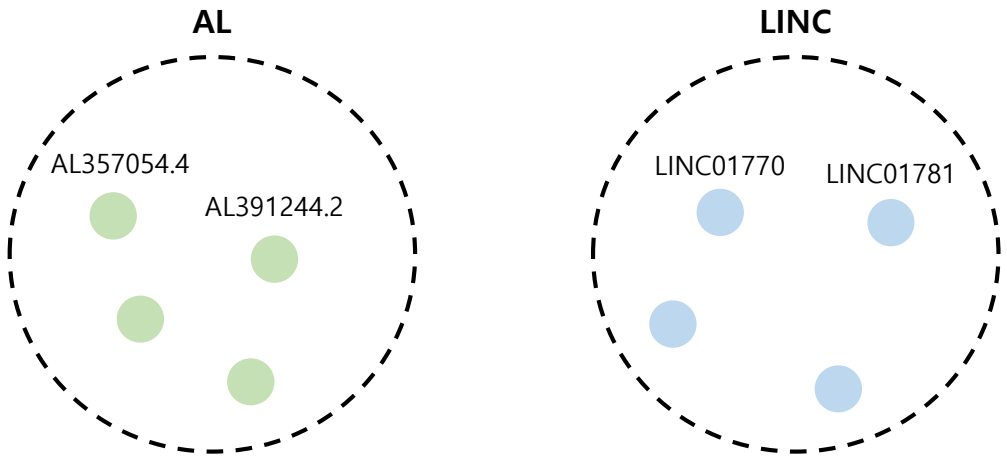
목적: 접두사가 암시하는 **유전체적 특성, 계통적 유사성, 구조적 연관성**을 바탕으로 군집화

방식:

- 특정 접두사를 가진 유전자들이 **공통된 유전체적 특성을 반영**
- 동일 접두사를 공유하는 유전자들을 하나의 그룹으로 묶음

기대효과:

- 유전자 간의 유전적 연관성을 반영

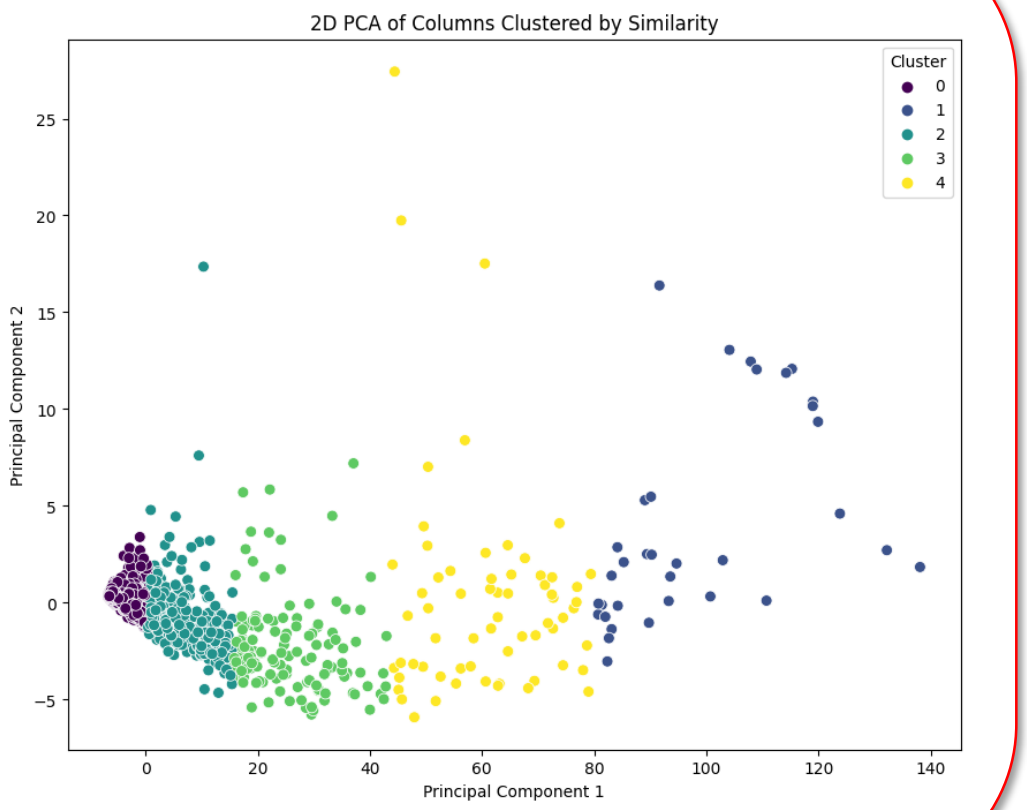


상관관계 고려 모델 학습 전략
- 유사도 기반 접근 -

상관관계 고려 : 유사도 기반 접근

타겟 데이터의 유사도를 기반으로 유전자를 군집화

3. 거리 기반 군집화 (Distance-based Clustering, Kmeans Clustering)



분석: 유전자들을 차원 축소 후 시각화

결과:

- 타겟 변수들 간 상관성이 존재하는 그룹이 있음을 발견
- 유전자 간 거리가 가까울 수록 유사한 상관관계를 가짐
- Cluster 알고리즘은 Kmeans가 가장 적절하게 그룹을 나누며,

  K=5로 했을 때 가장 의미 있는 패턴으로 군집화가 이루어짐

상관관계 고려 : 유사도 기반 접근

타겟 데이터의 유사도를 기반으로 유전자를 군집화

3. 데이터 거리 기반 군집화 (Distance-based Clustering, Kmeans Clustering)

목적: 데이터 포인트 간의 **유사성(거리)을 중심으로 군집화**

방식:

- 각 유전자의 **발현값을 바탕으로 유사한 발현 패턴을 가진 유전자들을 같은 그룹**으로 묶음
- 각 데이터를 일정한 수의 중심점 중 하나에 할당

기대효과:

- 발현 패턴 및 스케일이 유사한 유전자들이 같은 군집에 속할 가능성 증가
- 일부 유전자들에 대한 모델 과적합 방지

학습 과정 최적화

학습 과정 최적화 : 이미지 증강

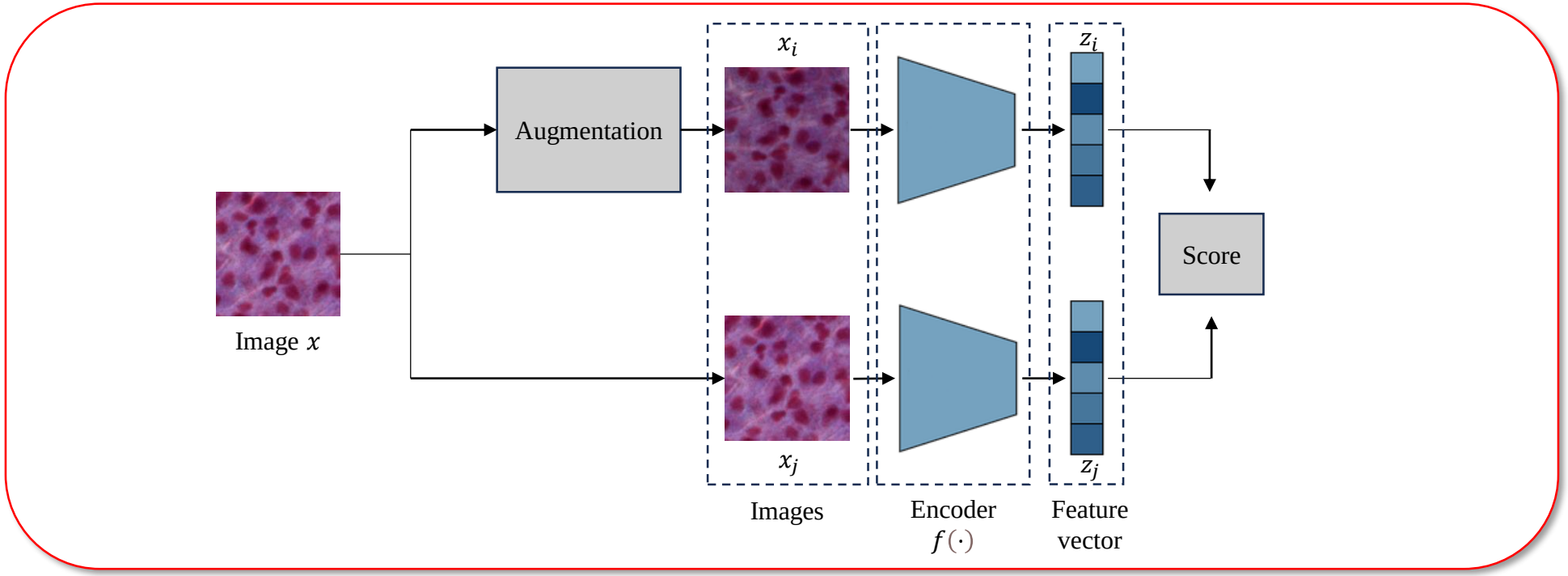
유전자의 특징을 변형시키지 않는 증강 선정

이미지 증강 기법 선정(DINO & LPIPS Score)

- 증강 전후 2개의 이미지에 대한 DINO & LPIPS Score 측정 후 특정 임계값보다 낮은 방법들로 선정

- DINO & LPIPS Score는 이미지 생성 논문에서 많이 쓰이는 Metric으로
  생성 이미지와 원본 이미지의 변형된 정도를 측정

  ➢ 염색의 변형 정도를 최소화하며 이미지를 증강시키는 방법을 탐색하는데 활용



학습 과정 최적화 : 이미지 증강

유전자의 특징을 변형시키지 않는 증강 선정

이미지 증강 기법 선정(DINO & LPIPS Score)

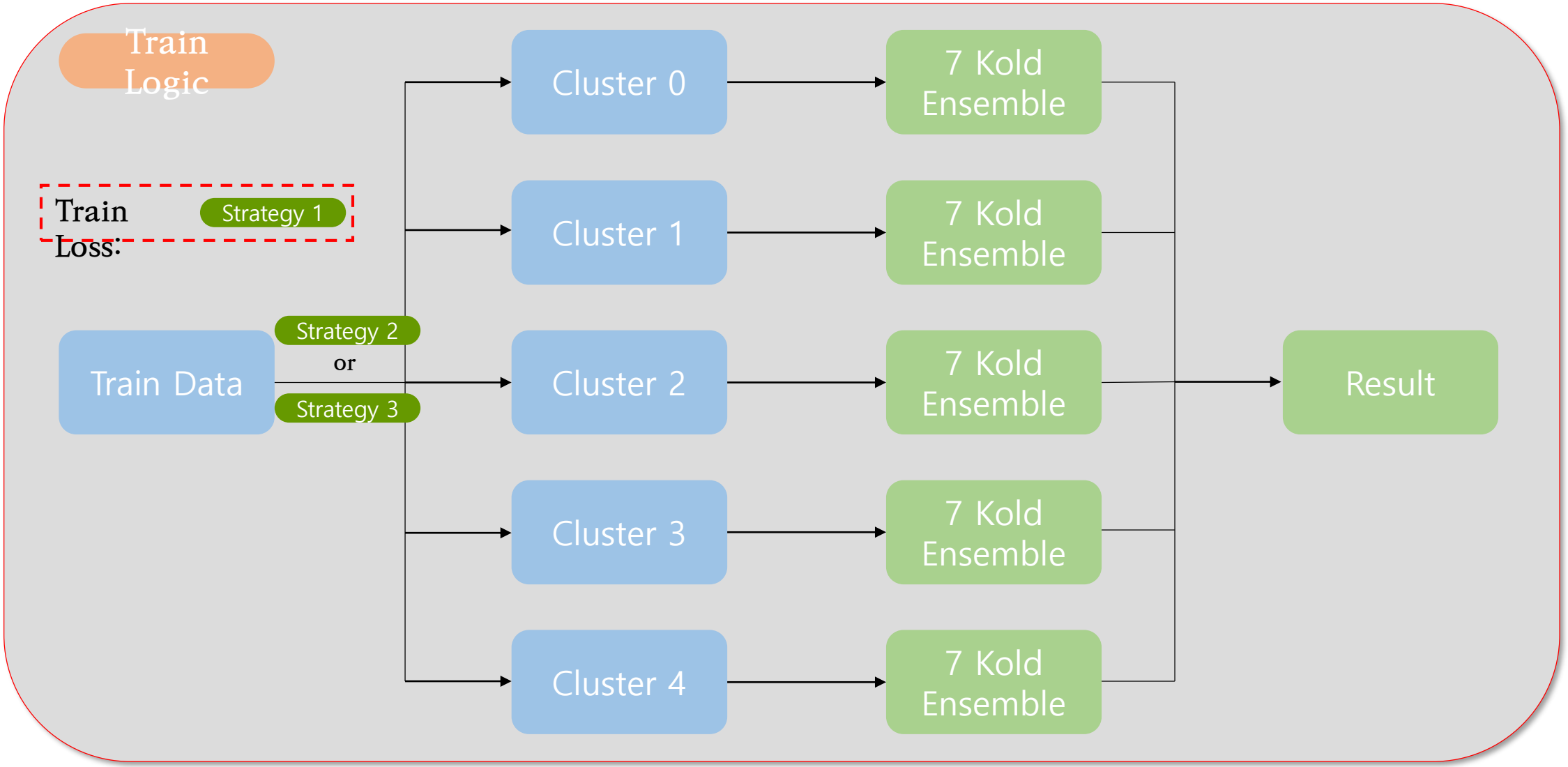
- 증강 전후 2개의 이미지에 대한 DINO & LPIPS Score 측정 후 특정 임계값보다 낮은 방법들로 선정

HorizontalFlip
VerticalFlip
ShiftScaleRotate
RandomBrightnessContrast
ColorJitter
Equalize
HueSaturationValue
Mixup
…

Filtering →

HorizontalFlip
VerticalFlip
ShiftScaleRotate
RandomBrightnessContrast

학습 과정 최적화 : 학습 프로세스

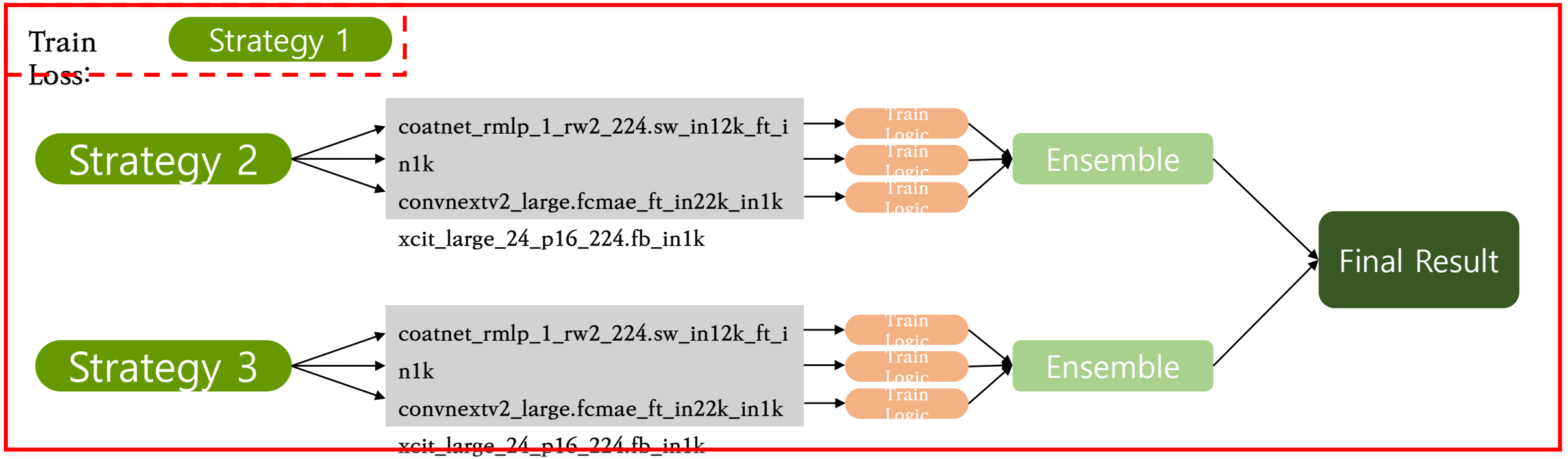


학습 과정 최적화 : 학습 프로세스

다양한 관점에서 Feature 추출 및 학습하기 위해 3가지 Backbone 모델 선정

❖ 선정 기준: Train_text_split=0.2로 Validation Score를 측정했을 때 상위 모델들

- coatnet_rmlp_1_rw2_224.sw_in12k_ft_in1k
- convnextv2_large.fcmae_ft_in22k_in1k
- xcit_large_24_p16_224.fb_in1k



학습 과정 최적화 : 학습 결과

학습(실험) 결과

	Public Score
Coatnet	0.5201
Coatnet + Strategy (2,3)	0.5400
Coatnet + Strategy (1,2,3)	0.5417
3 models + Strategy (1,2,3)	0.5428

- Strategy 2,3 군집화 방법들을 적용했을 때 성능이 크게 증가
- Strategy 1과 3개의 모델 앙상블을 하여 일반화 성능 향상

발전 가능성

발전 가능성

손실 함수 및 유전체 군집화 개선 방안

1. 손실 함수 개선

- 현재 상태

  모델이 개별 유전자 발현값과 유전자 간의 관계성을 모두 학습

- 아이디어

  KEGG, Reactome, BioCarta 등 **생물학적 경로 정보 통합**:
  - 경로 기반 가중치 적용: **중요한 경로의 유전자에 가중치 부여**로 모델이 예측 집중
  - 경로 간 상호작용 반영: 상호작용 경로 유전자 간 **공동 손실 항목 추가**로 복잡한 상호작용 학습

2. 유전체 군집화 개선

- 현재 상태

  유전자의 기능적, 데이터 특성을 반영한 군집화

- 아이디어

  유전자들의 **GO 용어** 활용:
  - GO 용어 기반 군집화: 생물학적 과정, 분자 기능 등 **세포 구성 정보 반영**
  - 다차원 스케일링: **GO 용어 간 유사성을 저차원 공간에 매핑**해 정교한 군집화 수행


DNA
MRNA
감사합니다!