JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 2, pp. 557-568, Feb. 2026

Check for updates

기능적 유전자 집합 기반 Cross-Attention과 지도 대조 학습을 이용한 항암제 반응 예측

송 종 웅[1] · 유 선 용[2,3*]
[1]전남대학교 지능전자컴퓨터공학과 석사과정
[2]전남대학교 지능전자컴퓨터공학과 교수
[3]주식회사 마틸로에이아이 대표이사

Anti-Cancer Drug Response Prediction via Functional Gene Set Cross-Attention and Supervised Contrastive Learning

Jongung Song[1] · Sunyong Yoo[2,3*]

[1]Master's Course, Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, Korea
[2]Professor, Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, Korea
[3]CEO, R&D Center, MATILO AI Inc., Gwangju 61186, Korea

[요 약]

정밀의학과 신약 재창출에서 항암제-암 세포주의 약물 반응(IC50) 예측은 중요하다. 하지만 기존 모델은 약물과 세포를 독립적으로 인코딩한 뒤 단순 결합함으로써, 약물–세포 상호작용과 기능적 유전자 집합(gene set)의 생물학적 정보를 충분히 활용하지 못한다. 본 연구는 ChemBERTa 약물 임베딩과 GDSC RNA-seq 기반 949개 유전자 발현을 MSigDB Hallmark gene set으로 집약한 세포 표현 위에, 약물 조건부 gene set 게이팅과 gene set 수준 cross-attention을 적용하고, TARGET_PATHWAY 레이블을 이용한 supervised contrastive learning으로 약물 임베딩을 정규화하는 모델을 제안한다. GDSC 데이터셋에서 제안 모델은 ChemBERTa+MLP 베이스라인(PCC 0.911, RMSE 1.133) 대비 PCC 0.922, RMSE 1.069를 달성하였으며, gene set 기반 표현과 경로 지식, 약물 조건부 게이팅 및 cross-attention 통합이 약물 반응 예측의 정확도와 경로 수준 해석 가능성을 동시에 향상시킬 수 있음을 보였다.

[Abstract]

Predicting the drug response (IC50) of cancer cell lines is crucial in precision oncology and drug repurposing. However, many existing models independently encode drugs and cells and simply concatenate their representations, thus under-utilizing drug–cell interactions and the biological structure of functional gene sets. We propose a Gene Set Cross-Attention model, GS-ChemDRP, that applies drug-conditioned cross-attention to cell representations obtained by aggregating 949 Genomics of Drug Sensitivity in Cancer (GDSC) database gene expression features into MSigDB Hallmark gene sets, while regularizing ChemBERTa drug embeddings via supervised contrastive learning. Using the GDSC dataset, the proposed model achieved a test Pearson's correlation coefficient (PCC) of 0.922 and a root mean square error (RMSE) of 1.069, improving over the ChemBERTa + MLP baseline (PCC, 0.911; RMSE, 1.133). These data suggest that this integrated design can enhance both the accuracy and pathway-level interpretability of anti-cancer drug response predictions.

색인어 : 약물 반응 예측, Gene Set, Cross-Attention, 지도 대조 학습, GDSC

Keyword : Drug Response Prediction, Gene Set, Cross-Attention, Supervised Contrastive Learning, GDSC

http://dx.doi.org/10.9728/dcs.2026.27.2.557

Received 23 December 2025; **Revised** 30 January 2026
Accepted 02 February 2026

***Corresponding Author; Sunyong Yoo**

Tel: ※ 개인정보 표시제한
E-mail: syyoo@jnu.ac.kr

Ⅰ. 서 론

암 치료제에 대한 세포 수준의 약물 반응성을 정량적으로 예측하는 일은 정밀의학의 핵심 과제이다[1]. 동일한 항암제라도 종양의 분자생물학적 특성에 따라 치료 효과와 부작용 양상이 크게 달라지므로, 개별 환자의 종양 유전체 및 전사체 특성을 반영하여 어떤 약물이 어떤 세포에서 어느 정도의 민감도(IC50)를 보일지를 사전에 *in silico*로 예측하는 기술이 필수적이다[1]. 그러나 수백 개의 항암제와 수천 개의 암 세포주 조합에 대해 실험적으로 스크리닝을 수행하는 것은 막대한 비용과 시간이 소요되므로, 실험 설계상 고려 가능한 용량·노출 시간 조건에도 한계가 있다[2]-[4]. Genomics of Drug Sensitivity in Cancer (GDSC)와 같은 대규모 데이터베이스는 다양한 암 세포주에 대한 IC50과 분자 프로파일을 제공하며, 데이터 기반 약물 반응 예측 모델 개발을 위한 표준 벤치마크로 널리 활용되고 있다[2]-[4].

초기 약물 반응 예측 연구들은 주로 약물 분자 지문(fingerprint)과 세포주 유전자 발현 프로파일을 입력으로 하는 전통적 기계학습 모형(랜덤포레스트, 서포트 벡터 머신 등)에 기반하여 IC50을 회귀하는 접근을 사용하였다[5],[6]. 이후 딥러닝이 도입되면서, 다층 퍼셉트론(MLP)과 합성곱 신경망(CNN), 나아가 그래프 신경망(Graph Neural Network, GNN)을 이용해 약물을 분자 그래프로 표현하고 세포주 전사체와 결합하는 방법들이 제안되었고, 기존 방법 대비 예측 성능 향상을 보고하였다[7],[8]. 그럼에도 불구하고 많은 모델이 약물과 세포를 각각 독립적으로 인코딩한 뒤 단순히 표현을 연결(concatenation)하여 사용하는 구조를 취해, 약물-세포 상호작용의 조건 의존적 패턴을 충분히 포착하지 못한다. 세포 측에서는 수백에서 수천 개의 유전자를 평탄한 벡터로만 다루어, 기능적 유전자 집합(gene set)이나 신호전달 경로 수준의 구조적 정보를 충분히 활용하지 못한다는 한계가 있다.

한편, 자연어처리 분야에서 발전한 사전학습 언어 모델(pretrained language model)을 SMILES(Simplified Molecular Input Line Entry System) 문자열에 적용한 화학 언어 모델(chemical language model)의 등장으로, 그래프 변환 없이도 대규모 비라벨 분자 데이터로부터 화학 구조 표현을 학습하는 접근이 활발히 연구되고 있다[9],[10]. ChemBERTa와 유사한 계열의 모델들은 self-supervised 학습을 통해 SMILES 수준에서 일반적인 화학 패턴을 학습한 뒤, 다운스트림 과제(IC50, 독성 예측 등)에 전이 학습하여 데이터 효율성과 성능을 동시에 확보할 수 있음을 보였다[9]. 그러나 기존 ChemBERTa 기반 약물 반응 예측 연구의 상당수는 "SMILES → ChemBERTa → 세포 표현과 단순 결합 → IC50 회귀" 구조에 머물러, 약물이 표적하는 신호전달 경로, 표적 단백질 결합, 또는 기전(annotation)과 같은 생물학적 사전 지식을 약물 임베딩의 구조화된 제약으로 활용하지 못하는 공통의 한계를 가진다.

세포주 표현 측면에서도, 고차원 유전자 발현 벡터를 MSigDB(Molecular Signatures Database) Hallmark와 같은 기능적 gene set으로 집약하면 잡음을 줄이고 해석 가능성을 높일 수 있음이 다양한 오믹스 분석에서 보고되어 왔다[11],[12]. 그럼에도 다수의 약물 반응 예측 모델은 gene set을 단순 집계 피처로만 사용하거나, 약물 표현과의 상호작용을 충분히 모델링하지 못한다. 즉, 특정 약물이 어떤 gene set(예: DNA repair, apoptosis, PI3K/MTOR signaling)의 활성 패턴에 주로 의존하는지를 약물 조건부(drug-conditioned)로 학습하는 구조는 상대적으로 부족하다. 또한 GDSC와 같이 약물에 대해 TARGET_PATHWAY와 같은 기전 주석이 부여된 경우에도, 이를 단순 분류 문제에만 사용하거나 완전히 무시하는 경우가 많아, 사전 지식을 표현 학습 단계에서 강하게 반영하는 "pathway-aware" 약물 임베딩 설계는 여전히 연구 여지가 크다[2]-[4].

본 연구에서는 이러한 한계를 보완하기 위해, ChemBERTa 기반 약물 임베딩과 기능적 gene set 수준의 세포 표현, 그리고 기전 정보 기반 대조 학습을 통합한 항암제 반응 예측 프레임워크를 제안한다. 구체적으로, (i) GDSC RNA-seq에서 LINCS(Library of Integrated Network-based Cellular Signatures) L1000과의 교집합으로 얻은 949개 유전자 발현을 MSigDB Hallmark gene set(및 GO 기반 gene set 클러스터) 수준의 경로별 활성도 벡터(pathway-level representation)로 집약하여 세포 상태를 표현하고[11], (ii) SMILES로부터 얻은 ChemBERTa 약물 임베딩을 이용해 gene set 별 가중치를 조절하는 약물 조건부 gene set 게이팅(drug-conditioned gene set gating) 및 drug-to-gene set cross-attention을 적용함으로써, 약물-세포-경로 간 상호작용을 명시적으로 모델링한다. 더불어 (iii) GDSC가 제공하는 TARGET_PATHWAY 주석을 supervised contrastive learning의 레이블로 활용하여, 동일 기전을 공유하는 약물들이 임베딩 공간에서 서로 가깝게 클러스터링되도록 유도함으로써 pathway-aware 약물 표현을 학습한다[2]-[4].

Ⅱ. 본 론

2-1 실험 데이터 및 전처리

1) GDSC 데이터셋과 전처리

본 연구는 Genomics of Drug Sensitivity in Cancer (GDSC) 데이터베이스를 활용하여 항암제-세포주 조합에 대한 약물 민감도(IC50)를 학습·예측하였다. GDSC는 수백 개의 항암제와 약 1,000개의 암 세포주에 대한 대규모 약물 스크리닝 결과를 제공하며, 각 조합에 대해 반수 최대 억제 농도(IC50)를 측정하여 약물 반응의 정량적 지표로 사용한다

[3],[4]. IC50 값은 넓은 농도 범위를 다루기 위해 로그 변환(log10)을 적용하여 정규화하였으며, 이는 약물 효능을 안정적으로 비교하기 위한 표준적인 전처리 절차이다[2],[13].

세포주 표현에는 GDSC에서 제공하는 RNA-seq 기반 전사체 데이터를 사용하였다. 계산 효율과 재현성을 고려하여, 전체 유전자 중 LINCS L1000 landmark 유전자 집합과의 교집합을 취해 총 949개의 유전자를 선별하였다[4],[14]. 유전자 발현값은 TPM(Transcripts Per Million) 단위로 정규화된 값을 기반으로 하였으며, 분포 안정화를 위해 log2(TPM+1) 변환을 추가 적용하였다. 이러한 RNA-seq 기반 전사체 정량 및 정규화 전략은 대규모 암 세포주 분자 특성 분석 연구에서도 널리 사용되어 왔다[15]-[17]. 또한 RPKM(Reads Per Kilobase of transcript per Million mapped reads)은 샘플 간 비교에서 일관성이 떨어질 수 있으므로 TPM 기반 정규화를 사용하였다[18]. 유전자 식별자는 Entrez Gene ID(NCBI에서 제공하는 유전자 식별자)를 기준으로 통일하였고, 결측값이 존재하는 샘플은 제거함으로써 데이터 품질을 확보하였다[14].

약물 구조 정보는 canonical SMILES 표기를 사용하였다. GDSC 메타데이터에서 제공하는 약물 식별자를 기준으로 SMILES를 매핑하고, RDKit를 이용해 SMILES 유효성을 검증하였다[19]. 파싱 불가능하거나 구조적 결함이 있는 분자는 제거하였으며, 염(salt) 형태로 제공된 경우 주요 분자 골격만을 추출하여 사용하였다. 최종적으로 유효한 SMILES, 전사체 데이터, IC50 값을 모두 갖춘 약물-세포주 조합만을 분석에 포함하였다.

데이터 분할은 train/validation/test = 8:1:1의 비율로 무작위 분할을 적용하였다. 데이터 누수(data leakage)를 방지하기 위해 동일한 (세포주, 약물) 조합은 하나의 분할에만 속하도록 그룹화하였으며, 세포주와 약물의 분포가 세 분할 간 과도하게 편향되지 않도록 계층화(stratification)를 적용하였다. 무작위 시드는 고정하여 재현성을 확보하였고, 이후 제시하는 모든 실험은 동일한 분할을 사용하여 공정한 비교가 가능하도록 하였다.

2) 기능적 유전자 집합 기반 세포 표현

세포주는 앞 절에서 선정한 949개 유전자 발현값을 기능적 유전자 집합(gene set) 차원으로 요약하여 표현하였다. 구체적으로, MSigDB의 Hallmark 컬렉션에서 인간(human)에 해당하는 gene set을 선택하고, 각 gene set에 포함된 유전자를 GDSC RNA-seq 데이터의 유전자 식별자(Entrez ID 및 gene symbol)와 매핑하였다[11],[12]. 이 과정에서 GDSC 발현 데이터와 교집합이 전혀 없거나 포함 유전자 수가 매우 적은 gene set은 분석에서 제외하며, 충분한 정보량을 갖는 48개의 Hallmark gene set만을 최종적으로 사용하였다.

세포주 c에서 gene set s의 활성도 $h_{c,s}$는 해당 집합에 포함된 유전자 발현값의 산술 평균으로 정의하였다. 세포주 c에서 유전자 g의 정규화된 발현값을 $x_{c,g}$, gene set s에 포함된 유전자 집합을 G_s라고 할 때 활성도는 다음과 같다.

$$h_{c,s} = \frac{1}{|G_s|} \sum_{g \in G_s} x_{c,g} \quad (1)$$

각 gene set에 대해 모든 세포주의 $h_{c,s}$ 분포를 기준으로 z-score 표준화를 추가 적용하여, 특정 gene set의 수치 범위 차이만으로 학습이 편향되는 현상을 완화하였다[15]. 이렇게 얻어진 $h_c \in R^s$는 고차원 유전자 발현 벡터를 기능적 경로 단위의 저차원 활성도 벡터로 투영한 것으로, 노이즈를 줄이면서 세포 내 주요 경로 패턴을 보존하는 것을 목표로 한다.

추가 분석에서는 Hallmark 이외에 Gene Ontology(GO) 기반 gene set 클러스터를 정의하여, gene set 구성 방식에 따른 성능 차이를 비교하였다[20]. GO 기반 집합은 biological process term을 중심으로 유전자 목록을 수집한 뒤, 유전자 겹침 비율을 기준으로 유사 term을 군집화하여 과도한 중복을 줄이는 방식으로 구성하였다.

3) TARGET_PATHWAY를 이용한 약물 기전 라벨링

GDSC는 각 약물에 대해 주요 표적 신호전달 경로를 나타내는 TARGET_PATHWAY 주석을 함께 제공한다[2]-[4]. 본 연구에서는 GDSC (Release 8.4) drug annotation(약물 메타데이터)에서 제공되는 TARGET_PATHWAY 컬럼을 수집하고, DRUG_NAME-TARGET_PATHWAY 매핑 테이블을 약물 임베딩에 기전 수준의 구조를 부여하기 위한 감독 신호로 활용하기 위해 구축하였다. 우선 GDSC 메타데이터에서 TARGET_PATHWAY 항목이 정의된 약물만을 추출한 뒤, 표기 차이 또는 의미가 중복되는 항목(예: PI3K/mTOR signaling, PI3K/MTOR pathway)은 수동 검토를 통해 소수의 대표 경로명으로 통합하였다. 하나의 약물에 여러 경로가 주석된 경우에는 GDSC가 제시하는 주 기전(primary mechanism)을 우선적으로 선택하였고, 기전이 모호하거나 “other”와 같이 비특이적으로 표기된 약물은 대조 학습에서 제외하였다. 정제 결과, TARGET_PATHWAY 라벨이 적용된 고유 약물은 총 352개이며 최종 경로 클래스 수는 22개이다. 정제된 TARGET_PATHWAY 목록에 대해 고유한 경로 인덱스를 부여하고, 동일 경로를 공유하는 약물은 동일한 정수 라벨을 갖도록 구성하였다. 이후 학습 단계에서 이 라벨을 supervised contrastive loss의 클래스 정보로 사용하여, 동일 기전이 갖는 약물들이 임베딩 공간에서 서로 가깝게 위치하도록 유도하였다[21]. IC50 회귀 손실은 모든 약물-세포주 조합에 대해 계산하되, contrastive loss는 TARGET_PATHWAY 라벨이 정의된 약물에 대해서만 추가로 적용하였다.

2-2 방법

1) 모델 개요

본 연구에서 제안하는 GS-ChemDRP 모델은 SMILES 기반 약물 인코더, gene set 기반 세포주 인코더, 그리고 두 임베딩을 결합하여 IC50을 회귀하는 상호작용 모듈로 구성된다. 약물은 canonical SMILES를 ChemBERTa 계열 사전학습 화학 언어 모델에 입력하여 고정 길이 분자 임베딩 $z_d \in R^{d_d}$로 표현하고(Drug token)[9], 세포주는 GDSC RNA-seq에서 전처리한 949개 유전자 발현을 MSigDB Hallmark gene set에 매핑하여 gene set 수준의 활성도 벡터 $a = (a_1, ..., a_K)$로 요약한다[11],[12]. 이렇게 얻은 Hallmark 기반 벡터 a는 다층 퍼셉트론(MLP)을 통해 저차원 세포 임베딩 공간으로 투영되며, 동시에 각 gene set k에 대한 임베딩 $h_{gk} \in R^{K \times d_g}$를 계산하여 $H_g = [h_{g1}, ..., h_{gK}] \in R^{K \times d_g}$라는 gene set 토큰 시퀀스를 구성한다.

약물 임베딩 z_d과 gene set 토큰 H_g는 gene set 게이팅(gene set gating)과 multi-head cross-attention으로 이루어진 상호작용 모듈에서 결합된다. 이 모듈은 세포 측 gene set들을 하나의 토큰 시퀀스로 간주하고, 약물 임베딩 z_d를 질의(query) 벡터로 사용하여 특정 약물이 어떤 gene set · 경로 축을 상대적으로 더 강하게 활용하는지를 나타내는 주의 가중치와 게이트 값을 학습한다. 이 과정을 통해 약물 조건부 gene set 표현이 형성되며, 최종적으로 약물 임베딩과 함께 회귀 헤드에 입력되어 IC50 예측에 사용된다.

약물 인코더에서 얻은 z_d는 IC50 회귀 손실과 함께 TARGET_PATHWAY 레이블을 활용한 supervised contrastive loss L_{con}로 동시에 학습된다[21]. 동일한 기전을 공유하는 약물들이 임베딩 공간에서 서로 가깝게 위치하고, 다른 기전을 갖는 약물들과는 분리되도록 학습함으로써, 예측 성능뿐 아니라 pathway-aware한 해석 가능성을 갖춘 약물 표현을 얻고자 하였다.

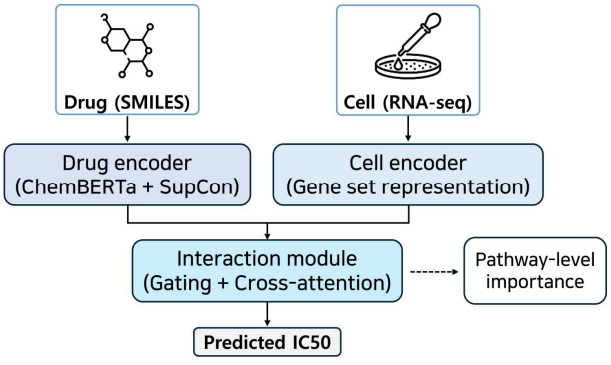

그림 1. 제안하는 GS-ChemDRP 프레임워크의 개요
Fig. 1. Overview of the proposed GS-ChemDRP framework

2) 약물 임베딩 추출

약물 표현은 SMILES 문자열에 대한 사전학습 화학 언어 모델(chemical language model)을 활용하여 구성하였다. 각 약물의 canonical SMILES는 RDKit를 이용해 표준화하고, 염(salt)과 용매 정보를 제거한 뒤, 구조적으로 유효하지 않은 분자는 제외하였다. 전처리된 SMILES는 ChemBERTa 계열의 Transformer 기반 언어 모델에 입력하여 고정 길이 분자 임베딩 벡터를 추출하였다.

사전학습된 ChemBERTa 모델의 토크나이저를 사용해 SMILES를 서브워드(subword) 토큰 시퀀스로 변환하고, 전체 시퀀스 앞에 [CLS] 토큰(서열 전체를 대표하는 토큰)을 추가하였다. 이후 Transformer 인코더의 마지막 층에서 [CLS] 토큰에 해당하는 은닉 상태를 해당 약물의 임베딩으로 사용하였으며, 이를 약물 임베딩 z_d(Drug token)으로 표기한다.

이 약물 임베딩 z_d는 상단의 IC50 회귀 헤드와 supervised contrastive learning 모듈로 동시에 전달되며, 약물 구조에 내재된 화학 패턴과 기전 정보를 함께 반영한다. ChemBERTa의 사전학습 가중치는 본 연구 모델의 초기값으로 사용하였으며, GDSC 기반 IC50 예측 과제에 대해 end-to-end 미세조정(fine-tuning)을 수행하였다. 이를 통해 대규모 비라벨 분자 데이터에서 학습된 일반적 화학 표현을 유지하면서, 항암제 반응 예측에 특화된 표현으로 적응시키고자 하였다.

3) 세포주 임베딩 추출

세포주 인코더는 2-1의 2)에서 정의한 기능적 유전자 집합 기반 활성도 벡터를 입력으로 사용하였다. 각 세포주는 949개 유전자 발현값 $\{x_g\}$로 표현되고, 이를 Hallmark(또는 GO 클러스터) gene set에 매핑하여 gene set별 활성도 a_k를 계산하였다. gene set k에 포함된 유전자 집합을 G_k라 할 때, 세포주의 gene set 활성도는 다음과 같이 정의하였다.

$$a_k = \frac{1}{|G_k|} \sum_{i \in G_s} x_g \tag{2}$$

각 세포주는 K개의 gene set에 대한 활성도 벡터 $a = (a_1, ..., a_K) \in R^K$로 표현되며, gene set별 분포 차이를 보정하기 위해 세포주 전체에 대해 각 성분 a_k를 기준으로 z-score 표준화를 적용하였다.

전역적 세포 상태를 요약하는 임베딩을 얻기 위해, 활성도 벡터 a 전체를 입력으로 하는 얕은 다층 퍼셉트론(MLP)을 적용하였다. 이 MLP는 입력 차원 K를 모델 내부에서 사용하는 세포 임베딩 차원 d_c로 투영하는 2층 완전연결 신경망으로 구성되며, 중간 층에 비선형 활성 함수와 드롭아웃을 포함한다. 이렇게 얻어진 벡터를 전역 세포 임베딩 $c \in R^{d_c}$로 정의

하며, 이후 약물 임베딩과 함께 상호작용 모듈의 입력으로 사용된다.

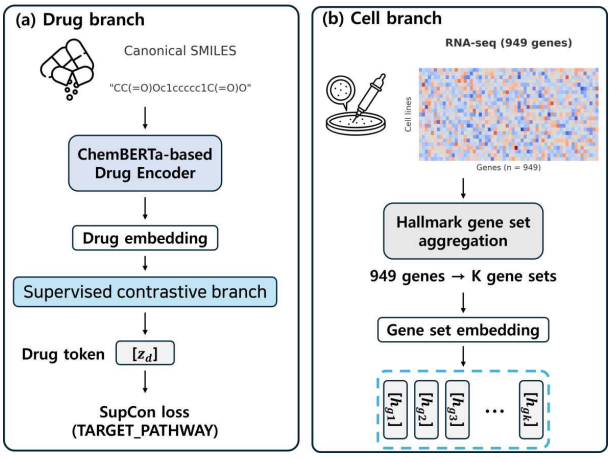

그림 2. GS-ChemDRP의 입력 인코더 구조
Fig. 2. Input encoder architecture of the GS-ChemDRP framework

동시에, gene set 수준의 세부 정보를 보존하기 위해 활성도 벡터 a의 각 성분 a_k에 대해 별도의 선형 변환(또는 소규모 MLP)을 적용하여 차원 d_g의 gene set 임베딩 벡터 $h_{gk} \in R^{d_g}$를 계산하였다. 즉,

$$h_{gk} = f_g(a_k), \quad k = 1,\ldots,K, \tag{3}$$

여기서 f_g는 1-2층의 얕은 MLP를 의미한다. 모든 gene set에 대해 $H_g = [h_{g1},\ldots,h_{gK}] \in R^{K \times d_g}$라는 길이 K의 gene set 임베딩 시퀀스를 구성하였으며, 그림 2(b)에서 각 토큰은 $[h_{gk}]$로 표시된다. 이 시퀀스는 이후 약물-세포 상호작용 단계에서 약물을 조건으로 한 gene set 게이팅과 cross-attention의 입력으로 사용되어, 특정 약물이 어떤 gene set·경로 축을 주로 활용하는지를 학습하도록 한다. Hallmark 기반 gene set과 GO 클러스터 기반 gene set 모두 동일한 인코더 구조를 사용하였으며, 두 표현 방식의 성능 차이는 결과(III장)에서 비교하였다.

4) 약물-세포 상호작용 모듈

약물-세포 상호작용 모듈은 2-2의 2)와 3)에서 얻은 약물 임베딩 z_d와 gene set 임베딩 시퀀스 H_g를 결합하여, 특정 약물이 특정 세포주에서 어떤 경로 축을 통해 작용하는지를 모델링하고 최종 IC50을 예측하는 역할을 한다. 필요에 따라 전역 세포 임베딩 c 역시 회귀 헤드 입력에 포함하였다.

먼저 약물-세포 간 전역 상호작용을 반영하기 위해, 약물 임베딩 z_d와 세포 전역 임베딩 c를 단순 연결(concatenation)한 벡터 $[z_d;c]$를 소규모 다층 퍼셉트론(MLP)에 통과시켜 통합 표현 u를 계산하였다. 이 전역 표현 u는 해당 약물-세포 조합의 전반적인 민감도 수준을 요약하는 정보로 사용되며, 이후 회귀 헤드의 입력으로 활용된다.

동시에, gene set 수준에서의 세밀한 상호작용을 모델링하기 위해, gene set 임베딩 시퀀스 H_g에 대해 약물 조건부(drug-conditioned) 게이팅과 cross-attention을 적용하였다. 각 gene set k에 대해 약물 임베딩 z_d와 gene set 임베딩 h_{gk}를 결합한 후, 소규모 신경망을 통해 스칼라 게이트 값 s_k를 계산하고, 시그모이드 함수를 통해 [0,1] 범위로 정규화하여 gene set별 중요도를 추정하였다.

$$s_k = \sigma(w_g^T[z_d;h_{gk}] + b_g), \quad s_k \in (0,1) \tag{4}$$

이렇게 얻어진 게이트 s_k는

$$\widetilde{h_{hk}} = s_k h_{gk} \tag{5}$$

와 같이 원래 gene set 임베딩에 곱해져, 특정 약물이 특정 세포에서 상대적으로 더 강하게 활용하는 gene set·경로 축을 강조하는 역할을 한다. 정보 손실을 줄이기 위해, 필요 시 잔차(residual) 경로를 함께 두어 $h'_{gk} = h_{gk} + \widetilde{h_{gk}}$ 형태로 사용하는 구조를 적용하였다.

이후 약물 임베딩 z_d를 질의(query), 게이팅된 gene set 임베딩 $\widetilde{H_g} = (\widetilde{h_{g1}},\ldots,\widetilde{h_{gK}})$를 키/값(key/value)로 입력하는 multi-head cross-attention 층을 적용하여, 약물 관점에서 재가중된 gene set 요약 벡터 v을 계산하였다[22]. 이 벡터 v는 해당 약물이 이 세포주에서 어느 gene set·경로 축을 통해 주로 작용하는지를 요약하는 표현으로 해석할 수 있다.

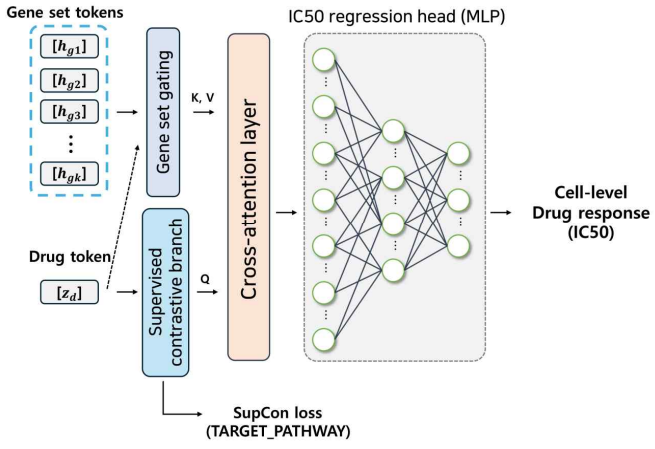

그림 3. Gene set gating과 cross-attention을 이용한 약물-세포 상호작용 모듈
Fig. 3. Drug-cell interaction module based on gene set gating and cross-attention

최종적으로는 전역 표현 u, gene set 기반 표현 v, 그리고 원래 약물 임베딩 z_d를 적절히 결합하여 $[u;v;z_d]$ 형태의 통합 벡터를 구성하고, 이를 회귀 헤드에 입력하여 IC50 값을 예측하였다. 회귀 헤드는 1-2층의 완전연결 신경망으로 구성하였으며, 출력 차원은 스칼라 IC50 예측값 하나 $\hat{y}$로 설정하였다.

5) 학습 및 손실 함수

모델 학습에서는 IC50 회귀 손실과 TARGET_PATHWAY 기반 supervised contrastive loss, 그리고 gene set 게이팅에 대한 규제 항을 함께 사용하였다.

우선, 약물-세포 조합 i에 대해 실측 IC50(log-변환 값)을 y_i, 모델이 예측한 값을 $\hat{y}_i$라 두면, 기본 회귀 손실은 평균제곱오차(Mean Squared Error, MSE)로 정의하였다.

$$L_{reg} = \frac{1}{|N|}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{6}$$

보고 시에는 루트평균제곱오차(RMSE)를 사용하여 예측 오차 크기를 해석하였고, 테스트 단계에서는 예측값과 실측값 간의 피어슨 상관계수(Pearson correlation coefficient, PCC)를 추가로 산출하여 IC50 순서·방향 보존 정도를 평가하였다.

약물 임베딩 z_d가 TARGET_PATHWAY 정보를 반영하도록 하기 위해, GDSC에서 정의한 TARGET_PATHWAY 레이블을 활용한 supervised contrastive learning을 적용하였다. 미니배치 내에서 동일 TARGET_PATHWAY를 공유하는 약물 임베딩 z_d들을 양성(positive) 쌍, 다른 경로에 속하는 약물들을 음성(negative) 쌍으로 간주하고, 약물 임베딩 간 코사인 유사도를 기반으로 하는 대조 손실을 계산하였다. 직관적으로, 이 손실 L_{con}은 동일 기전을 갖는 약물 임베딩은 서로 가깝게, 다른 기전을 갖는 약물 임베딩은 떨어지도록 임베딩 공간을 재구성하는 역할을 하며, 구체적인 수식은 supervised contrastive learning에서 사용되는 표준 형태를 따른다. TARGET_PATHWAY가 정의되지 않았거나 "other"로만 표기된 약물은 대조 손실 계산에서 제외하고, 회귀 손실만으로 학습하였다.

gene set 게이팅 모듈에 대해서는, 소수의 gene set만 강하게 선택되도록 하는 희소성(sparsity) 규제와, 극단적으로 한두 개의 gene set만 비정상적으로 큰 값을 갖지 않도록 하는 완화(smoothness) 규제를 추가하였다. 희소성 규제는 gene set 게이트 값 s_k에 대한 L1 노름을 최소화하는 항

$$L_{sp} = \frac{1}{K}\sum_{k=1}^{K}|s_k| \tag{7}$$

로 정의하였고, 완화 규제 L_{sm}는 인접하거나 기능적으로 유사한 gene set 간 게이트 값이 과도하게 들쭉날쭉하지 않도록, 게이트 분포의 변동성을 억제하는 형태로 정의하였다.

최종적으로 모델은 다음과 같은 합성 목적함수를 최소화하도록 학습하였다.

$$L_{total} = L_{reg} + \alpha L_{sp} + \beta L_{con} + \gamma L_{sm} \tag{8}$$

여기서 α,β,γ는 검증 세트 성능을 기준으로 설정한 하이퍼파라미터이며, 회귀 성능(RMSE, PCC)과 게이트의 해석 가능성(소수의 의미 있는 gene set만 선택되는지)을 함께 고려하여 선택하였다. 최적화는 Adam(Adaptive Moment Estimation) 계열 옵티마이저를 사용하였고, 학습률과 배치 크기 등 나머지 하이퍼파라미터는 II장의 서두에서 기술한 설정에 따라 고정하였다.

2-3 평가 방법

본 연구는 약물-세포주 조합에 대한 IC50을 연속값으로 회귀하는 모델을 다루므로, 분류 지표는 사용하지 않았다. 예측값이 실측값과 얼마나 가까운지(오차 규모), 그리고 IC50의 상대적 크기·순서를 얼마나 잘 보존하는지(상관성)를 평가하기 위해, 평균제곱오차(MSE)와 루트평균제곱오차(RMSE), 그리고 피어슨 상관계수(Pearson correlation coefficient, PCC)를 사용하였다.

1) MSE / RMSE

샘플 I에 대해 실측 IC50을 y_i, 모델 예측값을 $\hat{y}_i$라 두고, 전체 테스트 샘플 집합을 D_{test}라 할 때 평균제곱오차(MSE)는

$$MSE = \frac{1}{|D_{test}|}\sum_{i\in D_{test}}(\hat{y}_i - y_i)^2 \tag{9}$$

로 정의하였다. 루트평균제곱오차(RMSE)는 $\mathrm{RMSE} = \sqrt{MSE}$로 계산하였으며, 값이 작을수록 예측 오차가 작음을 의미한다.

2) 피어슨 상관계수 (PCC)

예측값과 실측값의 선형 상관성을 평가하기 위해, 테스트 집합에서 피어슨 상관계수를

$$PCC = \frac{\sum_{i\in D_{test}}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i\in D_{test}}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i\in D_{test}}(y_i - \bar{y})^2}} \tag{10}$$

와 같이 계산하였다. 여기서 $\hat{y}$와 $\bar{y}$는 각각 예측값과 실측값의 평균이다. PCC 값이 1에 가까울수록, 약물-세포주 조합 간 IC50의 상대적인 크기와 변동 방향을 더 잘 보존함을 의미한다.

3) 데이터 분할 및 재현성

모든 실험은 2-1절에서 기술한 것과 동일하게, GDSC 데이터셋을 train/validation/test = 8:1:1 비율로 무작위 분할하여 수행하였다. 데이터 누수를 방지하기 위해 동일한 (세포주,약물) 조합은 하나의 분할에만 속하도록 그룹화 분할을 적용하였다. 무작위 시드는 고정하여 재현성을 확보하였으며, 세포주와 약물의 분포가 세 분할 간 과도하게 치우치지 않도록 각 분할의 샘플 수와 IC50 분포를 확인하였다. 이후 제시하는 모든 성능 평가는 이 고정된 분할을 기준으로 산출하였다.

Ⅲ. 실험결과

3-1 약물 반응 예측 성능 비교

제안 모델의 성능을 평가하기 위해, GDSC 데이터셋에서 구축한 train/validation/test = 8:1:1 분할(2-1절 참조)을 모든 비교 대상 모델에 공통으로 적용하였다. 평가는 IC50의 회귀 문제의 특성을 고려하여 루트평균제곱오차(root mean squared error, RMSE)와 피어슨 상관계수(Pearson correlation coefficient, PCC)를 지표로 사용하였다.

표 1. GDSC 데이터셋에서의 약물 반응 예측 성능 비교
Table 1. Comparison of drug response prediction performance on the GDSC dataset

Category	Model	RMSE	PCC
Traditional ML	Ridge	1.2828	0.8840
	RF	1.7768	0.7625
	XGBoost	1.2404	0.8923
	MLP	1.2157	0.8972
Deep models (existing)	CSG2A	1.0698	0.9213
Proposed	ChemBERTa+MLP (Baseline)	1.1327	0.9112
	GS-ChemDRP + Gating	1.0812	0.9201
	GS-ChemDRP (Full, Gating+SCL)	**1.0682**	**0.9220**

표 1은 동일한 데이터 분할에서 전통적인 회귀 모형(Ridgeregression, Random Forest, XGBoost, 다층 퍼셉트론), 딥러닝 기반 벤치마크 모델(CSG2A), 그리고 제안 모델(GS-ChemDRP)의 IC50 예측 성능을 비교한 결과를 요약한다. 전통적인 모델들 중에서는 MLP가 RMSE 1.2157, PCC 0.8972로 가장 우수한 성능을 보였으며, XGBoost 또한 RMSE 1.2404, PCC 0.8923으로 비교적 양호한 결과를 나타냈다. 반면, 랜덤 포레스트는 RMSE 1.7768, PCC 0.7625로 다른 모델에 비해 오차가 크고 상관계수가 낮아, 고차원 연속형 발현 피처에서 과적합과 일반화 성능 저하가 두드러지는 양상을 보였다.

선행 연구 모델인 CSG2A는 동일한 분할에서 RMSE 1.0698, PCC 0.9213을 기록하여 강한 기준선(baseline)을 제공하였다[23].

이에 비해, ChemBERTa 약물 임베딩과 gene set 기반 세포 표현을 결합한 제안 모델 ChemBERTa+MLP(baseline)는 RMSE 1.1327, PCC 0.9112를 기록하여, 전통적인 모든 회귀 모형(Ridge, RF, XGBoost, MLP)을 상회하는 성능을 보였다. 특히, MLP와 비교했을 때 RMSE는 약 6.8% 감소(1.2157→1.1327), PCC는 약 0.014p 증가(0.8972→0.9112)하여, 단순 유전자 벡터 입력 대신 기능적 유전자 집합 표현과 화학 언어 모델을 결합하는 접근이 약물 반응 예측에 유의미한 이득을 제공함을 확인할 수 있었다.

약물 조건부 gene set gating과 supervised contrastive learning을 도입한 GS-ChemDRP(Full, Gating+SCL (supervised contrastive learning))는 RMSE 1.0682, PCC 0.9220으로, CSG2A와 유사한 수준의 성능을 달성하였다. 즉, 제안 프레임워크는 전통적인 기계학습 모델 대비 일관된 성능 개선을 보이는 동시에, 강한 딥러닝 기준선과 비교 가능한 예측 성능을 유지하면서 경로 수준 해석 가능성을 구조적으로 도입한다는 점에서 의의가 있다.

표 2. GS-ChemDRP 구성 요소별 ablation 실험 결과(GDSC 데이터셋)
Table 2. Ablation study of GS-ChemDRP components on the GDSC dataset

Model	RMSE	PCC
ChemBERTa+MLP(Baseline)	1.1327	0.9112
GS-ChemDRP + Contrastive	1.1090	0.9148
GS-ChemDRP + Gating	1.0812	0.9201
GS-ChemDRP (Full, Gating+SCL)	**1.0682**	**0.9220**

3-2 GS-ChemDRP 구성 요소별 ablation 분석

제안 프레임워크의 각 구성 요소가 성능에 미치는 영향을 확인하기 위해, 동일한 ChemBERTa 약물 인코더와 gene set 기반 세포 표현을 유지한 채 손실 함수 및 gating 구조만 달리한 변형 모델들을 비교하였다.

ChemBERTa+MLP(Baseline)는 Hallmark/GO 기반 gene set 활성도 벡터를 MLP로 변환한 세포 임베딩과 약물 임베딩을 단순 결합하여 IC50을 회귀하는 구조이다. GS-ChemDRP + Contrastive는 baseline 구조에

TARGET_PATHWAY 레이블을 활용한 supervised contrastive loss를 추가하여, 동일 기전을 공유하는 약물 임베딩이 표현 공간에서 서로 가깝게 위치하도록 유도한 변형이다. GS-ChemDRP + Gating은 contrastive loss 없이, gene set별 스칼라 게이트를 학습하여 약물 조건부 gene set 재가중(drug-conditioned gating)을 수행하도록 한 변형이다. 마지막으로 GS-ChemDRP(Full, Gating+ SCL)는 gene set gating과 supervised contrastive loss를 모두 결합한 완전한 모델이다.

표 2에서 보듯이, baseline에 contrastive loss만 추가한 GS-ChemDRP + Contrastive는 RMSE 1.1090, PCC 0.9148로 baseline 대비 소폭의 성능 향상을 보였다. 반면, gene set gating만을 추가한 GS-ChemDRP + Gating은 RMSE 1.0812, PCC 0.9201로, baseline과 비교하여 오차를 더 크게 줄이고 상관계수를 유의하게 개선하였다. 이는 gene set 기반 세포 표현 자체뿐 아니라, 특정 약물이 어떠한 gene set·경로 축을 더 중요하게 사용하는지를 학습하는 gating 메커니즘이 IC50 예측에 직접적인 기여를 한다는 점을 시사한다.

두 요소를 모두 결합한 GS-ChemDRP(Full, Gating+ SCL)는 RMSE 1.0682, PCC 0.9220로, 개별 요소를 도입한 변형들보다 일관되게 더 나은 성능을 보였다. 즉, (1) TARGET_PATHWAY 정보를 이용해 약물 임베딩을 기전 수준에서 정렬시키는 contrastive 학습과, (2) 약물 조건부 gene set gating을 통한 세포 표현의 재가중이 상호 보완적으로 작용하여 최종 IC50 예측 성능을 개선한 것으로 해석할 수 있다.

3-3 Gene set 기반 모델 해석 가능성 평가

본 절에서는 제안한 GS-ChemDRP 모델이 학습한 gene set 수준 표현이 실제 약물 작용 기전을 얼마나 반영하는지 평가하였다. 이를 위해 (i) drug-conditioned gene set gating 모듈이 학습한 gene set 가중치 분포를 분석하고, (ii) 최종 예측 단계에서 사용되는 drug–gene set cross-attention 가중치를 추출하여 약물–경로 중요도 행렬을 구성·해석하였다.

먼저, GS-ChemDRP + Gating 및 GS-ChemDRP(Full, Gating+ SCL) 모델에서 전체 약물–gene set 쌍에 대해 추정된 gate 값 분포를 분석하였다. 학습이 수렴한 후 gate 값의 범위는 약 0.0–0.9991, 평균 0.30, 표준편차 0.24로 나타났다. 이는 모델이 48개 Hallmark gene set 토큰에 대해 균일한 가중치를 부여하는 것이 아니라, 일부 gene set에는 0에 가까운 값을, 일부에는 1에 가까운 값을 부여하며 선택적 재가중 패턴을 학습하고 있음을 의미한다.

다음으로, gene set gating으로부터 얻은 약물–gene set 중요도 행렬에 대해 약물 간(행 기준)·경로 간(열 기준) 분산을 계산하였다. gene set별 약물 중요도 분산은 약 0.0044, 약물별 gene set 중요도 분산은 약 0.0052로 비교적 낮은 값으로 나타났다. 이는 개별 약물마다 전혀 다른 경로 집합을 선택하기보다는, 일부 공통 Hallmark 경로(예: HEDGEHOG_SIGNALING, NOTCH_SIGNALING, ESTROGEN_RESPONSE_EARLY, WNT_BETA_CATENIN_SIGNALING 등)에 전반적으로 높은 가중치를 부여하는 경향이 있음을 의미한다[11],[12]. 다시 말해, gene set gating만을 기준으로 보면 모델이 부분적으로는 의미 있는 선택을 하고 있으나, 강하게 drug-specific한 pathway signature를 형성하는 수준까지는 제한이 있음을 확인하였다.

한편, 본 연구의 최종 예측 단계에서는 gene set gating으로 재가중된 gene set 임베딩을 key/value로 두고, 약물 임베딩을 query로 하는 cross-attention을 적용하여 세포 표현을 추가로 보정한다. 따라서 drug-specific pathway 중요도를 보다 직접적으로 평가하기 위해, 학습이 완료된 Full 모델에서 cross-attention 모듈의 set_attn 가중치를 추출하여 약물–gene set 중요도 행렬을 구성하였다. 각 약물에 대해 테스트 셋에서 해당 약물이 포함된 모든 샘플의 set_attn 가중치를 평균하여 약물별 pathway 중요도를 산출하고, 상위 10개 Hallmark gene set을 도출하였다. 작용 기전이 비교적 명확한 10종 대표 약물을 대상으로 문헌 기반으로 정리한 주요 작용 기전과, cross-attention 모듈이 예측한 상위 10개 Hallmark pathway 간의 일치 여부를 평가한 결과, Top-10 기준 50%(5/10)의 hit rate를 기록하였다.

표 3은 각 약물의 알려진 기전, 예측된 상위 경로, 일치 여부를 요약한다. 예를 들어 EGFR(Epidermal Growth Factor Receptor 억제제 Erlotinib, Gefitinib, Lapatinib는 KRAS_SIGNALING_UP 또는 EMT과 같은 Hallmark 경로가 상위 순위에 포함되었고, 다중 키나아제 억제제 Sunitinib은 KRAS_SIGNALING_UP를 높은 순위로 예측하였다. 또한 프로테아좀 억제제 Bortezomib은 UPR를 상위 10위 내에 포함하여 알려진 주요 기전과 정합한 결과를 보였다.

그림 4는 EGFR 억제제 Erlotinib의 상위 10개 pathway attention weight를 제시한다. EMT이 2순위로 예측되었으며, 이는 Erlotinib이 EMT 경로를 통해 작용한다는 문헌 보고와 일치한다. 가장 높은 attention weight를 보인 경로는 COMPLEMENT였으며, E2F(E2F transcription factors) targets가 3순위로 나타났다. 이는 cross-attention 모듈이 약물의 주요 작용 경로를 pathway-level에서 포착할 수 있음을 시사한다.

그러나 cross-attention 분석에서도 한계가 확인되었다. 여러 약물에서 MYC_TARGETS_V1/V2와 같이 범용적인 전사 프로그램과 연관된 경로가 상위 순위에 반복적으로 등장하는 경향이 있었고, gene set gating 기반 중요도와 cross-attention 기반 중요도 사이의 상관이 낮게 나타나 두 모듈이 서로 다른 관점에서 정보를 활용하고 있음을 보여준다. 이는 두 모듈이 일부 biologically plausible한 신호를 포착함에도 불구하고, 더 뚜렷한 drug-specific 경로 분기를

위해서는 보다 정교한 pathway 주석과 세분화된 기전 레이블, 그리고 손실 가중치 및 구조 설계의 추가 튜닝이 필요함을 시사한다.

표 3. 대표 약물의 상위 10개 예측 경로 및 알려진 기전 비교
Table 3. Comparison of top-10 predicted pathways and known mechanisms for representative drugs

Drug	Known Pathways	Top Predicted (Hit)	Hit Rate
Erlotinib	KRAS_UP(KRAS signaling up), EMT (epithelial–mesenchymal transition), HYPOXIA(hypoxia)	EMT (rank 2)	1/3
Gefitinib	KRAS_UP, HYPOXIA	KRAS_UP (rank 6)	1/2
Lapatinib	KRAS_UP, IFN_GAMMA (interferon-gamma response)	KRAS_UP (rank 5)	1/2
Sunitinib	ANGIO (angiogenesis), KRAS_UP, HYPOXIA	KRAS_UP (rank 1)	1/3
Sorafenib	ANGIO, APOPTOSIS(apoptosis)	–	0/2
Paclitaxel	G2M(G2/M checkpoint), SPINDLE(mitotic spindle), APOPTOSIS	–	0/3
Docetaxel	G2M, SPINDLE	–	0/2
Rapamycin	MTORC1(mTOR complex 1 signaling), GLYCOLYSIS (glycolysis)	–	0/2
AZD8055	MTORC1, GLYCOLYSIS	–	0/2
Bortezomib	UPR(unfolded protein response), APOPTOSIS	UPR (rank 6)	1/2

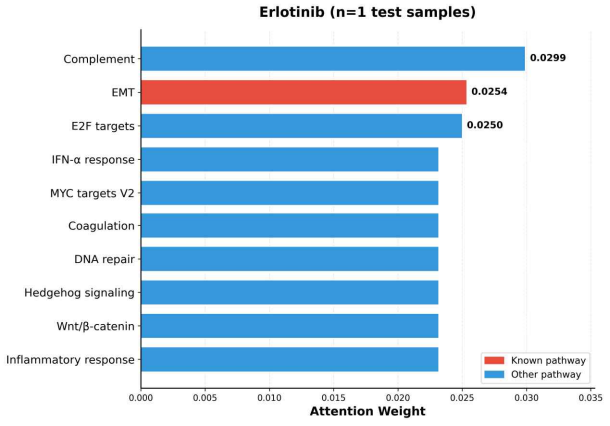

그림 4. Erlotinib의 pathway-level attention weights
Fig. 4. Pathway-level attention weights for Erlotinib

정리하면, 제안한 GS-ChemDRP는 (i) gene set-level gating을 통해 원시 유전자 발현을 약물 조건에 맞추어 조정하고, (ii) drug-gene set cross-attention을 통해 기능적 경로 수준에서 일정 수준의 기전 정합성을 확보함으로써, 기존 "black-box" IC50 예측 모델 대비 경로 수준 해석 가능성을 구조적으로 강화한 것으로 평가된다.

Ⅳ. 결 론

암 치료제에 대한 세포 수준 약물 민감도(IC50)를 정량적으로 예측하는 일은 정밀의학의 핵심 과제이지만, 약물의 화학 구조와 세포 내 신호전달 네트워크, 전사체 특성이 복합적으로 얽혀 있어 여전히 어려운 문제로 남아 있다. 본 연구에서는 GDSC 데이터셋을 기반으로, ChemBERTa 약물 임베딩과 기능적 유전자 집합(gene set) 기반 세포 표현을 결합한 GS-ChemDRP 프레임워크를 제안하고, IC50 회귀 문제에서의 예측 성능과 해석 가능성을 함께 평가하였다 [2]-[4].

전통적인 회귀 모형(Ridge, Random Forest, XGBoost, MLP)과 비교했을 때, GS-ChemDRP는 동일한 데이터 분할에서 일관되게 더 낮은 RMSE와 더 높은 PCC를 기록하였다. 특히 gene set 기반 세포 표현과 ChemBERTa 약물 임베딩만을 사용한 baseline 모델만으로도 MLP 대비 RMSE를 약 6~7% 감소시키고 PCC를 소폭 향상시키는 결과를 확인하였다. 또한 gene set gating과 supervised contrastive learning을 결합한 최종 GS-ChemDRP(Full, Gating+SCL)는 최신 딥러닝 기반 약물 반응 예측 모델(CSG2A)과 유사한 수준의 RMSE·PCC를 달성하여, 경쟁력 있는 예측 성능을 유지함을 확인하였다.

구성 요소별 ablation 분석을 통해, 제안 프레임워크의 설계 선택이 성능에 미치는 기여도도 확인하였다. TARGET_PATHWAY 레이블을 이용한 supervised contrastive loss는 약물 임베딩을 기전별로 정렬시켜 baseline 대비 성능을 소폭 개선하였고, 약물 조건부 gene set gating은 동일한 세포 입력을 사용하면서도 약물이 어떤 gene set · 경로 축을 더 강하게 활용하는지를 학습하도록 함으로써 성능을 보다 크게 향상시켰다. 두 요소를 결합한 Full 모델이 가장 우수한 성능을 보인다는 점은, (1) 기전 수준에서 정렬된 약물 표현과 (2) drug-conditioned gene set 재가중이 서로 보완적으로 작용해 IC50 예측력을 끌어올린다는 해석을 뒷받침한다.

해석 가능성 측면에서, 본 연구는 gene set gating 및 drug-gene set cross-attention 분석을 통해 제안 모델이 학습한 경로 수준 신호를 정량·정성적으로 확인하였다. gate 값 분포 분석 결과 모델은 일부 gene set을 선택적으로 강조/억제하는 재가중 패턴을 학습하였고, cross-attention 기반

분석에서는 작용 기전이 비교적 명확한 10종 대표 약물에 대해 문헌 기반 기전과의 Top-10 일치율이 50%(5/10)로 나타나, 일부 약물에서 생물학적으로 타당한 경로가 상위 순위로 회복되는 양상을 확인하였다. 이는 제안 프레임워크가 예측 성능을 유지하면서도 경로 수준 해석 단서를 구조적으로 제공한다는 점에서 의의가 있다.

약물 기전과의 정합성은 gene set gating과 drug-gene set cross-attention 두 수준에서 평가하였다. 첫째, gene set gating에서 유도한 약물–경로 중요도 행렬을 이용해 대표 약물 집합에서 상위 gene set과 문헌 기반 표적 경로 간의 일치도를 평가한 결과, 무작위 선택 대비 소폭 향상된 수준의 타당성을 보였다. 둘째, 최종 예측 단계에서 사용되는 drug-gene set cross-attention 모듈의 주의(attention) 가중치를 별도로 추출하여 동일한 분석을 수행한 결과, 작용 기전이 비교적 명확한 10종 대표 약물에 대해 Top-10 기준 50%(5/10)의 일치율을 확인하였다. 예를 들어, EGFR 억제제(Erlotinib, Gefitinib, Lapatinib)의 경우 KRAS_SIGNALING_UP 또는 EPITHELIAL_MESENCHYMAL_TRANSITION이 Hallmark gene set 상위 순위에 포함되었고, 다중 키나아제 억제제 Sunitinib에서는 ANGIOGENESIS 또는 KRAS_SIGNALING_UP이 상위 10위 내에 위치하는 등, 일부 대표 약물에서는 알려진 작용 기전과 일치하는 경로가 model-predicted 상위 gene set으로 회복되는 양상이 관찰되었다. 이는 약물 임베딩을 query로 사용하는 cross-attention이 gene set gating만을 사용할 때보다 기전 수준의 해석 가능성 측면에서 더 직접적인 단서를 제공함을 시사한다.

그럼에도 불구하고, 해석 가능성에는 분명한 한계도 존재한다. gene set gating과 cross-attention 모두에서 약물–경로 중요도 행렬의 약물 간·경로 간 분산이 전반적으로 낮게 나타났으며, 특히 cross-attention 분석에서는 MYC_TARGETS_V1/V2와 같이 범용적인 전사 프로그램 관련 Hallmark 경로가 많은 약물에서 상위 순위에 반복적으로 등장하는 경향이 있었다. 이는 모델이 일부 biologically plausible한 패턴을 포착하면서도, 강하게 drug-specific한 pathway signature를 뚜렷하게 분리해내는 수준까지는 도달하지 못했음을 의미한다. 이러한 한계는 GDSC의 TARGET_PATHWAY 레이블이 비교적 거친 범주의 기전 분류에 머무른다는 점, Hallmark gene set 정의만으로는 약물별 미세한 작용 기전 차이를 충분히 표현하기 어렵다는 점, 그리고 contrastive loss 가중치를 예측 성능 저하를 피하기 위해 보수적으로 설정한 점 등이 복합적으로 작용한 결과로 해석할 수 있다. 그럼에도 gene set gating과 cross-attention이 예측 성능을 유의하게 저해하지 않는 범위 내에서 기능 수준의 추가 정보를 제공하고, 일부 대표 약물에서 생물학적으로 타당한 경로를 상위 순위로 회복한다는 점은, 제안 프레임워크가 성능과 해석 가능성 간 균형을 향한 유의미한 출발점에 해당함을 시사한다.

앞서 논의한 내용을 종합하면, 본 연구의 한계와 향후 연구 과제는 다음과 같다.

(1) 본 연구는 GDSC 단일 데이터셋의 혼합 분할(mixed split) 설정에 기반하므로, drug-blind 및 cell line-blind 분할에서의 일반화 성능을 추가로 검증할 필요가 있다.

(2) 본 연구의 경로 해석은 MSigDB Hallmark gene set에 기반하므로 경로의 세분성이 제한될 수 있으며, 향후 KEGG·Reactome·OncoKB 등 보다 세분화된 경로 체계로 확장할 필요가 있다. 또한 LINCS L1000 교란 전사체(Perturbed Gene Expression) 기반 사전학습 후 GDSC IC50 미세조정을 통해 조건-특이적 표현 학습을 강화할 필요가 있다.

(3) TARGET_PATHWAY 라벨은 공개 주석 기반의 상위 수준 기전 분류로 노이즈/불확실성이 존재할 수 있어, 향후 더 정교한 기전 라벨링 또는 다중 라벨 학습 설정을 고려할 필요가 있다. 더불어 희소성(sparsity)·군집 구조를 반영한 gating 정규화를 통해 drug-specific pathway 선택의 분해능과 해석 가능성을 강화할 필요가 있다.

(4) 본 연구는 전사체 기반 세포 표현에 집중하였으나, 향후 유전체 변이, 복제수 변이 등 다중 오믹스 통합과 외부 데이터셋 검증을 통해 생물학적 근거와 적용 범위를 강화할 필요가 있다.

요약하면, 본 연구의 GS-ChemDRP는 ChemBERTa 기반 약물 임베딩과 Hallmark gene set 수준 세포 표현, 약물 조건부 gene set gating, drug-gene set cross-attention, 및 pathway-aware contrastive 학습을 결합함으로써, GDSC IC50 예측에서 기존 기계학습 및 딥러닝 모델과 경쟁력 있는 예측 성능을 달성하는 동시에 일정 수준의 생물학적 해석 가능성을 확보하였다. 이러한 결과는 기능적 경로 수준 표현과 화학 언어 모델을 접목한 접근이 정밀의학에서의 약물 반응 예측을 고도화하는 데 유효한 방향임을 보여주며, 향후 더 풍부한 생체 데이터와의 통합 및 실험적 검증을 통해 확장 가능한 연구 기반을 제공한다.

감사의 글

본 연구는 2025년도 식품의약품안전처 연구개발비(RS-2025-02215961), 과학기술정보통신부의 재원으로 한국연구재단 바이오·의료기술개발사업의 지원(RS-2025-16063391), 그리고 2025년도 교육부 및 광주광역시의 재원으로 광주RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE) 사업(2025-RISE-05-011)의 지원을 받아 수행되었으며, 이에 감사드립니다.

참고문헌

[1] D. M. Hyman, B. S. Taylor, and J. Baselga, “Implementing

Genome-Driven Oncology," *Cell*, Vol. 168, No. 4, pp. 584-599, February 2017. https://doi.org/10.1016/j.cell.2016.12.015

[2] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, ... and M. J. Garnett, "A Landscape of Pharmacogenomic Interactions In Cancer," *Cell*, Vol. 166, No. 3, pp. 740-754, July 2016. https://doi.org/10.1016/j.cell.2016.06.017

[3] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, ... and M. J. Garnett, "Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells," *Nucleic Acids Research*, Vol. 41, No. D1, pp. D955-D961, January 2013. https://doi.org/10.1093/nar/gks1111

[4] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, ... and C. H. Benes, "Systematic Identification of Genomic Markers of Drug Sensitivity in Cancer Cells," *Nature*, Vol. 483, No. 7391, pp. 570-575, March 2012. https://doi.org/10.1038/nature11005

[5] M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester, and J. Saez-Rodriguez, "Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties," *PLoS One*, Vol. 8, No. 4, e61318, April 2013. https://doi.org/10.1371/journal.pone.0061318

[6] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, ... and G. Stolovitzky, "A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms," *Nature Biotechnology*, Vol. 32, No. 12, pp. 1202-1212, December 2014. https://doi.org/10.1038/nbt.2877

[7] Q. Liu, Z. Hu, R. Jiang, and M. Zhou, "DeepCDR: A Hybrid Graph Convolutional Network for Predicting Cancer Drug Response," *Bioinformatics*, Vol. 36, No. Suppl_2, pp. i911-i918, December 2020. https://doi.org/10.1093/bioinformatics/btaa822

[8] T. Nguyen, G. T. T. Nguyen, T. Nguyen, and D.-H. Le, "Graph Convolutional Networks for Drug Response Prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 19, No. 1, pp. 146-154, January-February 2022. https://doi.org/10.1109/TCBB.2021.3060430

[9] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction," arXiv:2010.09885, 2020. https://arxiv.org/abs/2010.09885

[10] S. Honda, S. Shi, and H. R. Ueda, "SMILES Transformer: Pre-Trained Molecular Fingerprint for Low Data Drug Discovery" arXiv:1911.04738, 2019. https://arxiv.org/abs/1911.04738

[11] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, ... and J. P. Mesirov, "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102, No. 43, pp. 15545-15550, October 2005. https://doi.org/10.1073/pnas.0506580102

[12] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P Mesirov, and P. Tamayo, "The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection," *Cell Systems*, Vol. 1, No. 6, pp. 417-425, December 2015. https://doi.org/10.1016/j.cels.2015.12.004

[13] G. W. Caldwell, Z. Yan, W. Lang, and J. A Masucci, "The IC50 Concept Revisited," *Current Topics in Medicinal Chemistry*, Vol. 12, No. 11, pp. 1282-1290, 2012. https://doi.org/10.2174/156802612800672844

[14] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, ... and T. R. Golub, "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles," *Cell*, Vol. 171, No. 6, pp. 1437-1452, e17, November 2017. https://doi.org/10.1016/j.cell.2017.10.049

[15] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "Analysis of Microarray Data Using Z Score Transformation," *The Journal of Molecular Diagnostics*, Vol. 5, No. 2, pp. 73-81, May 2003. https://doi.org/10.1016/S1525-1578(10)60455-2

[16] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, ... and L. A. Garraway, "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity," *Nature*, Vol. 483, No. 7391, pp. 603-607, March 2012. https://doi.org/10.1038/nature11003

[17] M. Ghandi, F. W. Huang, J. Jané-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald, ... and W. R. Sellers, "Next-Generation Characterization Of The Cancer Cell Line Encyclopedia," *Nature*, Vol. 569, No. 7757, pp. 503-508, May 2019. https://doi.org/10.1038/s41586-019-1186-3

[18] G. P. Wagner, K. Kin, and V. J. Lynch, "Measurement of mRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent Among Samples," *Theory in Biosciences*, Vol. 131, No. 4, pp. 281-285, December 2012. https://doi.org/10.1007/s12064-012-0162-3

[19] A. P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton,

F. Atkinson, ... and A. R. Leach, “An Open Source Chemical Structure Curation Pipeline Using RDKit,” *Journal of Cheminformatics*, Vol. 12, No. 1, 51, September 2020. https://doi.org/10.1186/s13321-020-00456-1

[20] The Gene Ontology Consortium, “The Gene Ontology Resource: 20 Years And Still GOing Strong,” *Nucleic Acids Research*, Vol. 47, No. D1, pp. D330-D338, January 2019. https://doi.org/10.1093/nar/gky1055

[21] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, ... and D. Krishnan, “Supervised Contrastive Learning,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver: Canada, pp. 18661-18673, December 2020. https://doi.org/10.48550/arXiv.2004.11362

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, “Attention Is All You Need,” in Proceedings of the 31th International Conference on *Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, pp. 6000-6010, December 2017. https://doi.org/10.48550/arXiv.1706.03762

[23] D. Bang, B. Koo, and S. Kim, “Transfer Learning of Condition-Specific Perturbation in Gene Interactions Improves Drug Response Prediction,” *Bioinformatics*, Vol. 40, No. Suppl_1, pp. i130-i139, July 2024. https://doi.org/10.1093/bioinformatics/btae249

송종웅(Jongung Song)

2024년 : 전남대학교 (이학사)

※관심분야 : 생명정보학(bioinformatics), 인공지능(artificial intelligence)

유선용(Sunyong Yoo)

2012년 : 한국항공대학교 정보통신공학과 (공학석사)

2018년 : 한국과학기술원 바이오및뇌공학과 (공학박사)

2018년～2019년: 국민건강보험공단 빅데이터실 부연구위원

2019년～현 재: 전남대학교 지능전자컴퓨터공학과 교수

※관심분야 : 생명정보학(bioinformatics), 인공지능(artificial intelligence), 빅데이터(big data) 등