공학박사 학위논문

신약탐색을 위한 다중채널 기반의 인공지능 구조 설계

A study on multi-channel based AI architecture design for drug discovery

2023년 2월

서울대학교 대학원

치의과학과 의료경영과정보학

이문환

신약탐색을 위한 다중채널 기반의 인공지능 구조 설계

A study on multi-channel based AI architecture design for drug discovery

지도교수 김홍기

이 논문을 공학박사 학위논문으로 제출함
2023년 1월

서울대학교 대학원
치의과학과 의료경영과정보학
이문환

이문환의 공학박사 학위논문을 인준함
2023년 1월

위 원 장 _____ 장 병 탁 _____ (인)

부위원장 _____ 김 홍 기 _____ (인)

위 원 _____ 허 충 길 _____ (인)

위 원 _____ 김 필 종 _____ (인)

위 원 _____ 안 진 현 _____ (인)

초 록

새로운 약물이 원하는 효능을 갖도록 설계하는 것은 제약 산업에서 여전히 어려운 과제로 남아있으며 비용 집약적인 과정이 요구된다. 특히 Eroom의 법칙으로 대표되는 신약 개발의 비효율성 문제는 타분야의 급속한 기술발전과 매우 대조적이다. 1950년대 이래로 10억 달러의 연구개발 비용으로 FDA에서 허가된 신약의 수는 9년마다 약 절반으로 감소해왔다. 비록 질병 생물학의 발전과 생물정보학 기반의 의생명빅데이터 활용을 통해서 신약개발의 효율성을 개선하려 노력하고 있으나, 현재 10억 달러의 연구개발 비용으로 FDA에서 허가된 신약의 수는 1개 이하에 머무르고 있는 실정이다.

최근 신약개발의 연구개발 효율성을 높이기 위하여 인공지능 기반의 예측모델을 활용하는 연구가 증가하고 있다. 예를 들어서, 약물-표적 단백질 상호작용 식별은 약물 후보물질 발굴의 기초단계로써, 신약후보 선도물질 탐색, 약물 용도 변경, 오프 타겟 또는 부작용 예측과 같은 다양한 응용분야에서 중요한 역할을 한다. 이를 위해 전통적인 기계학습 모델부터 최신의 신경망 모델이 약물-표적 상호작용을 예측하기 위해 활용되고 있다. 그러나 약물 후보물질 또는 표적 단백질을 하나의 특질로만 표현한다는 점에서 여전히 개선의 여지가 남아있다.

이에 더해서, 암 환자 마다 상이한 항암제의 효능과 이질성은 항암제 신약 개발에서 해결해야할 도전적인 과제이다. 특히 유전체, 전사체 및 후성 유전체 전반에 걸친 종양 이질성은 항암제 치료 효능을

손상시킬 수 있다. 이를 극복하기 위해 최근 다양한 층위의 생물학적 데이터를 활용하는 다중 오믹스 통합 모델이 개발되고 있다. 그러나 기존의 통합 기법들은 다양한 오믹스 데이터를 동일한 차원의 특질로 구축하기 때문에 생물학 데이터 특유의 복잡성과 노이즈에 취약해진다는 한계가 있다. 이에 더해서, 단일 오믹스 데이터를 활용할 때보다 성능이 나빠지는 결과도 보고되고 있다.

본 연구에서는 각 데이터가 가진 특질을 다양한 채널로 구축하여 인공지능 모델이 다양한 생물학적 측면을 총체적으로 학습하도록 제안한다. 첫째, 약물-타겟 단백질 상호작용의 식별을 위해서 다중 채널 기반의 쌍입력 신경망(MCPINN)을 구축했다. MCPINN은 신경망의 3가지 활용 기법인 분류기, 특질 추출기, 그리고 종단 간 학습기를 활용하여 표현학습 능력을 극대화한다. MCPINN은 특질의 다양한 표현형을 다중 채널에 입력하여 활용하고 그 특질을 상보적으로 통합했다. MCPINN은 모델의 성능과 학습속도에서 가장 높은 성능을 보였다.

이에 더해서, 항암제 반응성 예측을 위해서 유전자 중심의 다중 채널(GCMC)을 구축했다. GCMC는 다중 오믹스 데이터를 3차원 텐서로 변환하며 새로운 차원은 오믹스 타입을 표현한다. GCMC는 각 유전자에 대한 모든 오믹스 채널의 특질을 통합하여 유전자 중심의 새로운 특질을 추출할 수 있다. GCMC는 265개의 암세포주 데이터와 TCGA 환자데이터, 그리고 PDX 환자 유래 생쥐 모델 데이터에서 기존의 최고 성능 모델보다 더 나은 성능을 보여주었다. 또한 GCMC는 다중 오믹스

프로파일을 균형있게 조합하여 예측 성능을 향상시킬 수 있다. 이러한 결과는 GCMC가 유전자 중심 방식으로 다중 오믹스 프로파일을 통합하여 성능 및 특질 추출 기능을 향상시킬 수 있음을 시사한다.

주요어 : 인공지능, 기계학습, 딥러닝, 신약탐색, 약물-타겟 연관관계 예측, 다중 오믹스 데이터 통합
학 번 : 2015-23266

목 차

표 목차

그림 목차

I. 서 론

1.1 연구의 배경

새로운 약물이 원하는 효능을 갖도록 설계하는 것은 제약 산업에서 여전히 어려운 과제로 남아있으며 비용 집약적인 과정이 요구된다. 특히 논문 “제약 R&D 효율성 저하 진단”[1]에서 제시된 Eroom의 법칙에 따르면, 약물 개발의 효율성은 시간이 지날수록 악화되고 있으며 이는 타 분야의 급속한 기술 발전과 대조적이다. 이를테면 신약 탐색과 관련된 기술들인 고효율 스크리닝(high-throughput screening) 기법, 생명공학, 화학 합성(chemical synthesis) 기법 및 계산 약물 설계 기법이 발전했음에도 불구하고, 그림 1에서 보이듯이, 10억 달러의 연구개발 비용으로 FDA에서 허가된 신약의 수는 1950년 이래로 9년마다 약 절반씩 감소해왔다. 이와 같은 연구 개발비용의 비효율성을 강조하기 위하여 Eroom의 법칙은 Moore의 법칙[①]을 거꾸로 써서 명명되었다.

Eroom의 법칙을 야기하는 4가지 요소로는 (1)‘비틀즈 보다 낫다’ 문제, (2) ‘신중한 규제자’ 문제, (3) ‘돈 던지기’ 경향, 그리고 (4) ‘기초연구의 무차별 대입’ 편향이 제시된다. 특히 ‘비틀즈보다 낫다’ 문제는 제약 산업만의 독특한 문제를 음악 산업에 빗대어 설명한

---

① 1964년 인텔의 공동 설립자인 고든 얼 무어(Gordon Earle Moore)가 경험적인 관찰을 바탕으로 제시한 법칙으로, 반도체 집적회로의 성능이 24개월에 2배씩 향상된다는 법칙이다.

요소이다. 이를테면, 제약 산업은 기존의 약품이 수십년간 널리 쓰이는 상황에서 신약이 기존의 약품보다 효능이 좋아야 한다는 특성이 있다. 이를 마치 많은 사람들이 오래된 비틀즈 음악을 수십년간 즐겨 듣는 상황에서 비틀즈보다 더 나은 신곡이 있어야만 성공한다는 상황으로 비유한다. 따라서 기존의 유망한 치료법을 대체할 신약을 개발하는 것은 매우 어려운 실정이다.

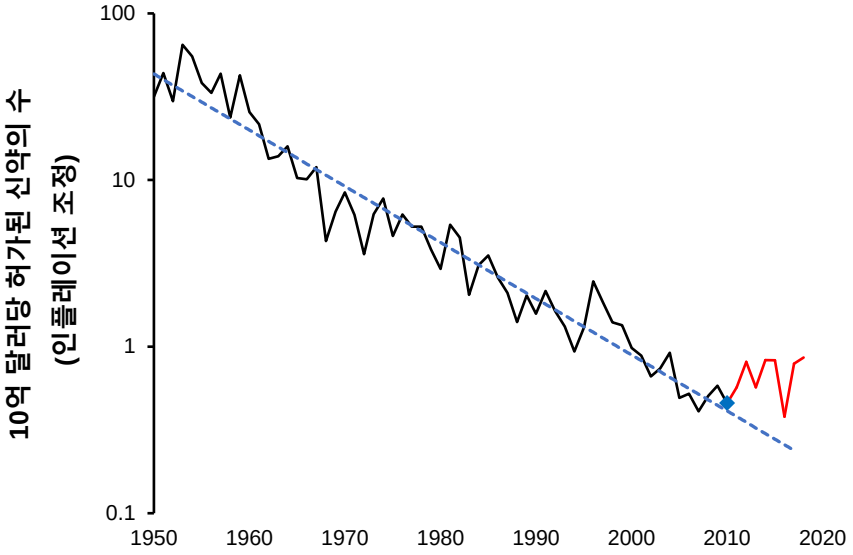

그림 1. 10억 달러의 연구개발 비용으로 허가된 신약의 수

반면에 최근 연구[2]에 따르면, 생물학의 발전과 적용 덕분에 신약개발의 효율성이 Eroom의 법칙을 따르지 않고 그 생산성이 증가되고 있다는 가능성이 드러나고 있다. 그림 1에서 붉게 표현된 2010년 이후의 경향성에서 보이듯이, 2010년부터 2018년까지 10억 달러의 연구개발 비용으로 FDA에서 승인된 신약 개체 수는 기존의 추세선을 상회했다. 해당 연구의 저자들은 이런 개선에 영향을 준

세가지 요인이 있다고 설명한다. 그 중에서 질병 생물학 또는 이를 조절하는 기법들을 통해서 질병에 대한 더 나은 정보를 얻는 것이 첫 요소로 꼽혔다. 이 정보를 활용해서 아직 효과적인 치료법이 없는 질병을 위한 신약을 탐색하도록 의사결정을 내릴 수 있다. 이는 '비틀즈보다 낫다' 문제를 회피하는 것에 큰 도움을 준다. 특히 저자들은 전장유전체 연관분석(GWAS)의 활용을 그 신약개발연구의 효율성 향상과 연관 지었다. 최근에는 차세대 염기서열 분석법(NGS)을 통해서 의생명빅데이터가 구축되고 있기에 이에 기반한 분석 및 활용의 잠재력이 더욱 커지고 있다.

최근 인공지능을 활용하여 신약개발의 효율성을 높이려는 시도가 증가하고 있다. 이를 위해서 전통적인 기계학습 모델인 나이브 베이즈, 랜덤 포레스트, 서포트 벡터 머신부터 신경망 모델에 이르는 다양한 모델이 활용된다. 특히 신경망 모델은 높은 성능을 위해서는 많은 수의 데이터가 요구되는데, 다양한 의생명빅데이터의 활용을 통해서 신경망 모델들의 성능이 개선되고 있다. 예를 들어서, ChEMBL[3]은 많은 수의 약물후보물질과 타겟 단백질, 그리고 생체활성도 실험 데이터를 제공한다. 이에 더해서 항암제의 경우에는 GDSC[4]가 약 1000여개의 암세포주에 대해서 265개의 항암제를 실험한 결과를 제공한다. 이처럼 의생명빅데이터는 약물-타겟 단백질 상호작용(drug-target interaction, DTI) 및 항암제 반응성(cancer drug response, CDR)을 위한 인공지능 모델 개발에 도움을 주고 있다.

다중채널 모델은 다양한 분야에서 높은 성능을 보이고 있으며,

모델의 유연성이 높은 신경망 구조를 기반으로 구축되고 있다. 예를 들어서, 컴퓨터 비전 분야에서 사진은 빨강, 초록, 파랑의 3가지 레이어를 가진 RGB 채널 데이터로 표현된다. 다중 채널 데이터의 통합을 위해서 컨볼루션 신경망(convolution neural networks, CNN)은 각 픽셀에 해당하는 모든 채널 데이터를 동시에 학습한다. 유사한 기법으로 다중 양식 학습(multimodal learning)이 있으나, 다중 채널은 동일한 데이터를 여러 채널들을 통해 다양한 특질로 표현하는 반면, 다중 양식(multi-modality)는 영상, 음성, 텍스트와 같은 서로 다른 양식의 데이터를 기반으로 한다는 점에서 차이가 있다.

1.2 연구의 목적

본 연구는 의생명 데이터를 다양한 채널의 특질으로 추출하고 각 특질을 상보적으로 학습하는 인공지능 구조를 설계하는 것을 목적으로 한다. 이를 위해서 다중채널 기반의 약물-표적 상호작용 모델과 약물 반응성 모델을 제안한다. 첫째로, 약물-표적 상호작용 모델은 약물 후보물질과 표적 단백질을 다중 채널로 구축하기 위해서 다양한 특질 추출 모듈을 활용한다. 특질 추출 모듈들이 입력 데이터에서 추출한 상보적인 특질을 통합하여 약물-표적 상호작용을 학습한다. 두번째로, 약물 반응성 모델은 세포주 데이터가 가진 다양한 생물학적 층위의 오믹스 데이터를 활용한다. 다층적인 오믹스 데이터를 다중 채널 특질로 구축하여 통합하고, 이를 기반으로 약물의 반응성을 학습한다.

이에 더해서, 본 연구에서 구축하는 두 모델들은 신약개발 프로세스의 관점에서도 상보적인 관계를 갖는다. 그림 2에서 볼 수 있듯이, 약물 반응성 모델은 약물 반응성 모델에 비해 선행적으로 활용되며, 단일한 타겟 단백질과 약물 후보물질의 반응성에 집중한다. 반면에 약물 반응성 시험은 세포주 및 동물 실험을 통해서 시행되므로, 약물이 세포 내의 많은 유전자들과 단백질들에 미치는 약물 반응성을 총체적으로 측정할 수 있다.

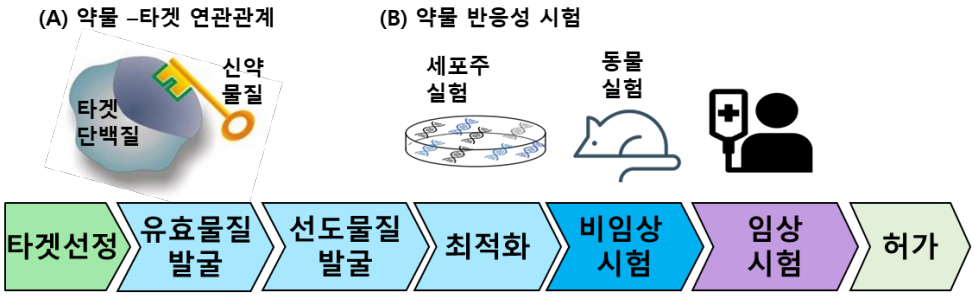

그림 2. 신약 탐색 프로세스 개요

따라서, 본 연구에서는 의도한 타겟(on-target)의 약물 반응성 뿐만 아니라, 의도하지 않은 타겟(off-target)의 약물 반응성을 분석한다. 이를 통해서 잠재적인 약물의 타겟 및 생물학적 크로스토크(cross talk)를 탐색할 수 있다. 특히, 암 세포의 유전자는 그 돌연변이로 인하여 단일한 타겟을 규정하는 것은 어렵기에, 새로운 타겟 및 다중 타겟 탐색은 항암치료에 도움이 될 것으로 기대된다.

1.2.1 약물-타겟 단백질 상호작용 예측

약물-타겟 단백질 상호작용(drug-target interaction, DTI)식별은 신약후보 선도물질 발견, 약물 용도 변경, 가능한 오프 타겟 또는 부작용 설명과 같은 약물 개발의 다양한 응용 분야에서 중요한 역할을 한다[5]-[7]. 그러나 DTI 검출을 위한 전통적인 생물학적 실험은 일반적으로 비용과 시간이 많이 소요된다. 예를 들어서, 현재 합성 가능한 신약 잠재 물질들은 약 10^8 개[8], [9]이지만, 그들은 약물 후보물질의 작은 일부분이다. 이론적으로 약물 후보물질의 숫자는 약 10^{24}에서 10^{60}개에 이른다[9], [10]. 단백질의 탐색 공간으로는 약 20만개의 검증된 인간의 단백질이 있다[11]. 시험관 내 실험(*in vitro*) 모델은 약물과 타겟 단백질의 연결성을 밝히기 위해서 사용되는 일반적인 방법이지만, 앞서 설명한 DTI의 탐색 영역을 실험적인 기법으로 모두 식별하는 것은 비용과 소요 시간의 측면에서 rm 실현가능성이 매우 낮다. 따라서, 컴퓨터 시뮬레이션에 기반한 *in silico* 모델은 실험의 탐색 범위를 줄이거나 높은 잠재력을 가진 약물 후보물질의 우선순위를 정의하는 방법으로 전통적인 실험기반의 기법을 보조하기 위해 활용된다[12]-[15].

일반적인 in silico 모델은 구조 기반 방법[16]-[18]과 리간드 기반 방법[19]-[21]이 있다. 이에 더해서 단일 모델 내에서 리간드와 표적 단백질의 데이터를 모두 통합하여 약물과 타겟 단백질의 연결성을 예측하는 단백질 화학 측정(proteochemometrics, PCM) 방법[22]-[25]이 제안되었다. PCM 방법은 최근 다양한 작업들에서 성능을

입증했다. 예를 들어서, 새로운 약물 조합의 식별[26], 약물과 표적 간의 상호작용 예측[27] 그리고, G 단백질 결합 수용체(GPCR) 및 단백질 키나제 표적에 대한 DTI 예측[28]에서 PCM 방법의 가능성을 보여주었다. 또한 PCM은 다양한 다종의 정보를 단일 모델로 활용할 수 있는 잠재력[22], [29]을 가지고 있기에 DTI 문제를 해결하는 방법[24]으로 주목을 받고 있다. 따라서 본 연구에서는 PCM 기법을 기반으로 다중채널 모델을 구축하고 DTI 예측 성능을 높이는 것을 연구의 세부 목표로 수행한다.

1.2.2 다중 오믹스 기반의 항암제 반응성 예측

암 환자 마다 상이한 항암제의 효능과 이질성은 항암제 신약 개발에서 해결해야할 도전적인 과제이다 [30]. 예를 들어서, 유전체, 전사체 및 후성 유전체 전반에 걸친 종양 이질성(tumor heterogeneity)은 항암제 치료 효능을 손상시킬 수 있다[31]-[35]. 따라서, 이를 극복하는 것은 항암 후보물질에 대한 환자의 임상적 이질성을 해결하는 것에 큰 영향을 미칠 수 있다. 고효율 시퀀싱(HTS) 기술의 출현은 악성 종양의 발생과 진행에 영향을 미치는 다양한 생물학적인 측면을 다양한 관점에서 측정할 수 있게 되었다[36]. 각 유형의 다중 오믹스 데이터는 특정한 생물학적 측면의 관점에서 전체적인 생물학적 시스템에 대한 추가적인 정보를 제공한다[37].

다중 오믹스 프로파일은 다양한 유형의 데이터로 구성된다.

유전체학 및 전사체학은 암 연구 및 임상 적용을 위해 광범위하게 연구된 오믹스 데이터이다. 인간 발암에 대한 대부분의 연구는 체세포 돌연변이에 초점을 맞춰 왔다[38]. 체세포 돌연변이는 체세포의 모든 변화로 설명된다. 이는 종양 유전자, 종양 억제 유전자 및 DNA 복구 시스템이 종양 생존을 향상시키기 위해 변경될 수 있기 때문이다[39]. 유전자 복제수 변이는 게놈의 섹션이 반복되는 구조적 변이 유형이다[40]. 여러 연구에서 특정 유전자의 복제수 변이가 다양한 유형의 암의 발생, 발달 및 진행에 역할을 하는 것으로 나타났다[40]. 유전자 발현 데이터는 유전자가 단백질 또는 RNA 구조로 번역되는 정보이다. 암 유전자는 변이된 발현에 의해 확인되며, 이는 악성 종양의 상당 부분에서 비정상적인 표현형을 유발한다[41]. 체세포 돌연변이, 유전자 복제수 변이 및 유전자 발현 데이터는 암 연구를 위해 분석된 다중 오믹스 프로파일의 대표적인 사례이다. 따라서 본 연구에서는 다중 오믹스 프로파일을 기반으로 다중채널 모델을 구축하여 항암제 반응성(cancer drug response, CDR)의 예측 성능을 높이는 것을 연구의 세부 목표로 수행한다.

1.3 연구의 구성

본 연구는 다음과 같이 구성되어 있다. 2장에서는 약물-타겟 단백질 상호작용 예측을 위한 쌍기반 다중채널(Multi-channel Pairwise Input

Neural Network, MCPINN) 모델을 제안한다. MCPINN은 특질의 다양한 표현형을 다중 채널에 입력하여 활용하고 그 특질을 상보적으로 통합하여 모델의 성능을 높인다. 모델의 성능을 평가하기 위해서 생체활성도에 기반한 DTI 예측 성능 및 학습 속도를 측정했다. 이에 더해서 본 연구는 전이학습을 통해서 일반화된 정보를 추출하고, 이를 활용하여 약물 후보물질의 독성 예측 성능을 높이는 방안을 살핀다.

3장에서는 다중 오믹스 기반의 항암제 반응성 예측을 위한 유전자 중심의 다중채널(Gene-centric multi-channel, GCMC) 모델을 제안한다. GCMC는 CDR 예측 모델 중 최초로 유전자의 순서에 불변성(order-invariant)을 가진 컨볼루션 신경망 구조이다. GCMC는 다중 오믹스 프로파일을 3차원의 표현형으로 변환하고 컨볼루션 인코더를 사용하여 유전자 수준에서 다중 오믹스 특질을 통합한다. GCMC를 평가하기 위해서 세포주 데이터뿐만 아니라 암환자 및 환자유래 암조직 이종이식동물모델 데이터를 사용했다. 이에 더해서, 설명가능한 인공지능(explainable AI, XAI) 알고리즘을 통해서 예측 모델의 입력 데이터 중에서 어떤 오믹스 유형이 성능에 가장 많이 기여하는지 분석한다. 이것은 CDR의 예측 결과를 오믹스 유형의 비율로 분석한 첫 번째 연구이다. 마지막으로 4장에서는 본 연구를 종합하여 결론을 제시한다.

II. 다중 채널 기반의 쌍 입력 신경망(MCPINN)

2.1 서론

약물-타겟 단백질 상호작용(drug-target interaction, DTI)을 분석하는 것은 신약 탐색과 신약 재창출을 위한 필요조건으로써 그 중요성이 점차 증가되고 있다[5]-[7]. 그러나 신약 화합물과 타겟 단백질의 가능한 모든 경우의 수를 모두 탐색하는 것은 매우 어렵고 비용 집약적인 작업이다. 이 탐색 공간은 넓고 이질적이며, 무엇보다도 약물과 타겟 단백질의 연결성은 아직 탐색되지 않은 영역이 매우 넓다. 예를 들어서, 현재 합성 가능한 신약 잠재 물질들은 약 10^8 개[8], [9]이지만, 그들은 약물 후보물질의 작은 일부분에 불과하다. 이론적으로 약물 후보물질의 숫자는 약 10^{24}에서 10^{60}개에 이르기 때문이다[9], [10]. 또한, 단백질은 약 20만개의 검증된 인간의 단백질이 있다[11]. 시험관 내 실험(*in vitro*) 모델은 약물과 타겟 단백질의 연결성을 밝히기 위해서 사용되는 일반적인 방법이지만, 앞서 설명한 DTI 탐색 공간을 전통적인 실험법으로 식별하는 것은 그 실현가능성이 매우 낮다. 컴퓨터 시뮬레이션에 기반한 *in silico* 모델은 높은 잠재력을 가진 약물 후보물질의 우선순위를 정의하거나 실제 실험의 탐색 범위를 줄이는 방법으로 전통적인 실험 기법을 보조하기 위해 활용되고 있다[12]-[15].

일반적인 *in silico* 모델은 구조 기반 방법[16]-[18]과 리간드 기반

방법[19]-[21]이 있다. 이에 더해서 단일 모델 내에서 리간드와 표적 단백질의 데이터를 모두 통합하여 약물과 타겟 단백질의 연결성을 예측하는 단백질 화학 측정(proteochemometrics, PCM) 기법[22]-[25]이 제안되었다. 첫째, 구조 기반 방법은 합리적인 예측 성능과 시각적으로 해석 가능한 결과를 제공한다. 구조 기반 방법은 분자 도킹에 3차원 시뮬레이션을 사용하여 DTI를 발견한다. 예를 들어서, AutoDock[42], Glide[43], Fred[44] 및 AtomNet[45]은 도킹 기법을 활용한 모델들이다. 그러나 이 방법에는 두 가지 주요 제한 사항이 있다. 매우 복잡하고 높은 계산량과 부족한 숫자의 3차원 구조 데이터가 그 한계이다. 따라서 리간드 기반의 기법과 PCM 기법이 대부분의 경우에서 선호된다.

둘째, 리간드 기반 방법은 분자 유사성(molecular similarity) 원리[46]을 기본 가정으로 삼는다. 이 가정에 따르면 유사한 화합물은 유사한 단백질과 상호작용하기 위해 사용되며, QSAR(Quantitative Structure-Activity Relationship) 모델이 이 가설의 대표적인 사례 중 하나이다. 이 접근 방법에 기반하야 다양한 기계학습(machine learning, ML) 알고리즘이 활용되었다. 예를 들어서, 전통적인 기계학습 모델인 나이브 베이즈[47], [48], 랜덤 포레스트[49], 서포트 벡터 머신[50]에서부터 최근 높은 성능을 보이고 있는 딥러닝 모델[51]과 이에 기반한 다중 작업 학습(multi-task learning)[52], [53]도 높은 성능을 보이고 있다. 그러나 이 접근법은 단일적인 분자 활성 정보만 활용하는 반면에 생물학적인 정보는 활용하지 않는다는 한계가 있다.

마지막으로, 리간드 기반 방법과 달리 PCM 기법은 각 화합물과 단백질 쌍을 입력 데이터로 사용하여 모델을 구축하여 단백질 및 분자 공간을 모두 활용한다. PCM 기법은 최근 다양한 작업들에서 성능을 입증했다. 예를 들어서, 새로운 약물 조합의 식별[26], 약물과 표적 간의 상호작용 예측[27] 그리고, G 단백질 결합 수용체(GPCR) 및 단백질 키나제 표적에 대한 DTI 예측[28]에서 PCM 방법의 가능성을 보여주었다. 또한 PCM은 다양한 다종의 정보를 단일 모델로 활용할 수 있는 잠재력[22], [29]을 가지고 있다. 따라서 PCM 방법은 DTI 문제를 높은 성능으로 해결할 수 있는 방법[24]으로 주목을 받았다.

최근 DNN 알고리즘은 DTI 쌍을 예측하기 위해 적용되고 있으며 랜덤 포레스트, 나이브 베이즈 및 서포트 벡터머신[54], [55]과 같은 전통적인 분류기보다 우수한 성능을 보이고 있다. 순방향 신경망(FFNN)과 같은 기본적인 구조를 사용하는 것 외에도 이전 연구[56]에서는 쌍입력 신경망(PINN)을 제안했다. PINN은 순방향 신경망의 변형이며 분리 레이어와 통합 레이어로 구성된다. 각 분리 레이어에는 각기 다른 특질(즉, 화합물 및 단백질)이 입력되고 각 레이어는 독립적으로 구성되며, 추후 통합 레이어에서 각기 학습된 특질이 통합된다. 이 구조를 통해 PINN은 성능 저하 없이 기존 순방향 신경망에 비해 네트워크의 총 매개변수 수를 약 50% 줄일 수 있다. 이 구조는 화합물 및 단백질 특질을 모두 활용하는 PCM 방법에도 적합하다. 그러나 일반적으로 대부분의 DNN은 각 훈련 대상에 대해 상당한 양의 데이터를 필요로 한다. 공개적으로 사용 가능한 DTI

데이터의 수가 빠르게 증가했지만 DTI의 탐색 공간을 모델링하는 데는 여전히 충분하지 않다[3].

게다가 최근 Lenselink의 연구[54]에서 지적했듯이 공개 데이터는 서로 다른 과학적 프로토콜을 사용하기 때문에 많은 오류가 있을 수 있다. 저자들은 고품질 벤치마크 데이터 세트를 제시하고 다양한 기계학습 알고리즘 및 검증 분할의 다양한 조합 간의 성능을 비교했다. 연구에 따르면 PCM 모델은 일반적으로 동일한 조건에서 QSAR 모델을 능가했다. PCM 기반 DNN 알고리즘은 두 평가 세트(임시 검증 및 무작위 검증)에서 다른 모델보다 성능이 뛰어났다.

PCM 기반의 DNN 모델은 DTI의 탐색 공간을 모델링[24]을 하기 위해 표현 학습(representation learning) 측면에서 여전히 개선의 여지가 남아있다. 왜냐하면 DNN은 분류기, 특질 추출기[57], [58], 종단 간(end-to-end) 학습 모델을 포함한 세 가지 접근 방식으로 활용될 수 있기 때문이다. 첫째로, 약물 발견을 위한 DNN 분류 모델은 일반적으로 도메인 전문가에게 추출된 특질을 기반으로 생체 활성 여부를 예측한다. 둘째로, DNN은 대규모 데이터 베이스의 정보를 활용하는 특질 추출기로도 사용할 수 있다. 예를 들어서, 화합물은 ZINC[59] 데이터 베이스를 활용하고 단백질은 UniProt[60] 데이터 베이스를 활용하여 학습 데이터 보다 더 많은 데이터를 활용한 특질 추출기를 구축할 수 있다. 셋째로, 종단 간 학습(end-to-end learning) 모델로서 DNN은 SMILES 화합물 문자열 및 단백질 아미노산 서열과 같은 원본 데이터에서 특질을 학습할 수 있다. 이 모델은 특질 추출에서

분류에 이르는 학습 프로세스를 단일 학습 모델이 시작부터 끝까지 관리할 수 있다.

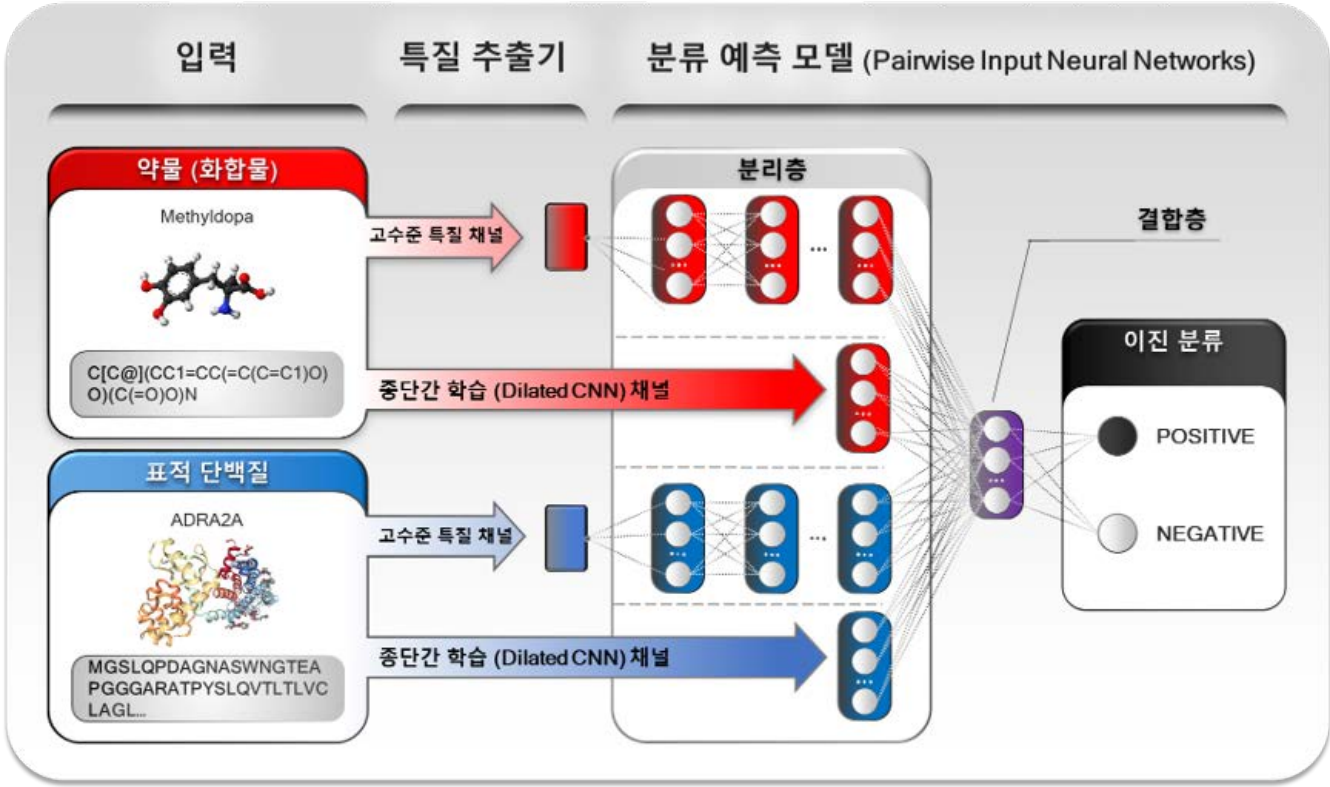

그림 3. 다중채널 기반의 쌍 입력 신경망(MCPINN)의 개요

본 논문에서는 다중 채널 기반의 쌍 입력 신경망(Multi-channel Pairwise Input Neural Network, MCPINN)을 제안한다. MCPINN은 PCM 기반의 신경망 모델이며, 주어진 데이터를 최대한 활용하기 위해 DNN의 세가지 접근 방식인 분류기, 특질 추출기 및 종단 간 학습기의 접근 방식을 모두 사용한다. MCPINN은 다양한 특질의 조합을 활용하여 이를 단일 모델에서 상보적으로 통합한다. 예를 들어서, 약물 데이터는 SMILES 문자열, ECFPs[61], 그리고 Mol2vec[57]으로 추출된 특질로 표현되고, 단백질 데이터는 아미노산 서열과 ProtVec[58]으로 추출된 특질로 표현된다. MCPINN은 추출된 화합물과 단백질의 다양한 특질들을 학습하고 DTI의 탐색 공간을 모델링한다. 이에 더해서, 본 연구는 전이학습(transfer learning)을 적용하여 MCPINN의 특질학습 능력과 일반화 성능을 탐구한다. MCPINN은 양성과 음성의 비율이 균형 잡힌 생체활성(bioactivity) 데이터에서 일반화된 정보를 추출하여 불균형한 독성 데이터 세트에 그 정보를 적용한다.

MCPINN의 성능을 평가하기 위해서 MCC 및 ROC를 평가 척도로 사용하여 ChEMBL에서 얻은 표준화된 벤치마크 데이터 세트[54]를 사용했다. 먼저 각 특질의 효과를 조사하기 위해 MCPINN은 단일 채널 특질 쌍의 6개 조합으로 평가되었다. 이에 더해서 다양한 수준의 특질 표현형의 시너지 효과를 탐색하기 위해 MCPINN은 9개의 특질 조합으로 성능이 평가되었다. 모델은 최고 성능뿐만 아니라 초기 성능 및 수렴 속도 측면에서도 평가되었다. 또한, MCPINN은 전이학습을 통해서 화합물 및 단백질의 일반화된 특질을 새로운 작업으로 전송할 수

있는지 테스트되었다. 이를 위해서 MCPINN을 생체활성 벤치마크 데이터 세트에서 사전 훈련시킨 다음 Tox21 데이터 세트[62]에서 사전 훈련된 모델을 미세 조정했다. MCPINN의 일반화 정보의 적용 가능성은 초기 성능, 수렴 속도[63], [64] 및 최고 성능 측면에서 매튜 상관계수(Matthews Correlation Coefficient, MCC)와 정밀도-재현율 곡선 (Precision Recall Curve, PRC) 지표를 통해서 평가했다. 따라서 본 연구는 분류기, 특질 추출기, 종단 간 학습기로서의 DNN의 특질 추출 능력을 최대한 활용하고 이를 전이학습에도 적용함으로써, DTI 모델링[24]을 탐색한다.

2.2 연구 방법

2.2.1 데이터 및 전처리

생체 활성(bioactivity) 벤치마크 데이터 세트로 Lenselink의 연구[54]에서 개발한 고품질 데이터 세트를 사용했다. 이 벤치마크 데이터 세트에서는 15개의 특질 조합에 기반한 MCPINN 모델들이 평가된다. 데이터세트는 ChEMBL에서 조합 가능한 총 생체 활성 값의 0.13%를 포함한다. ChEMBL은 204,085개의 화합물과 1,227개의 단백질 표적에 의해 총 250,412,295개의 조합 가능한 데이터 포인트가 있다. 그 중에서 벤치마크 데이터셋에 314,767개의 관찰 값이 포함되어있으며, 이진 값으로 구성된 데이터의 양성 비율은 54.7%이다.

전이 학습을 위한 독성 데이터는 Tox21이며 12개의 다른 대상에서 8,014개의 화합물에 대한 79,585개의 측정값이 있다. Tox21 데이터 세트의 양성 비율은 7.5%으로 양성과 음성의 데이터 불균형이 심하다.

본 연구에서는 특질 추출을 위하여 단백질과 약물 후보물질의 데이터 세트를 구축했다. 첫째, 단백질을 위한 데이터 세트를 위해서 UniProt과 ChEMBL 데이터 베이스[3]를 사용하였다. 모든 중복된 단백질은 삭제하였으며 결과적으로 553,195개의 단백질과 그 아미노산 서열 데이터를 추출했다. 둘째, 약물 후보물질을 위한 데이터 세트를 위해서 ZINC와 ChEMBL에 포함된 약 19,900,000 개의 약물 후보 물질을 사용하였다. 이 물질들은 Mol2Vec 모델에서 제안한 방법으로 전처리 되었으며 SMILES 서열로 표현되었다. 아미노산 서열과 SMILES 서열은 종단간 학습 모델에 입력되어 저수준 표현형 특질로 변환된다.

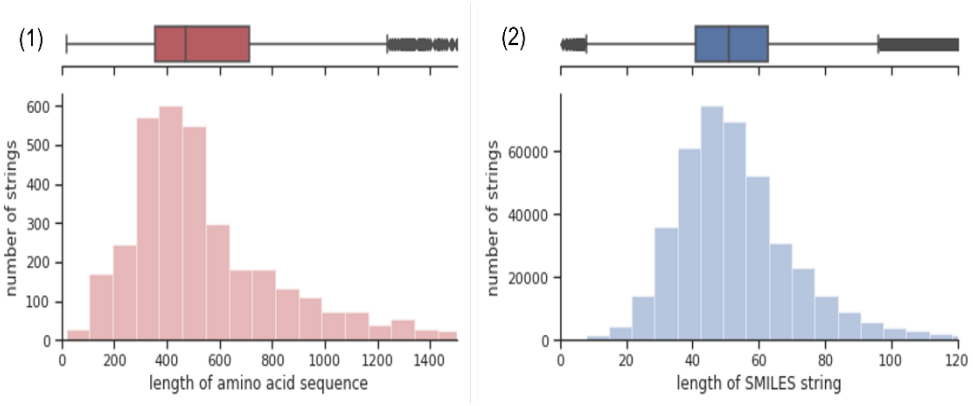

그림 4. 아미노산 서열과 SMILES 서열의 길이 분포

약물과 단백질의 아미노산 서열은 문자 단위로 토큰화하고 고정

길이의 원 핫 (one-hot) 이진 벡터로 인코딩되었다. 예를 들어서, 단백질은 사전 연구[65]를 따라서 아미노산 서열을 24개의 문자로 변환했고 약물은 SMILES 서열을 57개의 문자로 변환했다. 이 문자열들은 대응되는 문자에 1을 할당하고 나머지는 0을 할당하는 원 핫 인코딩으로 변환되었다. 변환된 서열들이 기계학습모델의 입력으로 사용되기 위해서 서열 데이터를 사후 절단 또는 제로 패딩(zero padding)을 사용하여 고정 길이로 변환한다. 해당 서열이 고정 길이보다 길면 그 이상의 서열을 제거하고, 반대로 고정길이보다 짧으면 시퀀스 끝에서 고정 길이까지 0으로 값을 채워 넣는다. 이때, 서열의 고정 길이를 결정할 때 정보 보존과 계산 효율성 사이에는 절충점이 필요하다. 그림 4과 같이 SMILES 문자열에 대한 75%의 백분위수는 63이고 아미노산 서열의 경우 75%의 백분위수는 712이기에 본 연구에서는 화합물에 대해 고정 길이 100, 단백질에 대해 700을 선택했다.

2.2.2 고수준 특질 추출 및 후처리

본 연구에서는 화합물과 단백질의 고수준 표현형의 특질 추출을 위해서 Mol2vec[57]과 ProtVec[58]을 활용했다. 또한, 추출된 특질에 후처리 과정을 적용해서 기존의 기법들이 반영하지 못한 단백질과 약물 데이터의 특질을 적용시켰다. 첫째, ProtVec과 Mol2vec은 자연어처리기법 (NLP)에 기반한 비지도 학습 기반의 특질 추출 모델이다. 이 알고리즘들은 Word2Vec 라는 기법에 기반한다.

자연어처리 모델의 관점에서 단백질과 약물은 문장으로 간주되며 그 하위구조는 단어로 간주된다. 예를 들어서, 약물은 Morgan 알고리즘을 활용하여 2차원 그래프를 더 작은 하위 그래프로 나눌 수 있으며, 단백질은 N-gram을 사용하여 N개의 아미노산 서열의 집합으로 표현할 수 있다. 동일한 관점에서 ZINC[59], ChEMBL[3] 및 UniProt[60]와 같은 대규모 데이터베이스는 말뭉치(corpus) 데이터 셋으로 간주된다.

둘째, 단백질 특질 추출 모델은 Skip-gram 모델[66]을 기반으로 하며 특질 차원은 300, 창 크기는 35, 최소 개수는 2의 하이퍼 파라미터로 학습된다. Tox21에서 SR-MMP 에는 아미노산 서열이 없으므로 벡터 값은 0으로 처리되었다. 약물 특질 추출 모델도 Skip-gram 모델을 기반으로 구축되었으며, Mol2vec의 저자가 제안한 사전 훈련 모델을 사용했다. 이에 더해서, 화학정보학에서 전통적으로 사용하는 기법인 ECFP으로도 특질을 추출하여, Mol2vec의 상대적인 성능 차이를 측정하였다.

마지막으로, 문장의 특질 벡터는 각 단어의 특질 벡터들의 구성으로 표현된다. 본 연구에서는 산술 평균과 TF-IDF 기법[67]을 활용한 가중 평균법을 제안하여 기존의 단순 합계 기법을 개선한다. 기존의 방법은 단어 특질 벡터의 단순 합계이며, 이를 수식으로 표현하면 다음과 같다.

$$S^{sum} = \sum_{i=1}^{N} w_i$$

S^{sum}은 문장 특질이고 N은 해당 문장 내의 단어 숫자, 그리고 w_i는

i번째 단어의 특질 벡터이다. 기존의 기법은 문장 내의 단어의 수가 많을수록 합계 연산이 증가하므로, 문장 특질 벡터 값이 단어의 수에 비례적으로 증가한다는 단점이 있다. 이것은 한 문장 내의 단어의 수가 20개 내외라는 자연어처리모델의 가정을 그대로 사용했기 때문에 생긴 한계점이다.

이와 같은 문제점을 해결하기 위해서 산술평균과 TF-IDF 가중 평균 기법을 제안한다. 산술평균기법을 수식으로 표현하면 다음과 같다.

$$S^{mean} = \frac{1}{N}\sum_{i=1}^{N} w_i$$

이 기법은 각 단어의 가중치를 문장 내의 단어 숫자의 역수로 정의한다. 따라서 한 문장 내의 모든 단어들은 모두 같은 가중치를 갖게 된다. 따라서 같은 단어의 가중치가 소속된 문장에 따라서 다른 가중치를 가질 수 있다는 특징을 갖는다.

반면에 TF-IDF 가중 평균 기법은 각 단어의 가중치가 모든 문장에서 동일하다는 특징이 있다. 이를 수식으로 표현하면 다음과 같다.

$$S^{tf-idf} = \sum_{i=1}^{N} t_{w\_i} w_i$$

t_{w_i}는 단어 w_i의 가중치 값이다. 이 가중치 값은 TF-IDF 알고리즘을 통해서 추출되며, 문장의 단어 수에 상관없이 모든 문장에서 동일한 가중치를 갖는다.

2.2.3 MCPINN 모델 설계

본 연구는 MCPINN 모델은 쌍입력신경망(PINN) 구조에 기반하여 단백질과 약물의 특질의 다양한 수준의 표현형을 학습하도록 설계되었다. 첫째, PINN은 FFNN의 변형이며 단백질과 약물의 특질을 모두 활용하는 PCM 기법에 적합하다. PINN은 두 개의 각 입력층이 분리층으로 이어지고 이후 분리층이 결합층으로 합쳐지는 구조이다. 각각의 분리층들은 다른 특질이 입력된 분리층과의 연결없이 독립적으로 구성되어 있으며, 각 입력 채널층은 독립적으로 특질을 학습하고 고도화할 수 있다. 또한 이 구조는 마지막으로 분리된 레이어의 노드 수를 제어하여 각 특질의 비율을 균형 있게 조정할 수 있다. 예를 들어, ECFP의 입력 차원이 1024이고 ProtVec의 입력 차원이 300으로 다를 수 있다. FFNN은 차원이 높은 특질 벡터에 편향된 학습이 진행될 수 있으나, PINN에서는 분리층의 노드 개수를 조절하여서 특질의 차원을 동일하게 통일할 수 있다.

둘째, 저수준 특질 채널을 위한 종단 간 학습 모델은 확장된 컨볼루션 신경망 (Dilated CNN)[68]을 사용하였다. Dilated CNN은 입력 데이터 및 연산은 동일하지만 더 넓은 범위를 학습할 수 있다. 이 기법은 기존의 커널이 입력 받는 범위 사이에 간격을 만들어서 보다 넓은 수용 범위를 갖도록 했다. 이를 통해서 단백질과 약물의 서열 데이터의 장기적인 종속성을 효율적으로 학습할 수 있다. 기존 연구에서는 작은 커널 크기(예. 3)를 추천하였다[69]. 그러나 본 연구의

고수준 특질 추출 단계에서 이미 좁은 범위의 특질을 추출하였기에 커널의 크기를 상대적으로 큰 크기(예. 12)로 사용했다. 그 외의 하이퍼 파라미터로 필터의 수는 16, 기본 임베딩의 차원은 16, 그리고 유효패딩(valid padding)을 사용하여서 모델의 중간 특질의 차원을 줄였다. Dilated CNN 이외에도 RNN 기반의 LSTM[70]과 BLSTM[71]도 실험되었으나 Dilated CNN이 가장 높은 성능을 보였기에 본문에서는 생략한다.

셋째, 고수준 특질 채널은 FFNN으로 구성되었으며 과적합을 방지하기 위해서 dropout이 정규화 기법으로 적용되었다. 초기층에는 10%과 이후 은닉층에서는 50%의 비율로 drop이 적용된다. 또한, 고수준 특질을 안정적으로 학습하기 위하여 특질의 평균을 0으로 단위 분산을 1로 정규화했다. 이는 딥러닝을 포함한 구배(gradient) 기반의 알고리즘들은 입력 데이터가 일반적으로 표준화된 데이터라는 가정으로 설계되는 경우가 많기 때문이다. 고수준 특질 채널의 최적화를 위해서 레이어와 노드의 개수를 포함한 파라미터를 최적화했다. 예를 들어서, 분리층은 2이며 각 층의 노드 수는 1024와 256이다. 결합층은 1이며 노드는 256, 그리고 drop 비율은 0.5이다.

마지막으로, 모델의 최적화를 위한 활성함수와 기법들을 소개한다. 테스트된 모든 모델은 최적의 네트워크 구성을 찾기 위해 조기 중지(early stopping) 기법과 함께 400 epoch 동안 검증 데이터(훈련 데이터의 20%)에 대한 5중 교차 검증으로 검증되었다. 이후, 전체 학습 데이터에 대해 최적의 모델을 학습하고 테스트 데이터에 대해 평가했다.

Adam은 일반적으로 학습율의 크기에 민감하지 않기 때문에 효율적이고 빠른 훈련 성능으로 인해 DNN에서 사용된다. β_1은 0.9, β_2은 0.999를 제안된 연구[72]에 따라서 사용했다. 벤치마크 데이터 세트가 매우 희박하기 때문에 작은 배치 크기는 학습 모델을 지역 최적점(local minima)로 오도할 수 있다. 따라서 일반화된 성능을 위해 사전 연구[73]에서 제안한 대로 미니 배치 크기를 1,024로 설정했다. 모든 가중치와 편향 값은 Lecun 균일 분포로 초기화 했으며, 이는 효율적인 역전파 계산으로 이어진다고 알려져 있다[74]. ReLU 활성함수는 훈련 속도가 빠르기 때문에 DNN에서 일반적으로 사용되지만, 음수 값을 무시한다는 정보손실이 있다[75]. 따라서 지수 선형 단위(ELU) 활성함수[76]를 사용했다.

2.2.4 전이학습 파이프라인

전이 학습은 기계 학습 모델이 훈련 작업에서 다른 관련 테스트 작업으로 일반화된 특질을 전송할 수 있는지 여부에 중점을 둔다. 미세 조정 방법에 영향을 미치는 몇 가지 요소가 있지만 일반적으로 두 가지 중요한 요소가 고려된다[63], [64]. 그 요소들은 테스트 작업 데이터 세트의 크기와 유사성(예, 데이터의 내용 및 균형) 이다. 이 요소의 조합에 의해서 사전 훈련된 모델을 미세 조정하는 네 가지 기본 전략이 있다. 첫째, 테스트 데이터 셋이 크고 그 데이터의 유사성이 사전 학습 데이터와 매우 유사한 경우이다. 이 경우에는 과적합의 위험이 낮기

때문에 모델 전체를 미세 조정할 수 있다. 이러한 경우에는 전이학습을 통해서 유망한 성과를 보여줄 것으로 기대된다.

둘째, 테스트 데이터 셋이 크고 그 데이터의 유사성이 사전 학습 데이터와 매우 다른 경우이다. 이 경우에는 사전 훈련의 횟수를 줄이고 전체 모델을 미세 조정하거나, 사전 훈련없이 테스트 데이터 셋으로만 학습을 진행하는 방법이 있다. 실제적인 전략으로는 전체적인 학습 시간을 줄이기 위해서 전자의 방법을 하는 것이 추천된다.

셋째, 테스트 데이터 셋이 작고 그 데이터의 유사성이 사전 학습 데이터와 매우 유사한 경우이다. 과적합의 위험 때문에 모델의 전체 네트워크를 미세 조정하는 것은 권장되지 않는다. 대신에 모델의 분류기에 해당되는 일부분의 네트워크만 미세 조정하고 그 외의 네트워크는 학습하지 않는 전략이 추천된다.

넷째, 테스트 데이터 셋이 작고 그 데이터의 유사성이 사전 학습 데이터와 매우 다른 경우이다. 이 경우에는 모델의 전체 네트워크를 미세 조정하는 것이 권장되지 않을 뿐만 아니라, 모델의 분류기만 미세 조정하는 것도 권장되지 않는다. 따라서, 상위 레이어를 초기화하고 그 외의 레이어를 학습에서 배제하는 전략이 사용된다. 이를 통해서 사전 학습 데이터에서 추출된 일반적인 정보를 활용하면서도 그 정보를 테스트 데이터에 알맞게 적용하는 것이 가능해진다.

본 연구는 네번째 시나리오에 해당되는 테스트 및 사전학습 데이터로 구성되어 있다. 테스트 데이터인 Tox21은 사전학습 데이터보다 데이터의 크기가 1/4 수준으로 작고, 데이터의 종류와

양성-음성의 비율도 매우 다르기 때문이다. 예를 들어서, Tox21의 독성 데이터는 생리학적인 측면에서 다뤄지지만, 사전학습 데이터의 생체 반응 데이터는 생물 물리학적인 수준에서 다뤄진다[77]. Tox21의 양성비율은 7.5% 인 반면에 사전학습 데이터는 약 54.7%이다. 최적의 사전훈련 및 미세 조정 횟수를 찾아내기 위해서 5회의 사전훈련마다 미세 조정이 진행되었다.

2.3 연구 결과 및 고찰

모델의 예측 성능에 대한 각 특질의 기여도를 파악하기 위해 MCPINN은 단일채널모델의 6종류와 다중채널모델 9종류를 포함하는 15개의 특질 조합으로 평가되었다. 각 모델들은 단백질과 약물 특질들의 조합을 기반으로 하며, 표 1에 단축된 이름과 특질의 조합이 나열되어 있다. 약물을 위한 다중채널 모델 중 Mol2vec과 ECFP 모델의 조합은 성능의 향상이 없었기에 본문에서는 생략되었다.

표 1. 채널별 특질 조합에 따르는 모델명

모델명	채널 타입	약물 특질	단백질 특질
SC_1	단일채널	SMILES 서열	아미노산 서열
SC_2			ProtVec
SC_3		Mol2vec	아미노산 서열

SC_4			ProtVec
SC_5		ECFP	아미노산 서열
SC_6			ProtVec
MC_1	다중채널 (단백질)	SMILES 서열	아미노산 서열과 Protvec
MC_2		Mol2vec	
MC_3		ECFP	
MC_4	다중채널 (약물)	SMILES 서열과 Mol2vec	아미노산 서열
MC_5			ProtVec
MC_6		SMILES 서열과 ECFP	아미노산 서열
MC_7			ProtVec
MC_8	다중채널 (단백질과 약물)	SMILES 서열과 Mol2vec	아미노산 서열과 Protvec
MC_9		SMILES 서열과 ECFP	아미노산 서열과 Protvec

2.3.1 단일 채널 모델 및 특질 추출 개선

저수준 및 고수준의 조합을 포함한 단일채널모델 (SCPINN)의 성능을 비교하고 이를 개선한 기법의 성능도 평가한다. 단일채널모델은 단백질과 화합물 특질을 한 가지만 사용하는 전통적인 구조의 모델이다. 단백질 특질은 아미노산 서열과 ProtVec 특질이 사용되었고, 약물

특질은 SMILES 서열, Mol2vec, 그리고 ECFP 특질이 사용되었다. 단일채널모델들의 평균 성능은 MCC가 0.636이고 ROC가 0.892이다. 단일채널모델 간의 성능에서 가장 큰 차이는 약물 특질에서 저수준 표현형 특질 (SMILES 서열) 대신 고수준 표현형 특질 (ECFP 및 Mol2vec)을 사용한 것으로 관찰되었다. 예를 들어, 화합물에 대해 ECFP 및 Mol2vec를 사용하는 모델의 평균 성능은 MCC가 0.66 이며 ROC는 0.90인 반면 SMILES 서열을 사용하는 모델의 평균 성능은 MCC가 0.60이고 ROC는 0.87이었다.

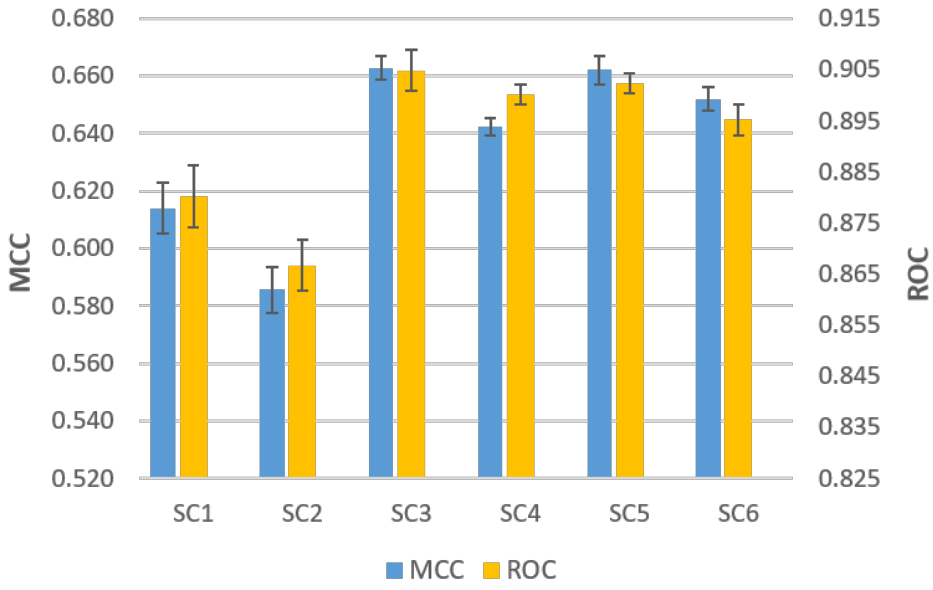

그림 5. 단일채널 모델의 성능 비교

반면에 ProtVec을 사용한 모델들은 사용된 화학적 특질의 유형에 관계없이 아미노산 서열을 사용한 모델보다 성능이 좋지 않았다.

아미노산 서열을 사용한 모델의 평균 MCC는 0.646이며 평균 ROC는 0.896이지만 반면에 ProtVec을 사용한 모델의 평균 MCC는 0.627이고 평균 ROC는 0.887이었다. 이러한 성능 차이는 특질 벡터에서 아미노산 서열의 순서를 포착할 수 있는지 여부에 따라 결정되는 것으로 보인다. Dilated CNN 모델은 아미노산 서열의 내용과 순서를 특질로 추출할 수 있는 반면에, ProtVec은 특질 벡터에서 서열의 순서를 반영할 수 없다는 한계가 있다. 왜냐하면 ProtVec은 시퀀스를 N-gram으로 나누어 단어 단위를 만들고, 각 N-gram 단어에 대해 개별 임베딩을 수행한 다음, 그 순서에 관계없이 모든 임베딩 벡터를 합산하기 때문이다. 따라서 동일한 N-gram이 사용된다면 실제 서열의 내용이 다른 단백질들도 결과적으로 동일한 임베딩 벡터를 가진다는 한계점이 있다. 이와 같은 한계를 개선하기 위해서 Mol2vec과 ProtVec의 특질 추출 기법을 개선할 필요가 있다.

표 2. 고수준 특질 추출기 개선에 따르는 성능 비교

Mol2vec	ProtVec	MCC	ROC
Mean	Sum	0.652 (± 0.004)	0.905 (± 0.002)
	Mean	0.648 (± 0.003)	0.902 (± 0.003)
	TF-IDF	0.678 (± 0.002)	0.912 (± 0.002)
TF-IDF	Sum	0.651 (± 0.003)	0.904 (± 0.003)
	Mean	0.644 (± 0.002)	0.901 (± 0.002)

	TF-IDF	0.674 (± 0.004)	0.905 (± 0.002)
Sum	Sum	0.642 (± 0.005)	0.900 (± 0.003)
	Mean	0.636 (± 0.003)	0.898 (± 0.003)
	TF-IDF	0.668 (± 0.002)	0.911 (± 0.002)

Mol2vec 및 ProtVec에서 단순 합계 기법을 개선하기 위해서 2가지 가중 평균 기법의 성능을 평가했다. 표 2는 Mol2vec와 ProtVec 모두에 대해 산술평균과 TF-IDF 가중 평균 기법을 추가적으로 적용한 9가지 조합의 예측 성능을 보여준다. 가장 성능이 높은 기법의 조합은 Mol2vec의 산술 평균과 ProtVec의 TF-IDF 가중 평균 방법 조합이다. 기존 기법과 성능을 비교하자면, MCC는 0.642에서 0.678로, ROC는 0.900에서 0.912로 성능이 증가했다.

개선된 방법들은 각 화합물 및 단백질의 함량을 보다 정확하게 포착할 수 있음을 보여준다. 데이터의 각 문서(예, 생체 활성 데이터 세트) 내의 문장(예: 화합물 및 단백질)에는 특정한 맥락과 그 특성이 있다. 특히 TF-IDF는 문장의 각 단어에 가중치를 부여하기 때문에 TF-IDF 가중평균법이 문서 고유의 특성과 맥락을 보다 세밀하게 포착할 수 있다. 반면에, Mol2vec의 경우 TF-IDF 가중 평균 방법이 산술 평균 방법보다 약간 낮은 성능을 보인다. 이는 특정 문서의 TF-IDF 가중치는 해당 문서에 편향된 정보를 제공하여 일반화 성능을 저하시킬 위험을 내포하는 것으로 보인다.

2.3.2 다중채널 특질 조합 비교

특질의 저수준 및 고수준 표현형이 조합된 시너지 효과를 파악하기 위해, 9개의 다중채널모델의 성능을 평가했다. 본문의 가독성을 높이기 위해서 다음과 같이 세 가지 다중 채널 특질의 명명법을 축약한다. 예를 들어서, 아미노산 서열 특질을 함께 사용하는 ProtVec은 $ProtVec_{AA}$이고, SMILES 문자열이 있는 Mol2vec는 $Mol2vec_{SS}$, 그리고 ECFP의 경우는 $ECFP_{SS}$이다.

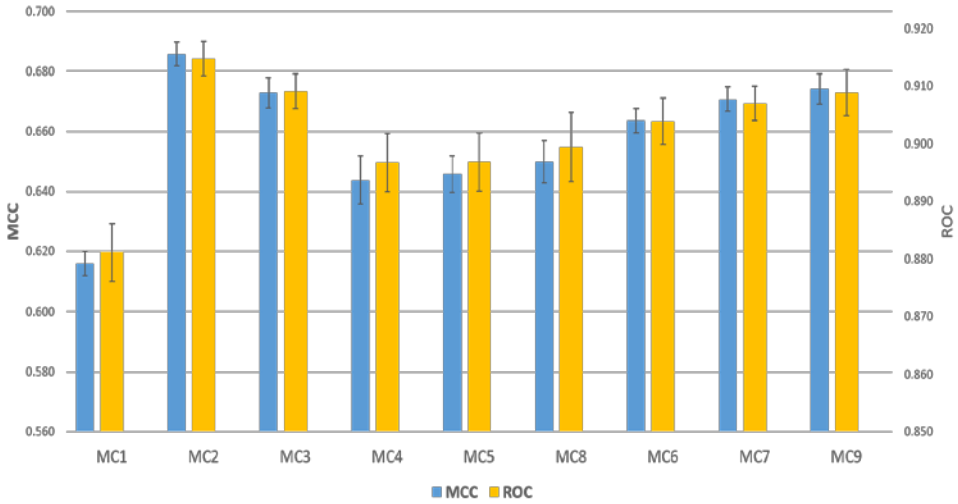

그림 6. 다중채널 모델의 성능 비교

첫째, 다채널의 성능 차이는 단백질과 화합물 사이에 차이가 있다. 그림 6에서 볼 수 있듯이, 단백질 특질의 경우 $ProtVec_{AA}$의 사용이 단일채널모델들보다 우수한 성능을 보이는 것으로 관찰되었다. 다중채널모델들의 평균 MCC는 0.658을 보이는 반면에, 단일채널모델들의 평균 MCC는 0.649을 보였다. 마찬가지로,

다중채널모델들의 평균 ROC 0.902이지만, 단일채널모델들의 평균 ROC는 0.897이다. Dilated CNN을 사용한 End to End 학습 채널은 주로 아미노산 서열 전체와 그 순서를 특질화 하는 것으로 보이며, ProtVec 채널은 아미노산 서열의 조각에 집중하여 특질화 하는 것으로 보인다. 이는 제안된 다중채널구조가 두 채널을 모두 활용하여 단백질에 대한 문장 및 단어 관점 모두에서 상호보완적으로 특질을 추출할 수 있음을 시사한다.

둘째, 화합물에 대한 다중 채널 모델은 $Mol2vec_{SS}$ 와 $ECFP_{SS}$ 중 어떤 특질을 사용했는가에 따라서 매우 다른 결과를 보여주었다. 예를 들어, $ECFP_{SS}$의 사용은 ECFP의 단독 활용과 거의 비슷한 성능을 보였다. 예를 들어서, $ECFP_{SS}$ 기반 모델의 MCC는 0.670이고, ECFP 기반 모델의 MCC는 0.669으로 거의 비슷했다. ROC도 마찬가지로 0.907와 0.906으로 그 성능이 거의 비슷했다. 반면에 $Mol2vec_{SS}$를 사용한 모델은 Mol2vec를 단독으로 사용한 모델보다 성능이 더 낮았다. 평균 MCC는 0.68에서 0.65로 떨어졌고 평균 ROC는 0.91에서 0.89로 떨어졌다. 따라서 더 나은 성능을 얻으려면 특질의 조합을 신중하게 선택하는 것이 필요해 보인다.

이러한 결과는 화합물의 특질의 성능이 실제 데이터의 표현형에 크게 영향을 받는다는 것을 시사한다. 예를 들어, 화합물은 ECFP 및 Mol2vec에 대한 실제 데이터의 표현형은 2차원 그래프의 형태로 표시된다. 그래프 데이터를 벡터로 변환하기 위해서 그래프를 하위 구조로 나누고 각 부분을 재정의하여 화합물 특질을 추출한다.

대조적으로, Dilated CNN은 1차원 SMILES 문자열에서 일반화된 특질을 추출하는 것에 어려움을 겪는 것으로 보인다. 본 연구에서는 시퀀스 데이터를 사용했으나, 새로운 채널로 활용가능한 더 다양한 데이터 유형이 있다. 예를 들어서, 그래프, 이종 네트워크 및 노드, 그리고 더 많은 생물학적 및 분자적 정보(예: 경로 및 약물-약물 상호작용)는 약물 발견, 다중 약리학, 부작용 예측 및 약물 내성에도 활용될 수 있다.

2.3.3 모델의 성능 순위 및 학습 속도 비교

위에서 소개한 6개의 단일채널모델들과 9개의 다중채널모델들을 포함하여 총 15개의 모델을 비교하고 순위를 매겼다. 모델 간의 성능차이를 비교하기 위해 각 모델과 메트릭(MCC 및 ROC)에 대해 두 개의 표준점수를 계산하고 그림 7 및 표 3와 같이 그 평균을 계산했다. 이에 더해서 표준점수 차이의 유효성을 확인하기 위해 스튜던트 t-테스트와 f-테스트를 수행했다.

표 3. 표준점수로 표현된 모델 간의 성능 비교

Model	MCC	ROC	평균	표준편차
MC_2	1.22	1.22	1.22	0.001
SC_4	0.91	0.95	0.93	0.020
MC_9	0.77	0.73	0.75	0.017
MC_3	0.72	0.75	0.74	0.018
SC_3	0.69	0.65	0.67	0.020
MC_7	0.64	0.58	0.61	0.027
SC_6	0.64	0.58	0.61	0.030
MC_6	0.36	0.32	0.34	0.027

SC_5	0.30	0.20	0.25	0.050
MC_8	−0.18	−0.04	−0.11	0.069
MC_5	−0.34	−0.26	−0.30	0.038
MC_4	−0.42	−0.27	−0.34	0.074
MC_1	−1.50	−1.55	−1.53	0.027
SC_1	−1.58	−1.63	−1.60	0.027
SC_2	−2.24	−2.25	−2.24	0.004

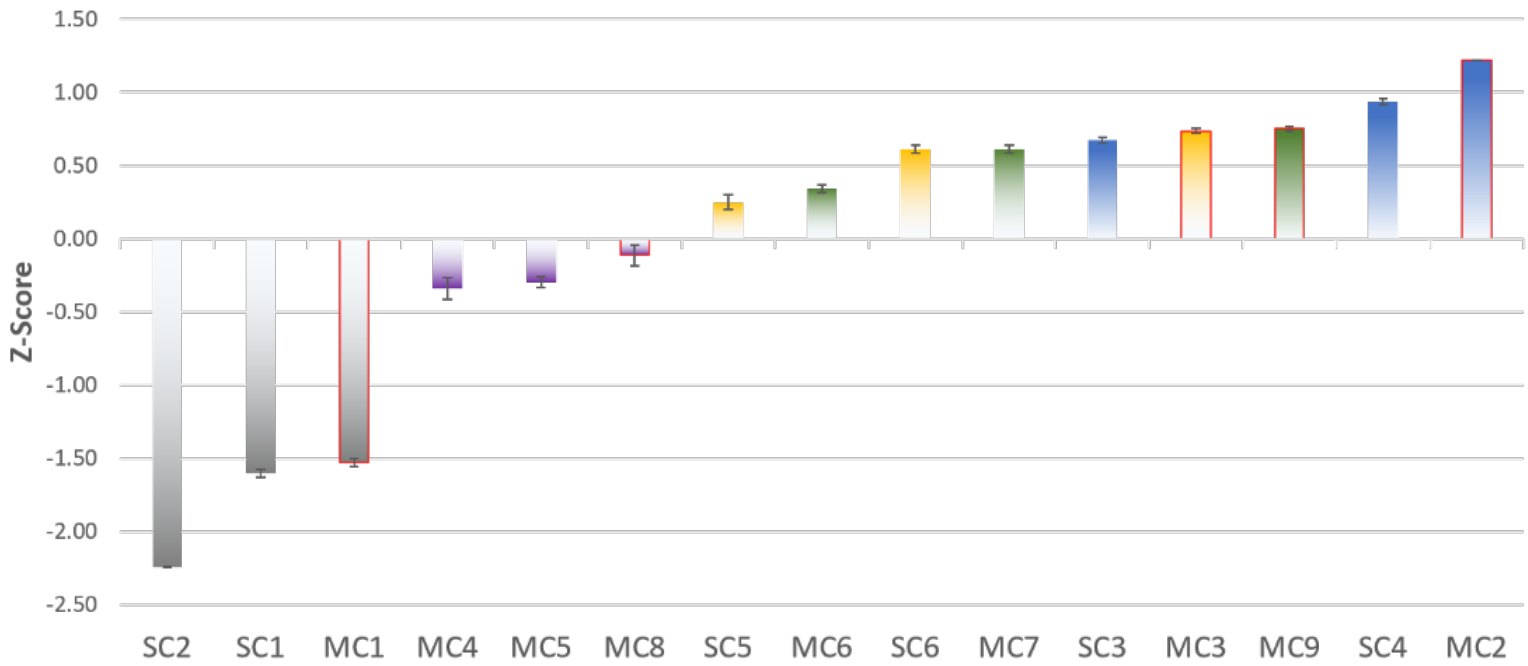

그림 7. 표준점수 기반의 모델 성능순위 비교

첫째, 약물 특질 중 Mol2Vec의 사용은 가장 좋은 성능을 나타냈다. Mol2Vec의 평균 표준점수는 0.94(±0.01) 이며, 그 다음으로 높은 $ECFP_{SS}$은 0.57(±0.02), ECFP는 0.53(±0.02), $Mol2vec_{SS}$은 −0.25(±0.06), 그리고 SMILES 서열은 −1.79(±0.02) 이었다. 통계검정으로 약물 특질을 기준으로 모델들의 성능차이를 측정한 결과, Mol2vec과 $Mol2vec_{SS}$는 유의한 성능 차이를 보였으며 ECFP와 $ECFP_{SS}$도 유의미한 성능 차이를 보였다.

둘째, 단백질 특질 중에서는 $ProtVec_{AA}$의 사용이 가장 높은 성능을 보였다. $ProtVec_{AA}$의 평균 표준점수는 0.21이며, ProtVec은 −0.14, 그리고 아미노산 서열은 −0.08 이었다. 통계검정으로 단백질 특질을 기준으로 모델 사이의 성능차이를 측정한 결과, $ProtVec_{AA}$기반의 모델들이 타모델과의 p−value가 0.05 이하로 낮았으며, 이는 단백질 특질에서 다중채널의 시너지 효과가 통계적으로 큰 차이를 일으키는 것으로 보인다.

셋째, 가장 좋은 모델은 Mol2vec과 $ProtVec_{AA}$을 사용하는 다중채널모델인 MC_2이며 이 모델은 표준점수의 평균이 1.22이다. MC_2는 다른 14개의 모델과 비교한 통계검정에서 p−value가 모두 0.05보다 작았으며, 이는 MC_2의 높은 성능이 통계적으로 유의미함을 의미한다.

넷째, 초기 성능과 수렴 속도에서 다중채널구조의 영향에 대해 살펴본다. 초기 성능은 첫 번째 epoch에서의 성능으로 측정하였고, 수렴 속도는 모델의 최고 성능의 98%에 이르는 실제 실행 시간으로

측정하였다. 각 모델의 수렴 속도를 보다 정확하게 비교하기 위해 실제 실행 시간을 측정했다. 그림 8에서와 같이 상위 3개 모델(MC_2, SC_4 및 MC_9)과 기준 모델(SC_1)을 비교했다. 다중채널모델이 초기 성능과 수렴 속도 측면에서 단일채널모델보다 우수한 성능을 보이는 것으로 관찰되었다. 초기 성능을 MCC로 측정했을 때, MC_9은 0.47, MC_2은 0.43, SC_1은 0.40, 그리고 SC_4은 0.38이었다. 성능에 더해서, 최고 성능의 98%에 이르는 수렴 속도를 측정했을 때, MC_9은 11분(18 Epoch), MC_2은 41분(113 Epoch), SC_1은 50분(102 Epoch), 그리고 SC_4은 55분 (201 Epoch)이 소요되었다.

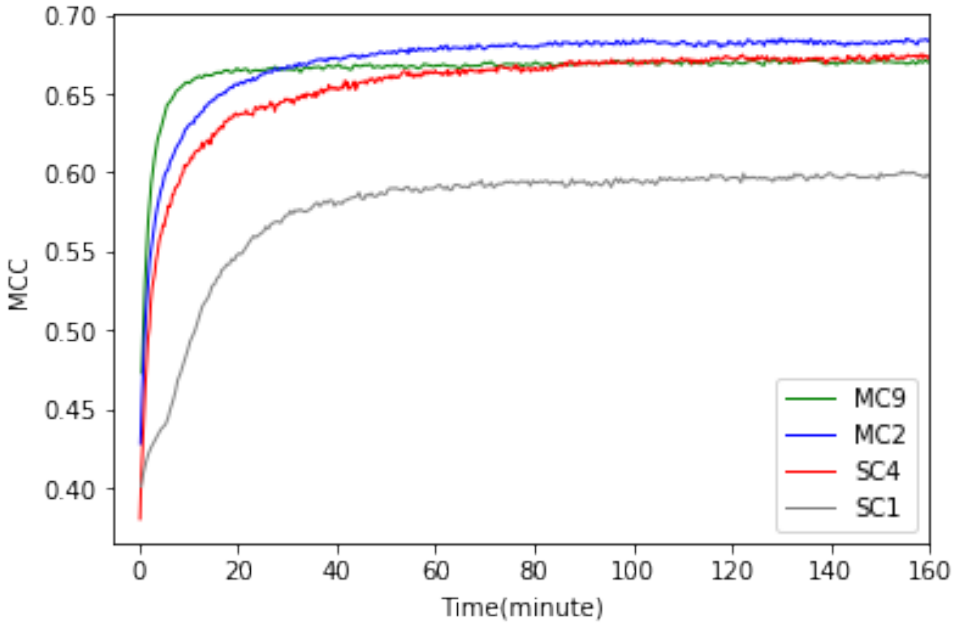

그림 8. 모델의 학습 수렴속도 비교

특히, SC_4와 MC_9의 최고 성능과 수렴속도에서 대조적인 차이를

보였다. MCC가 0.678 대 0.674로 단일채널모델인 전자가 다중채널모델인 후자보다 약간 더 나은 성능을 보이지만, 수렴 속도에서는 전자가 후자를 능가하는 데 104분이 추가적으로 소모되었다. 비록 이러한 차이의 정확한 원인을 입증할 수는 없지만, Dilated CNN의 저수준 표현이 이 차이에 기여한 것으로 보인다. 이러한 현상을 딥러닝의 정보병목현상 (IB) 이론[78]의 관점에서 논의해본다. IB이론의 저자들은 "DNN은 학습 초기에는 학습(fitting) 및 기억 단계의 단계를 가지며, 학습 후기에는 정보압축과 망각 단계로 구성된 두 가지의 학습 단계를 가진다. 이는 DNN의 높은 일반화 성능과 관련이 있다."라고 설명한다. 이러한 관점에서 다음 설명은 수렴 속도의 차이를 설명하는 데 도움이 될 수 있다. (1) 다중 채널 아키텍처는 일반화된 특질이 많기에 새롭게 압축하거나 잊어버릴 정보가 적기 때문에 학습을 적게 할 수 있다. (2) 단일 채널 아키텍처는 일반화 특질이 충분하지 않기 때문에 피팅 단계와 압축 단계 모두에 대한 적절한 표현을 찾기 위해 더 많은 학습이 필요하다. 요약하면, 다중 채널 아키텍처는 성능과 수렴 속도를 모두 향상시킬 수 있다.

2.3.4 전이학습 적용

다중채널모델이 생체 반응 데이터에서 독성 데이터에 적용 가능한 일반화된 정보를 추출가능한지에 대한 평가를 수행한다. 다중채널모델이 약물과 단백질에 대한 일반화된 정보를 추출할 수 있는지 테스트하기 위해서, MC_2는 벤치마크 데이터 세트에서 사전 훈련된 후, Tox21

데이터 세트 미세 조정(fine-tuning) 되었다. 본문의 가독성을 높이기 위해 사전 훈련된 모델은 다음과 같이 축약된다. PM_i 에서 i는 학습 작업에 대해 사전 학습된 횟수이며, 사전 학습되지 않은 기본 모델은 PM_0으로 표기된다. MCPINN의 전이학습 능력을 평가하기위해, 최고 성능, 초기 성능, 수렴 속도 측면에서 서로 다른 횟수의 사전 훈련된 모델들을 비교했다.

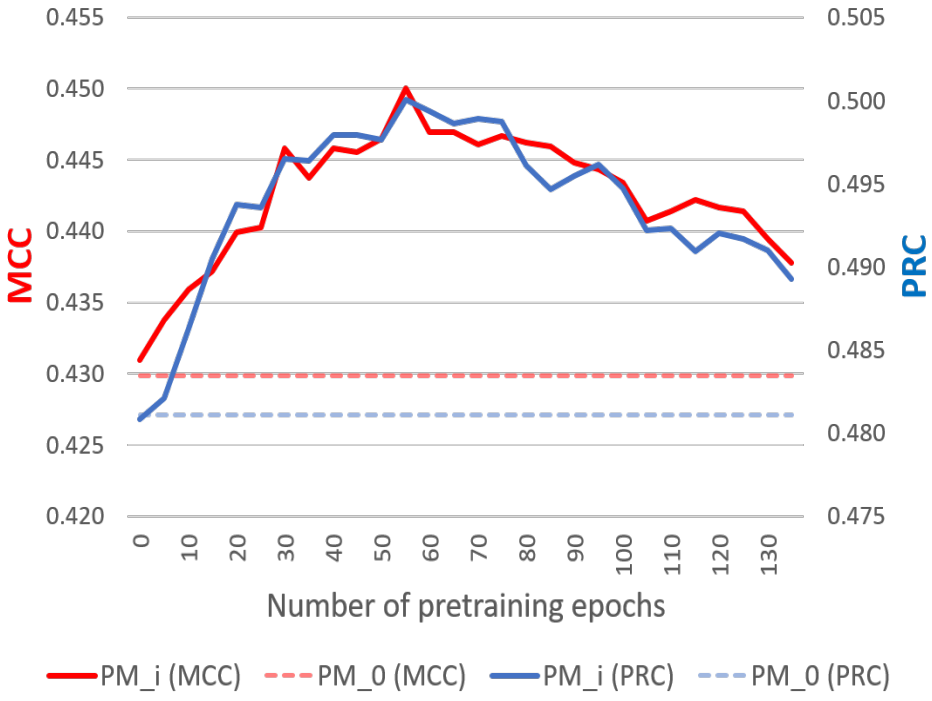

그림 9. 사전훈련 횟수에 따른 성능비교

첫째, 사전 훈련된 모델들(PM_i)은 사전 훈련되지 않은 모델(PM_0)보다 더 높은 성능을 보였다. 그림 9에서 보이듯이, PM_0의 MCC는 0.43이고 PRC가 0.48인 반면에, PM_{30}에서 PM_{110}까지 사전

훈련된 모델들은 MCC와 PRC 모두에 대해 통계적으로 유의미한 성능차이를 보였다. 단, PM_{85}은 p-value가 0.053으로 통계적인 차이를 보이지 못했다. 학습모델의 성능이 사전학습 55회까지는 상승하지만, 그 이후로는 성능이 감소하는 것으로 관찰되었다. 성능 감소의 이유는 학습 데이터 세트에 과적합(overfit) 되었기 때문으로 보인다.

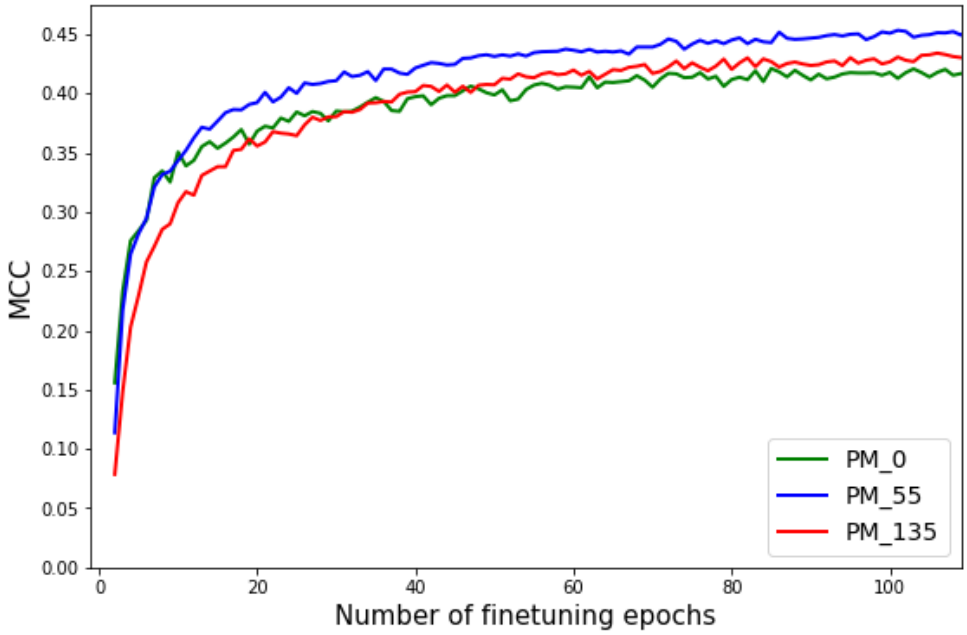

그림 10. 사전훈련 횟수에 따른 수렴속도 비교

둘째, 초기 성능은 미세조정학습의 횟수에 따라서 다른 양상을 보였다. 예를 들어서, PM_0, PM_{55}, PM_{135} 모델들의 미세 조정 단계를 살펴보고 사전학습 수에 따른 일반화 성능 차이를 비교해보았다. 그림 10 에서 보이듯이, PM_0는 미세 조정 epoch 10회 까지는 다른 모델보다 약간 나은 성능을 보였지만 미세 조정이 계속되면서 다른 모델보다

성능이 저하되었다. 초기 성능을 MCC로 측정했을 때, PM_0은 0.16, PM_{55}은 0.11, PM_{135}은 0.08이었다. 반면에, 미세 조정 epoch 11회 이후로는 PM_{55}이 PM_0의 성능을 능가하기 시작했다. 또한, 미세 조정 epoch 40회 이후로는 PM_{135}가 PM_0 보다 높은 성능을 보였다. 이와 마찬가지로, 최고 성능의 95%에 도달하기 위한 수렴 속도에서도 성능차이가 관찰되었다. PM_{55}은 epoch 46회가 필요하고, PM_{135}은 56회, 그리고 PM_0은 60회가 필요했다.

2.4 소결론

본 연구에서는 DTI 데이터를 최대한 활용하기 위해 PCM 방법에 기반한 MCPINN구조를 제안했다. MCPINN은 DNN의 3가지 접근 방식인 분류기, 특질 추출기, 종단 간 학습기를 활용하여 표현 학습 능력을 극대화한다. 각 쌍의 효과를 조사하기 위해 특질 쌍의 전체 조합을 평가했다. 초기 성능과 수렴 속도 측면에서 단일채널모델과 다중채널모델을 비교했다. 이에 더해서, 전이학습을 통해서 MCPINN의 일반화 정보 추출 능력을 평가했다. 이전의 절에서 논의한 바와 같이, 단일채널 및 다중채널에서 성능의 개선을 이루었다. Mol2vec 및 ProtVec의 경우 가중 평균 연산이 화합물 및 단백질을 나타내는 합 연산보다 더 나은 대안임을 제안했다. $ProtVec_{AA}$ 및 Mol2vec의 특질 쌍을 사용하는 다중채널모델은 통계적으로 유의한 차이로 다른 모든 모델보다 성능이 뛰어났다. $ProtVec_{AA}$의 사용은 단일채널모델보다 더

나은 성능을 보여주었다. MCPINN은 고수준 표현형과 저수준 표현형을 모두 활용하여 단백질 정보의 순서와 내용의 상보적인 특질을 모두 추출할 수 있다.

III. 유전자 중심의 다중 채널 구조(GCMC)

3.1 서론

새로운 항암제가 원하는 효능을 갖도록 설계하는 것은 제약 산업에서 여전히 어려운 과제로 남아 있다[30]. 특히 유전체, 전사체 및 후성 유전체 전반에 걸친 종양 이질성(tumor heterogeneity)은 항암제 치료 효능을 손상시킬 수 있으며 이를 극복하는 것은 항암 후보물질에 대한 환자의 임상적 이질성을 해결하는 중요한 요소로 꼽히고 있다[31]-[35]. 고효율 시퀀싱(HTS) 기술의 출현으로 악성 종양의 발생과 진행에 영향을 미치는 다양한 생물학적 측면을 측정할 수 있게 되었다[36]. 각 유형의 다중 오믹스 데이터는 특정 생물학적 정보 계층을 반영하며, 이를 통해서 생물학적 시스템에 대한 추가적인 관점을 제공할 수 있다[37].

다중 오믹스 프로파일은 다양한 유형의 데이터로 구성된다. 유전체학 및 전사체학은 암 연구 및 임상 적용을 위해 광범위하게 연구된 일반적인 오믹스 데이터이다. 첫째, 인간 발암에 대한 대부분의 연구는 체세포 돌연변이(somatic mutation)에 초점을 맞춰 왔다[38]. 이는 종양 유전자, 종양 억제 유전자 및 DNA 복구 시스템이 모두 성장 이점과 종양 생존을 향상시키기 위해 변경될 수 있기 때문이다[39]. 둘째, 유전자 복제수 변이(CNV)은 게놈의 섹션이 반복되는 구조적 변이 유형이다[40]. 여러 연구에서 특정 유전자의 CNV가 다양한

유형의 암의 발생, 발달 및 진행에 역할을 하는 것으로 나타났다[40]. 셋째, 유전자 발현 데이터는 유전자가 단백질 또는 RNA 구조로 번역되는 정보이다. 암 유전자는 변이된 발현에 의해 확인되며, 이는 악성 종양의 상당 부분에서 비정상적인 표현형을 유발한다[41]. 암 연구를 위해 분석된 체세포 돌연변이, 유전자 복제수 변이 및 유전자 발현 데이터는 다중 오믹스 데이터의 대표적인 사례이다.

암 세포주는 세포 암 모델의 환경에 대한 통찰력을 제공하기 때문에 약물유전체학(pharmacogenetics) 연구에도 중요하다. CCLE(Cancer Cell Line Encyclopedia)는 유전자 발현, 체세포 돌연변이, 복제수변이를 포함한 대규모 다중 오믹스 데이터를 제공한다. 암 세포주는 항암제 반응성(cancer drug response, CDR)을 식별하기 위해 일반적으로 사용되는 모델이다. 최근 암 세포주에 대한 대규모 항암제 스크리닝은 Gene of Drug Sensitivity in Cancer(GDSC) 및 환자 유래 이종이식편(PDX) 마우스 모델[79]과 같은 여러 데이터베이스가 구축되었다. 이러한 리소스는 in silico 접근 방식의 예측 모델의 발전을 촉진시켰다. 예를 들어서, 전통적인 기계학습 모델인 로지스틱 회귀[79], 랜덤 포레스트[80], 서포트 벡터 머신[81] 부터 딥 러닝[82]-[84]에 이르는 최신 인공지능 모델들이 CDR 예측의 성능을 높이고 있다.

이에 더해서, 인공지능 모델들의 임상적인 적용 및 유용성 여부도 중요해지고 있다. 이를 위해서는 실제 환자 데이터의 약물 반응성 데이터를 사용하여 예측 모델을 훈련시키는 것이 이상적이지만, TCGA(The Cancer Genome Atlas) 데이터 세트와 같은 환자 데이터

세트는 신경망 모델을 포함한 기계학습 모델을 학습시키기에 데이터의 양이 부족하다. 대안적인 접근법으로, 일부 연구[85], [86]에서는 세포주 데이터 세트에 먼저 예측 모델을 학습시킨 뒤에 TCGA 환자 데이터 세트에 적용하여 유망한 예측 결과값을 얻었다. 이와 같은 전략은 정밀 종양학으로의 발전을 가속화할 수 있는 가능성을 보여주고 있다.

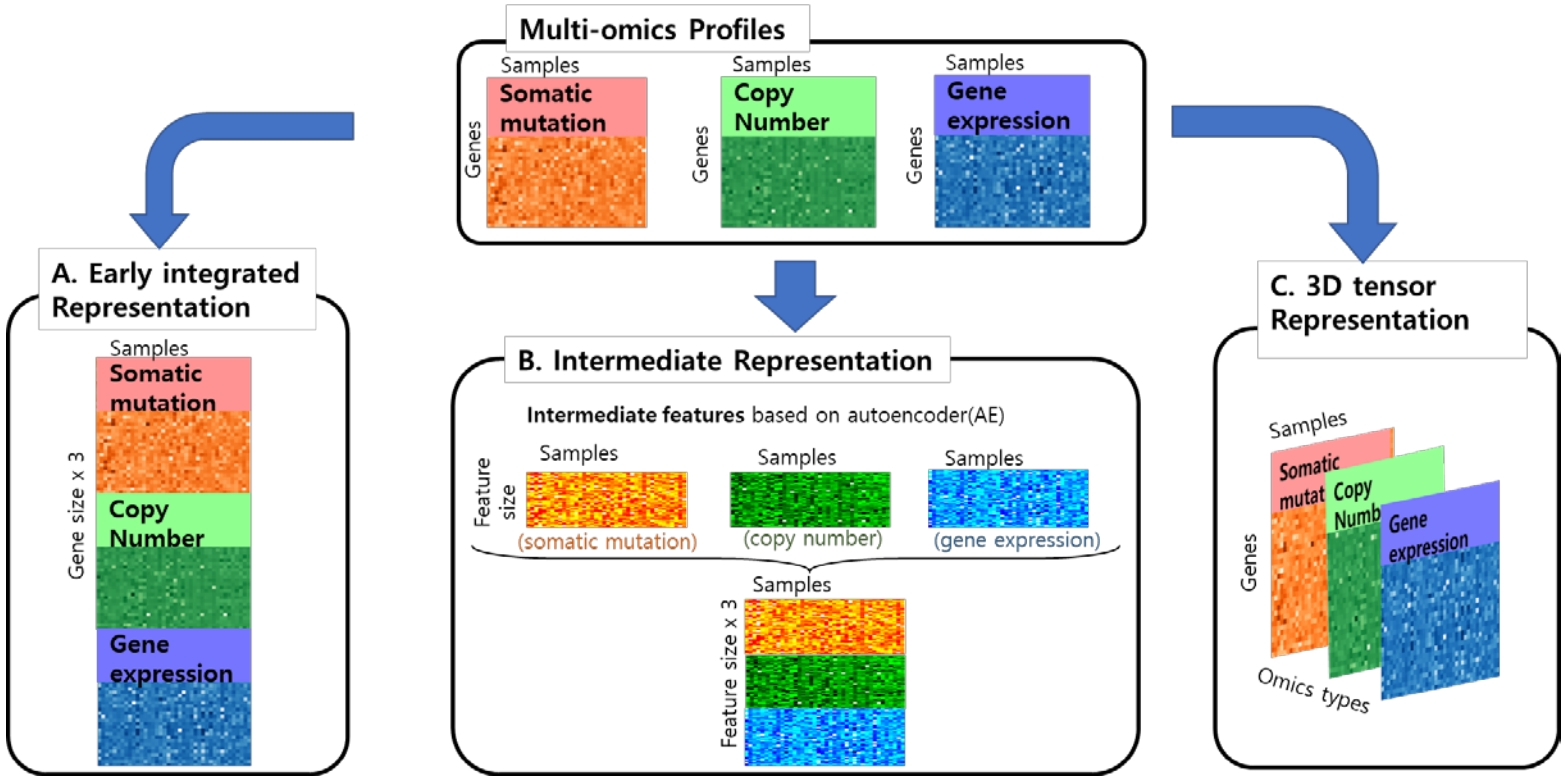

그림 11. 다중 오믹스 통합 기법 개요

다중 오믹스 데이터 분석에서 가장 중요한 과제는 데이터 통합이다. 오믹스 데이터는 잡음이 많고 복잡하기 때문에 모델의 CDR 예측 성능은 통합 전략에 따라 달라질 수 있다. 이 기법들은 통합이 발생하는 시점에 따라서 초기 통합과 중기 통합 방식으로 분류된다[87], [88]. 첫째, 초기 통합(early integration) 기법에서 다중 오믹스 데이터 세트는 그림 11(A)와 같이 큰 행렬로 통합된다. 이때 데이터의 샘플 숫자 자체는 동일하게 유지되지만 통합된 다중 오믹스 특질의 차원 수는 증가한다. 결국 이 통합기법은 데이터 세트를 더 복잡하고 오류를 증가시켜서 학습에 어려움을 초래한다[89]. 또한, 이 접근 방식은 각 오믹스 유형의 고유한 데이터 분포를 무시하기 때문에 예측모델이 그 특질이 소속된 오믹스 유형과 같은 의미없는 패턴에 매몰되어 학습하게 될 수 있다[89].

이 문제를 해결하기 위해 중기 통합(intermediate integration) 기법은 그림 11(B)와 같이 각 오믹스 특질을 중간 특질(intermediate feature)으로 변환한다. 이 기법은 각 오믹스 유형의 데이터 분포를 유지하면서 중간 특질의 노이즈를 줄이고 차원 수를 줄이는 것을 목표로 한다[88]. 이러한 중간 특질은 CDR 예측의 성능을 향상시키는 것으로 나타났다[89]. 그러나 중간 특질은 다중 오믹스 프로파일이 각 오믹스 별로 변환되기 때문에, 각 유전자 단위의 정보의 손실 및 노이즈, 그리고 중복된 정보가 포함되는 잠재적인 위험이 있다[89]. 이상적으로는 전체 학습 과정에서 각 오믹스 프로파일이 추출될 때 다른 오믹스의 프로파일도 고려되는 것이 필요하다. 특히 더 미세한

입도(granularity)의 관점에서는 각 오믹스 내부의 유전자 수준에서 중간 표현을 생성하여서 보다 자세한 정보로 표현할 필요가 있다.

다단계 통합(multi-staged integration) 접근법은 유전자 수준에 영향을 미치는 오믹스 구성 요소를 발견하는 데 중점을 둔다[90]. 이 접근 방식에서 다중 오믹스 통합을 위해서는 각 오믹스 프로파일이 동일한 개수의 유전자와 샘플 수를 가져야 한다. 이를 통해, 다중 오믹스 데이터는 그림 11(C)와 같이 3차원 텐서로 변환될 수 있다. 텐서의 새로운 차원은 오믹스 프로파일의 타입을 나타내며, 이는 생물학적 채널으로써 이미지 데이터가 빨강, 녹색 및 파랑 색상 채널로 표현되는 방식과 유사하다. 컴퓨터 비전 분야에서 컨볼루션 신경망(CNN)은 이미지 데이터에서 잠재적인 특질을 추출하는 탁월한 아키텍처이다[64]. CNN의 구조는 주어진 픽셀 영역 내의 모든 채널 정보를 동시에 추출하므로 다중 오믹스 프로파일을 통합하는 데 적합하다. 그러나 이미지 데이터와는 다르게 유전자 데이터의 서열은 순서가 존재하지 않으므로, CNN의 구조를 이에 알맞게 수정하는 것이 필요하다. 예를 들어서, 일반적인 CNN 기반 모델들은 다중 오믹스 프로파일 내의 유전자 순서가 변경된 경우, 동일한 특질을 결과물로 추출하지 못할 수 있다[91]. 왜냐하면 CNN은 커널과 보폭(stride)의 크기에 따라 입력 데이터의 공간 패턴을 추출하므로 유전자 순서의 변화에 따라서 결과가 달라질 수 있기 때문이다. 따라서 CNN의 구조는 유전자의 순서에 불변성(order-invariant)을 갖도록 신중하게 설계되어야 한다.

또한 생물의학 연구에서는 높은 예측 성능만으로는 충분하지 않다[92], [93]. 다중 오믹스 프로파일이 있는 예측 모델은 입력과 예측 결과 간의 관계를 설명하는 것이 필요하다. 기존의 연구에서는 ablation study[94]를 사용하여, 특정 오믹스 프로파일을 제거 유무의 성능차이를 비교함으로써 해당 오믹스 타입이 모델의 성능에 미치는 영향을 이해하려 했다[95]-[97]. 이를 통해 기존의 연구들은 유전자 발현 데이터를 CDR 예측을 위한 가장 유망한 오믹스 유형으로 제안했다[4], [98], [99]. 그러나 다중 오믹스 통합 연구 중 각 오믹스 유형의 기여도를 분석하는 것은 아직 탐구되지 않았다. ablation study는 특정 오믹스 유형을 사용하는 모델 간의 성능 차이를 조사할 수 있는 반면에, 각 오믹스 유형 조합의 기여 비율을 직접 분석할 수는 없다는 한계가 있다.

현재 설명가능한 인공지능(explainable AI, XAI) 알고리즘을 통해서 인공지능 모델들이 보다 해석 가능하고 설명 가능한 결과를 생성하기 위한 연구가 진행 중이다[100]-[102]. 특히 Integrated Gradients (IG) 알고리즘[103]은 이론적 토대와 다양한 응용분야의 적용으로 인해 예측 결과를 해석하는 방법으로 인기를 얻었다. IG 알고리즘은 각 특질의 중요도를 정량화하여 특질 기여 비율을 추출할 수 있다. 이를 활용한다면, 다중 오믹스 통합 모델에서 입력과 예측된 결과 간의 관계에 대한 추가 통찰력을 얻을 수 있다.

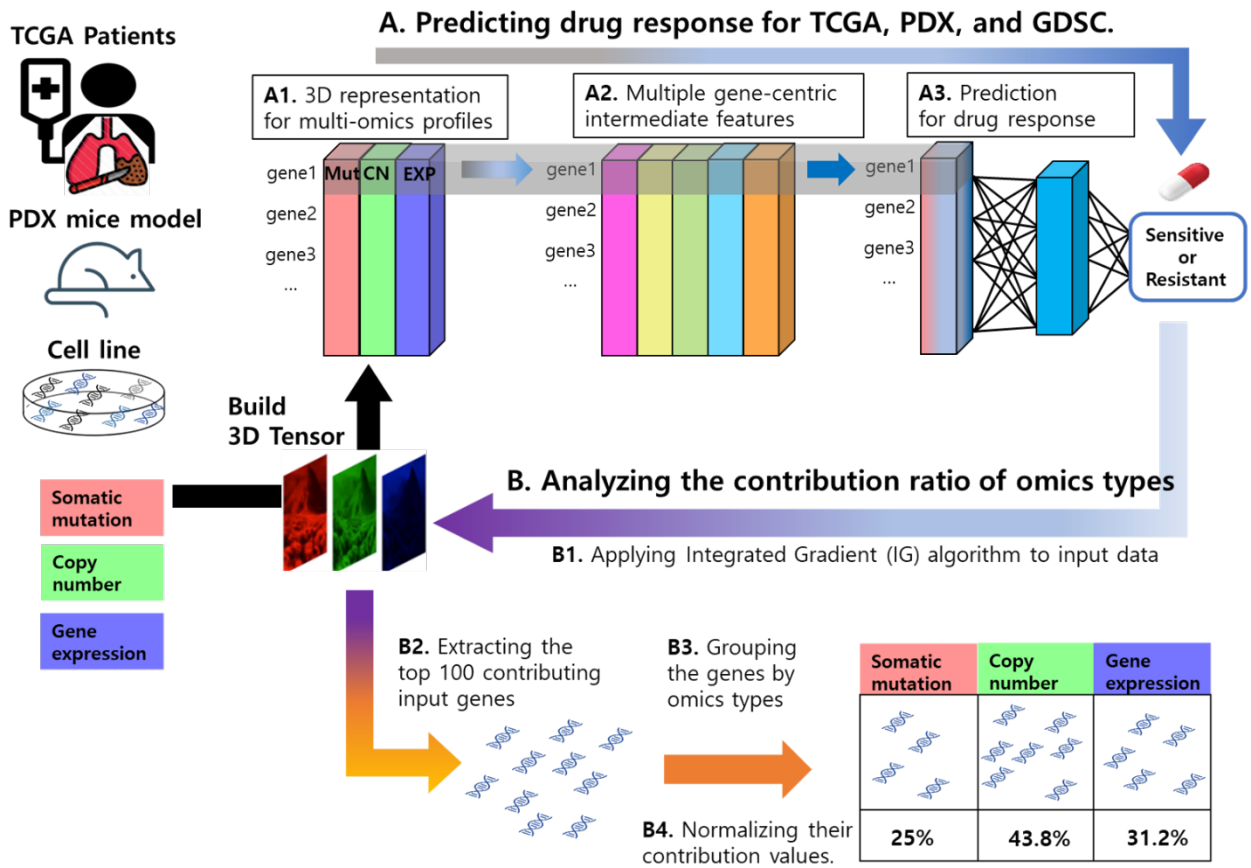

그림 12. 유전자 중심 다중 채널(GCMC) 구조의 개요.

본 연구에서는 CDR 예측을 위한 다중 오믹스 프로파일을 통합하기 위해 CNN을 기반으로 하는 유전자 중심의 다중 채널 구조(GCMC)를 제안한다. GCMC는 CDR 예측 모델 중 처음으로 유전자의 순서에 불변성(order-invariant)을 가진 CNN 구조이다. 그림 12와 같이 GCMC는 다중 오믹스 타입을 새로운 차원을 가진 3차원 텐서로 변환한다. GCMC의 컨볼루션 인코더(CE)는 각 유전자에 대한 모든 오믹스 채널을 통합하여 유전자 중심의 특질을 추출한다. 유전자 중심 특질 벡터를 추출하기 위해 CE는 각 유전자의 다중 오믹스 프로파일을 통합하여 여러 개의 중간 특질을 생성한다. 예측 모듈에서 GCMC는 축소된 차원에 대한 유전자 정보의 가중치를 학습하고 이를 사용하여 CDR을 예측한다. 추가적인 기여로써 본 연구는 IG 알고리즘을 통해 어떤 오믹스 유형이 예측 성능에 가장 많이 기여하는지 분석한다. 이것은 예측 모델에 사용된 다양한 오믹스 유형의 기여 비율을 분석한 첫 번째 연구이다.

GCMC는 TCGA 암환자, PDX 마우스 모델 및 GDSC 세포주 데이터 세트를 포함한 다양한 데이터 세트에서 평가되었다. 본 연구는 GCMC를 초기 및 중기 통합 모델과 비교했다. GCMC는 GDSC 세포주 데이터 세트의 265개 약물 중 75% 이상에서 단일 오믹스 모델을 포함한 기준 모델보다 더 나은 성능을 달성했다. 또한 GCMC가 AUPR 및 ROC 측면에서 TCGA 및 PDX 데이터 세트에서 최고의 성능을 달성했으며, 이를 통해 임상적인 활용가능성에서도 높은 잠재력을 지녔음을 보였다. 또한 본 연구에서는 예측모델의 성능에 기여하는

오믹스 유형의 비율을 측정하여 다중 오믹스 통합 모델들을 분석했다. GCMC는 다중 오믹스 프로파일을 균형있게 조합하여 예측성능을 높였다.

본 연구의 주요 기여를 요약하면 다음과 같다. (1) CNN을 사용하여 다중 오믹스 프로파일을 통합하는 유전자 중심 다중 채널(GCMC) 아키텍처를 제안했다. GCMC는 CDR 예측 모델 중 처음으로 유전자의 순서에 불변성(order-invariant)을 가진 CNN 기반의 모델이다. (2) TCGA, PDX 및 GDSC 데이터 세트에서 GCMC가 평가되었다. GCMC는 비교 모델보다 통계적으로 유의한 차이로 더 나은 성능을 달성했다($p<0.05$). (3) IG 알고리즘을 통해 오믹스 유형 별로 예측 성능에 얼마나 기여했는지 분석한다. GCMC는 다중 오믹스 프로파일을 균형있게 조합하여 예측성능을 높였다. 이것은 다중 오믹스 통합 연구 중에서 다중 오믹스 유형의 기여 비율을 분석한 첫 번째 연구이다.

3.2 관련 연구

본 연구에서는 그림 11에서와 같이 다양한 표현형을 기반으로 다중 오믹스 데이터셋을 통합하는 몇 가지 방법을 소개한다. 다중 오믹스 통합을 강조하기 위해 생물학적 데이터만 사용하는 연구에 중점을 두었다.

3.2.1 초기 및 중기 통합 기반 모델

먼저 CDR 예측을 위한 초기 통합 기반 모델을 알아본다. [104]는 다양한 분산 메트릭으로 필터링된 정보 유전자를 선택했다. 이 방법은 AE와 elastic net에 다중 오믹스 데이터가 통합된 특질을 입력하여 CDR 예측을 수행한다. [105]은 예측 가능성이 높은 유전자 세트를 선택하기 위해 Neighborhood Component Analysis[106]를 사용했다. 특질들은 또한 유방암의 CDR을 예측하기 위해 DNN에 입력되었다.

둘째, CDR을 예측하기 위한 몇 가지 중기 통합 기반 모델을 제시한다. DeepDR[107]은 각 오믹스 유형에 대해 TCGA 데이터를 사용하여 오토인코더(autoencoder, AE) 모델을 사전 훈련했고, AE를 DeepDR의 예측 모듈에 연결하여 세포주 데이터에서 CDR을 학습했다. DeepCDR[95]은 FFNN을 사용하여 각 오믹스 데이터에 대한 100차원의 특질을 추출했다. DeepCDR은 중간 특질을 통합한 뒤 CDR 예측을 수행한다. MOLI[86]는 TCGA 및 PDX 데이터 세트를 사용하여 각 오믹스 유형 별로 AE를 학습하여 중간 특질을 추출했다. MOLI는 현재 최고의 성능(state of art, SOTA)을 가진 모델로써, [86]에서 단일 오믹스 및 초기 통합 방법 기반 방법을 포함한 기존 모델의 성능을 능가했음을 보였다. 하지만 중기 통합 기반의 모델들은 종종 중간 특질을 추출하기 위해 AE에 의존하는데, AE의 많은 매개변수는 고차원 데이터 세트의 노이즈에 대한 과적합에 취약하게 될 위험이 존재한다.

3.2.2 3차원 텐서 기반 모델

3차원 텐서 표현을 기반으로 한 암 관련 예측 방법을 살펴본다. 먼저 일부 연구에서는 텐서 분해 방법[108]을 활용하여 유전자, 샘플(환자 또는 세포주) 및 오믹스 유형에 대한 중간 특질을 추출하는 3차원의 다중 오믹스 텐서를 구축했다. MONTI[108]는 암의 아형별 특질을 추출하기 위해 유전자 중심 방식으로 다중 오믹스 프로파일을 통합했다. MONTI는 텐서 분행 방법으로 분리된 환자 기반의 특질을 사용하였고, 이를 통해 추출된 특질을 활용하여 암의 하위 유형 성능이 개선되었다. [109]은 베이지안 텐서 분해를 통해 중간 특질을 추출했으며, 이 기법은 텐서의 최고 순위(rank)를 자동으로 선택할 수 있다. 분해된 특질은 환자의 암 하위 유형 수를 결정하고 각 하위 유형에 구성원을 할당하기 위해 합의 클러스터링 기법에 활용되었다. 그러나 텐서 분해 방법은 비지도(unsupervised) 학습 기법에 기반하기 때문에 상당한 양의 기존 생물학적 지식을 통합하는 데 어려움이 있다[89].

둘째, 소수의 연구에서는 중간 특질을 추출하기 위해 3차원 다중 오믹스 텐서에서 CNN을 사용했다. [110]는 유방암 하위 유형을 예측하기 위해서 PAM50의 유전자 발현과 복사 수 데이터를 통합했다. 그러나 [91]에서 강조한 것처럼, 이 방법은 유전자의 순서에 대한 불변성(order-invariant)을 갖지 못했다. 이 문제를 해결하기 위해 [111]은 다중 오믹스 데이터 세트를 유전자 유사성 네트워크 (gene similarity network, GSN)라고 하는 다중 계층 네트워크로 변환했다.

그러나 GSN의 핵심 템플릿인 자기 조직화 지도(self-organized map, SOM)를 개발하기 위해 올바른 수의 유전자를 직접 선택하는 것은 여전히 해결하기 어려운 과제로 남아 있다. 예를 들어, 저자는 실험적인 방법으로 14개의 유전자를 최종적으로 선택했으나, 이것은 CDR 예측을 포함한 다른 작업에 동일하게 적용하기는 어렵다는 한계가 있다.

위의 연구들은 암 관련 예측 작업을 기반으로 하는 3D 텐서 표현을 위해 몇 가지 방법이 개발되었음을 보여준다. 또한 몇 가지 연구에서 CNN을 통해 3D 텐서 표현을 통합하려고 시도했지만, 유전자의 순서에 불변성을 갖도록 예측 모델을 설계하는 것은 여전히 어려운 과제로 남아있다.

3.3 연구 방법

3.3.1 GCMC 구조

3.3.1.1 3차원 데이터 표현형 변환

GCMC는 다중 오믹스 프로파일을 그림 12(A1)과 같이 오믹스 유형을 나타내는 새로운 차원의 3차원 텐서로 변환했다. GCMC는 유전자 발현, 체세포 돌연변이, 카피 수 데이터를 포함한 3가지 오믹스 프로파일을 사용했다. 각 오믹스 프로파일에는 각 약물에 대해 동일한 유전자와 샘플이 있다. 데이터 세트는 유전자 발현, 체세포 돌연변이 및

카피 수 데이터에 대해 각각 다음의 2차원 행렬로 나타낼 수 있다. $\boldsymbol{D_{exp}} \in \mathbb{R}^{G\times S}, \boldsymbol{D_{mut}} \in \mathbb{R}^{G\times S}, \boldsymbol{D_{cn}} \in \mathbb{R}^{G\times S}$. 여기서 G는 유전자의 개수를 나타내고, S는 데이터 샘플의 숫자를 나타낸다.

둘째, 다중 오믹스 행렬을 동일한 유전자 및 샘플로 정렬하여 3차원 텐서를 구축했다. 각 오믹스 프로파일의 유전자는 서로 일치해야 하는 반면, GCMC는 유전자의 순서에 불변성을 가지므로 초기 유전자의 순서 자체는 임의적일 수 있다.

셋째, 2차원 오믹스 행렬을 오믹스 유형을 나타내는 새로운 차원의 3차원 텐서로 변환하여, O개의 오믹스 유형을 갖는 다중 오믹스 텐서 $\mathrm{D} \in \mathbb{R}^{\mathrm{G\times S\times O}}$ 를 생성했다. 결과적으로 각 유전자에 대한 다중 오믹스 프로파일을 동시에 분석하여 그 정보를 상보적으로 활용할 수 있다.

3.3.1.2 특질 추출 모듈

본 연구에서는 그림 12(A2)와 같이 유전자 순서 불변인 유전자 중심의 특질을 추출하기 위해 컨볼루션 인코더(convolutional encoder, CE)를 구성했다. 일반적으로 CE는 커널 $\mathbf{K}$를 사용하여 위에서 정의한 3차원 다중 오믹스 텐서 D의 특질을 추출한다. 커널 $\mathbf{K}$는 M개의 필터로 중간 특질 $\mathbf{Z} \in \mathbb{R}^{\mathrm{G\times S\times M}}$를 생성한다. 커널 $\mathbf{K} \in \mathbb{R}^{\mathrm{M\times F\times 1\times O}}$ 는 훈련 가능한 매개변수의 숫자이며 여기서 M 와 F 는 각각 필터의 개수와 커널의 크기를 나타낸다.

특히, CE의 커널 $\mathbf{K}$ 는 유전자 순서 불변이 되도록 신중하게

설계했다. 커널 크기 F는 입력으로 수신되는 유전자 수를 결정하고 보폭 크기(stride) T는 커널이 한 번에 건너뛰는 유전자 수를 결정한다. 결과적으로 특질지도 $\mathbf{Z}$의 유전자 수는 $(G-F)/T+1$로 변한다. 따라서 본 연구에서는 유전자의 순서에 불변성(order-invariant)이 되도록 CE를 구축하기 위해 커널과 보폭의 크기를 1로 설정했다. 만약 그렇지 않으면 유전자의 순서에 따라서 추출된 특질의 차원과 값에 대한 변형이 발생한다.

위의 논의를 종합하여 커널 $\mathbf{K}$의 m 번째 필터를 사용하여 n 번째 데이터 샘플(D_n)의 모든 유전자에 걸쳐 중간 특질 $Z_{n,m}$를 추출하는 수식은 다음과 같다.

$$Z_{n,m} = \begin{bmatrix} \sigma(\sum_j^C D_{1,n,j} K_{m,j}) \\ \sigma(\sum_j^C D_{2,n,j} K_{m,j}) \\ \cdots \\ \sigma(\sum_j^C D_{G,n,j} K_{m,j}) \end{bmatrix}$$

여기서 C는 입력 데이터에서 오믹스 유형의 수를 나타내며, n, m, j는 샘플, 커널 및 채널의 인덱스를 나타내고, $\sigma(\cdot)$는 ReLU(Rectified Linear Unit)와 같은 활성화 함수를 나타낸다.

마지막 레이어에서는 커널의 필터 크기 M은 1으로 설정하여 여러 개의 중간 특질을 단일 벡터로 압축한다. 결과적으로 이 모듈은 $Z_{last} \in \mathbb{R}^{G \times S \times 1}$의 최종 유전자 중심 특질을 생성한다.

또한 커널 $K_{m,j}$는 j번째 오믹스 유형에 대해 1부터 G까지 모든 유전자에 걸쳐 매개변수를 공유하며, 이 기법은 매개변수의 수를 크게

줄인다. 예를 들어, GCMC는 각 오믹스 유형에 대한 모든 유전자의 매개변수를 공유하여 매개변수의 0.1%에 CE를 할당했다. 이것은 다중 오믹스 프로파일을 특질로 추출할 때 과적합을 방지하기 위해 CE를 정규화하는 효과를 발생시킨다[64].

3.3.1.3 예측 모듈

본 연구에서는 그림 12(A3)과 같이 유전자 중심 특질을 통합한 다음 CDR 값을 예측하는 예측 모듈을 구축했다. 이 예측 모듈은 모든 유전자 정보 간의 가중치를 학습하기 위해 완전 연결 신경망(fully connected neural networks, FCNN)으로 구성되었다. 이 모듈은 특질을 저차원 공간에 매핑하고 CDR 예측을 위한 정보를 추출했다. 약물 반응을 예측하기 위해 마지막 레이어를 단일 노드와 시그모이드(sigmoid) 활성화 함수로 설계했다.

이 모듈에는 복잡한 약물 반응 패턴을 학습하기 위한 GCMC의 모델 파라미터가 99.9% 포함되어 있다. 이 많은 양의 파라미터들은 예측 모듈을 고차원 데이터 세트의 노이즈에 대한 과적합에 취약하게 만들 수 있다[64]. 따라서 과적합을 방지하기 위해 레이어 사이에 드랍아웃[112]을 사용했다.

3.3.1.4 최적화 및 하이퍼 파라미터 탐색

다양한 방법론과 하이퍼 파라미터를 활용하여 GCMC를 최적화했다. 손실 함수는 이진 교차 엔트로피이고 최적화 기법은 AdamW[113]이다. AdamW는 가중치 감쇠를 손실 함수에 대한 최적화 단계에서 분리하여 가중치 감쇠를 L_2 정규화와 동일하게 만든다. 가중 샘플링은 MOLI[86]에 제안된 대로 TCGA 데이터 세트에 적용된다. 학습률을 제어하기 위해서 코사인 담금질(consine annealing) 기반의 학습 스케줄러[114]를 사용했다 . 여기서 워밍업 단계는 전체 학습 단계의 5%이고 최소 학습률은 초기 학습률의 10%이다. PReLU(Parametric Rectified Linear Unit)[115]이 활성화 함수로 사용된다.

3.3.2 데이터 구축 및 전처리

3.3.2.1 TCGA 및 PDX 데이터

본 연구에서는 예측 모델의 임상 적용 가능성을 위해 PDX 마우스 모델과 TCGA 환자 데이터 세트를 사용했다. PDX 데이터 세트에서 약 300개의 PDX 모델이 30개 이상의 약물로 스크리닝 되었다[79]. TCGA 데이터[79]에는 다양한 암 유형을 가진 10,000명 이상의 환자로부터 얻은 종양 샘플의 프로파일이 포함되어 있으며 임상 주석에서 일부 환자에 대한 약물 반응을 사용했다[98].

SOTA 모델인 MOLI의 방법론에 따라서 다중 오믹스 데이터

세트와 약물 반응 데이터를 얻었다. 다중 오믹스 프로파일은 각 약물에 대한 유전자 발현, 체세포 돌연변이 및 유전자 복제수 변이 프로파일로 구성되었다. 각 오믹스 프로파일 데이터 테이블의 행은 유전자를 나타내고 열은 환자 또는 마우스 모델을 나타내며 데이터 값은 유전자 수준의 특질을 나타낸다. 모든 오믹스 프로파일에 공통적으로 사용되는 유전자를 활용하여 TCGA에서 약 15,000개의 유전자와 PDX에서 13,000개의 유전자를 생성했다. PDX에는 4가지 약물이 있고 TCGA의 3가지 약물이 있으며, Gemcitabine은 TCGA와 PDX 모두에서 사용된다.

3.3.2.2 암세포주 데이터

훈련 및 검증 세트에 CCLE와 GDSC를 사용했다. C C L E 데이터베이스는 수천 개의 암세포주에 대한 유전자 발현, 체세포 돌연변이, 복제 수 데이터를 제공한다. 본 연구에서는 체세포 돌연변이 및 유전자 수준 복사 수 데이터를 DeMap 포털에서 다운로드 했다. 다중 배열 평균(RMA) 방법으로 정규화된 유전자 발현 프로파일은 Paccmann pytoda[116]에서 다운로드 되었다.

Scikit-learn 라이브러리[117]를 통해 유전자 발현 값의 평균을 제거하고 단위 분산으로 조정하여 표준 점수로 변환시켰다. 체세포 돌연변이 데이터는 돌연변이가 있는 유전자에 1을 할당하고 나머지에 0을 할당했다. 다중 오믹스 프로파일에서 유전자를 교차시켜 14,070개의 유전자를 추출했다. 또한 유전자 발현 데이터 세트에서

631개의 세포주를 추출하고 누락된 세포주가 0으로 채워진 데이터를 추출했다. 그 결과, 각 오믹스 프로파일에는 14,070개의 유전자 수준 특질을 가진 631개의 세포주가 표 형식으로 표현되었다.

GDSC 데이터베이스는 대규모 약물 스크리닝 데이터에 대한 약물 반응성 데이터를 제공한다. 이 데이터는 265개 약물과 1001개 암세포주에서 이진화된 IC50 값으로 구성되어 있다. 다중 오믹스 프로파일과 약물 반응에서 추출된 세포주를 기반으로 총 631개 세포주에서 265개 약물로 테스트한 134,846개의 IC50 값이 사용되었다.

3.3.3 모델 평가 방법

3.3.3.1 대조 모델

본 연구에서는 기준 모델을 구축하기 위해서 초기 및 중기 통합 모델을 사용했다. 첫째, MOLI는 PDX 및 TCGA에 대한 CDR 예측을 위한 SOTA 모델이다. MOLI는 $\mathrm{D_{exp}}, \mathrm{D_{mut}}$ 및 $\mathrm{D_{cn}}$를 사용하고 각 오믹스 프로파일에 대한 중간 특질을 생성하기 위해 AE를 구축했다. MOLI는 중간 특질을 연결하여 예측 모듈에 제공했다. MOLI는 성능을 향상시키기 위해 삼중 손실(triple loss) 기법을 사용한다. 이 방법은 유사한 데이터와 유사하지 않은 데이터 사이에 최소한의 거리를 두도록 한다. 특히 MOLI의 삼중 손실 방식은 한 배치 내에서 적어도 하나의 클래스가 다른 경우에만 작동하기 때문에 이 손실 기법의 샘플링

단계에서는 두 개의 클래스가 필요하다. 따라서 본 연구에서는 MOLI를 훈련하는 이 기법의 효율성을 극대화하기 위해서 배치 크기를 최대한 크게 설정했고 이를 통해 각 배치 세트에서 두 클래스를 검색할 확률을 높였다. 본 연구는 MOLI의 TCGA 및 PDX의 ROC와 AUPR 성능을 참조하기 위해서 해당 연구자가 출간한 성능 결과를 그대로 사용했다.

둘째, 초기 통합 모델 Early_multi로 $\mathbf{D_{exp}}, \mathbf{D_{mut}}$ 및 $\mathbf{D_{cn}}$를 통합했다. 이 모델에 GCMC의 모든 최적화 방법을 테스트하여 모델 아키텍처에 따른 성능 차이를 정확하게 비교했다. 예를 들어 최적화 방법에는 학습률 스케줄러, 최적화기법, 활성화 함수 및 드롭아웃이 있다. 또한 모델의 복잡도를 다른 모델과 유사하게 만들기 위해 첫 번째 은닉층의 노드를 256개로 하였다.

셋째, 모든 모델에 대한 단일 오믹스 모델을 구축하여 통합 오믹스 모델과의 성능차이를 살펴본다. 많은 연구[4], [98], [99]에 따르면 유전자 발현 데이터는 CDR 예측을 위한 가장 유망한 오믹스 유형이다. 기존 모델의 구조를 동일하게 보존하기 위해 다른 오믹스 특질은 0으로 상쇄(masking)하여 단일 오믹스 모델을 구성했다. 단일 오믹스 모델은 GCMC_exp, MOLI_exp 및 Early_exp와 같이 모델 이름에 '_exp'를 접미사로 사용했다.

3.3.3.2 성능 지표 및 모델 평가

두 가지 성능 지표를 사용하여 모델의 성능을 평가했다. ROC-AUC과 정밀도-재현율(precision-recall) AUC (AUPR)이다. 만약 데이터의 긍정과 부정의 비율이 불균형 하다면 AUPR이 ROC보다 성능을 평가할 때 더 정확하게 성능차이를 측정할 수 있다. 모델 성능 간의 차이가 통계적으로 유의한지 여부를 확인하기 위해 paired t-테스트를 수행했다.

TCGA, PDX 및 GDSC를 포함한 다양한 데이터 세트에서 모델에 대한 포괄적인 평가를 수행했다. 각 약물에 따라서 모델을 따로 구축했기 때문에, 모든 데이터 세트는 약물 유형을 기반으로 구축되었으며 모델은 한 번에 각 약물 유형에 대해 학습되었다. 첫째, TCGA 및 PDX 데이터 세트의 테스트 세트에 대한 모델을 평가했다. 모델의 하이퍼 파라미터를 탐색하기 위해 GDSC 및 CCLE의 훈련 세트를 사용하여 계층화된 5중(stratified 5-fold) 교차 검증을 수행했다. 그런 다음 모든 훈련 세트에서 모델을 훈련하고 테스트 세트에서 평가했다.

둘째, GDSC 세포주 데이터 세트에 대한 모델을 평가했다. 계층화된 5중 교차검증을 10회 수행하고 결과를 평균화하여 성능을 평가했다. 각 약물 작업에 대해 최고 성능의 모델을 평가했다. 최적의 하이퍼 파라미터를 효율적으로 발견하기 위해 각 약물에 대한 양성률을 기준으로 265개의 약물을 8개의 하위 그룹으로 분류했다.

셋째, 단일 오믹스 프로파일에 대한 다중 오믹스 프로파일 통합의 성능 향상 효과를 평가했다. 유전자 발현 데이터가 CDR 예측에 가장 유망한 오믹스 유형이기 때문에 유전자 발현 프로파일을 사용하여 단일 오믹스 모델을 구축하고 성능을 측정했다. 예를 들어, GCMC, MOLI 및 Early_multi는 ROC와 AUPR으로 각각 GCMC_exp, MOLI_exp 및 Early_exp와 비교되었다.

3.3.3.3 오믹스 종류별 기여비율

본 연구에서는 각 약물의 반응성을 예측하는 과정에서 오믹스 유형의 기여 비율을 측정했다. 먼저 IG 알고리즘을 사용하여 모델의 입력 데이터와 예측 결과 간의 관계를 나타낸다. IG 알고리즘은 모델의 특질에 대한 기여도 값을 산출했으며, 이를 활용하여 오믹스 유형의 비율을 측정할 수 있다.

기여 비율 행렬 $\mathrm{R}^{\mathrm{D}} \in \mathbb{R}^{\mathrm{D}\times\mathrm{O}}$은 기여도가 가장 높은 유전자 100개가 소속된 오믹스 타입을 기준으로 생성되었으며, D개의 약물과 O개의 오믹스 타입으로 구성된다. 또한, 특정 오믹스 타입 k에 대한 기여 비율 벡터도 $\mathrm{r}_{\mathrm{k}}^{\mathrm{D}} \in \mathbb{R}^{\mathrm{D}}$으로 표현할 수 있다. 예를 들어서, $\mathrm{r}_{\mathrm{exp}}^{\mathrm{D}}$, $\mathbf{r}_{\mathbf{mut}}^{\mathbf{D}}$, $\mathbf{r}_{\mathbf{cn}}^{\mathbf{D}}$ 은 각각 유전자 발현, 체세포 돌연변이, 복제수변이를 위한 기여 비율 벡터를 나타낸다. 이에 더해서, $\mathrm{ratio}_{\mathrm{exp}}$, $ratio_{mut}$, $ratio_{cn}$ 은 특정 단일 약품에 대한 각 오믹스 타입의 기여율을 나타낸다.

또한 R^{D}의 IQR(interquantile range)을 측정하여 모델이 각 약물의

성능을 향상시키기 위해 최적의 오믹스 유형을 얼마나 유연하게 활용했는지 평가한다. 높은 IQR은 작업이 변경될 때 모델이 오믹스 유형의 기여 비율을 크게 조정한다는 것을 나타낸다. 또한 오믹스 유형의 성능과 기여 비율 간의 관계를 시각화하기 위해 산점도를 생성했다. y축에는 오믹스 유형의 기여 비율을, x축에는 성능을 표시했다. 산점도는 D 약물에 대해 $D \times O$ 개의 데이터 포인트가 표현되며 각 약물에 대해서 O 개의 오믹스 타입의 데이터 포인트가 기여도의 비율을 나타낸다.

3.4. 연구 결과

3.4.1 TCGA 및 PDX 데이터 기반의 성능 비교

TCGA 및 PDX 데이터 세트를 사용하여 GCMC를 기준 모델과 비교했다. 표 4에서 보는 바와 같이, GCMC는 ROC와 AUPR 모두에서 최고의 성능을 달성했다. 특히, PDX 데이터 세트에서 GCMC와 MOLI 사이에는 높은 성능 격차가 있다. PDX 데이터 세트는 평균 양성 비율의 16%로 매우 불균형 했으며, 여기서 양성 비율은 표 4의 Random classifier로 표시된다. 예를 들어, PDX의 Gemcitabine, Erlotinib, Paclitaxel 및 Cetuximab에 대해 양성률이 0.28에서 0.08로 감소하는 반면에 AUPR의 성능 격차는 1.1배에서 4.3배로 증가했다. 표 4에서 보는 바와 같이 GCMC와 비교모델의 성능차이는 $p-value<0.05$로

통계적인 유의성이 있다.

둘째, 다중 및 단일 오믹스 모델을 비교했다. GCMC는 ROC 및 AUPR 측면에서 GCMC_exp를 능가했으며, 그 성능 차이는 통계적인 유의성을 보였다. 그러나 다른 모델에서는 단일 오믹스 모델이 때때로 다중 오믹스 모델보다 높은 성능을 보였다. 예를 들어 MOLI에서 TCGA의 Docetaxel과 Cisplatin의 다중 오믹스 모델의 ROC 값은 단일 오믹스 모델의 ROC 값보다 작다. Early_multi에서 Docetaxel과 Paclitaxel의 다중 오믹스 모델의 ROC와 AUPR 값은 단일 오믹스 모델보다 작았다.

표 4. TCGA 및 PDX에서 평가된 모델의 성능 비교

AUPR	TCGA (patients)			PDX (mice models)				p-value
	Docetaxel	Cisplatin	Gemcitabine	Paclitaxel	Gemcitabine	Cetuximab	Erlotinib	
Random classifier	0.50	0.91	0.37	0.12	0.28	0.08	0.14	0.001
Early_exp	0.51	0.89	0.44	0.11	0.34	0.09	0.22	0.001
GCMC_exp	0.55	0.92	0.43	0.32	0.43	0.31	0.24	0.005
Early	0.49	0.92	0.44	0.10	0.37	0.09	0.29	0.001
MOLI	0.49	0.93	0.45	0.24	0.49	0.11	0.33	0.005
GCMC	**0.67**	**0.98**	**0.68**	**0.45**	**0.53**	**0.47**	**0.60**	

ROC	TCGA (patients)			PDX (mice models)				p-value
	Docetaxel	Cisplatin	Gemcitabine	Paclitaxel	Gemcitabine	Cetuximab	Erlotinib	
Early_exp	0.58	0.68	0.62	0.58	0.48	0.48	0.45	0.005
MOLI_exp	0.63	0.75	0.64	0.69	0.52	0.51	0.39	0.034
GCMC_exp	0.61	0.68	0.63	0.72	0.50	0.74	0.52	0.011
Early	0.55	0.63	0.63	0.55	0.62	0.51	0.61	0.003
MOLI	0.58	0.66	0.65	0.74	0.64	0.53	0.63	0.019
GCMC	**0.66**	**0.78**	**0.70**	**0.77**	**0.69**	**0.83**	**0.80**	

3.4.2 세포주 데이터 기반의 성능 비교

265개 약물의 대규모 검증 세트에서 GCMC를 평가했다. 첫째, GCMC는 그림 13(A)와 같이 대부분의 약물에 대해 가장 높은 성능의 모델이었다. 예를 들어, GCMC는 AUPR, ROC 및 F1 점수 측면에서 각각 204개 약물(77.0%), 200개 약물(75.5%) 및 212개 약물(79.2%)에서 비교 모델의 성능을 능가했다. GCMC와 타 모델들 간의 성능차이는 통계적으로 유의한 차이가 있었다.

둘째, 다중 오믹스 모델과 단일 오믹스 모델을 비교하여 다중 오믹스 통합의 성능 향상 효과를 측정했다. 그림 13(B)에서 볼 수 있듯이 GCMC는 ROC, AUPR 및 F1의 성능에서 GCMC_exp보다 높은 성능을 보였다. 반면에 MOLI는 AUPR과 F1 점수에서는 통계적으로 유의한 차이가 없었다. 특히 Early_multi와 Early_exp의 성능은 모든 지표에서 비슷했으며 통계적으로 유의한 차이도 없었다. 그림 13(B)에서 볼 수 있듯이, 모델 간의 상자 플롯 위의 기호 *는 통계적으로 유의성 있는 차이를 나타내는 p-value가 0.05 미만임을 나타낸다.

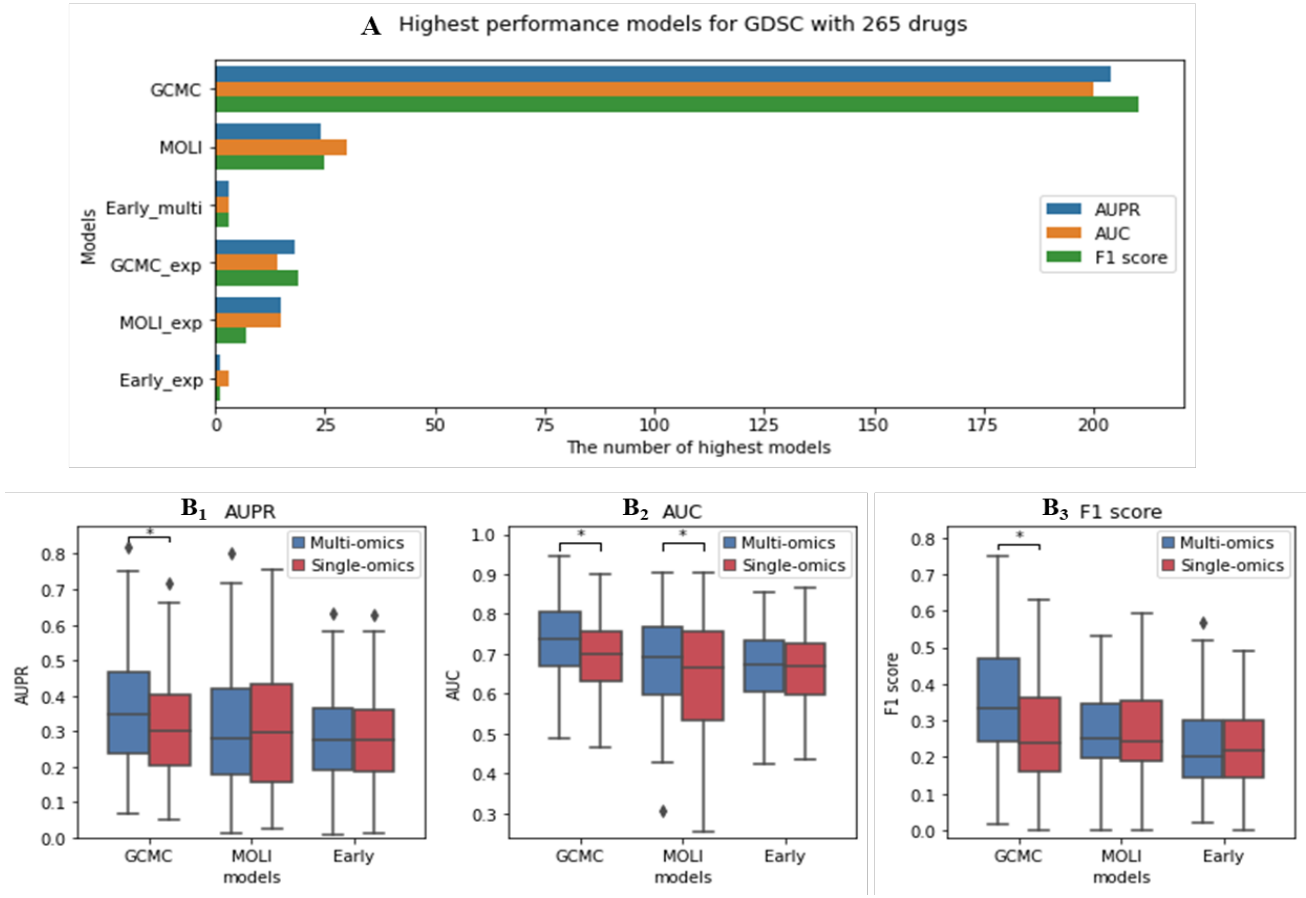

그림 13. 265개의 약물로 평가된 모델의 성능 비교

3.4.3 오믹스 종류별 학습 기여도 비교

본 연구는 IG 알고리즘을 통해 오믹스 유형의 기여도를 분석하여 오믹스 유형을 적용하기 위한 모델 유연성을 평가했다. GDSC 데이터 세트는 265개 약품에 대해 $\mathbf{R}^{265}$의 기여 비율 행렬이 추출되었고 TCGA 및 PDX 데이터 세트는 7개 약품에 대해 $\mathbf{R}^{7}$ 의 기여 비율 행렬이 추출되었다.

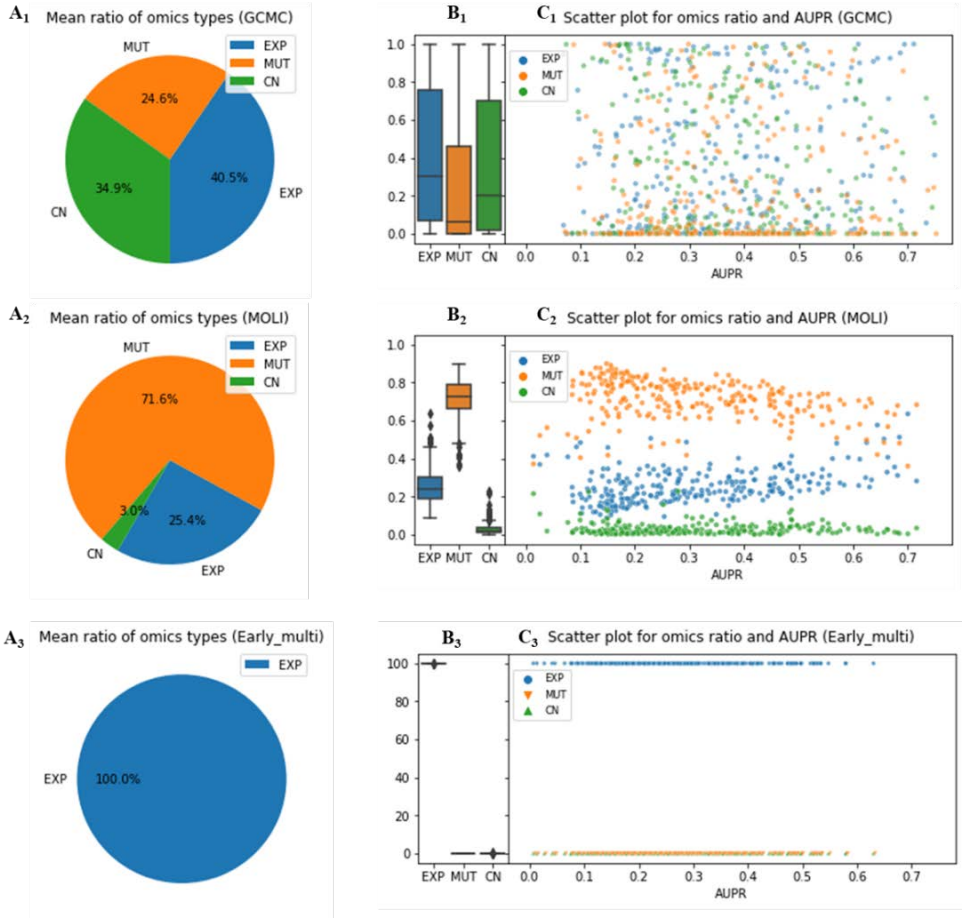

그림 14. 오믹스 유형별 기여 비율 비교

첫째, 본 연구에서는 각 모델들이 얼마나 유연하게 오믹스 유형을 조합하여 예측 성능을 향상시키는지 분석했다. 그림 14에서 $\mathbf{R^{265}}$를 파이 플롯과 상자 플롯으로 요약했다. GCMC에서는 그림 14(A1)에서와 같이 각 오믹스 유형의 평균 기여율이 50%를 넘지 않았다. 반면에 다른 모델에서는 평균적으로 단일 오믹스 유형이 우세했다. 예를 들어, MOLI의 평균 $\mathbf{r_{mut}^{265}}$는 그림 14(A2)와 같이 71.6%이고 Early_multi의 평균 $\mathbf{r_{exp}^{265}}$는 그림 14(A3)와 같이 약 100%였다. 또한, $\mathbf{R^{265}}$의 IQR을 측정하여 모델 유연성을 분석했다. 각 약물 종류가 달라질 때 모델이 사용하는 오믹스 유형의 비율도 크게 조정되면, 높은 IQR 값이 측정된다. 그림 14(B)와 같이, GCMC의 IQR은 최소 0.46인 반면 MOLI의 IQR은 최대 0.12였으며 Early_multi는 모두 0이었다.

둘째, 오믹스 유형과 성능 간의 관계를 조사했다. 그림 14(C)에서 보여지듯이 $\mathbf{R^{265}}$과 모델의 AUPR 성능을 통합하여 분석하고 그것을 산점도로 시각화했다. GCMC의 기여 비율은 단일 오믹스 유형으로 편향되지 않은 반면 MOLI 및 Early_multi는 단일 오믹스 유형으로 편향되었다. 예를 들어 MOLI는 거의 모든 약물의 AUPR 성능 지표에서 $\mathbf{r_{mut}^{265}}$가 가장 컸다. Early_multi는 모든 약물의 모든 AUPR 성능 지표에서 $\mathbf{r_{exp}^{265}}$가 가장 더 높았다. 이러한 결과는 Early_multi와 Early_exp 간의 성능 차이가 ROC 및 AUPR 측면에서 p-value>0.05로 통계적 유의성이 없다는 실험 결과를 뒷받침한다. GCMC는 각 약물 종류에 대한 성능을 최적화하기 위해 다양한 비율의 다중 오믹스 유형을 유연하게 활용했다.

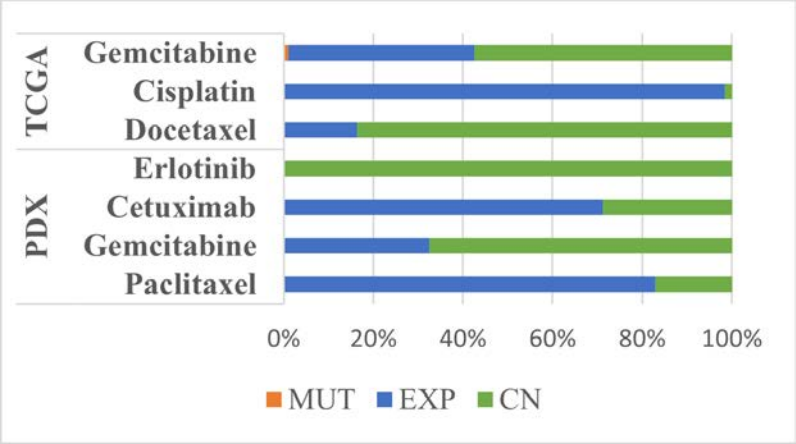

그림 15. TCGA 및 PDX의 오믹스 유형별 기여 비율

셋째, TCGA와 PDX의 7개 약품을 기준으로 GCMC의 기여율 행렬 $\mathbf{R}^7$ 를 측정하였으며, 그림 15와 같이 $ratio_{exp}$ 와 $ratio_{cn}$ 가 주로 사용되었다. 예를 들어 Paclitaxel, Cetuximab 및 Cisplatin에서 $ratio_{exp}$는 70% 이상이었고 $ratio_{cn}$는 Erlotinib 및 Docetaxel에서 80% 이상이었다. 특히, TCGA와 PDX의 Gemcitabine에서 $ratio_{exp}$ 와 $ratio_{cn}$사이의 균형 잡힌 기여 비율이 관찰되었다. 예를 들어, TCGA의 Gemcitabine에서 비율은 57.4% $ratio_{cn}$ 및 41.7% $ratio_{exp}$ 로 구성되었다. PDX의 Gemcitabine에서 비율은 67.5% $ratio_{cn}$ 및 32.4% $ratio_{exp}$로 구성된다.

3.5 고찰

본 연구에서는 CDR 예측을 위한 다중 오믹스 데이터를 통합하기 위한 GCMC 구조를 제안한다. GCMC는 이전 CNN 기반 모델들에서 발생했던 유전자의 순서 불변 문제[91]를 해결하도록 신중하게 설계되었다. GCMC는 CDR 예측 모델 중 처음으로 유전자의 순서에 불변성을 가진 CNN 구조이다. GCMC는 다중 오믹스 프로파일을 3D 표현형으로 변환하고 컨볼루션 인코더(CE)를 통해 유전자 수준에서 다중 오믹스 특질을 통합한다. GCMC는 GDSC의 265개 약물 중 75% 이상에서 비교 모델보다 우수한 성능을 보였다. 또한 GCMC는 ROC와 AUPR 모두에서 TCGA와 PDX의 비교모델보다 우수한 성능을 보였다. 이에 더해서, GCMC는 다양한 비율의 다중 오믹스의 유형을 유연하게 활용하여 각 약물의 성능을 최적화했다. 이러한 결과는 GCMC가 유전자 중심 방식으로 다중 오믹스 프로파일을 통합하여 성능 및 특질 추출 기능을 향상시킬 수 있음을 시사한다.

GCMC와 비교 모델 간의 성능 및 기여도의 차이는 추출된 특질의 세분화의 정도(granularity) 측면에서 설명할 수 있다. GCMC는 CE를 사용하여 각 유전자에 대한 중간 특질을 추출한다. CE의 커널 텐서는 그림 12(A2)와 같이 한 번에 하나의 유전자의 특질을 추출했다. 그와 대조적으로 비교 모델들은 한 번에 전체 유전자 세트의 특질을 포착했다. Early_multi는 그림 11(A)와 같이 모든 오믹스 유형에 대한 모든 유전자로 구성된 연결 벡터를 활용했다. MOLI는 그림 11(B)에서와 같이 중간 표현형을 구축하기 위해 각 오믹스 프로파일에 대한 모든

유전자를 사용했다. 따라서 GCMC의 유전자 중심적인 특질은 성능과 기여도를 향상시키기 위해 경쟁 모델들에 비해 더 세분화된 정보를 제공한다고 볼 수 있다.

본 연구의 실험 결과에 따르면 GCMC는 경쟁 모델들보다 다중 오믹스 데이터 세트의 노이즈와 복잡성을 처리하는데 더 탁월한 것으로 보인다[89]. 예를 들어서, Early_multi는 이 문제에 취약하며, MOLI는 AE를 활용해서 노이즈와 복잡성을 처리하려 했다. 그러나 AE는 FCNN을 기반으로 했기 때문에 MOLI 모델 매개변수의 99.9%를 포함한다. 높은 모델 복잡도의 AE는 노이즈에 취약하게 될 수 있으며 이로 인해 MOLI의 기여 비율이 $ratio_{mut}$ 에 편향된 것으로 보인다. GCMC의 매개변수의 총량은 약 1000만개로 MOLI와 유사하지만, GCMC는 과적합 문제를 해결하기 위해 매개 변수의 비율을 다르게 배분했다. GCMC는 매개변수의 0.1%에 CE를 할당했으며, 이는 각 오믹스 유형에 대한 모든 유전자에 걸쳐 매개변수를 공유한 덕분이다. 매개변수 공유(parameter sharing)는 딥러닝 모델을 정규화하는 효과적인 방법이다[64]. 따라서 GCMC는 CE가 유전자 중심의 특질이 더적은 노이즈를 갖도록 정제하였고, 예측 모듈에서 남은 99.9%의 매개변수를 활용하여 유전자 특질과 약물 반응 사이의 매우 복잡한 패턴을 학습했다. 요약하면 GCMC는 CNN의 개념과 아키텍처를 다중 오믹스 프로파일에 신중하게 적용하여 더 나은 성능과 기여도를 달성한 것으로 보인다.

또한 본 연구에서는 새로운 생물의학 발견으로 이어질 수 있는 탐색

전략을 제안한다. 유전자 중심 특질은 잠재적 표적 단백질, 생물학적 크로스토크 (cross talk) 및 약물 조합과 같은 유전자 또는 약물 간의 새로운 상호 작용을 탐색하는 데 사용할 수 있다. 먼저 IG 알고리즘을 활용하여 유전자의 우선 순위를 지정하고, 높은 순위의 유전자와 항암 약물을 쌍으로 입력하여 문헌 검색을 수행했다. 예를 들어, GCMC는 Erlotinib에 대해 PFKFB3 및 EGFR과 같은 최상위 유전자를 추출했다. Erlotinib은 그 타겟인 EGFR에 고유 또는 후천 저항[120], [121]이 있으나, 최근 연구[118], [119]에서는 PFKFB3와 EGFR 사이의 새로운 크로스토크를 확인하여 대한 Erlotinib의 감도를 향상시키는 것으로 확인했다. 또한, GCMC는 AKR1B10과 FGFR3을 Dasatinib의 상위 유전자로 선정했다. 최근 연구[122]에서는 AKR1B10이 다우노루비신 (daunorubicin)의 AKR1B10 매개 대사를 억제하여 오프 타겟 효과를 갖는 Dasatinib의 새로운 표적으로 확인했다. 또 다른 연구[123]는 FGFR3 변이를 가진 요로상피암 환자의 내재적 약물 내성을 극복하기 위해서 Dasatinib과 FGFR 억제제의 약물 조합에 대한 전임상 증거를 제공한다.

마지막으로 다음과 같은 향후 연구 방향을 제안한다. 첫째, 생물학적 네트워크를 사전 지식으로 사용할 수 있다. GCMC는 유전자 중심의 특질을 추출하지만, 모든 유전자가 서로 직접 상호작용할 수 있다는 가정 하에 예측 모듈에 병합된다. 그러나 유전자는 생물학적 네트워크에서 서로 상호 작용한다. 따라서 그 네트워크를 유전자 특질 추출을 위한 템플릿으로 사용하여서 다중 오믹스 프로파일을 다중

오믹스 네트워크로 구축할 수 있다. 예를 들어서, 해당 네트워크는 단백질-단백질 상호작용(protein-protein interaction, PPI)과 같은 생물학적 도메인 지식을 활용한다. 또한, 다중 채널 오믹스 네트워크를 통합하기 위해 그래프 신경망(graph neural networks, GNN) 모델을 활용할 수 있다. 최근에는 PPI[124], [125] 에 대한 노드 분류와 같은 생물학적 네트워크를 평가하기 위해 다양한 GNN 모델이 개발되었다. 예를 들어서, PPI[126]에 대한 링크 예측 및 연결 속성 예측[127], 질병 경로 발견[128], 그리고 다세포 기능 예측[129] 이 있다.

둘째, GCMC는 multi-task learning(MTL)으로 모델을 구성하여 학습의 효율을 높일 수 있다. MTL은 동일한 입력 데이터를 학습하여 다양한 목표를 학습한다[130]. 현재의 벤치마크 데이터 셋은 약물마다 학습 데이터가 구성되어 있고, 각 약물마다 예측모델을 학습해야 하기에 학습의 효율이 낮다. 따라서, 동일한 세포주 데이터를 입력하고 그에 해당되는 다양한 약물 반응성을 예측하는 MTL 기반의 모델은 약물의 수만큼 학습의 효율성을 높일 수 있다.

3.6 소결론

본 연구에서는 CDR 예측을 위한 다중 오믹스 통합 모델로 GCMC를 제안했다. GCMC는 다중 오믹스 프로파일을 3차원 표현형으로 변환하고 컨볼루션 인코더(CE)를 사용하여 유전자 수준에서 다중 오믹스 특질을 통합한다. CE는 유전자 중심 특질 벡터를 추출하기 위해 각 유전자에 대한 다중 오믹스 레이어를 통합하여 여러 중간 특질을

생성한다. 예측 모듈은 유전자의 가중치를 학습하고 CDR 예측을 위해 축소된 차원에 유전자의 특질을 삽입한다. GCMC는 GDSC의 265개 약물 중 75% 이상에서 단일 오믹스 모델을 포함한 기준 모델보다 우수한 성능을 보였고 TCGA와 PDX의 데이터에서 비교모델보다 우수한 성능을 보였다.

이에 더해서 본 연구는 설명가능한 인공지능(XAI) 기법을 활용하여 약물 반응성에 대한 입력 데이터의 기여도를 분석했다. GCMC는 최적의 오믹스 유형을 유연하게 활용하여 각 약물 작업에 대한 성능을 향상시켰다. 이러한 결과는 GCMC가 유전자 중심 방식으로 다중 오믹스 프로파일을 통합하여 성능 및 특질 추출 기능을 향상시킬 수 있음을 보인다. 또한, GCMC의 유전자 중심 특질 추출 기법은 주어진 약물에 중요한 유전자를 선별하였으며, 해당 유전자들이 잠재적인 약물의 타겟 또는 생물학적 크로스토크(cross talk)가 될 수 있음을 문헌적으로 검증했다.

IV. 결　　론

새로운 약물이 원하는 효능을 갖도록 설계하는 것은 제약 산업에서 여전히 어려운 과제로 남아있으며 비용 집약적인 과정이 요구된다. 약물-타겟 단백질 상호작용을 정의하는 것은 신약 탐색과 신약 재창출을 위해 중요한 분석이지만 모든 가능한 모든 경우의 수를 탐색하는 것은 매우 비용 집약적인 작업이다. 이에 더해서 항암제 개발에서는 종양 이질성이 미치는 임상적인 차이를 극복하기 위해서 종양의 다양한 오믹스 데이터의 층위를 종합하여 고려하는 것이 필요하다. 이를 개선하고 약물 탐색의 효율성을 높이기 위해 인공지능 모델들이 활용되고 있으나, 기존의 모델들은 특질의 단일한 표현형을 사용하거나 단일한 오믹스 층위를 사용한다는 한계가 있었다.

본 연구에서는 이를 해결하기 위해서 약물-타겟 단백질 예측을 위한 쌍기반 다중채널(MCPINN) 모델과 항암제 반응성 예측을 위한 유전자 중심의 다중채널(GCMC) 모델을 제안했다. MCPINN은 DNN의 3가지 접근 방식인 분류기, 특질 추출기, 종단 간 학습기를 활용하여 표현 학습(representation learning) 능력을 극대화한다. MCPINN은 특질의 다양한 표현형을 다중 채널에 입력하여 활용하고 그 특질을 상보적으로 통합했다. MCPINN은 모델의 성능과 학습속도에서 가장 높은 성능을 보였다. 이에 더해서, MCPINN은 전이학습을 활용하여 독성 예측의 성능을 개선했다.

GCMC는 CDR 예측에 적용된 최초로 유전자의 순서에

불변성(order-invariant)을 가진 컨볼루션 신경망 구조이다. GCMC는 다중 오믹스 프로파일을 3D 표현형으로 변환하고 컨볼루션 인코더를 사용하여 유전자 수준에서 다중 오믹스 특질을 통합한다. GCMC는 GDSC의 265개 약물 중 75% 이상에서 단일 오믹스 모델을 포함한 기준 모델보다 우수한 성능을 보였다. 또한 모델의 임상적 적용 가능성에서 ROC와 AUPR 모두에서 GCMC가 TCGA와 PDX에 대한 비교 모델보다 우수한 성능을 보였다. 이에 더해서 GCMC는 최적의 오믹스 유형을 유연하게 활용하여 각 약물 작업에 대한 성능을 향상시킬 수 있다. 이러한 결과는 GCMC가 유전자 중심 방식으로 다중 오믹스 프로파일을 통합하여 성능 및 특질 추출 기능을 향상시킬 수 있음을 시사한다.

본 연구에서 구축하는 두 모델은 신약개발 프로세스의 관점에서도 상보적인 관계를 갖는다. MCPINN은 의도한 타겟(on-target)의 상호작용을 예측하는 반면, GCMC는 의도하지 않은 타겟(off-target)의 약물 반응성을 탐색할 수 있다. 예를 들어서, GCMC는 세포 내의 많은 유전자들과 단백질들에 미치는 약물 반응성을 분석하여 해당 약물에 중요한 유전자를 추출한다. 이에 기반하여 잠재적인 약물의 타겟 및 생물학적 크로스토크(cross talk)를 탐색하고 이를 문헌적으로 검증했다. 본 연구결과로 다중채널 인공지능 구조가 신약개발 프로세스를 개선하고, 항암 치료를 위한 새로운 바이오마커 탐색에 도움이 될 것으로 기대한다.

참고 문헌

[1] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington, “Diagnosing the decline in pharmaceutical R&D efficiency,” *Nat. Rev. Drug Discov.*, vol. 11, no. 3, pp. 191–200, 2012.

[2] M. S. Ringel, J. W. Scannell, M. Baedeker, and U. Schulze, “Breaking Eroom’s law,” *Nat. Rev. Drug Discov.*, vol. 19, no. 12, pp. 833–835, 2020.

[3] G. Papadatos, A. Gaulton, A. Hersey, and J. P. Overington, “Activity, assay and target data curation and quality in the ChEMBL database,” *J. Comput. Aided. Mol. Des.*, vol. 29, no. 9, pp. 885–896, 2015.

[4] F. Iorio *et al.*, “A Landscape of Pharmacogenomic Interactions in Cancer,” *Cell*, vol. 166, no. 3, pp. 740–754, Jul. 2016.

[5] A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Do, “Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases,” *Brief. Bioinform.*, pp. 1–36, 2018.

[6] R. Chen, X. Liu, S. Jin, J. Lin, and J. Liu, “Machine Learning for Drug-Target Interaction Prediction,” *Molecules*, vol. 23, no. 9, p. 2208, 2018.

[7] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, “Similarity-

based machine learning methods for predicting drug−−target interactions: a brief review," *Brief. Bioinform.*, vol. 15, no. 5, pp. 734–747, 2013.

[8] S. Kim *et al.*, "PubChem substance and compound databases," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1202−−D1213, 2015.

[9] W. P. Walters, "Virtual Chemical Libraries: Miniperspective," *J. Med. Chem.*, vol. 62, no. 3, pp. 1116–1124, 2018.

[10] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.−L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database GDB−17," *J. Chem. Inf. Model.*, vol. 52, no. 11, pp. 2864–2875, 2012.

[11] U. Consortium, "UniProt: a hub for protein information," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D204−−D212, 2014.

[12] D.−S. Cao *et al.*, "Large−scale prediction of drug−−target interactions using protein sequences and drug topological structures," *Anal. Chim. Acta*, vol. 752, pp. 1–10, 2012.

[13] M. Gönen, "Predicting drug−−target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.

[14] T. Scior *et al.*, "Recognizing pitfalls in virtual screening: a

critical review," *J. Chem. Inf. Model.*, vol. 52, no. 4, pp. 867–881, 2012.

[15] J.-L. Reymond, R. Van Deursen, L. C. Blum, and L. Ruddigkeit, "Chemical space as a source for new drugs," *Medchemcomm*, vol. 1, no. 1, pp. 30–38, 2010.

[16] H. Li *et al.*, "TarFisDock: a web server for identifying drug targets with docking approach," *Nucleic Acids Res.*, vol. 34, no. suppl_2, pp. W219--W224, 2006.

[17] L. Xie, T. Evangelidis, L. Xie, and P. E. Bourne, "Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir," *PLoS Comput. Biol.*, vol. 7, no. 4, p. e1002037, 2011.

[18] L. Yang *et al.*, "Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome—clozapine-induced agranulocytosis as a case study," *PLoS Comput. Biol.*, vol. 7, no. 3, p. e1002016, 2011.

[19] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nat. Biotechnol.*, vol. 25, no. 2, p. 197, 2007.

[20] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork,

"Drug target identification using side-effect similarity," *Science (80-.).*, vol. 321, no. 5886, pp. 263–266, 2008.

[21] A. Koutsoukas *et al.*, "From in silico target prediction to multi-target drug design: current databases, methods and applications," *J. Proteomics*, vol. 74, no. 12, pp. 2554–2574, 2011.

[22] G. J. P. van Westen, J. K. Wegner, A. P. IJzerman, H. W. T. van Vlijmen, and A. Bender, "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets," *Medchemcomm*, vol. 2, no. 1, pp. 16–30, 2011.

[23] G. J. V Westen and others, "Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): Comparative study of 13 amino acid descriptor sets," *J Cheminform*, vol. 5, p. 41, 2013.

[24] I. Cortés-Ciriano *et al.*, "Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects," *Medchemcomm*, vol. 6, no. 1, pp. 24–50, 2015.

[25] T. Qiu *et al.*, "The recent progress in proteochemometric modelling: focusing on target descriptors, cross-term descriptors and application scope," *Brief. Bioinform.*, p. bbw004, 2016.

[26] H. Iwata, R. Sawada, S. Mizutani, M. Kotera, and Y. Yamanishi, “Large−scale prediction of beneficial drug combinations using drug efficacy and target profiles,” *J. Chem. Inf. Model.*, vol. 55, no. 12, pp. 2705–2716, 2015.

[27] Z. Li *et al.*, “In silico prediction of drug−target interaction networks based on drug chemical structure and protein sequences,” *Sci. Rep.*, vol. 7, no. 1, p. 11174, 2017.

[28] H. Yabuuchi *et al.*, “Analysis of multiple compound−−protein interactions reveals novel bioactive molecules,” *Mol. Syst. Biol.*, vol. 7, no. 1, p. 472, 2011.

[29] M. Lapinsh, P. Prusis, T. Lundstedt, and J. E. S. Wikberg, “Proteochemometrics modeling of the interaction of amine G−protein coupled receptors with a diverse set of ligands,” *Mol. Pharmacol.*, vol. 61, no. 6, pp. 1465–1475, 2002.

[30] J.−K. Lee *et al.*, “Pharmacogenomic landscape of patient−derived tumor cells informs precision oncology therapy,” *Nat. Genet.*, vol. 50, no. 10, pp. 1399–1411, 2018.

[31] D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: the next generation,” *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[32] M. J. Garnett *et al.*, “Systematic identification of genomic markers of drug sensitivity in cancer cells,” *Nature*, vol. 483, no. 7391, pp. 570–575, 2012.

[33] J. Qu, X. Chen, Y.−Z. Sun, J.−Q. Li, and Z. Ming, “Inferring potential small molecule−−miRNA association based on triple layer heterogeneous network,” *J. Cheminform.*, vol. 10, no. 1, pp. 1–14, 2018.

[34] X. Chen, N.−N. Guan, Y.−Z. Sun, J.−Q. Li, and J. Qu, “MicroRNA−small molecule association identification: from experimental results to computational models,” *Brief. Bioinform.*, vol. 21, no. 1, pp. 47–61, 2020.

[35] C.−C. Wang, X. Chen, J. Yin, and J. Qu, “An integrated framework for the identification of potential miRNA−disease association based on novel negative samples extraction strategy,” *RNA Biol.*, vol. 16, no. 3, pp. 257–269, 2019.

[36] G. de Anda−Jáuregui and E. Hernández−Lemus, “Computational oncology in the multi−omics era: state of the art,” *Front. Oncol.*, vol. 10, p. 423, 2020.

[37] B. B. Misra, C. Langefeld, M. Olivier, and L. A. Cox, “Integrated omics: tools, advances and future approaches,” *J. Mol. Endocrinol.*, vol. 62, no. 1, pp. R21−−R45, 2019.

[38] B. Miles and P. Tadi, “Genetics, somatic mutation,” 2020.

[39] J. Vijg, “Somatic mutations, genome mosaicism, cancer and aging,” *Curr. Opin. Genet. Dev.*, vol. 26, pp. 141–149, 2014.

[40] A. J. Sharp *et al.*, “Segmental duplications and copy−number

variation in the human genome," *Am. J. Hum. Genet.*, vol. 77, no. 1, pp. 78–88, 2005.

[41] R. Sager, "Expression genetics in cancer: shifting the focus from DNA to RNA," *Proc. Natl. Acad. Sci.*, vol. 94, no. 3, pp. 952–955, 1997.

[42] G. M. Morris *et al.*, "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function," *J. Comput. Chem.*, vol. 19, no. 14, pp. 1639–1662, 1998.

[43] R. A. Friesner *et al.*, "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy," *J. Med. Chem.*, vol. 47, no. 7, pp. 1739–1749, 2004.

[44] M. McGann, "FRED pose prediction and virtual screening accuracy," *J. Chem. Inf. Model.*, vol. 51, no. 3, pp. 578–596, 2011.

[45] I. Wallach, M. Dzamba, and A. Heifets, "AtomNet: A deep convolutional neural network for bioactivity prediction in structure−based drug discovery," *arXiv Prepr. arXiv1510.02855*, 2015.

[46] A. Bender and R. C. Glen, "Molecular similarity: a key technique in molecular informatics," *Org. Biomol. Chem.*, vol.

2, no. 22, pp. 3204–3218, 2004.

[47] F. Nigsch, A. Bender, J. L. Jenkins, and J. B. O. Mitchell, "Ligand−target prediction using Winnow and naive Bayesian algorithms and the implications of overall performance statistics," *J. Chem. Inf. Model.*, vol. 48, no. 12, pp. 2313–2325, 2008.

[48] R. Lowe, H. Y. Mussa, F. Nigsch, R. C. Glen, and J. B. O. Mitchell, "Predicting the mechanism of phospholipidosis," *J. Cheminform.*, vol. 4, no. 1, p. 2, 2012.

[49] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, 2003.

[50] R. Lowe, H. Y. Mussa, J. B. O. Mitchell, and R. C. Glen, "Classifying molecules using a sparse probabilistic kernel binary classifier," *J. Chem. Inf. Model.*, vol. 51, no. 7, pp. 1539–1544, 2011.

[51] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure−−activity relationships," *J. Chem. Inf. Model.*, vol. 55, no. 2, pp. 263–274, 2015.

[52] G. E. Dahl, N. Jaitly, and R. Salakhutdinov, “Multi-task neural networks for QSAR predictions,” *arXiv Prepr. arXiv1406.1231*, 2014.

[53] B. Ramsundar *et al.*, “Is multitask deep learning practical for pharma?,” *J. Chem. Inf. Model.*, vol. 57, no. 8, pp. 2068–2076, 2017.

[54] E. B. Lenselink *et al.*, “Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set,” *J. Cheminform.*, vol. 9, no. 1, p. 45, 2017.

[55] A. Koutsoukas, K. J. Monaghan, X. Li, and J. Huan, “Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data,” *J. Cheminform.*, vol. 9, no. 1, p. 42, 2017.

[56] C. Wang, J. Liu, F. Luo, Y. Tan, Z. Deng, and Q.-N. Hu, “Pairwise input neural network for target-ligand interaction prediction,” in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, 2014, pp. 67–70.

[57] S. Jaeger, S. Fulle, and S. Turk, “Mol2vec: unsupervised machine learning approach with chemical intuition,” *J. Chem. Inf. Model.*, vol. 58, no. 1, pp. 27–35, 2018.

[58] E. Asgari and M. R. K. Mofrad, “Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics,” *PLoS One*, vol. 10, no. 11, p. e0141287, Nov. 2015.

[59] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, “ZINC: a free tool to discover chemistry for biology,” *J. Chem. Inf. Model.*, vol. 52, no. 7, pp. 1757–1768, 2012.

[60] U. Consortium and others, “UniProt: the universal protein knowledgebase,” *Nucleic Acids Res.*, vol. 46, no. 5, p. 2699, 2018.

[61] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies, “How similar are similarity searching methods? A principal component analysis of molecular descriptor space,” *J. Chem. Inf. Model.*, vol. 49, no. 1, pp. 108–119, 2009.

[62] N. T. Program, “Tox21 Challenge.” 2014.

[63] E. S. Olivas, *Handbook of Research on Machine Learning Applications and Trends*. Hershey, PA: IGI Global, 2009.

[64] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[65] R. Gómez-Bombarelli *et al.*, “Automatic Chemical Design

Using a Data-Driven Continuous Representation of Molecules," *ACS Cent. Sci.*, vol. 4, no. 2, pp. 268–276, 2018.

[66] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Nips*, pp. 1–9, 2013.

[67] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 28, no. 1, pp. 11–21, 1972.

[68] A. Van Den Oord *et al.*, "WaveNet: A generative model for raw audio."

[69] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[70] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[71] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, 2005.

[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[73] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv Prepr. arXiv1609.04836*, 2016.

[74] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, Springer, 2012, pp. 9–48.

[75] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv Prepr. arXiv1505.00853*, 2015.

[76] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv Prepr. arXiv1511.07289*, 2015.

[77] Z. Wu *et al.*, "MoleculeNet: a benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, no. 2, pp. 513–530, 2018.

[78] A. M. Saxe *et al.*, "On the information bottleneck theory of deep learning," 2018.

[79] H. Gao *et al.*, "High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response," *Nat. Med.*, vol. 21, no. 11, pp. 1318–1325, 2015.

[80] S. Haider, R. Rahman, S. Ghosh, and R. Pal, "A copula based

approach for design of multivariate random forests for drug sensitivity prediction," *PLoS One*, vol. 10, no. 12, p. e0144490, 2015.

[81] Z. Dong *et al.*, "Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection," *BMC Cancer*, vol. 15, no. 1, pp. 1–12, 2015.

[82] Y. Chang *et al.*, "Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature," *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, 2018.

[83] M. Manica, A. Oskooei, J. Born, V. Subramanian, J. Sáez-Rodr\'\iguez, and M. Rodr\'\iguez Mart\'\inez, "Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders," *Mol. Pharm.*, vol. 16, no. 12, pp. 4797–4806, 2019.

[84] P. Liu, H. Li, S. Li, and K.-S. Leung, "Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–14, 2019.

[85] P. Geeleher *et al.*, "Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies," *Genome Res.*, vol. 27, no. 10, pp.

1743–1751, 2017.

[86] H. Sharifi-Noghabi, O. Zolotareva, C. C. Collins, and M. Ester, “MOLI: multi-omics late integration with deep neural networks for drug response prediction,” *Bioinformatics*, vol. 35, no. 14, pp. i501--i509, 2019.

[87] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, “Methods of integrating data to uncover genotype--phenotype interactions,” *Nat. Rev. Genet.*, vol. 16, no. 2, pp. 85–97, 2015.

[88] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities,” *Inf. Fusion*, vol. 50, pp. 71–91, Oct. 2019.

[89] M. Picard, M.-P. Scott-Boyer, A. Bodein, O. Périn, and A. Droit, “Integration strategies of multi-omics data for machine learning analysis,” *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 3735–3746, 2021.

[90] A. Sathyanarayanan, R. Gupta, E. W. Thompson, D. R. Nyholt, D. C. Bauer, and S. H. Nagaraj, “A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping,” *Brief. Bioinform.*, vol. 21, no. 6, pp. 1920–1936, 2020.

[91] M. Kang, E. Ko, and T. B. Mersha, “A roadmap for multi-omics data integration using deep learning,” *Brief. Bioinform.*, vol. 23, no. 1, p. bbab454, 2022.

[92] J. Martorell-Marugán *et al.*, “Deep learning in omics data analysis and precision medicine,” *Exon Publ.*, pp. 37–53, 2019.

[93] A. Talukder, C. Barham, X. Li, and H. Hu, “Interpretation of deep learning in genomics and epigenomics,” *Brief. Bioinform.*, vol. 22, no. 3, p. bbaa177, 2021.

[94] P. R. Cohen and A. E. Howe, “How evaluation guides AI research: The message still counts more than the medium,” *AI Mag.*, vol. 9, no. 4, p. 35, 1988.

[95] Q. Liu, Z. Hu, R. Jiang, and M. Zhou, “DeepCDR: a hybrid graph convolutional network for predicting cancer drug response,” *Bioinformatics*, vol. 36, no. Supplement_2, pp. i911--i918, 2020.

[96] T. Wang *et al.*, “MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–13, 2021.

[97] X. Li *et al.*, “MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis,” *Front. Genet.*, p. 127, 2022.

[98] Z. Ding, S. Zu, and J. Gu, "Evaluating the molecule-based prediction of clinical drug responses in cancer," *Bioinformatics*, vol. 32, no. 19, pp. 2891–2895, 2016.

[99] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitrodrug sensitivity in cell lines," *Genome Biol.*, vol. 15, no. 3, pp. 1–12, 2014.

[100] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *Ieee Access*, vol. 8, pp. 42200–42216, 2020.

[101] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Natl. Acad. Sci.*, vol. 116, no. 44, pp. 22071–22080, 2019.

[102] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, 2018.

[103] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*, 2017, pp. 3319–3328.

[104] M. Q. Ding, L. Chen, G. F. Cooper, J. D. Young, and X. Lu, "Precision oncology beyond targeted therapy: combining

omics data with machine learning matches the majority of cancer cells to effective therapeutics," *Mol. cancer Res.*, vol. 16, no. 2, pp. 269–278, 2018.

[105] V. Malik, Y. Kalakoti, and D. Sundar, "Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer," *BMC Genomics*, vol. 22, no. 1, pp. 1–11, 2021.

[106] Y. Goldberg and O. Levy, "word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method," *arXiv Prepr. arXiv1402.3722*, no. 2, pp. 1–5, 2014.

[107] Y.-C. Chiu *et al.*, "Predicting drug response of tumors from integrated genomic profiles by deep neural networks," *BMC Med. Genomics*, vol. 12, no. 1, pp. 143–155, 2019.

[108] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[109] Q. Liu, B. Cheng, Y. Jin, and P. Hu, "Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data," *J. Biomed. Inform.*, vol. 125, p. 103958, 2022.

[110] J. Zeng, H. Cai, and T. Akutsu, "Breast cancer subtype by imbalanced omics data through a deep learning fusion model," in *Proceedings of the 2020 10th International Conference on*

Bioscience, Biochemistry and Bioinformatics, 2020, pp. 78–83.

[111] N. Fatima and L. Rueda, "iSOM−GSN: an integrative approach for transforming multi−omic data into gene similarity networks via self−organizing maps," *Bioinformatics*, vol. 36, no. 15, pp. 4248–4254, 2020.

[112] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[113] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv Prepr. arXiv1711.05101*, 2017.

[114] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv Prepr. arXiv1608.03983*, 2016.

[115] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human−level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[116] J. Born, M. Manica, A. Oskooei, J. Cadow, G. Markert, and M. Rodríguez Martínez, "PaccMann^{RL}: De novo generation of hit−like anticancer molecules from transcriptomic data via reinforcement learning," *iScience*, vol. 24, no. 4, p. 102269, 2021.

[117] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in {P}ython,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[118] N. Lypova, S. M. Dougherty, L. Lanceta, J. Chesney, and Y. Imbert-Fernandez, “PFKFB3 Inhibition Impairs Erlotinib-Induced Autophagy in NSCLCs,” *Cells*, vol. 10, no. 7, p. 1679, 2021.

[119] N. Lypova, S. Telang, J. Chesney, and Y. Imbert-Fernandez, “Increased 6-phosphofructo-2-kinase/fructose-2, 6-bisphosphatase-3 activity in response to EGFR signaling contributes to non--small cell lung cancer cell survival,” *J. Biol. Chem.*, vol. 294, no. 27, pp. 10530–10543, 2019.

[120] E. Santoni-Rugiu *et al.*, “Intrinsic resistance to EGFR-tyrosine kinase inhibitors in EGFR-mutant non-small cell lung cancer: differences and similarities with acquired resistance,” *Cancers (Basel).*, vol. 11, no. 7, p. 923, 2019.

[121] Y. Mu *et al.*, “Acquired resistance to osimertinib in patients with non-small-cell lung cancer: mechanisms and clinical outcomes,” *J. Cancer Res. Clin. Oncol.*, vol. 146, no. 9, pp. 2427–2433, 2020.

[122] N. Büküm *et al.*, “Inhibition of AKR1B10-mediated metabolism of daunorubicin as a novel off-target effect for the Bcr-Abl

tyrosine kinase inhibitor dasatinib," *Biochem. Pharmacol.*, vol. 192, p. 114710, 2021.

[123] N. C. Lima, E. Atkinson, T. D. Bunney, M. Katan, and P. H. Huang, "Targeting the Src pathway enhances the efficacy of selective FGFR inhibitors in urothelial cancers with FGFR3 alterations," *Int. J. Mol. Sci.*, vol. 21, no. 9, p. 3214, 2020.

[124] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," in *International Conference on Machine Learning*, 2020, pp. 1725–1735.

[125] M. Ding *et al.*, "VQ-GNN: A Universal Framework to Scale up Graph Neural Networks using Vector Quantization," *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.

[126] K. C. Kishan, R. Li, F. Cui, and A. R. Haake, "Predicting biomedical interactions with higher-order graph convolutional networks," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 19, no. 02, pp. 676–687, 2022.

[127] M. Zhang, P. Li, Y. Xia, K. Wang, and L. Jin, "Labeling Trick: A Theory of Using Graph Neural Networks for Multi-Node Representation Learning," *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.

[128] M. Agrawal, M. Zitnik, and J. Leskovec, "Large-scale analysis of disease pathways in the human interactome."

[129] M. Zitnik and J. Leskovec, “Predicting multicellular function through multi-layer tissue networks,” *Bioinformatics*, vol. 33, no. 14, pp. i190--i198, 2017.

[130] S. Ruder, “An overview of multi-task learning in deep neural networks,” *arXiv Prepr. arXiv1706.05098*, 2017.

Abstract

A study on multi-channel based AI architecture design for drug discovery

Munhwan Lee

Healthcare Management and Informatics

The Graduate School

Seoul National University

Designing new drugs with desired efficacy remains a challenge for the pharmaceutical industry and requires a cost-intensive process. In particular, the efficiency problem of new drug development represented by Eroom's law is in stark contrast to the rapid technological development in other fields. Since the 1950s, the number of new drugs approved by the FDA at $1 billion in R&D costs has halved about every nine years. Recent advances in disease biology and the use of bioinformatics-based biomedical big data have slightly increased the efficiency of new drug development. However, the number of new drugs approved by the FDA with a R&D cost of $1 billion is still limited to less than one.

Recently, attempts to increase the R&D efficiency of new drug development by using predictive models based on artificial

intelligence are increasing. For example, the identification of drug-target protein interactions (DTI) is a basic step in the discovery of drug candidates. DTI plays an important role in various applications such as discovery of new drug candidates, repurposing of drugs, and prediction of off-target or side effects. For this purpose, from traditional machine learning models to modern neural network models are being utilized to predict DTI. However, there is still room for improvement in that the drug candidate or target protein is expressed with only one type of features for each drug and target protein.

In addition, the efficacy and heterogeneity of different anticancer drugs for different cancer patients is a challenging task to be solved in the development of new anticancer drugs. In particular, tumor heterogeneity across the genome, transcriptome and epigenome can impair the efficacy of anticancer drugs. To overcome this, artificial intelligence models for multi-omics integration have been proposed to utilize various levels of biological data. However, existing AI models have limitations in that they are vulnerable to the inherent complexity and noise of biological data because different types of omics data are constructed as a single, one-dimensional feature. Therefore, it is also reported that performance can be worse than when using single omics data.

Therefore, in this study, we propose that the AI model learn

various biological aspects by building the features of each data through muilti-channels. First, for the identification of DTI, a multi-channel paired input neural network (MCPINN) was proposed. MCPINN maximizes representation learning ability by utilizing three approaches of DNN: classifier, feature extractor, and end-to-end learner. MCPINN utilized various levels of features as input into multiple channels and incorporated those features. MCPINN showed the highest performance in performance and training speed. In addition, MCPINN utilizes transfer learning to improve the performance of toxicity prediction.

In addition, in this study, a gene-centric multi-channel (GCMC) architecture was constructed for predicting anticancer drug responsiveness. GCMC transforms multi-omics data into a three-dimensional tensor, and a new dimension expresses the omics type. GCMC can extract gene-centric new features by integrating multi-omics profiles for each gene. GCMC showed better performance than the previous best performing model in 265 cancer cell line data, TCGA patient data, and PDX patient-derived mouse model data. In addition, GCMC can flexibly utilize optimal omics types to improve performance for each drug task. These results suggest that GCMC can integrate multiple omics profiles in a gene-centric manner to improve performance and feature extraction capabilities.

Keywords : Artificial Intelligence, Machine learning, Deep learning, Drug discovery, Drug-target interaction prediction, multi-omics integration
Student Number : 2015-23266