

약물 재창출 가능성을 고려한
생성 및 예측 모델 기반
신약후보물질 발굴 방법론 개발

이규민(석박통합과정)¹, 이창현(석박통합과정)²
울산과학기술원 ¹경영과학과, ²산업공학과
{optimist, messy92}@unist.ac.kr

Abstract (한글)

최근, 인공지능을 활용한 약물개발 연구가 활발히 진행중에 있다. 대표적으로, 신약발굴과 약물재창출 분야는 이미 많은 딥러닝 기반의 연구가 성공적으로 수행되었으며 괄목할 만한 성과들을 배출하고 있다. 해당 분야에서의 딥러닝 연구는 각각 약물 시퀀스 생성모델과 DTI (Drug-Target Interaction) 예측모델 개발이라는 주제로 독립적으로 발전하였다. 그러나, 기존 연구들은 두 분야 사이에 존재하는 학습구조의 유사성과 이를 반영하였을 때 창출될 수 있는 시너지를 간과하였다는 한계가 존재한다. 이에 본 연구는 두 분야의 시너지를 극대화하기 위해 시퀀스 생성과 DTI 예측을 동시에 수행하는 시퀀스-투-시퀀스 생성 모델 및 다층 퍼셉트론 예측 모델 기반 학습 프레임워크를 제안한다. 해당 학습 프레임워크를 통해 우리는 약물 발견 및 재배치 작업을 동시에 처리할 수 있으며, 두 작업이 엔드 투 엔드로 학습되는 과정에서 그라디언트 공유를 통해 개별 작업에서 성능을 향상시켰다.

Abstract (영문)

Recently, research on drug development using artificial intelligence is actively underway. Representative, many deep-learning-based studies have already been successfully carried out in the field of drug discovery and drug repositioning, and remarkable achievements are being produced. Deep learning research in these fields has been independently proceeded by developing drug sequence generation models and DTI (Drug-Target Interaction) prediction models, respectively. However, existing studies have a limitation in that they overlooked the similarity of the learning structure between the two fields thus not considering their possible synergies. This study proposes a learning framework based on the generative model using Sequence-to-Sequence and predictive model using multi-layer perceptron that simultaneously generates a sequence and predicts DTI maximizing the synergy. With this learning framework, we can handle drug discovery and repositioning together, and we have improved performance in individual tasks based on multi-modal end-to-end learning.

목 차

I. 서론	4
II. 관련 연구	6
III. 연구 방법	7
1. 데이터 수집 및 전처리	
2. 신약후보물질 발굴 모델	
3. 모델 성능 평가 방법	
IV. 연구 결과	16
1. 데이터 수집 및 전처리 결과	
2. 신약후보물질 생성 결과 및 모델 성능 평가	
V. 토의 및 결론	21

I. 서론

➤ 연구 배경

- 인공지능 기반 약물개발 (Drug development) 연구 규모의 급격한 증가
 - 스탠포드 인공지능 인덱스 보고서 (The AI index 2021 Annual Report)에 따르면 20년도 가장 많은 민간 AI 투자를 받은 분야는 "약물 설계 및 발견" 분야로 그 규모가 19년도 대비 4.5 배 증가하여 미화 138억 달러에 준함
 - 향후 민간 투자에 더해 정부 투자까지 추가되어 더욱 많은 AI 기반 약물개발 연구가 이루어지리라 예상됨
- 현재 딥러닝 기반 약물개발 연구는 크게 두 분야로 분류됨
 - 신약 발굴 (Drug discovery) 분야 – 기존 타겟 질병의 신규 후보약물 생성에 초점 (생성모델 기반)
 - 약물 재창출 (Drug repositioning) 분야 – 기존 약물의 신규 타겟질병 예측에 초점 (예측모델 기반)

➤ 연구의 필요성

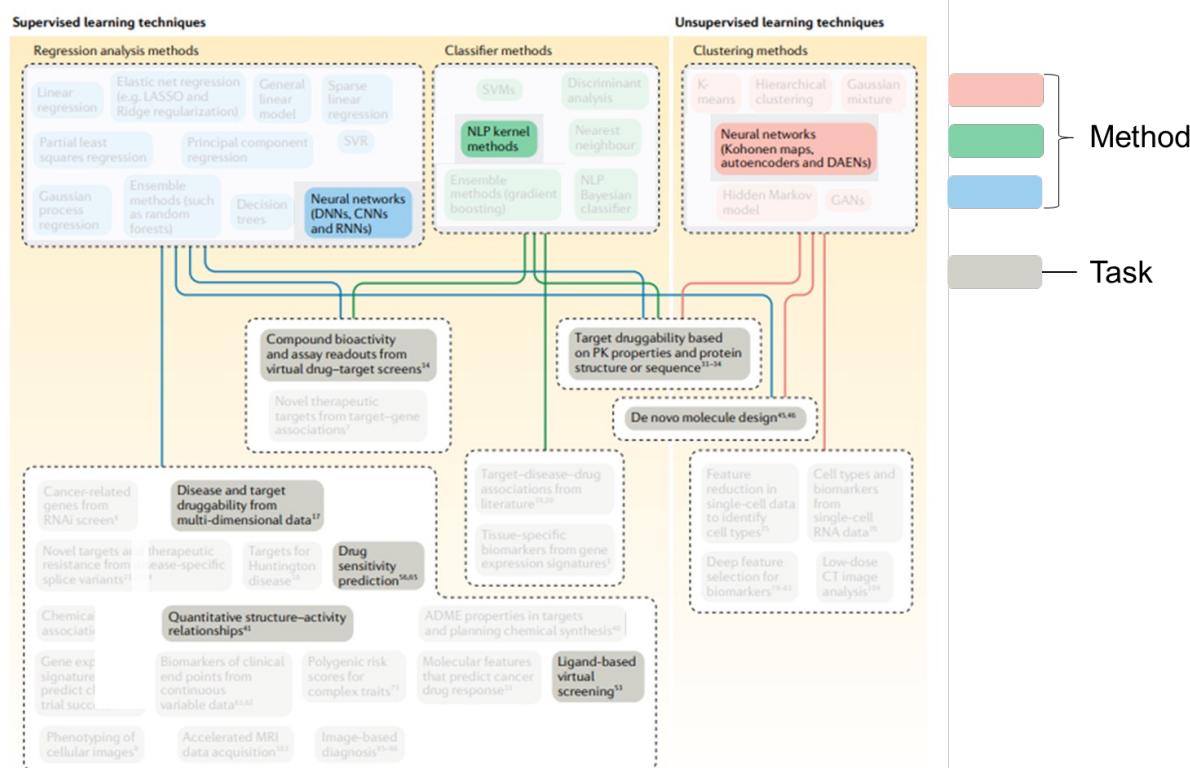
- 신약발굴과 약물 재창출을 동시에 고려하여 약물개발 연구 시너지 창출 필요
 - 지금까지 신약 발굴과 약물 재창출 분야의 딥러닝 기반 모델들은 별개로 연구 및 개발되어 왔음
 - 그러나, 신약 발굴과 약물 재창출 연구는 "약물과 타겟 질병 간 관계를 학습"한다는 점에서 교집합이 존재하며 이는 두 연구분야 간 시너지 창출의 가능성을 시사함
 - 신약 발굴 연구 – 타겟 질병을 정한 뒤 해당 타겟 질병을 공유하는 약물들의 공통적인 화합물 구조를 학습하여 약물 시퀀스 생성모델 개발
 - 약물 재창출 연구 – 약물들간 타겟 질병에 대한 반응 패턴의 유사성을 학습하여 DTI (Drug-Target Interaction) 예측모델 개발
 - 약물 시퀀스 생성모델은 타겟 질병의 단백질 시퀀스와 반응약물 분자시퀀스 간 시퀀스 분포의 유사성을 통해 질병 – 약물 간 구조적 관계 학습
 - DTI 예측모델은 타겟 질병의 단백질 시퀀스와 약물의 반응도 (결합친화도) 패턴을 통해 질병 – 약물 간 반응적 관계를 학습

I. 서론

➤ 연구 목적

- 약물재창출을 고려한 신규 후보약물 생성 모델 개발
 - 각 약물이 공유하는 타겟 질병의 범위를 (1) 실제 반응한 질병 뿐만 아니라 (2) 반응이 예상되는 질병들까지 확장하여 (즉, 신규 타겟 질병을 예측하여), 각 질병에 유효한 신규 후보약물을 생성하는 방법을 제안하고자 함

➤ 연구 키워드



II. 관련 연구

➤ 선행연구 요약

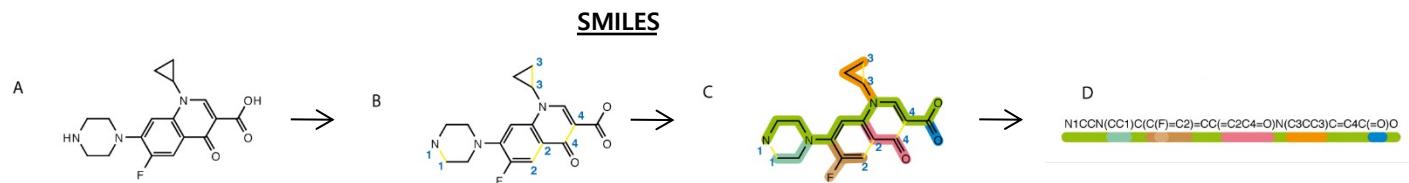
분야	논문명	데이터	제안 모델	세부 요소
신약발굴	Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models	ZINC	ORGAN	<ul style="list-style-type: none">GANREINFORCE
	Junction tree variational autoencoder for molecular graph generation	ZINC	JT-VAE	<ul style="list-style-type: none">Tree DecompositionMessage PassingVAE
	Graph convolutional policy network for goal-directed molecular graph generation	ZINC	GCPN	<ul style="list-style-type: none">GCNPPO
	Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules	ZINC, QM9	-	<ul style="list-style-type: none">VAEGaussian Process
약물 재창출	Deep-Learning-Based Drug–Target Interaction Prediction	DrugBank	DeepDTI	<ul style="list-style-type: none">DBN (Deep Belief Network)
	DeepDTA: deep drug–target binding affinity prediction	Davis, KIBA	DeepDTA	<ul style="list-style-type: none">CNN
	deepDR: a network-based deep learning approach to in silico drug repositioning	DrugBank, repoDB, ClinicalTrials.gov	DeepDR	<ul style="list-style-type: none">cVAERandom walk representationMonte Carlo gradient estimator
	Interpretable Drug Target Prediction Using Deep Neural Representation	BindingDB	-	<ul style="list-style-type: none">gCNNRNNAttention mechanismSiamese Network

III. 연구 방법

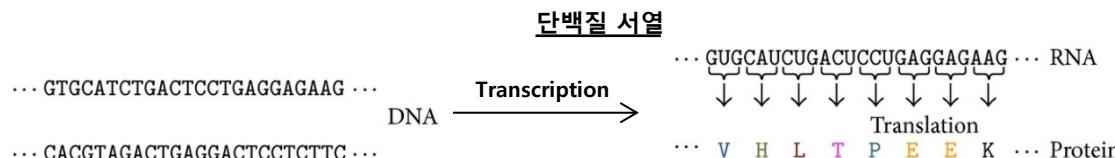
➤ 데이터 수집 및 전처리

• 데이터 수집

- 화합물 구조 표현 방식: **SMILES** (Simplified Molecular-Input Line-Entry System) 문자열 (Weininger, 1988)
 - 문자 내에 포함된 원자를 선형으로 표시하여 원자의 배열과 결합을 나타냄으로써 3차원 화합물 구조를 문자열로 표현하는 방식
 - ✓ 화합물 구조를 이루는 각 원소가 하나의 문자로 표현되고, 이러한 문자들이 순차적으로 연결되어 있는 형태
 - 해석이 쉽고 간단하며, 데이터 형태의 변환이 용이하므로 기계학습 모델의 입력으로 적합함. 이에 따라 화합물 구조와 화합물의 특성의 관계를 정량적으로 파악하기 위한 QSAR (Quantitative Structure-Activity Relationship) 모델에 주로 활용되고 있음
 - ✓ 특히, **LSTM 또는 GRU**와 같이 시계열 데이터의 학습에 강점이 있는 기계학습 모델에 적용하기 용이함
 - 예시



- 표적 단백질 표현 방식: **단백질 서열**(Protein sequence) (Sanger, 1952)
 - 표준 아미노산의 선형 서열을 나열하여 표적 단백질을 문자열로 표현하는 방식
 - ✓ 단백질을 이루는 표준 아미노산이 하나의 문자로 표현되고, 이러한 문자들이 순차적으로 연결되어 있는 형태
 - SMILES와 마찬가지로, 시계열 형태의 데이터이므로 **LSTM 또는 GRU** 모델에 적용하기 적합함
 - 예시



III. 연구 방법

➤ 데이터 수집 및 전처리

- 데이터 수집
 - 화합물 구조-표적 단백질 쌍에 대한 결합 친화도(Binding affinity)
 - 화합물 구조가 특정 표적 단백질과 반응하는 정도로서, 특정 질병에 대해 원하는 효능을 얻기까지 필요한 약물의 농도를 나타내는 약물 효력(Potency) 또는 약물-표적 단백질 간 결합의 세기를 나타내는 해리 상수를 통해 측정됨
 - 약물 효력
 - ✓ IC50: 화합물이 억제제인 경우, 표적 단백질과의 생물학적 반응을 50% 낮추는 데 필요한 화합물의 농도
 - ✓ EC50: 화합물이 약물인 경우, 표적 단백질과의 최대 생물학적 반응의 50%에 도달하는 데 필요한 화합물의 농도
 - 평형 해리 상수
 - ✓ Kd: 단일 분자(표적 단백질)와 결합 파트너(약물) 간 결합 상호 작용의 세기를 나타내며, 값이 작을수록 높은 결합 친화도를 나타냄
 - 데이터 수집 방법
 - 공개 화합물 데이터베이스인 PubChem을 활용하여 표적 단백질과 그에 반응하는 것으로 알려진 화합물 구조 데이터를 수집함
 - ✓ PubChem: 여러 데이터 소스로부터 취합된 1억 개 가량의 화합물 데이터를 제공하는 공개 데이터베이스로서, 각 화합물에 대한 SMILES 및 여러 표적에 대한 생물 검정(Bio assay) 결과를 제공하므로, 약물-표적 쌍(Drug-target pair)를 확보하기 용이함 (Kim et al., 2019)

III. 연구 방법

➤ 데이터 수집 및 전처리

• 데이터 전처리

- 입력 데이터 형태 변환

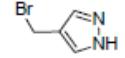
- 각 화합물 구조 또는 표적 단백질 문자열의 길이는 가변적이므로, 기계학습 모델에 적용하기 위해 **최대 문자열 길이를 제한함**
- 각각의 문자열 형태의 데이터를 **학습 가능한 행렬 형태로 변환함**
 - ✓ 등장 가능한 문자의 집합(Vocabulary)을 열(column)로, 문자열 내 순서대로 나타나는 각각의 문자를 행(row)으로 간주하는 행렬을 만들고, 각 행에 대해, 그에 해당하는 문자의 위치에만 1을 할당하고 나머지 위치에는 0을 할당하는 **one-hot encoding**을 수행하여 학습 가능한 행렬 형태의 데이터로 변환함

$$\mathbf{x}^{(i)} = \mathbb{1}_C(x) := \begin{cases} 1, & \text{if } x = C \\ 0, & \text{if } x \neq C \end{cases}$$

$\mathbf{x}^{(i)}$: i 번째 행 벡터, x : 각각의 문자에 대응되는 위치 C : 해당 문자

✓ 예시

One-hot encoding

Graph																																																																																																														
SMILES		BrCc1c[nH]nc1																																																																																																												
One-hot encoding	$\left\{ \begin{array}{c} \text{Br} \\ \text{C} \\ \text{H} \\ \text{c} \\ \text{n} \\ \text{1} \\ [\\] \end{array} \right\}$	<table border="1" style="margin-left: auto; margin-right: auto;"><thead><tr><th></th><th>Br</th><th>C</th><th>c</th><th>1</th><th>[</th><th>n</th><th>H</th><th>]</th><th>n</th><th>c</th><th>1</th></tr></thead><tbody><tr><th>Br</th><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><th>C</th><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><th>H</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><th>c</th><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr><tr><th>n</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr><tr><th>[</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><th>]</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr></tbody></table>		Br	C	c	1	[n	H]	n	c	1	Br	1	0	0	0	0	0	0	0	0	0	0	C	0	1	0	0	0	0	0	0	0	0	0	H	0	0	0	0	0	0	1	0	0	0	0	c	0	0	1	0	1	0	0	0	0	1	0	n	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	[0	0	0	0	0	1	0	0	0	0	0]	0	0	0	0	0	0	0	1	0	0	0
	Br	C	c	1	[n	H]	n	c	1																																																																																																			
Br	1	0	0	0	0	0	0	0	0	0	0																																																																																																			
C	0	1	0	0	0	0	0	0	0	0	0																																																																																																			
H	0	0	0	0	0	0	1	0	0	0	0																																																																																																			
c	0	0	1	0	1	0	0	0	0	1	0																																																																																																			
n	0	0	0	0	0	1	0	0	1	0	0																																																																																																			
1	0	0	0	1	0	0	0	0	0	0	1																																																																																																			
[0	0	0	0	0	1	0	0	0	0	0																																																																																																			
]	0	0	0	0	0	0	0	1	0	0	0																																																																																																			

III. 연구 방법

➤ 신약후보물질 발굴 모델 개발

- 모델 개요
 - 시퀀스-투-시퀀스(sequence-to-sequence – seq2seq) 모델과 다층 퍼셉트론(multilayer perceptron – MLP) 모델을 결합하여 입력된 표적 단백질과 반응할 가능성이 있는 화합물 구조를 생성하고, 결합 친화도를 예측하여 신약후보물질을 발굴함
- 화합물 구조 생성 모듈
 - 시퀀스-투-시퀀스 모델
 - LSTM, GRU 등의 순환 신경망을 기반으로 하여, 특정 도메인의 시퀀스 데이터를 입력으로 받아 그에 대응하는 또 다른 도메인의 시퀀스 데이터로 출력하는 모델로써, 주로 챗봇 또는 기계 번역에 활용됨
 - 순차적으로 입력되는 데이터를 은닉층(hidden layer)을 통해 압축하여 잠재 벡터(latent vector)로 변환하는 인코더와 잠재 벡터를 전달받아 새로운 데이터를 순차적으로 생성하는 디코더로 구성됨
 - 예시: 질의 응답 모델, 영어-독일어 자동 번역 모델
- 결합 친화도 예측 모듈
 - 다층 퍼셉트론 모델
 - 기본적인 인공신경망 구조로서, 1차원 벡터 형태의 데이터를 입력으로 받아 여러 개의 은닉층을 통해 해당 데이터의 특징을 추출함으로써 분류 또는 예측을 수행하는 모델
 - 예시: MNIST 손글씨 숫자 예측 모델

III. 연구 방법

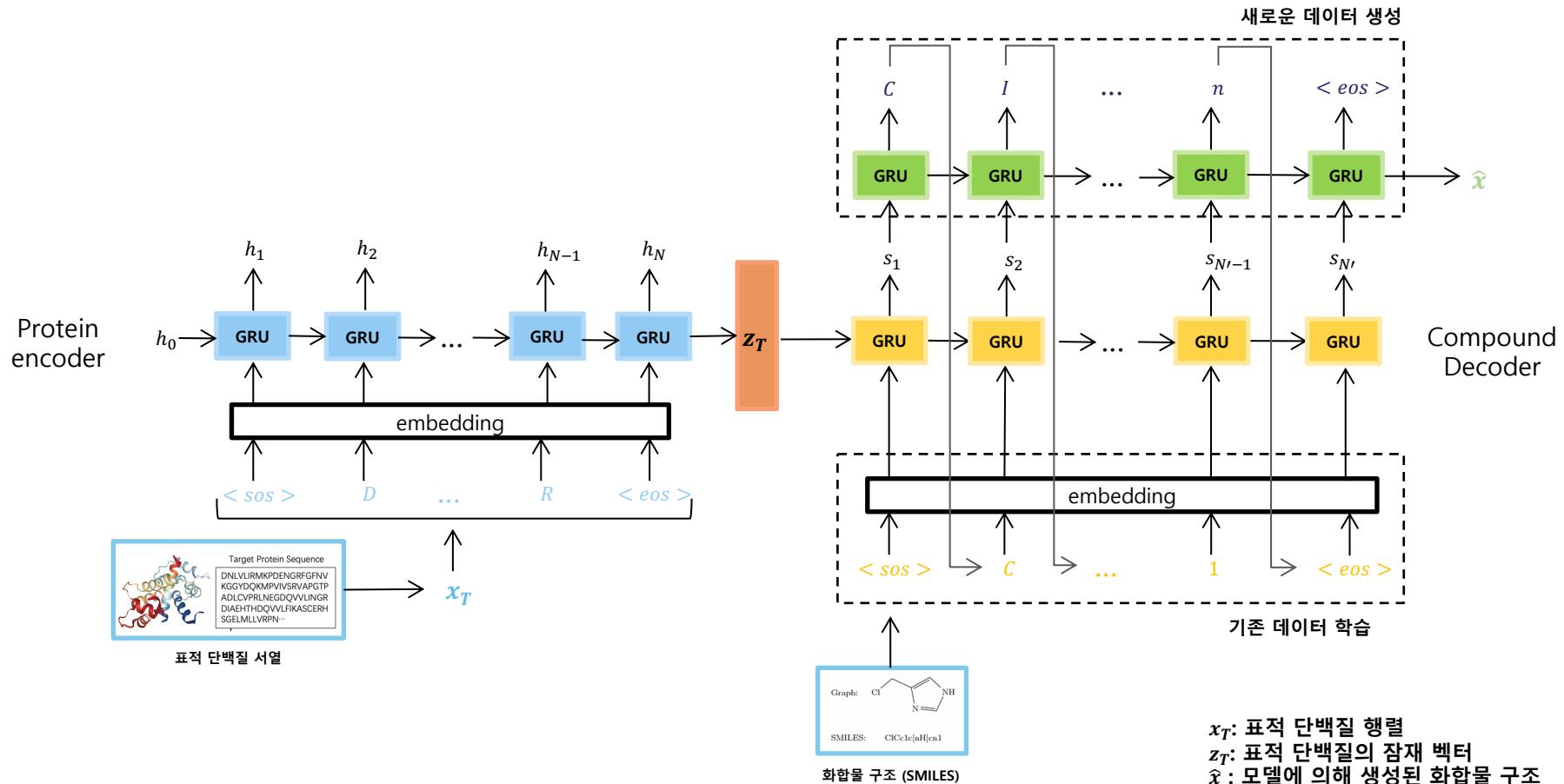
➤ 신약후보물질 발굴 모델 개발

- 신약후보물질 발굴 모델 학습 과정
 - 1) 표적 단백질의 아미노산 서열 내 문자를 생성 모듈의 인코더에 순차적으로 입력하여 문자열 **임베딩층(embedding layer)**을 거쳐 **임베딩 벡터로 변환함**
 - 2) 임베딩 벡터는 **인코더의 은닉층을 통해 잠재 벡터로 변환됨**
 - 3) 생성 모듈의 **디코더는** 잠재 벡터를 전달받아 해당 표적 단백질과 반응할 가능성이 있는 **화합물 구조를 순차적으로 생성함**
 - 4) 표적 단백질과 반응하는 것으로 알려진 실제 약물의 화합물 구조와 비교하여 생성 모듈의 **재현 오차(reconstruction error)**를 계산함
 - 5) 결합 친화도 예측을 위해 또 다른 인코더를 통해 실제 약물의 화합물 구조를 잠재 벡터로 변환하고, 표적 단백질의 잠재 벡터와 연결(concatenate)시켜 하나의 **1차원 벡터로 통합함**
 - 6) 통합된 잠재 벡터를 입력하여 **결합 친화도를 예측**하고, 실제 값과 비교하여 예측 모듈의 **예측 오차(prediction error)**를 계산함
 - 7) 재현 오차와 예측 오차를 합하여 최종 오차를 측정하고, **오차 역전파(error backpropagation)** 기법을 통해 각 모듈의 가중치를 갱신함으로써 신약후보물질 발굴 모델의 학습을 진행함

III. 연구 방법

➤ 신약후보물질 발굴 모델 개발

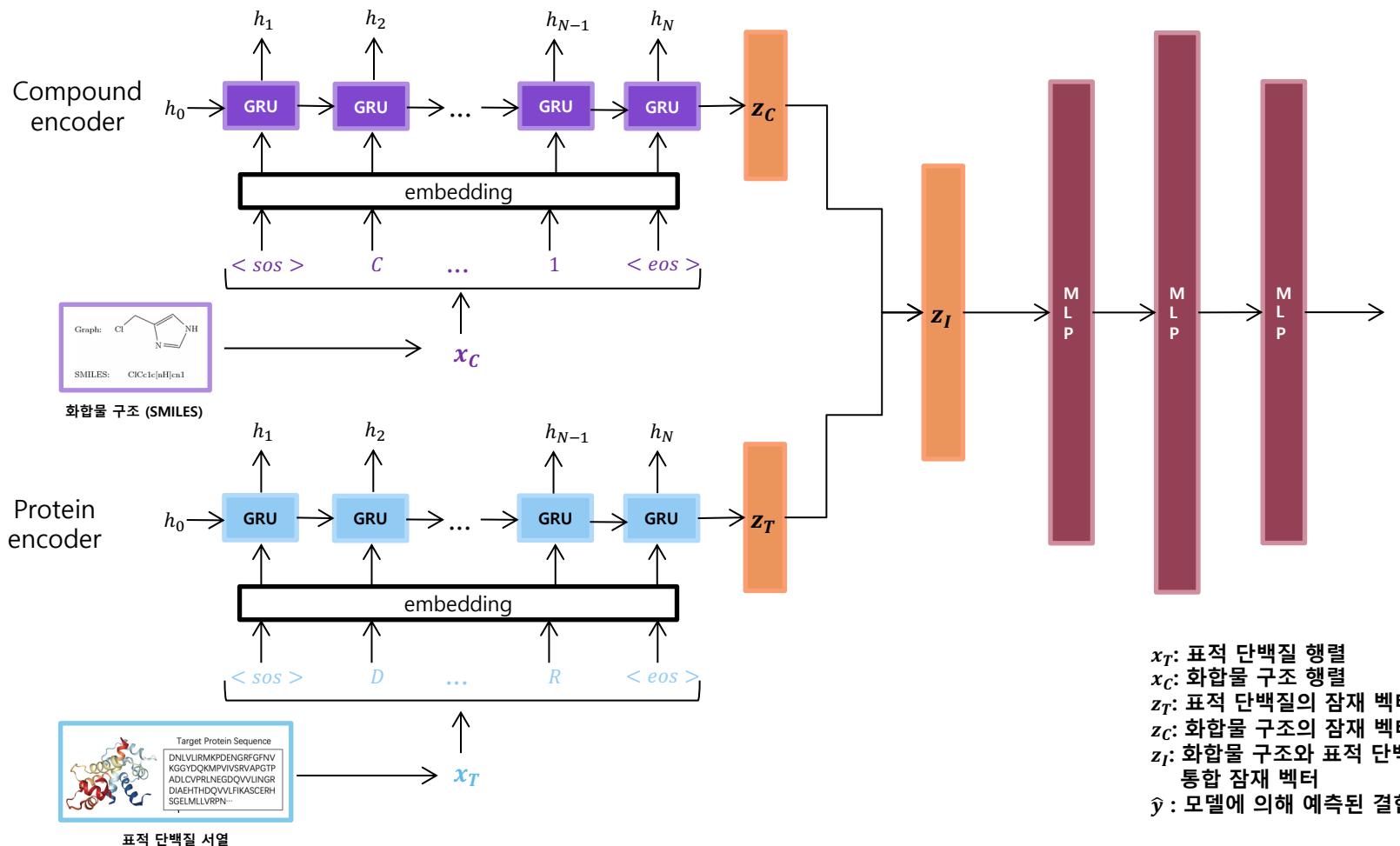
- 화합물 구조 생성 모듈 구조



III. 연구 방법

➤ 신약후보물질 발굴 모델 개발

- 결합 친화도 예측 모듈 구조

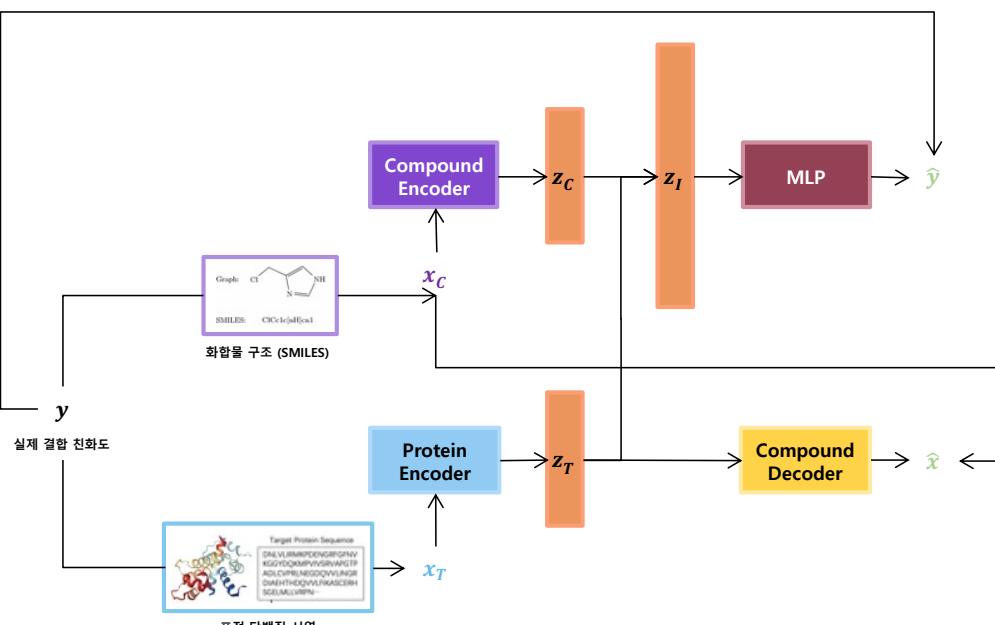


III. 연구 방법

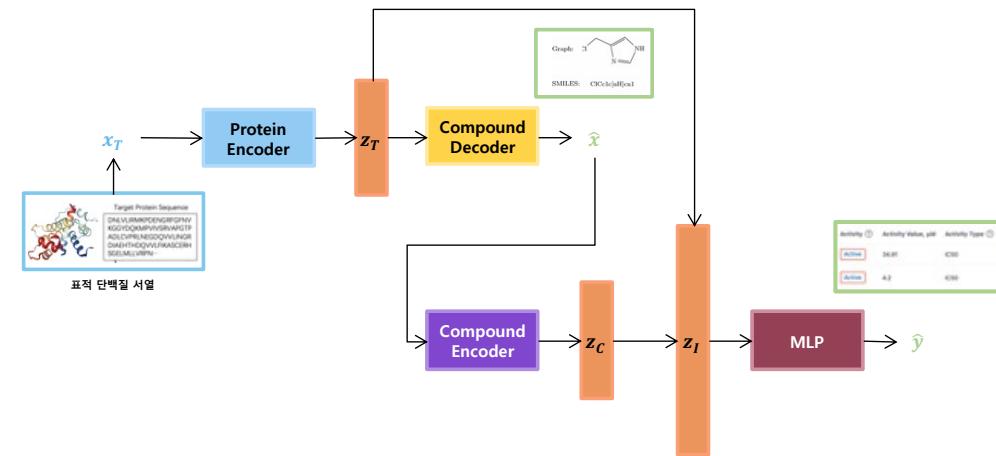
➤ 신약후보물질 발굴 모델 개발

- 신약후보물질 발굴 과정 예시

신약후보물질 모델 학습 과정



신약후보물질 발굴 과정



x_T : 표적 단백질 행렬
 x_C : 화합물 구조 행렬
 z_T : 표적 단백질의 잠재 벡터
 z_C : 화합물 구조의 잠재 벡터
 z_I : 화합물 구조와 표적 단백질을 연결한 통합 잠재 벡터
 \hat{x} : 모델에 의해 생성된 화합물 구조
 \hat{y} : 모델에 의해 예측된 결합 친화도

III. 연구 방법

➤ 모델 평가

- 화합물 구조 생성 모델

- **타니모토 계수**(Tanimoto coefficient): 두 집합 사이 유사도를 측정하는 방법으로써, 문자열 간 유사도 계측이 가능하므로 SMILES 문자열에 적용하여 화합물 구조의 유사도 측정에 주로 활용되고 있음. 두 문자열이 같은 문자를 더 많이 공유할수록 유사한 것으로 평가함

$$\text{Tanimoto coefficient} = \frac{|A \cap B|}{|A \cup B|}$$

- **BLEU 점수**(Bilingual evaluation understudy – BLEU): 기계 번역 모델의 대표적인 성능 평가 지표로써, 실제 번역 결과와 모델이 예측한 결과의 유사도를 기반으로 하여 생성 모델의 성능을 평가함

$$\text{BLEU score} = \min \left(1, \frac{\text{output length(예측 문장)}}{\text{reference length(실제 문장)}} \right) \left(\prod_{i=1}^N \text{precision}_i \right)^{1/N}$$

- 결합 친화도 예측 모델

- **평균 제곱근 오차**(Root mean squared error – RMSE): 회귀 예측에 대한 대표적인 성능 평가 지표로써, 실제값과 예측값 간의 평균 제곱근 오차를 통해 예측 모델의 성능을 평가함

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}$$

IV. 연구 결과

➤ 데이터셋 구축

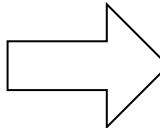
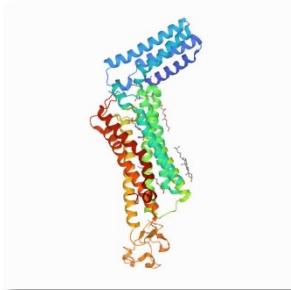
- 데이터 수집
 - PubChem 데이터베이스를 활용하여 주요 표적 단백질 352개에 대한 단백질 서열과 그에 반응하는 약물 679개의 화합물 구조를 수집함. 총 표적 단백질 – 약물 쌍의 수는 1,000개
 - PubChemPy: PubChem에서 제공하는 데이터 수집 API를 파이썬(Python)에서 활용할 수 있도록 하는 파이썬 라이브러리
- 등장 가능한 문자 집합의 크기
 - 표적 단백질: 42 (대/소문자 구분된 표준 아미노산 20개 및 패딩을 위한 공백 문자)
 - 화합물 구조: 59 (화합물 구조의 원소를 나타내기 위한 56개 문자 및 시작, 끝, 패딩을 표시하는 3가지 문자)
- 문자열 최대 길이
 - 표적 단백질: 518
 - 화합물 구조: 56
 - 전체 데이터셋에 포함된 표적 단백질/화합물 구조의 평균 길이로 설정함
 - 최대 길이보다 짧은 샘플의 경우 나머지 부분은 공백 문자로 패딩을 수행, 최대 길이보다 긴 샘플의 경우 최대 길이만큼만 잘라서 활용함

IV. 연구 결과

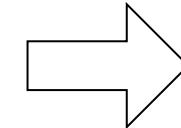
➤ 데이터셋 구축

- 데이터 샘플 예시
 - 표적 단백질
 - 단백질 이름: Thromboxane A2 receptor
 - 관련 질병: 출혈 장애(bleeding disorder)

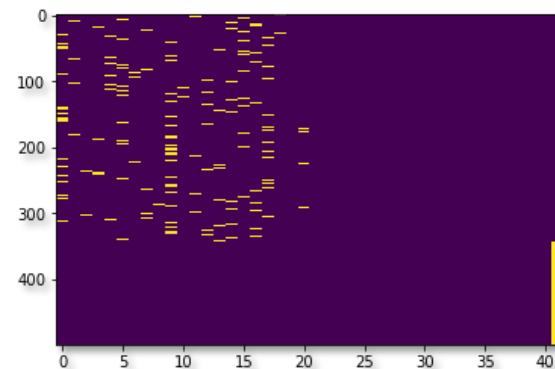
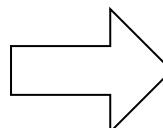
학습 가능한 행렬 형태로의 변환 예시



```
MWPNGSSLGPCFRPTNITLEERRLIASPWFAA  
SFCVVGLASNLLALSVLAGARQGGSHTRSSFL  
TFLCGLVLTDLGLLVGTIVSQHAALFEWH  
AVDPGCRLCRFMGVVMIFFGLSPLLLGAAMA  
SERYLGITRPFSPRAVASQRRAWATVGLWVA  
AALALGLLPLLGVRGRYTVQYPGSWCFLTGAE  
SGDVAFGLLFSMLGGLSVGLSFLLNTVSATL  
CHVYHGQEAAQQRPRDSEVEMMAQLLGIM  
VVASVCWLPLLVFIAQTVLRNPPAMSPAGQL  
SRTTEKELLIYLRVATWNQILDPWVYILFRRAV  
LRLQPRLSTRPRLSLSLPQLTQRSGLQ
```



```
[10 18 12 11 5 15 15 9 5 12 1 4 14 12 16 11 7 16 9 3 3 14 14 9  
7 0 15 12 18 4 0 0 15 4 1 17 17 5 9 0 15 11 9 9 0 9 15 17  
9 0 5 0 14 13 5 5 15 6 16 14 15 15 4 9 16 4 9 1 5 9 17 9  
16 2 4 9 5 9 17 16 5 16 7 17 17 15 13 6 0 0 9 4 3 18 6  
0 17 2 12 5 1 14 9 1 14 4 10 5 17 17 10 7 4 4 5 9 15 12 9  
9 9 5 0 0 10 0 15 3 14 19 9 5 7 16 14 12 4 15 14 12 0 17 0  
15 13 14 14 0 18 0 16 17 5 9 17 18 0 0 0 9 0 9 5 9 9 12 9  
9 5 17 5 14 19 16 17 13 19 12 5 15 18 1 4 9 16 9 5 0 3 15 5  
2 17 0 4 5 9 9 4 15 10 9 5 5 9 15 17 5 9 15 4 9 9 11 16  
17 15 17 0 16 9 1 6 17 19 6 5 13 3 0 0 13 13 14 12 14 2 15 3  
17 3 10 10 0 13 9 9 5 7 10 17 17 0 15 17 1 18 9 12 9 9 17 4  
7 0 13 16 17 9 14 11 12 12 0 10 15 12 0 5 13 9 15 14 16 16 3 8  
3 9 9 7 19 9 14 17 0 16 18 11 13 7 9 2 12 18 17 19 7 9 4 14  
14 0 17 9 14 14 9 13 12 14 9 15 16 14 12 14 15 9 15 9 13 12 13 9  
16 13 14 15 5 9 13 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20  
20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20  
20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20  
20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20  
20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20]
```



IV. 연구 결과

➤ 데이터셋 구축

- 데이터 샘플 예시

- 화합물 구조

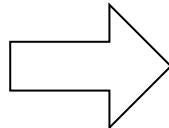
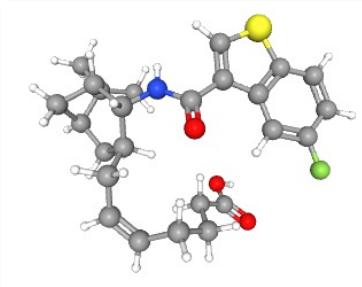
- 약물 이름: (Z)-7-[(1R,2R,3S,5S)-2-[(5-fluoro-1-benzothiophene-3-carbonyl) amino]-6,6-dimethyl-3-bicyclo[3.1.1]heptanyl]hept-5-enoic acid

- 분자식: C₂₅H₃₀FNO₃S

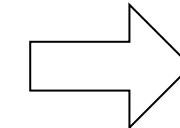
- 생물 검정 결과: 표적 단백질 Thromboxane A2 receptor에 대해 **0.43 (IC50)** 결합 친화도가 측정됨

- ✓ 생물 검정 이름: Inhibition of [3H]- (+)-S-145 specific binding to human platelet membranes in TXA2 receptor (TP) assay

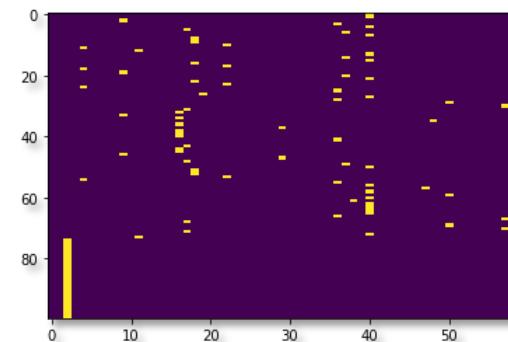
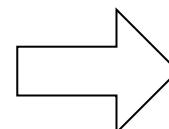
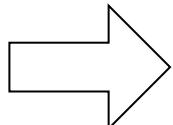
학습 가능한 행렬 형태로의 변환 예시



```
CC1(C)[C@H]2C[C@H]1[C@H](NC(=O)c1csc3ccc(F)cc13)[C@H](C)(C\CC=C/CCCC(O)=O)C2
```



```
[ 0 40 40 9 36 40 17 37 40 18 18 22 4 11 40 37 40 18 22 4 9 37 40 18  
22 4 36 19 40 36 50 57 17 16 9 16 48 16 29 16 16 16 36 58 17 16 16 16 9  
29 1 ]
```



IV. 연구 결과

➤ 신약후보물질 생성 결과 및 모델 성능 평가

- 실험 설계
 - 데이터셋 분할
 - 훈련 데이터셋: 전체 데이터셋의 **90%** → 900개 샘플
 - 테스트 데이터셋: 전체 데이터셋의 **10%** → 100개 샘플
 - 하이퍼 파라미터 설정
 - 모델 구조 관련
 - ✓ 임베딩층 크기: 256
 - ✓ 은닉층 크기: (Encoder) 128, (Decoder) 256-128, (Predictor) 128-256
 - 모델 학습 관련
 - ✓ 훈련 배치 크기(batch size): 25개 샘플
 - ✓ 학습률(learning rate): 0.005
 - ✓ 최적화 함수(optimizer): Adam (Adaptive Moment Estimation)
 - ✓ 손실 함수(loss function): Categorical cross-entropy

IV. 연구 결과

➤ 신약후보물질 생성 결과 및 모델 성능 평가 결과

- 모델 성능 평가
 - 학습 결과
 - 최종 손실: **0.6734**
 - 성능 평가 결과

모델 구분	성능 평가 지표	
화합물 구조 생성 모델	평균 Tanimoto coefficient	0.3100
	평균 BLEU 점수	0.3458
결합 친화도 예측 모델	평균 제곱근 오차	0.3360

- 신약후보물질 생성 결과 예시

표적 단백질	실제 반응 약물 (SMILES)	실제 결합 친화도 (IC50)	생성된 화합물 구조 (SMILES)	예측된 결합 친화도 (IC50)
Cytokinin dehydrogenase 2	COc1ccccc(Nc2nc(Cl)nc3nc[nH]c23)c1	3974.28	=n1c1ccccc1-c1c1cc1cc2c1cc(c1	4654.07
Secreted frizzled-related protein 1	CNCC(=O)N1CCC(CC1)NS(=O)(=O)C2=C(C=CC(=C2)S(=O)(=O)C3=CC=CC=C3)C(F)(F)F	201.88	2c1c1cc1ccc(O[Cn(=O)NC(=O)c1(=@H)c1c1ccc(F)C(=S)N[C@@H]1	513.15

V. 토의 및 결론

➤ 연구 요약

- 본 연구는 시퀀스-투-시퀀스 생성 모델과 다층 퍼셉트론 예측 모델을 활용하여 표적 단백질에 적합한 화합물 구조 생성 및 결합 친화도 예측을 수행하였고, 이를 통해 신약후보물질 발굴과 약물 재창출을 동시에 고려하여 약물 개발에 있어 시너지를 극대화할 수 있는 연구로서의 발판을 마련함

➤ 연구의 중요성 및 기대효과

- 최근 코로나-19의 확산으로 인해 약물 개발 과정의 가속화가 요구되고 있으며, 이는 완전히 새로운 신약발굴뿐만 아니라 약물 재창출 가능성 또한 고려해야함을 의미함. 이러한 상황에서 본 연구는 신약발굴과 약물 재창출을 동시에 고려하여 약물개발의 두 가지 연구 흐름의 시너지를 창출하는 것을 목표로 함으로써 시의적으로 적절한 것으로 판단됨
- 본 연구에서 제안하는 생성 및 예측 모델 통합 학습 프레임워크는 타겟 질병에 적합한 새로운 후보 약물을 생성함으로써 신약후보물질 가상 탐색(virtual screening)에 활용이 가능하며, 기존 약물의 화합물 구조와 다른 타겟 질병의 표적 단백질 간 결합 친화도를 예측함으로써 약물 재창출 가능성 판별에 활용이 가능할 것으로 기대됨

➤ 연구의 한계점

- 데이터 측면에서, 현재 약 70만 개의 화합물 구조-표적 단백질 쌍 데이터를 확보하였지만, 모델 학습 환경의 한계로 인해 실제로 모델 학습에 활용한 것은 1,000개 정도로 적은 수준임. 추후 더 많은 데이터 샘플을 활용할 수 있도록 모델 학습 환경을 개선할 예정임
- 모델 측면에서, 평가 결과 모델의 성능이 높지 않아서 실제로 활용하기에는 부족한 수준임. 추가적인 모델 구조 개선 및 조정(fine tuning)을 수행하여 성능을 높일 예정임
- 실용성 측면에서, 생성된 화합물 구조의 실제 존재 가능성 또는 합성 가능성 등 약물의 유효성(validity)에 대한 검증이 필요함

약물 재창출 가능성을 고려한
생성 및 예측 모델 기반
신약후보물질 발굴 방법론 개발

이규민(석박통합과정)¹, 이창현(석박통합과정)²
울산과학기술원 ¹경영과학과, ²산업공학과
{optimist, messy92}@unist.ac.kr

참고 문헌

1. Zhang, Daniel, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons et al. "The AI Index 2021 Annual Report." *arXiv preprint arXiv:2103.06312* (2021).
2. Guimaraes, Gabriel Lima, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. "Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models." *arXiv preprint arXiv:1705.10843* (2017).
3. Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." In *International Conference on Machine Learning*, pp. 2323-2332. PMLR, 2018.
4. You, Jiaxuan, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. "Graph convolutional policy network for goal-directed molecular graph generation." *arXiv preprint arXiv:1806.02473* (2018).
5. Gómez-Bombarelli, Rafael, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. "Automatic chemical design using a data-driven continuous representation of molecules." *ACS central science* 4, no. 2 (2018): 268-276.
6. Wen, Ming, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. "Deep-learning-based drug–target interaction prediction." *Journal of proteome research* 16, no. 4 (2017): 1401-1409.
7. Öztürk, Hakime, Arzucan Özgür, and Elif Ozkirimli. "DeepDTA: deep drug–target binding affinity prediction." *Bioinformatics* 34, no. 17 (2018): i821-i829.
8. Zeng, Xiangxiang, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, and Feixiong Cheng. "deepDR: a network-based deep learning approach to in silico drug repositioning." *Bioinformatics* 35, no. 24 (2019): 5191-5198.
9. Gao, Kyle Yingkai, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. "Interpretable Drug Target Prediction Using Deep Neural Representation." In *IJCAI*, vol. 2018, pp. 3371-3377. 2018.
10. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1), 31-36.
11. Sanger, F. (1952). The arrangement of amino acids in proteins. *Advances in protein chemistry*, 7, 1-67.
12. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1), D1388–D1395.