Pusan National University

Computer Science and Engineering

Technical Report 2023-10

신약 개발을 위한 binding affinity prediction

박한얼

임연후

김선아

지도교수 송길태

목 차

1. 서론

1.1. 연구 배경

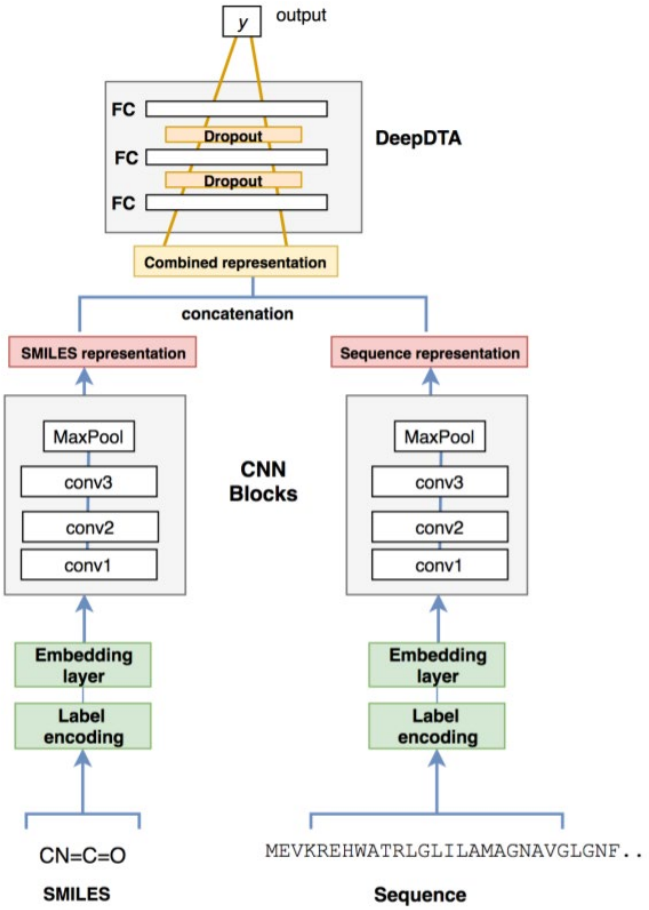


그림 1 SMILES와 단백질 서열로부터 학습하는 DeepDTA 모델[2]

예측을 이용한 새로운 약물-표적 상호작용(Drug Target Interaction, DTI) 식별은 신약 개발에 있어 중요한 과정이다. 이러한 예측을 위해선 단백질과 리간드 간의 상호 작용을 모델링하는 과정이 필요하다.

SMILES(Simplified Molecular Input Line Entry System) 표기법은 분자 구조를 문자열로 표현하는 방법 중 하나로, 분자의 화학 구조 정보를 담고 있다. 단백질 서열은 단백질의 구조와 기능을 나타내는 아미노산의 나열로 이루어진다. 이러한 SMILES와 단백질 서열 정보를 활용하여 단백질과 리간드 간의 상호 작용을 예측할 수 있다.

DeepDTA (Deep Drug-Target Interaction Prediction)는 이러한 약물-표적 상호작용 분야에서 주목받고 있는 모델 중 하나로, 본 과제는 단백질과 리간드 서열을 학습하여 약물-표적 쌍의 결합 친화도 값을 예측하는 DeepDTA 모델을 제작하는 것을 목적으로 한다.

1.2. 기존 문제점

과거 연구에서는 부족한 컴퓨팅 자원으로 인하여 한정된 양의 데이터로 모델을 학습하는 경우가 일반적이었다. 이로 인해 모델은 학습 데이터에 과도하게 적응하는 Overfitting이 발생하여 새로운 데이터에 대한 예측 능력이 떨어지는 경향이 있었다.

하지만, 컴퓨팅 성능의 발전으로 현재는 더욱 큰 데이터셋을 다룰 수 있게 되었고, 더욱 고도화된 모델을 이용하여 학습을 할 수 있게 되었다. 본 과제에서는 Overfitting 문제를 극복하기 위해 가능한 큰 데이터셋을 확보하고, 모델의 성능을 향상하기 위해 CNN 레이어에서의 parameter과 FC 레이어에서의 parameter, learning rate 등의 Hyperparameter 최적화를 수행하여 최적화된 모델을 구현한다.

1.3. 연구 목표

본 연구는 ligand 서열의 representation과 protein 서열의 representation, 총 두 가지의 output을 위한 CNN(Convolution Neural Network)을 지닌 DeepDTA 모델을 구현하고, CNN으로 학습시킨 두 개의 1D 서열 representation을 통해 약물-표적 결합 친화도를 예측하는 것을 목표로 한다. 더불어 Hyperparameter 최적화를 통한 모델의 성능을 향상시키는 것으로 학습 과정을 효율화해 보다 정확하고 빠른 약물-표적 상호작용 예측을 가능하게 한다.

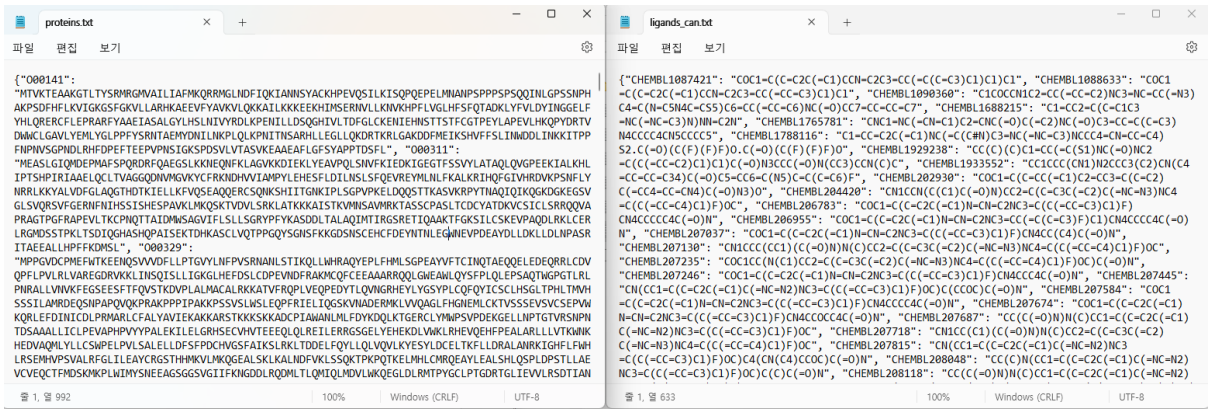
2. 연구 배경

2.1. 데이터 수집

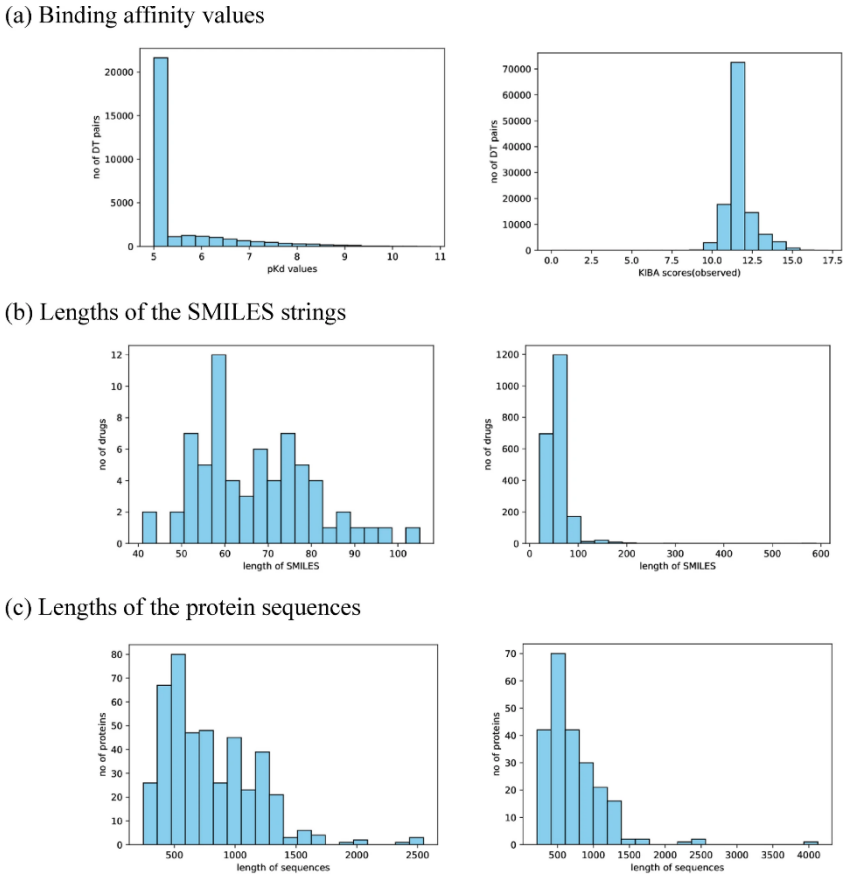
기존에는 BindingDB[3]의 binding affinity data from PubChem을 사용하려고 하였으나 여러 문제가 발생했다. 첫 번째로, binding affinity data를 포함한 파일들의 용량이 커 전 처리 프로세스 과정에 상당히 많은 시간이 소요되었다. 두 번째로, binding affinity 값을 나타내는 지표인 Ki, Kd, IC50 데이터가 균일하지 않아 학습이 어려웠다. Ki, Kd, IC50 값은 그 값 자체로는 통일되지 않아서 데이터셋을 사용할 수 없었다.

따라서 약물-표적 상호작용 데이터를 포함하고 binding affinity value 측정 기준이 통일된 Davis[2] 데이터셋과 다양한 binding affinity value 측정 기준을 묶어 bioactivity 값인 KIBA value 로 산출한 Kiba[3] 데이터셋으로 데이터셋을 변경했다. Davis 데이터셋은 442 개의

Protein 과 68 개의 화합물을 포함 총 30,056 개의 쌍과 그에 관한 binding affinity(K_d) 값을 지니고 있다. 반면 KIBA 데이터셋은 원래 52,498 개의 화합물과 467 개의 Protein 으로 이루어져 있으나, 상호작용이 10 이하인 약물-타겟을 제거한 결과 229 개의 Protein 과 2,111 개의 화합물 포함 총 118,254 개의 쌍과 그에 관한 binding affinity 값을 사용할 수 있었다.[10]



2 KIBA 데이터셋의 SMILES와 단백질 서열



3 Davis 데이터셋(왼쪽 패널) 과 KIBA 데이터셋(오른쪽 패널) 요약

2.2. 개발 환경

파이썬 환경을 효율적으로 관리하고 필요한 라이브러리와 패키지를 설치하기 위해 Anaconda와 주피터 노트북을 사용했다. Anaconda를 통해 파이썬 가상 환경을 생성하여 DeepDTA의 의존성을 관리하고 별도의 환경에서 프로젝트를 실행하였다.

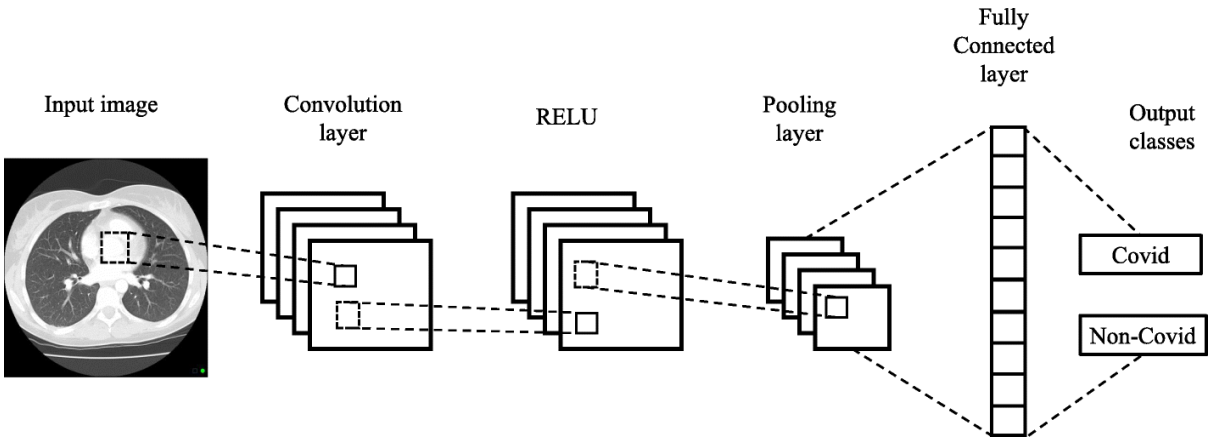
DeepDTA 모델의 개발과 학습에 있어 파이토치(PyTorch) 라이브러리를 사용했다. 파이토치를 사용하여 그래디언트 계산과 역전파를 자동으로 처리해 모델을 가중치를 업데이트하고 손실을 최소화하는 학습 과정을 구현할 수 있었다. 또한 모델을 학습 설계와 최적화를 위해 파이토치의 데이터로더(DataLoader), 손실 함수(MSELoss), 옵티마이저(Adam)등이 사용되었다.

3. 연구 내용

3.1. 모델 설계

DeepDTA는 약물과 단백질 각각에 대해 1D 컨볼루션 신경망을 적용해 단백질 및 약물 데이터의 특성을 추출하고 그 결과를 결합해 결합 친화도를 예측하는 모델이다. 본 과제에서는 DeepDTA 모델을 구현하기 위해 1D 컨볼루션 신경망을 정의하는 Conv1d와 FC layer인 DeepDTA 두 가지의 클래스를 사용하였다.

3.1.1. Conv1d



4 CNN 구조[9]

컨볼루션 신경망은 [그림 3]과 같은 구조로 이루어져 있다. CNN(Convolutional neural

network)는 주로 이미지와 같은 2차원 형태의 입력 데이터에 사용된다. 본 과제에서 사용하는 데이터셋은 텍스트로 주어진 SMILES 표현법과 단백질 서열 정보로, 이를 학습하기 위해 2D 컨볼루션 신경망을 1차원으로 변경한 1D 컨볼루션 신경망(Conv1d)를 사용했다.

PyTorch를 사용하여 1D 컨볼루션 신경망을 정의하였다. 입력 데이터는 3개의 1D 컨볼루션 신경망을 통과한다. 이때 비선형성을 추가하기 위한 활성화 함수로 ReLU(Rectified Linear Unit)가 사용되었다. 최종적으로 글로벌 맥스 풀링을 통해 최종 특징 벡터를 반환한다.

3.1.2. DeepDTA

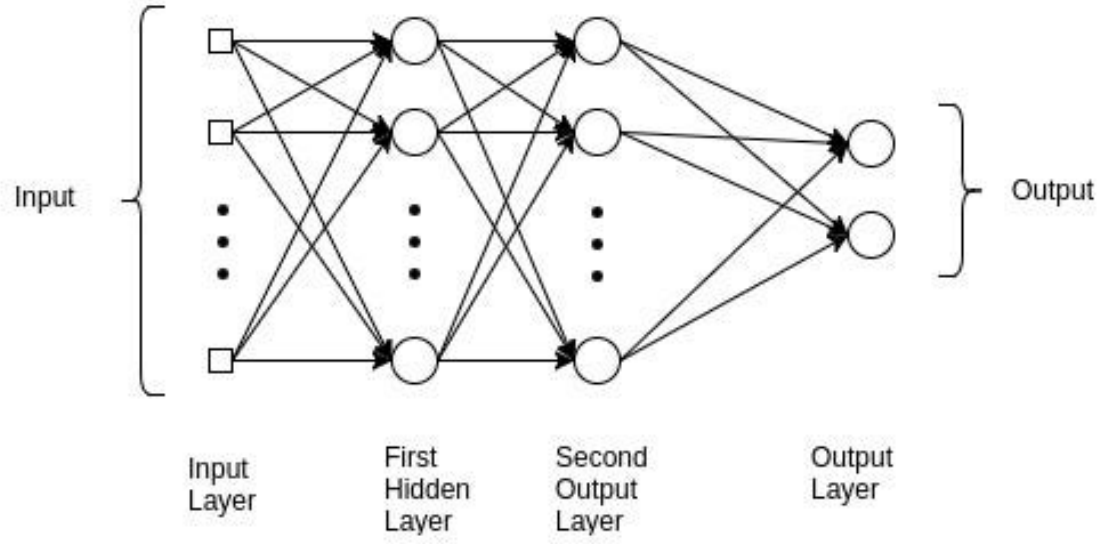


그림 5 다층 퍼셉트론의 예시

DeepDTA 클래스는 DeepDTA 모델의 주요 아키텍쳐 클래스로, 입력으로 1D 컨볼루션 레이어를 통과한 단백질 및 리간드의 특징 벡터가 주어진다. DeepDTA 클래스에선 여러 개의 완전 연결 레이어(Fully Connected layer)로 구성된 다층 퍼셉트론(MLP)이 사용된다.

다층 퍼셉트론은 [그림 5]와 같이 단일 퍼셉트론의 중첩 구조로 이루어져 있다. 다층 퍼셉트론의 각 레이어 사이에 활성화 함수 ReLU(Recitified Linear Unit)와 Dropout 레이어를 적용해 모델을 학습시키고 결합 친화도를 예측했다.

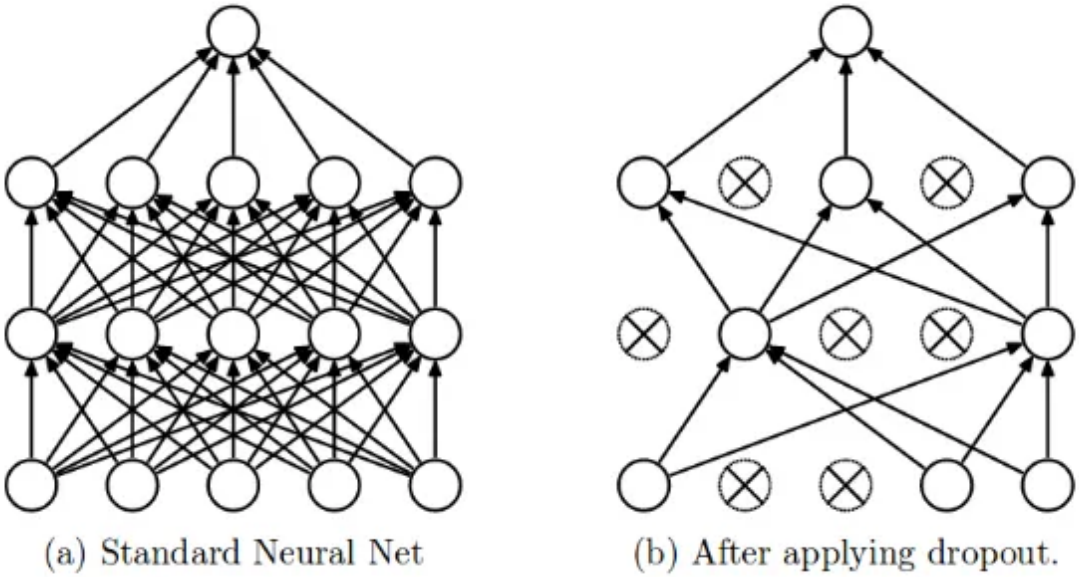


그림 6 (a) Drop-out이 적용되지 않은 신경망 (b) Drop-out이 적용된 신경망

Overfitting 현상을 제거하기 위하여 Drop-out을 사용하였다. Drop-out은 [그림 6(a)]와 같이 통해 노드를 랜덤하게 학습에서 제외하는 과정으로, 이를 통해 모델은 특정 가중치에 치우치지 않고 학습할 수 있다. 본 과제에서는 0.1의 비율로 Drop-out을 사용하여 10%의 확률로 노드를 비활성화 시켜 Overfitting을 제거하였다.

3.1.3. 하이퍼파라미터 튜닝: Grid Search

```
fp = './Davis/'

model = DeepDTA
channel = 32
protein_kernel = [8, 12]
ligand_kernel = [4, 8]

for prk in protein_kernel:
    for ldk in ligand_kernel:
        trainer = Trainer(fp, model, channel, prk, ldk, "training_logs-prk{}-ldk{}.log".format(prk, ldk))
        trainer.train(num_epochs=10, batch_size=250, lr=0.01, save_path='training_result-prk{}-
ldk{}.pt'.format(prk, ldk))
```

코드 1 그리드 검색 코드

최적의 하이퍼파라미터 조합을 찾기 위해 그리드 검색(Grid Search)을 사용했다. protein_kernel 및 ligand_kernel은 Conv1d 레이어의 커널 크기를 결정하며, 적절한 커널의 크기 설정으로 모델의 예측 정확도를 향상할 수 있다.

3.1.4. 학습 수행

```
k_folds = 6
kf = KFold(n_splits=k_folds, shuffle=True, random_state=42)

validation_losses = []

optimizer = torch.optim.Adam(self.model.parameters(), lr=lr, weight_decay=1e-5)
criterion = nn.MSELoss()

for fold, (train_indices, val_indices) in enumerate(kf.split(X_train)):
    print(f"============== Fold {fold + 1}/{k_folds} ==============")

    train_data = Subset(X_train, train_indices)
  val_data = Subset(X_train, val_indices)
```

코드 2 6-fold 교차검증을 통한 모델 학습

K-fold 교차 검증을 통해 모델 학습이 수행되었다. 각 폴드에선 훈련 데이터를 미니 배치로 나누어 모델을 훈련하고, 각 에포크 후 검증 데이터를 사용해 모델의 성능을 평가하여 검증 손실이 더 이상 개선되지 않을 때 훈련을 중단한다. 또한 모델의 가중치 업데이트를 관리하기 위해 Adam 최적화 알고리즘을 사용했다.

3.1.5. 테스트 및 평가

```
test_loader = DataLoader(X_test, batch_size= batch_size, drop_last = False, collate_fn=collate_fn)
        test_losses = []
        with torch.no_grad():
            for protein, ligand, target in test_loader:
                protein, ligand, target = protein.to(self.device), ligand.to(self.device),
target.to(self.device)

                output = self.model(protein, ligand)

                test_loss = criterion(output, target)
```

```
            test_losses.append(test_loss.item())


        avg_test_loss = np.mean(test_losses)
        print(f"Average Test Loss: {avg_test_loss}")
```

코드 3 테스트 및 평가 코드

K-fold 교차 검증이 완료된 후 테스트 데이터에 대한 성능을 평가한다. 이때 최종 테스트 데이터셋에서 모델의 성능이 평가되며 평가 지표로 평균 제곱 오차(MSE)를 사용하였다.

4. 연구 결과 분석 및 평가

화합물과 표적 단백질 간의 binding affinity 정보가 포함된 데이터셋의 부족으로 본 과제에 적합한 데이터셋의 확보가 어려웠다. 대부분의 데이터셋은 제한적인 종류의 표적 단백질에 대한 binding affinity만을 포함하고 있었으며, 정형화되지 않은 Ki/Kd/IC50 등의 데이터 형식을 제공하여 학습에 어려움을 겪었다.

따라서, 여전히 많은 연구자가 데이터 분석에 활용하는 KIBA와 Davis 데이터셋을 사용했으며, k-fold cross validation을 통해 데이터셋을 여러 fold로 나누어 여러 종류의 test set에 대한 평가를 진행했다. 데이터셋 자체는 한정적이었으나, binding affinity 예측 모델은 시중에 풍부해 Kiba dataset과 Davis dataset에 대한 성능 비교가 가능했다.

이 연구에서는 모델과 비교를 위해 KronRLS와 SimBoost를 채택하였다. 모델의 성능 평가 기준으로는 Concordance Index(CI) 와 Mean Squared Error(MSE)를 사용하였다.

	CI	MSE
KronRLS	0.890	0.380
SimBoost	0.877	0.282
DeepDTA	0.865	0.272

5. 결론 및 향후 연구 방향

본 과제에서는 표적 단백질과 화합물의 1D 서열에 대한 one-hot-encoding을 사용해 각각에 대한 representation을 생성하고, CNN을 통한 딥러닝으로 단백질과 화합물 간의 결합 친화도를 예측하는 모델을 만들었다. 이를 위해 KIBA dataset과 Davis dataset을 활용하였으며, 파라미터를 최적화하여 학습시켰다. 결과적으로, 기존의 결합 친화도 예측을 위한 머신러닝 모델들보다 나은 MSE 점수를 기록하였고, CI 점수 또한 좋은 결과를 보였다.

더욱 정확한 예측을 위해 개발 중 one-hot encoding가 아닌 transformer encoder[7]를 활용하자는 아이디어가 제시되었다. Transformer encoder를 사용하면 embedding에 대한 representation을 생성할 수 있으며, 이를 통해 토큰과 토큰 간의 sequential 데이터를 포함할 수 있게 되어 염기 서열로 이루어진 단백질의 구성에서 더 많은 정보를 학습할 수 있을 것으로 예상된다. 그러나 이에 따라 모델의 복잡도가 증가해 학습 시간이 더 길어질 수 있기에 앞으로의 연구가 필요하다.

6. 참고 문헌

[1] 대웅제약 - "인공지능의 시대, 물질 발굴부터 임상 설계까지" AI신약팀에게 듣는 'AI 신약개발은 지금' Available: https://newsroom.daewoong.co.kr/archives/19864

[2] Hakime Ozturk, Arzucan Ozgur, Elif Ozkirimli. (2018) DeepDTA: deep drug-target binding affinity prediction., *Bioinformatics*, Volume 34, Issue 17, September 2018, Pages i821-i829

[3] Binding DB. Available: https://www.bindingdb.org/rwd/bind/chemsearch/marvin/Download.jsp

[4] Davis M.I. et al.. (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol., 29, 1046-1051.

[5] Tang J. et al.. (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J. Chem. Inf. Model., 54, 735-743.

[6] Pubchem. Available: https://pubchem.ncbi.nlm.nih.gov/

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. (2017) Attention is All You Need. arXiv:1706.03762

[8] Pytorch documentation. Available: https://pytorch.org/docs/stable/index.html

[9] Sneha Kugunavar, C.J. Prabhakar *Convolutional neural networks for the diagnosis and prognosis of the coronavirus disease pandemic,* Visual Computing for Industry, Biomedicine, and Art volume 4, Article number: 12, 2021

[10] Jooyong Shim, Zhen-Yu Hong, Insuk Sohn, Changha Hwang *Prediction of drug–target binding affinity using similarity-based convolutional neural network, Scientific Reports* volume 11, Article number: 4416, 2021