인공지능을 이용한 신약개발 플랫폼

모델 설명서

KAIDD | 인공지능 신약개발플랫폼

Contents

Part 1 전체 모델 소개 6

Part 2 AD3 단백질구조기반

1) GalaxyTBM 8
2) Redesigns 9
3) GalaxyDock3 10
4) AK-Scores 11
5) SPICA 12

Part 3 CSK Studio 신경퇴행성

1) NetExp 14
2) Molecule Generation 16
3) BioActivity Prediction 17
4) Drug Neighbor 18

Part 4 MiLearn™ 항암신약

1) AiCAD 20
2) AiGPro 21
3) AiKPro 22
4) AiP450 23
5) CRX4 24

Part 5 AIDrug 빅데이터/AI 신약

1) ADMET 26
2) Toxicity 등 28
3) De novo Design 30
4) Virtual Screening 32
5) 기타 33

Part 6 Synbi 약물재창출

1) SynbiDrug-R 36

Part 7 Smart PV 스마트약물감시

1) Smart PV irAE 40
2) CDM Based ADR screening tool 42

Part 1

전체 모델 소개

1 전체 모델 소개

Platform	AI model	Description
AD3 Structural Protein 단백질구조기반	GalaxyTBM	단백질 아미노산 서열로부터 단백질 3차원 구조 예측
	Redesigns	새로운 후보물질 생성
	GalaxyDock3	단백질-리간드 결합가능성 예측
	AK-Scores	단백질-리간드 복합체구조의 결합 친화도 예측
	SPICA	약물의 ADME 특성 예측
CSK Studio Neurodegenerative 신경퇴행성	NetExp	유전자 질환 등 다양한 데이터를 사용하여 표적 예측
	Molecule Generation	화합물 예측
	BioActivity Prediction	기초 독성 예측
	Drug Neighbor	ChemMap 상에서 가까운 거리 내에 존재하는 물질들을 drug bank 데이터베이스에서 검색
MiLearn™ Anticancer Drug 항암신약	AiCAD	표현형 기반 항암타겟 치료제 스크리닝 모델
	AiGPro	다중서열 정렬 기반 타겟-항암제 가상 스크리닝 모델 (GPCR)
	AiKPro	다중서열 정렬 기반 타겟-항암제 가상 스크리닝 모델 (Kinase)
	AiP450	5종의 CYP450 억제능 예측
	CRX4	Kinase inhibitor likeness, GPCR ligand likeness 예측
AIDrug Big Data/AI Drug 빅데이터/AI 신약	ADMET	ADMET 예측 AI 모델 12건
	Toxicity 등	독성 등 예측 AI 모델 16건
	De novo Design	Scaffold, 물성, 단백질 기반 구조 생성 등 AI 모델 3건
	Virtual Screening	약물-단백질 상호작용 예측 등
	기타	빅데이터 검색 등
Synbi Drug Repurposing 약물재창출	SynbiDrug-R	약물 다중 특성 기반 승인 약물들의 항암 표적 및 효능 예측
Smart PV Smart Pharmacovigilance 스마트약물감시	Smart PV irAE	인공지능 기반 면역관련 부작용 예측 모델
	CDM Based ADR screening tool	환자 정보 및 바이오마커 기반 부작용 예측

Part 2

AD3 단백질구조기반

1) GalaxyTBM
2) Redesigns
3) GalaxyDock3
4) AK-Scores
5) SPICA

2 AD3 단백질구조기반

AD3는 표적 단백질 중심 신약 개발 플랫폼입니다. 인공지능과 표적 단백질 3차원 구조를 기반으로 후보 물질 결합 가능 위치 예측, 결합 특성 분석, 후보 물질 디자인, 결합 가능성 예측 및 신약 가능성 분석을 통해 빠른 신약 후보 물질 개발을 가능하게 합니다.

1) GalaxyTBM (Galaxy Template-Based Modeling)

가) 모델 설명

AD3는 표적 단백질 중심 신약 개발 플랫폼입니다. 인공지능과 표적 단백질 3차원 구조를 기반으로 후보 물질 결합 가능 위치 예측, 결합 특성 분석, 후보 물질 디자인, 결합 가능성 예측 및 신약 가능성 분석을 통해 빠른 신약 후보 물질 개발을 가능하게 합니다.

나) 학습 데이터

GalaxyTBM은 다음 정보를 이용하여 학습데이터를 생성하며 다양한 인공지능 모델을 학습합니다.

- 단백질 구조 정보 : RCSB PDB (https://www.rcsb.org/)
- 단백질 서열 DB1 : UniProt (https://www.uniprot.org/)
- 단백질 서열 DB2 : BFD (https://bfd.mmseqs.com/)

다) Input/Output

Input

Protein Sequence : FASTA Format File

Output

- ***.fa** : Query Sequence (Input File)
- ***.pdb** : Predicted Structures
- ***.domain.txt** : Domain Prediction Report
- ***.domain.png** : Domain Prediction Report

라) 결과 해석

단백질 구조는 구조 및 기능의 단위인 Domain 단위로 예측되며, 입력 서열에서 Domain을 예측한 뒤, 각 Domain 별 MSA 생성, Contact 예측, 3차원 구조 예측을 수행합니다.

결과 파일은 예측된 구조를 담고 있으며, 단백질 구조를 저장하는 대표적인 형식인 PDB형식의 파일로 받을 수 있습니다.

2) Redesigns

가) 모델 설명

표적 단백질에 결합하는 화합물을 생성하는 인공지능 모델입니다. 표적 단백질의 3차원 구조와 3결합한 화합물 3차원 결합 구조를 기반으로 상호작용을 수치화 시킨 뒤 결합 강도를 예측할 수 있는 인공지능 모델을 생성한 뒤 이를 이용하여 강화 학습을 통해 새로운 후보 물질을 생성합니다.

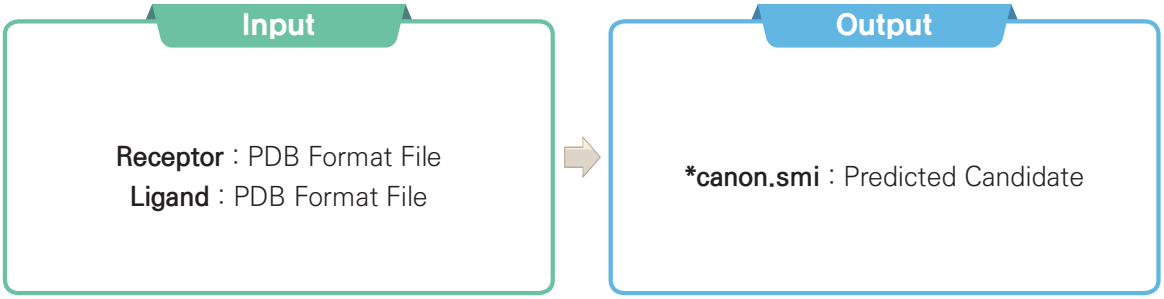
만약 결합 구조가 알려져 있는 화합물이 없을 경우 GalaxyDock3를 이용하여 새로운 결합 구조를 생성한 뒤 결과 파일을 이용하여 Redesigns를 이용할 수 있습니다.

나) 학습 데이터

Redesigns 모델은 약물 특성이 있는 보편적인 생성 모델로 pre-training 한 뒤, 주어진 표적 화합물에 맞도록 transfer-learning을 합니다. 이를 위해서 다음 데이터를 활용하고 있습니다.

- ChEMBL (https://www.ebi.ac.uk/chembl/)
- ZINC (https://zinc.docking.org/)
- RCSB PDB (https://www.rcsb.org/)

다) Input/Output



- Receptor : ATOM의 정보만 포함되어 있어야 합니다.
- Ligand : 결합되어 있는 화합물 ATOM, CONNECT 정보만 포함되어 있어야 합니다.

라) 결과 해석

입력된 화합물의 결합 정보를 이용하여 유사한 모양과 결합 특성을 가지는 후보 물질을 생성합니다. 결과 파일은 생성된 화합물의 SMILES 형식의 구조를 담고 있습니다.

인공지능 모델이 예측한 화합물은 약물 개발을 위한 특성 분석과 결합 강도의 정밀 예측이 되지 않은 후보 물질이기 때문에 이후 SPICA와 GalaxyDock3, AK-Scores등을 이용하여 특성을 분석하고 선별하는 과정을 수행해야 합니다.

3) GalaxyDock3

가) 모델 설명

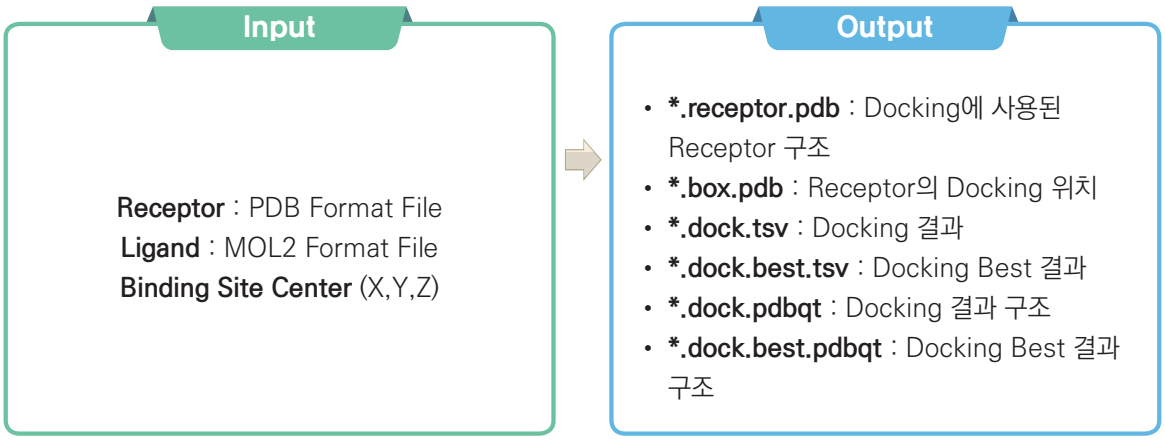
표적 단백질과 후보 화합물의 3차원 결합 구조와 결합 강도를 측정하는 모델입니다. 이 모델은 물리화학적인 계산 방법을 이용하여 결합 강도를 예측합니다.

나) 학습 데이터

실험적으로 결합 강도가 알려져 있는 단백질-화합물 결합 구조를 학습 데이터로 이용합니다.

- PDBbind (http://pdbbind.org.cn/)
- RCSB PDB (https://www.rcsb.org/)

다) Input/Output



- Receptor : ATOM의 정보만 포함되어 있어야 합니다.
- Ligand : ATOM, CONNECT 정보만 포함되어 있어야 합니다.

라) 결과 해석

GalaxyDock3의 결과는 다음과 같은 정보를 담고 있습니다.

- 입력으로 받은 화합물의 결합 3차원 구조
- 각 3차원 구조 별 결합 강도

각 화합물은 1개 이상의 결합 구조를 가지며, 각 구조별로 다른 결합 에너지를 가집니다. 일반적으로는 가장 좋은 결합 에너지를 가지는 구조(Best)를 정답으로 선택하지만, 다양한 가능성을 가질 수 있으며, 이를 위해서 AK-Scores 같은 결합 강도만을 정확하게 예측하기 위한 방법을 사용합니다.

4) AK-Scores

가) 모델 설명

단백질-화합물의 결합 구조에서 정확한 결합 에너지를 예측하는 인공지능 모델입니다. 단백질-화합물 결합 구조와 결합 강도를 예측하기 위해서 사용하는 GalaxyDock3의 경우 결합 강도의 정밀도가 상대적으로 높지 않을 수 있습니다.

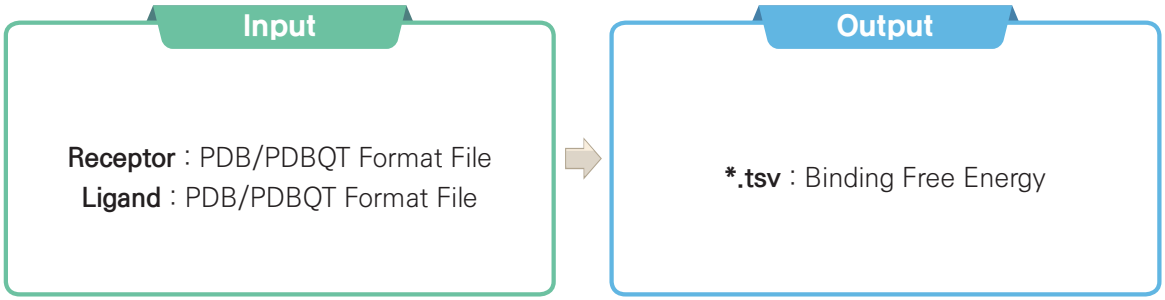
AK-Scores는 단백질-리간드 복합체 구조의 친화성을 예측하는 새로운 신경망 모델입니다. 이 모델은 여러 독립적으로 훈련된 네트워크들(3D 컨볼루션 신경망 계층의 여러 채널로 구성됨)의 조합을 사용하여 복합체의 친화도를 예측합니다. 이 모델은 3740개의 정제된 PDBbind 데이터세트 단백질-리간드 복합체로 훈련되었으며, 270개의 core 세트로부터 온 270개의 복합체를 이용하여 테스트하였습니다. 벤치 마크 결과는 AK-Scores 모델에 의해 예측된 결합 친화도와 실험 데이터 사이의 상관 계수가 0.72 보다 높음을 보여주며, 이는 최첨단 결합 친화도 예측 방법과 비교할 수 있습니다. 또한, 우리의 방법은 단백질의 가능한 다중 결합제의 상대적 결합 친화도를 매우 정확하게 순위를 매깁니다. 마지막으로 결합 친화도를 예측하는 데 중요한 구조 정보를 측정했습니다.

나) 학습 데이터

AK-Scores는 결합 강도와 결합 3차원 구조가 알려진 데이터셋을 이용합니다.

- Training Set : 3740개의 정제된 PDBbind 데이터세트 단백질-리간드 복합체
- Test Set : 270개의 core 세트로부터 온 270개의 복합체

다) Input/Output



- Receptor : ATOM의 정보만 포함되어 있어야 합니다.
- Ligand : ATOM, CONNECT 정보만 포함되어 있어야 합니다.

라) 결과 해석

AK-Scores로 예측된 결과 파일은 각 화합물의 결합 자유 에너지를 기술하고 있으며, 낮을 수록 더 강력한 결합을 합니다.

5) SPICA

가) 모델 설명

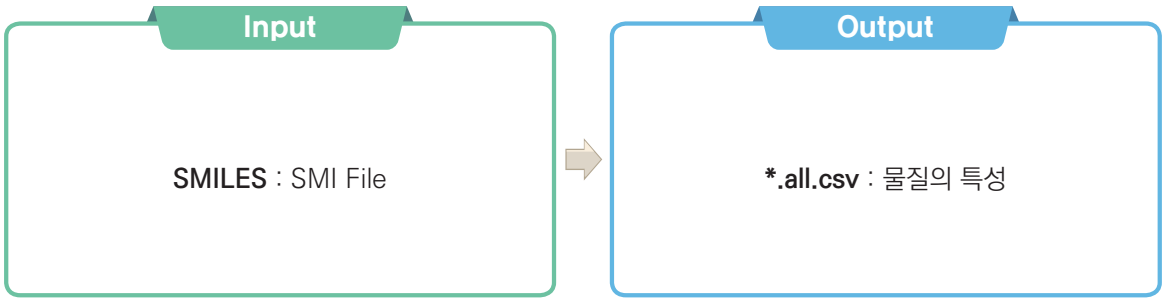
Redesigns에 의해서 생성된 화합물의 다양한 특성을 분석합니다. 이를 통하여 독성을 가지지 않는 화합물의 선별과 약물 후보물질의 더 좋은 특성을 가지는 화합물의 선택이 가능합니다.

SPICA는 SMILES형식의 화합물 정보를 입력으로 다양한 화합물 특성을 분석합니다. 이를 위해서 딥러닝 모델 중 BERT 모델과 ChEMBL에 등록되어 있는 Drug-Likeness Compounds를 이용하여 pre-training을 수행하였으며, 이후 DNN 모델을 추가하여 각 특성별로 전이 학습을 하였습니다.

나) 학습 데이터

- ChEMBL (https://www.ebi.ac.uk/chembl/)
- MoleculeNet (https://moleculenet.org/)

다) Input/Output



- Input : SMILES 형식의 화합물 구조를 담고 있는 SMI 파일

라) 결과 해석

SPICA 결과 파일은 각 화합물에 대한 다음 특성을 가질 확률 값과 분류 결과를 담고 있습니다.

- CYP1A2, CYP2C19, CYP2C9, CYP2D6, CYP3A4, CYP3A4
- BBB permeability
- ESOL
- hERG

Part 3

CSK Studio 신경퇴행성

1) NetExp
2) Molecule Generation
3) BioActivity Prediction
4) Drug Neighbor

3 CSK Studio 신경퇴행성

CSK Studio는 인공지능을 기반으로 한 퇴행성 뇌질환 선도물질 탐색 플랫폼입니다.

새로운 약물의 표적을 찾고, 잠재적 분자를 생성 그리고 뇌질환 신약개발을 위해 그 분자들의 속성을 분석할 수 있습니다.

1) NetExp

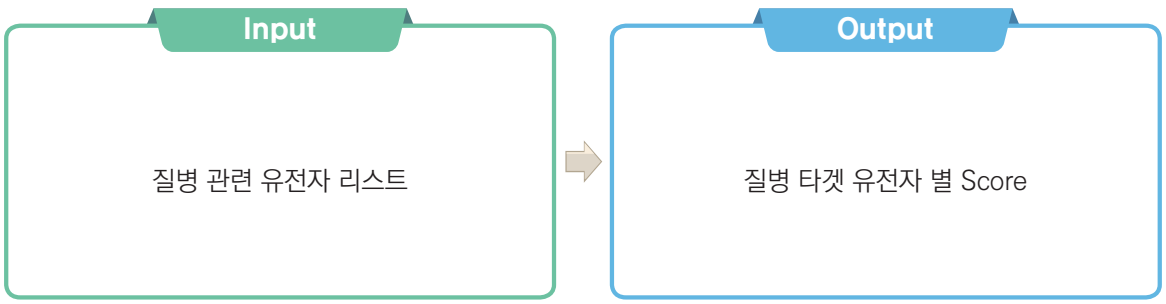
가) 모델 설명

Protein 네트워크 기반의 표적 예측 모델로 뇌질환과 관련되어 있으며 현재 1개의 질병에 대한 표적을 예측하는 모델입니다. 뇌 관련 transcriptome, proteome, transcription regulation 등의 데이터 및 annotation, pathway, domain 등의 meta-data 등을 기반으로 신규 표적을 예측합니다.

나) 학습 데이터

GeneOntology, Human transcriptome database, KEGG database, Interpro Database, STRING (protein-protein interaction database)를 사용합니다.

다) Input/Output



- Input:

  Genes 쉼표로 구분된 유전자 이름의 문자열, 일반적인 유전자 이름으로 사용 가능하며, Uniprot accession number, STRING 아이디 사용이 가능합니다.

  특히, Uniprot accession number 사용을 추천합니다. 최소 10개 이상의 유전자를 input으로 사용해야 합니다.

  Input 유전자는 Textmining을 통해 관련 유전자를 찾아도 되고, transcriptome 분석 결과로 얻은 유전자 리스트를 사용해도 됩니다.

- **Output:**

  중요도(p-value 기준)에 따라 정렬된 유전자 리스트를 반환합니다.

  Query : O의 경우 input에 포함된 것이며, - 의 경우 포함되지 않았다는 의미를 가집니다.

  유전자 이름: 예측된 표적의 유전자 이름을 나타냅니다.

  Score : p-value 값이며, 낮을수록 좋다는 의미를 나타냅니다.

라) 결과 해석

결과 유전자 중 상위 10개 혹은 20개 유전자에 대해 추가 조사를 통하여 표적을 결정하면 됩니다.

결과 유전자는 p-value (0-1사이의 값)으로 제시되며, p-value 값이 낮을 수록 우선순위가 높은 유전자를 의미합니다.

계산과정은 Input 유전자를 기준으로 표적 가능성이 높은 유전자를 예측하는 방식이며, 결과 유전자 중에는 Input에 포함된 유전자가 포함될 수도 있습니다.

2) Molecule Generation (Standigm BEST MolGen)

가) 모델 설명

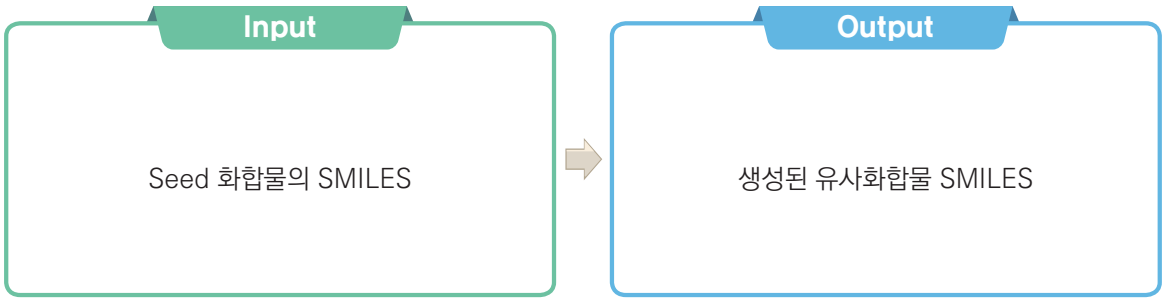
구조 생성모델 화합물 예측 모델로 입력으로 주어진 화합물을 ChemMap에 표시하고 그와 가까운 화합물을 생성합니다.

나) 학습 데이터

ZINC, MOSES 등 공개 data등을 curation한 standigm dataset 사용하여 학습하였습니다.

* 참고 –〉 ZINC : http ://zinc15.docking.org/ , MOSES : https://github.com/molecularsets/moses

다) Input/Output



- Input은 SMILES로 표현된 분자 1개, duration은 계산시간, max count는 생성할 분자 수, sample range는 분자 생성 시 사용할 distance 값입니다.
- KAIDD Browser에서 이용 시 duration=120, max count=100, sample range=2.5로 고정되어있으며 Swagger를 통해 이용할 경우 사용자가 조정하여 사용가능 합니다.

라) 결과 해석

Input은 입력분자의 구조(SMILES), 생성에 사용할 시간(샘플링 하는 시간), 생성할 화합물의 최대 개수, 샘플링거리(seed 로부터 얼마나 가까운 거리에 있는 화합물을 샘플링 할 것 인지, 값이 높을수록 더 넓은 범위에서 화합물을 생성할 수 있고 더 다양한 구조를 얻을 수 있게 됩니다.)

입력한 Input 구조에 따라 생성된 유사 화합물 구조 SMILES가 출력됩니다.

샘플링 방법이므로 오랜 시간 샘플링을 하면 많은 화합물을 얻을 수 있습니다.

생성화합물의 구조를 통해서 새로운 아이디어를 얻거나 새로운 종류의 일들을 수행할 수 있습니다.

* 단, sampling_range는 1~3 사이의 값에서 sampling이 제일 잘됩니다.

3) BioActivity Prediction

가) 모델 설명

심장독성, 간독성, BBB투과 여부를 예측하는 모델입니다.

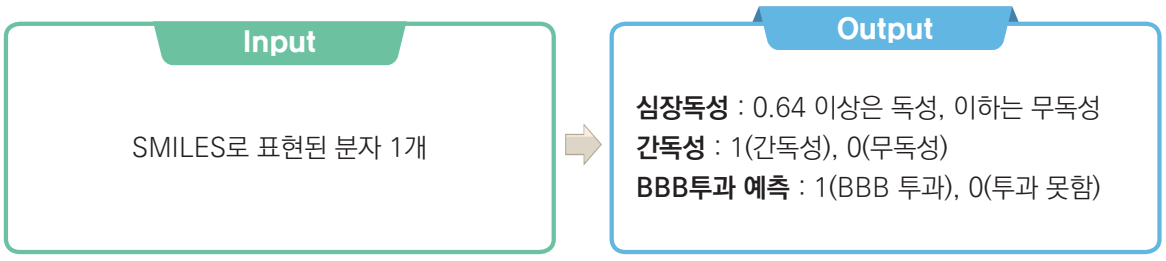
- **심장독성 예측** : 채널 단백질에 대해서 blocking을 안 할 수 있는 약물을 예측(독성이 없는지를 예측하는데 특화된 모델) 3개의 딥러닝 모델(화합물의 특성을 수치로 변환하여 descriptor를 이용한 모델, Fingerprint를 이용하여 학습한 모델, 화합물을 그래프 모델)을 합쳐서 hERG 심장독성을 일으키는지 일으키지 않는지를 알 수 있는 모델을 구현하였습니다. 특히 독성이 없는 것을 잘 예측하도록 특화된 모델입니다. 애매한 것들은 독성이 있는 것으로 분류합니다.
- **간독성 예측** : 기존에 알려진 머신러닝 알고리즘(12가지)를 이용한 앙상블 모델(Ensemble Model)을 사용하였습니다.
- **BBB 투과여부 예측** : light gradient boosting algorithm 사용하였습니다.

나) 학습 데이터

- **심장독성 예측** : power blocker 심장독성을 일으킨다고 알려진 7000여 종의 약물들, 8000여 종의 무반응 약물들에 대해 학습을 진행하였습니다.
- **간독성 예측** : 1000여 종의 독성을 일으키는 물질, 1000여종의 독성을 일으키지 않는 물질을 사용하였습니다.
- **BBB 투과도 예측** : BBB를 투과한다고 알려진 분자 5천여개, BBB를 통과하지 않는 분자 1700여개에 대해서 학습을 진행하였습니다. (데이터들은 logBB 값이 알려진 분자들의 logBB 값을 구해서 학습을 진행)

다) Input/Output

* output 순서 : 심장 독성, 간독성, BBB 투과 예측



- Receptor : ATOM의 정보만 포함되어 있어야 합니다.
- Ligand : ATOM, CONNECT 정보가 포함되어 있어야 합니다.

라) 결과 해석

- **심장독성 예측(hERG blocker)** : 실수 값으로 결과가 반환되며, 0.64이상은 독성, 이하는 무독성입니다.
- **간독성 예측** : 1과 0으로 결과가 반환되며, 1은 간독성이고 0은 무독성입니다.
- **BBB 투과도 예측** : 1과 0으로 결과가 반환되며, 1은 BBB투과, 0은 BBB 투과 못하는 구조로 예측합니다.

4) Drug Neighbor (Standigm BEST neighbor)

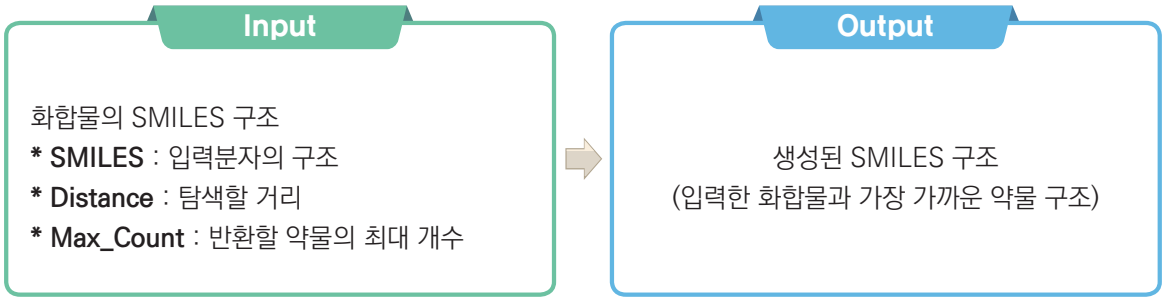
가) 모델 설명

Molecule과 ChemMap 상에서 가까운 거리내에 존재하는 물질들을 drug bank 데이터베이스에서 검색합니다.

나) 학습 데이터

ChemMap 모델을 이용하여 drug bank의 모든 data를 encoding 하여 database를 구성하였습니다.

다) Input/Output



- Input으로는 SMILES 구조를 1개 입력합니다. distance는 탐색할 거리이며, max_count는 반환할 약물의 최대 개수입니다. KAIDD Browser에서 이용할 시 distance=10, max count=20으로 고정되어 있으며 Swagger 페이지에서 사용자가 자유롭게 조절하여 이용 가능합니다.

라) 결과 해석

Input값으로 구조(SMILES), distance=10, max_count=20를 입력할 경우 drug bank에서 seed와 거리상으로 10내에 있는 물질을 최대 20개까지 검색한다는 의미입니다.

입력분자의 구조(SMILES), 탐색할 거리, 반환할 약물의 최대 개수를 조절가능 합니다.

Output으로는 생성된 화합물의 구조(SMILES)를 확인하실 수 있습니다. 결과로 약물을 반환하므로 약물을 drugbank에 들어있는 화합물로 정의하고 drugbank에 있는 화합물들을 반환하며, 가까운 약물의 구조(SMILES), 입력한 분자와의 거리, 반환된 약물의 drugbank ID를 제공합니다.

* drugbank ID를 이용하여 drugbank 사이트에 접속을 하면 해당되는 약물에 대한 자세한 정보를 얻을 수 있습니다.

* Drug Neighbor에 있는 화합물은 수 천개밖에 없으므로 max_count를 크게 지정하여도 반환 받는 화합물의 개수는 많지 않습니다.

입력으로 주어진 화합물을 ChemMap에 표시하고 그와 가까운 약물을 반환합니다. 입력으로 주어진 화합물과 유사한 약물이 어떤 것들이 있는지 확인하고 이를 통하여 입력으로 주어진 화합물이 어떠한 성질을 가지고 있을지 역으로 추적해볼 수 있습니다.

입력한 화합물과 가까이에 있는 약물이 어떤 것인지를 보고 약물들이 가지고 있는 정보들을 역으로 활용함으로써 입력한 화합물들의 정보, 그 화합물들에 대한 성질을 이해하는데 도움이 될 수 있습니다.

Part 4

MiLearn™ 항암신약

1) AiCAD
2) AiGPro
3) AiKPro
4) AiP450
5) CRX4

4 MiLearn™ 항암신약

MiLearn™은 Kinase, GPCR 표적 항암신약 개발을 위한 항암신약 개발 혁신 인공지능 플랫폼입니다.

1) AiCAD (Ai for Cancer Drug Discovery)

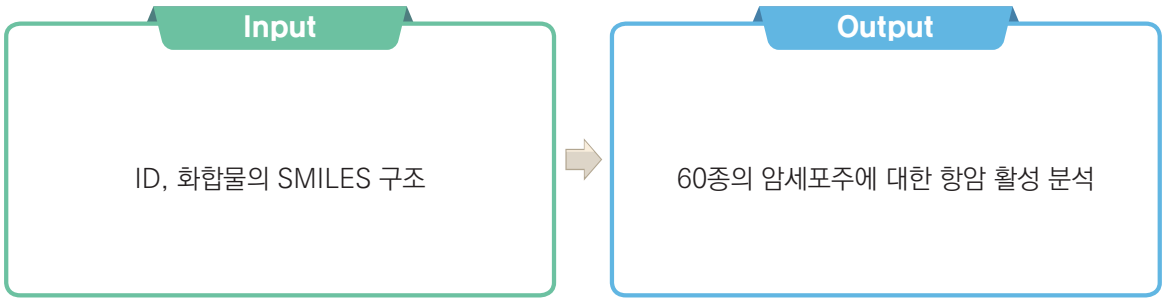
가) 모델 설명

표현형 기반 항암물질 검색 모델입니다.

나) 학습 데이터

NCI60 제공 60종의 암세포주와 3,000종 화합물의 GI50 사용하였습니다.

다) Input/Output



라) 결과 해석

- 60종의 암세포주의 조직변이 단백질, 발현량에 선택적으로 작용하는 약물을 탐색합니다.
- 60종 세포주의 GI_{50} 10uM 기준으로 활성을 가질 확률 예측합니다.
- 결과값은 0~1의 범위로 %inhibition (예) 〉0.5 이면 50% 저해 가능)

2) AiGPro (Ai for GPCR Profiling)

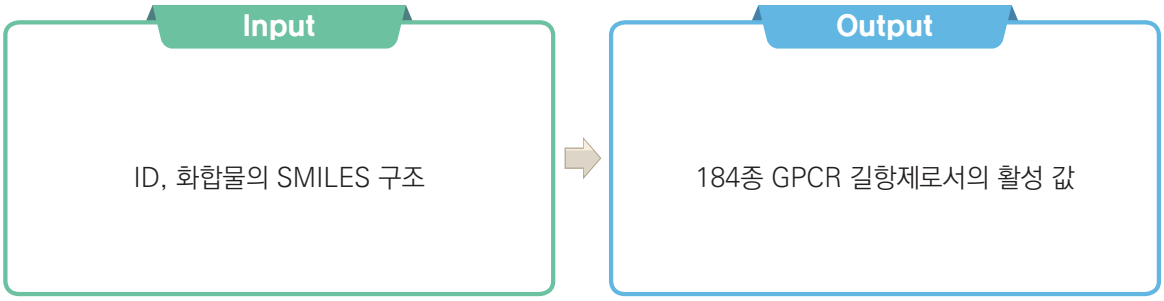
가) 모델 설명

다중서열 정렬 기반 GPCR 길항 활성 예측 모델입니다.

나) 학습 데이터

GLASS 데이터에서 약물활성 500개 이상인 184종의 GPCR 데이터를 확보하였습니다.

다) Input/Output



라) 결과 해석

- 184종의 GPCR에 대해 IC_{50} 1uM 기준으로 길항활성을 가질 확률을 예측합니다.
- 결과값은 0~1의 범위로 %inhibition (예) >0.5 이면 50% 저해 가능)

3) AiKPro (Ai for Kinase Profiling)

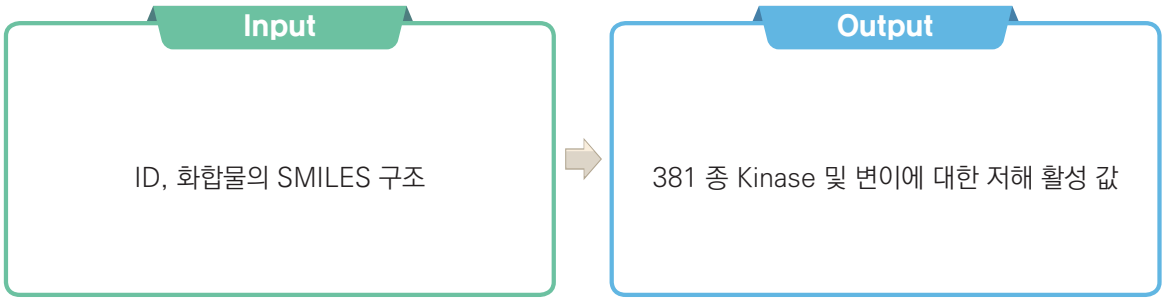
가) 모델 설명

다중서열 정렬 기반 Kinase 저해 활성 예측 모델입니다.

나) 학습 데이터

- GSK, Takeda, Pfizer에서 공개한 Published Kinase Inhibitor Set 2(PKIS2) 데이터로 학습하였습니다.
- Kinase 서열 정보 확보: Kinase 381종

다) Input/Output



라) 결과 해석

- 381종의 kinase에 대해 IC_{50} 1uM 기준으로 활성을 가질 확률을 예측합니다.
- 결과값은 0~1의 범위로 %inhibition (예) 〉0.5 이면 50% 저해 가능)

4) AiP450 (Ai for CYP450 Profiling)

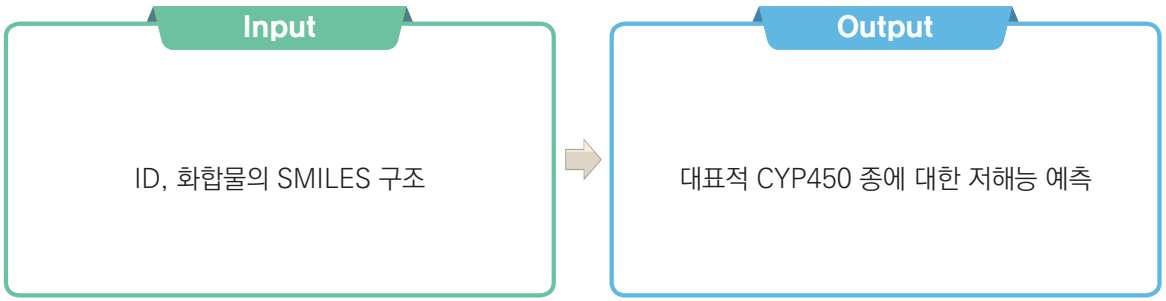
가) 모델 설명

5종의 CYP450 저해 가능성 예측 모델입니다.

나) 학습 데이터

- 공개된 Pubchem library에서 대표적인 5개의 CYP450 type 저해 활성을 수집하였습니다. (1a2, 2c9, 2c19, 3a4, 2d6)
- K-MEDI hub 신약개발지원센터에서 생산한 1,000종의 화합물에 대한 5개의 CYP450 type 데이터로 학습하였습니다.

다) Input/Output



라) 결과 해석

- 5개의 CYP450 type IC_{50} 1uM 기준으로 저해 확률을 예측합니다.
- 결과값은 0~1의 범위로 %inhibition (예) 〉0.5 이면 50% 저해 가능)

5) CRX4(Kinase and GPCR likeness model)

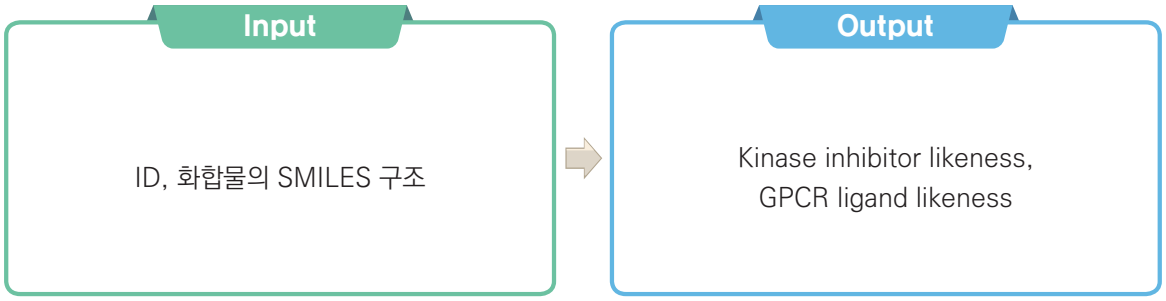
가) 모델 설명

Kinase inhibitor likeness, GPCR ligand likeness를 예측합니다.

나) 학습 데이터

ChEMBL에 있는 화합물과 Kinase & GPCR 데이터를 사용하였습니다.

다) Input/Output



- Input에 입력하는 ID는 물질 고유 ID가 아닌 사용자가 임의로 지정하는 ID입니다.

라) 결과 해석

- 예측 값은 얼마나 Kinase inhibitor 혹은 GPCR ligand 와 구조적으로 유사한지를 의미합니다.
- 예측 값의 숫자가 높을 수록 유사성이 높습니다. ( )6 이면 IC_{50} 1uM 이하 예상)
- 본 모델은 타겟에 대한 결합력 예측이 아니라 Kinase 와 GPCR 약물의 구조적 유사성 예측을 통해 각 타겟에 대한 적용범위 (applicability domain)를 제한함으로써 타겟 적합성을 높이는데 활용 가능합니다.

Part 5

AIDrug 빅데이터/AI 신약

1) ADMET
2) Toxicity 등
3) De novo Design
4) Virtual Screening
5) 기타

5 AIDrug 빅데이터/AI 신약

AIDrug는 빅데이터/인공지능 기반 신약개발 플랫폼입니다.

AI 기반 약물설계와 선도/후보물질 발굴 및 최적화와 검증을 수행하고 인공지능 원천기술 개발 및 딥러닝 기반 예측 시스템 구축을 하여, 이를 신약개발 플랫폼 및 고성능 컴퓨팅 클라우드로 구축했으며, 개방형 API와 지속적인 서비스를 제공합니다.

1) Absorption, Distribution, Metabolism, Excretion & Toxicity (ADMET)

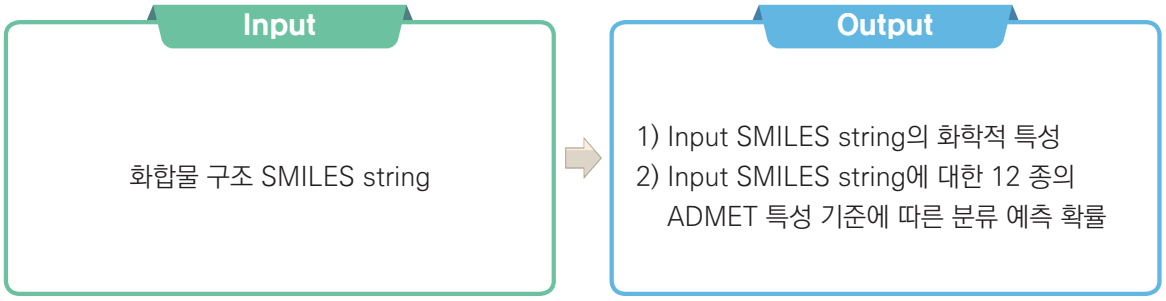
가) 모델 설명

흡수, 분포, 대사, 배설(ADME) property와 독성(T)을 빠르고 정확하게 예측하는 인공지능 기반 분석 모델입니다. 본 모델은 인공지능 기법을 활용한 12 종의 ADMET 예측 모델로서 주어진 분자 구조 정보를 이용해서 흡수, 분포, 대사, 배설 특성 및 독성을 예측합니다.

나) 학습 데이터

Category	Model	Criteria	Dataset Size
흡수 (Absorption)	Passive absorption (permeability)	Permeability in cells	3,482
분포 (Distribution)	Blood–brain barrier (BBB) penetration	logBB ≥ −1	2,080
	P-gp substrates	$CaCO_2$ ratio ≥ 2	3,017
대사 (Metabolism)	CYP1A2 inhibition	IC_{50} < 10μM	14,469
	CYP2C9 inhibition	IC_{50} < 10μM	10,524
	CYP2C19 inhibition	IC_{50} < 10μM	14,654
	CYP2D6 inhibition	IC_{50} < 10μM	16,914
	CYP3A4 inhibition	IC_{50} < 10μM	17,261
배설 (Excretion)	Human liver microsomal stability	HCLint < 20ml/min/kg	2,245
	Mouse liver microsomal stability	MCLint < 90ml/min/kg	475
	Rat liver microsomal stability	RCLint < 85ml/min/kg	911
독성 (Toxicity)	hERG inhibition	IC_{50} < 10μM	4,678

다) Input/Output



라) 결과 해석

1) Input SMILES string의 화학적 특성을 보여줍니다.

2) Input SMILES string의 화학적 구조와 특성을 기반으로 흡수, 분배 대사, 배출, 독성과 관련된 특성을 분류할 확률을 수치적으로 나타냅니다. 12 종의 ADMET 예측 모델이 제공하는 결과 확률 값이 각각 100%에 가까울수록, Input SMILES string이 12 종 모델의 특성으로 분류될 가능성이 높아집니다.

2) Toxicity 등

가) 모델 설명

약물 분자들의 Atom Meta feature, edge type feature 들을 사용하여 graph encoder 구조 데이터 활용한 그래프 뉴럴 네트워크(Graph Neural Network)와 기존 모델보다 더 robust하고 overconfidence problem를 해소하는 효과가 있는 베이지안 딥러닝(Bayesian Deep Learning)들을 동시에 적용한 16 가지의 독성 등 예측 모델입니다.

나) 학습 데이터

카테고리	독성 예측 항목	(Full name, 설명)	데이터 수 (toxic/ non-toxic)	DB
Tox21 (12)	ER	Estrogen receptor alpha	6,193 (793/5,400)	MoleculeNet – Tox21 Data
	ER-LBD	Estrogen receptor alpha, ligand-binding domain	6,955 (350/6,605)	
	Aromatase	Aromatase	5,821 (300/5,521)	
	AHR	Aryl hydrocarbon receptor	6,549 (768/6,781)	
	AR	Androgen receptor	7,265 (309/6,956)	
	AR-LBD	Androgen receptor, ligand-binding domain	6,758 (237/6,521)	
	PPAR-gamma	Peroxisome proliferator-activated receptor gamma (PPAR-gamma)	6,450 (186/6,264)	
	ARE	Antioxidant Responsive Element	5,832 (942/4,890)	
	HSE	Heat shock factor response element	6,467 (372/6,095)	
	ATAD5	ATPase Family AAA Domain Containing 5	7,072 (264/6,808)	
	MMP	Mitochondrial membrane potential (MMP)	5,810 (918/4,892)	
	P53	항암 관련 p53 단백질	6,774 (423/6,351)	
Distribution	BBBP	Blood-brain barrier (BBB) penetration	2,039 (1,560/479)	MoleculeNet – BBBP 데이터
Target Activity	EGFR	항암 관련 Epidermal Growth Factor Receptor	43,082 (4729/38,353)	CheEMBL – EGFR관련 항암제 active/ inactive assay data
	TGFR1	항암 관련 TGF beta receptor 1	9,577 (647/8,930)	CheEMBL – TGFR1관련 항암제 active/ inactive assay data
	VGFR2	항암 관련 Vascular endothelial growth factor receptor 2	33,611 (5855/27,756)	CheEMBL – VGFR2관련 항암제 active/ inactive assay data

- 학습 데이터의 경우 위 테이블에 독성 데이터를 기반으로 물질에 해당하는 “SMILES”와 독성/비독성 또는 Active/Inactive 해당하는 “1/0”를 큐레이션 된 학습데이터들을 사용하였고, MoleculeNet과 CheEMBL에서의 Binary label 데이터를 사용하였습니다.
- 학습에 사용되는 SMILES 데이터를 통해 물질 구조의 Atom Meta feature (Degree of molecule, Total number of hydrogen, Implicit valence, Aromatic or not 등)와 Edge feature (single, double, triple, aromatic bonds)들을 활용하였습니다.
- 학습을 위한 label 데이터의 경우 MoleculeNet은 NIH, EPA, FDA가 협력하여 화합물의 독성을 예측할 수 있는 새로운 방법을 만들기 위한 컨소시엄으로 전문가들이 검토한 label 데이터들을 활용하였습니다.

다) Input/Output

Input

화합물 구조 SMILES string

Output

- 독성 등 16가지 모델의 예측 확률 값(0~1)
- 항목 : ER, ER-LBD, Aromatase, AhR, AR.AR-LBD, PPAR gamma, Nrf2/ARE, HSE, ATAD5, MMP, p53, BBBP, EGFR, TGFR1, VGFR2

라) 결과 해석

- 독성을 일으킬 수 있는 생물학적 경로를 교란시키는 잠재성과 관련한 수많은 세포 및 생화학 분석을 통해 잠재적 독성을 예측합니다.
- 0 ~ 1사이의 확률 값으로 1에 가까울 수록 잠재적 독성이 있을 가능성이 높습니다.

Output Data:

1. ER: 유방암 치료 항암 관련
2. ER-LBD: 유방암 치료 항암 리셉터에서 리간드 결합 도메인 활성
3. Aromatase: 유방암과 관련
4. AhR: 면역독성과 관련
5. AR: 전립선암과 관련
6. AR-LBD: 전립선암과 관련 항암 리셉터에서 리간드 결합 도메인 활성
7. PPAR gamma : 심장독성과 관련
8. ARE : 항산화작용과 관련된 Pathway로 신경퇴행과 관련
9. HSE: 외부 스트레스로 인한 세포 손상과 관련
10. ATAD5: DNA 손상과 관련
11. MMP: 에너지 대사와 관련하여 세포독성과 관련
12. p53: 항암 작용 메커니즘과 관련
13. BBBP: 뇌 속의 endothelial cell(내피세포)를 compound가 통과 관련
14. EGFR: 항암 치료에 큰 영향을 주는 관련된 receptor 활성 관련
15. TGFR1: 항암 치료에 큰 영향을 주는 관련된 receptor 활성 관련
16. VGFR2: 항암 치료에 큰 영향을 주는 관련된 receptor 활성 관련

3) De novo Design

가) 모델 설명

표적단백질의 active/binding site에서의 fragments 기반 De novo drug design을 위한 인공지능 기반의 생성 모델입니다.

나) 학습 데이터

- 강화 학습 데이터: 강화학습에 적용할 리워드인 Ligand-based drug design (LBDD) & Structure-based drug design (SBDD)을 활용한 스코어 계산을 위해 Zinc DB를 통해 학습데이터를 위한 fragments dataset 또는 분자의 모티프 dataset을 만들었습니다.
- 학습데이터를 만들기 위한 fragmentation이 상당히 오래 걸리는 작업이고, 분자의 개수가 많을수록 학습 시간이 오래 걸리기 때문에, 랜덤하게 선택된 2,500여개 분자들을 학습데이터로 활용합니다.

다) Input/Output

Input

- 단백질 구조와 리간드 구조 데이터 (데이터형식: pdbqt)

or

- PDB에서 단백질과 리간드의 PDBID

or

- 알파폴드2나 로제타폴드를 통해 예측된 단백질 구조 데이터(데이터형식: pdbqt)

Output

- ***_년-월-일_시:분:초.csv** : 생성된 모든 물질 SMILES 데이터
- ***_rule_filtered.csv** : 약물 특성과 관련된 규칙를 기반으로 평가된 물질의 SMILES 데이터
- **results.txt** : 역합성 분석을 적용한 결과로 SMILES 데이터와 관련 합성 정보

라) 결과 해석

Input Data:

단백질 구조를 기반으로 fragments를 확장하면서 약물 디자인합니다.

1) active/binding site가 있는 단백질 구조 데이터와 관련 있다고 알려진 리간드 구조 데이터입니다.
   - 리간드가 반응하는 active/binding site를 중심으로 약물을 디자인합니다.

2) active/binding site가 있는 기존에 알려진 단백질 또는 알파폴드2나 로제타폴드를 통해 예측된 단백질 구조 데이터입니다. (리간드 구조 데이터가 없어도 가능)
   - active/binding site를 탐색하면서 약물을 디자인하기 때문에, 수행 시간이나 결과물에 있어 리간드 정보가 있을 때 보다는 좋지 않습니다.

Output Data:

1) 생성 모델을 통해 만들어진 SMILES

- 강화학습 기반의 생성 모델을 활용한 약물 구조 데이터로 LBDD와 SBDD를 동시에 고려한 생성된 물질 구조 정보 데이터입니다.

2) Post hoc filtering

- 생성모델이 생성한 분자(SMILES) 중에는 약이 되기에 적합한 분자와 그렇지 않은 분자가 섞여 있으므로 post hoc filtering을 잘하는 것이 매우 중요합니다. 의약화학에서 자주 고려되는 조건은 pharmacochemical (medicinal chemistry) filter, Lipinski Rule of Five 등의 예측한 정보와 이를 기반으로 필터링 된 결과입니다.

3) 역합성 분석(선택)

- 인공지능을 활용한 역합성 분석으로 인공지능 생성학습을 통해 만들어진 물질의 경우 실제 합성이 불가능하거나 존재하지 않은 것들이 생성되기 때문에 합성이 가능하며 합성 비용 및 최적화를 위한 합성 계획 정보입니다.

4) Virtual Screening

가) 모델 설명

약물-단백질 상호작용을 예측하는 앙상블 모델입니다. 본 모델은 딥러닝을 포함한 7종의 최신 알고리즘으로 구조식이 주어진 약물과 400여종의 kinase의 결합도를 각각 예측하여서, 그 값을 종합하여 최종 결합 여부를 판단합니다.

나) 학습 데이터

28,033개의 SMILES 구조식과 252개의 kinase 단백질 쌍으로 학습하였습니다.

다) Input/Output

Input

SMILES string

Output

- 426종의 kinase에 대한 결합도 예측
- Kinase 그룹(disease/origin/kinase family)별 예측
- 예측의 신뢰도(confidence)를 약물의 ECFP 클러스터별로 제공
- 트레이닝 데이터의 하부 알고리즘의 결과를 PCA로 출력하여 신뢰도/약물의 ATC code/단백질의 EC number과 비교
- Masking 기법으로 결합부위 예측

라) 결과 해석

kinase를 표적으로 하는 약물을 디자인할 때 일반적으로 400여종의 kinase 단백질에 대해 결합 여부를 실험으로 측정하게 됩니다. 이를 표적 kinase 단백질에 효율적으로 결합하는 선도물질을 선택할 수 있을 뿐만 아니라, 표적 외 kinase 단백질에 대한 결합 또한 예측할 수 있기 때문에 잠재적인 독성에 대해서도 확인할 수 있습니다. 따라서, 본 약물-단백질 상호작용 예측 모델으로 가상 스크리닝을 통해 신약 개발에 소요되는 비용과 시간을 줄일 수 있습니다.

5) 기타

가) 모델 설명

Transformer 기반의 Self-Supervised learning을 이용하여 대규모 화합물 데이터를 사전 학습(pre-training) 하는 인공지능 모델입니다. 자연어 처리에서 뛰어난 성능을 보여준 BERT 모델 기반의 masked language model 학습 방법에 착안하여 화합물 표현방식 중 하나인 SMILES에 적용했으며 추가적으로 화합물의 QED 값과 인접 행렬 정보를 같이 학습합니다. 플랫폼에는 이미 사전 학습이 완료된 모델이 탑재되어 있으며, 사용자가 자신의 데이터셋을 제출하면 탑재된 모델에 대해 fine-tuning이 진행됩니다. 학습이 완료되면 사용자는 자신이 제출한 데이터셋으로 학습된 새로운 AI 모델을 얻게 됩니다.

나) 학습 데이터

약 900만 건의 ZINC 데이터를 사용하여 사전 학습을 진행하였습니다. 사전 학습이 완료된 후 다양한 벤치마크 데이터에 대하여 fine-tuning을 통해 평가를 진행하였으며 평균적으로 state-of-art 수준의 성능을 보여주었습니다. 벤치마크 데이터는 아래와 같습니다.

- BBBP: Blood-brain barrier penetration. Prediction of the barrier permeability.
- Tox21: Qualitative toxicity measurements on 12 different targets, including nuclear receptors and stress response pathways.
- ToxCast: Data collection providing toxicology data based on in vitro high-throughput screening.
- SIDER: Drug side-effects into 27 system organ classes following MedDRA classification.
- ClinTox: It contains clinical trial toxicity and FDA approval status.
- MUV: Benchmark dataset from PubChem BioAssay. It contains 17 challenging tasks for validation of virtual screening techniques.
- HIV: The ability to inhibit HIV replication.
- BACE: Binding results for a set of inhibitors of human β-secretase 1.
- Esol: Water solubility data.
- Freesolv: Hydration free energy of small molecules in water

다) Input/Output

Input

- **Dataset.csv** : SMILES와 label이 담긴 csv 파일
- **Task** : Classification / Regression 선택
- **Split ratio** : 0~1 사이의 train set, test set 비율
- **Time** : 학습 제한시간(분)

Output

- **Dm.json** : 모델 파라미터 및 학습 결과 정보
- **Loss_graph.png** : Epoch에 따른 loss값 변화 그림
- **Test_score.png** : Test set에 대한 평가 결과 그림
- **Finetuned_model.pt** : 학습된 모델 파일
- **Result.csv** : Epoch에 따른 loss 및 성능 변화

라) 결과 해석

Epoch에 따른 train/valid loss의 변화 정보를 그림과 csv 파일로 확인 가능합니다. Test set에 대한 evaluation score도 그림으로 확인 가능합니다. 또한 사용자의 데이터셋으로 fine-tuning된 모델을 다운로드 받을 수 있으며 해당 모델을 학습하는데 사용한 정보를 JSON 파일에서 확인할 수 있습니다. JSON 파일은 아래와 같은 정보를 담고 있습니다.

변수 명	의미
optimizer	사용된 optimizer 종류
model	사용된 기본 모델
batch_size	학습에 사용된 batch 크기
dropout	적용된 dropout 비율
learning rate	적용된 learning rate
epoch	전체 학습 epoch 수
metric	학습 평가 기준
best_score	최고 성능

Part 6

Synbi 약물재창출

1) SynbiDrug-R

6 Synbi 약물재창출

Synbi 는 인공지능 기반 약물 재창출 플랫폼입니다.

Synbi 약물재창출 플랫폼은 기존에 임상에서 승인된 약물 빅데이터의 다양한 특성에 기반하여 새로운 항암 표적, 작용 기전, 적응증으로의 재창출 가능성을 예측하는 인공지능 모델입니다.

1) SynbiDrug-R

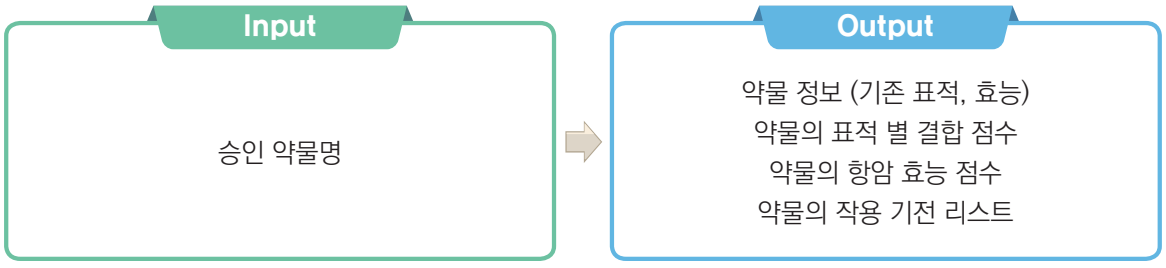
가) 모델 설명

Synbidrug-R은 기존에 임상에서 사용되고 있는 승인 약물들의 새로운 항암 표적, 기전 및 적응증으로의 재창출 가능성을 예측하는 인공지능 플랫폼으로 승인 약물을 포함한 화합물과 표적 단백질의 상호 작용 관계 정보를 기반으로 훈련된 항암 표적 예측 모델과 약물의 작용 기전 예측을 기반으로 승인 약물의 항암 효능 예측 모델로 구성되어 있습니다. 입력 특성으로는 약물의 화학 구조, 약물 처리 시 유전자 발현 변동 패턴 및 세포 기전 활성, 부작용, 표적 여부 및 아미노산 서열 정보를 사용합니다. 최적화하기 위해 심층 신경망(DNN), Convolution 신경망(CNN), 장단기 기억 신경망(LSTM), Capsule Network 등 다양한 딥러닝 모델을 활용해 모델을 최적화하였습니다.

나) 학습 데이터

총 4,845개의 승인 약물을 포함해 약 14,000여개의 화합물과 약 5,800개의 단백질 표적간 상호 작용 관계를 기반으로 표적 예측 모델을 훈련시켰으며, 항암제 304개, 비항암제 약 4,500여개를 대상으로 암 효능 예측 모델, 약물 약 2,800여개에 대한 암 세포 어세이 정보를 활용하여 작용 기전 예측 모델을 훈련시켰습니다.

다) Input/Output



- 승인 약물명의 입력 값은 약물의 국제 일반 명 (INN, International Nonproprietary Name)으로 한정되며, 대/소문자 구분 없이 영문으로만 입력 가능합니다.
- 플랫폼 웹 페이지 (http://synbi.kaist.ac.kr/synbidrug/) 또는 KAIDD 홈페이지 Platform〉 Library 내에 플랫폼에서 검색 가능한 약물명 목록을 확인하실 수 있습니다.

라) 결과 해석

약물-표적 결합 예측 확률, 약물-표적 결합 유의성/약물-항암 효능 확률, 약물-항암 효능 유의성을 나타냅니다.

• Drug Information: 약물정보

입력한 약물의 기본 정보를 나타냅니다.

PubChem CID	Drug (INN)	Original Targets (Gene Symbol)	Original Indications
약물명에 해당하는 PubChem CID	입력한 약물명	기존약물 허가 시, 알려진 표적단백질	허가 적응증 정보
60838	irinotecan	NDUFS5, CYP2D6, SDHD, FBP1, CACNG4, ALOX5, POLA1, P16094, TOP1, TLR4, CACNG5, P1	Metastatic colon cancer, Adenocarcinoma of pancreas

예) 식물성 알칼로이드 계열의 항암제 irinotecan 국제일반명 (INN, International Nonproprietary Name)을 검색하면 기존 표적의 Gene Symbol 정보 및 기존 적응증 정보 표시

• Predicted Targets: 약물의 표적 별 결합 점수

입력한 약물의 표적 단백질들에 대한 결합 점수를 나타냅니다. 결과는 표적 별 약물 결합 점수 (Prediction Score)는 0에서 1 사이의 확률 값으로써 점수가 1에 가까울수록 입력 약물이 해당 표적에 결합할 가능성이 높다는 의미이며, 이를 기준 값 0.6565를 기반으로 양성과 음성으로 분류했다. (Binding or Not) 또한 예측 양성 표적 중 기존에 알려지지 않은 것에 대해서는 신규 예측 (Novel Positives) 컬럼에 1로 표시했습니다. 플랫폼 서비스에서 총 1,149개의 항암제 단백질 표적에 대한 결합 예측 결과를 제공합니다.

Drug (INN)	Target (UNIPROT AC)	Binding Score	Binding or Not	Novel Positives
입력 약물	결합 점수 내림차순 된 표적 리스트	입력 약물과 표적의 결합 점수 (0~1)	기준 값에 따라 결합 여부를 'Positive' 또는 'Negative'로 분류한 결과	예측 양성 표적 중 기존에 알려진 관계 '0'인지 신규 예측된 관계 '1'인지 분류한 결과 예측 음성의 경우 '-'로 표시
irinotecan	ABL1 (P00519)	0.7821	Positive (〉Threshold: 0.6565)	1
irinotecan	STK10 (O94804)	0.5576	0 (〈 Threshold: 0.6565)	-

예) irinotecan의 경우, Tyrosine-protein kinase ABL1에 대한 표적 결합 점수가 기준 값보다 높게 예측되어 결합 가능성이 있는 표적으로 추정되며, 기존에 ABL1 표적에 대한 결합이 알려진 바 없기 때문에 신규 예측 후보임. STK10 표적에 대한 결합 점수는 기준 값보다 낮게 예측됨.

• Predicted Mechanisms: 약물의 작용 기전 리스트

입력한 약물의 암 연관 기전 대상, 세포 작용 기전 리스트를 나타냅니다. 결과는 세포 작용 기전 활성 예측 점수가 0에서 1 사이의 확률 값으로써 기준 값 0.5 이상일 경우, 입력 약물의 세포 작용 기전으로 예측되어 해당 리스트가 점수와 함께 제공됩니다. 플랫폼 서비스에서 항암 대표 기전 18개 중 점수가 0.5 이상인 기전에 대한 결과를 제공합니다.

Drug (INN)	Drug Mechanism	Mechanism Score
입력 약물	약물의 항암 대표 기전 GO term	약물의 세포 기전 활성 예측 점수
irinotecan	GO:0006281 DNA repair	1
irinotecan	GO:0008283 cell population proliferation	1

예) irinotecan의 경우, 암 세포에 처리 시 DNA repair 및 cell population proliferation 기전에 대한 활성 가능성이 기준값 보다 높게 예측됨.

• Predicted Cancer Indication Score: 약물의 항암 효능 점수

입력한 약물의 항암 효능 점수를 나타냅니다. 결과는 항암 효능 점수가 0에서 1사이의 확률 값으로써 기준 값 0.051 이상일 경우, 항암 효능 약물로 예측되어 플랫폼 서비스에서 결과를 제공합니다.

Drug (INN)	Cancer Type	Indication Score	Cancer Drug or Not	Novel Positives
입력 약물	항암 효능 예상 암 종 (현재 개발 중)	입력 약물의 항암 효능 점수 (0~1)	기준 값에 따라 항암 효능 여부를 'Positive' 또는 'Negative'로 분류한 결과	항암 효능 예측 양성 중 기존에 알려진 항암제 '0'이지, 신규 예측된 약물 '1'인지 분류한 결과 예측 음성의 경우 '-'로 표시
irinotecan	Cancer (General)	0.5007	Positive (〉 Threshold: 0.051)	0

예) irinotecan의 경우, 일반 전체 암에 대하여 항암 효능 가능성 점수가 모델 스코어 기준보다 높게 예측되어 (0.5007〉0.5), 항암 효능을 가질 가능성이 있는 약물로 추정되나, 기존에 항암제로 승인된 바 있기 때문에 Novel Positives는 아님.

Part 7

Smart PV 스마트약물감시

1) Smart PV irAE

2) CDM Based ADR screening tool

7 Smart PV 스마트약물감시

Smart PV 는 면역항암제 대상 다기관 임상연구를 통한 스마트 약물 감시를 목표로 제작된 플랫폼입니다.

인구학적 정보, 새로운 진단, 새롭게 투약된 약제, 혈액검사 기반 정보, 의무기록 정보,유전체 정보, 환자보고결과(PRO), 전향적 연구 추가 수집 자료를 기반으로 면역항암제 부작용과 관련된 바이오마커를 발굴 및 사전에 부작용을 예측할 수 있습니다.

1) Smart PV irAE: 인공지능 기반 면역관련 부작용 예측 모델

가) 모델 설명

Smart PV irAE는 면역항암치료 이전에 환자의 임상정보와 유전적 요인을 종합적으로 활용하여 치료에 따른 면역관련 부작용 발생을 예측하는 플랫폼으로, 다양한 면역 관련 부작용(irAE)에 대한 정확한 예측을 목적으로 개발되었습니다.

환자에게서 발생한 부작용을 흔히 발생하는 장기 및 임상적 중요성을 고려한 총 12개의 주요 카테고리로 분류하였으며, 입력된 환자의 임상 정보와 유전체 정보를 기반으로 각 카테고리별 부작용에 대한 해당 환자의 위험도를 도출해내게 됩니다.

보다 정확한 예측과 변수들간 포괄적 고려를 위해 심층 신경망 모델 (Deep neural network model)을 사용하였으며, average precision을 기반으로 최적화되어 보다 sensitive한 예측이 가능합니다.

나) 학습 데이터

국내 아산병원을 필두로 한 9개의 기관에서 등록된 685명의 환자를 대상으로 수집된 임상정보와, 그 중 608명에 대해서 수행된 blood WES 기반의 지표들을 기반으로 irAE와 유의미한 연관성을 지닌 지표들을 학습 모델의 feature로 사용하였습니다.

다) Input/Output

Input

CBC와 chemistry를 포함한
laboratory test 결과
환자의 Germline WES기반 variant VCF 및
CNV 값환자의 HLA Type 정보

Output

12개 irAE 대한 발생확률 예측 값

- 환자 CBC의 경우 정해진 단위를 맞춰주시길 바랍니다.
- 정확한 input의 형태는 example을 참고하시길 바랍니다.

라) 결과 해석

모델이 도출해낸 발생확률 값은 해당 irAE 발생확률의 예측 값이며, 0에서 1까지의 값을 가집니다.

해당 부작용이 발생하는 환자의 경우에는 1에 가까운 값을, 발생하지 않는 환자의 경우 0에 가까운 값을 보입니다.

1과 0의 극단에 있는 값이 아닌 중간에 위치하는 값의 경우, 해당 부작용에 대한 위험도를 내포한 상태라고 이해할 수 있습니다.

2) CDM Based ADR screening tool

가) 모델 설명

플랫폼 사용자는 일반인이 아닌 CDM을 이용하는 병원 대상으로 합니다.

플랫폼을 이용하기 위한 사용 환경은 R-Studio, Atlas, Achilles 등의 시스템 적인 부분이 설치된 후에 API를 사용하실 수 있습니다.

나) 학습 데이터

학습 데이터는 기존 CDM 구조에서의 약물 관련 정보 테이블(Drug exposure list)과 부작용 정보 테이블 (Drug reaction) 입니다.

부작용 정보 테이블에서는 부작용 용어에 대해 MedDRA 라고 하는 용어를 맵핑하여 사용하게 됩니다.

또한, 부작용 정보 테이블 사용시 사용하는 변수로는 부작용 반응 시작(reaction_start_date), 종료일 (reaction_end_date), 등급(grade), 부작용 지속여부 (reaction_current_status) 컬럼을 포함합니다.

다) Input/Output

- **Chi-square test API**

Input

Ohdsi CDM table에서 각 해당되는 변수

- **Cohort table** : subject_id, cohort_definition_id, cohort_start_date, cohort_end_date
- **Death table** : death_date, person_id
- **Drug_reaction table** : reaction_start_date, reaction_end_date, person_id
- **Condition_occurrence table** : person_id, condition_start_date, condition_end_date, condition_concept_id
- **Procedure table** : person_id, procedure_date, procedure_concept_id
- **Drug_exposure table** : person_id, drug_eposure_start/end_date, drug_concept_id
- **Cohort_definition table** : cohort_name
- **Concept table** : concept_id, concept_name
- **Concept_relationship table** : concept_id_1, concept_id_2
- **Concept_ancestor table** : ancestor/descendant_concept_id

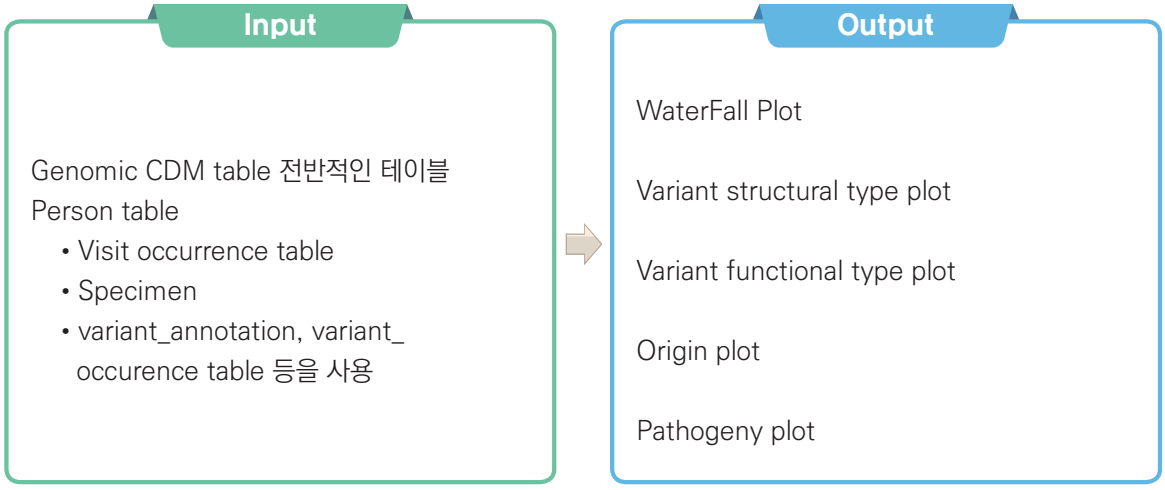
Output

Case age/gender summary

Control age/gender summary

Case/comtrol의 chi-square test의 결과 테이블

- **Genomic CDM 데이터의 특성 API**



라) 결과 해석

Chi-square test API의 경우 해당 약물을 먹었을 경우 부작용이 일어날 비율이 target, comparator 군의 차이가 있는지를 보기 위한 부분인데. 기존 target 군은 Smart PV에서 연구 대상자 (target)인 IO를 복용한 군이고, comparator 군은 비교군 (comparator)인 chemo를 복용한 환자라고 하고 예를 들어 관련 결과 변수는 Acidosis 라고 한다면, chemo 군 대비하여 IO 군에서 Acidosis가 발생할 확률의 차이가 있다 (p-value 기준 0.05 미만 인 경우, 5% 유의수준 기준)고 결론을 내릴 수 있습니다.

Genomic CDM은 전반적인 CDM 데이터 구조의 Visualization을 강조하기 위한 플랫폼으로 관련 Variant pathogeny나 Variant Origin의 비율 등을 파악하실 수 있습니다. 분석 기법이라기 보다는 Data를 한눈에 살펴보기 위한 API입니다.