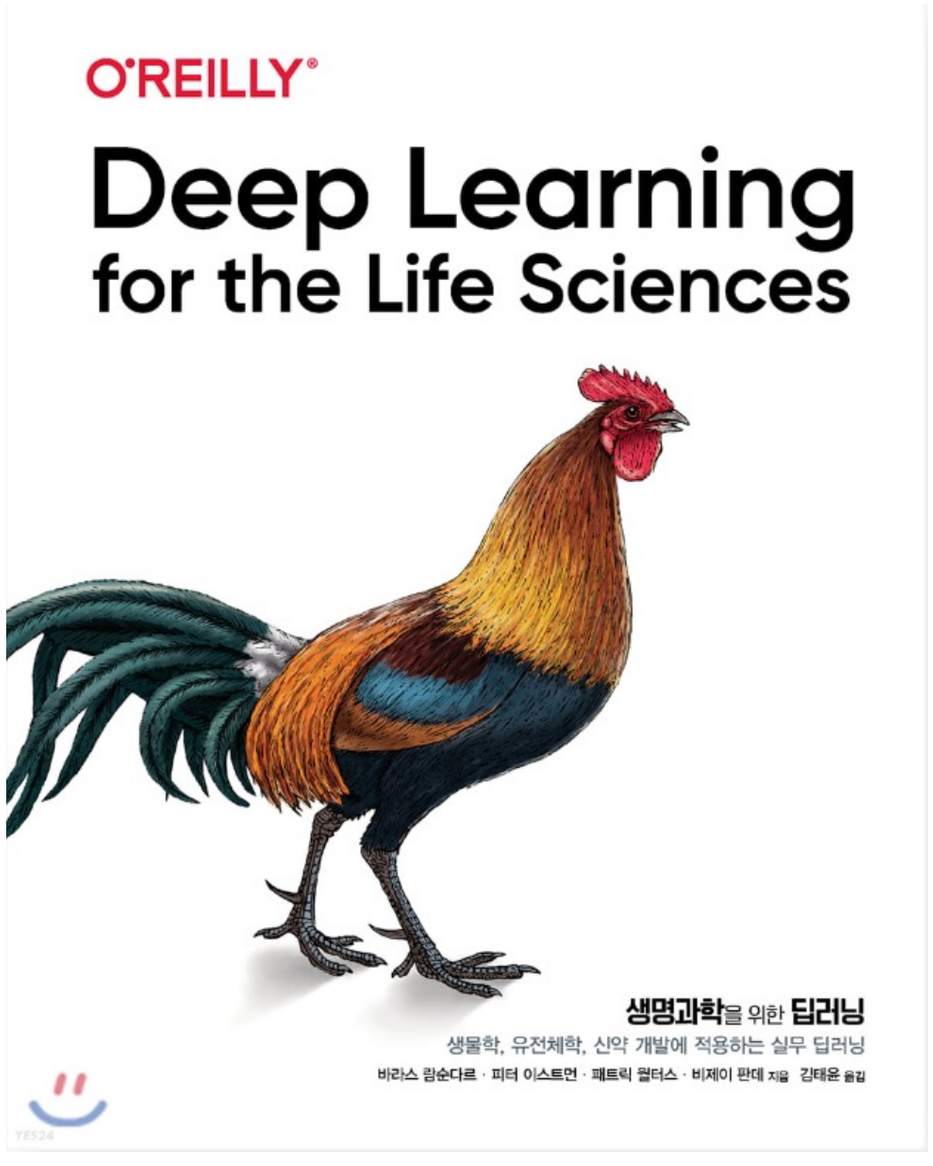
Deep Learning for The Life Science



Bio Big Data

- 현대 생명과학은 빅데이터를 다룬다
  - 수십년 걸려서 구축할 실험 데이터를 하루만에 얻기도 한다
  - 염기서열 분석 (sequencing)으로 유전자와 질병의 관계를 파악한다
  - 세포의 이미지(사진)정보가 방대하게 축적된다
  - 구조-활성관계 (SAR)은 화학 정보학의 기초가 되었다
    - Strcture-activity relationship
  - 신약개발에서 빅데이터 분석을 사용하기 시작했다

4. 분자 다루기

DeepChem

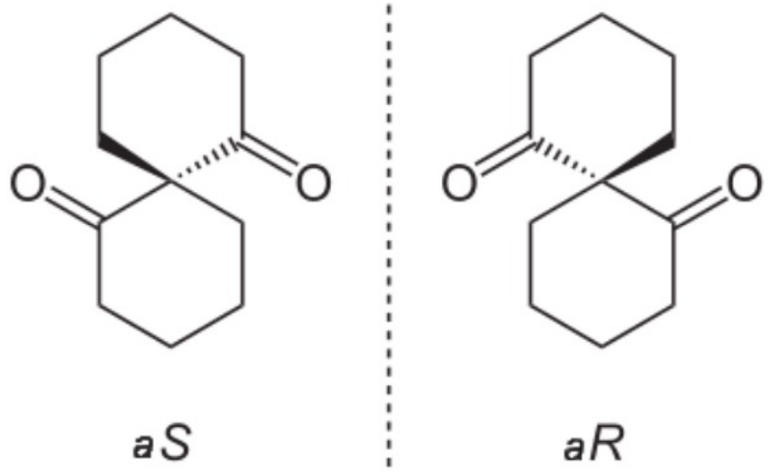
- 생명과학에 특화된 모델과 데이터셋을 제공한다
  - 텐서플로우를 기반으로 동작한다
  - 케라스는 일반 데이터의 처리에 사용되는 범용 ML 라이브러리
  - DeepChem은 분자 데이터, 유전체 데이터 등을 다루는데 특화되었다

예제

- 독성 예측 (MLP)
- MNIST (CNN)
- 용해도 예측 (GCN)
- SMART (SMILES의 확장 표현으로 검색에 사용)
- 결합의 자유에너지 예측
- RNA 전사인자 결합 예측 (CNN)
- 새로운 분자 생성 모델 (VAE)
- 리간드 기반 가상선별(VS) - GCN

분자 구조

- 분자 그래프(molecular graph)는 화학결합을 표현해준다
- 분자 구조(molecular conformation)는 3차원 공간에서 원자의 배치를 나타낸다
- 이성질체
  - 공유결합에서도 결합 축으로 회전하면 이성질체(isotype)가 형성된다
- 카이랄성(chirality)
  - 거울상 이성질체 (R, S)
- 라세미 혼합물(racemic mixture)을 만들게 되는데 물리적으로 구분하기가 어렵지만 생물학적으로는 부작용이 발생할 수 있다
  - 1975년 탈리도마이드 진정제 기형아 부작용 발생



aS aR

분자 표현법

- Molecular Featurizations (분자 피처화)
  - 분자 구조를 벡터로 표현하는 방법
  - 머신러닝의 입력으로 사용하기 위한 표현법
- SMILE (Simplified Molecular-Input Line-Entry System)
  - DeepChem 라이브러리에서 사용된다
  - RDKit은 SMILE을 다루는 다양한 함수를 제공한다
- Chemical Fingerprints
  - 분자의 여러가지 특성 유무를 1, 0으로 표현하는 방식 (화학 지문)
  - Extended-connectivity fingerprints (ECFP4)
    - FP의 몇가지 유용한 특성을 결합하여 고정 길이의 벡터로 표현
    - 두 분자의 속성이 유사하면 겹치는 부분이 많다
    - 계산이 빠르다
    - (단점) 일부 정보는 손실된다, 두 분자의 지문이 동일할 수 있다.
- 분자 표현자 (molecular descriptor)
  - 분자 구조를 설명하는 분배계수, 극성표면적 등으로 구성된다

그래프 컨볼류션 모델(GCN)

- CNN에서 커널 계수를 학습으로 찾아내듯이 분자 구조를 기술하는 계수를 학습으로 찾는다
  - 분자 그래프의 노드와 엣지를 벡터로 변환한다
- 다양한 변형
  - 그래프 컨볼류션 (GraphConvModel),
  - 위브 모델 (Weave model)
  - 메시지 전달 신경망 (MPNNModel),
  - 딥 텐서 신경망 (DTNNModel)
- 단점
  - 분자 그래프만 사용하므로 분자 구조에 대한 정보가 사라진다
    - 거대 분자에는 잘 동작하지 않는다

5. 생물물리학

BioPhysics

- 생물물리학
  - 약물이 체내에서 단백질과 어떤 결합을 하는지를 예측
  - 예: 노바티스의 imatinib-만성 골수성 백혈병 항암제, BCR-ALB 유전자 translocation으로 과발현된 tyrosine kinase와 결합해 효소의 활성을 저해
- 다중약리학(Polypharmacology)
  - 약이 한가지 타겟에만 작용하지 않는 부작용
  - 현재는 동물과 임상실험으로만 확인할 수 있다
- 화합물의 타겟 친화도 예측이 필요
  - 머신러닝 모델을 만들려면 많은 학습데이터가 필요하다
  - 적은 양의 데이터로도 동작하는 모델이 필요 → 단백질 물리학 필요

BioPhysics

- 단백질 물리학
  - 단백질의 3차원 구조를 알면 약물의 결합력을 예측하는데 유용
  - 단백질 리간드의 결합 친화도를 예측할 수 있다
    - 단백질 리간드 3차원 구조의 집합인 PDBBind 데이터셋을 사용
- 단백질은 다른 분자와 결합함으로써 특정한 동작을 수행한다
  - 약과 독성도 단백질과 분자와 결합으로 작용한다

단백질의 구조 분석

- X선 결정학 crystallography
  - 널리 사용되고 있으나 결정을 만들어야 하므로 한계가 있다
- NMR 핵자기공명
  - 다양한 구조를 파악할 수 있으나 작은 크기의 단백질 분석만 가능
- 저온 전자현미경 (cryo-EM)
  - 다수의 사진을 합성하면 해상도가 개선된다
- PDB(단백질 정보 은행)
  - 단백질의 3차원 구조를 좌표상에 표현
  - 실험으로 충분한 해상도의 데이터를 얻지 못해 단백질 구조 정보의 일부가 누락되어 있다
  - 단백질은 활성화/비활성화 상태에 따라 구조가 달라진다

단백질 서열

- 단백질은 20가지 아미노산들의 체인으로 구성
  - 각 아미노산은 공통부분과 서로 다른 사슬 residue로 구성된다
  - N-terminus에서 시작하여 C-terminus로 끝난다
- 펩타이드
  - 100개 이하의 아미노산으로 구성된 짧은 서열
- 구조 예측 방법
  - 상동성 모델링 homology modeling - 성격이 비슷한 단백질은 구조가 비슷할 것이라는 가정
  - 물리적 모델링 physical modeling – 물리적 법칙에 따른 여러 형태중 가장 안정적인 모델을 예측한다. 많은 계산이 필요하다.

단백질-리간드 결합

- 특정 단백질의 주된 기능은 특정 분자와 결합하는 것
  - 세포내의 신호전달(반응 intracellular response)은 분자에 결합하는 단백질을 통해 이루어진다
  - 약의 치료효과와 독성 작용
  - 신호전달 메커니즘을 이해하면 원하는 약효를 얻을 수 있다

생물물리학 피처화 featurization

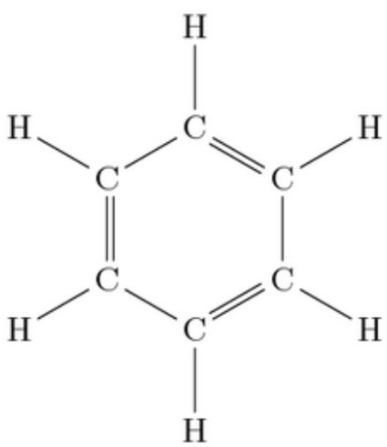
- 분자 데이터의 피처화(2차원)를 사용하면 3차원에서는 중요한 정보가 사라진다
- 피처화 기법
  - grid featurization
    - 단백질 구조를 결정하는데 중요한 역할을 하는 수소결합과 이온결합 같은 상호작용을 3차원 구조에 명시적으로 나타낸다 (비공유결합 정보)
  - atomic featurization
    - 모든 원자를 식별하고 3차원 좌표로 표현한다

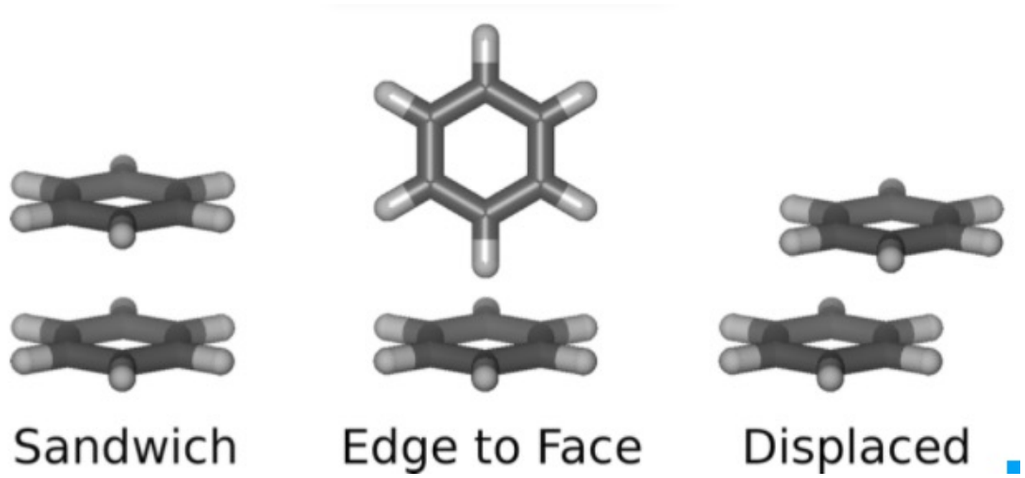
그리드 피쳐화

- 특징
  - 신뢰할 수 있는 물리학적인 정보를 얻는다
  - 알려진 물리적 특성만 탐색 가능하며 새로운 정보는 다루지 못한다
  - 수소가 띈 여분의 + 전하(*excess positive charge*)에 의해 주변의 여분의 - 전하 성분의 원자와 결합하려는 힘 (물의 결합력 예)
- 동작
  - RdkitGridFeaturizer (DeepChem) 함수 사용
  - 주어진 생물물리학 데이터에서 화학적 상호작용의 존재를 탐색하고 상호작용의 수를 포함한 피처 벡터를 만든다
  - 가까운 거리에 있는 수소결합 갯수를 측정한다.
    - 수소결합: 수소의 양전하와 산소 등의 음전하가 결합하는 힘 (물의 끓는점)
    - 이온결합: 아미노산 사이의 비공유 결합 (양전하를 띈 아미노산 잔기와 음전하를 띈 아미노산 사이에 형성) – 전체 단백질 구조를 안정화 한다
    - 파이겹침 결합:

파이겹침 결합

- **Pi-stacking interactions**
  - 방향족(아로마틱 링) 고리간의 비공유 결합
    - 방향족 고리의 모든 탄소는 파이 결합을 형성한다
    - 파이 결합: 두개의 전자의 궤도가 겹쳐지는 화학결합이며 안정적이다
  - 방향족은 평면상의 6각구조를 갖는다. DNA, RNA,등 여러 생물학적 분자에 존재한다
  - 거대분자의 구조를 안정화 한다
  - 단백질-리간드 사이의 결합에도 나타난다.
  - 그리드 피처기는 방향족을 감지하고, 거리, 각도를 통해 결합의 강도를 계산한다





원자 피처화

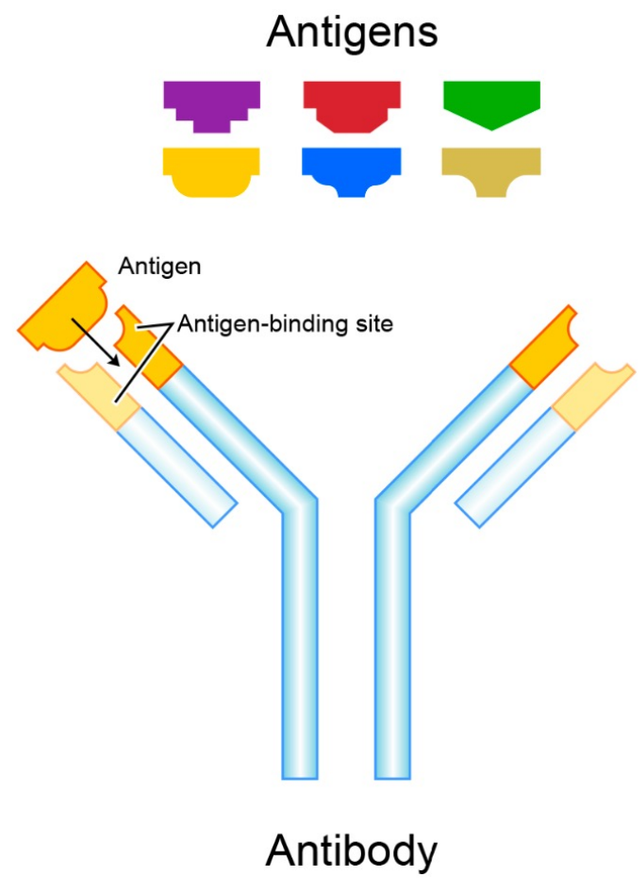
- N개 원자로 구성된 분자에 대해 (N,3) 배열을 만들고 다양한 계산을 한다
- 특징
  - (사람이 개입하지 않고) 모델이 스스로 중요한 피처(상호작용)를 선택하므로 계산량이 많다
  - 기존에 알려지지 않은 상호작용을 학습으로 찾을 수 있다
  - (위치 정보만 있으므로) 원자번호를 구분하는 별도의 배열이 필요하다
  - 이웃한 원자들의 목록을 만들어야 한다

PDBBind Dataset

- 생체 분자의 3차원 구조와 이들의 결합 친화력을 포함
  - 단백질-단백질, 단백질-리간드, 단백질-DNA, DNA-리간드 등
- 단백질 정보 은행 (PDB)에서 수집되었다
  - 잘 정리된 데이터: 핵심 core 데이터만 사용하겠다
  - 결합 친화력이 클수록 두 분자가 복합체 형태로 더 많이 존재한다
- 모든 분자생물학적 동작은 고정형이 아니라 thermodynamic behavior를 가진다
- 잔기 Residue
  - 아미노산들이 결합할 때 2 H, O가 없어진다. 남는 아미노산을 residue라고 부른다
- 시각화 도구
  - NGLView(노트북에서 동작), VMD, PyMOL, Chimera
  - (참고) docking: 단백질-리간드의 결합 부위를 예측하는 도구

항체-항원 결합

- 중요한 생물학적 상화작용중 하나
- 항체 antibody는 항원 antigen과 특이적으로 결합하는 Y모양의 단백질이다
  - 항원은 주로 병원균에서 발견되는 분자이다
  - 단일클론 항체: 세포 배양을 통해 특정 항원을 표적으로 하는 항체를 대량 생산하는 것
  - Antigen-binding site가 두개이다



6. 유전학과 딥러닝

유전체학의 딥러닝

- 모든 생물은 유전체 genome을 가지고 있다
  - 유전체를 구성하는 DNA에는 생물이 살아가는데 필요한 모든 정보가 들어있다 (소프트웨어)
- 두 가지 영역
  - Genetics(유전학) - DNA를 추상적인 정보원으로 다룬다. 유전의 패턴을 보거나, DNA 서열과 신체적 특성과의 연관성을 분석한다 (집단간 상관관계를 찾는 것)
  - Genomics(유전체학) - genome을 물리적인 머신으로 간주해 유전체를 구성하는 조각과 유전체들이 작동하는 방식을 파악한다

DNA

- DNA는 네가지 단위(A,C,G,T)가 반복적으로 연결되어 있는 긴 사슬 chain의 중합체 polymer이다
  - DNA 분자를 코로모좀이라고 한다
  - 진핵동물의 경우 DNA는 histone에 포장되어 있다 (히스톤이 언제 어떻게 풀리는지는 알려져 있지 않다)
- DNA가 소프트웨어이면 단백질은 하드웨어
  - 단백질은 세포의 모든 동작을 처리한다
  - 단백질을 구성하는 20종의 아미노산 배열 정보가 DNA에 의해서 정해진다
    - 3개 염기의 연결(codon)이 특정 아미노산을 결정한다
    - AAA는 라이신, GCG는 알라닌 등

RNA

- DNA 서열에서 단백질을 만드는 중간 과정에 정보를 전달하는 분자
  - 티민 대신 유사실 Uracil을 포함한 네개의 염기로 구성된다
  - DNA로부터 단밸질을 만들려면 정보를 전달하는 mRNA를 두 단계를 거쳐 사용한다 (센트럴 도그마)
    - DNA sequence is *transcribed(전사)* into an equivalent mRNA
    - mRNA molecule is *translated(번역)* into a protein molecule
- RNA splicing
  - 진핵생물에서, mRNA로 전사된 후 인트론을 제거하고 엑손만 연결하는 과정을 말한다
  - 전사과정에서 엑손들을 선택적으로 이어 맞춤으로써(alternative splinibg) DNA가 다양한 단백질을 발현할 수 있다 (splice variants)

RNA의 종류

- RNA의 동작은 ribosomal RNA와 transfer RNA의 영향을 받는다
  - rRNA는 아미노산의 펩타이드 결합을 만드는데 필요하다
  - tRNA는 DNA의 코돈을 인지하고 올바른 아미노산을 추가하는 역할을 한다
  - microRNA는 짧은 RNA로 mRNA에 결합해 단백질로 번역되는 것을 막는다
- siRNA도 microRNA와 유사한 동작을 하나, 이중가닥 형태이다
- Ribozyme은 화학반응을 촉매하는 효소로 작용한다 (효소는 일반적으로 단백질에 의해 이루어진다)
- Riboswitch는 두부분으로 구성되는데 한 부분은 mRNA로 작용하고 다른 부분은 특정 대사산물과 결합하여 mRNA의 번역을 조절한다

단백질 구조

- 리보솜이 mRNA를 단백질로 번역한 후, 일부 단백질은 스스로 3차원 구조를 형성하지만 대부분 chaperone 단백질의 도움을 받아 3차원 구조를 만든다
- 단백질이 번역된 후 일반적으로 PTM(Post-Translational Modification)이라는 추가적인 화학전 변형이 일어나 올바른 위치로 단백질이 옮겨진다
- 더 이상 필요하지 않게 되면 아미노산으로 분해된다

Transcription Factor(전사인자)

- 특정 단백질의 생성을 조절하는 단백질
  - 전사인자가 특정 DNA서열에 결합하는 위치에 따라 DNA전사속도가 조절된다
  - DNA는 단백질 정보를 암호화하고, 전사인자는 mRNA의 전사를 조절한다
  - 메틸화로 전사 가능성을 줄여 DNA 복제 속도가 조절되지만 동작 메카니즘은 잘 알려져 있지 않다

전사인자 결합

- 결합부위 모티프
  - 전사인자는 DNA 서열상에 존재하는 다양한 binding site motif 에 결합된다.이는 10개 정도의 염기서열로 구성된다
  - 전사인자는 유사한 모티프에도 결합할 수 있다. 즉, 위치에 따라 중요도가 다르게 작용한다 (position weighted matrix로 모델링한다)
  - 이외에도 DNA 이중나선의 물리적인 구조와, 메틸화에 따라서 TF의 동작이 영향을 받는다
- 전사인자는 히스톤이 풀려있는 DNA에만 결합할 수 있다

염색질 접근성

- Chromatin accessibility
  - 염색질 chromatin은 염색체를 구성하는 단위로 DNA, 히스톤, RNA로 구성된 거대 분자 복합체이다
  - 염색질 접근성은 염색체의 각 부분이 외부 분자에 얼마나 쉽게 접근할 수 있는지를 나타낸다
- 세포가 유전자의 발현을 조절하는데 사용하는 도구이다 (환경의 영향을 많이 받는다)
  - 앞에서 사용한 JunD 전사인자 결합 데이터는 HepG2 세포주를 사용한 실험을 통해 얻은 것이다
  - 염색질 접근성 데이터는 각각의 염색체 부위에 대한 실별자와 염색질 접근성 수치로 구성된다

RNA Interference

- RNA 간섭
  - 작은 RNA 조각인 siRNA가 상보적인 mRNA에 결합한 후 비활성화해서(silence) 단백질로의 번역을 막는 현상 (2006년 노벨상)
- 동작
  - 먼저 siRNA가 RISC(RNA-Induced Silencing Complex) 단백질 복합체에 결합하는 것으로 시작한다
  - RISC는 세포내에서 siRNA와 일치하는 mRNA를 찾아 분해한다
  - 이는 유전자 발현 조절 과정인 동시에 바이러스에 대한 방어기작으로 동작한다
  - siRNA 분자를 만들어 단백질의 발현을 조작할 수가 있다 (질병 치료나 유전자가 비활성화될 때 일어나는 일을 연구하는데 사용된다)
- 좋은 siRNA 서열을 선택하는 방법
  - 예를 들어 첫번째 염기는 A 또는 C여야 하고, 전체 염기중 GC 비율은 30~50%여야 한다
  - 이러한 경험적 지식보다 딥러닝 모델로 예측할 수 있다

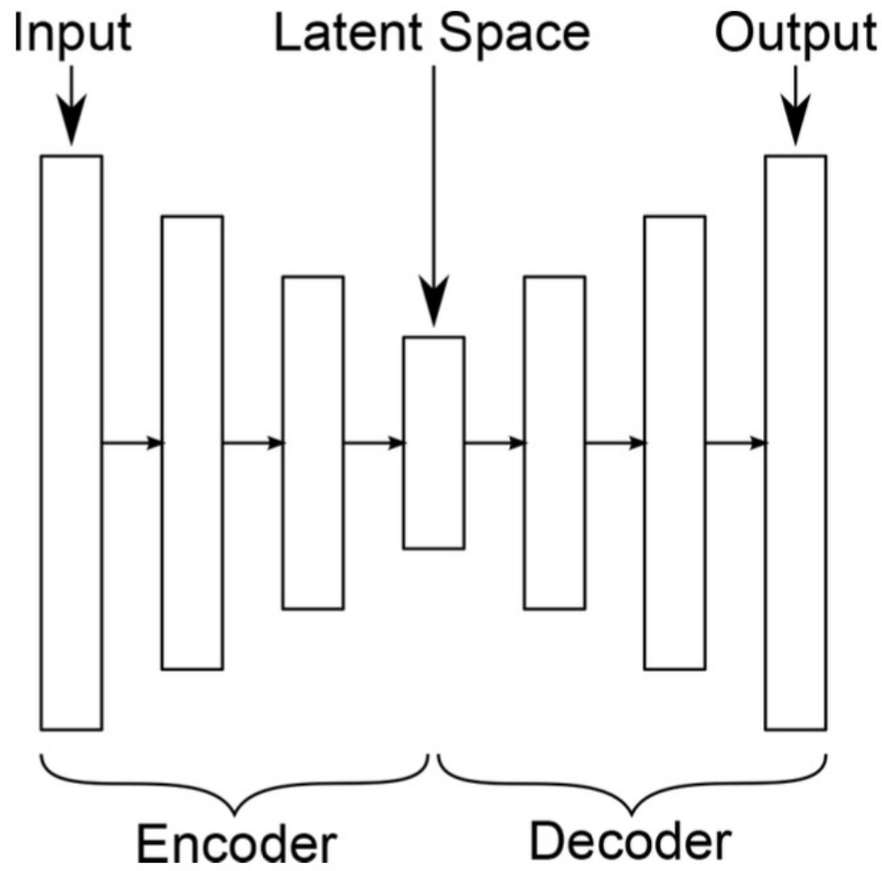
9. 생성 모델

생성 모델

- Generative model은 데이터를 입력으로 주면 새로운 데이터를 생성하는 모델이다
  - 데이터셋의 확률분포를 학습하고 새로운 샘플을 확률적으로 만든다
- VAE 모델
  - 고품질 분포를 생성하는데 유리하다
  - 즉, 생성된 샘플의 전체적인 확률 분포가 원복과 유사하다
- GAN 모델
  - 고품질 샘플을 생성하는데 유리하다
  - 즉, 생성된 각각의 샘플은 원본과 매우 유사하게 된다

VAE (Variable Auto Encoder)

- AE: 병목현상이 있도록 latent space를 사용
  - 입력 데이터와 비슷하지만 다른 데이터를 만들 수 있다
- VAE
  - 잠재영역의 벡터가 특정 분포를 갖도록
    손실함수에 새로운 항을 추가한다
  - 잠재공간에 노이즈를 추가해 학습 능력을 높인다 (잠재 벡터의 세부사항에 민감하지 않게 된다)

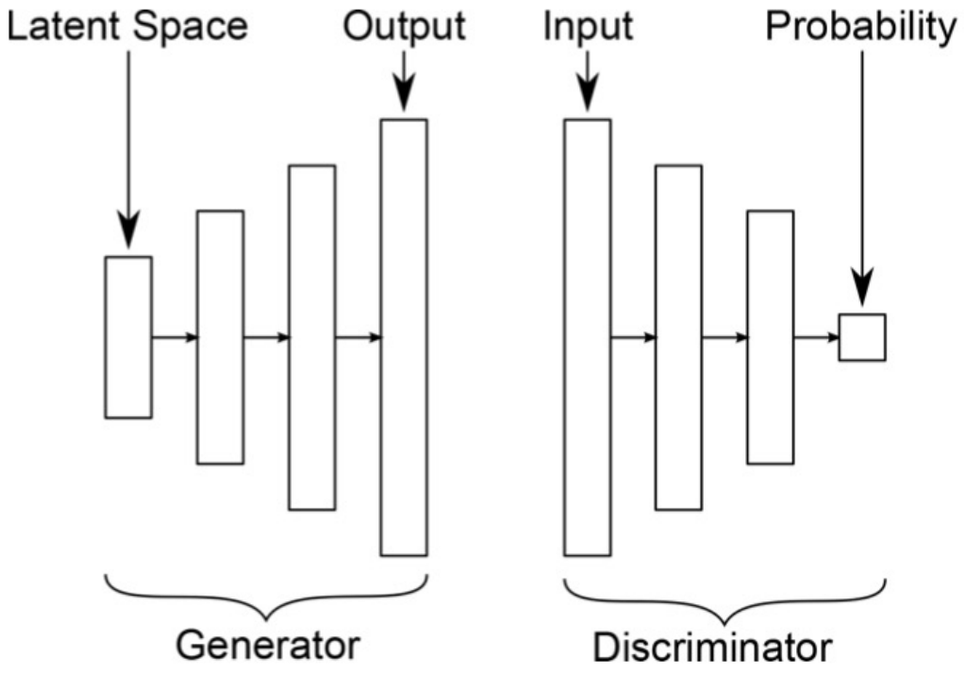


신약후보 분자 생성

- VAE를 이용한 생성 모델
  - 새로운 분자의 SMILES를 출력
  - MolrculeNet이 제공하는 SMILES 데이터셋 MUV 사용 (약 75000개)
- 생성된 SMILES 의 유효성 검증
  - 화학적으로 안정적인가? – 가상선별이 필요
  - 합성 가능한가? – 합성 단계를 학습하는 방법 시도
  - 분자의 크기를 확인한다
    - 예: 10보다 작으면 상호작용에 필요한 에너지가 불충분하고, 50 이상이면 분자의 용해도가 너무 낮다
- QED (Quantitave Estimate of Drugness)
  - 약물과 얼마나 유사한지를 판단하기 위해서 사용
  - 계산된 속성 집합과 판매된 약물의 동일한 특성 분포를 정량화 한 것
    - 예: QED > 0.5 인 분자만 고른다

GAN

- 잠재공간 벡터를 데이터로 변환하기 위해서 VAE의 디코더 대신에 생성자 네트워크를 사용한다
  - 생성된 데이터가 훈련 데이터가 얼마나 유사한지를 측정하는 손실함수를 만들기가 어렵다 (레이블이 없으므로)
  - GAN에서는 데이터에서 손실함수를 학습하는 방법을 사용한다
    - 구분자 discriminator가 생성 데이터와 학습데이터를 구분하는 시도를 하고 진위 여부 확률을 구해서 이를 생성자의 손실함수로 사용한다
    - 생성자는 구분자의 출력이 0이되게, 구분자는 구분자의 출력이 1이되게 학습한다
    - 이를 적대적 경쟁이라고 한다



생명과학의 생성 모델

- 신약 후보물질 찾기
  - 지금까지는 과학자의 지식에 크게 의존하여 한계 존재
  - 생성모델로 새로운 후보 물질을 찾을 수 있다
    - 그러나 학습에 필요한 데이터가 특정 조직의 자산이므로 공유활용이 어렵고, 유효성 검사가 어려우며, 합성의 가능성 판단이 어렵다
- 단백질 엔지니어링
  - 딥러닝 생성 모델로 원하는 특성의 단백질 서열을 얻을 수 있다.
    - 세탁의 첨가제로 때를 잘 분해하는 단백질 등
  - 고분자 화합물은 사람이 설계하는 것이 거의 불가능하다
- 복잡한 (생물학적) 시스템 모델링
  - 조직발달 과정을 예측하기 위해서 다양한 환경 조건에서 빠른 시뮬레이션 수행 (합성시험)
  - 복잡한 생리학적 과정의 모델링 또는 진화의 가설 검증에 사용될 것

11. Virtual Screening

Virtual screening

- 신약 후보 물질의 약리 활성 및 독성을 평가하는 방법
  - 실험을 통한 HTS 대신 가상선별 검사를 도입하고 있다
- structure-based virtual screening
  - 단백질의 결합 부위에 최적으로 결합하는 분자를 찾는 방법
  - 단백질의 동작을 억제하게 하여 종양, 염증, 감염의 치료제로 사용
- ligand-based virtual screening
  - 이미 알려진 분자와 유사한 동작을 하는 화합물을 찾는다
  - 기존 약물의 기능을 향상시키거나 부작용을 줄이는 것이 목표
  - 이미 효과가 확인된 분자에서 시작하여 새로운 대상을 예측한다