Artificial Intelligence for Autonomous Molecular Design: A Perspective

Category	Molecule Review paper
BCI 관련	X
Conference / Journal	MDPI Molecules (IF 4.42)
Year	2021
dataset	Review Paper
작성 완료 or 작성중	작성완료
작성자	한지웅

* Abstract

1. Aims to identify the most recent technology and breakthrough achieved by each of the components and discusses how such autonomous AI and ML workflows can be integrated to radically accelerate the protein target or disease model-based probe design that can be iteratively validated experimentally
각 구성 요소에 의해 달성 된 최신 기술과 문제를 해결하고자 하는 돌파구(breakthrough)를 식별하는 것을 목표로 하며, 그러한 자동 AI 및 ML 워크플로우를 통합 & 반본적인 실험으로 검증 될 수 있는 표적 단백질 또는 질병 모델 기반 프로브 설계를 발전하는 방법을 논의한다.
2. Our article serves as a guide for medicinal, computational chemistry and biology, analytical chemistry, and the ML community to practice autonomous molecular design in precision medicine and drug discovery.
본 논문은 정밀 의학 및 약물 발견에서 자동화 되는 분자 설계를 실천하기 위한 의학, 컴퓨터 화학 및 생물학, 분석 화학 및 ML 커뮤니티의 가이드 역할을 한다.

1. Introduction

1. Until recently, experimental laboratories have been mostly human operated;
   최근까지도 많은 연구소의 실험들은 사람의 지시에 의해 이루어져 왔고
2. The high-impact materials of today come from exploring only a fraction of the known chemical space.
   오늘날 논의되고 있는 (큰 포텐셜을 가지고 있는) 물질들은 알려진 화학 공간의 극히 일부만을 탐사하는 것으로부터 나온다.
3. Larger portions of the chemical space are still uncovered, and it is expected to contain exotic materials with the potential to bring unprecedented advances to state-of-the-art technologies.
   화학 공간의 더 큰 부분은 여전히 연구되어있지 않고, 그것은 최첨단 기술에 여태까지 없었던 발전을 가져올 잠재력을 가진 이질적인 물질들을 포함하고 있을 것으로 기대된다.

   ==========

   이러한 기술을 연구하기 위해서 아래와 같은 방법이 있는데,

4. High throughput quantum mechanical calculations, such as (density functional theory
   (DFT)) → Are the first step, towards
   a. providing insight into larger chemical space and
   b. have shown some promise in accelerating **novel molecule discovery.**
5. still requires human intelligence for different decision-making processes, (사람에게 많이 의지하고 있음) → Thus slowing down the entire process.

   ==========

6. One such particular example is the challenge associated with identifying **new metabolites in a biological sample** from mass spectrometry data → 이게 어떤 novel molecule 과 연관이 있는지, library 가 필요하다. 하지만 ↓↓↓
7. Reference libraries do not exist, and an **ML-integrated, automated workflow could be an ideal choice** → 참고 할 수 있는 db 가 매우 적다. Machine Learning 과 연관된 workflow 가 하나의 해답이 될 지도 모른다

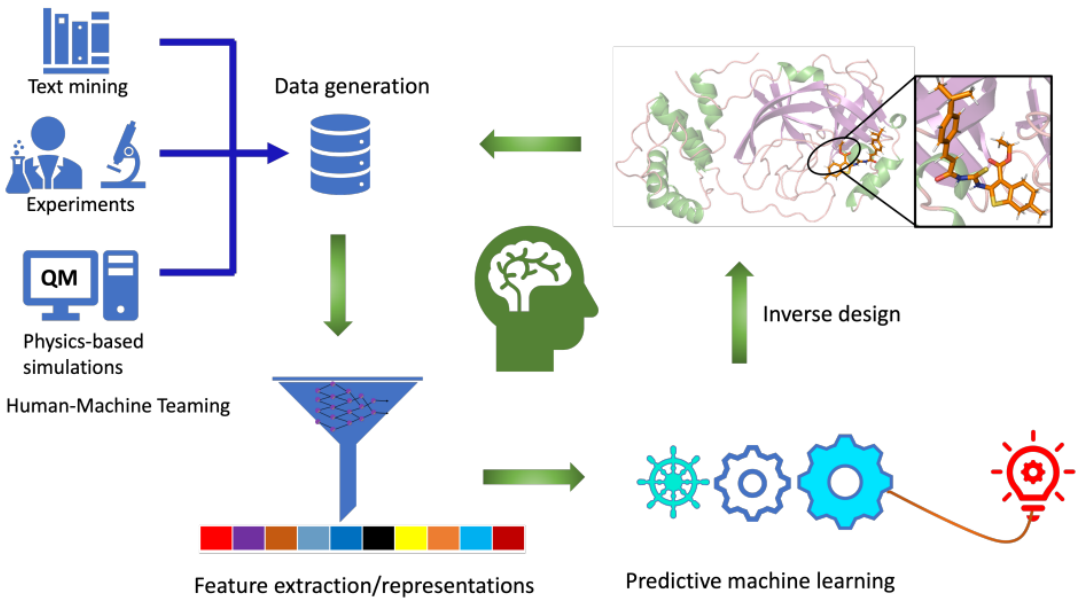
   ==========

8. We discuss how computational workflows for autonomous molecular design can **guide the bigger goal of laboratory automation through active learning approaches**.
    a. At first, we assess the performance of current state-of-the-art artificial intelligence (AI)-guided molecular design tools, mainly focusing on small molecule for therapeutic design and discovery
    처음에는 치료 및 발견을 위한 **작은 분자**에 주로 초점을 맞춘 현재의 SOTA (AI) 유도 분자 설계 도구의 성능을 평가한다.
    b. Popular molecular representation with various generation tools
    c. Data generation tools (ML, DL)
    d. we highlighted the cutting edge (최첨단의) AI tools to utilize these ML models

2. Results and Highlights

2.1 Components of Computational Autonomous Molecular Design Workflow - CAMD Workflow 의 구성 요소

1. Computational autonomous molecular design (CAMD)
    - 아래 그림과 같이 통합되었으며, 닫혀있는 형태의 loop 여야만 한다
    - 4가지의 구성요소를 만족시켜야 한다.



1. Efficient data generation and extraction tools
2. Robust data representation techniques
3. Physics-informed predictive machine learning models
4. Tools to generate new molecules using the knowledge learned from steps i–iii

2. For data generation in CAMD
    a. 주로 사용되는 것은 DFT
        i. High-throughput **density functional theory (DFT)** is a common choice mainly because of its reasonable accuracy and efficiency
        ii. Typically **feed in 3D structures** to predict the properties of interest (관심 가는 것의 특징을 예측하기 위해 3D 구조를 집어 넣는다)
        iii. **Extract the more relevant structural** and **properties data**, which are then either used as input to **learn the representation** or **as a target** required for the ML models
    b. Data generation 은 2가지 다른 방법으로 사용 된다.
        i. 새로운 분자의 특성 예측
        ii. Inverse design 을 통한, 연구자가 원하는 특징을 가지고 있는, 새로운 분자를 생성
3. Database 와 같은 부수적인 요소와 엮일 수 있음
4. CAMD 는 molecular design 을 위한 자동화 workflow 의 발전의 첫 번째 발걸음이다
5. Such an automated pipeline will **not only accelerate the hit identification** and **lead optimization** for the **desired therapeutic candidates** but can actively be used for machine reasoning to develop transparent and interpretable ML models. (해석 가능한 ML 모델을 개발하기 위해 사용이 가능)
6. Inverse Design 으로 부터의 data generation은 아래의 두 가지 방법으로 검증한다
    a. 우리가 원하는 특성에 잘 맞는 DFT 방법을 사용하거나
    b. closed-loop system 에서 affinity (분자끼리 붙는 강도) 를 확인하기 위해, 표적 단백질과의 대량의 도킹에 의해

    이에 따라 나머지 CAMD 를 업데이트 하도록 한다

7. 이 스텝은 closed loop 에서 반복되는데, data representation, 예측 특성, 새로운 데이터 생성을 improving 하고 최적화 한다.
8. 일단 제대로 훈련되고 나면, DFT 를 이용한 검증 단계가 ML 모델로 대체되며, 컴퓨터를 이용한 방법이 좀 더 효과적이게 된다.

==========

CAMD 의 주요 요소를 뜯어 볼 시간이다

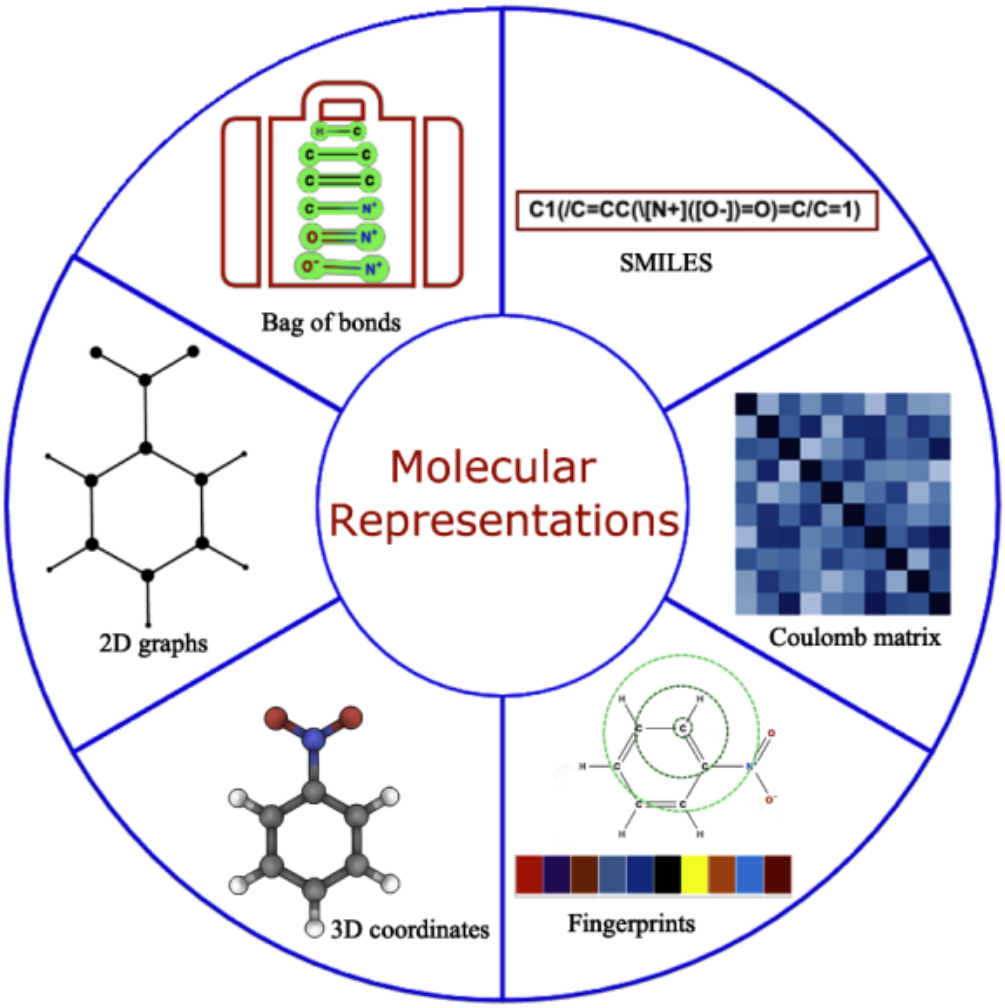
2.2 Data Generation and Molecular Representation

1. ML models are data-centric—the more data, the better the model performance 이.지.만, (특히) 윤리적인 문제로 인해, 생리적 생물학적 데이터는 매우 부족하다.
2. 일부 하위 도메인의 경우에는, 주로 데이터베이스의 물리학 기반 시뮬레이션 또는 NIST 와 같은 실험 데이터베이스에서 나오는 제한된 양의 데이터가 존재한다.
3. 생화학적으로는 free energy of reaction database 가 존재하지만, empirical method 로 인해 얻어진 것이기 때문에 ML 의 ground truth 로 사용하기엔 무리가 있다
4. 즉, 진짜 문제는 → **For many domains, accurate and curated data does not exist.**
5. 이러한 시나리오에서, 출판된 과학 문헌 (literature)과 ML에 대한 특허? (patents)로부터 **데이터를 생성 (Creating data)** 하는 다소 비정상적이지만, 매우 효과적인 접근법이 최근 채택되었다.
6. **NLP 를 바탕으로 하고 있음**
7. 즉, NLP 를 기반으로 데이터를 학습하고 그것을 이용해 대용량 데이터의 연구 디자인의 시간을 상당히 줄여준다는 뜻 인 듯 하다

2.3 Molecular Representation in Automated Pipelines

1. Robust representation of molecules is required for accurate functioning of the ML
2. An ideal molecular representation should be **unique, invariant**
   with respect to
   1. different symmetry operations (rotation, reflection...), 2. invertible, 3. efficient to obtain,
   4. capture the physics, 5. stereo chemistry (입체 구조), 6. structural motif.

3. 위의 조건을 충족시키기 위해 아래 그림과 같이 다양한 방법이 제안 되었음



4. 그.러.나, Still a challenging task, and any gold standard method is yet.....
5. There are **two broad categories** in the literature to represent the Molecules
    a. 시뮬레이션 및 실험으로 부터 얻은 Properties 를 포함하도록 유지한 채 도메인별 지식을 사용하는 experts가 설계한 1D 및/또는 2D 표현
    ==========
        i. includes properties of the molecules
        ii. molecular fingerprints
        iii. structured text sequences (SMILES, InCHI)
            1. The SMILES representation of molecules is the main workhorse as a **starting point** for both representation learning as well as for generating expert-engineered molecular descriptors.

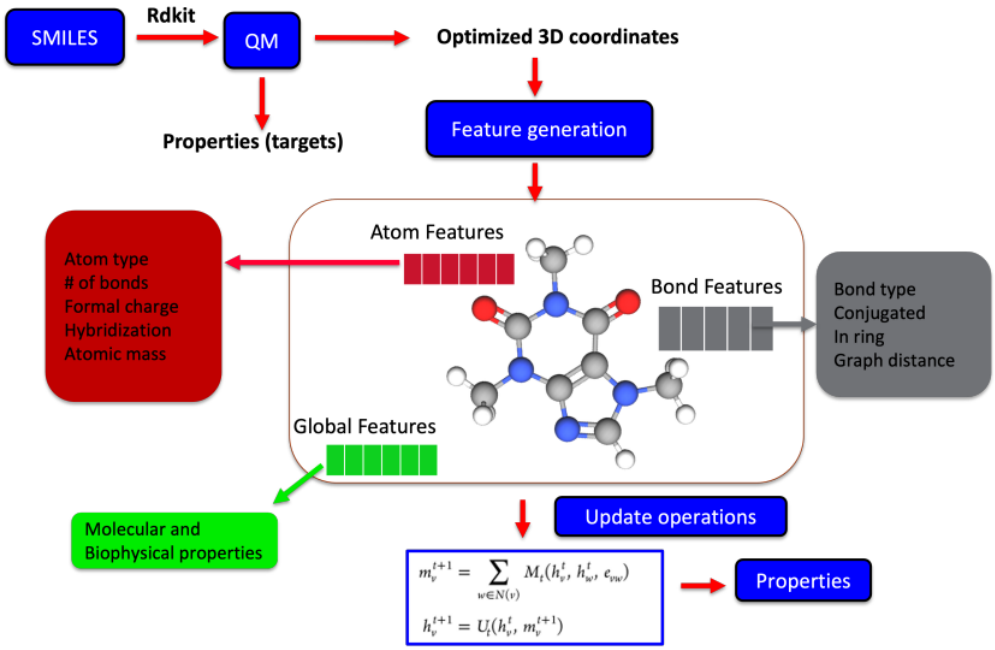
SMIES 표현은 representation learning 과 전문가에 의해 공정된 molecular discriptor 둘 다를 생성하기 위한 출발점으로서 주요 작업물이다.

2. 후자의 경우 SMIES 문자열은 지문을 계산하거나 하나의 핫 인코딩 벡터로 직접 사용될 수 있으므로 더 빠른 속도와 다양한 적절한 특성을 제공함으로써 양자 화학/실험에서 feature generation 을 유도할 수 있다.
3. 또한 SMIES는 분자가 2D 그래프로 쉽게 변환할 수 있다.
4. 주로 SMIES 문자열을 사용하여 분자 생성 모델링에서 상당한 진전이 있었지만, they often lead to the generation of syntactically **invalid molecules** and are synthetically **unexplored**.
5. SMIES의 확장은 높은 의미론적 및 구문론적 타당성을 가진 보다 구체적인 표현을 찾기 위해 분자의 rings & branches 를 더욱 robust 하게 인코딩 하는 방법으로 시도되었는데, SMILES, InChI, SMARTS, DeepSMILES, DESMILES, etc. 등이 있으며
6. 예측 및 생성 모델링에 점점 더 사용되고 있는 SELFIES 로 알려진 100% syntactically correct 한 분자의 정확하고 강력한 문자열 기반 표현을 제안했다.

b. ML 프레임워크 내의 3D nuclear coordinates **(이거 잘 이해가 안 가네요)**/properties 로부터 직접 분자 표현을 반복적으로 학습 시키는 것

==========

6. 최근에는 아래 그림과 같이 다양한 범위의 properties 를 위한 화학적인 정확도를 높이기 위한, 분자 예측 구조를 모델링 하기 위한, 좀 더 직접적으로 분자 자체를 학습하는 방법이 이루어지고 있다. 이 방법은 더욱 robust 하고 outperform 하다!



7. 우리가 예측하듯이 다양한 방식의 표현법을 가지고 있는 GNN 이 일반적인 선택법이다.
    a. Starts with generating the atom (node) and bond (edge) features for all the atoms and bonds within a molecule,
    b. Iteratively updated using graph traversal algorithms
    c. Taking into account the **chemical environment information** to learn a robust molecular representation
    d. The starting atom and bond features of the molecule may just be **one hot encoded vector** to only **include atom-type, bond-type, or a list of properties of the atom and bonds** derived from SMILES strings.
8. 또한, 분자는 3D 3D multiconformational entities (???) 이므로 물리학-기반의 분자 시뮬레이션의 경우와 마찬가지로 nuclear coordinate로 잘 표현될 수 있다고 가정하는 것이 당연하다.
   → (하지만) However, non-invariant, non-invertible, and non-unique in nature
   → and hence not commonly used in conventional machine learning
9. 위의 단점과 계속 이어서, 좌표 자체는 bond types, symmetry, spin states, charge 와 같은 분자의 주요 속성에 대한 정보를 가지고 있지 않다.
   → 이것을 극복하기 위해 atom-centered Gaussian functions, tensor field networks 를 사용하고, 그림 3과 같이 표현 학습 기법을 사용하였다.
   → (그림 3에서) RDkit - 오픈소스 플랫폼, empirical properties 와 같은 것을 계산 가능하게 해줌 → Quantum Mechanical (QM)
   the combined quantum mechanical (QM) and Molecular Mechanical (MM)

approach (QM/MM) is a popular method to study reactions in biochemical macromolecules.

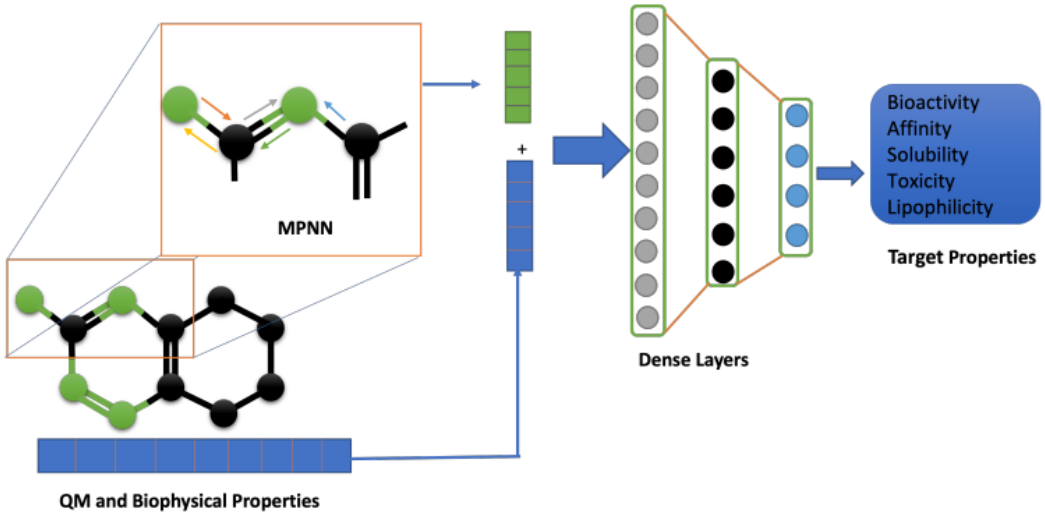
2.4 Physics-Informed Machine Learning

1. Physics-informed machine learning (PIML) is the most widely studied area of applied mathematics in molecular modeling, drug discovery, and medicine
2. ML 아키텍처가 input feature로써 pre-defined input representation 을 요구하는지 혹은 자체 적으로 input representation 을 학습하는지에 따라, 두 가지 하위 범주로 광범위 하게 분류가 가능
    a. 전자는 다른 참고 문헌을 잘 읽어보시고
    b. 후자에 집중 하겠다

    ==========

후자를 살펴보자 (자체 적으로 input representation 을 학습)

3. feature/ property 를 예측하는 학습을 위한 많은 관련된 방식이 **최근 그래프 기반 모델, GNN 이라는 포괄적 용어로 제안되어** 다양한 양자 화학 벤치마크 데이터 셋에서 광범위 하게 테스트 되어왔다.
4. GNN은 두 가지 단계로 구성된다
    a. representation learning
    b. property prediction
5. 어떠한 분자의 의미있는 representation 을 배우고 동시에 어떻게 학습된 feature 를 이용할 것인지에 대해 예측하며 이는, 정확한 property prediction 을 목적으로 한다.
6. Feature -learning 단계에서, robust 한 화학적인 인코딩을 위해 layer 의 sequence 를 통과하여 nuclear coordinate 혹은 graph input 로 부터 얻은 원자와 bond 결합 정보가 업데이트 되며, property prediction 에 사용 된다.
7. 학습된 특징들은 아래 그림과 같이 차원 감소 기법을 사용하여 처리될 수 있다. (MPNN - Message Passing Neural Network)


MPNN
+
Dense Layers
Bioactivity
Affinity
Solubility
Toxicity
Lipophilicity
Target Properties
QM and Biophysical Properties