Deep-learning Prediction Based Molecular Structure Virtual Screening

LG화학 기술 연구원, Yerin Jeon, Kyu-Hwang Lee, Hokyung Lee

abstract

- 분자 구조와 물성 정보가 축척된 데이터베이스를 기반으로 구조와 물성 간의 관계식을 찾는 딥러닝 모형을 구축한 후에, 새로운 분자 구조에 대한 물성 예측 값을 얻을 수 있다.

introduction

- 1950년대 이후에 계산 능력의 발전에 따라서 분자 내부에 전자가 들어 있는 모양과 그 에너지를 양자 역학으로 계산하는 범밀도함수(`DFT`, Density Functional Theory)나 원자와 분자의 물리적인 움직임을 해석하는 분자 동력학(`MD`, molecular dynamics)을 통해 분자 구조의 특정 값을 추출하는 방법이 발전되어 왔다.

- 특히 2000년대 부터는 축척된 데이터 기반의 물질 특성 연구가 활발하게 진행되고 있고 이러한 분야를 `물질 정보학`이라고 부른다.
    - 새로운 재료의 개발과 생산에 소요되는 시간을 획기적으로 단축하는 것을 목표로 하며, 여러 재료의 특성 데이터를 관리 및 분석하여 신재료 개발 뿐만 아니라 공정 모델링, 제품의 수명 주기 관리까지 다양하게 활용이 가능하다.
- 물질 정보학 연구에서는 그 과정을 8단계로 설명하고 있다.
    - 먼저 개별 구조 pool을 바탕으로, 모델링과 계산을 통해서 추출한 특성 값을 저장한 데이터베이스를 구축
    - 구축한 데이터 베이스는 머신러닝 또는 딥러닝 등의 다양한 통계적 기법을 통해서 물성 예측 모형을 세우는 데 사용된다. 예측 모형을 바탕으로 개별 물질에 대한 물성을 예측한다.
    - 예측된 결과는 실험을 통해 검증되며, 실제 물성 평가 결과와 예측 결과가 유사하도록 지속적으로 모형을 업데이트한다.

    - 이렇게 하여 신뢰도 확보 후, 최종적으로 예측 물성을 기반으로 물성이 우수할 것으로 예상되는 구조를 선정하여 일부에 대해 실험 및 평가를 수행한다.

- 재료 ai 관련 논문들의 출판이 증가 중
    - 머신러닝 / 딥러닝을 활용한 재료 연구에서 분자 구조로부터 추출된 특성 값과 실험 조건 등을 가지고 에너지 레벨, 밴드 갭 등의 양자 특성을 예측한다.
        - OLED 재료의 TADF(Themally Activated Delayed Fluorescence) 상수를 예측하여 효율이 좋은 구조를 스크리닝
        - SOC(Spin-Orbital Coupled) 상수를 예측하여 수명이 긴 구조를 가상으로 평가하는 연구
    - RNN을 활용하여 분자 구조를 문자로 변환한 SMILES(Simplified Molecular Input Line Entry System)을 벡터화한 후 물성을 예측하는 방법
    - LSTM 기반으로 원자 간의 거리 정보를 반영하여 분자 구조의 특성을 추출한 후 물성을 예측하는 방법

- 모두 `딥러닝 알고리즘을 적용하여 분자 구조를 수치로 표현하고 물성과의 관계를 탐색`하는 연구

2. 인공지능 방법론을 이용한 신물질 후보의 탐색

2-1. 물성 예측 과정

- 인공지능을 이용하여 신물질 후보를 탐색하는 과정은 `QSAR`(Quantitative Structure Activity Relationship)과 유사하다.
    - QSAR : 물질의 화학 구조로부터 물성 또는 독성을 예측
- `분자 구조가 비슷하면 물성이 비슷할 것`이다.

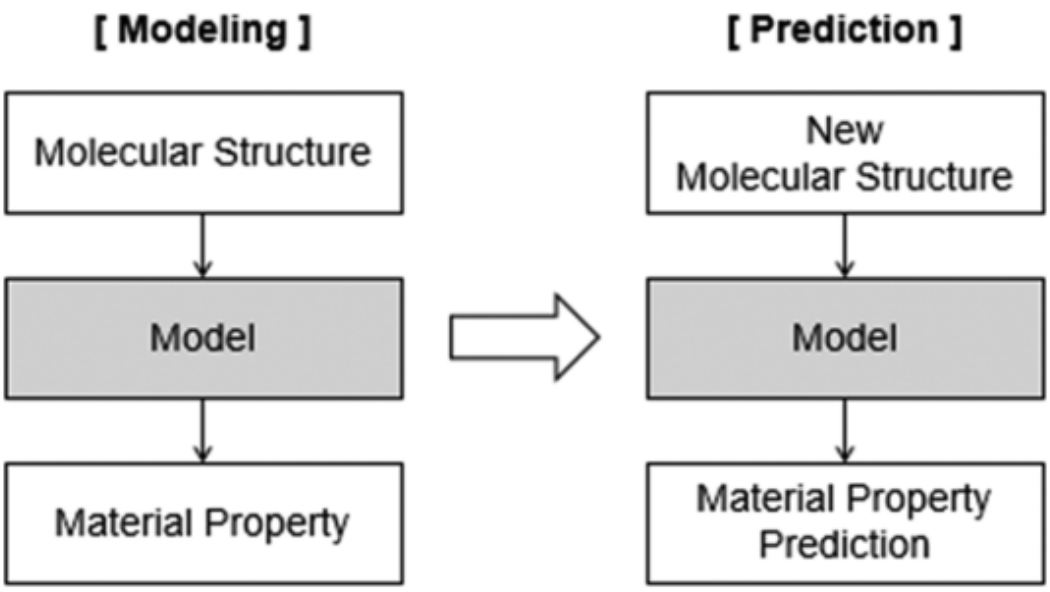


Fig. 1. Material property prediction model and methodology.

- QSAR에서 분자 구조를 설명인자로 사용하는 방법
  1. 문자열로 표현하는 방법, 분자 구조를 1차원 문자열로 표시하여 사용하는 방법 (`SMILES`)
  2. 분자 구조를 그림 자체로 해석하는 경우.
  3. 숫자 데이터인 분자의 양자 역학적 계산 결과인, 2차원 또는 3차원 특성 값을 나타내는 방법.(`DFT`, `MD` 활용)

- 개발하고자 하는 대상이 자주 바뀌게 되는 현실에서는, 적은 실험 데이터를 가지고 분자 구조로부터 실험 특성을 예측하기 위해서는 의미가 있는 분자 구조의 특성 값을 사전에 추출하여 사용해야 한다.
- 논문에서는 3번 째 방법인 분자 구조로부터 양자 역학적인 계산 결과를 기반으로 추출된 구조 및 양자 특성 값을 설명 인자로 사용하여 실험 특성을 예측하는 모형을 구성하였다.
  - 구조 특성 값 : 원자 및 Bond의 개수, Density, Charge, Weight, Surface area, Solubility
  - 양자 특성 값 : Band Gap, Energy Level, Triplet

- 물질을 빠른 시간에 선별하기 위해서 `분자 구조에서 빠른 시간에 계산이 가능한 구조 특성값을 가지고` 우수한 실험 특성을 보일 것으로 예상되는 후보를 1차 선별
- `긴 시간이 필요한 양자 특성을 계산 및 실험 특성을 검토`하여 2차 선별
- 데이터베이스는 개별 구조에 대한 정보와 구조를 설명할 수 있는 특성 값, 그리고 실제 실험에서 측청된 물성 값을 포함한다.
- SMILES 활용

2-2 모형화 방법

- **QSAR과 동일한 아이디어를 차용하되 구조와 물성 간 관계를 모형화하는 방법**으로 딥러닝을 활용한다.
- 구조 및 양자 특성 값으로 실험 특성 값을 바로 예측하는 것이 가능하지만, 정확도의 향상을 위해 구간 추정 후에, 절댓값을 추정하는 2단계로 예측 알고리즘을 사용하였다.
- 구간 예측을 한 결과를 추가 인자로 사용하여 절댓값을 예측한다
  - 먼저 예측하고자 하는 물성의 분포를 등구간으로 나누어 범주를 구성한다.
  - EX) 물성이 0 ~ 1 사이의 값을 가질 때 5 구간으로 나눈다면, `[0, 0.2]`, `[0.2, 0.4]`, `[0.4, 0.6]`, `[0.6, 0.8]`, `[0.8, 1]`로 세분화된다. 0.3인 구조는 두 번째 구간에 속하므로 label 정도는 `(0, 1, 0, 0, 0)`이 된다.
  - 기존 실험 예측 56% → 논문 실험 예측 성능 78%

2-3. 신규 분자의 선별

- 예측 모형이 확보되면, 신규 후보 물질 탐색을 위한 가상 구조를 생성하여 예측한다.
- 신규 분자 구조는 **OLED 분자를 구성하는 세부 그룹으로 나누어**서 각 그룹마다 가능한 분자 구조 집합을 사전에 정의하고, 각 그룹에서 선택된 분자를 결합하여 생성한다. 이를 위한 구조에 대한 데이터베이스 또한 구축해 놓았다.

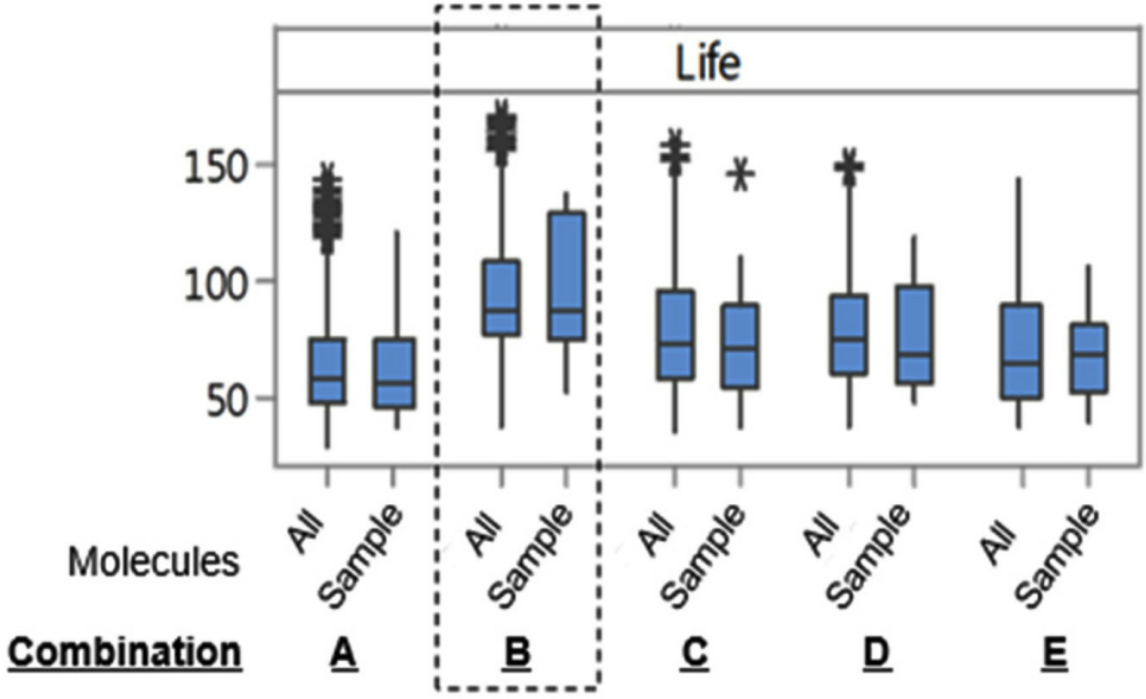
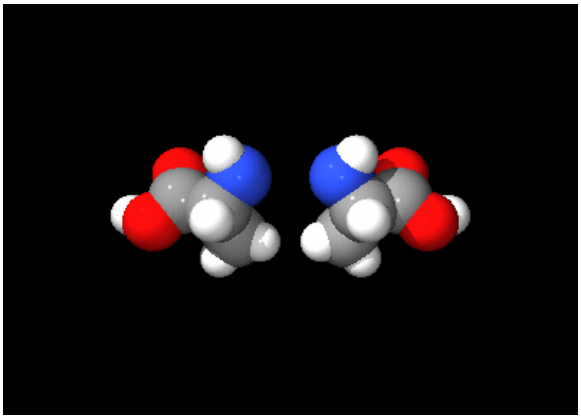


Fig. 4. Selection of molecular structure through random sampling.

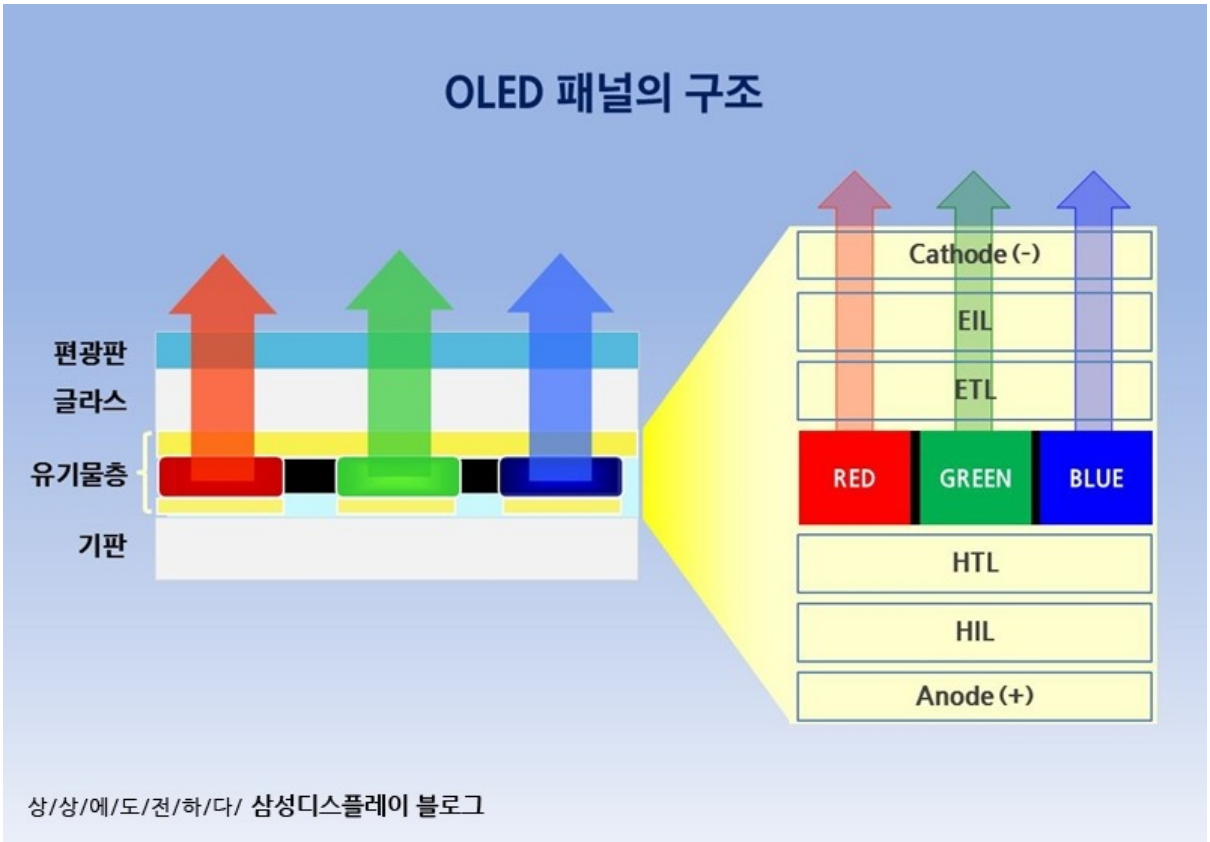
- 이성질체 : 화학의 용어로 같은 원소 배합으로 이루어지며 그 배치에 따라서 성질이 달라지는 물질을 말한다. 더 간단히 말해 분자식은 같은데 다른 물질이다.



- 각 그룹에서 선택된 분자가 결합하여 생성될 수 있는 분자의 개수도 이성질체가 생기는 관계로 그 숫자가 증가하게 되어 계산 시간이 상당히 증가하게 된다.
- 가능한 모든 이성질체를 계산하는 것이 아닌, **그룹별로 선택된 분자로 만들 수 있는 소수의 분자들로 선택된 분자의 우수성을 판단하고, 우수한 조합에 대하여 가능한 모든 이성질체를 탐색**한다.

3. 적용 사례

- OLED는 전기를 주면 스스로 빛을 내는 자체 발광형 유기 물질로 구성이 되어 있으며, 전류를 가하면 봘광층에서 음극과 양극을 통하여 전달된 전자와 전공이 만나서 빛을 내는 구조로 되어 있다.
- OLED의 구조는 HIL(정공 주입), HTL(정공 이동), EML(발광), ETL(전자이동), EIL(전자 주입)층으로 구성이 되어있다.



- 연구에서는 다른 층은 정해져 있다고 가정하고, ETL 층의 물질 변화에 대한 OLED 성능인 전압, 수명, 효율을 대상으로 우수한 후보 물질을 발굴하고자 하였다.
    - 전압이 낮으면서도, 효율과 수명이 높은 재료를 찾는 것을 목표
    - 효율과 수명은 반비례 관계, 전압은 기존 수준을 유지하면서 효율이 기존 대비 10% 이상 높고 수명이 기존 수준의 절반 이하로 떨어지지 않는 재료를 발굴하고자 하였다.

3-1. OLED 재료용 데이터베이스 구축

- 분석에 필요한 데이터베이스 확보를 위하여 과거 축적된 실험 데이터 중 구조 정보와 실험 특성을 수집하였다.

- 구조 및 양자 특성은 직접 알고리즘을 통해 계산
- 후보 탐색 POOL이 될 가상 물질 데이터베이스는 기존의 실험 구조를 바탕으로 확장하는 방식.
- **OLED 구조 전체를 부분으로 나누어 각 세부 그룹 별로 가능한 분자 구조(Building Block)를 정의**하고, 이들을 조합하여 가상의 신규 구조를 형성하였다.
- 결합 위치에 따라서도 최종 물성이 달라지므로, `신규 후보 물질을 조합할 때는 분자 구조의 종류와 결합 위치를 동시에 고려`하였다.
- OLED 후보 물질을 4개의 세부 그룹으로 나누고, 각 세부 그룹마다 15 ~ 20여 개의 분자 구조를 사용하였으며, 이를 조합하여 약 5800만 개의 신규 구조를 생성하였다.

3-2. OLED 물성 예측 및 스크리닝

- 실험 특성 예측을 위해서는 1, 2차 스크리닝 용 모형을 각각 구축하였다.
  - 1차 스크리닝 모형에서는 구조 특성과 일부 주요 양자 특성에 대한 예측 값을 고려하였다.
  - 2차 스크리닝 모형에서는 구조 특성과 양자 특성을 모두 설명 인자로 고려하였다.

- 양자 특성 중 `에너지 레벨과 관련된 인자들이 OLED 물성과 직접적인 상관성`을 가진다.
- 1차 스크리닝 과정에서 에너지 레벨에 대한 양자 계산 값이 없더라도 예측값으로 대체하여 활용하고자 하였다.

Table 1. R-square of energy level predicted from molecular structure

	Training Data		Test Data	
	Orbital Gap	LUMO	Orbital Gap	LUMO
Prediction R-square	77%	82%	60%	64%

- 물성 예측이 필요한 5800만 개의 신규 구조에 대해 1차 스크리닝을 먼저 거친 후, 1차로 예측된 물성치가 특정 조건을 만족하는 수백 개의 구조에 대해서만 양자 계산을 수행 후 2차 스크리닝을 진행하도록 하였으며, 이로 인해 양자 계산에 필요한 시간을 수만 년에서 1개월 이내로 줄일 수 있었다.

- 딥러닝을 활용하여 물성 예측 모형을 구축하는 과정이 필요하다.

- 모형 구축을 위한 라이브러리에는 실험 데이터가 150여개 축적된 상황이며, 모형 적합에 활용할 인자는 1차 스크리닝 모형은 구조 특성과 양자 특성 예측값 100여개, 2차 스크리닝 모형은 구조 특성과 양자 특성 계산값 150여개이다.
- 이들 이자가 Input, Output으로는 전압, 효율, 수명을 개별로 예측하는 모델을 각각 구축하였다.

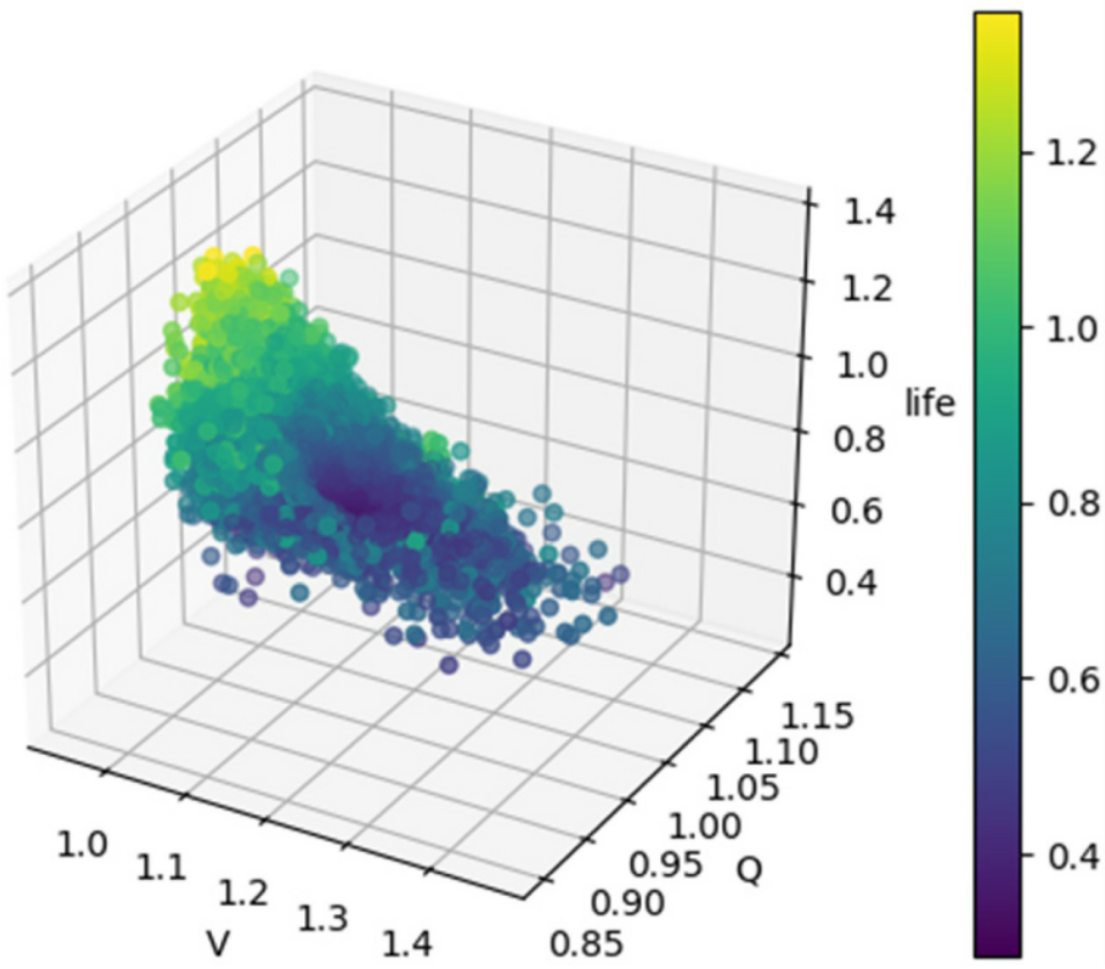


Fig. 5. Life Prediction Map with Voltage (V)-Efficiency (Q).

Table 2. Comparison of actual and predicted value for properties

ID	Voltage		Efficiency		Life	
	Real	Predicted	Real	Predicted	Real	Predicted
1	1.01	0.98	1.10	1.11	0.59	0.58
2	0.98	0.97	1.09	1.11	0.67	0.63
3	1.06	1.02	1.07	1.10	0.59	0.58
4	1.12	1.01	1.04	1.10	0.65	0.55
5	0.97	0.98	1.12	1.10	0.55	0.61
6	1.03	0.96	1.10	1.11	0.51	0.57
7	1.01	0.99	1.11	1.12	0.52	0.60