Korean Chem. Eng. Res., 58(2), 230-234 (2020)
https://doi.org/10.9713/kcer.2020.58.2.230
PISSN 0304-128X, EISSN 2233-9558

딥러닝 예측 기반의 OLED 재료 분자구조 가상 스크리닝

전예린 · 이규황† · 이호경

LG화학 기술연구원
34122 대전광역시 유성구 문지로 188
(2020년 1월 29일 접수, 2020년 2월 24일 수정본 접수, 2020년 3월 5일 채택)

Deep-learning Prediction Based Molecular Structure Virtual Screening

Yerin Jeon, Kyu-Hwang Lee† and Hokyung Lee

LG Chem Research Park, 188, Moonji-Ro, Yuseong-Gu, Daejeon, 34122, Korea
(Received 29 January 2020; Received in revised form 24 February 2020; accepted 5 March 2020)

요 약

딥러닝 기법을 활용하여 분자 구조로부터 물성을 예측하는 시스템은 화학, 생물학, 재료 연구에 적용하기 위해 개발되었다. 분자 구조와 물성 정보가 축적된 데이터베이스를 기반으로, 구조와 물성간의 관계식을 찾는 딥러닝 모형을 구축한 후 최종적으로는 새로운 분자 구조에 대한 물성 예측값을 제공할 수 있다. 또한 선정된 분자 구조의 실제 물성값에 대한 실험을 병행하여 지속적인 검증 및 모형 업데이트를 수행하게 된다. 이를 통해 다량의 분자구조로부터 물성이 우수한 분자 구조를 빠른 시간 안에 스크리닝할 수 있으며, 연구의 효율성 및 성공률을 높일 수 있다. 본 논문에서는 딥러닝을 활용한 물성 예측 시스템의 전반적인 구성과 LG화학에서 실제 신규 구조 발굴에 적용된 사례를 중심으로 소개하고자 한다.

Abstract – A system that uses deep-learning techniques to predict properties from molecular structures has been developed to apply to chemical, biological and material studies. Based on the database where molecular structure and property information are accumulated, a deep-learning model looking for the relationship between the structure and the property can eventually provide a property prediction for the new molecular structure. In addition, experiments on the actual properties of the selected molecular structure will be carried out in parallel to carry out continuous verification and model updates. This allows for the screening of high-quality molecular structures from large quantities of molecular structures within a short period of time, and increases the efficiency and success rate of research. In this paper, we would like to introduce the overall composition of the materiality prediction system using deep-learning and the cases applied in the actual excavation of new structures in LG Chem.

Key words: Deep-learning, Prediction, Property, Molecular structure, Virtual screening

1. 서 론

새로운 물질이나 재료의 개발에 대한 연구는 끊임없이 진행되고 있으며, 일반적으로 원하는 물성을 얻기 위해서 분자 구조를 바꿔가며 합성하고 평가하는 과정이 반복되지만 유의미한 결과를 얻어내는 경우는 많지 않다. 이러한 시행착오적 접근 방식에서 벗어나 최근에는 이론적 배경을 바탕으로 한 논리적 접근을 시도하는 경우가 증가하고 있다[1]. 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 통하여 분자 구조의 특성 값을 추출하는 방법이 발전하게 되었으며, 특히 2000년대부터는 축적된 데이터 기반의 물질 특성 연구가 활발하게 진행[2]되고 있고, 이러한 분야를 물질 정보학(Material Informatics)이라고 부르고 있다.

물질 정보학은 새로운 재료의 개발과 생산에 소요되는 시간을 획기적으로 단축하는 것을 목표로 하며, 여러 재료의 특성 데이터를 관리 및 분석하여 신재료 개발뿐만 아니라 공정 모델링, 제품의 수명 주기 관리까지 다양하게 활용이 가능하다. 데이터 기반의 물질 정보학 연구에서는 그 과정을 8단계로 설명하고 있다[3]. 먼저 개별 구조 Pool을 바탕으로, 모델링과 계산을 통해서 추출한 특성값을

†To whom correspondence should be addressed.
E-mail: amine@lgchem.com
‡이 논문은 POSTECH 이인범 교수님의 정년을 기념하여 투고되었습니다.

저장한 데이터베이스를 구축한다. 이렇게 구축한 데이터베이스는 머신러닝 또는 딥러닝 등의 다양한 통계적 기법을 통해서 물성 예측 모형을 세우는 데 사용되며, 예측 모형을 바탕으로 개별 물질에 대한 물성을 예측한다. 예측된 결과는 실험을 통해 검증되며, 실제 물성 평가 결과와 예측 결과가 유사하도록 지속적으로 모형을 업데이트한다. 위 과정의 반복을 통해 예측 모형에 대한 신뢰도가 확보된 후에, 최종적으로는 예측 물성을 기반으로 물성이 우수할 것으로 예상되는 구조를 선정하여 일부에 대해 실험 및 평가를 수행한다. 이런 과정을 통하면, 실험 이전 단계에서의 물질 스크리닝이 가능하며, 연구 비용을 절감 및 효율성을 높이는 효과를 기대할 수 있다.

국내외적으로 수많은 기관 및 학교에서 데이터베이스 및 예측모형 구축과 관련한 연구가 진행 중이며, 2010년 이후 관련 연구 수가 급격히 증가하고 있다. Citrine informatics에서는 소재 정보 관련 학술 논문이 2010년 대비 2016년에 4배 규모로 증가한 수준이라고 분석한 바 있다[4]. 대표적으로는 머신러닝/딥러닝을 활용한 재료 연구에서 분자 구조로부터 추출된 특성값과 실험조건 등을 가지고 에너지레벨, 밴드갭 등의 양자특성을 예측하는 과정에 사용하고 있다. OLED 재료의 TADF (Thermally Activated Delayed Fluorescence) 상수를 예측하여 효율이 좋은 구조를 스크리닝하거나[5], SOC (Spin-Orbital Coupled) 상수를 예측하여 수명이 긴 구조를 가상으로 평가하는 연구가 이에 해당한다[6]. 또한, Recurrent Neural Network을 활용하여 분자 구조를 문자로 변환한 SMILES (Simplified Molecular Input Line Entry System)를 벡터화한 후 물성을 예측하는 방법[7], LSTM (Long Short-Term Memory) 기반으로 원자간의 거리 정보를 반영하여 분자 구조의 특성을 추출한 후 물성을 예측하는 방법[8] 등 다양한 딥러닝 알고리즘을 적용하여 분자 구조를 수치로 표현하고 물성과의 관계를 탐색하는 연구가 활발히 진행되고 있다.

본 연구에서는 이를 확장하여 계산 결과인 양자특성값이 아닌 재료 평가의 직접적인 지표가 되는 실험 물성을 예측하고, 예측된 물성을 바탕으로 재료를 선별하는 방법론 및 시스템을 구축하였다. 여기서는 LG화학에서 데이터 기반의 연구 방식을 적용하여 새로운 물질이나 재료를 발굴하는 데 어떻게 활용되고 있는지 실제 사례를 중심으로 소개하고자 한다.

2. 인공지능 방법론을 이용한 신물질 후보의 탐색

2-1. 물성 예측 과정

인공지능 방법론을 이용하여 신물질 후보를 탐색하는 과정은 이전부터 QSAR (Quantitative Structure Activity Relationship)이라는 이름으로 알려진, 물질의 화학 구조로부터 물성 또는 독성을 예측하는 것과 유사하다[9]. 이는 분자구조가 유사하다면 물성도 유사하다는 전제를 바탕으로 하고 있으며, 분자의 구조로부터 실험 특성을 모형화한 이후에, 신규 분자 구조에 대하여 구축된 모형을 통하여 실험 특성을 예측하는 과정을 거치게 된다.

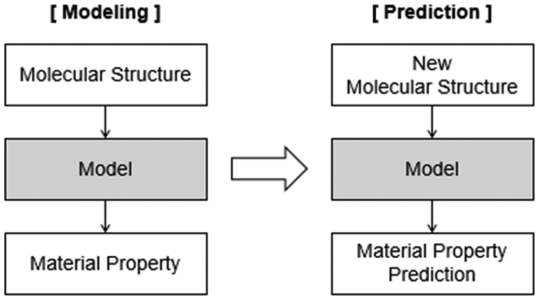

Fig. 1. Material property prediction model and methodology.

QSAR에서 분자 구조를 설명 인자로 사용하는 방법은 데이터의 형태에 따라서 3가지로 구분할 수가 있다. 첫째는 문자열로 표현하는 방법으로, 분자 구조를 1차원 문자열(SMILES, Simplified Molecular Input Line Entry System)로 표시하여 사용하는 방법이 있다. 두 번째는 분자 구조를 그림 자체로 해석하는 경우가 있다. 별도의 프로그램을 통하여 2차원이나 3차원 분자 구조를 만든 다음에 그림의 축을 움직이면서 다양한 그림을 추출하여 사용하는 방법이 있다. 세 번째는 숫자 데이터인 분자의 양자 역학적 계산 결과인, 2차원 또는 3차원 특성값을 나타내는 방법이 있다. 본 방법은 특히 DFT나 MD와 같은 고급 계산화학 방법을 사용하여 계산된 결과로 분자의 특성을 잘 나타내는 것으로 알려져서 널리 활용이 되고 있으며, 이를 사용하여 실험 특성값을 설명할 수 있는 다양한 설명 인자를 발굴하는 연구가 진행되고 있다. 그러나, 1차원이나 2차원의 특성값은 계산에 1초 이내가 소요되지만, 3차원 특성값의 경우에는 12시간씩 소요되는 문제가 있어서, 고려해야 하는 신규 분자 구조가 많은 경우에는 계산 시간의 제약으로 원하는 모든 탐색이 가능하지 못하게 되는 문제가 있다.

방대한 데이터가 있는 경우에는 분자의 화학 구조식에서 실험 특성값을 바로 예측하는 방법을 검토할 수가 있다. 하지만, 개발하고자 하는 대상이 자주 바뀌게 되는 현실에서는 축적된 실험 데이터가 많지는 않기 때문에, 적은 실험 데이터를 가지고 분자 구조로부터 실험 특성을 예측하기 위해서는 의미가 있는 분자 구조의 특성값을 사전에 추출하여 사용해야만 하였다. 그래서, 본 연구에서는 세 번째 방법인 분자 구조로부터 양자 역학적인 계산 결과를 기반으로 추출된 구조 및 양자 특성값을 설명 인자로 사용하여 실험 특성을 예측하는 모형을 구성하였다. 이때 구조 특성값은 원자 및 Bond의 개수, Density, Charge, Weight, Surface Area, Solubility 등에 관한 정보를, 양자 특성값은 Band-gap, Energy Level, Triplet 등에 관한 정보를 포함한다.

여기서 하고자 하는 것은, 가능한 모든 가상의 물질을 대상으로 실험 성능이 우수할 것으로 예측되는 물질을 찾는 것이다. 하지만, 양자 특성값이 실험 특성을 가장 잘 설명하는 설명 인자인 것은 맞지만, 계산하는 데에 시간이 많이 필요하기 때문에 모든 가능한 물질에 대하여 양자 특성값을 계산하는 것은 불가능하다는 문제가 있다. 따라서, 물질을 빠른 시간에 선별하기 위하여 분자 구조에서 빠

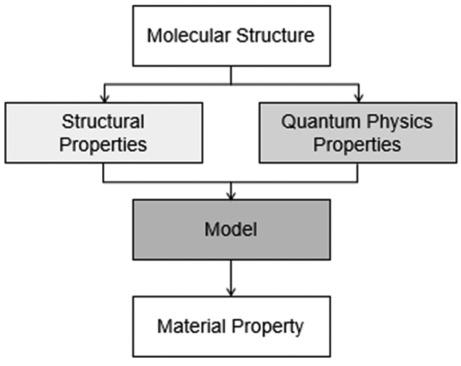

Fig. 2. Prediction model using structural/quantum physics properties.

른 시간에 계산이 가능한 구조 특성값을 가지고 우수한 실험 특성을 보일 것으로 예상되는 후보를 1차 선별하고, 긴 시간이 필요한 양자 특성을 계산 및 실험 특성을 검토하여 2차 선별하는 과정을 사용하였다. 특히, 1차 선별에서는 구조 특성만을 가지고 실험 특성을 모형화하는 것이 아니라, 구조 특성으로 양자 특성을 예측하는 별도의 모형을 구축한 이후에, 구조 특성으로 양자 특성을 예측하고 두 특성을 사용하여 실험 특성을 예측하였다.

구조 특성으로 양자 특성을 예측하는 모형을 구축하기 위해서는 데이터베이스가 필요하며, 이는 실제 실험 결과와 같이 저장되어 있다. 본 데이터베이스는 개별 구조에 대한 정보와 구조를 설명할 수 있는 특성값, 그리고 실제 실험에서 측정된 물성값을 포함한다. 일반적인 경우에 개별 구조 정보는 분자구조 이미지로 존재하지만, 데이터 관리 측면에서 다루기 쉽도록 분자구조를 단순화하여 문자열로 변환하는 SMILES를 활용하였다[10]. 또한 SMILES로부터 구조의 특성 및 양자특성을 설명할 수 있는 설명 인자들을 별도의 DFT 계산을 통해 추출하거나, rdkit 또는 Mordred와 같이 공개된 라이브러리로부터 해당되는 구조에 대한 특성을 수집하여 데이터베이스를 구성하였다[11,12].

2-2. 모형화 방법

본 연구는 QSAR과 동일한 아이디어를 차용하되 구조와 물성 간 관계를 모형화하는 방법으로 딥러닝을 활용하였다. 딥러닝은 Neural Network을 이용하는 방법론이며, 잘 알려진 회귀분석과 기본적인 원리는 동일하다. 과거의 데이터를 학습하고 수학적으로 잘 설명할 수 있는 모형을 만들어 새로운 데이터에 대해 정확하게 예측 또는 분류하는 것을 목표로 하며, 여러 비선형 관계식을 활용하여 다량의 데이터에 내재된 복잡한 관계나 핵심적인 특징을 요약하는 기법이다. 딥러닝 예측 모형을 세우기 위해 본 연구에서는 Google에서 배포한 Python 기반의 Tensorflow 라이브러리를 활용하였다.

본 연구에서는 구조 및 양자 특성값으로 실험 특성값을 바로 예측하는 것이 가능하지만, 정확도의 향상을 위하여 구간 추정을 한 이후에, 절대값을 추정하는 2단계로 예측 알고리즘을 사용하고 있다. 일반적으로 절대값 자체를 예측하는 것보다 구간의 범주를 예측하는 것의 정확도가 높기 때문에, 구간 예측을 한 결과를 추가 인자로 사용하여 절대값을 예측하는 방법을 사용하였다. 먼저 예측하고자 하는 물성의 분포를 등구간으로 나누어 범주를 구성한다. 예를 들면, 물성이 0~1 사이의 값을 가질 때 5구간으로 나눈다면 [0,0.2], [0.2,0.4], [0.4,0.6], [0.6,0.8], [0.8,1]로 세분화될 것이며, 이중 물성이 0.3인 구조는 두번째 구간에 속하므로 Label 정보는 (0,1,0,0,0)이 된다. 같은 방식으로 모든 개별 구조에 대한 구간 정보를 부여한 다음, 구조로부터 추출한 특성값을 활용하여 각 구조가 몇 번째 구간에 속할 것인지 예측하는 딥러닝 분류모형을 구축하였다. 필요에 따라서는, 구간 예측에서 예측된 구간 자체의 값을 인자로 사용할 수도 있으며, 범주별로 예측되는 확률(예를 들어, 직전 예시는 (0.1,0.6,0.2,0.1,0))을 인자로 사용할 수도 있다. 그 다음으로 예측된 구간 정보와 특성값을 함께 활용하여 최종 실험 물성을 예측하는 딥러닝 모형을 구축하였으며, 3개 물성의 검증 Data에 대한 예측 정확도 평균을 기준으로 비교했을 때 특성값으로부터 실험 물성을 바로 예측하는 것보다 약 56%에서 약 78%로 정확도가 향상됨을 확인하였다.

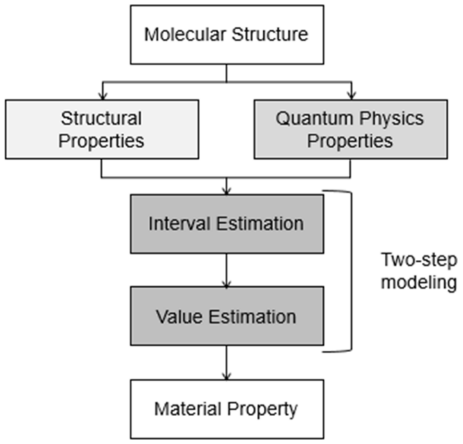

Fig. 3. Two-step modeling.

2-3. 신규 분자의 선별

예측 모형이 확보되면, 신규 후보 물질 탐색을 위한 가상 구조를 생성하여 예측하는 과정을 진행하게 된다. 신규 분자 구조는 OLED 분자를 구성하는 세부 그룹으로 나누어서 각 그룹마다 가능한 분자 구조 집합을 사전에 정의하고, 각 그룹에서 선택된 분자를 결합하여 생성하는 것으로 하였다. 현재는 가능한 분자 구조의 개수는 각 그룹에 각각 수십에서 수백 개를 고려하고 있어서 수백만 개가 되며, 조합이 가능한 이성질체를 고려한 그 수는 더 많이 증가하게 된다. 따라서, 이 수많은 가상의 구조에 대하여 매번 특성값을 추출할 필요가 없도록 하기 위하여, 앞서 실험 데이터를 포함한 데이터베이스와는 별도로, 가상의 구조를 대상으로 하는 데이터베이스를 운영하고 있다. 이는 앞서의 데이터베이스와 동일하게 구조 정보와 특성값을 포함하나, 실제 물성 값은 없어도 무방하다. 이렇게 구성된 데이터베이스에서 신규 구조의 정보로 기 구축된 실험 특성 예측 모형을 사용하여 우수한 성능을 보일 것으로 예상되는 후보 물질을 선별하는 과정을 거치게 된다.

신규 분자 구조를 생성하는 과정에서 각 그룹에서 선택된 분자가 결합하여 생성될 수 있는 분자의 개수도 이성질체가 생기는 관계로 그 숫자가 증가하게 되어, 계산 시간이 상당히 늘어나게 된다. 따라서, 여기서는 각 그룹별로 우수한 분자를 우선 선별하고자 하여, 가능한 모든 이성질체를 계산하는 것이 아니라, 그룹별로 선택된 분자로 만들 수가 있는 소수의 분자들로 선택된 분자의 우수성을 판단하고, 우수한 조합에 대하여 가능한 모든 이성질체를 탐색하는 과정을 실행하고 있다. 그림에서는 랜덤 샘플링을 통해 임의로 선

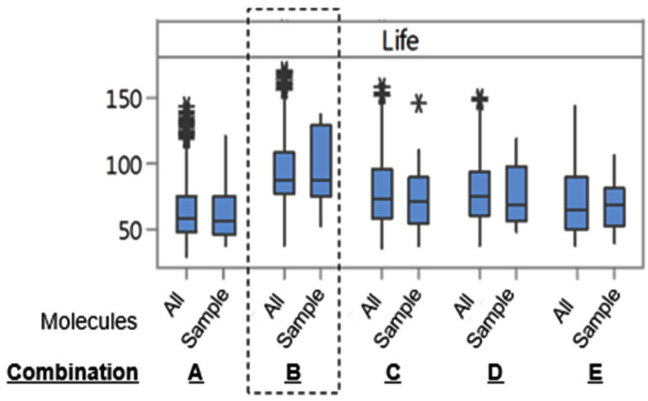

Fig. 4. Selection of molecular structure through random sampling.

택된 구조와, 가능한 전체 조합을 사용한 경우에 Life의 평균적인 변화가 유사함을 확인할 수 있다.

연구 초기에는 연구자가 우선적으로 추천한 분자 Pool에서 분자를 선택하여 조합하는 과정을 통해, 50만개의 가상의 분자에 대하여 우선적으로 성능을 예측하여 보았으나, 원하는 수준을 달성할 수가 없었다. 따라서, 가능한 조합에서 랜덤 샘플링을 통하여 우수한 조합을 탐색한 이후에, 상세 탐색을 하는 과정을 거쳐서 신규 후보를 발굴하고 있다.

3. 적용 사례

본 연구에서 제안된 인공지능 방법론을 이용하여 OLED(Organic Light-Emitting Diode) 신물질 후보 발굴에 적용을 하였다. OLED는 전기를 주면 스스로 빛을 내는 자체 발광형 유기 물질로 구성이 되어 있으며, 전류를 가하면 발광층에서 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 되어 있는데, 본 연구에서는 다른 층은 정해져 있다고 가정하고, ETL 층의 물질 변화에 대한 OLED의 성능인 전압, 수명, 효율을 대상으로 우수한 후보 물질을 발굴하고자 하였다. 특히, 전압이 낮으면서도 효율과 수명이 높은 재료를 찾는 것을 목표로 하였다. 이때 효율과 수명 간 반비례 관계가 존재함을 고려하여, 구체적으로는 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 수집하였고 구조 및 양자 특성은 내부 알고리즘을 통하여 직접 계산하였다. 추가로 후보 탐색 Pool이 될 가상 물질 데이터베이스는 기존의 실험 구조를 바탕으로 확장하는 방식으로 구축하였다. OLED 구조 전체를 부분으로 나누어 각 세부 그룹별로 가능한 분자 구조(Building Block)을 정의하고, 이들을 조합하여 가상의 신규 구조를 형성하였다. 각 세부 그룹별 분자 구조는 결합 위치에 따라서도 최종 물성이 달라지므로, 신규 후보 물질을 조합할 때에는 분자 구조의 종류와 결합 위치를 동시에 고려하였다. 본 연구에서는 OLED 후보 물질을 4개의 세부 그룹으로 나누고, 각 세부 그룹마다 15~20여개의 분자 구조를 사용하였으며, 이를 조합하여 약 5800만개의 신규 구조를 생성하였다.

3-2. OLED 물성 예측 및 스크리닝

실험 특성 예측을 위해서는 1, 2차 스크리닝용 예측 모형을 각각 구축하였으며, 1차 스크리닝 모형에서는 구조 특성과 일부 주요 양자 특성에 대한 예측값을, 2차 스크리닝 모형에서는 구조 특성과 양자 특성을 모두 설명 인자로 고려하였다. 양자 특성 중 에너지레벨과 관련된 인자들이 OLED물성과 직접적인 상관성을 가진다는 것이 알려져 있기 때문에, 1차 스크리닝 과정에서 에너지 레벨에 대한 양자 계산값이 없더라도 예측값으로 대체하여 활용하고자 하였다. 과거 실험을 통하여 축적된 데이터베이스에는 약 1만 5천여 개의 물질에 대한 구조 및 양자 특성 계산 결과가 있으며, 이를 사용하여 100여개의 구조 특성으로 에너지 레벨을 예측하는 딥러닝

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%

모형을 별도로 구축하여 활용하였고, 이때 예측 정확도는 검증 데이터 기준 60% 수준으로 나타났다. 물성 예측이 필요한 5800만개의 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.

다음은 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 인자가 딥러닝 구조의 Input Node로 들어가게 되며, Output Node는 1개로 전압, 효율, 수명을 개별로 예측하는 모형을 각각 구축하였다. 한편, 인자 수에 비해 관측치 수가 상대적으로 적은 상황에서 모형의 예측력을 확보하기란 쉽지 않은 문제이다. 특히, 과적합이 발생하여 학습된 데이터가 아닌 새로운 재료에 대한 예측 오차가 커질 우려가 있다. 따라서 유사한 데이터를 가상으로 생성하여 데이터베이스를 확장하였고, 딥러닝 모형 구조를 최적화하면서 과적합을 줄이고자 하였다. 첫째로, 가상 데이터 베이스 확장은 기존 데이터를 복제하되 Input Feature에 약간의 Noise를 더해주는 방식인 Noise-Based Augmentation 방법을 활용하였다. 이때 Noise는 특정 변수 j의 편차를 σ_j라고 할 때, $N(0, (0.5\sigma_j)^2)$에서 임의로 생성하여 원래 데이터의 100배 크기로 확장하였다. 둘째로, 딥러닝 구조 최적화는 Hidden Layer, Hidden Node 및 Dropout Rate 등을 활용하였다. Input Feature의 수가 p개라고 할 때, Hidden Layer 수는 (2, 3, 4), Hidden Node 수는 (p/2, p, 2p)개, Dropout Rate은 0.25로 구성된 모든 가능한 조합의 수를 생성한 후, 검증 데이터에 대한 예측 성능이 가장 좋은 구조를 채택하도록 하였다.

모형 구축 및 실험을 통한 검증 과정을 3번 거친 후, 최종 확정된

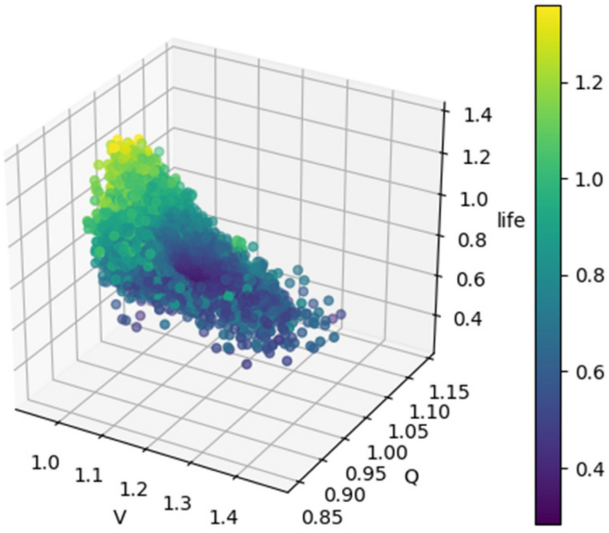

Fig. 5. Life Prediction Map with Voltage (V)-Efficiency (Q).

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

예측모형으로 후보 물질 중 7건을 선별하였다. 실험을 통해 7건에 대한 소자 평가를 완료하였으며, 그 중 1, 2번 물질이 목표하던 물성 조건을 만족함을 확인하였다(Table 2). 물성 예측 모형 구축부터 실험 검증까지는 약 2개월이 소요되었으며, 신재료 발굴 기간 단축과 효율성의 측면에서 충분히 의미 있는 연구라고 판단된다.

4. 결 론

여기서는 딥러닝을 활용하여 분자 구조로부터 물성을 예측하여 신물질 탐색에서의 연구 비용 절감 및 효율성을 높이는 데 기여할 수 있음을 소개하였다. 재료 분야뿐만 아니라 화학, 생명과학 분야에서도 활용 가능하며, 특히 신약 개발 및 타겟 발굴 연구에서 대용량의 유전자 스크리닝, 실험 비용 발생으로 인해 어려움을 겪고 있음을 고려하면 기대효과가 매우 클 것으로 예상된다.

딥러닝 예측 모형의 특성상 물성 예측 과정에서 구조의 어떤 특성이 어떤 방향으로 영향을 미치는지 알 수 없다는 점에서 한계가 있지만, 반대로 생각하면 특성값에 대한 물리화학적 지식이 없더라도 분자구조에 대한 정보만 있으면 물성 예측이 가능하다는 점은 장점으로 작용할 수 있다. 다만, 실험 결과를 반영하여 모형을 수정 및 업데이트하는 과정에서 물리화학적 지식과 노하우를 가진 실험자와의 충분한 의견 교환을 통해 모형에 대한 신뢰도를 확보하는 과정은 분명히 필요하다.

여기서는 OLED 중에서도 ETL 단분자 재료를 타겟으로 하여, 다른 Layer의 재료는 고정하고 ETL 재료만 변화했을 때 최종 OLED 물성이 어떻게 변동할지를 예측하였다. 추가적인 연구를 통해 다른 Layer 재료와의 상호작용을 고려하여 OLED 물성 예측이 가능하다면 훨씬 효율적인 연구를 수행하는 데 도움이 될 것이라 생각한다.

References

1. Youn, Y. and Han, S., "Paradigm Shift in Material Research: from Edisonian to Mechanistic to Data-driven," Physics & High Technology, Sep(2017).
2. Agrawal, A. and Choudhary, A., "Perspective: Materials Informatics and Big Data: Realization of the "fourth paradigm" of Science in Material Science," *APL Materials* **5**, 053208(2016).
3. Takahashi, K. and Tanaka, Y., "Materials Informatics: a Journey Towards Material Design and Synthesis," *Dalton Transactions*, **45**(26), 10497-10499(2016).
4. Citrine informatics, "Material informatics: Artificial intelligence driven materials development and optimization" (2016).
5. Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M. A. and Markopoulos, G., "Design of Efficient Molecular Organic Light-emitting Diodes by a High-throughput Virtual Screening and Experimental Approach," *Nature Materials*, **15**(10), 1120(2016).
6. Kwak, H. S., Giesen, D. J., Hughes, T. F., Goldberg, A., Cao, Y., Gavartin, J. and Halls, M. D., "In Silico Evaluation of Highly Efficient Organic Light-emitting Materials," *Organic Light Emitting Materials and Devices XX*, **9941**, 994119(2016).
7. Goh, G. B., Hodas, N. O., Siegel, C. and Vishnu, A., "Smiles2vec: An Interpretable General-purpose Deep Neural Network for predicting Chemical Properties," arXiv preprint arXiv:1712.02034 (2017).
8. Altae-Tran, H., Ramsundar, B., Pappu, A. S. and Pande, V., "Low Data Drug Discovery with One-shot Learning," *ACS Central Science*, **3**(4), 283-293(2017).
9. Leo, A. and Hoekman, D. H., "Exploring QSAR," American Chemical Society(1995).
10. Weininger, D., "SMILES, a Chemical Language and Information System, 1. Introduction to Methodology and Encoding Rules," *Journal of Chemical Information and Computer Sciences*, **28**(1), 31-36(1988).
11. Landrum, G., "Rdkit", Open-source cheminformatics(2006).
12. Moriwaki, H., Tian, Y. S., Kawashita, N. and Takagi, T., "Mordred: a Molecular Descriptor Calculator," *Journal of Chem-informatics*, **10**(1), 4(2018).