기계 학습 이론을 활용한 재료 연구: 현황과 전망

DOI: 10.3938/PhiT.27.003

김 치 호

Machine Learning in Materials Science: Status and Prospects

Chiho KIM

Dramatic improvements in hardware and algorithmic techniques have advanced computing performance, which has opened 'data-driven' opportunities in materials science. The Materials Genome Initiative (MGI), launched by the US government in 2011, has accelerated materials design by changing the underlying philosophy. Under political and technological circumstances, researchers can now successfully employ the 'rational data-driven design' strategy in their work rather than using the former Edisonian-style approach. This article provides an overview of a recent success story, the Polymer Genome Project, in which the emerging machine learning method was actively utilized. The discussion also points towards some challenge, that might be encountered in the future and the advanced strategies to overcome them.

서 론

우리의 인지 체계는 과거의 '경험'에 비추어 새로 맞는 상황을 '해석'한다. 과거의 풍부한 경험은 새로운 문제를 합리적으로 판단하는데 도움을 주고 경험이 부족하다면 다소 부정확한 결론을 내놓기도 한다. 인지 체계는 성공한 해석과 실패한 해석 모두에서 교훈을 얻고, 다음에 마주할 새로운 상황을 더 합

저자약력

김치호 박사는 한양대학교 신소재공학부 박사(2009)로서 디스플레이공학연구소 박사 후 연구원(2009-2010), 삼성전자 반도체연구소 책임연구원(2010-2014), University of Connecticut 박사 후 연구원 및 Research Associate(2014-2018)를 거쳐 현재 Georgia Institute of Technology 에서 Research Engineer(2018-현재)로 재직 중이다.
(chiho.kim@mse.gatech.edu)

Fig. 1. Learning and decision making process based on past experience, intuition, and common sense.

리적으로 해석하게끔 서서히 발달한다.(그림 1) 축적된 경험에서 찾은 규칙이 우리에게 합리적인 해석을 제공하는 사례들은 역사 속에서 쉽게 찾아볼 수 있다. 케플러는 브라헤가 평생 동안 기록한 천체 관측 자료를 분석하여 행성 운동 법칙을 정리하였다. 이종 금속의 고용체 형성 조건을 기술하는 홈-로슬리의 법칙이나 재료의 결정립 크기가 기계적 강도에 미치는 영향을 정량화한 홀-페치 관계도 경험에 근거한 과학적 발견의 좋은 예이다.^[1-3] 현대에 이르러 통계학과 인공 지능 분야의 전문가들은 경험이라는 데이터를 수학적으로 분석하는 몇 가지 이론들을 정립했다. 그 이론들은 다시 알고리즘으로 구현되어 게임, 기상, 운전, 학습, 쇼핑, 언론, 금융 등 '예측'을 필요로 하는 다양한 분야에서 합리적 추론의 근거를 제공하고 있다.

기계 학습(Machine Learning)은 현상을 분석하거나 추론 모델을 작성하기 위한 대표적인 '데이터 기반(Data-driven)' 접근 방법이다. 따라서 학습의 근거로 사용할 양질의 데이터를 확보하는 것이 매우 중요하다. 2011년부터 미국 정부 주도하에 추진되어 온 소재 게놈 이니셔티브(MGI, Materials Genome Initiative)는 소재 개발의 효율성 고취를 위해서 데이터에 기반한 연구 수행을

REFERENCES

- [1] W. Hume-Rothery, J. Less Common Met. **3**, 264 (1961).
- [2] E. O. Hall, Phys. Soc. B **64**, 747 (1951).
- [3] N. J. Petch, Acta Metall. **34**, 1387 (1986).

강조하고 있다.^[4,5] 이에 따라 OQMD(The Open Quantum Materials Database, www.oqmd.org)와 Materials Project(www.materialsproject.org) 같은 거대한 재료 정보 서비스가 등장하였으며, 이들은 모든 자료를 대가 없이 대중에게 공개하고 있다. 또한 전 세계의 과학자들이 유럽의 NOMAD 연구소(www.nomad-coe.eu)에 자신들이 생성한 소재 시뮬레이션 자료를 제출하여 데이터베이스 구축에 기여하는가 하면, 일본의 PolyInfo(mits.nims.go.jp)는 방대한 고분자 정보를 대중에게 공유하고 있다. 공개된 소재 정보는 연구자들에 의해 기계 학습 알고리즘으로 분석되어, 각 응용 분야의 연구 패러다임을 바꿀만한 값진 결과물로 재창조되고 있다. 이 기고문은 기계 학습 이론과 재료 정보를 활용하여 수행된 고분자 관련 연구 사례를 소개하고, 향후에 직면할 수 있는 과제와 극복 방안에 대해서 논의하고자 한다.

소재 연구와 기계 학습

재료 과학은 물질의 구조와 성분이 물성에 미치는 영향을 파악하여, 더 나은 성능의 새로운 소재를 개발하는데 필요한 실험을 찾는 연구 분야이다. 물질은 실험실에서 합성되거나 가상으로 디자인될 수 있고, 물질 평가는 장비를 이용한 실험과 컴퓨터를 이용한 시뮬레이션 등을 통해 이루어진다. 예를 들어 반도체 소자의 절연막으로 이용할 산화물이 필요한 경우, 소재를 합성하여 밴드갭을 측정하고, 목표치에 비해 그 값이 작다면 조성을 달리해서 합성과 측정을 반복한다. 그림 2에 보인 것과 같이 시행착오(Trial and Error)법에 기반하는 Edisonian 접근 방식은 문제 해결을 위한 실험 모수가 자칫 지나치게 증가할 부담이 있다. 반면 물질 탐색 과정에서 산화물의 조성 and 밴드갭의 상관관계를 체계적으로 파악하고 그 정보를 새로운 실험 디자인에 활용하면 목표 달성을 위한 시간과 비용을 효과적으로 줄일 수 있을 것이다.^[6] 이런 이유로 소재 연구에 기계 학습 방법을 활용하는 이른바 데이터 기반 합리적 재료 설계(Rational Design)의 사례들이 급격히 증가하는 추세다.^[7-11]

반도체 소자 절연막의 예에서 산화물의 조성 and 밴드갭은 알고리즘의 용어로 각각 입력(Input)과 출력(Output)이라고 말할 수 있다. 기계 학습 분야에서는 입력을 디스크립터(Descriptor) 혹은 피쳐(Feature)라고 부르고, 출력은 특성(Property)이라고 부른다. 우리는 기계 학습 방법을 활용할 때 우선 디스크립터와 특성을 결정해야 한다. 특성은 예측하고자 하는 물성으로 처음부터 명백하게 결정되지만, 디스크립터를 정하는 것은 연구자의 주관적인 판단에 크게 의존한다. 만약 밴드갭의 예에서 산화물 조성 외에 시편의 모양, 결합 종류별 농도 등을 디스크립터 구성 요소로 함께 고려했다면 예측 정확도가 더욱 높은 다른 형태의 관계 모델을 세울 수도 있었을 것이다. 반면 오늘 측정된 밴드갭이 어

Fig. 2. Edisonian style approach vs machine learning based rational design of materials. (idea of illustration from Ref. 6)

제의 측정치보다 더 크다고 해서 측정 날짜나 요일을 디스크립터로 이용하면 엉뚱한 모델이 만들어진다. 이처럼 디스크립터를 선정할 때는 반드시 전문적 기술 지식(Domain Expertise)이 활용되어야 한다. 다음 절에서는 기계 학습 방법이 재료 연구에 활용된 예를 구체적인 단계(데이터 수집, 디스크립터 선정, 예측 모델 훈련)를 따라 소개하고, 이어서 예측 모델을 이용한 신 재료 디자인의 성공 사례와 물성 예측 플랫폼 구축 사례를 소개한다.

1. 기계 학습 이론을 활용한 고분자 물성 예측 모델 개발

고분자는 가공이 쉽고 가벼우며 내화학성, 열 및 전기적 절연성

REFERENCES

- [4] Materials Genome Initiative, <https://www.mgi.gov/>.
- [5] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, APL Mater. **1**, 011002 (2013).
- [6] G. Ceder and K. Persson, Scientific American, December, 36 (2013).
- [7] V. Sharma, C. C. Wang, R. G. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, S. A. Boggs and R. Ramprasad, Nat. Commun. **5**, 4845 (2014).
- [8] A. F. Baldwin, T. D. Huan, R. Ma, A. Mannodi-Kanakithodi, M. Tefferi, N. Katz, Y. Cao, R. Ramprasad and G. A. Sotzing, Macromolecules **48**, 2422 (2015).
- [9] A. Mannodi-Kanakithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, Scientific Reports **6**, 20952 (2016).
- [10] C. Kim, G. Pilania and R. Ramprasad, Chemistry of Materials **28**, 1304 (2016).
- [11] A. Mannodi-Kanakithodi, G. Pilania, R. Ramprasad, T. Lookman and J. E. Gubernatis, Computational Materials Science **125**, 92 (2016).

Fig. 3. Comparison of DNA and polymer structures.

이 우수하여 다양한 분야에 널리 응용되는 소재이다. 플라스틱 병과 같은 생활 용품부터 전기 자동차의 전지 속 유전막 같은 특수한 곳까지, 도처에서 고분자 재료를 발견할 수 있다. 고분자는 단위체(Monomer)가 중합을 이루어 형성하는 사슬 형태의 물질로, 사슬을 이루는 화학적 빌딩 블록(Building Block)의 종류와 연결 순서가 전반적인 물성을 결정짓는다. 마치 생물 유전자의 본체인 DNA가 네 가지 단위체(구아닌, 시토신, 아데닌, 티민) 짝의 배열 순서에 따라 다르게 만들어지는 것과 유사하다.(그림 3) 이런 유사성에서 착안하여 이름 지어진 Polymer Genome은 ‘누구나 쉽고 빠르게 원하는 고분자를 디자인’하도록 돕기 위해 개발되었다. 우선 시뮬레이션과 문헌 조사를 통해 기계 학습에 활용할 벤치마크 데이터셋(Benchmark Dataset)을 구축하였고, 이어서 데이터셋에 포함된 모든 고분자들을 지문화(Fingerprinting)하여 기계 학습 알고리즘이 인식할 수 있는 수치 언어, 즉 고분자 지문(Polymer Fingerprint)으로 변환하였다. 벤치마크셋의 고분자 지문과 물성 정보는 기계 학습 알고리즘에 각각 입력과 출력 항으로 적용되어 물성 예측 모델을 구축하는데 활용되었다.

(1) 데이터셋 구성

데이터셋을 확보하는 것은 기계 학습 예측 모델을 구축하기 위한 중요한 사전 단계이다. 예측 모델을 훈련(Train)하기 위한 벤치마크 정보가 바로 데이터셋이기 때문이다. 일반적으로 정보의 규모가 클수록 예측 모델 성능은 향상된다. Polymer Genome은 다양한 고분자 정보를 확보하기 위해서 9가지의 빌딩 블록, 즉 CH₂, CF₂, CHF, CO, O, N, CS, C₆H₄, C₄H₂S를 선정하여 이들 중 4~8개를 조합해 만들 수 있는 고분자들을 데이터셋에 포함하였다.^[7,9,12-14] 예를 들어 CH₂ 빌딩 블록으로만 이루어진 폴리에틸렌(Polyethylene)은 Polymer Genome 데이터셋을 구성하는 가장 단순한 고분자이다. 유전 상수, 밴드갭 등을 제일원리 방법으로 계산하여 데이터셋의 ‘특성’ 항목에 담았다. 다양한 문헌(예, CRC

Handbook^[15])에도 매우 체계적으로 고분자의 물성들이 정리되어 있는데, 여기에는 유리 전이 온도(Glass Transition Temperature), 녹는점, 용해도 파라미터(Solubility Parameter)와 같이 제일원리 방법으로 얻기 어렵지만 합성을 위해서 반드시 알아야 하는 중요한 정보들이 포함된다. Polymer Genome은 문헌에서 얻은 고분자와 해당 물성 정보들도 데이터셋에 포함하여, 현재 약 900 종류의 고분자를 예측 모델 구축용 벤치마크 데이터로 활용한다.

(2) 재료 지문화

기계 학습 모델 구축 과정에서 입력 값으로 쓰일 지문은 각 재료마다 고유해야 한다. 모델 훈련 알고리즘이 지문에 새겨진 수치, 즉 디스크립터 값들을 이용해서 고분자를 서로 구분하기 때문이다. Polymer Genome의 지문 체계는 크게 세 가지로 구성된다.(그림 4) 첫 번째는 Atomic-triples로서, 고분자를 이루는 원자들의 결합 구조를 기술하는 디스크립터이다. 결합으로 이어진 세 개 원자의 원소 기호(A, B, C)와 각 원자에 결합한 다른 원자들의 개수(a, b, c)를 엮어서 Aa-Bb-Cc 형식으로 적는다. 예를 들어 폴리에틸렌에는 세 가지 유효 Atomic-triples(C₄-C₄-H₁, C₄-C₄-C₄, H₁-C₄-H₁)가 부여된다.^[12,16-17] 두 번째 지문 요소는 QSPR(Quantitative Structure Property Relationship)에 근거해서 생성되었다. Atom-triples보다 큰 범위의 기하(Topological) 및 전자(Electronic) 구조 특징을 기술한다. MQNs(Molecular Quantum Numbers),^[18] TPSA(Topological Polar Surface Area), 벤젠이나 타이오피엔(Thiophene) 같은 방향족 고리의 수 등이 이에 해당되며, 고분자 구성 원자들의 국부적 연결 관계만을 기술하는 Atom-Triples의 한계를 보완해준다. QSPR 디스크립터는 공개된 Python 모듈 패키지인 RDKit을 이용해서 생성하였다.^[19] 세 번

REFERENCES

- [12] J. Hill, A. Mannodi-Kanakkithodi, R. Ramprasad and B. Meredig, *In Computational Materials System Design*, D. Shin, J. Saal, Eds. (Springer International Publishing, Cham, 2017).
- [13] T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania and R. Ramprasad, *Scientific Data* **3**, 160012 (2016).
- [14] C. C. Wang, G. Pilania, S. A. Boggs, S. Kumar, C. Breneman and R. Ramprasad, *Polymer* **55**, 979 (2014).
- [15] A. F. M. Barton, *Handbook of Solubility Parameters and Other Cohesion Parameters* (CRC Press, Inc., Florida, USA, 1983).
- [16] T. D. Huan, A. Mannodi-Kanakkithodi and R. Ramprasad, *Phys. Rev. B* **92**, 14106 (2015).
- [17] A. Mannodi-Kanakkithodi, T. D. Huan and R. Ramprasad, *Chem. Mater.* **29**, 9001 (2017).
- [18] K. T. Nguyen, L. C. Blum, R. van Deursen and J. L. Reymond, *Chem. Med. Chem.* **4**, 1803 (2009).
- [19] RDKit, Open source toolkit for cheminformatics. <http://www.rdkit.org/>.

Fig. 4. Hierarchical descriptors introduced for fingerprinting the polymeric materials.

째 그룹은 가장 큰 규모의 구조적 특징을 기술하는 형태학적 (Morphological) 디스크립터들이다. 고분자 반복 단위의 길이, 결 가지 사슬(Side-chain)의 길이 등이 여기에 포함된다. 유리 전이 온도와 용해도 파라미터처럼 이웃 사슬과의 상호 작용이 크게 영향 주는 물성은 기계 학습 훈련시 반드시 형태학적 디스크립터를 비중 있게 고려해야 정확도 높은 예측 모델을 얻을 수 있다. 세 가지 그룹에 포함된 디스크립터들의 총 수는 약 200여 개이다. 900여 종 고분자 각각의 디스크립터를 산출해서 벡터 형태로 집약함으로써 지문화 과정을 마치게 된다.

(3) 물성 예측 모델 구축

예측 모델 구축의 핵심은 고분자 지문과 물성 사이의 상관관계를 찾는 것이다. 마치 x 축과 y 축의 2차원 좌표 상에 산재한 여러 개의 점들을 가장 잘 이어주는 함수 찾기 문제와 유사하다. 함수를 찾는 목적은 나중에 새로운 점의 x 값만을 알 때 y 값을 예상하기 위함이다. 디스크립터와 물성간 관계를 통계적으로 분석하기 위한 기계 학습 모델 훈련 알고리즘으로 KRR(Kernel Ridge Regression), RFR(Random-forest Regression), LASSO (Least Absolute Shrinkage and Selection Operator), GPR (Gaussian Process Regression) 등이 널리 활용되며, 원래 기계 학습 방법의 일종이었던 딥러닝(Deep Learning)은 독립된 줄기로 급격히 발전하는 추세이다.^[20-22] Polymer Genome은 RBF(Radial Basis Function)와 White Noise 함수의 조합을 식별 함수로 선정하고, 모델 훈련 알고리즘으로 GPR을 적용했다. 모델 훈련이란 벤치마크 데이터들 각각의 200여 개 지문 요소와 물성 사이의 관계를 잘 기술해 주도록 식별 함수의 하이퍼 파라미터(Hyper-parameter)를 최적화하는 과정을 말한다. 상기 식별 함수에는 신호 분산 상수(Signal Variance), 지문 벡터들 간의 거리 보정 상수(Length Scale Parameter), 노이즈 상수(Noise Parameter)가 하이퍼 파라미터로 쓰였다. 900여 고분자 중 80%를 모델 훈련 데이터(Train Set)로 사용하고 나머지 20%는 모델 검증용 테스트 데이터(Test Set)로 이용했다. 모델의 성능은 일반적으로 테스트 데이터의 실제 물성 값과 모델을 통한 예

Fig. 5. Parity plots showing the performance of the machine learning models for bandgap (left) and glass transition temperature (right).

측 값의 R^2 혹은 RMSE(Root Mean Square Error)로 정량화한다. 밴드갭 모델의 경우 테스트셋 RMSE가 0.37 eV로, 제일원리 HSE06^[23] 수준 이론치에 매우 근접한 예측 성능을 보여준다. 그림 5에 밴드갭과 유리 전이 온도 예측 모델의 성능을 패리티 도식으로 나타냈다.

2. 고분자 물성 예측 모델 활용 및 대중화

(1) 기계 학습과 실험 연구 병행을 통한 시너지 창출

에너지 저장 분야에서는 제품 디자인 시 불량으로 인한 유전 파괴(Dielectric Breakdown)나 폭발을 항상 염려한다. 고분자는 갑작스러운 유전 파괴 대신 느린 속도로 열화(Degradation)되고, 폭발의 위력이 낮은 장점이 있다. 합성이 쉽고 제조 단가가 낮은 BOPP(Biaxially Oriented Polypropylene)는 이미 하이브리드 자동차의 커패시터 뱅크(Capacitor Bank)나 군용 장비 혹은 각종 생활 용품 등에 활용되고 있다. 그러나 BOPP는 에너지 밀도가 낮다.(~ 5 J/cc) 물질의 에너지 밀도는 유전 파괴 강도(혹은 밴드갭)의 제곱과 유전 상수의 곱에 비례한다. 일반적으로 유전 상수와 밴드갭은 상쇄 관계에 있기 때문에 두 물성이 동시에 높은 고분자를 찾는 것은 쉽지 않다. 그러나 데이터셋 사전 분석만으로 3 eV 이상의 밴드갭과 4 이상의 유전 상수를 동시에 나타내는 고분자들은 적어도 하나 이상의 극성 빌딩 블록(CO, NH, O)들로 구성되며, 동시에 한 개 이상의 방향족 고리(C_6H_4 , C_4H_2S)도 포함한다는 규칙을 찾을 수 있었다. 밴드갭이 일정 수준 이상

REFERENCES

- [20] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).
- [21] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective* (Academic Press, 2015).
- [22] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2013).
- [23] J. Heyd, G. E. Scuseria and M. J. Ernzerhof, *Chem. Phys.* **118**, 8207 (2003).

으로 유지되면서 유전 상수가 높을 수 있던 것은 극성 블록들에 의해 유전 상수의 이온 운동 기여 부분이 높게 발현했기 때문으로 판단된다. 실험 연구팀은 Polymer Genome이 제안한 폴리아우레아(Polyurea) 계열 고분자 $-CO-NH-CO-C_6H_4-$ 와 폴리이미드(Polyimide) 계열 고분자 $-CO-NH-CO-C_6H_4-$ 등을 직접 제조하여 밴드갭과 유전 상수가 예측치와 매우 유사한 것을 확인하였다.^[7]

그러나 이러한 고분자들은 잘 구부러지지 않아서 필름 형태의 유전막으로 활용할 수가 없다. Polymer Genome 팀은 폴리아우레아와 폴리이미드 계열 고분자에 구조적 변화를 준 새로운 고분자들을 디자인하였다. 기계 학습 예측 모델을 통해 밴드갭과 유전 상수가 매우 높은 후보 물질들을 선정하여 합성 후 측정된 결과, BOPP에 비하여 최대 3배 이상의 높은 에너지 밀도를 띠는 새로운 고분자를 찾을 수 있었다. 이들 중 하나인 BTDA-HK511은 반복 단위 내에 CH_2 빌딩 블록들이 길게 이어진 구간이 포함되어 있어서 유연성도 높다.^[24] 본 연구는 데이터 해석, 기계 학습 이론, 그리고 실험을 통한 합성과 측정이 조화를 이룬 성공적인 고분자 연구 사례라고 할 수 있다.

(2) 사용자 친화적 어플리케이션 개발

개발된 기계 학습 예측 모델을 활용하려면 한 가지 작업을 반드시 거쳐야 한다. 모델이 이해할 수 있는 입력 값, 즉 '재료 지문화' 절에서 소개한 지문을 만들어 내는 일이다. 그러나 새로운 물질을 들고 온 연구자에게 200여 가지의 디스크립터로 구성된 지문을 직접 만들도록 요구할 수는 없다. Polymer Genome 팀은 지문화 및 물성 예측 과정을 자동으로 처리하는 웹 어플리케이션(www.polymergenome.org)^[25]을 개발하였다. 현재 Polymer Genome은 900종 고분자의 물성 예측 값과 지문을 데이터베이스에 담고 있으며, 밴드갭, 유전 상수, 굴절률(Refractive Index), 원자화(Atomization) 에너지, 이온화 에너지, 전자 친화도, 유리 전이 온도, 용해도 파라미터 및 밀도 예측 모델을 탑재하고 있다.

사용자는 고분자 연구 분야에서 사용하는 언어인 구조 기반 이름, 화학식, SMILES,^[26] 구조 스케치, 빌딩 블록을 이어 만든 반복 단위 등을 GUI 인터페이스를 통해 제출하는 것만으로 즉시 물성 요약표를 얻을 수 있다. 고분자 이름과 화학식은 직접 데이터베이스 검색 쿼리 작성에 이용되고, SMILES, 빌딩 블록, 스케치는 지문 생성 모듈로 지문화한 뒤 데이터베이스를 검색한다. 'C'와 'CC'는 모두 폴리에틸렌의 SMILES 표기다. 빌딩 블록도 ' CH_2-CH_2 ' 혹은 ' $CH_2-CH_2-CH_2-CH_2$ '처럼 서로 다르게 적을 수 있다. 그러나 지문화 모듈은 입력된 SMILES나 빌딩 블록 분석을 통해서 폴리에틸렌의 고유한 고분자 지문을 생성해 낸다. 따라서 동일한 고분자를 지칭하는 서로 다른 키워드가 입력되더라도 Polymer Genome 어플리케이션은 일관성 있는 물성을 내놓는다. 기계 학습 모델의 진정한 힘은 데이터베이스에 존재하지

Fig. 6. Glimpse of Polymer Genome web application. (available at www.polymergenome.org)

않는 고분자의 물성도 예측이 가능하다는 점이다. 만약 사용자가 입력한 키워드 혹은 생성된 지문을 데이터베이스에 찾지 못하면 사전에 탑재된 기계 학습 모델이 즉시 가동되어 예측 결과를 보여준다. 연구자가 시각화 도구 혹은 시뮬레이션 프로그램에 쉽게 활용할 수 있도록 사슬의 3차원 구조와 원자 좌표 목록을 함께 제공한다. Polymer Genome은 새로운 고분자 데이터셋을 확보할 때마다 기계 학습 모델을 새로 훈련하여 모델 식별 함수의 하이퍼 파라미터를 손쉽게 업데이트 하고, 지문 정보와 물성을 데이터베이스에 추가할 수 있도록 설계되었다. 지속적인 모델 개선과 데이터베이스 확장은 현재의 데이터셋이 대변하지 못하는 새로운 고분자의 물성 예측을 가능하게 할 것이다.

진화하는 기계 학습 방법들

1. 다중 피델리티 학습(Multi-fidelity Learning)

연구자들은 종종 벤치마크 데이터셋이 부족해서 기계 학습 모델을 구축하는데 어려움을 겪는다. 데이터를 만드는 것은 실험과 시뮬레이션을 막론하고 시간과 노력이 필요한 일이기 때문이다.

REFERENCES

- [24] A. Mannodi-Kanakithodi, G. M. Treich, T. D. Tran, R. Ma, M. Tefferi, Y. Cao, G. A. Sotzing and R. Ramprasad, *Advanced Materials* **28**, 6277 (2016).
- [25] Polymer Genome, <https://www.polymergenome.org/>.
- [26] D. Weininger, *Journal of Chemical Information and Computer Sciences* **28**, 31 (1988).

그렇다고 해서 부족한 양의 데이터만으로 모델 학습을 진행하면 성능이 만족스럽지 않을 수도 있다. 이 문제는 다중 피델리티 학습을 이용해서 효과적으로 극복할 수 있다.^[27] 재료 연구 분야에서 찾을 수 있는 좋은 예로 다시 물질의 밴드갭을 들 수 있다. 연구자들은 매우 정확한 밴드갭 혹은 밴드 구조를 필요로 할 때가 있는데, 이를 위해 GW 혹은 하이브리드 기반 제일원리 시뮬레이션을 수행할 수 있다. 그러나 이 방법은 많은 계산 자원과 시간을 필요로 한다. 우선 일부 재료들의 HSE06 수준 밴드 구조를 많은 시간을 쏟아 확보하였다고 가정하자. 이 정보는 다중 피델리티 학습 방법에서 높은 피델리티(High-fidelity) 특성(P_{HF})으로 쓰일 것이다. 한편 연구자는 제일원리 계산을 통해 PBE 수준 밴드 구조를 어렵지 않게 계산해 낼 수 있다. 일반적으로 밴드갭이 실험치에 비해서 작지만 재료의 실제 밴드 특성을 대략적으로 추정하기에는 충분한 정보를 제공한다. 이 정보들은 다중 피델리티 방법에서 낮은 피델리티(Low-fidelity) 특성(P_{LF})으로 쓰인다. 쉽게 생각할 수 있는 활용 방법으로, 기본 지문 요소와 P_{LF} 를 함께 디스크립터 삼아 P_{HF} 예측 모델을 작성하는 수가 있는데, 만약 예측할 물질이 너무나 많은 경우 P_{LF} 를 확보하는 것조차 부담될 수 있다. 다중 피델리티 학습은 주어진 P_{HF} 와 P_{LF} 를 Co-kriging 기법으로 조합하여 P_{HF} 예측 모델을 구축한다. 확보한 모델은 기본적인 지문 요소를 최소한의 입력 값으로 요구하고 P_{LF} 는 선택적 입력으로 두기 때문에, 밴드 구조의 예에서 PBE 수준 결과 확보 여부에 관계없이 P_{HF} (GW 혹은 HSE06 수준 밴드 구조)를 얻을 수 있게 된다. 재료 분야에서는 다중 피델리티 학습 방법 응용 사례를 찾아 볼 수 없었지만 유체 역학 문제나 지구통계학(Geostatistics) 분야에서 자료 확보의 어려움을 극복하기 위해 활용된 바 있다.^[27,28]

2. 어댑티브 학습(Adaptive Learning)

GPR 기반 기계 학습 모델이 특성을 예측할 때마다 함께 제공하는 예측 불확실성, 즉 각 예측의 1σ 값은 모델을 개선하기 위한 방향 설정에 도움을 준다. 예를 들어 모델 훈련 과정에서 예측 불확실성이 큰 테스트 데이터가 발견되면 훈련 데이터로 옮긴 뒤 재훈련함으로써 모델의 전반적인 성능을 향상시킬 수 있다. 또한 모델 구축 후에는 새로운 데이터의 예측 불확실성 정보를 모델 개선에 활용하기도 한다. 가령 새로운 재료의 물성을 예측한 결과 불확실성이 벤치마크 데이터셋의 평균 불확실성보다 확연히 크다면, 이는 모델이 충분한 예측 능력을 가지지 못했다는 것을 의미한다. 따라서 연구자는 반드시 해당 재료 정보를 실측이나 시뮬레이션을 통해 확보한 뒤 새로운 모델 훈련에 포함시켜야 한다. 이러한 과정을 포함하는 기계 학습 방법을 어댑티브 학습이라고 부른다. 넷플릭스(Netflix)나 아마존(Amazon)이 소비자 맞춤형 콘텐츠(비디오 스트리밍 혹은 상품) 추천 모델을 만들고 개선하는데 실제로 활용되는 방법이다. 재료 연구 분야에는 열 이력(Thermal

Hysteresis)이 낮은 형상 기억 합금 개발에 어댑티브 학습 방법을 이용한 사례가 있다.^[29] 기 확보한 22종의 벤치마크 데이터셋으로 최초의 예측 모델을 훈련한 뒤, 열 이력 예측치와 불확실성을 참고해서 새로운 합금을 선정 및 실측하였다. 초기 모델은 정확도 높은 예측보다는 주로 새로운 재료 후보를 탐색하는 목적으로 활용되었다. 실측 결과는 새로운 모델 훈련을 위한 벤치마크 데이터셋에 포함시킴으로써, 모델의 전반적인 불확실성은 점차 감소되고, 새로 선정하는 후보 재료의 물성도 실측치와 유사해졌다. 이 절차를 단 9회 반복한 끝에 유력한 조성 36종을 선정하였고, 실측을 통해 기존 재료보다 우수한 14개 조성을 찾을 수 있었다. 약 800,000가지 가능성 중 우수한 경우 14개를 단 9회의 실험 끝에 찾은 것으로서, 어댑티브 학습 방법을 매우 잘 활용한 대표 사례라고 할 수 있다. 본 방법을 초기 재료 선정 단계 및 실험 설계 과정에 활용한다면 불필요한 재료 탐색과 실험을 피할 수 있어서 연구의 효율성을 매우 높여줄 것으로 판단된다.

결론

이상으로 고분자 소재 개발 사례를 통해 기계 학습 방법을 재료 연구에 활용하기 위한 기술적 단계들(데이터셋 구성, 재료 지문화, 물성 예측 모델 구축, 사용자 친화적 웹 플랫폼 개발)을 소개하였다. 또한 다중 피델리티 학습과 어댑티브 학습 등과 같은 발전된 방법들이 연구자가 향후에 직면할 수 있는 문제들(불완전한 벤치마크 데이터셋, 지속적 모델 개선이 필요한 문제 등)을 효과적으로 해결할 수 있음을 보였다. 실험과 시뮬레이션을 이용한 소재 정보 생산이 점차 용이해짐에 따라 재료 분야에서 기계 학습의 활용도는 더욱 증가할 것이다. 이제 우리는 어떤 문제에 이 방법을 적용할지에 대한 질문을 남겨두고 있다. 너무 복잡해서 분석적 해법을 찾기 어려운 문제, 혹은 기존의 탐구 방식으로 풀어내기에는 큰 시간과 비용이 소요되는 문제 등이 데이터에 근거한 기계 학습 방법을 활용해서 풀기에 적합할 것이다. 그 이유는 통계적 접근 방법으로 도출한 현상학적 모델(Phenomenological Model)이 문제를 직관적으로 해석하거나 결론을 빠르게 추론하는데 도움을 주기 때문이다. 실험과 이론 그리고 기계 학습 방법을 재료 연구에 함께 활용하여, 각 방법론의 취약점을 상호 보완하고 장점을 극대화한다면 앞으로 더욱 많은 성공 사례들이 발표될 것으로 전망된다.

REFERENCES

- [27] A. I. J. Forrester, A. Söbester and A. J. Keane, Proc. R. Soc. A **463**, 3251 (2007).
- [28] P. Perdikaris, D. Venturi, J. O. Royset and G. E. Karniadakis, Proc. Math. Phys. Eng. Sci. **471**, 20150018 (2015).
- [29] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue and T. Lookman, Nat. Commun. **7**, 11241 (2016).