Polym. Korea, Vol. 46, No. 4, pp. 515-522 (2022)
https://doi.org/10.7317/pk.2022.46.4.515

ISSN 0379-153X(Print)
ISSN 2234-8077(Online)

고분자 엉킴 분자량 예측을 위한 심층 학습 모델 연구

박지훈 · 방준하† · 허 준†

고려대학교 화공생명공학과

(2022년 4월 7일 접수, 2022년 5월 12일 수정, 2022년 5월 13일 채택)

Deep Learning Model for Prediction of Entanglement Molecular Weight of Polymers

Jihoon Park, Joona Bang†, and June Huh†

Department of Chemical and Biological Engineering, Korea University, Seoul 02841, Korea

(Received April 7, 2022; Revised May 12, 2022; Accepted May 13, 2022)

초록: 엉킴 분자량은 고분자 사슬의 엉킴에 관한 성질로 고분자의 기계적 물성 및 동적 성질에 관련된 중요성에도 불구하고 광범위한 고분자 종에 따른 측정 데이터가 많이 부족한 실정이다. 본 연구에서는 이러한 문제점을 극복하기 위하여 심층 학습 기법을 활용하여 엉킴 분자량을 예측하고자 하였으며, 분자를 그래프 구조로 변환하여 다루는 그래프 합성곱 신경망을 이용하여 엉킴 분자량을 성공적으로 학습 및 예측하였다. 또한 데이터 수 부족으로 인한 성능의 한계점을 극복하기 위하여, 대용량 데이터를 통해 학습한 지식을 재활용하는 전이 학습 기법을 도입하여 높은 예측 성능 개선을 이루었다. 학습된 인공신경망 모델은 기존 예측 기법보다 높은 예측 성능을 보였다. 본 연구에서 사용된 기계학습 기법은 고분자 역설계 및 물성 예측 기법에 많은 도움을 줄 수 있을 것으로 기대된다.

Abstract: Entanglement molecular weight is one of the key polymer properties strongly related to many mechanical and dynamic behaviors of polymers. Despite its importance, the data for entanglement molecular weight by either measurements or predictions are still far from covering a wide range of polymer species. To address this issue, we employed the deep learning technique to predict the entanglement molecular weight of polymers using graph convolutional neural networks that convert molecules into graph structures. In addition, to overcome the limitation due to the lack of data, the transfer learning technique, which transfers knowledge learned through large-scale datasets, was also introduced to improve the performance. The trained neural network model showed higher prediction performance than the conventional prediction methods.

Keywords: entanglement molecular weight, machine learning, transfer learning, graph convolutional neural network, quantitative structure property relationship.

서 론

엉킴 분자량 M_e(entanglement molecular weight)는 고분자 사슬의 엉킴에 관한 고유한 성질로, 서로 이웃한 두 엉킴 점 사이 고분자 사슬의 평균적인 분자량이다.[1] 고분자 사슬의 길이가 짧을 경우, 사슬 내/간 엉킴이 발생하지 않지만 사슬이 길어질수록 엉킴이 발생하여 고분자는 다른 거동을 보이게 된다. 이는 고분자의 용융 점도(melt viscosity)와 분자량 간 관계로 나타나는데, 고분자의 분자량이 엉킴 분자량의 두배인 $2M_e$ 미만일 경우, 용융 점도는 분자량에 선형적으로 비례하며 그 이상일 경우, 용융 점도는 분자량의 3.4제곱에 비례하여 급격히 증가하는데, M_e는 플라스틱의 다양한 기계적 물성에 관계되어 있다. 따라서 엉킴 분자량은 원하는 물성을 갖는 고분자 역설계시 고려해야 할 중요한 요소이다. 엉킴 분자량은 용융 또는 고무 상태(rubbery state)에 있는 고분자의 고원 탄성률(plateau modulus) G_N^o 을 측정, 다음 관계식을 이용하여 구할 수 있다.[2]

$$G_N^o = \frac{\rho RT}{M_e} \tag{1}$$

이 때, ρ는 고분자의 용융 밀도, R는 기체 상수, T는 절대 온도이다.

그러나 이러한 중요성에도 불구하고 측정의 어려움에[3] 기인하여 엉킴 분자량의 측정 데이터는 매우 부족하다. CROW

†To whom correspondence should be addressed.
joona@korea.ac.kr, ORCID 0000-0002-2301-6190
junehuh@korea.ac.kr, ORCID 0000-0002-4610-4521

사에서 운영하는 polymerdatabase에[4] 60종 정도의 엉킴 분자량 데이터가 수록되어 있으며 최대의 고분자 데이터베이스인 PolyInfo에[5,6] 엉킴 분자량 데이터가 기록되어 있지 않다. 따라서 고분자 역설계에 많은 어려움이 발생한다. 이를 극복하기 위해선 많은 측정을 통한 데이터 축적도 중요하지만, 고분자의 구조를 통하여 예측할 수 있는 예측 기법의 개발이 필요하다.

측정된 데이터를 통해 주어진 구조로부터 물성을 예측하는 구조-물성의 정량적 관계(quantitative structure property relationship, QSPR)[7] 기법이 등장하면서 고분자의 다양한 물성도 예측하려는 시도가 있어왔다. 이 중, Van Krevelen의 group contribution method[8] 및 Bicerano의 topological method가[9] 엉킴 분자량을 예측하는 방법을 기술하였으나, 다른 물성의 예측 방법과 비교하였을 시 비교적 정확도가 높지 않은 단점을 가지고 있다.

또한 4차 혁명 이후 심층 학습(deep learning) 기법이 대두되면서[10-12] 고분자의 물성도 이를 통하여 예측하려는 시도가[13-17] 많이 있어왔다. 데이터 수가 많은 유리 전이온도(glass transition temperature) 등의 물성은 다양한 심층 학습 기법들을 통해 매우 높은 정확도로 예측 및 보고되었다.[18-20] 그러나 데이터가 매우 적은 물성들의 경우 전이 학습(transfer learning)을 이용한 시도[21,22]가 보고되었으나 그 사례가 많지 않으며, 특히 엉킴 분자량은 사례가 전무하다.

심층 학습에 사용되는 인공신경망 모델은 단순한 fully-connected layer부터 합성곱 신경망까지 다양하게 제시되고 있다. 특히 그래프 합성곱 신경망은 그래프 구조를 다루는 신경망으로서 분자 구조를 다루기에 적합한 특성을 갖고 있다. 신경망 내에서 분자는 원자가 꼭짓점(vertex), 원자간 결합은 변(edge)인 무방향 그래프(undirected graph)와 같이 다루어진다. 그래프 신경망은 Monfardini 그룹에[23] 의해 발표된 이후 분자 표현에 적합한 특성에 의해 Adams 그룹에[24] 의하여 분자 지문(molecular fingerprints)으로 사용함에 있어 월등한 성능이 보고되었다. 이 후 DeepChem 라이브러리의[25] message passing neural network(MPNN)[26] 등 개선을 통해 QSAR 및 QSPR의 다양한 분야에 활용되어 강력한 성능이 보고되고 있다.[27,28]

엉킴 분자량과 같이 데이터의 수가 적은 목표에 대해 심층 학습을 수행할 경우, support vector machine(SVM) 및 random forest와 같은 전통적인 알고리즘에 비해 더 예측 성능이 떨어짐이 일반적으로 알려져 있다.[29,30] 따라서 데이터의 수가 적을 경우, 전이 학습(transfer learning)을[31] 통해 이를 보완하는 기법이 사용된다. 전이 학습은 관련된 대용량 데이터 집합을 학습 후, 해당 학습 모델을 목표하는 학습에 재활용하는 기법으로, 소규모 데이터 집합 학습에 적합하며, 처음부터 학습하는 것보다 빠른 시간 내에 높은 예측 성능을 달성할 수 있다.

본 연구에서는 엉킴 분자량의 예측 성능 개선을 위하여 심층 학습을 이용, MPNN을 활용한 그래프 합성곱 인공신경망을 통해 CROW polymerdatabase로부터 수집한 용융 상태 고분자의 엉킴 분자량을 학습하였다. 학습을 위하여 고분자의 구조를 표현한 문자열을 인공신경망에 입력, 엉킴 분자량 값이 최종 출력되도록 회귀 지도 학습을 수행하였다. 입력된 고분자의 구조는 인공신경망에서 분자의 구조를 학습하기에 용이한 그래프 자료구조로 변환되었으며, 서로 결합된 원자-원자간 상호작용을 학습하도록 하였다. 인공신경망에서 출력된 예측 값 및 목표 값 사이의 오차를 인공신경망에 다시 역전파(backpropagation)하여, 오차를 감소시키는 방향으로 인공신경망 내의 상태 파라미터인 가중치(weight) 및 편차(bias)를 조정함으로써 엉킴 분자량을 학습하였다. 또한 적은 데이터 수로 인하여 발생하는 예측 성능의 한계를 극복하기 위하여 추가적으로 전이 학습 기법을 도입하였다. 목표 물성과 관련도가 높은 물성을 가진 대용량 데이터 집합인 QM9 데이터 집합에 대하여 먼저 학습 후, 해당 인공신경망에 고분자 구조-엉킴 분자량 데이터 집합을 다시 학습시켜 예측 성능을 높은 폭으로 개선하였다.

데이터 가공 및 심층 학습

데이터 집합 준비. 인공신경망 학습을 위해 두가지 데이터 집합을 준비하였다. 사전 학습을 위한 데이터 집합은 QM9[32,33] 데이터 집합을 사용하였다. QM9 데이터 집합은 H, C, N, O, F로 구성된 작은 유기 분자들의 성질을 양자 계산을 통해 축적한 데이터 집합으로, 총 13만개의 데이터를 가지고 있어 사전 학습에 사용할 빅데이터 집합으로 적합하다. 중복된 항목 및 라디칼을 제거 후 데이터 집합에 기록된 분자 구조를 simplified molecular input line entry system(SMILES)[34] 문자열로 변환하였다.

고분자의 엉킴 분자량 학습을 위하여 고분자 구조-엉킴 분자량 데이터 집합을 CROW polymerdatabase에서[4] 수집하였으며, 해당 엉킴 분자량 데이터들은 실험을 통해 용융 상태의 고분자로부터 측정된 값이다. 엉킴 분자량 데이터들은 실험을 통해 QM9 데이터 집합이 H, C, N, O, F로만 이루어진 분자들만 수록하고 있으므로 이와의 일관성을 유지하기 위하여 반복단위의 원소가 H, C, N, O, F로만 이루어진 데이터만 사용하였으며, 총 51개의 데이터를 수집하였다. 반복단위간의 작용기를 표현하기 위하여 반복단위를 4번 반복한 올리고머의 SMILES를 엉킴 분자량과 함께 입력-출력 데이터 집합으로 사용하였다.

분자의 그래프화. 인공신경망을 통해 학습할 수 있도록 SMILES 문자열은 분자 구조로 변환 및 문자열에서 생략된 수소 원자 추가 후, 원자는 꼭짓점, 원자-원자간 결합은 변인 무방향 그래프 $G=(V, E)$로 변환하였다. 이 때 $V \in \mathbb{R}^{N \times D}$는 원자 표현 행렬을, $E \in \mathbb{R}^{N \times N \times F}$는 원자간 결합이 포함된 그래

Table 1. Atom Representations of Molecular Graph

Feature	Description
Atom Type	H, C, O, N, F (one-hot)
Bond Degree	The number of directly bonded neighbors (one-hot)
Formal Charge	Formal Charge (integer)
Hybridization	sp, sp^2, sp^3 (one-hot)
IsAromatic	Is atom an aromatic system (binary)

Table 2. Bond Representations of Molecular Graph

Feature	Description
Bond Type	single, double, triple, aromatic (one-hot)
IsConjugated	Is bond conjugated (binary)
IsInRing	Is bond in a ring (binary)

프 인접 행렬을 나타내며, N, D, F는 각각 분자 내의 원자 수, 원자 표현(atom representations)의[26] 길이, 결합 표현(bond representations)의[26] 길이를 나타낸다. 원자 표현 및 결합 표현은 RDKit Python 라이브러리를[35] 활용, Table 1 및 Table 2와 같이 추출하였다.

인공신경망 모델. 학습에서 사용한 그래프 합성곱 인공신경망은 MPNN의[26,36,37] 디자인을 이용하였으며, 인공신경망 모델의 개요도를 Figure 1에 나타냈다. 인공신경망 내에서 입력된 그래프의 원자 표현 행렬 V는 fully-connected layer(FC)인 특징(feature) 추출 레이어를 통해 각각 초기 원자 특징 $X^0(X^t \in \mathbb{R}^{N\times L})$로 변환된다.

$$X^0 = \mathrm{ReLU}(VW_{FC,V} + B_{FC,V}) \tag{2}$$

$W_{FC,V}$ 및 $B_{FC,V}$는 원자 특징 추출 레이어의 가중치(weight) 및 편향(bias)를, L은 특징의 길이를 나타낸다. 그 후, 초기 원자 특징 X^0는 MPNN 모듈 내에서 직접 연결된 다른 원자로부터 발생되는 메시지 M^t를 통해 $t < T$인 동안 반복적으로 T번 갱신된다. u번째 원자에 대한 t번째 메시지 M_u^t는 MPNN 모듈 내의 NNConv[26] 레이어를 통해 다음과 같이 연산된다.

$$M_u^t = W_{NC}X_u^t + \sum_{v\in\mathbb{N}(u)} X_v^t \phi(E_{uv}) \tag{3}$$

$W_{NC} \in \mathbb{R}^{L\times L}$는 NNConv 레이어의 가중치 행렬, $\mathbb{N}(u)$는 u번째 원자에 직접적으로 연결된 원자들의 집합이며, $\phi(E_{uv}) \mapsto \mathbb{R}^{L\times L}$는 결합 표현 E_{uv}를 결합 별 가중치로 대응시키는 함수이다. $\phi(\cdot)$는 다층 fully connected layer로 구현되어 학습되도록 하였다. u번째 원자의 특징 X_u^t는 전달된 메시지 M_u^t를 통해 MPNN 모듈 내의 GRU[38] 레이어에서 갱신된다.

$$X_u^{t+1} = \mathrm{GRU}(M_u^t, H_u^t) \tag{4}$$

H_u^t는 GRU 레이어의 u번째 원자에 대한 t번째 은닉 상태(hidden state)이다. 최종적으로 분자의 특징을 나타내는 출력 특징 $\bar{X}$는 전역 합 풀링 레이어(global add pooling layer)를 통해 출력된다.

$$\bar{X} = \sum X_u^T \tag{5}$$

분자의 특징 $\bar{X}$는 다층 fully-connected layer를 거쳐 목표 물성 값 P로 출력된다. 인공신경망의 구현은 Python 라이브러리 중 기계학습 라이브러리인 PyTorch와[39] 그래프 인공신경망 라이브러리인 PyTorch Geometric을[40] 활용하여 구현하였다.

인공신경망 학습. 전이 학습의 효과를 비교하기 위하여 두

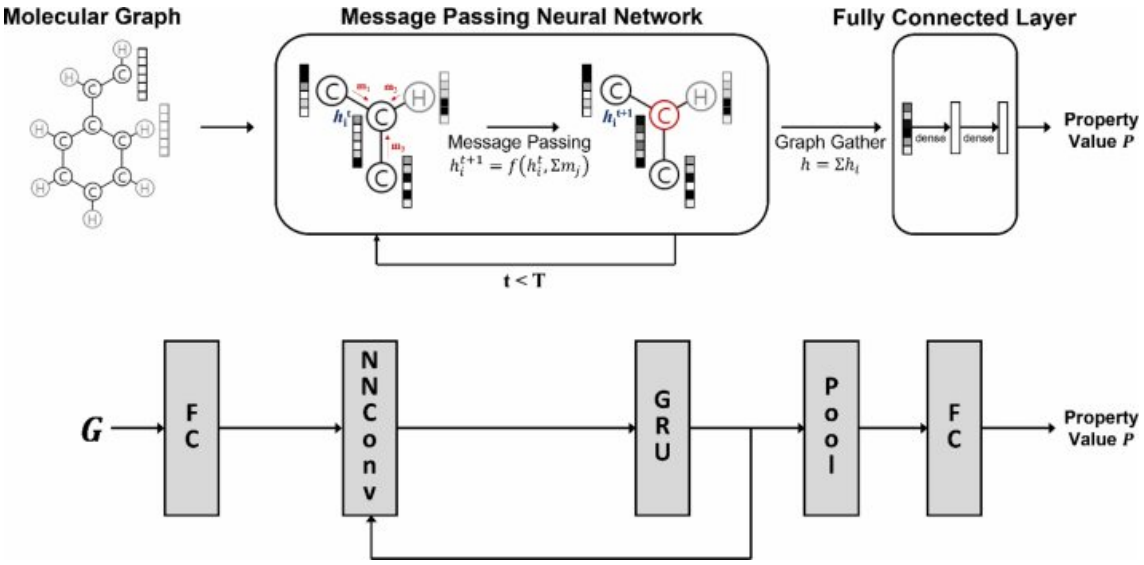

Figure 1. Overview of the graph convolutional neural network model. Each neural layer is explained in eq. (2), (3) and (4).

가지 방법으로 고분자의 엉킴 분자량을 학습 후, 성능을 측정하였다. 첫번째로 전이 학습을 수행하지 않고 처음부터 고분자 구조-엉킴 분자량을 학습하였으며, 500 epoch 동안 학습하여 시험 집합 예측 성능이 가장 높았던 epoch의 모델을 이용하여 성능을 측정하였다. 두번째로 QM9 데이터 집합을 먼저 학습한 모델을 재사용하는 전이 학습을 통해 고분자 구조-엉킴 분자량을 학습하였다. 이 과정에서 사전 학습 과정으로 QM9 데이터 집합을 먼저 200 epoch 동안 학습하였다. 이후 학습한 모델의 말단 레이어인 다층 fully-connected layer를 제외한 모든 레이어를 학습이 되지 않도록 고정(freeze) 및 말단 다층 fully-connected layer를 초기화 후 고분자 구조-엉킴 분자량으로 200 epoch 동안 학습하였다. 이후 고분자 데이터 집합에 적합하도록 모델의 미세 조정(fine-tuning)을 위하여 모든 레이어의 고정 상태를 해제한 후, 기존 학습률(learning rate) lr의 0.01배인 $0.01lr$을 새로운 학습률로 설정하여 300 epoch 동안 학습하였다. 미세 조정 과정 동안 시험 집합 예측 성능이 가장 높았던 epoch의 모델을 이용하여 전이 학습 모델의 성능을 측정하였다. 학습에 사용한 최적화알고리즘(optimizer)은 Adam optimizer를[41] 사용하였으며, 초 매개 변수(hyper parameter)인 학습률 lr=$1e^{-4}$, 특징 길이 L=128, 메시지 전달 반복 횟수 T=6, 배치의 크기는 16으로 설정하였다. 학습을 위한 손실 함수(loss function) 및 시험 집합에 대한 모델의 성능 평가 지표로 평균 제곱 오차(MSE, mean square error)를 사용하였다. 또한 회귀 문제를 학습하는 모델이므로 출력 값은 scikit-learn Python 라이브러리의[42] StandardScaler를 사용, 0의 평균 및 1의 표준편차를 갖도록 표준화(standardization)하여 예측하도록 하였다.

인공신경망 성능 검증. 수집한 고분자 구조-엉킴 분자량 데이터의 개수는 매우 적으므로, 일반적인 방법을 사용하여 훈련 집합(train set) 및 시험 집합(test set)을 분리 시 모델의 성능 검증에 문제점이 발생한다. 먼저 학습 시 분리된 시험 집합에 따른 예측 성능의 편차가 매우 크며, 학습 시 인공신경망 모델이 훈련 집합에 대하여 과적합(overfitting)되어 실제 예측 성능이 매우 저하될 가능성이 크다. 따라서 모델의 성능 검증을 위하여 K-겹 교차 검증(K-fold cross validation)을[43] 사용하여 이러한 문제점을 극복하고자 하였다. K-겹 교차 검증에서 데이터는 집합을 서로 겹치지 않는 K개의 부분 집합으로 나뉜다. i($i \in [1, K]$)번째 실험에서 i번째 부분 집합을 시험 집합으로, 나머지 부분 집합을 훈련 집합으로 사용하여 총 K번 실험을 진행하였다. 이 후 총 K개의 결과를 취합하여 전체적인 성능을 평가하였으며, 피어슨 상관 계수(pearson correlation coefficient) r,[44] 평균 절대 오차(mean absolute error, MAE), 평균 제곱근 오차(root mean square error, RMSE)를 평가 지표로 이용하였다. K-겹 교차 검증 방법은 scikit-learn Python 라이브러리를 사용하여 구현하였으며, K=5로 하여 실험을 진행하였다.

결과 및 토론

데이터 시각화를 통한 학습 전략 수립. QM9 데이터 집합은 총 12 종류의 다양한 성질을 기록하고 있다. 따라서 사전 학습에 사용하기 위해 고분자의 엉킴 분자량과 상관 관계를 갖는 적합한 성질 탐색이 필요하였다. Figure 2에 나타낸 것과 같이 고분자의 물성 중 몰 부피(molar volume)가 엉킴 분자량과 상관성을 보임을 확인할 수 있었으며, QM9의 성질 중 분자의 크기와 관련 있는 전자 공간 범위($\langle R^2 \rangle$, electronic spatial extent)가 이와 유사한 성질로 판단되어[45] 사전 학습의 목표 예측 성질로 사용하였다.

다만, Figure 2에서 확인할 수 있듯이 모든 데이터가 양의 상관 관계를 보이지 않았다. 데이터는 두개의 군집(cluster)으로 나뉘어 한 군집은 양의 상관 관계를 보였으며 다른 군집은 상관 관계를 보이지 않았다. 따라서 데이터를 나누지 않고 학습하였을 경우, 다른 군집의 데이터로 인해 학습에 악영향을 끼칠 수 있어 이를 극복하고자 데이터를 나누어 각각 학습을 진행하는 전략을 수립하였다.

데이터를 두개의 군집으로 나누기 위하여 군집 분석(clustering analysis)을 시행하였다. OSIRIS Datawarrior 프로그램을[46] 이용하여 분자의 분자 지문을 통해 군집 분석을 수행하였으며, 사용한 분자 지문은 FragFp 설명자(descriptor)[47]이다. 군집 분석을 수행한 결과, 고분자 데이터를 총 5개의 군집으로 나눌 수 있었으며, 이는 Figure 2의 마커 색깔 및 범례를 통해 나타냈다. 이를 통해 acrylate, methacrylate, ethylene 구조를 갖는 고분자 데이터를 양의 상관 관계를 갖는 펜던트 그룹으로, 나머지 데이터를 그 상관 관계를 가지지 않는 사슬 그룹으로 나눌 수 있었으며 각각의 그룹에 대하여 실험을 진행하였다.

사전 학습. 전이 학습에 사용하기 위하여 QM9 데이터 집

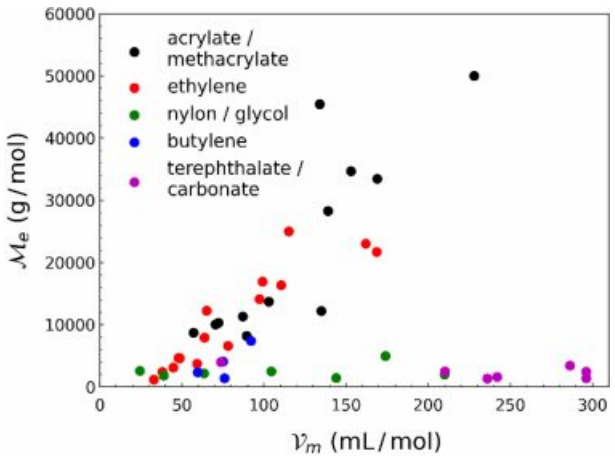

Figure 2. Scatter plot of molar volume of polymers versus entanglement molecular weight of polymers. Each color represents molecular groups from clustering analysis.

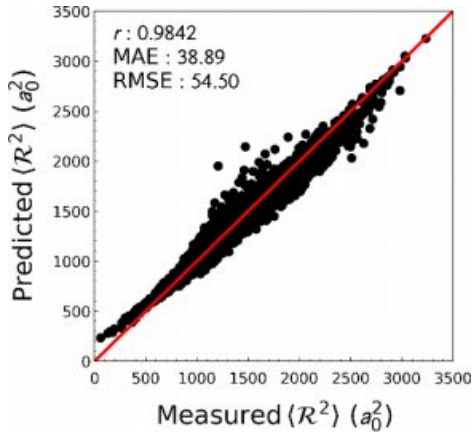

Figure 3. Prediction performance of an electronic spatial extent from pre-training.

합을 이용, 사전 학습을 수행하였다. 총 데이터 집합 중 훈련 집합으로 80%, 나머지는 시험 집합으로 나누어 학습 및 검증을 진행하였다. 목표 예측 성질인 전자 공간 범위는 데이터의 범위 및 편차가 매우 크고, 양의 값 만을 예측해야 하므로 해당 값을 그대로 학습하지 않고, 상용로그 연산을 통해 변환된 값을 학습 및 예측하도록 하였다. 그 결과, 시험 집합에 대하여 높은 예측 성능을 보임을 확인할 수 있었다. (Figure 3).

학습 결과. 사전 학습과 마찬가지로, 고분자의 엉킴 분자량은 데이터의 범위 및 편차가 매우 크고 양의 값 만을 예측해야 하므로 상용로그 연산을 통해 변환된 값을 학습 및 예측하도록 하였다. 또한 전이 학습 도입의 효과를 알아보기 위하여, 전이 학습을 하지 않은 경우 및 전이 학습을 한 경우에 대하여 고분자의 엉킴 분자량을 예측한 결과를 비교하였다. 먼저 사전 학습 데이터와 양의 상관 관계를 갖는 펜던트 그룹에 대한 학습 결과를 Figure 4에 나타냈다. 펜던트 그룹의 경우 전이 학습 여부에 관계없이 높은 예측 성능을 보여주었으나, 전이 학습을 수행하였을 때 모든 지표에서 더 높은 예측 성능을 보여주었다. 따라서 사전 학습에서 유사한 성질을 학습, 전이되는 지식을 이용하여 본 학습에 도움을 받

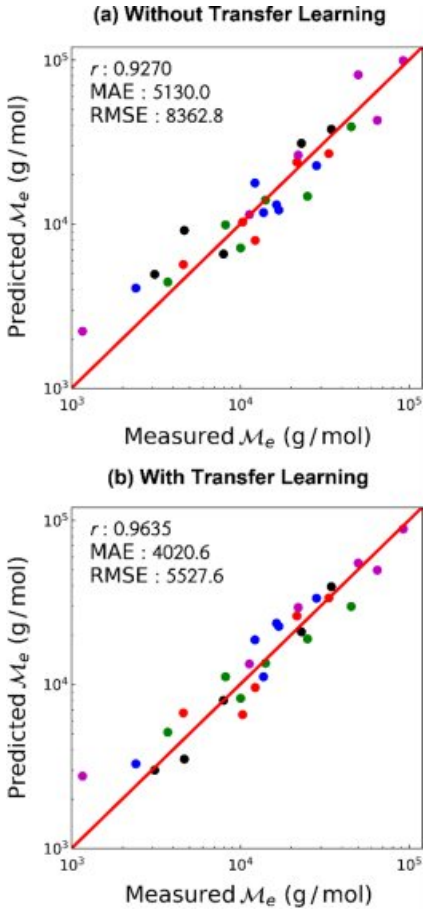

Figure 4. Prediction performance of the pendant group entanglement molecular weight (a) without transfer learning; (b) with transfer learning. Each color represents different folds of cross-validation.

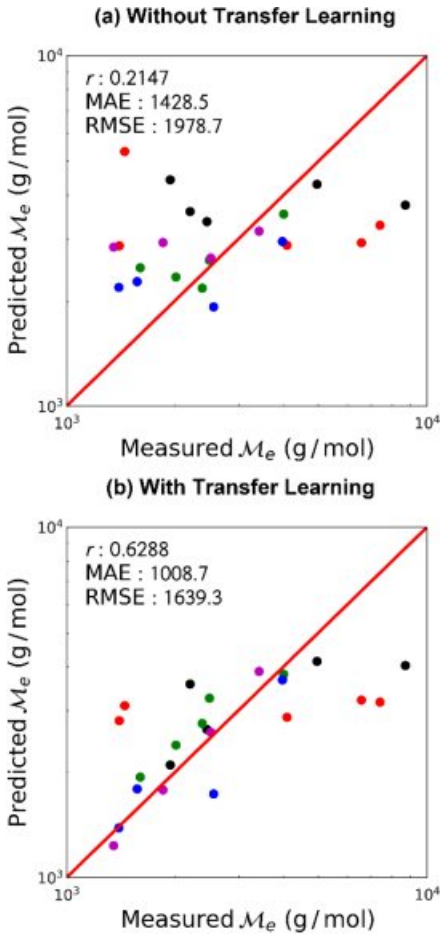

Figure 5. Prediction performance of the chain group entanglement molecular weight (a) without transfer learning; (b) with transfer learning. Each color represents different folds of cross-validation.

을 수 있음을 확인하였다.

사전 학습 데이터와 상관 관계를 갖지 않는 사슬 그룹에 대한 학습 결과를 Figure 5에 나타냈다. 전이 학습을 하지 않은 경우, 피어슨 상관 계수 r=0.2147로서 분자 구조에 관계없이 비슷한 값을 예측, 낮은 예측 성능을 보였다. 이는 펜던트 그룹의 경우 주쇄의 구조가 비슷하여 학습이 비교적 수월한 반면, 사슬 그룹은 매우 다양한 구조를 가지고 있어 적은 데이터 수로는 학습에 어려움을 겪었기 때문으로 해석된다. 그러나 전이 학습을 수행한 경우, 전이 학습을 하지 않은 경우에 비하여 월등히 예측 성능이 개선되었다. 사슬 그룹의 목표 물성 데이터는 사전 학습 시 사용한 목표 물성 데이터와 상관 관계를 보이지 않음에도 전이되는 지식이 본 학습에 도움을 주었는데, 이는 사용한 인공신경망 모델의 구조적 특성 때문으로 해석된다. 사용한 인공신경망 모델은 그래프 합성곱 인공 신경망인 MPNN 모듈 및 말단의 다층 fully-connected layer 두가지로 나뉘어진다. MPNN 모듈에서는 입력된 분자로부터 분자 구조를 학습하며, 다층 fully-connected layer에서는 MPNN 모듈에서 출력되는 값과 최종 출력 값 사이의 관련성을 학습한다. 따라서 설령 출력 데이터의 상관 관계가 부족할지라도, MPNN 모듈에서 학습하는 분자 구조는 화학 공간(chemical space) 내의 일반적인 지식으로 취급되어 전이 학습 시 도움을 줄 수 있다. 그럼에도 불구하고, 펜던트 그룹의 경우에 비하여 떨어지는 예측 성능은 적절한 상관 관계를 가지는 대리 물성(proxy property) 설정의 중요성을 보여준다.

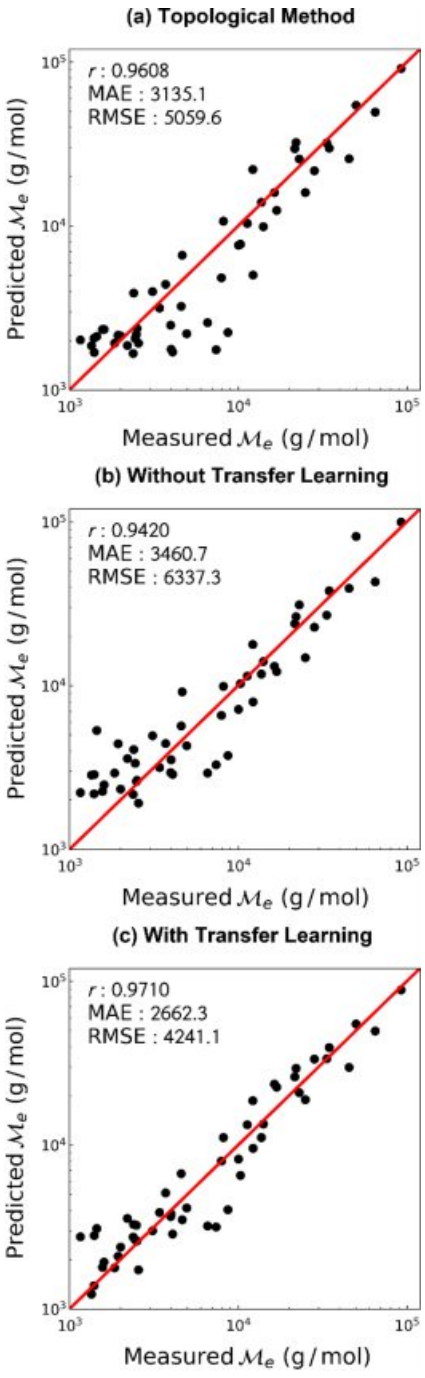

Figure 6. Prediction performance of the entanglement molecular weight (a) from topological method; (b) without transfer learning; (c) with transfer learning.

기존 예측 방법과의 비교. 기존에 널리 사용되던 예측 방법과의 예측 성능 비교를 위하여 Bicerano의 topological method를[9] 이용하였다. 해당 방법을 이용하여 예측한 고분자의 엉킴 분자량을 대조군으로 설정, 비교 결과를 Figure 6에 나타냈다. 인공신경망 학습 결과는 펜던트 그룹 및 사슬 그룹의 결과를 통합하여 그래프에 도시하였다. Topological method와 비교하였을 시 전이 학습을 하지 않은 경우 예측 성능이 떨어졌지만, 전이 학습을 사용한 경우 모든 지표 측면에서 예측 성능이 개선됨을 확인하였다. 또한 topological method를 사용하여 예측하였을 때 비교적 낮은 엉킴 분자량을 갖는 고분자에 대해서는 예측 성능이 좋지 않았으나, 전이 학습을 사용하였을 경우 모든 값 범위에서 좋은 예측 성능을 보여주었다.

결 론

그래프 합성곱 신경망에 전이 학습을 도입하여 매우 적은 데이터 수를 가지는 용융 상태 고분자의 엉킴 분자량을 높은 성능으로 예측하였다. 먼저 MPNN 모듈을 기반으로 한 그래프 합성곱 신경망을 통해 엉킴 분자량을 성공적으로 예측하였다. 이후 엉킴 분자량과 연관을 가지는 대용량 데이터 집합을 통해 사전 학습을 수행한 뒤, 엉킴 분자량에 대해 전이 학습을 수행하여 예측 성능을 큰 폭으로 개선하였다. 특히 사슬 그룹 데이터에 대한 전이학습 사례로 미루어 보아, 직접적인 연관성을 갖지 않는 데이터라도 화학 공간에 대한 일반적인 지식 전이를 통하여 목표 학습에 많은 도움을 줄 수

있음을 확인하였으며 이는 분자 구조를 직접적으로 학습하는 MPNN 모듈에 의한 효과로 해석된다. 전이 학습을 수행한 결과, 기존 사용되던 topological method에 비해서도 높은 예측 성능을 보였으며, 전 데이터 범위에서 고른 예측 성능을 보였다. 본 연구에서 사용한 엉킴 분자량 데이터는 용융 상태로부터 측정된 데이터에 한정되어 있으나, 온도 또는 용액의 농도 등에 따른 엉킴 분자량을 추가적으로 수집 및 학습한다면 다양한 조건에 따른 엉킴 분자량도 예측할 수 있을 것으로 생각된다. 따라서 본 연구는 원하는 물성을 갖는 고분자의 역설계에 많은 도움이 될 것으로 생각되며, 데이터 수가 적은 다른 성질의 예측 전략 수립에도 많은 기여를 할 것으로 기대된다.

이해상충: 저자들은 이해상충이 없음을 선언합니다.

참 고 문 헌

1. Wool, R. P. Polymer Entanglements. *Macromolecules* **1993**, 26, 1564-1569.
2. Eckstein, A., Suhm, J.; Friedrich, C.; Maier, R. D.; Sassmannshausen, J.; Bochmann, M.; Mülhaupt, R. Determination of Plateau Moduli and Entanglement Molecular Weights of Isotactic, Syndiotactic, and Atactic Polypropylenes Synthesized with Metallocene Catalysts. *Macromolecules* **1998**, 31, 1335-1340.
3. Cho, K. S. Viscoelastic Measurement and Structure of Polymeric Materials. *Polym. Sci. Tech*. **2008**, 19, 170-176.
4. CROW. Polymer Database. https://polymerdatabase.com (accessed Feb 17, 2022).
5. National Institute for Materials Science. PoLyInfo. https://polymer.nims.go.jp/en/ (accessed Feb 17, 2022).
6. Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer Database for Polymeric Materials Design. *In 2011 International Conference on Emerging Intelligent Data and Web Technologies*, **2011**, 22-29.
7. Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, 24, 279-287.
8. Van Krevelen, D. W. *Proerties of Polymers, Their Estimation and Correlation with Chemical Structure*, 2nd ed; Elsevier: Amsterdam, 1976.
9. Bicerano, J. *Prediction of Polymer Properties*, 1st ed; Marcel Dekker: New York, 1993.
10. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, 521, 436-444.
11. Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design–A Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, 4, 828-849.
12. Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model* **2019**, 59, 2545-2559.
13. Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-Learning Predictions of Polymer Properties with Polymer Genome. *J. Appl. Phys.* **2020**, 128, 171104.
14. Kuenneth, C.; Rajan, A. C.; Tran, H.; Chen, L.; Kim, C.; Ramprasad, R. Polymer Informatics with Multi-task Learning. *Patterns* **2021**, 2, 100238.
15. Sha, W.; Li, Y.; Tang, S.; Tian, J.; Zhao, Y.; Guo, Y.; Zhang, W.; Zhang, X.; Lu, S.; Cao, Y.; Cheng, S. Machine Learning in Polymer Informatics. *InfoMat*. **2021**, 3, 353-361.
16. Kim, C.; Batra, R.; Chen, L.; Tran, H.; Ramprasad, R. Polymer Design Using Genetic Algorithm and Machine Learning. *Compu. Mater. Sci.* **2021**, 186, 110067.
17. Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, 6, 1-10.
18. Cassar, D. R.; de Carvalho, A. C.; Zanotto, E. D. Predicting Glass Transition Temperatures Using Neural Networks. *Acta Mater.* **2018**, 159, 249-256.
19. Alcobaca, E.; Mastelini, S. M.; Botari, T.; Pimentel, B. A.; Cassar, D. R.; de Leon Ferreira, A. C. P.; Zanotto, E. D. Explainable Machine Learning Algorithms for Predicting Glass Transition Temperatures. *Acta Mater.* **2020**, 188, 92-100.
20. Chen, G.; Tao, L.; Li, Y. Predicting Polymers' Glass Transition Temperature by a Chemical Language Processing Model. *Polymers* **2021**, 13, 1898.
21. Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent. Sci.* **2019**, 5, 1717-1730.
22. Wu, S.; Kondo, Y.; Kakimoto, M. A.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-learning-assisted Discovery of Polymers with High Thermal Conductivity Using a Molecular Design Algorithm. *Npj Comput. Mater.* **2019**, 5, 1-11.
23. Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **2008**, 20, 61-80.
24. Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Adv. Neural Inf. Process. Syst.* **2015**, 28.
25. Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. O'Reilly Media: Sebastopol, 2019.
26. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In International Conference on Machine Learning. *PMLR* **2017**, 1263-1272.
27. Sun, M.; Zhao, S.; Gilvary, C.; Elemento, O.; Zhou, J.; Wang, F. Graph Convolutional Networks for Computational Drug Development and Discovery. *Brief. Bioinform.* **2020**, 21, 919-935.

28. Wang, X.; Li, Z.; Jiang, M.; Wang, S.; Zhang, S.; Wei, Z. Molecule Property Prediction Based on Spatial Graph Embedding. *J. Chem. Inf. Model* **2019**, 59, 3817-3828.
29. Zhu, X.; Vondrick, C.; Fowlkes, C. C.; Ramanan, D. Do We Need More Training Data? *Int. J. Comput. Vis.* **2016**, 119, 76-92.
30. Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M. M. A.; Yang, Y.; Zhou, Y. Deep Learning Scaling is Predictable, Empirically. *arXiv* **2017**, arXiv:1712.00409.
31. Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2009**, 22, 1345-1359.
32. Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model* **2012**, 52, 2864-2875.
33. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, 1, 1-7.
34. Weininger, D. SMILES, A Chemical Language anD Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28, 31-36.
35. RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org (accessed Feb 22, 2022).
36. Li, Z.; Wellawatte, G. P.; Chakraborty, M.; Gandhi, H. A.; Xu, C.; White, A. D. Graph Neural Network Based Coarse-grained Mapping Prediction. *Chem. Sci.* **2020**, 11, 9524-9531.
37. Vinyals, O.; Bengio, S.; Kudlur, M. Order Matters: Sequence to Sequence for Sets. *arXiv* **2015**, arXiv:1511.06391.
38. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
39. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. Pytorch: An Imperative Style, High-performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, 32.
40. Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. *arXiv* **2019**, arXiv:1903.02428.
41. Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
42. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V., Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, 12, 2825-2830.
43. Mosteller, F.; Tukey, J. W. Data Analysis, Including Statistics. *Handbook of Social Psychology* **1968**, 2, 80-203.
44. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing*, Springer: Berlin, 2009; pp 37-40.
45. Hait, D.; Liang, Y. H.; Head-Gordon, M. Too Big, Too Small, or Just Right? A Benchmark Assessment of Density Functional Theory for Predicting the Spatial Extent of the Electron Density of Small Chemical Systems. *J. Chem. Phys.* **2021**, 154, 074109.
46. Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: an Open-source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, 55, 460-473.
47. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 1273-1280.