KIPS Trans. Softw. and Data Eng.
Vol.11, No.1 pp.11~18
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.1.11

De Novo Drug Design Using Self-Attention Based Variational Autoencoder

Piao Shengmin[†] · Jonghwan Choi[††] · Sangmin Seo[†††] · Kyeonghun Kim[††††] · Sanghyun Park[†††††]

ABSTRACT

De novo drug design is the process of developing new drugs that can interact with biological targets such as protein receptors. Traditional process of de novo drug design consists of drug candidate discovery and drug development, but it requires a long time of more than 10 years to develop a new drug. Deep learning-based methods are being studied to shorten this period and efficiently find chemical compounds for new drug candidates. Many existing deep learning-based drug design models utilize recurrent neural networks to generate a chemical entity represented by SMILES strings, but due to the disadvantages of the recurrent networks, such as slow training speed and poor understanding of complex molecular formula rules, there is room for improvement. To overcome these shortcomings, we propose a deep learning model for SMILES string generation using variational autoencoders with self-attention mechanism. Our proposed model decreased the training time by 1/26 compared to the latest drug design model, as well as generated valid SMILES more effectively.

Keywords : De Novo Drug Design, SMILES, Deep Learning, Self-attention, Variational Autoencoder

Self-Attention 기반의 변분 오토인코더를 활용한 신약 디자인

Piao Shengmin[†] · 최 종 환[††] · 서 상 민[†††] · 김 경 훈[††††] · 박 상 현[†††††]

요 약

신약 디자인은 단백질 수용체와 같은 생물학적 표적과 상호작용할 수 있는 약물 후보물질을 식별하는 과정이다. 전통적인 신약 디자인 연구는 약물 후보 물질 탐색과 약물 개발 단계로 구성되어 있으나, 하나의 신약을 개발하기 위해서는 10년 이상의 장시간이 요구된다. 이러한 기간을 단축하고 효율적으로 신약 후보 물질을 발굴하기 위하여 심층 학습 기반의 방법들이 연구되고 있다. 많은 심층학습 기반의 모델들은 SMILES 문자열로 표현된 화합물을 재귀신경망을 통해 학습 및 생성하고 있으나, 재귀신경망은 훈련시간이 길고 복잡한 분자식의 규칙을 학습시키기 어려운 단점이 있어서 개선의 여지가 남아있다. 본 연구에서는 self-attention과 variational autoencoder를 활용하여 SMILES 문자열을 생성하는 딥러닝 모델을 제안한다. 제안된 모델은 최신 신약 디자인 모델 대비 훈련 시간을 1/26로 단축하는 것뿐만 아니라 유효한 SMILES를 더 많이 생성하는 것을 확인하였다.

키워드 : 신약 디자인, SMILES, 심층학습, 셀프 어텐션, 변분 오토인코더

1. 서 론

신약 디자인(de novo drug design)은 단백질과 같은 생물학적 표적(biological target)과 상호작용(interaction)할 수 있는 새로운 약물을 개발하는 과정을 의미한다[1]. 전통적인 신약 디자인 연구는 약물 후보 물질 탐색(hit to lead) 및 약물 개발(drug development) 단계로 구성되며, 하나의 신약을 만들기 위해서는 10~20년의 긴 시간이 요구된다[2]. 장기간의 신약 개발 과정을 단축 및 더 효율적인 신약 디자인은 지속적으로 요구되고 있으며 최근에 많은 주목을 받고 있다. 최근 연구에서는 여러 심층 학습(deep learning)을 활용한 신약 디자인 기법이 제안했으며, 실제로 약물 후보 물질 탐색 시간 및 경제적 비용을 많이 감소시킬 수 있음을 보여주었다[2-10].

약물 혹은 화합물을 표현하는 방법에는 여러 가지가 있으며[11-14], 그 중에서 ASCII 문자열을 사용하여 화합물의 분자식을 나타내는 표기법인 simplified molecular-input line-entry system (SMILES)가 많이 사용되고 있다[12-14]. SMILES에서 대소문자 영문 알파벳은 원자(atom)를 나타내고, 특수기호들은 원자 간의 결합(bond) 또는 특수한 화학구조를 표현하는데 사용된다.

예를 들어, 진통제 중 하나인 아스피린(aspirin)의 SMILES은

※ 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2019R1A2C3005212, 딥러닝을 이용한 간암 표적항암제 내성기전 규명 및 이를 극복할 새로운 표적항암제 탐색)과 국토교통부의 스마트시티 혁신인재육성사업의 지원을 받아 수행된 연구임.
※ 이 논문은 2021년 한국정보처리학회 춘계학술발표대회의 우수논문으로 "신약 디자인을 위한 Self-Attention 기반의 SMILES 생성자"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 연세대학교 컴퓨터과학과 석사과정
†† 준 회 원 : 연세대학교 컴퓨터과학과 박사과정
††† 비 회 원 : 연세대학교 컴퓨터과학과 박사과정
†††† 비 회 원 : 연세대학교 컴퓨터과학과 석사과정
††††† 종신회원 : 연세대학교 컴퓨터과학과 교수
Manuscript Received : June 23, 2021
First Revision : August 23, 2021
Accepted : August 26, 2021
* Corresponding Author : Sanghyun Park(sanghyun@yonsei.ac.kr)

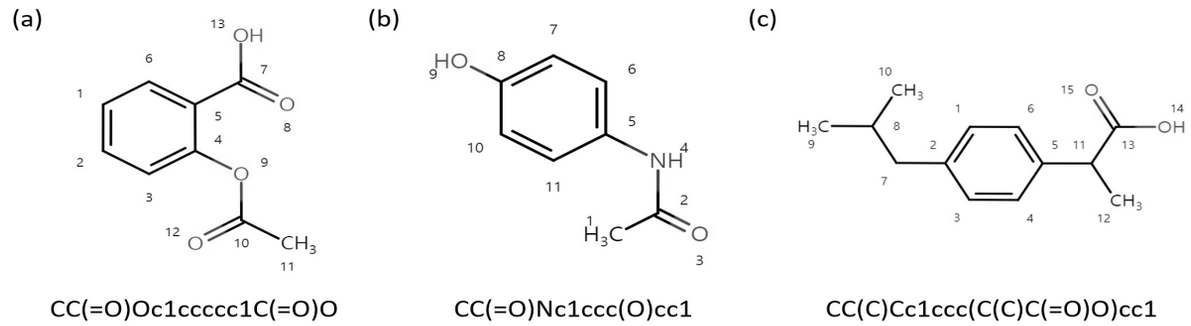

Fig. 1. Molecular Structure and SMILES of (a) Aspirin, (b) Acetaminophen, and (c) Ibuprofen

“CC(=O)Oc1ccccc1C(=O)O”와 같이 표현된다(Fig. 1a). SMILES에는 화합물의 구조적 특징, 분기(branches), 방향족(aromaticity), 고리(rings) 등을 나타내기 위한 여러 가지 규칙들이 존재한다. 대표적으로 분기 규칙은 화합물에서 분기되어 있는 구조를 표현하기 위해서 사용된다. 예를 들어, 아스피린에서 10번 탄소에 결합한 첫 번째 산소의 분기 구조는 “CC(=O)O”와 같이 표기된다(Fig. 1a). 이와 유사하게 아세트아미노펜(acetaminophen)은 2번 탄소에 결합한 첫 번째 산소의 분기 구조를 "CC(=O)N"와 같이 표기하였고, 이부프로펜(ibuprofen)은 11번 탄소에 결합한 12번 탄소의 분기 구조를 “CC(C)C”와 같이 표기하였다(Fig. 1b-c).

SMILES 문자열 생성을 위하여 재귀신경망(recurrent neural network; RNN)에 기반한 심층 생성 신경망 기법들이 제안되었다[2-10]. 대표적으로 ReLeaSE[9] 모델은 stack-augmented RNN (stackRNN)[15]을 사용하여 SMILES 문법에 견고한 SMILES생성 방법을 보여준다. StackRNN은 스택(stack) 자료구조를 활용하는 괄호 검사 알고리즘에서 아이디어를 착안하여, RNN에 스택 역할을 갖는 레이어를 재귀신경망에 추가한 모델이다. 그러나 ReLeaSE는 몇 가지 한계점을 갖고 있다. 첫 번째는 RNN 기반의 모델을 사용해서 훈련 속도를 개선할 여지가 있다. RNN은 문자열을 순차적으로 처리하므로, 훈련 시간이 문자열 길이에 종속적이다. 두 번째는 괄호와 같이 쌍으로 이어지는 SMILES 문법 규칙에는 stackRNN의 스택이 효과적일 수 있지만, 원자가(valence), 결합 종류 등과 같이 스택으로 정보 저장할 수 없는 SMILES 문법 규칙에는 효과적이기 어렵다. 예를 들어, 산소 원자가는 2로 다른 원자들과 최대 2개의 결합을 가질 수 있어서, 2개의 단일 결합(single bond) 또는 1개의 이중 결합(double bond)만을 가질 수 있다. 탄소 원자가는 4로 최대 4개의 결합을 가질 수 있어서, 단일 결합, 이중 결합, 삼중 결합(triple bond) 등으로 여러 가지 조합을 가질 수 있다. 원자가 제약에 따른 결합을 표현하기 위해 SMILES에서는 “-”, “=”, “#”과 같은 결합 기호를 사용하고 있다. Fig. 1a의 아스피린에서 7번 탄소와 8번 산소의 이중 결합이 “C(=O)”와 같이 표기된 것을 볼 수 있다. 적절한 결합 표기를 위해서는 나열된 원자들 간의 관계 정보가 필요하지만, stackRNN의 스택은 이러한 정보를 저장하도록 설계되지 않아, 원자 간의 관계에 기인하는 SMILES 문법 규칙에는 효과적이기 어렵다.

복잡한 SMILES 규칙을 효과적으로 학습하기 위해, 문자열에 나열된 단어들을 서로 비교하여 특징을 추출하는 self-attention 기법이 활용될 수 있다. Self-attention 기법은 자연어처리 분야에서 기계번역 성능을 개선하기 위한 transformer 모델에서 처음 제시된 방법으로, 많은 자연어 처리 영역에서 기존보다 우수한 성능을 보여주었다[16]. Self-attention은 한 문장에 포함된 단어 간의 연관성을 계산하고, 계산된 결과를 바탕으로 단어 별 특징을 추출하는 기법이다. SMILES를 구성하는 원자, 결합 등의 기호를 단어로 취급하면, self-attention이 원자들을 비교하여 원자가에 따른 알맞은 결합 기호를 예측하고, 나아가 나열된 단어들의 전체적인 관계를 통해 괄호의 여닫힘 문제도 해결할 것으로 기대할 수 있다. 또한, self-attention 계산은 병렬적인 처리가 가능하다는 특징이 있어서, 순차적인 계산이 강제되는 RNN을 사용하는 것보다, self-attention 기반의 모델을 사용하는 것이 SMILES 데이터 훈련 속도를 크게 향상시킬 수 있다.

본 연구에서는 SMILES 구성요소 간의 관계를 학습하기 위한 self-attention 기법과 SMILES 데이터의 잠재 분포를 학습할 수 있는 variational autoencoder(VAE) 모델을 접목하여 신약 디자인을 위한 SMILES 생성 모델을 제안한다. 제안하는 모델은 훈련 시간 및 유효한 SMILES 문자열 생성 측면에서 ReLeaSE 보다 우수한 성능을 보여주었고, 통계적인 분석을 통해 제안하는 모델이 SMILES의 규칙을 잘 학습하여 새로운 SMILES를 많이 생성할 수 있는 것을 확인하였다.

2. 관련 연구

2.1 RNN와 신약디자인

SMILES의 순차적인 언어 특성으로 많은 연구들은 RNN 기반의 생성 모델을 사용했다. RNN과 언어 모델(language model)을 결합해서 복잡한 SMILES문법 규칙을 학습 및 새로운 SMILES를 생성한다[3,5,7,9]. 모델이 SMILES의 규칙을 더 잘 학습하기 위해 전통 RNN이 아닌 양방향(bidirectional)

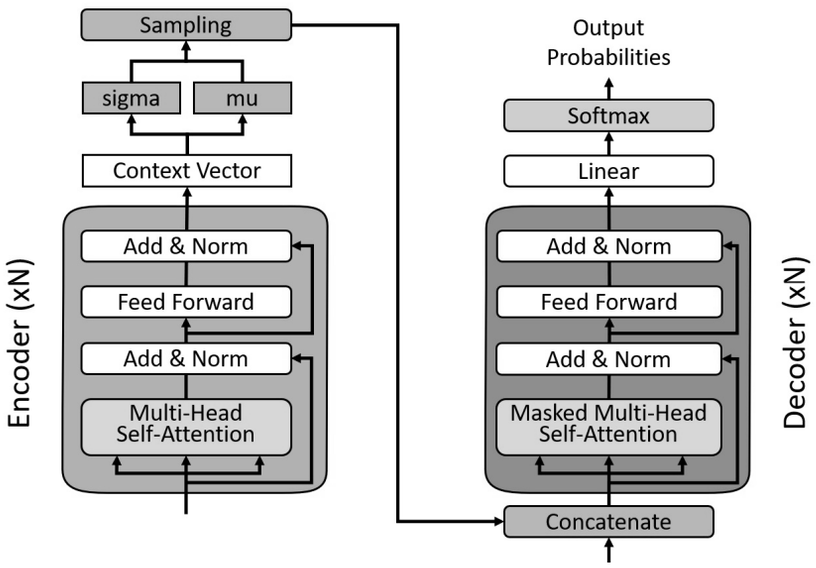

Fig. 2. Model Architecture of Proposed Model

RNN을 사용해서 전방과 후방 두 방향을 통해 SMILES를 분석하는 방법도 있다[7,8]. 또 다른 접근법으로 스택(stack) 자료구조를 이용하여 괄호와 같이 쌍으로 이어지는 규칙에 더 집중적으로 학습하는 모델도 제안되었다[8,9].

2.2 VAE와 신약디자인

VAE는 심층 생성 모델(deep generative model)의 일종으로 확률 모델을 통해 입력 데이터에 대한 잠재공간(latent space)을 학습할 수 있다[17]. VAE에서 잠재공간은 정규분포로 표현되며, 정규분포상에서의 임의 추출을 통해 새로운 데이터를 생성할 수 있다. [4]는 VAE를 사용해서 훈련 데이터와 유사한 성질을 가진 SMILES를 많이 생성했다. 최근 연구에서는 특정 화학적 성질을 만족하는 SMILES을 생성하기 위해 SSVAE기법[18]을 사용해서 SMILES를 생성하거나[6], 두 개의 서로 다른 데이터에서 훈련한 VAE를 결합하여 특정 화학적 성질을 만족하는 SMILES을 생성하는 방법이 제안되고 있다[8]. 또 다른 접근법으로 잠재공간에 대한 페널티를 추가하여 약물에 대한 화학적 성질 또는 생물학적 성질이 반영되도록 제약을 부여하는 Penalized VAE 기법도 제안되었다[10].

3. 모 델

3.1 Self-attention 기법

Self-attention은 3개의 입력, 쿼리 행렬 Q, 키 행렬 K, 값 행렬 V을 이용하여 Equation (1)과 같이 계산된다.

$$Attention(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

Equation (1)에서 $Q, K \in \mathrm{R}^{n \times d_k}$, $V \in \mathrm{R}^{n \times d_v}$, n은 입력 데이터의 최대길이다. Self-attention은 Q와 K의 내적 연산에 softmax 함수를 적용해서 V에 대해 가중치를 계산한다. 가중치는 입력에 포함된 단어들 사이의 연관성으로 해석할 수 있다. Q와 K의 내적 연산 결과가 커질수록 softmax 함수를 적용할 시에 반환되는 가중치가 작아진다. 가중치 값이 작아져서 소실되는 문제를 방지하기 위해 내적 된 결과에 $\sqrt{d_k}$를 나누어 준다.

3.2 VAE 기법

VAE는 생성 모델의 일종으로 훈련 데이터의 확률 분포와 유사한 데이터를 생성하는 것을 목표로 한다. VAE의 구조는 encoder와 decoder로 구성 되어있으며 encoder에 입력된 데이터의 정보를 압축해서 잠재공간으로 사상한다. VAE는 이 잠재공간을 표준정규분포로 가정하고, encoder로 계산된 잠재 벡터의 평균과 표준편차를 사용하여 표준정규분포와 근사한 정규분포를 생성한다. Decoder는 정규분포로부터 임의 추출된 값을 입력 받아서 encoder의 입력 값을 복원한다. 훈련 단계에서는 encoder와 decoder가 동시에 사용되고, 추론 단계에서는 decoder만을 활용하여 새로운 SMILES을 생성한다.

3.3 모델 구조 설명

본 연구에서 제안하는 모델은 일반적인 VAE와 같이 encoder-decoder 구조를 갖는다(Fig. 2). Encoder와 decoder는 각각 N개의 블록으로 구성되며, 각 블록은 2개의 레이어를 포함한다. 첫 번째 레이어는 multi-head self-attention 레이어로, 여러 개의 self-attention을 수행하여 주어진 SMILES의 다양한 특징을 효과적으로 학습 및 추출한다. 두 번째 레이어는 feed-forward 레이어로, 앞선 self-attention 레이어의 결과에 비선형 변환을 적용하여 복잡한 규칙을 더 잘 학습할 수 있도록 하는

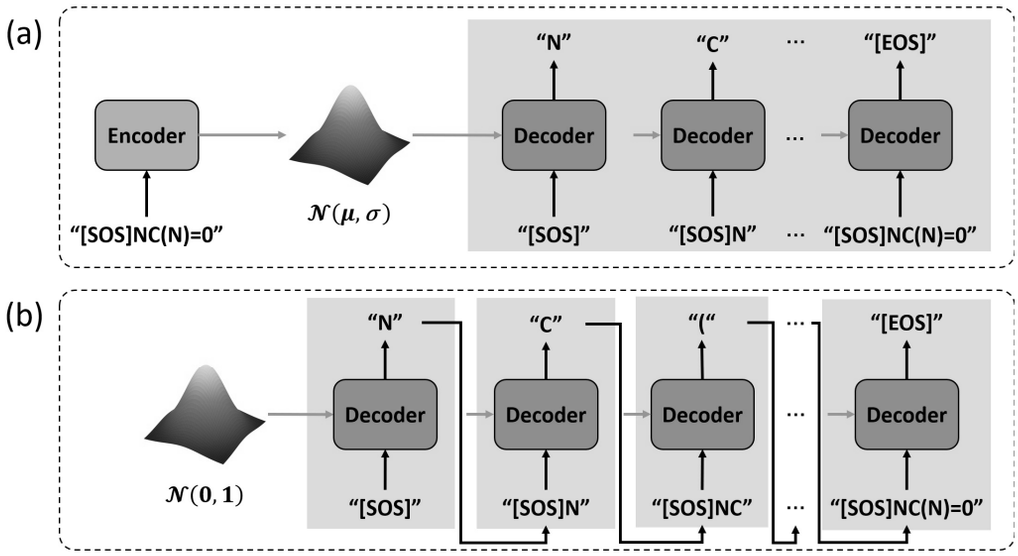

Fig. 3. Overall Process of Model; (a) Training Phase; (b) Inference Phase

역할을 갖는다. 각 레이어는 residual connection[19]과 layer normalization[20] 기법들을 사용해서 모델의 학습 성능을 더 높이도록 설계되었다.

3.4 훈련 단계

Encoder는 입력된 SMILES의 규칙과 화합물의 특성에 대한 정보 혹은 잠재 벡터를 계산한다. 모델은 계산한 잠재 벡터를 통해 입력 데이터 분포의 평균과 표준편차를 구하고, 재매개변수화(reparamterization trick)를 통해 잠재공간으로 사상하는 특징 벡터를 계산한다. 특징 벡터는 encoder에 입력된 SMILES와 연결해서 decoder에 입력되고 이를 기반으로 데이터를 생성한다(Fig. 3a).

제안 모델은 학습한 잠재 공간과 표준정규분포와의 차이를 Kullback-Leibler (KL) divergence로 계산하고, cross-entropy를 사용해서 decoder가 입력 SMILES를 잘 복원했는지를 평가한다. 훈련 단계는 KL divergence 함수와 cross entropy 함수로 구성된 손실함수(loss function)로 역전파(back-propagation)를 수행한다. 손실함수는 Equation (2)와 같다.

$$L = CrossEntropy(x, \hat{x}) + D_{KL}(q(z \mid x) \mid \mid p(z)) \quad (2)$$

Equation (2)에서 x는 encoder의 입력 값이고 $\hat{x}$는 decoder의 출력 값이다. $q(z \mid x)$는 입력 데이터가 사상하는 잠재 공간이고, $p(z)$는 표준정규분포다. 계산된 손실함수를 기반으로 Adam[21] 최적화 알고리즘을 통해 제안 모델의 가중치를 업데이트한다. 본 연구는 교사 강제(teacher-forcing) 기법을 적용하여 모델을 훈련시킨다. 훈련 단계에서 decoder는 먼저 encoder로 받은 특징 벡터를 사용해서 첫 단어를 예측한다. 전통적인 언어 생성 모델은 이 단어를 기반으로 다음 단어를 예측하고 decoder는 이 예측한 단어를 사용해서 또 다음 단어를 예측한다. 즉 훈련 단계에서 decoder는 전 시점에서 예측한 단어를 기반으로 다음 단어를 예측하는 과정을 반복한다. 이와 다르게 교사 강제 기법은 전 시점에서 예측한 단어를 사용하지 않고 예측해야 할 단어 즉 정답을 사용해서 다음 단어를 예측한다[22]. 즉 훈련 단계에서 decoder는 계속 정답을 보면서 다음 단어를 예측한다는 것이다. 이런 기법은 전통적인 방법보다 더 빠른 훈련 속도를 보여주었다[23].

3.5 추론 단계

새로운 SMILES 문자열 생성을 위한 추론 단계는 encoder를 사용하지 않고, 표준정규분포로부터 잠재 벡터를 무작위 추출하여 SMILES를 생성한다(Fig. 3b). 추론 단계에서는 decoder를 반복적으로 실행하여 SMILES를 한 단어씩 생성한다. 생성 과정은 문자열의 끝을 나타내는 [EOS](end of sentence) 기호가 나오거나 혹은 생성된 SMILES의 길이가 최대생성 길이에 도달하면 종료한다.

4. 실험 및 결과

4.1 실험 환경

본 연구는 ubuntu 18.04.5 LTS 서버에서 실험을 진행하였다. 실험 서버의 메모리는 64GB이고, i7-6700k CPU 및 GeForce RTX 3090 GPU가 탑재 되어있다. 본 연구에서 제안하는 모델 및 비교 모델의 훈련 및 성능평가는 모두 상기 환경에서 진행되었다.

4.2 실험 데이터

본 연구는 여러 연구에서[3,5,7-9] 사용한 ChEMBL[24] 데이터베이스에서 제공하는 150만여 개의 SMILES를 벤치마킹

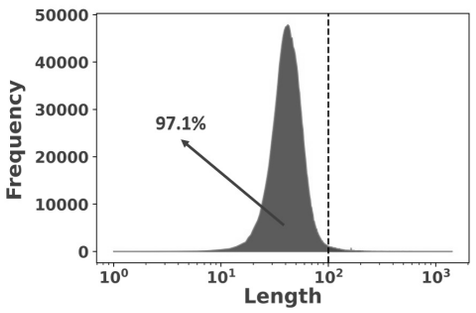

Fig. 4. Distribution of the Length of SMILES in ChEMBL Database

Table 1. Training Time of Proposed Model and ReLesSE

Model	Training time
Ours	**1h 23min**
ReLeaSE	36h 21min

데이터로 사용한다. ChEMBL 데이터베이스는 정기적으로 약물과 유사한 특성을 가진 생체 활성 분자를 수동으로 선별해서 만든 화학 데이터베이스다[25]. ChEMBL 데이터베이스에서 제공하는 SMILES의 최대 길이는 1423이지만, (Fig. 4)와 같이 길이 100 이상의 길이를 가지는 SMILES는 극소수로, 안정적인 모델 훈련을 위해 SMILES의 길이가 100 이하인 데이터만 선별하였다. 훈련데이터는 64개의 SMILES를 갖는 batch 단위로 SMILES 데이터를 학습하도록 훈련 단계를 설계하였다.

4.3 실험 설계

본 연구는 제안 모델의 성능을 평가하기 위해서 두 가지 실험을 진행하였다. 첫 번째로 서론 부분에서 지적하였던 RNN 계열의 한계점을 극복하였는지 확인하기 위하여 ReLeaSE와 훈련 시간 및 유효한 SMILES 생성 효율을 비교하였다. 두 번째로 훈련된 모델이 생성하는 SMILES의 구조적 다양성과 훈련데이터의 화학적 성질 재현성을 분석한다. 구체적으로, 제안 모델로 생성된 SMILES와 훈련 데이터의 SMILES의 분자 유사도와 화학적 유사도를 측정하여 다양성 및 재현성 평가를 진행한다. 분자 유사도는 SMILES의 구조 특성에 대한 유사도를, 화학적 유사도는 SMILES의 물리 화학적 특성에 대한 유사도를 의미한다[26].

4.4 훈련 시간 비교

본 연구는 제안 모델의 훈련 시간 단축 성능을 평가하기 위해 RNN 기반의 대표 신약 디자인 모델인 ReLeaSE와 비교실험을 수행하였다. 150만여 개의 SMILES 데이터에 대한 두 모델의 훈련 시간은 Table 1과 같이, RNN의 순차적 실행으로 인해 36시간 21분이 소요되는 ReLeaSE 와 달리, 본 연구의 제안 모델은 self-attention의 병렬 처리 특성을 통해 훈련 시간을 1시간 23분으로 크게 단축했다. 이는 ReLeaSE의 1/26에 해당하여, 제안 모델의 훈련 속도가 기존 모델보다 빨라진 것을 보여준다.

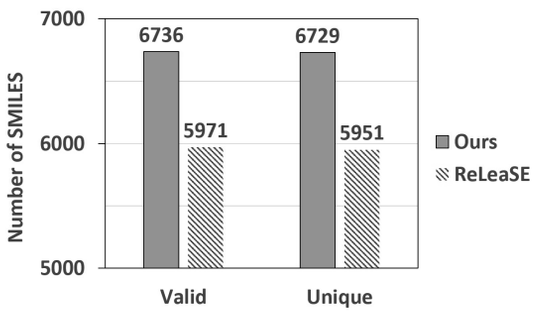

Fig. 5. Number of Valid SMILES and Unique SMILES in the SMILES Generated by Proposed Model and ReLeaSE

4.5 생성 품질 비교

훈련된 모델은 효과적인 신약 후보 물질을 탐색하기 위해, SMILES 규칙 준수 및 다양한 SMILES를 생성할 수 있는 능력을 갖추어야 한다. 제안 모델의 SMILES 생성 품질을 평가하기 위해, 제안 모델과 ReLeaSE로부터 각각 무작위로 1만 개의 SMILES를 생성하였고, 대표적인 화합물 분석도구인 RDKit 오픈소스 소프트웨어를 활용하여 두 모델로부터 생산된 SMILES의 유효성과 다양성을 비교 분석하였다. (Fig. 5)는 1만 개의 생성 결과 비교를 보여준다. ReLeaSE는 1만 개 중 5,971개가 유효한 SMILES이지만, 제안 모델은 ReLeaSE보다 765개 더 많은 6,736개의 유효한 SMILES를 생성하였다. 또한, 유효한 SMILES 중에서 제안 모델은 중복되지 않은 SMILES가 6,729개(67.29%)인 반면에, ReLeaSE는 5,951개(59.51%)로, 제안 모델이 다양성 또한 우수한 것을 보여주었다.

4.6 분자 유사도 분석

제안한 모델이 생성한 SMILES의 다양성을 평가하기 위해 본 연구는 제안 모델로부터 1 만개의 유효한 SMILES를 생성 및 훈련 데이터와의 구조적 유사도를 분석했다. 구조적 유사도를 계산하기 위해 모든 SMILES는 Morgan fingerprints로 변환되었다. Morgan fingerprints는 분자구조의 특징을 표현하는 방법 중의 하나로 화합물에 특정한 구조의 존재 여부를 이진 벡터(binary vector)로 표현하는 방법이다.

타니모토 계수 (Tanimoto coefficient) T는 두 개의 분자식 사이의 구조적 유사도를 측정하는 지표로 fingerprints의 원소와 비교하여 어느 정도 공통된 원소를 가지는지를 계산하여 유사성을 평가하는 방법이다. 타니모토 계수는 0에서 1 사이의 값을 가지며, 일반적으로 0.85보다 크면 두 개 분자식의 구조는 유사하다고 판단한다[27]. 본 연구는 생성한 데이터와 훈련 데이터의 모든 분자식들 간의 타니모토 계수를 계산하고, 그 분포를 조사하여 훈련데이터에 없는 새로운 구조의 SMILES를 얼마나 생성하였는지를 평가하였다.

Fig. 6(a)와 같이 생성한 대부분의 SMILES 은 훈련 데이터와 0.3~0.6 사이의 유사도로 갖고 있다. 예를 들어, Fig. 6(b-d)와 같이 타니모토 계수가 0.44, 0.57, 0.33인 유효한 SMILES가 모델로 생성하는 것을 볼 수 있다. 또한, 0.85 보다

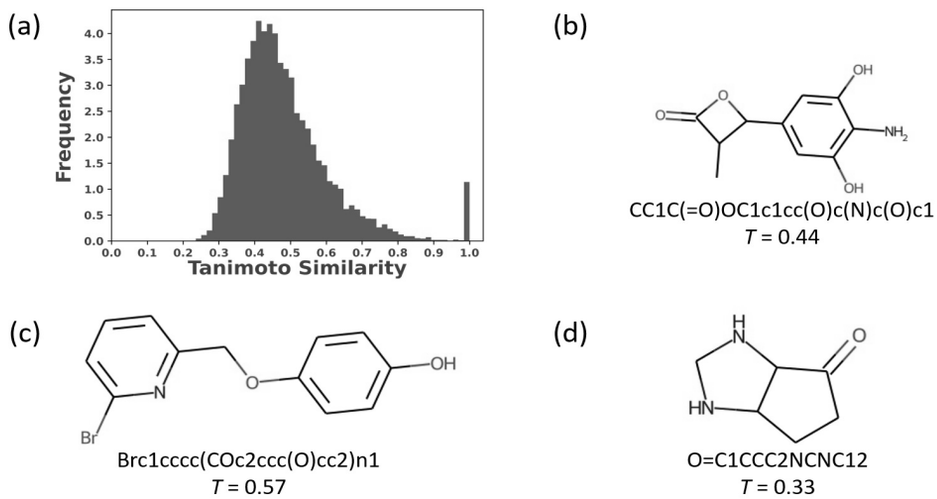

Fig. 6. Diversity Analysis of Proposed Model Compared to ChEMBL Data; (a) the Distribution of Tanimoto Similarity between Training Data and Generated Data; (b-d) Three Examples of Generated Novel Compounds with Corresponding SMILES and Tanimoto Similarity

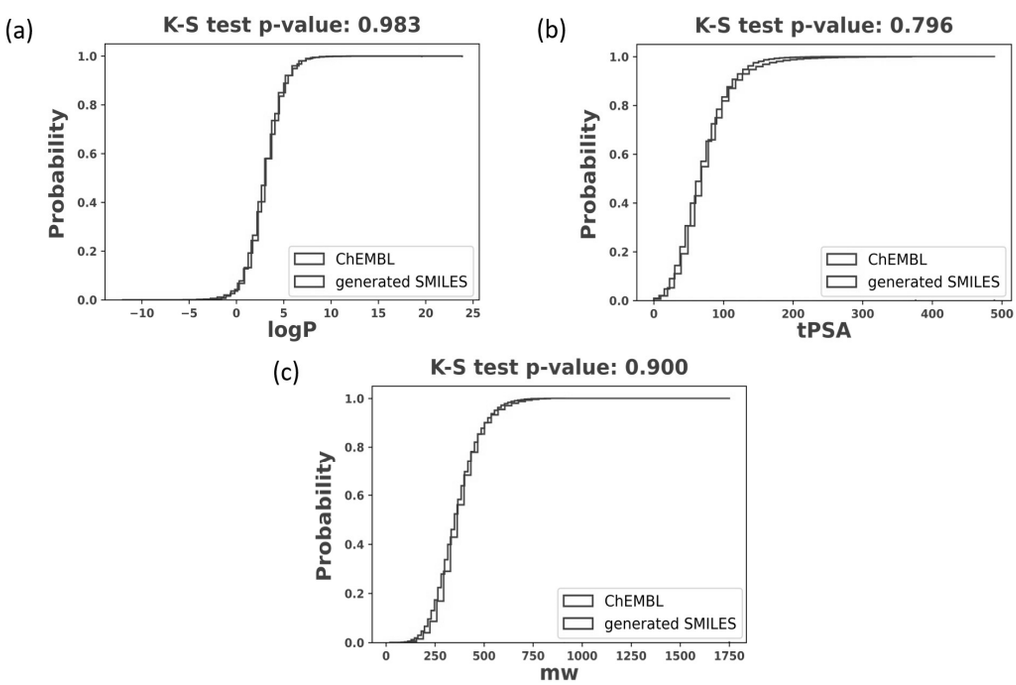

Fig. 7. K-S Test on the CDF of Three Chemical Properties (a) logP, (b) tPSA and (c) mw of Training Data and Generated Data

큰 유사도를 갖는 SMILES도 생성한 것을 Fig. 6(a)에서 확인할 수 있다. 즉, 제안한 모델은 다양한 SMILES를 생성하는 동시에 훈련 데이터와 유사한 SMILES도 생성할 수 있다는 것을 증명했다.

4.7 화학적 유사도 분석

화학적 유사도 분석을 통해 훈련된 모델이 훈련 데이터의 화학적 특성을 잘 학습했는지 평가한다. 본 연구는 Octanol-water partition coefficient (logP), topological polar surface area (tPSA), molecular weight (mw) 세 가지 화학적 특성을 사용해서 모델의 화학적 유사도를 분석했다.

각 화학적 특성에 대한 생성된 SMILES 및 훈련 데이터의 누적확률분포는 Fig. 7(a-c)과 같다. 두 확률분포의 유사성을 평가하기 위해 Kolmogorov-Smirnov 검정 (K-S 검정)으로 통계분석을 진행하였다. K-S 검정은 2개 누적분포가 같은 분포인지를 검정하는 통계 방법이다. 각 특성 분포에 대한 K-S 검정의 p-값은 0.983, 0.796, 0.900이고, 이는 유의 수준 0.05보다 높아서 두 분포가 서로 동일하다는 귀무가설(null hypothesis)을 기각할 수 없다. 따라서 제안한 모델은 훈련 데이터의 화학적 특성을 잘 학습 및 화학적으로도 유효한 SMILES를 생성할 수 있음을 확인하였다.

5. 결 론

본 연구에서는 효과적인 SMILES 규칙 학습을 위한 self-attention 기법과 다양한 SMILES 생성을 위한 VAE 기법을 활용하여 신약 디자인을 위한 SMILES 생성 기법을 제안하였다. 제안하는 모델은 최신 모델인 ReLeaSE보다 더 빠른 훈련 속도 및 더 높은 유효 SMILES 생성율을 보여주었다. 또한, 분자 유사도 분석의 결과와 화학적 유사도 분석의 결과를 통해 제안한 모델이 훈련 데이터의 특성을 잘 학습한 것도 증명하였다.

References

[1] S. K. Jain and A. Agrawal, "De novo drug design: An overview," *Indian Journal of Pharmaceutical Sciences*, Vol.66, No.6, pp.721, 2004.

[2] A. Zhavoronkov, et al., "Deep learning enables rapid identification of potent DDR1 kinase inhibitors," *Nature Biotechnology*, Vol.37, No.9, pp.1038-1040, 2019.

[3] M. Olivecrona, T. Blaschke, O. Engkvist, and H. Chen, "Molecular de-novo design through deep reinforcement learning," *Journal of Cheminformatics*, Vol.9, No.1, pp.1-14, 2017.

[4] R. Gomez-Bombarelli, et al., "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Central Science*, Vol.4, No.2, pp.268-276, 2018.

[5] M. H. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," *ACS Central Science*, Vol.4, No.1, pp.120-131, 2018.

[6] S. Kang and K. Cho, "Conditional molecular design with deep generative models," *Journal of Chemical Information and Modeling*, Vol.59, No.1, pp.43-52, 2018.

[7] F. Grisoni, M. Moret, R. Lingwood, and G. Schneider, "Bidirectional molecule generation with recurrent neural networks," *Journal of Chemical Information and Modeling*, Vol.60, No.3, pp.1175-1183, 2020.

[8] R. Martínez, "PaccMannRL: Designing anticancer drugs from transcriptomic data via reinforcement learning," *arXiv preprint arXiv:1909.05114*, 2019.

[9] M. Popova, O. Isayev, and A. Tropsha, "Deep reinforcement learning for de novo drug design," *Science Advances*, Vol.4, No.7, pp.eaap7885, 2018.

[10] S. Mohammadi, B. O'Dowd, C. Paulitz-Erdmann, and L. Goerlitz, "Penalized variational autoencoder for molecular design," *ChemRxiv. 10.26434/chemrxiv. 7977131,v2*, 2021.

[11] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "InChI, the IUPAC international chemical identifier," *Journal of Cheminformatics*, Vol.7, No.1, pp.1-34, 2015.

[12] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, Vol.28, No.1, pp.31-36, 1988.

[13] D. Weininger, A. Weininger, and J. L. Weininger, "SMILES. 2. Algorithm for generation of unique SMILES notation," *Journal of Chemical Information and Computer Sciences*, Vol.29, No.2, pp.97-101, 1989.

[14] D. Weininger, "SMILES. 3. DEPICT. Graphical depiction of chemical structures," *Journal of Chemical Information and Computer Sciences*, Vol.30, No.3, pp.237-243, 1990.

[15] A. Joulin and T. Mikolov, "Inferring algorithmic patterns with stack-augmented recurrent nets," *arXiv preprint arXiv:1503.01007*, 2015.

[16] A. Vaswani, et al, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[17] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, Vol.2, No.1, pp.1-18, 2015.

[18] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *In Advances in Neural Information Processing Systems*, pp.3581-3589, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

[20] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, Vol.1, No.2, pp.270-280, 1989.

[23] G. A. Bekey and K. Y. Goldberg, eds, "Neural networks in robotics," *Springer Science & Business Media*, Vol.202, 2012.

[24] Bento, A. Patricia, et al., "The ChEMBL bioactivity database: An update," *Nucleic Acids Research*, Vol.42, No.D1, pp.D1083-D1090, 2014.

[25] A. Gaulton, et al., "ChEMBL: A large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, Vol.40, No.D1, pp.D1100-D1107, 2012.

[26] G. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, "Molecular similarity in medicinal chemistry: Miniperspective," *Journal of Medicinal Chemistry*, Vol.57, No.8, pp.3186-3204, 2014.

[27] D. E., Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, and L. E. Weinberger, "Neighborhood behavior: A useful concept for validation of 'molecular diversity' descriptors," *Journal of Medicinal Chemistry*, Vol.39, No.16, pp.3049-3059, 1996.

Piao Shengmin

https://orcid.org/0000-0003-2358-9797
e-mail : sungmin630@yonsei.ac.kr
2016년 Yanbian University Computer Science and Technology(학사)
2020년 ~ 현 재 연세대학교 컴퓨터과학과 석사과정
관심분야 : 기계학습, 심층학습, 자연언어처리

최 종 환

https://orcid.org/0000-0002-8429-4135
e-mail : mathcombio@yonsei.ac.kr
2016년 인천대학교 수학과(이학사)
2018년 인천대학교 컴퓨터공학과(공학석사)
2019년 ~ 현 재 연세대학교 컴퓨터과학과 박사과정
관심분야 : 바이오인포매틱스, 신약디자인, 기계학습, 강화학습, 데이터마이닝

서 상 민

https://orcid.org/0000-0003-4883-3987
e-mail : ssm6410@yonsei.ac.kr
2018년 인천대학교 컴퓨터공학과(학사)
2020년 인천대학교 컴퓨터공학과(공학석사)
2020년 ~ 현 재 연세대학교 컴퓨터과학과 박사과정
관심분야 : 바이오인포매틱스, 신약디자인, 기계학습, 심층학습, 데이터베이스

김 경 훈

https://orcid.org/0000-0001-6250-0199
e-mail : kkh115505@yonsei.ac.kr
2020년 경희대학교 응용수학과, 컴퓨터공학과(학사)
2020년 ~ 현 재 연세대학교 컴퓨터과학과 석사과정
관심분야 : 기계학습, 자연언어처리

박 상 현

https://orcid.org/0000-0002-5196-6193
e-mail : sanghyun@yonsei.ac.kr
1989년 서울대학교 컴퓨터공학과(학사)
1991년 서울대학교 컴퓨터공학과(공학석사)
2001년 UCLA 컴퓨터공학과(공학박사)
2001년 ~ 2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow
2002년 ~ 2003년 포항공과대학교 컴퓨터공학과 조교수
2003년 ~ 2006년 연세대학교 컴퓨터과학과 조교수
2006년 ~ 2011년 연세대학교 컴퓨터과학과 부교수
2011년 ~ 현 재 연세대학교 컴퓨터과학과 교수
관심분야 : 데이터베이스, 데이터마이닝, 바이오인포매틱스, 빅데이터 마이닝 & 기계 학습