



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0015259
(43) 공개일자 2025년02월03일

(51) 국제특허분류(Int. Cl.)

G16C 20/30 (2019.01) G06N 3/08 (2023.01)
G16C 20/20 (2019.01) G16C 20/40 (2019.01)
G16C 20/70 (2019.01) G16C 20/80 (2019.01)

(52) CPC특허분류

G16C 20/30 (2019.02)
G06N 3/08 (2023.01)

(21) 출원번호 10-2023-0096464

(22) 출원일자 2023년07월25일

심사청구일자 2023년07월25일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

김태의

서울특별시 성북구 동소문로44길 17, 101동 1805호

손영한

서울특별시 동대문구 제기로6길 9-3 (제기동, 태양빌리지)301호

(74) 대리인

김준석, 박민욱

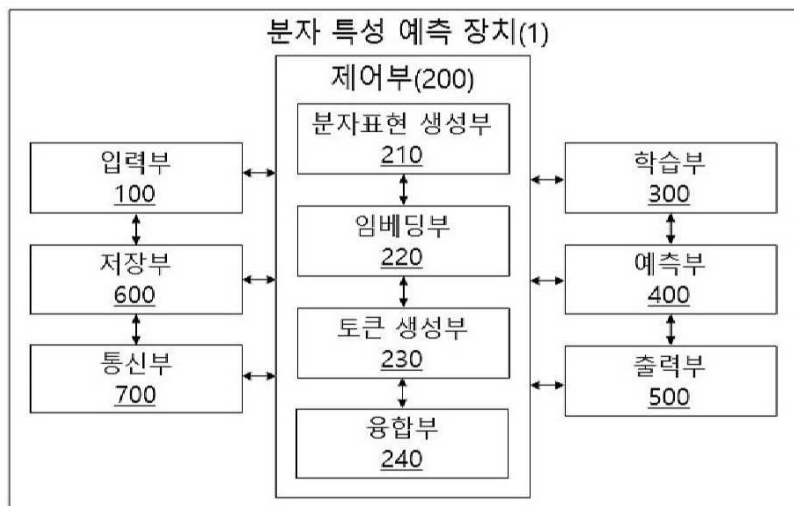
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 분자 특성 예측 장치 및 그 방법

(57) 요약

본 발명의 일 기술적 측면에 따른 분자 특성 예측 장치 및 그 방법에 관한 것으로, 분자표현 생성부가 미리 입력 받은 분자를 기초로 1D 분자표현, 2D 분자표현 및 3D 분자표현을 포함하는 다중 분자표현을 생성하는 단계; 토큰 생성부가 다중 분자표현을 기초로 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰을 포함하는 다중 입력 토큰을 생성하는 단계; 융합부가 다중 입력 토큰 중에서 적어도 하나의 입력 토큰을 제1 모델에 적용하여 복수의 단일 토큰 및 복수의 융합 토큰이 포함된 융합 출력 토큰 세트를 생성하는 단계; 학습부가 융합 출력 토큰 세트의 대조 손실을 연산하고, 대조 손실을 이용하여 조합 손실을 산출하여 조합 손실이 감소하도록 자기지도학습을 수행하는 단계; 및 예측부가 융합 출력 토큰 세트를 제2 모델에 적용하여 최종 분자특성을 예측하는 단계;를 포함한다.

대 표 도 - 도1



(52) CPC특허분류

G16C 20/20 (2019.02)

G16C 20/40 (2019.02)

G16C 20/70 (2019.02)

G16C 20/80 (2019.02)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193973
과제번호	2019-0-00079-005
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능대학원지원(고려대학교)
기 여 율	100/100
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

분자표현 생성부가 미리 입력받은 분자를 기초로 1D 분자표현, 2D 분자표현 및 3D 분자표현을 포함하는 다중 분자표현을 생성하는 단계;

토큰 생성부가 상기 다중 분자표현을 기초로 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰을 포함하는 다중 입력 토큰을 생성하는 단계;

융합부가 상기 다중 입력 토큰 중에서 적어도 하나의 입력 토큰을 제1 모델에 적용하여 복수의 단일 토큰 및 복수의 융합 토큰이 포함된 융합 출력 토큰 세트를 생성하는 단계;

학습부가 상기 융합 출력 토큰 세트의 대조 손실을 연산하고, 상기 대조 손실을 이용하여 조합 손실을 산출하여 상기 조합 손실이 감소하도록 자기지도학습을 수행하는 단계; 및

예측부가 상기 융합 출력 토큰 세트를 제2 모델에 적용하여 최종 분자특성을 예측하는 단계;

를 포함하는 분자 특성 예측 방법.

청구항 2

제1항에 있어서,

상기 융합 출력 토큰 세트를 생성하는 단계는,

상기 융합부가 상기 다중 입력 토큰을 정규화를 통해 상기 다중 입력 토큰 각각의 범위를 스케일링하는 단계;

상기 융합부가 상기 다중 입력 토큰 각각에 대하여 쿼리, 키, 밸류를 연산하는 단계;

상기 융합부가 상기 다중 입력 토큰 각각의 쿼리 및 키의 내적을 통해 상기 1D 입력 토큰, 상기 2D 입력 토큰 및 상기 3D 입력 토큰 사이의 어텐션 스코어를 연산하는 단계;

상기 융합부가 상기 어텐션 스코어, 상기 다중 입력 토큰 각각의 밸류를 이용하여 최종 어텐션 스코어를 연산하는 단계; 및

상기 융합부가 상기 최종 어텐션 스코어를 기초로 정규화를 통해 스케일링을 수행하고, 심층 신경망 및 글로벌 평균 풀링을 통해 상기 융합 출력 토큰 세트를 생성하는 단계;를 포함하는 것을 특징으로 하는 분자 특성 예측 방법.

청구항 3

제2항에 있어서,

상기 어텐션 스코어를 연산하는 단계는,

상기 융합부가 상기 다중 입력 토큰 각각의 쿼리, 키의 내적을 이용하여 상기 1D 입력 토큰의 어텐션 스코어, 상기 2D 입력 토큰의 어텐션 스코어, 상기 3D 입력 토큰의 어텐션 스코어, 상기 1D 입력 토큰과 상기 2D 입력 토큰 사이의 어텐션 스코어, 상기 2D 입력 토큰과 상기 3D 입력 토큰 사이의 어텐션 스코어, 상기 3D 입력 토큰과 상기 1D 입력 토큰 사이의 어텐션 스코어를 연산하는 것을 특징으로 하는 분자 특성 예측 방법.

청구항 4

제3항에 있어서,

상기 최종 어텐션 스코어를 연산하는 단계는,

상기 융합부가 상기 연산된 각각의 어텐션 스코어 및 상기 다중 입력 토큰 각각의 밸류를 이용하여 다음의 수식을 통해 상기 최종 어텐션 스코어를 연산하는 것을 특징으로 하는 분자 특성 예측 방법;

$$A = \text{Softmax}\left(\frac{a}{\sqrt{d}}\right) V$$

여기서, A는 상기 최종 어텐션 스코어이고, a는 어텐션 스코어이며, d는 입력 토큰의 길이이고, V는 밸류를 나타낸다.

청구항 5

제4항에 있어서,

상기 융합 출력 토큰 세트를 생성하는 단계는,

생성하고자 하는 상기 융합 출력 토큰 세트에 해당하는 입력 토큰의 상기 최종 어텐션 스코어를 심층신경망 및 정규화하고 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 상기 융합 출력 토큰 세트를 생성하는 것을 특징으로 하는 분자 특성 예측 방법.

청구항 6

제1항에 있어서,

상기 융합 출력 토큰 세트는,

1D 단일 토큰, 2D 단일 토큰, 3D 단일 토큰을 포함하는 상기 복수의 단일 토큰과, 1D 및 2D 융합 토큰, 2D 및 3D 융합 토큰, 3D 및 1D 융합 토큰을 포함하는 상기 복수의 융합 토큰으로 구성되는 것을 특징으로 하는 분자 특성 예측 방법.

청구항 7

제6항에 있어서,

상기 자기지도학습을 수행하는 단계는,

상기 학습부가 상기 단일 토큰과 상기 단일 토큰 사이의 대조 손실과, 상기 단일 토큰과 상기 융합 토큰 사이의 대조 손실을 연산하는 단계; 및

상기 학습부가 상기 단일 토큰과 단일 토큰 사이의 대조 손실과 상기 단일 토큰과 융합 토큰 사이의 대조 손실의 합인 상기 조합 손실을 산출하는 단계;를 포함하는 것을 특징으로 하는 분자 특성 예측 방법.

청구항 8

제7항에 있어서,

상기 조합 손실을 산출하는 단계는,

다음의 수학적식을 이용하여 상기 조합 손실을 산출하는 것을 특징으로 하는 분자 특성 예측 방법;

$$\begin{aligned} \mathcal{L}_{comb} = & \mathcal{L}_{cl}(z_s, z_t) + \mathcal{L}_{cl}(z_t, z_g) + \mathcal{L}_{cl}(z_g, z_s) \\ & + \mathcal{L}_{cl}(z_s, z_{tg}) + \mathcal{L}_{cl}(z_t, z_{gs}) + \mathcal{L}_{cl}(z_g, z_{st}) \end{aligned}$$

여기서, $\mathcal{L}_{cl}$ 은 대조 손실이고, z_s 은 1D 단일 토큰이며, z_t 은 2D 단일 토큰이고, z_g 은 3D 단일 토큰이며, z_{tg} 은 2D 및 3D 융합 토큰이고, z_{gs} 은 1D 및 3D 융합 토큰이며, z_{st} 은 1D 및 2D 융합 토큰이고, $\mathcal{L}_{comb}$ 는 조합 손실을 나타낸다.

청구항 9

제8항에 있어서,

상기 조합 손실을 산출하는 단계는,

상기 조합 손실이 감소되도록 상기 조합 손실을 산출하는 것을 특징으로 하는 분자 특성 예측 방법.

청구항 10

제1항에 있어서,

임베딩부가 상기 다중 분자표현을 임베딩하는 단계;를 더 포함하는 것을 특징으로 하는 분자 특성 예측 방법.

청구항 11

특성을 예측하고자 하는 분자를 입력받는 입력부;

상기 입력받은 분자를 기초로 1D 분자표현, 2D 분자표현 및 3D 분자표현을 포함하는 다중 분자표현을 생성하는 분자표현 생성부;

상기 다중 분자표현을 기초로 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰을 포함하는 다중 입력 토큰을 생성하는 토큰 생성부;

상기 다중 입력 토큰 중에서 적어도 하나의 입력 토큰을 제1 모델에 적용하여 복수의 단일 토큰 및 복수의 융합 토큰이 포함된 융합 출력 토큰 세트를 생성하는 융합부;

상기 융합 출력 토큰 세트의 대조 손실을 연산하고, 상기 대조 손실을 이용하여 조합 손실을 산출하여 상기 조합 손실이 감소하도록 자기지도학습을 수행하는 학습부; 및

상기 융합 출력 토큰 세트를 제2 모델에 적용하여 최종 분자특성을 예측하는 예측부;

를 포함하는 분자 특성 예측 장치.

청구항 12

제11항에 있어서,

상기 융합부는,

상기 다중 입력 토큰을 정규화를 통해 상기 다중 입력 토큰 각각의 범위를 스케일링하고, 상기 다중 입력 토큰 각각에 대하여 쿼리, 키, 밸류를 연산하며, 상기 다중 입력 토큰 각각의 쿼리 및 키의 내적을 통해 상기 1D 입력 토큰, 상기 2D 입력 토큰 및 상기 3D 입력 토큰 사이의 어텐션 스코어를 연산하고, 상기 어텐션 스코어, 상기 다중 입력 토큰 각각의 밸류를 이용하여 최종 어텐션 스코어를 연산하며, 상기 최종 어텐션 스코어를 기초로 정규화를 통해 스케일링을 수행하고, 심층 신경망 및 글로벌 평균 풀링을 통해 상기 융합 출력 토큰 세트를 생성하는 것을 특징으로 하는 분자 특성 예측 장치.

청구항 13

제12항에 있어서,

상기 융합부는,

상기 다중 입력 토큰 각각의 쿼리, 키의 내적을 통해 상기 1D 입력 토큰의 어텐션 스코어, 상기 2D 입력 토큰의 어텐션 스코어, 상기 3D 입력 토큰의 어텐션 스코어, 상기 1D 입력 토큰과 상기 2D 입력 토큰 사이의 어텐션 스코어, 상기 2D 입력 토큰과 상기 3D 입력 토큰 사이의 어텐션 스코어, 상기 3D 입력 토큰과 상기 1D 입력 토큰 사이의 어텐션 스코어를 연산하는 것을 특징으로 하는 분자 특성 예측 장치.

청구항 14

제13항에 있어서,

상기 융합부는,

상기 연산된 각각의 어텐션 스코어 및 상기 다중 입력 토큰 각각의 밸류를 이용하여 다음의 수학적식을 통해 상기 최종 어텐션 스코어를 연산하는 것을 특징으로 하는 분자 특성 예측 장치;

$$A = \text{Softmax}\left(\frac{a}{\sqrt{d}}\right) V$$

여기서, A는 상기 최종 어텐션 스코어이고, a는 어텐션 스코어이며, d는 입력 토큰의 길이이고, V는 밸류를 나타낸다.

청구항 15

제14항에 있어서,

상기 융합부는,

생성하고자 하는 상기 융합 출력 토큰 세트에 해당하는 입력 토큰의 상기 최종 어텐션 스코어를 심층신경망 및 정규화하고 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 상기 융합 출력 토큰 세트를 생성하는 것을 특징으로 하는 분자 특성 예측 장치.

청구항 16

제11항에 있어서,

상기 융합 출력 토큰 세트는,

1D 단일 토큰, 2D 단일 토큰, 3D 단일 토큰을 포함하는 상기 복수의 단일 토큰과, 1D 및 2D 융합 토큰, 2D 및 3D 융합 토큰, 3D 및 1D 융합 토큰을 포함하는 상기 복수의 융합 토큰으로 구성되는 것을 특징으로 하는 분자 특성 예측 장치.

청구항 17

제16항에 있어서,

상기 학습부는,

상기 단일 토큰과 상기 단일 토큰 사이의 대조 손실 및 상기 단일 토큰과 상기 융합 토큰 사이의 대조 손실을 연산하고, 상기 단일 토큰과 단일 토큰 사이의 대조 손실과 상기 단일 토큰과 융합 토큰 사이의 대조 손실의 합인 상기 조합 손실을 산출하는 것을 특징으로 하는 분자 특성 예측 장치.

청구항 18

제17항에 있어서,

상기 학습부는,

다음의 수학적식을 이용하여 상기 조합 손실을 산출하는 것을 특징으로 하는 분자 특성 예측 장치;

$$\begin{aligned} \mathcal{L}_{comb} = & \mathcal{L}_{cl}(z_s, z_t) + \mathcal{L}_{cl}(z_t, z_g) + \mathcal{L}_{cl}(z_g, z_s) \\ & + \mathcal{L}_{cl}(z_s, z_{tg}) + \mathcal{L}_{cl}(z_t, z_{gs}) + \mathcal{L}_{cl}(z_g, z_{st}) \end{aligned}$$

여기서, $\mathcal{L}_{cl}$ 은 대조 손실이고, z_s 은 1D 단일 토큰이며, z_t 은 2D 단일 토큰이고, z_g 은 3D 단일 토큰이며, z_{tg} 은 2D 및 3D 융합 토큰이고, z_{gs} 은 1D 및 3D 융합 토큰이며, z_{st} 은 1D 및 2D 융합 토큰이고, $\mathcal{L}_{comb}$ 는 조합 손실을 나타낸다.

청구항 19

제18항에 있어서,

상기 학습부는,

상기 조합 손실이 감소되도록 상기 조합 손실을 산출하는 것을 특징으로 하는 분자 특성 예측 장치.

청구항 20

제11항에 있어서,

상기 다중 분자표현을 임베딩하는 임베딩부;를 더 포함하는 것을 특징으로 하는 분자 특성 예측 장치.

발명의 설명

기술 분야

[0001] 본 발명은 분자 특성 예측 장치 및 그 방법에 관한 것이다.

배경 기술

[0002] 분자 특성 예측 (Molecular property prediction)은 분자의 성질을 계산하거나 예측하는 것으로, 컴퓨터 시뮬레이션을 통하여 실제 실험을 수행하지 않고 분자 특성을 예측할 수 있다.

[0003] 분자 특성 예측에는 분자의 다양한 표현 방법들이 사용되는데, 1D 표현, 2D 표현, 3D 표현이 있다.

[0004] 특히, 3D 분자 표현을 획득하는 데에 많은 시간과 비용이 소모되는 문제가 있고, 대부분의 분자 데이터는 3D 분자표현을 포함하고 있지 않은 문제가 있다.

[0005] 한편, 종래의 분자 특성 예측 모델은 단일 분자표현만을 사용하였지만, 두 개 이상의 분자표현을 결합하여 사용하면 단일 분자표현에서 나타나지 않는 정보들이 다중분자 표현을 통하여 나타날 수 있다는 것이 밝혀짐에 따라 다중 분자표현을 사용한 높은 정확도의 분자 특성 예측이 시도되었다.

[0006] 하지만, 서로 다른 분자표현은 이질적인 (heterogeneous) 특성을 가지기 때문에 분자표현의 결합에 방해가 되는 문제점이 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 대한민국 공개특허공보 제10-2022-0112692호 ("원자 구조를 이용한 분자 특성 예측 방법 및 이를 위한 장치", 공개일자: 2022.08.11)

발명의 내용

해결하려는 과제

[0008] 본 발명은 다중 분자표현을 이용하여 융합 트랜스포머를 통해 다중 분자표현을 결합하여 분자 특성을 예측하는 분자 특성 예측 장치 및 그 방법을 제공하고자 한다.

과제의 해결 수단

[0009] 상술한 과제를 해결하기 위하여 분자 특성 예측 장치 및 그 방법이 제공된다.

[0010] 분자 특성 예측 방법은 분자표현 생성부가 미리 입력받은 분자를 기초로 1D 분자표현, 2D 분자표현 및 3D 분자표현을 포함하는 다중 분자표현을 생성하는 단계; 토큰 생성부가 상기 다중 분자표현을 기초로 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰을 포함하는 다중 입력 토큰을 생성하는 단계; 융합부가 상기 다중 입력 토큰 중에서 적어도 하나의 입력 토큰을 제1 모델에 적용하여 복수의 단일 토큰 및 복수의 융합 토큰이 포함된 융합 출력 토큰 세트를 생성하는 단계; 학습부가 상기 융합 출력 토큰 세트의 대조 손실을 연산하고, 상기 대조 손실을 이용하여 조합 손실을 산출하여 상기 조합 손실이 감소하도록 자기지도학습을 수행하는 단계; 및 예측부가 상기 융합 출력 토큰 세트를 제2 모델에 적용하여 최종 분자특성을 예측하는 단계;를 포함한다.

[0011] 일 실시예에서, 상기 융합 출력 토큰 세트를 생성하는 단계는, 상기 토큰 생성부가 상기 다중 입력 토큰을 정규화를 통해 상기 다중 입력 토큰 각각의 범위를 스케일링하는 단계; 상기 토큰 생성부가 상기 다중 입력 토큰 각각에 대하여 쿼리, 키, 밸류를 연산하는 단계; 상기 토큰 생성부가 상기 다중 입력 토큰 각각의 쿼리 및 키의 내적을 통해 상기 1D 입력 토큰, 상기 2D 입력 토큰 및 상기 3D 입력 토큰 사이의 어텐션 스코어를 연산하는 단계; 상기 토큰 생성부가 상기 어텐션 스코어, 상기 다중 입력 토큰 각각의 밸류를 이용하여 최종 어텐션 스코어

를 연산하는 단계; 및 상기 토큰 생성부가 상기 최종 어텐션 스코어를 기초로 정규화를 통해 스케일링을 수행하고, 심층 신경망 및 글로벌 평균 풀링을 통해 상기 융합 출력 토큰 세트를 생성하는 단계;를 포함하는 것을 특징으로 할 수 있다.

[0012] 일 실시예에서, 상기 어텐션 스코어를 연산하는 단계는, 상기 토큰 생성부가 상기 다중 입력 토큰 각각의 쿼리, 키의 내적을 이용하여 상기 1D 입력 토큰의 어텐션 스코어, 상기 2D 입력 토큰의 어텐션 스코어, 상기 3D 입력 토큰의 어텐션 스코어, 상기 1D 입력 토큰과 상기 2D 입력 토큰 사이의 어텐션 스코어, 상기 2D 입력 토큰과 상기 3D 입력 토큰 사이의 어텐션 스코어, 상기 3D 입력 토큰과 상기 1D 입력 토큰 사이의 어텐션 스코어를 연산하는 것을 특징으로 할 수 있다.

[0013] 일 실시예에서, 상기 최종 어텐션 스코어를 연산하는 단계는, 상기 상기 연산된 각각의 어텐션 스코어 및 상기 다중 입력 토큰 각각의 밸류를 이용하여 다음의 수학적식을 통해 상기 최종 어텐션 스코어를 연산하는 것을 특징으로 할 수 있다.

[0014]
$$A = \text{Softmax}\left(\frac{a}{\sqrt{d}}\right) V$$

[0015] 여기서, A는 상기 최종 어텐션 스코어이고, a는 어텐션 스코어이며, d는 입력 토큰의 길이이고, V는 밸류를 나타낸다.

[0016] 일 실시예에서, 상기 융합 출력 토큰 세트를 생성하는 단계는, 생성하고자 하는 상기 융합 출력 토큰 세트에 해당하는 입력 토큰의 상기 최종 어텐션 스코어를 심층신경망 및 정규화하고 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 상기 융합 출력 토큰 세트를 생성하는 것을 특징으로 할 수 있다.

[0017] 일 실시예에서, 상기 융합 출력 토큰 세트는, 1D 단일 토큰, 2D 단일 토큰, 3D 단일 토큰을 포함하는 상기 복수의 단일 토큰과, 1D 및 2D 융합 토큰, 2D 및 3D 융합 토큰, 3D 및 1D 융합 토큰을 포함하는 상기 복수의 융합 토큰으로 구성되는 것을 특징으로 할 수 있다.

[0018] 일 실시예에서, 상기 자기지도학습을 수행하는 단계는, 상기 학습부가 상기 단일 토큰과 상기 단일 토큰 사이의 대조 손실과, 상기 단일 토큰과 상기 융합 토큰 사이의 대조 손실을 연산하는 단계; 및 상기 학습부가 상기 단일 토큰과 단일 토큰 사이의 대조 손실과 상기 단일 토큰과 융합 토큰 사이의 대조 손실의 합인 상기 조합 손실을 산출하는 단계;를 포함하는 것을 특징으로 할 수 있다.

[0019] 일 실시예에서, 상기 조합 손실을 산출하는 단계는, 다음의 수학적식을 이용하여 상기 조합 손실을 산출하는 것을 특징으로 할 수 있다.

$$\begin{aligned} \mathcal{L}_{comb} = & \mathcal{L}_{cl}(z_s, z_t) + \mathcal{L}_{cl}(z_t, z_g) + \mathcal{L}_{cl}(z_g, z_s) \\ & + \mathcal{L}_{cl}(z_s, z_{tg}) + \mathcal{L}_{cl}(z_t, z_{gs}) + \mathcal{L}_{cl}(z_g, z_{st}) \end{aligned}$$

[0020]

[0021] 여기서, $\mathcal{L}_{cl}$ 은 대조 손실이고, z_s 은 1D 단일 토큰이며, z_t 은 2D 단일 토큰이고, z_g 은 3D 단일 토큰이며, z_{tg} 은 2D 및 3D 융합 토큰이고, z_{gs} 은 1D 및 3D 융합 토큰이며, z_{st} 은 1D 및 2D 융합 토큰이고, $\mathcal{L}_{comb}$ 는 조합 손실을 나타낸다.

[0022] 일 실시예에서, 상기 조합 손실을 산출하는 단계는, 상기 조합 손실이 감소되도록 상기 조합 손실을 산출하는 것을 특징으로 할 수 있다.

[0023] 일 실시예에서, 임베딩부가 상기 다중 분자표현을 임베딩하는 단계;를 더 포함하는 것을 특징으로 할 수 있다.

[0024] 분자 특성 예측 장치는 특성을 예측하고자 하는 분자를 입력받는 입력부; 상기 입력받은 분자를 기초로 1D 분자표현, 2D 분자표현 및 3D 분자표현을 포함하는 다중 분자표현을 생성하는 분자표현 생성부; 상기 다중 분자표현을 기초로 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰을 포함하는 다중 입력 토큰을 생성하는 토큰 생성부; 상기 다중 입력 토큰 중에서 적어도 하나의 입력 토큰을 제1 모델에 적용하여 복수의 단일 토큰 및 복수의 융합 토큰이 포함된 융합 출력 토큰 세트를 생성하는 융합부; 상기 융합 출력 토큰 세트의 대조 손실을 연산하고, 상기 대조 손실을 이용하여 조합 손실을 산출하여 상기 조합 손실이 감소하도록 자기지도학습을 수행하는 학습부; 및 상기 융합 출력 토큰 세트를 제2 모델에 적용하여 최종 분자특성을 예측하는 예측부;를 포함한다.

[0025] 일 실시예에서, 상기 융합부는, 상기 다중 입력 토큰을 정규화를 통해 상기 다중 입력 토큰 각각의 범위를 스케일링하고, 상기 다중 입력 토큰 각각에 대하여 쿼리, 키, 밸류를 연산하며, 상기 다중 입력 토큰 각각의 쿼리 및 키의 내적을 통해 상기 1D 입력 토큰, 상기 2D 입력 토큰 및 상기 3D 입력 토큰 사이의 어텐션 스코어를 연산하고, 상기 어텐션 스코어, 상기 다중 입력 토큰 각각의 밸류를 이용하여 최종 어텐션 스코어를 연산하며, 상기 최종 어텐션 스코어를 기초로 정규화를 통해 스케일링을 수행하고, 심층 신경망 및 글로벌 평균 풀링을 통해 상기 융합 출력 토큰 세트를 생성하는 것을 특징으로 할 수 있다.

[0026] 일 실시예에서, 상기 융합부는, 상기 다중 입력 토큰 각각의 쿼리, 키의 내적을 통해 상기 1D 입력 토큰의 어텐션 스코어, 상기 2D 입력 토큰의 어텐션 스코어, 상기 3D 입력 토큰의 어텐션 스코어, 상기 1D 입력 토큰과 상기 2D 입력 토큰 사이의 어텐션 스코어, 상기 2D 입력 토큰과 상기 3D 입력 토큰 사이의 어텐션 스코어, 상기 3D 입력 토큰과 상기 1D 입력 토큰 사이의 어텐션 스코어를 연산하는 것을 특징으로 할 수 있다.

[0027] 일 실시예에서, 상기 융합부는, 상기 연산된 각각의 어텐션 스코어 및 상기 다중 입력 토큰 각각의 밸류를 이용하여 다음의 수학적식을 통해 상기 최종 어텐션 스코어를 연산하는 것을 특징으로 할 수 있다.

$$A = \text{Softmax}\left(\frac{a}{\sqrt{d}}\right) V$$

[0028] 여기서, A는 상기 최종 어텐션 스코어이고, a는 어텐션 스코어이며, d는 입력 토큰의 길이이고, V는 밸류를 나타낸다.

[0030] 일 실시예에서, 상기 융합부는, 생성하고자 하는 상기 융합 출력 토큰 세트에 해당하는 입력 토큰의 상기 최종 어텐션 스코어를 심층신경망 및 정규화하고 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 상기 융합 출력 토큰 세트를 생성하는 것을 특징으로 할 수 있다.

[0031] 일 실시예에서, 상기 융합 출력 토큰 세트는, 1D 단일 토큰, 2D 단일 토큰, 3D 단일 토큰을 포함하는 상기 복수의 단일 토큰과, 1D 및 2D 융합 토큰, 2D 및 3D 융합 토큰, 3D 및 1D 융합 토큰을 포함하는 상기 복수의 융합 토큰으로 구성되는 것을 특징으로 할 수 있다.

[0032] 일 실시예에서, 상기 학습부는, 상기 단일 토큰과 상기 단일 토큰 사이의 대조 손실 및 상기 단일 토큰과 상기 융합 토큰 사이의 대조 손실을 연산하고, 상기 단일 토큰과 단일 토큰 사이의 대조 손실과 상기 단일 토큰과 융합 토큰 사이의 대조 손실의 합인 상기 조합 손실을 산출하는 것을 특징으로 할 수 있다.

[0033] 일 실시예에서, 상기 학습부는, 다음의 수학적식을 이용하여 상기 조합 손실을 산출하는 것을 특징으로 할 수 있다.

$$\begin{aligned} \mathcal{L}_{comb} = & \mathcal{L}_{cl}(z_s, z_t) + \mathcal{L}_{cl}(z_t, z_g) + \mathcal{L}_{cl}(z_g, z_s) \\ & + \mathcal{L}_{cl}(z_s, z_{tg}) + \mathcal{L}_{cl}(z_t, z_{gs}) + \mathcal{L}_{cl}(z_g, z_{st}) \end{aligned}$$

[0034] 여기서, $\mathcal{L}_{cl}$ 은 대조 손실이고, z_s 은 1D 단일 토큰이며, z_t 은 2D 단일 토큰이고, z_g 은 3D 단일 토큰이며, z_{tg} 은 2D 및 3D 융합 토큰이고, z_{gs} 은 1D 및 3D 융합 토큰이며, z_{st} 은 1D 및 2D 융합 토큰이고, $\mathcal{L}_{comb}$ 은 조합 손실을 나타낸다.

[0036] 일 실시예에서, 상기 학습부는, 상기 조합 손실이 감소되도록 상기 조합 손실을 산출하는 것을 특징으로 할 수 있다.

[0037] 일 실시예에서, 상기 다중 분자표현을 임베딩하는 임베딩부;를 더 포함하는 것을 특징으로 할 수 있다.

발명의 효과

[0038] 본 발명의 일 실시형태에 따르면, 3D 분자표현 생성기를 이용하여 3D 분자표현을 생성할 수 있고, 다중 분자표현을 이용하여 대조학습 기반의 자기 지도학습을 통해 다중 분자표현의 이질성을 줄일 수 있다.

도면의 간단한 설명

- [0039] 도 1은 본 발명의 일 실시예에 따른 분자 특성 예측 장치를 설명하기 위한 블록도이다.
- 도 2는 분자표현을 설명하기 위한 도면이다.
- 도 3 및 도 4는 융합 트랜스포머 모델을 설명하기 위한 도면이다.
- 도 5는 어텐션 스코어를 설명하기 위한 도면이다.
- 도 6은 조합 손실을 이용한 자기지도학습을 설명하기 위한 도면이다.
- 도 7은 본 발명의 일 실시예에 따른 분자 특성 예측 방법을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0040] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.
- [0041] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 본 발명에 대해 구체적으로 설명하기로 한다.
- [0042] 본 발명에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.
- [0043] 명세서 전체에서 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다. 또한, 명세서에서 사용되는 "부", "모듈", "유닛" 등의 용어는 적어도 하나의 기능 또는 동작을 처리하는 단위를 의미하며, 소프트웨어, FPGA 또는 ASIC과 같은 하드웨어 구성요소, 또는 소프트웨어와 하드웨어의 결합으로 구현될 수 있다. 그렇지만 "부", "모듈", "유닛" 등의 용어가 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. "부", "모듈", "유닛" 등은 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 "부", "모듈", "유닛" 등의 용어는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다.
- [0044] 아래에서는 첨부한 도면을 참조하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략한다.
- [0045] "제1", "제2" 등과 같이 서수를 포함하는 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 구성 요소들은 용어들에 의해 한정되지는 않는다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. "및/또는" 이라는 용어는 복수의 관련된 항목들의 조합 또는 복수의 관련된 항목들 중의 어느 하나의 항목을 포함한다.
- [0047] 이하, 도면을 참조하여 본 발명의 일 실시예에 따른 분자 특성 예측 장치에 대하여 설명하도록 한다.
- [0048] 도 1은 본 발명의 일 실시예에 따른 분자 특성 예측 장치를 설명하기 위한 블록도이다.
- [0049] 도 1에서 도시한 바와 같이, 본 발명의 일 실시예에 따른 분자 특성 예측 장치(1)는 입력부(100), 제어부(200), 학습부(300), 예측부(400), 출력부(500), 저장부(600) 및 통신부(700)를 포함한다.
- [0050] 입력부(100)는 특성을 예측하고자 하는 분자를 입력받는다.

- [0051] 예를 들어, 입력부(100)는 카페인 분자를 입력받을 수 있다.
- [0052] 이때, 입력부(100)는 사용자가 소지하고 있는 사용자 단말 등을 통해 입력받을 수 있다.
- [0053] 여기서, 사용자 단말은 휴대가 가능한 사용자 단말기로서, 노트북 컴퓨터, 스마트패드 또는 스마트폰 등과 같이 유선 또는 무선으로 네트워크에 접속하여 정보를 주고받을 수 있는 기기로 구현되며, 입력부(100)는 유무선으로 연결된 사용자 단말을 통해 분자를 입력받을 수 있다.
- [0054] 제어부(200)는 분자표현 생성부(210), 임베딩부(220), 토큰 생성부(230) 및 융합부(240)를 포함한다.
- [0055] 분자표현 생성부(210)는 미리 입력받은 분자를 기초로 1D 분자표현, 2D 분자표현 및 3D 분자표현을 포함하는 다중 분자표현을 생성한다.
- [0056] 도 2는 분자표현을 설명하기 위한 도면이다.
- [0057] 예를 들어, 도 2에서 도시한 바와 같이, 분자표현 생성부(210)는 입력받은 카페인에 대한 분자를 1D 분자표현, 2D 분자표현, 3D 분자표현으로 생성할 수 있다.
- [0058] 이때, 분자표현 생성부(210)는 1D 분자표현을 SMILES 형식으로 된 문자열 데이터를 이용하여 생성할 수 있고, 2D 분자표현을 각 원자를 그래프의 노드로 하여 각 원자 간 결합을 그래프의 엣지로 만들어 분자를 그래프화하여 생성할 수 있으며, 3D 분자표현을 공간 정보를 이용하여 각 원자의 위치정보를 노드로 하여 원자 간 거리를 엣지로 하는 3D 공간 그래프를 생성할 수 있다.
- [0059] 이때, 3D 분자표현을 생성하기 전, 분자표현 생성부(210)는 각 원자의 3D 공간 위치가 없는 경우에 각 원자의 공간 정보에 대하여 3D 공간 정보 예측을 수행할 수 있다.
- [0060] 여기서, 분자표현 생성부(210)는 3D 공간 정보 예측하기 위해 오픈 소스 툴인 RDKit의 ETKDG를 사용하여 디퓨전 방식의 딥러닝 프레임워크를 이용할 수 있다.
- [0061] 또한, 분자표현 생성부(210)는 Candence 사의 OMEGA, MIT 팀이 개발한 딥러닝을 이용한 Torsional Diffusion 모델, MILA 연구진이 개발한 딥러닝을 이용한 GeoDiff 모델을 통해 3D 분자 구조를 예측할 수 있다.
- [0062] 즉, 분자표현 생성부(210)는 입력받은 분자를 1D 분자표현, 2D 분자표현 및 3D 분자표현으로 생성할 수 있다.
- [0063] 임베딩부(220)는 생성된 다중 분자표현을 각각의 분자표현에 맞는 임베딩 네트워크를 통해 임베딩을 수행한다.
- [0064] 더욱 자세하게는, 임베딩부(220)는 생성된 1D 분자표현을 1D 임베딩 네트워크를 통해 임베딩을 수행하고, 생성된 2D 분자표현을 2D 임베딩 네트워크를 통해 임베딩을 수행하며, 생성된 3D 분자표현을 3D 임베딩 네트워크를 통해 임베딩을 수행한다.
- [0065] 이때, 1D 임베딩 네트워크는 자연어 데이터를 처리하는 네트워크로, 장단기 메모리 신경망(LSTM), 양방향 장단기 신경망(Bi-LSTM), BERT 등을 포함할 수 있다.
- [0066] 또한, 2D 임베딩 네트워크는 그래프 신경망 모델(Graph Neural Network, GNN), 그래프 합성곱 신경망(Graph Convolutional Network, GCN), 그래프 어텐션 네트워크(Graph Attention Network, GAT), 메시지 패싱 신경망(Message Passing Neural Network, MPNN), 그래프 동형사상 네트워크(Graph Isomorphism Network, GIN) 모델 등을 포함할 수 있다.
- [0067] 또한, 3D 임베딩 네트워크는 구형 메시지 패싱 네트워크(Spherical Message Passing Network, SMP), 분극 가능 원자 상호작용 신경망(Polarizable atom interaction Neural Network, PaiNN) 등을 포함할 수 있으며, 3D 공간 그래프 데이터를 임베딩 할 수 있는 네트워크로 구현될 수 있다.
- [0068] 즉, 임베딩부(220)는 다중 분자표현을 각각의 분자표현에 적합한 임베딩 네트워크를 이용하여 임베딩을 수행할 수 있다.
- [0069] 토큰 생성부(230)는 임베딩된 다중 분자표현을 기초로 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰을 포함하는 다중 입력 토큰을 생성한다.
- [0070] 토큰 생성부(230)는 임베딩된 다중 분자표현을 이용하여 제1 모델에 입력하기 위한 다중 입력 토큰을 생성할 수 있다.
- [0071] 이때, 다중 입력 토큰은 1D 입력 토큰, 2D 입력 토큰, 3D 입력 토큰을 포함한다.

[0072] 토큰 생성부(230)는 임베딩된 다중 분자표현을 이용하여 다음의 수학적 식 1을 통해 다중 분자표현 각각에 대한 다중 입력 토큰을 생성할 수 있다.

수학적 식 1

$$\tau = LN(MLP(h))$$

[0073]

[0074] 여기서, h 는 임베딩된 다중 분자표현이고, τ 는 입력 토큰이며, MLP는 다층 퍼셉트론이고, LN은 레이어 정규화를 나타낸다.

[0075] 이때, 정규화와 다층 퍼셉트론을 이용한 방법을 토큰 투영(Token projection)이라 정의하고, 토큰 투영은 이질적인 임베딩을 공통의 공간으로 투영시킬 수 있다.

[0076] 즉, 토큰 생성부(230)는 임베딩된 다중 분자표현인 1D 분자표현, 2D 분자표현 및 3D 분자표현 각각에 대하여 토큰 투영을 통해 1D 입력 토큰, 2D 입력 토큰, 3D 입력 토큰을 포함하는 다중 입력 토큰을 생성할 수 있다.

[0077] 융합부(240)는 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 적어도 하나의 입력 토큰을 제1 모델에 적용하여 복수의 단일 토큰 및 복수의 융합 토큰이 포함된 융합 출력 토큰 세트를 생성한다.

[0078] 이때, 융합 출력 토큰 세트는 복수의 단일 토큰과 복수의 융합 토큰으로 구성되는 것으로, 단일 토큰은 1D 단일 토큰, 2D 단일 토큰, 3D 단일 토큰을 포함하고, 융합 토큰은 1D 및 2D 융합 토큰, 2D 및 3D 융합 토큰, 3D 및 1D 융합 토큰을 포함한다.

[0079] 도 3 및 도 4는 융합 트랜스포머 모델을 설명하기 위한 도면이다.

[0080] 융합부(240)는 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 1개의 입력 토큰을 제1 모델에 입력하여 단일 토큰을 출력할 수 있다.

[0081] 도 3에서 도시한 바와 같이, 융합부(240)는 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 1D 입력 토큰(τ_s)을 제1 모델에 입력하여 1D 단일 토큰(Z_s)을 출력하고, 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 2D 입력 토큰(τ_t)을 제1 모델에 입력하여 2D 단일 토큰(Z_t)을 출력하며, 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 3D 입력 토큰(τ_g)을 제1 모델에 입력하여 3D 단일 토큰(Z_g)을 출력할 수 있다.

[0082] 또한, 융합부(240)는 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 2개의 입력 토큰을 제1 모델에 입력하여 융합 토큰을 생성할 수 있다.

[0083] 이때, 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 선택된 2개의 입력 토큰은 컨кат(Concatenation)되는 것으로, 다음의 수학적 식 2와 같이 나타낼 수 있다.

수학적 식 2

$$\tau_s \oplus \tau_t = \tau_{st}$$

[0084]

[0085] 도 3과 같이, 융합부(240)는 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 1D 입력 토큰 및 2D 입력 토큰(τ_s , τ_t)을 제1 모델에 입력하여 1D 및 2D 융합 토큰(Z_{st})을 출력하고, 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 2D 입력 토큰 및 3D 입력 토큰(τ_{tg})을 제1 모델에 입력하여 2D 및 3D 융합 토큰(Z_{tg})을 출력하며, 생성된 다중 입력 토큰(τ_s , τ_t , τ_g) 중에서 3D 입력 토큰 및 1D 입력 토큰(τ_{gs})을 제1 모델에 입력하여 3D 및 1D 융합 토큰(Z_{gs})을 출력할 수 있다.

[0086] 더욱 자세하게는, 도 4에서 도시한 바와 같이, 융합부(240)는 토큰 생성부(230)로부터 생성된 다중 입력 토큰을 정규화를 통해 다중 입력 토큰 각각의 범위를 스케일링(Scaling)한다.

[0087] 융합부(240)는 다중 입력 토큰 각각에 대하여 쿼리(Query), 키(Key), 밸류(Value)를 다음의 수학식 3 내지 수학식 5를 통해 연산할 수 있다.

수학식 3

[0088]
$$Q = \tau * W_q$$

[0089] 여기서, Q는 쿼리이고, τ 는 입력 토큰이며, W_q 는 쿼리(Q)를 도출하기 위한 가중치를 나타낸다.

[0090] 예를 들어, 융합부(240)는 1D 입력 토큰(τ_s)과 1D 쿼리(Q_1)를 도출하기 위한 가중치(W_q)를 이용하여 1D 쿼리(Q_1)를 연산하고, 2D 입력 토큰(τ_l)과 2D 쿼리(Q_2)를 도출하기 위한 가중치(W_q)를 이용하여 2D 쿼리(Q_2)를 연산하며, 3D 입력 토큰(τ_g)과 3D 쿼리(Q_3)를 도출하기 위한 가중치(W_q)를 이용하여 3D 쿼리(Q_3)를 연산할 수 있다.

수학식 4

[0091]
$$K = \tau * W_k$$

[0092] 여기서, K는 키이고, W_k 는 키(K)를 도출하기 위한 가중치를 나타낸다.

[0093] 예를 들어, 융합부(240)는 1D 입력 토큰(τ_s)과 1D 키(K_1)를 도출하기 위한 가중치(W_k)를 이용하여 1D 키(K_1)를 연산하고, 2D 입력 토큰(τ_l)과 2D 키(K_2)를 도출하기 위한 가중치(W_k)를 이용하여 2D 키(K_2)를 연산하며, 3D 입력 토큰(τ_g)과 3D 키(K_3)를 도출하기 위한 가중치(W_k)를 이용하여 3D 키(K_3)를 연산할 수 있다.

수학식 5

[0094]
$$V = \tau * W_v$$

[0095] 여기서, V는 밸류이고, W_v 는 밸류(V)를 도출하기 위한 가중치를 나타낸다.

[0096] 예를 들어, 융합부(240)는 1D 입력 토큰(τ_s)과 1D 밸류(V_1)를 도출하기 위한 가중치(W_v)를 이용하여 1D 밸류(V_1)를 연산하고, 2D 입력 토큰(τ_l)과 2D 밸류(V_2)를 도출하기 위한 가중치(W_v)를 이용하여 2D 밸류(V_2)를 연산하며, 3D 입력 토큰(τ_g)과 3D 밸류(V_3)를 도출하기 위한 가중치(W_v)를 이용하여 3D 밸류(V_3)를 연산할 수 있다.

[0097] 즉, 융합부(240)는 정규화된 1D 입력 토큰에 대하여 1D 쿼리(Q_1), 1D 키(K_1), 1D 밸류(V_1)를 연산하고, 정규화된 2D 입력 토큰에 대하여 2D 쿼리(Q_2), 2D 키(K_2), 2D 밸류(V_2)를 연산하며, 정규화된 3D 입력 토큰에 대하여 3D 쿼리(Q_3), 3D 키(K_3), 3D 밸류(V_3)를 연산할 수 있다.

[0098] 융합부(240)는 연산된 다중 입력 토큰 각각의 쿼리와 키의 내적을 통해 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰 사이의 어텐션 스코어(Attention score)를 연산한다.

[0099] 융합부(240)는 다중 입력 토큰 각각의 쿼리, 키의 내적을 통해 1D 입력 토큰의 어텐션 스코어, 2D 입력 토큰의 어텐션 스코어, 3D 입력 토큰의 어텐션 스코어, 1D 입력 토큰과 2D 입력 토큰 사이의 어텐션 스코어, 2D 입력 토큰과 3D 입력 토큰 사이의 어텐션 스코어, 3D 입력 토큰과 1D 입력 토큰 사이의 어텐션 스코어를 연산할 수 있다.

[0100] 이때, 어텐션 스코어는 각 입력 토큰이 서로 얼마나 연관되어 있는지에 대하여 점수를 부여하는 것으로, 어텐션

스코어가 크다는 것은 2개의 입력 토큰이 서로를 중요하다고 인식하는 정도가 크다는 것을 의미한다.

[0101] 융합부(240)는 다음의 수학식 6을 이용하여 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰 사이의 어텐션 스코어(Attention score)를 연산할 수 있다.

수학식 6

[0102]
$$a_{nm} = Q_n * K_m$$

[0103] 여기서, a는 어텐션 스코어이고, Q는 쿼리이며, K는 키이고, n 및 m은 상수(1부터 3까지)를 나타낸다.

[0104] 도 5는 어텐션 스코어를 설명하기 위한 도면이다.

[0105] 예를 들어, 도 5에서 도시한 바와 같이, 융합부(240)는 1D 쿼리(Q₁)와 1D 키(K₁)의 내적을 통해 1D 입력 토큰의 어텐션 스코어(a₁₁)를 연산할 수 있고, 1D 쿼리(Q₁)와 2D 키(K₂)의 내적을 통해 1D 입력 토큰과 2D 입력 토큰 사이의 어텐션 스코어(a₁₂)를 연산할 수 있으며, 1D 쿼리(Q₁)와 3D 키(K₃)의 내적을 통해 1D 입력 토큰과 3D 입력 토큰 사이의 어텐션 스코어(a₁₃)를 연산할 수 있다.

[0106] 또한, 융합부(240)는 2D 쿼리(Q₂)와 1D 키(K₁)의 내적을 통해 2D 입력 토큰 및 1D 입력 토큰 사이의 어텐션 스코어(a₂₁)를 연산할 수 있고, 2D 쿼리(Q₂)와 2D 키(K₂)의 내적을 통해 2D 입력 토큰의 어텐션 스코어(a₂₂)를 연산할 수 있으며, 2D 쿼리(Q₂)와 3D 키(K₃)의 내적을 통해 2D 입력 토큰과 3D 입력 토큰 사이의 어텐션 스코어(a₂₃)를 연산할 수 있다.

[0107] 또한, 융합부(240)는 3D 쿼리(Q₃)와 1D 키(K₁)의 내적을 통해 3D 입력 토큰 및 1D 입력 토큰 사이의 어텐션 스코어(a₃₁)를 연산할 수 있고, 3D 쿼리(Q₃)와 2D 키(K₂)의 내적을 통해 3D 입력 토큰 및 2D 입력 토큰 사이의 어텐션 스코어(a₃₂)를 연산할 수 있으며, 3D 쿼리(Q₃)와 3D 키(K₃)의 내적을 통해 3D 입력 토큰의 어텐션 스코어(a₃₃)를 연산할 수 있다.

[0108] 이와 같은 방법으로, 융합부(240)는 쿼리와 키의 내적을 통해 1D 입력 토큰과 2D 입력 토큰 사이의 어텐션 스코어, 2D 입력 토큰과 3D 입력 토큰 사이의 어텐션 스코어, 3D 입력 토큰과 1D 입력 토큰 사이의 어텐션 스코어를 연산할 수 있다.

[0109] 이때, 어텐션 스코어는 다른 분자표현인 1D 입력 토큰과 2D 입력 토큰 사이, 2D 입력 토큰과 3D 입력 토큰 사이 또는 3D 입력 토큰과 1D 입력 토큰 사이에서 연산될 수도 있고, 같은 분자표현인 1D 입력 토큰, 2D 입력 토큰 또는 3D 입력 토큰 각각의 내부에서도 연산이 될 수 있다.

[0110] 즉, 융합부(240)는 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰 서로 간의 어텐션 스코어를 연산함으로써 서로를 얼마나 반영할지 결정할 수 있다.

[0111] 또한, 융합부(240)는 연산된 어텐션 스코어, 다중 입력 토큰 각각의 밸류를 이용하여 다음의 수학식 7을 통해 최종 어텐션 스코어를 연산할 수 있다.

[0112] 이때, 최종 어텐션 스코어는 다중 입력 토큰 각각에 대한 최종 어텐션 스코어일 수 있고, 입력 토큰과 입력 토큰 사이의 최종 어텐션 스코어일 수 있다.

수학식 7

[0113]
$$A = \text{Softmax}\left(\frac{a}{\sqrt{d}}\right) V$$

[0114] 여기서, A는 최종 어텐션 스코어이고, a는 어텐션 스코어이며, d는 입력 토큰의 길이이고, V는 밸류를 나타낸다.

- [0115] 이때, 입력 토큰의 길이(d)는 각 데이터의 길이로 분자 한개에 포함된 원소의 개수를 나타낸다.
- [0116] 예를 들어, 융합부(240)는 1D 입력 토큰의 어텐션 스코어(a_{11}) 및 1D 입력 토큰의 밸류(V_1)를 이용하여 1D 입력 토큰의 최종 어텐션 스코어(A_{11})를 연산할 수 있고, 1D 입력 토큰과 2D 입력 토큰 사이의 어텐션 스코어(a_{12}) 및 1D 입력 토큰의 밸류(V_1)를 이용하여 1D 입력 토큰 및 2D 입력 토큰의 최종 어텐션 스코어(A_{12})를 연산할 수 있으며, 1D 입력 토큰과 3D 입력 토큰 사이의 어텐션 스코어(a_{13}) 및 1D 입력 토큰의 밸류(V_1)를 이용하여 1D 입력 토큰 및 3D 입력 토큰의 최종 어텐션 스코어(A_{13})를 연산할 수 있다.
- [0117] 또한, 융합부(240)는 2D 입력 토큰 및 1D 입력 토큰 사이의 어텐션 스코어(a_{21}) 및 2D 입력 토큰의 밸류(V_2)를 이용하여 2D 입력 토큰 및 1D 입력 토큰의 최종 어텐션 스코어(A_{21})를 연산할 수 있고, 2D 입력 토큰의 어텐션 스코어(a_{22}) 및 2D 입력 토큰의 밸류(V_2)를 이용하여 2D 입력 토큰의 최종 어텐션 스코어(A_{22})를 연산할 수 있으며, 2D 입력 토큰과 3D 입력 토큰 사이의 어텐션 스코어(a_{23}) 및 2D 입력 토큰의 밸류(V_2)를 이용하여 2D 입력 토큰 및 3D 입력 토큰의 어텐션 스코어(A_{23})를 연산할 수 있다.
- [0118] 또한, 융합부(240)는 3D 입력 토큰 및 1D 입력 토큰 사이의 어텐션 스코어(a_{31}) 및 3D 입력 토큰의 밸류(V_3)를 이용하여 3D 입력 토큰 및 1D 입력 토큰의 최종 어텐션 스코어(A_{31})를 연산할 수 있고, 3D 입력 토큰 및 2D 입력 토큰 사이의 어텐션 스코어(a_{32}) 및 3D 입력 토큰의 밸류(V_3)를 이용하여 3D 입력 토큰 및 2D 입력 토큰의 최종 어텐션 스코어(A_{32})를 연산할 수 있으며, 3D 입력 토큰의 어텐션 스코어(a_{33}) 및 3D 입력 토큰의 밸류(V_3)를 이용하여 3D 입력 토큰의 최종 어텐션 스코어(A_{33})를 연산할 수 있다.
- [0119] 즉, 융합부(240)는 연산된 어텐션 스코어와 어텐션 스코어를 연산할 시 쿼리값으로 사용하는 입력 토큰의 밸류를 이용하여 다중 입력 토큰 각각의 최종 어텐션 스코어 또는 입력 토큰과 입력 토큰 사이의 최종 어텐션 스코어를 연산할 수 있다.
- [0120] 융합부(240)는 최종 어텐션 스코어를 기초로 정규화를 통해 스케일링을 수행하고, 심층 신경망과 글로벌 평균 풀링(Global Average Pooling, GAP)을 통해 서로의 중요도가 포함된 융합 출력 토큰 세트를 생성한다.
- [0121] 더욱 자세하게는, 융합부(240)는 1D 입력 토큰의 최종 어텐션 스코어(A_{11})를 심층신경망 및 정규화를 통해 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 1D 단일 토큰을 생성할 수 있다.
- [0122] 그리고, 융합부(240)는 2D 입력 토큰의 최종 어텐션 스코어(A_{22})를 심층신경망 및 정규화를 통해 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 2D 단일 토큰을 생성할 수 있다.
- [0123] 또한, 융합부(240)는 3D 입력 토큰의 최종 어텐션 스코어(A_{33})를 심층신경망 및 정규화를 통해 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 3D 단일 토큰을 생성할 수 있다.
- [0124] 또한, 융합부(240)는 1D 입력 토큰의 최종 어텐션 스코어(A_{11}), 2D 입력 토큰의 최종 어텐션 스코어(A_{22}), 1D 입력 토큰 및 2D 입력 토큰의 최종 어텐션 스코어(A_{12}), 2D 입력 토큰 및 1D 입력 토큰의 최종 어텐션 스코어(A_{21})를 심층신경망 및 정규화를 통해 글로벌 평균 풀링으로 4개의 최종 어텐션 스코어를 모두 더한 값의 평균을 통해 1D 및 2D 융합 토큰을 생성할 수 있다.
- [0125] 또한, 융합부(240)는 2D 입력 토큰의 최종 어텐션 스코어(A_{22}), 3D 입력 토큰의 최종 어텐션 스코어(A_{33}), 2D 입력 토큰 및 3D 입력 토큰의 어텐션 스코어(A_{23}), 3D 입력 토큰 및 2D 입력 토큰의 최종 어텐션 스코어(A_{32})를 심층신경망 및 정규화를 통해 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 2D 및 3D 융합 토큰을 생성할 수 있다.
- [0126] 또한, 융합부(240)는 1D 입력 토큰의 최종 어텐션 스코어(A_{11}), 3D 입력 토큰의 최종 어텐션 스코어(A_{33}), 1D 입력 토큰 및 3D 입력 토큰의 최종 어텐션 스코어(A_{13}), 3D 입력 토큰 및 1D 입력 토큰의 최종 어텐션 스코어(A_{31})를 심층신경망 및 정규화를 통해 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 3D 및 1D 융합 토큰을 생성할 수 있다.

- [0127] 즉, 융합부(240)는 생성하고자 하는 융합 출력 토큰 세트에 해당하는 입력 토큰의 최종 어텐션 스코어를 심층신경망 및 정규화하고 글로벌 평균 풀링으로 모두 더한 값의 평균을 통해 융합 출력 토큰 세트를 생성할 수 있다.
- [0128] 본 발명의 일 실시예에서 제1 모델은 융합 트랜스포머 모델로, ViT(Vision Transformer) 모델을 이용하였으나, 기존의 트랜스포머에서 메모리 사용량을 줄인 Reformer, 기존의 트랜스포머 구조에서 컨볼루션 모듈(Convolution module)을 추가한 Conformer를 사용할 수도 있다.
- [0129] 또한, 본 발명의 일 실시예에서 사용하는 융합 트랜스포머 모델은 일반적인 종래의 비전 트랜스포머 모델에 1D, 2D, 3D 형태의 토큰이 와도 입력받을 수 있도록 입력 차원을 3배로 늘리고, 받지 않은 입력 차원에 대해서 제로 패딩(zero-padding)을 사용한다.
- [0130] 학습부(300)는 복수의 단일 토큰 및 복수의 융합 토큰을 포함하는 융합 출력 토큰 세트의 조합 손실(Combinatorial Loss)을 이용하여 자가지도학습을 수행한다.
- [0131] 이때, 조합 손실은 단일 토큰과 단일 토큰 사이의 대조 손실(Contrastive Loss), 단일 토큰과 융합 토큰 간의 대조 손실을 포함할 수 있다.
- [0132] 학습부(300)는 융합 출력 토큰 세트의 대조 손실을 연산하고, 대조 손실을 이용하여 조합 손실을 산출하며, 조합 손실이 감소하도록 자가지도학습을 수행할 수 있다.
- [0133] 도 6은 조합 손실을 이용한 자가지도학습을 설명하기 위한 도면이다.
- [0134] 더욱 자세하게는, 도 6에서 도시한 바와 같이, 학습부(300)는 1D 단일 토큰과 2D 단일 토큰 사이의 대조 손실($\mathcal{L}_{cl}(z_s, z_t)$), 2D 단일 토큰과 3D 단일 토큰 사이의 대조 손실($\mathcal{L}_{cl}(z_t, z_q)$), 3D 단일 토큰과 1D 단일 토큰 사이의 대조 손실($\mathcal{L}_{cl}(z_q, z_s)$), 2D 및 3D 융합 토큰과 1D 단일 토큰 사이의 대조 손실($\mathcal{L}_{cl}(z_s, z_{tq})$), 1D 및 3D 융합 토큰과 2D 단일 토큰 사이의 대조 손실($\mathcal{L}_{cl}(z_t, z_{qs})$), 1D 및 2D 융합 토큰과 3D 단일 토큰 사이의 대조 손실($\mathcal{L}_{cl}(z_q, z_{st})$)을 다음의 수학적 식 8을 통해 연산할 수 있다.

수학적 식 8

$$\mathcal{L}_{cl}(a, b) = -\log \left(\frac{\exp(a^T b / \gamma)}{\sum_{i=1}^B \exp(a_i^T b_i / \gamma)} \right)$$

- [0135]
- [0136] 여기서, a와 b는 3개의 단일 토큰 및 3개의 융합 토큰 중 두개이며, T는 transpose로 a, b 벡터 간 연산을 위한 것이고, B는 미니배치 속 분자의 개수를 나타낸다.
- [0137] 여기서, 미니 배치는 모든 분자를 한꺼번에 예측하지 않고, 몇 개씩 나누어 순서대로 예측을 진행할 시에 몇 개씩 나누는 단위를 나타내는 것으로, 미니 배치 내부에 존재하는 분자의 개수를 B라 한다.
- [0138] 이때, $a_i^T \times b_i$ 는 미니 배치 안의 i번째 분자의 토큰 두개를 의미한다.
- [0139] 예를 들어, 1000개의 전체 데이터(배치)가 있다고 가정하면, 100개씩 나눌 수 있는데, 100개가 들어있는 한 묶음을 미니 배치라 할 수 있다.
- [0140] 학습부(300)는 연산된 각각의 단일 토큰과 단일 토큰 사이의 대조 손실 및 단일 토큰과 융합 토큰 사이의 대조 손실의 합을 조합 손실로 산출할 수 있다.
- [0141] 학습부(300)는 다음의 수학적 식 9를 이용하여 조합 손실을 산출할 수 있다.

수학식 9

$$\begin{aligned} \mathcal{L}_{comb} = & \mathcal{L}_{cl}(z_s, z_t) + \mathcal{L}_{cl}(z_t, z_g) + \mathcal{L}_{cl}(z_g, z_s) \\ & + \mathcal{L}_{cl}(z_s, z_{tg}) + \mathcal{L}_{cl}(z_t, z_{gs}) + \mathcal{L}_{cl}(z_g, z_{st}) \end{aligned}$$

[0142]

[0143] 여기서, $\mathcal{L}_{cl}$ 은 대조 손실이고, z_s 은 1D 단일 토큰이며, z_t 은 2D 단일 토큰이고, z_g 은 3D 단일 토큰이며, z_{tg} 은 2D 및 3D 융합 토큰이고, z_{gs} 은 1D 및 3D 융합 토큰이며, z_{st} 는 1D 및 2D 융합 토큰이고, $\mathcal{L}_{comb}$ 는 조합 손실을 나타낸다.

[0144] 이때, 조합 손실은 임베딩 네트워크, 트랜스포머 모델이 분자 표현 간의 이질성을 고려할 수 있도록 네트워크 학습을 유도하는 방법으로, 조합 손실의 값이 감소될수록 이질성이 감소하는 것으로 정의할 수 있다.

[0145] 즉, 학습부(300)는 융합 출력 토큰 세트에 포함된 단일 토큰과 단일 토큰 사이의 대조 손실과 단일 토큰과 융합 토큰 사이의 대조 손실의 합인 조합 손실이 감소되도록 조합 손실을 산출하여 자기지도학습을 수행할 수 있다.

[0146] 학습부(300)는 3개의 단일 토큰과 단일 토큰 사이의 대조 손실과, 3개의 단일 토큰과 융합 토큰 사이의 대조 손실을 포함하는 조합 손실을 이용하여 1D 분자표현, 2D 분자표현, 3D 분자표현 사이의 유사도를 증가시켜 각각의 분자표현 사이의 이질성을 낮출 수 있도록 학습할 수 있다.

[0147] 예측부(400)는 융합 출력 토큰 세트를 컨кат하여 제2 모델을 통해 최종 분자 특성을 예측할 수 있다.

[0148] 여기서, 제2 모델은 다층 퍼셉트론일 수 있다.

[0149] 즉, 예측부(400)는 융합 출력 토큰 세트를 제2 모델에 입력하여 최종 분자 특성을 예측할 수 있다.

[0150] 출력부(500)는 예측된 최종 분자 특성을 출력한다.

[0151] 출력부(500)는 예를 들어, 디스플레이, 프린터 장치, 스피커 장치, 영상 출력 단자, 데이터 입출력 단자 또는 통신 모듈 등을 포함할 수 있으나, 이에 한정되는 것은 아니다.

[0152] 필요에 따라 출력부(500)는 입력부(100) 또는 예측부(400)와 일체형으로 마련되는 것도 가능하다.

[0153] 저장부(600)는 입력부(100)로부터 입력받은 분자, 분자표현 생성부(210)로부터 생성된 1D 분자표현, 2D 분자표현 및 3D 분자표현, 토큰 생성부(230)로부터 생성된 1D 입력 토큰, 2D 입력 토큰 및 3D 입력 토큰, 융합부(240)로부터 출력된 융합 출력 토큰 세트, 예측부(400)에서 예측되는 분자 특성을 저장한다.

[0154] 저장부(600)에 저장된 프로그램은, 프로그래머 등의 설계자에 의해 직접 작성 또는 수정된 후 저장부(600)에 저장된 것일 수도 있고, 다른 물리적 기록 매체(외장 메모리 장치나 콤팩트 디스크(CD) 등)로부터 전달받아 저장된 것일 수도 있으며, 및/또는 유무선 통신 네트워크를 통해 접속 가능한 전자 소프트웨어 유통망을 통하여 획득 또는 갱신된 것일 수도 있다.

[0155] 저장부(600)는, 주기억장치 및 보조기억장치 중 적어도 하나를 포함할 수 있다. 주기억장치는 예를 들어, 롬(ROM) 및/또는 램(RAM)과 같은 반도체 저장 매체를 이용하여 구현된 것일 수 있고, 보조기억장치는, 플래시 메모리 장치(솔리드 스테이트 드라이브(SSD, Solid State Drive) 등), SD(Secure Digital) 카드, 하드 디스크 드라이브(HDD, Hard Disc Drive), 콤팩트 디스크, 디브이디(DVD) 또는 레이저 디스크 등과 같이 데이터를 영구적 또는 반영구적으로 저장 가능한 장치를 기반으로 구현될 수 있다.

[0156] 통신부(700)는 입력부(100), 제어부(200), 학습부(300), 예측부(400), 출력부(500), 저장부(600)를 서로 유무선으로 연결하여 데이터를 송수신할 수 있도록 한다.

[0157] 통신부(700)는 유무선 통신망을 모두 포함한다. 가령 통신부(700)로서 유무선 인터넷망이 이용되거나 연동될 수 있다. 여기서 유선망은 케이블망이나 공중 전화망(PSTN)과 같은 인터넷망을 포함하는 것이고, 무선 통신망은 CDMA, WCDMA, GSM, EPC(Evolved Packet Core), LTE(Long Term Evolution), 와이브로(Wibro) 망, 5G 통신망 등을 포함하는 의미이다. 물론 본 발명의 실시예에 따른 통신부(700)는 이에 한정되는 것이 아니며, 향후 구현될

차세대 이동통신 시스템의 접속망으로서 가령 클라우드 컴퓨팅 환경하의 클라우드 컴퓨팅망, 5G망 등에 사용될 수 있다. 가령, 통신부(700)가 유선 통신망인 경우 통신망 내의 액세스포인트는 전화국의 교환국 등에 접속할 수 있지만, 무선 통신망인 경우에는 통신사에서 운영하는 SGSN 또는 GGSN(Gateway GPRS Support Node)에 접속하여 데이터를 처리하거나, BTS(Base Station Transmission), NodeB, e-NodeB 등의 다양한 중계기에 접속하여 데이터를 처리할 수 있다.

- [0159] 이하, 도면을 참조하여 본 발명의 일 실시예에 따른 분자 특성을 예측하는 방법을 설명하도록 한다.
- [0160] 도 7은 본 발명의 일 실시예에 따른 분자 특성 예측 방법을 설명하기 위한 순서도이다.
- [0161] 입력부(100)는 분자를 입력받는다(S110).
- [0162] S110 단계는 도 1의 입력부(100)에서 이미 설명하였으므로, 중복된 설명은 생략한다.
- [0163] 분자표현 생성부(210)는 입력받은 분자를 1D 분자표현, 2D 분자표현 및 3D 분자표현으로 생성한다(S120).
- [0164] S120 단계는 도 1의 분자표현 생성부(210) 및 도 2에서 이미 설명하였으므로, 중복된 설명은 생략한다.
- [0165] 임베딩부(220)는 1D 분자표현, 2D 분자표현 및 3D 분자표현 각각을 임베딩한다(S130).
- [0166] S130 단계는 도 1의 임베딩부(220)에서 이미 설명하였으므로, 중복된 설명은 생략한다.
- [0167] 토큰 생성부(230)는 임베딩된 1D 분자표현, 2D 분자표현 및 3D 분자표현을 이용하여 다중 입력 토큰을 생성한다(S140).
- [0168] S140 단계는 도 1의 토큰 생성부(230)에서 이미 설명하였으므로, 중복된 설명은 생략한다.
- [0169] 융합부(240)는 다중 입력 토큰 중에서 적어도 하나의 입력 토큰을 제1 모델에 입력하여 융합 출력 토큰 세트를 생성한다(S150).
- [0170] S150 단계는 도 1의 융합부(240), 도 3 및 도 4에서 이미 설명하였으므로, 중복된 설명은 생략한다.
- [0171] 학습부(300)는 융합 출력 토큰 세트의 조합 손실을 이용하여 자기지도학습을 수행한다(S160).
- [0172] S160 단계는 도 1의 학습부(300) 및 도 5에서 이미 설명하였으므로, 중복된 설명은 생략한다.
- [0173] 예측부(400)는 융합 출력 토큰 세트를 제2 모델에 입력하여 최종 분자특성을 예측한다(S170).
- [0174] S170 단계는 도 1의 예측부(400)에서 이미 설명하였으므로, 중복된 설명은 생략한다.
- [0176] 이상에서 설명한 바와 같이, 본 발명에 따르면, 3D 분자표현 생성기를 이용하여 3D 분자표현을 생성할 수 있고, 다중 분자표현을 이용하여 대조학습 기반의 자기 지도학습을 통해 다중 분자표현의 이질성을 줄일 수 있다.
- [0178] 본원 발명의 실시예들과 관련된 기술 분야에서 통상의 지식을 가진 자는 상기 기재의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로, 개시된 방법들은 한정적인 관점이 아닌 설명적 관점에서 고려되어야 한다. 본 발명의 범위는 발명의 상세한 설명이 아닌 특허청구 범위에 나타나며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

- [0180] 1: 분자 특성 예측 장치
- 100: 입력부
- 200: 제어부
- 210: 분자표현 생성부 220: 임베딩부

230: 토큰 생성부 240: 융합부

300: 학습부

400: 예측부

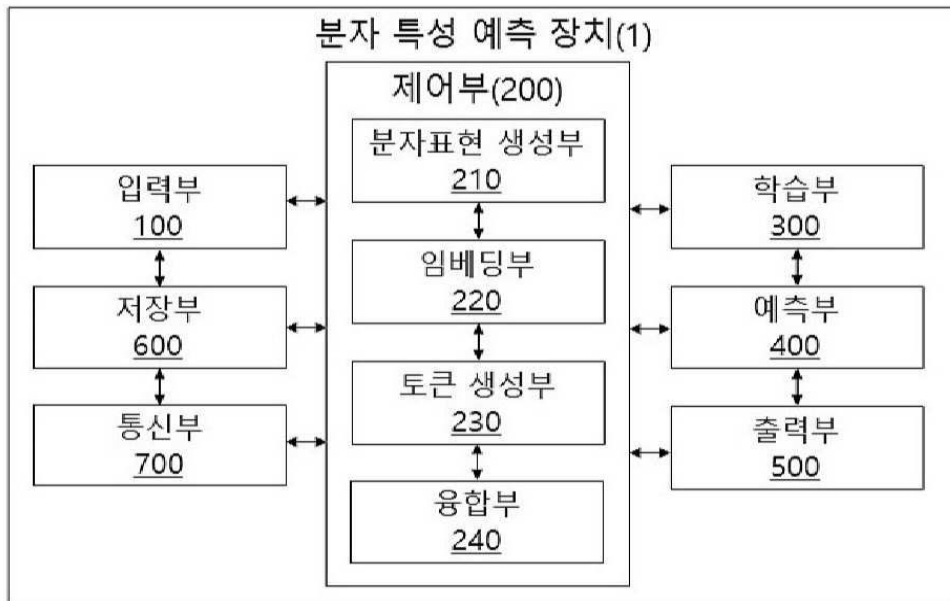
500: 출력부

600: 저장부

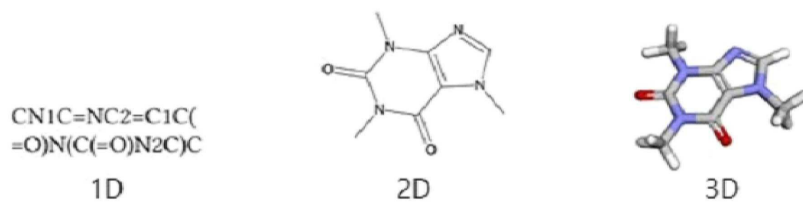
700: 통신부

도면

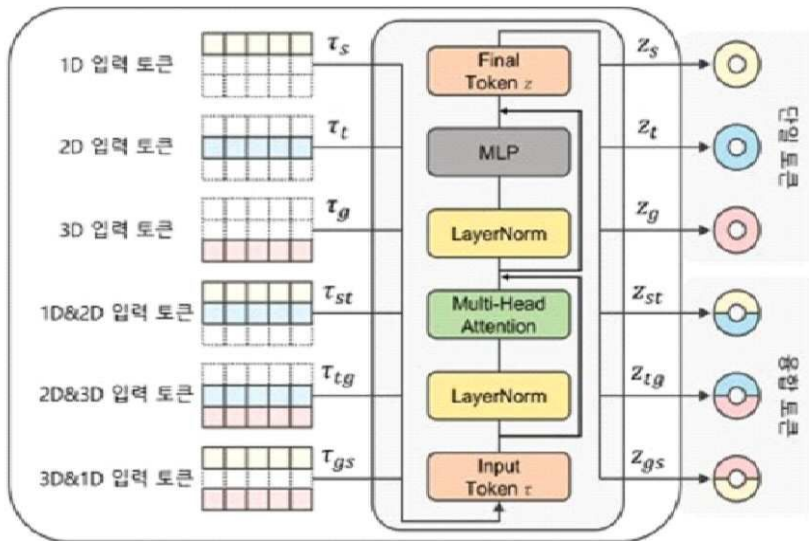
도면1



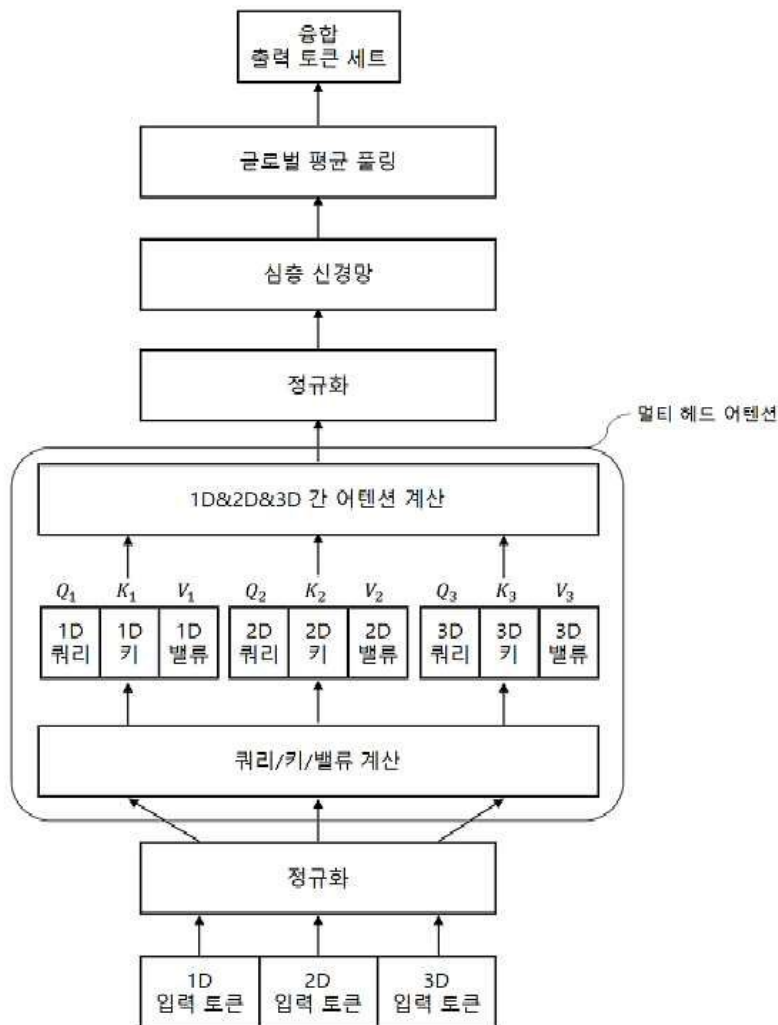
도면2



도면3



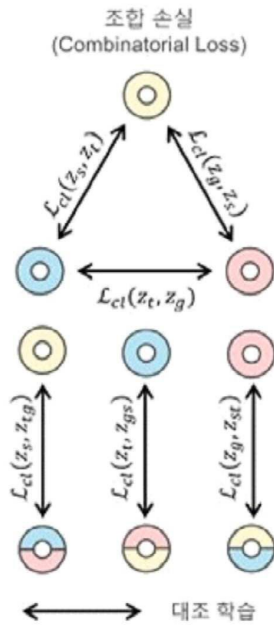
도면4



도면5

$$\begin{array}{lll} a_{11} = Q_1 * K_1, & a_{12} = Q_1 * K_2, & a_{13} = Q_1 * K_3 \\ a_{21} = Q_2 * K_1, & a_{22} = Q_2 * K_2, & a_{23} = Q_2 * K_3 \\ a_{31} = Q_3 * K_1, & a_{32} = Q_3 * K_2, & a_{33} = Q_3 * K_3 \end{array}$$

도면6



도면7

