



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0064094
(43) 공개일자 2023년05월10일

(51) 국제특허분류(Int. Cl.)

G16C 20/50 (2019.01) G06N 3/08 (2023.01)
G16C 20/70 (2019.01) G16C 20/90 (2019.01)

(52) CPC특허분류

G16C 20/50 (2019.02)
G06N 3/08 (2023.01)

(21) 출원번호 10-2021-0149307

(22) 출원일자 2021년11월03일

심사청구일자 2021년11월03일

(71) 출원인

서울대학교산학협력단
서울특별시 관악구 관악로 1 (신림동)

이화여자대학교 산학협력단

서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)

(72) 발명자

이원보

서울특별시 관악구 관악로 1 서울대학교

김현승

서울특별시 관악구 관악로 1 서울대학교

나종걸

서울특별시 서대문구 이화여대길 52 이화여자대학교

(74) 대리인

강문호

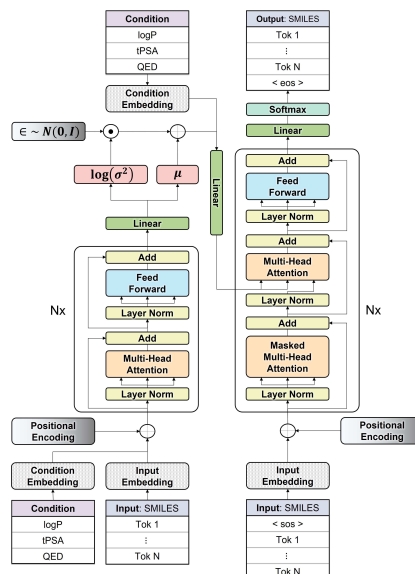
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 물질 설계 장치

(57) 요약

물질 설계 장치를 제공한다. 상기 물질 설계 장치는 생성적 화학 트랜스포머를 포함하고, 상기 생성적 화학 트랜스포머는 번역할 문장을 수신하고, 셀프 어텐션을 통해 수신된 문장을 이해하는 인코더 및 상기 인코더로부터 문장 이해에서 처리된 정보를 받아 이전 단계까지 번역된 문장 다음에 올 토큰을 선택하는 디코더를 포함한다.

대표도 - 도2



(52) CPC특허분류

G16C 20/70 (2019.02)

G16C 20/90 (2019.02)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711130934
과제번호	2018M3D1A1058633
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	나노및소재기술개발사업
연구과제명	하이퍼 이온이송채널의 멀티스케일 모델링, 분자디자인 기법 개발
기 여 율	1/1
과제수행기관명	서울대학교
연구기간	2021.01.16 ~ 2022.01.15

명세서

청구범위

청구항 1

생성적 화학 트랜스포머를 포함하고,

상기 생성적 화학 트랜스포머는,

번역할 문장을 수신하고, 셀프 어텐션을 통해 수신된 문장을 이해하는 인코더, 및

상기 인코더로부터 문장 이해에서 처리된 정보를 받아 이전 단계까지 번역된 문장 다음에 올 토큰을 선택하는 디코더를 포함하는 것을 특징으로 하는 물질 설계 장치.

청구항 2

제 1 항에 있어서,

상기 생성적 화학 트랜스포머는,

화학 언어에 대한 이해를 기반으로 목표 물성을 가진 분자를 생성하는 조건부 변화 오토인코더에 언어 모델의 트랜스포머를 내재한 것을 특징으로 하는 물질 설계 장치.

청구항 3

제 1 항에 있어서,

상기 생성적 화학 트랜스포머는,

상기 인코더와 상기 디코더 사이에 삽입된 조건부 가우시안 잠재 공간을 갖는 것을 특징으로 하는 물질 설계 장치.

청구항 4

제 1 항에 있어서,

상기 생성적 화학 트랜스포머는,

분자 구조와 다중 목표 물성을 학습하는 것을 특징으로 하는 물질 설계 장치.

청구항 5

제 1 항에 있어서,

상기 생성적 화학 트랜스포머의 손실 함수는 하기 식 3으로 표현되는 것을 특징으로 하는 물질 설계 장치.

[식 3]

$$L(\theta, \theta; x_{enc}, x_{<t}, x_t, c) = k_w D_{KL}(q_\theta(z|x_{enc}, c) \parallel p(z|c)) - E_{q_\theta(z|x_{enc}, c)}[\log g_\theta(x_t|z, x_{<t}, c)]$$

(상기 식 3에서, $D_{KL}(\cdot)$ 는 KL 발산(Kullback-Leibler divergence)을 나타내고, $E[\cdot]$ 은 $\cdot$ 의 기대값을 나타내며, q_θ 는 매개변수화된 인코더 함수를 나타내고, g_θ 는 매개변수화된 디코더 함수를 나타내며, $p(\cdot|c)$ 는 조건부 가우시안 사전분포를 나타내고, θ , θ , x_{enc} , $x_{<t}$, z , x_t , c , k_w 는 각각 인코더의 매개변수, 디코더의 매개변수, 인코더의 입력, 디코더의 입력, 잠재 변수, 복원 목표, 목표 물성 조건, KL 어닐링 계수를 나타냄)

청구항 6

제 1 항에 있어서,
 상기 생성적 화학 트랜스포머의 어텐션 메카니즘은,
 쿼리(query), 키(key), 및 밸류(value)의 세 가지 벡터를 사용하는 것을 특징으로 하는 물질 설계 장치.

청구항 7

제 6 항에 있어서,
 상기 어텐션 메카니즘은,
 상기 쿼리(query)와 상기 키(key) 간의 유사성을 계산하고, 계산된 유사성에 상기 키에 해당하는 값을 곱하여 어텐션 점수를 계산하는 것을 특징으로 하는 물질 설계 장치.

청구항 8

제 6 항에 있어서,
 상기 생성적 화학 트랜스포머는 멀티 헤드 어텐션을 사용하는 것을 특징으로 하는 물질 설계 장치.

청구항 9

제 1 항에 있어서,
 상기 인코더의 셀프 어텐션 블록은 SMILES 스트링과 다중 목표 물성의 연결 배열(concatenated array)을 획득하고,
 상기 디코더의 인코더-디코더 어텐션 블록은 상기 다중 목표 물성과 잠재 코드의 연결 배열(concatenated array)을 획득하며,
 상기 디코더의 셀프 어텐션 블록은 현재 단계까지 생성된 SMILES 스트링을 획득하는 것을 특징으로 하는 물질 설계 장치.

청구항 10

생성적 화학 트랜스포머를 포함하고,
 상기 생성적 화학 트랜스포머는,
 화학 언어에 대한 이해를 기반으로 목표 물성을 가진 분자를 생성하는 조건부 변화 오토인코더에 언어 모델의 트랜스포머를 내재한 것을 특징으로 하는 물질 설계 장치.

청구항 11

제 10 항에 있어서,
 상기 생성적 화학 트랜스포머는,
 번역할 문장을 수신하고, 셀프 어텐션을 통해 수신된 문장을 이해하는 인코더, 및
 상기 인코더로부터 문장 이해에서 처리된 정보를 받아 이전 단계까지 번역된 문장 다음에 올 토큰을 선택하는 디코더를 포함하는 것을 특징으로 하는 물질 설계 장치.

청구항 12

제 11 항에 있어서,
 상기 생성적 화학 트랜스포머는,
 상기 인코더와 상기 디코더 사이에 삽입된 조건부 가우시안 잠재 공간을 갖는 것을 특징으로 하는 물질 설계 장치.

발명의 설명

기술 분야

[0001] 본 발명은 물질 설계 장치에 관한 것이다.

배경 기술

[0002] 물질 발견은 특정 목적에 맞는 새로운 물질을 찾는 연구 분야이다. 화학 공간과 유기 분자의 수는 방대하여 여러 원하는 조건을 동시에 충족하는 분자 구조를 발굴하려면 높은 수준의 전문 지식과 상당한 시간 및 비용이 필요하다.

발명의 내용

해결하려는 과제

[0003] 본 발명은 우수한 성능을 갖는 물질 설계 장치를 제공한다.

[0004] 본 발명의 다른 목적들은 다음의 상세한 설명과 첨부한 도면으로부터 명확해 질 것이다.

과제의 해결 수단

[0005] 본 발명의 일 실시예에 따른 물질 설계 장치는 생성적 화학 트랜스포머를 포함하고, 상기 생성적 화학 트랜스포머는 번역할 문장을 수신하고, 셀프 어텐션을 통해 수신된 문장을 이해하는 인코더 및 상기 인코더로부터 문장 이해에서 처리된 정보를 받아 이전 단계까지 번역된 문장 다음에 올 토큰을 선택하는 디코더를 포함한다.

[0006] 상기 생성적 화학 트랜스포머는, 화학 언어에 대한 이해를 기반으로 목표 물성을 가진 분자를 생성하는 조건부 변화 오토인코더에 언어 모델의 트랜스포머를 내재한 것일 수 있다.

[0007] 상기 생성적 화학 트랜스포머는, 상기 인코더와 상기 디코더 사이에 삽입된 조건부 가우시안 잠재 공간을 가질 수 있다.

[0008] 상기 생성적 화학 트랜스포머는, 분자 구조와 다중 목표 물성을 학습할 수 있다.

[0009] 상기 생성적 화학 트랜스포머의 손실 함수는 하기 식 3으로 표현될 수 있다.

[0010] [식 3]

[0011] $L(\theta, \theta; x_{enc}, x_{<t}, x_t, c) =$

[0012] $k_w D_{KL}(q_\theta(z|x_{enc}, c) \parallel p(z|c)) - E_{q_\theta(z|x_{enc}, c)}[\log q_\theta(x_t|z, x_{<t}, c)]$

[0013] (상기 식 3에서, $D_{KL}(\cdot)$ 는 KL 발산(Kullback-Leibler divergence)을 나타내고, $E[\cdot]$ 은 $\cdot$ 의 기대값을 나타내며, q_θ 는 매개변수화된 인코더 함수를 나타내고, g_θ 는 매개변수화된 디코더 함수를 나타내며, $p(\cdot|c)$ 는 조건부 가우시안 사전분포를 나타내고, θ , θ , x_{enc} , $x_{<t}$, z , x_t , c , k_w 는 각각 인코더의 매개변수, 디코더의 매개변수, 인코더의 입력, 디코더의 입력, 잠재 변수, 복원 목표, 목표 물성 조건, KL 어닐링 계수를 나타냄)

[0014] 상기 생성적 화학 트랜스포머의 어텐션 메카니즘은, 쿼리(query), 키(key), 및 밸류(value)의 세 가지 벡터를 사용할 수 있다. 상기 어텐션 메카니즘은, 상기 쿼리(query)와 상기 키(key) 간의 유사성을 계산하고, 계산된 유사성에 상기 키에 해당하는 값을 곱하여 어텐션 점수를 계산할 수 있다. 상기 생성적 화학 트랜스포머는 멀티 헤드 어텐션을 사용할 수 있다.

[0015] 상기 인코더의 셀프 어텐션 블록은 SMILES 스트링과 다중 목표 물성의 연결 배열(concatenated array)을 획득할 수 있고, 상기 디코더의 인코더-디코더 어텐션 블록은 상기 다중 목표 물성과 잠재 코드의 연결 배열(concatenated array)을 획득할 수 있으며, 상기 디코더의 셀프 어텐션 블록은 현재 단계까지 생성된 SMILES 스트링을 획득할 수 있다.

[0016] 본 발명의 다른 실시예에 따른 물질 설계 장치는 생성적 화학 트랜스포머를 포함하고, 상기 생성적 화학 트랜스포머는 화학 언어에 대한 이해를 기반으로 목표 물성을 가진 분자를 생성하는 조건부 변화 오토인코더에 언어

모델의 트랜스포머를 내재한 것이다.

[0017] 상기 생성적 화학 트랜스포머는 번역할 문장을 수신하고, 셀프 어텐션을 통해 수신된 문장을 이해하는 인코더 및 상기 인코더로부터 문장 이해에서 처리된 정보를 받아 이전 단계까지 번역된 문장 다음에 올 토큰을 선택하는 디코더를 포함할 수 있다.

[0018] 상기 생성적 화학 트랜스포머는, 상기 인코더와 상기 디코더 사이에 삽입된 조건부 가우시안 잠재 공간을 가질 수 있다.

발명의 효과

[0019] 본 발명의 실시예들에 따른 물질 설계 장치는 우수한 성능을 갖는다. 예를 들어, 상기 물질 설계 장치는 다중 목표 물성을 충족하는 물질들을 빠른 시간에 생성할 수 있다. 또, 생성되는 물질들은 이미 존재하는 물질들뿐만 아니라 실현 가능한 새로운 물질들을 포함할 수 있다.

도면의 간단한 설명

[0020] 도 1은 언어 형태의 구조적 제한을 설명하기 위한 도면이다.

도 2는 본 발명의 일 실시예에 따른 생성적 화학 트랜스포머의 아키텍처를 나타낸다.

도 3은 생성적 화학 트랜스포머를 이용한 새로운 분자 생성을 설명하기 위한 도면이다.

도 4는 인코더 블록에서 어텐션 헤드의 셀프 어텐션 점수에 대한 시각화된 예를 보여준다.

도 5는 목표 물성(GCT에 주어진 전제 조건)과 30,000개의 생성된 분자의 물성 비교를 나타낸다.

도 6은 각 모델의 신규성 점수와 유효성 점수의 산점도를 보여준다.

발명을 실시하기 위한 구체적인 내용

[0021] 이하, 실시예들을 통하여 본 발명을 상세하게 설명한다. 본 발명의 목적, 특징, 장점은 이하의 실시예들을 통해 쉽게 이해될 것이다. 본 발명은 여기서 설명되는 실시예들에 한정되지 않고, 다른 형태로 구체화될 수도 있다. 여기서 소개되는 실시예들은 개시된 내용이 철저하고 완전해질 수 있도록 그리고 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 본 발명의 사상이 충분히 전달될 수 있도록 하기 위해 제공되는 것이다. 따라서, 이하의 실시예들에 의하여 본 발명이 제한되어서는 안 된다.

[0023] 본 발명의 일 실시예에 따른 물질 설계 장치는 생성적 화학 트랜스포머를 포함하고, 상기 생성적 화학 트랜스포머는 번역할 문장을 수신하고, 셀프 어텐션을 통해 수신된 문장을 이해하는 인코더 및 상기 인코더로부터 문장 이해에서 처리된 정보를 받아 이전 단계까지 번역된 문장 다음에 올 토큰을 선택하는 디코더를 포함한다.

[0024] 본 발명의 다른 실시예에 따른 물질 설계 장치는 생성적 화학 트랜스포머를 포함하고, 상기 생성적 화학 트랜스포머는 화학 언어에 대한 이해를 기반으로 목표 물성을 가진 분자를 생성하는 조건부 변화 오토인코더에 언어 모델의 트랜스포머를 내재한 것이다.

[0025] 상기 생성적 화학 트랜스포머는, 상기 인코더와 상기 디코더 사이에 삽입된 조건부 가우시안 잠재 공간을 가질 수 있다.

[0027] 도 1은 언어 형태의 구조적 제한을 설명하기 위한 도면이다. 도 1a는 논해밀턴 그래프(non-Hamiltonian graph)의 예를 나타낸다. 해밀턴 그래프에는 그래프의 모든 점을 한 번만 통과하는 경로가 있고, 논해밀턴 그래프에는 이러한 경로가 없다. 도 1b는 논해밀턴 분자 그래프와 이에 대응하는 SMILES 스트링의 예(4-(2-아미노프로필)-2-메톡시페놀)를 나타낸다. 각 원자는 원자로 표시된 숫자로 표시되고, 다른 색상은 다른 가치를 나타낸다.

[0028] 도 1을 참조하면, 화학 언어인 SMILES(Simplified Molecular-Input Line-Entry System)는 분자를 1차원 텍스트로 나타낸다. 하나의 SMILES 스트링이 하나의 정확한 분자로 변환되기 때문에 강력한 언어이다. 그러나 대부분의 분자는 논해밀턴 그래프(non-Hamiltonian graph)이므로 분자의 가치가 1차원 스트링으로 번역될 때마다 의미론적 불연속성이 발생할 수 있다. SMILES는 여는 괄호와 닫는 괄호로 각 가치를 구분하고 스트링 내에서 두 문

자의 거리와 분자 그래프에서 해당하는 두 원자의 거리 사이에 간격을 만든다. 원자 2의 C와 원자 13의 N은 기하학적으로 인접해 있지만 스트링에서는 서로 멀리 떨어져 있다. 즉, 가지가 길고 많을수록 화학 스트링을 순서대로 읽어서 분자 구조를 추론하는 것은 어렵다. 비슷한 현상이 자연어에서도 발생한다. 도 1c는 자연어에서 문장 내에서 의미적으로 가까운 단어가 문장 내에서 항상 구조적으로 가까운 것이 아님을 보여준다. 문장이 길수록 문장 내 단어와 의미적 유사성 사이의 간격이 커진다. 화학 언어에서는 의미 불연속성이 발생하는 영역이 더 많다. NLP(Natural Language Processing) 분야에서 이 문제에 대한 어텐션 메커니즘을 도입함으로써 언어 모델은 의미적으로 관련된 부분에 어텐션(attention)을 기울일 수 있다. 트랜스포머는 신경망 형태의 어텐션 메커니즘을 포함하는 아키텍처로, 어텐션 메커니즘이 화학 언어의 의미론적 불연속성을 넘어 구조적 이해에 도움이 될 수 있다.

[0030] SMILES는 분자를 구성하는 원자, 결합, 고리를 스트링으로 표현한다. 원자는 원소기호의 알파벳으로 나타내고, 결합은 종류에 따라 단일결합(-), 이중결합(=), 삼중결합(#)으로 나타낸다. 일반적으로 원자나 주변 원자의 고리 구조를 통해 쉽게 유추할 수 있는 결합은 생략한다. 단일 결합 수소가 화학적 원자가 규칙에 의해 명시적으로 유추될 수 있는 경우 SMILES 스트링에서도 수소 표기가 생략되지만, 결합이 암시적이면 [H]를 사용하여 단일 결합 수소를 명확하게 나타낼 수 있다. 수소 결합의 수를 명확하게 결정할 수 없는 하전된 원자의 경우, 원자와 전하는 대괄호 [] 안에 함께 표시된다. 각 고리의 시작과 끝은 동일한 숫자로 표시되며 쌍이 정확해야 한다. 고리가 열려 있으면 닫아야 한다. 방향족 고리에 존재하는 원자는 소문자로 표기하고 고리 외부의 원자는 대문자로 표기한다. 분자의 가지는 여는 괄호와 닫는 괄호로 표시된다.

[0031] SMILES 스트링을 언어 모델에 입력하기 위해서는 의미 단위로 토큰화하는 과정이 필요하다. SmilesPE 토큰라이저(tokenizer)는 MOSES(Molecular Sets benchmarking platforms: 분자 세트 벤치마킹 플랫폼)의 훈련 데이터에 포함된 SMILES 스트링을 토큰화하는 데 사용된다. 예를 들어, 특수 토큰 4개(<unknown>, <pad>, <sos>, <eos>), 원자 토큰 13개(<C>, <c>, <O>, <o>, <N>, <n>, <F>, <S>, <s>, <Cl>,
, <[nH]> 및 <[H]>), 결합 토큰 3개(<->, <=> 및 <#>), 가지 토큰 2개(<(> 및 <)>) 및 고리 토큰 6개(<1>, <2>, <3>, <4>, <5>, 및 <6>)로 총 28가지 유형의 토큰이 사용될 수 있다.

[0032] 하전된 원자와 관련된 토큰(예: <[O-]>, <[n+]>) 및 입체화학과 관련된 토큰(예: </>, <#>)은 MOSES 데이터베이스에서 다루지 않으므로 고려되지 않았다. SMILES 스트링을 구성하는 각 토큰은 28차원으로 원-핫 인코딩되고 512차원으로 임베딩된다. 생성적 화학 트랜스포머의 조건도 512차원에서 임베딩된다.

[0033] 어텐션 메커니즘은 트랜스포머의 언어 인지 능력의 핵심이다. 어텐션 메커니즘을 통해 트랜스포머는 주어진 작업을 더 잘 수행하기 위해 집중하는데 더 나은 입력 스트링의 토큰을 자체 학습할 수 있다. 어텐션 메커니즘은 쿼리(query) Q, 키(key) K, 및 밸류(value) V의 세 가지 벡터를 사용한다. 어텐션 메커니즘은 Q와 K의 모든 기간의 유사성을 계산하고 계산된 유사성에 키에 해당하는 값을 곱하여 어텐션 점수(attention score)를 계산한다. 트랜스포머에서 사용되는 스케일-닷 어텐션(scale-dot Attention)은 식 1과 같다.

[0034] [식 1]

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

[0035]

[0036] 식 1에서, d_k 는 K의 차원이고, d_v 는 Q의 차원인 d_q 에 대응해야 한다.

[0038] 트랜스포머는 단일 헤드 어텐션 대신 식 2의 멀티 헤드 어텐션을 사용한다.

[0039] [식 2]

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

[0040]

$$\text{where } \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

[0041]

$$\text{식 2에서, } W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_i^O \in \mathbb{R}^{h d_v \times d_{\text{model}}}$$

[0042]

$d_{model} = d_k \times h = d_v \times h$ 이다. h 는 멀티 헤드의 수이고 d_v 는 V의 차원이다. 각 헤드($head_i$)는 각 헤드에 속하는 다른 웨이트(weights)(W_i^q, W_i^k, W_i^v)를 사용하여 다른 관점에서 Q와 K 사이의 어텐션 점수를 계산한다.

[0044] 도 2는 본 발명의 일 실시예에 따른 생성적 화학 트랜스포머의 아키텍처를 나타내고, 도 3은 생성적 화학 트랜스포머를 이용한 새로운 분자 생성을 설명하기 위한 도면이다. 도 3a는 NMT용 트랜스포머 모델로 라틴어를 영어로 번역하는 예를 나타내며, 이전 단계까지 번역된 라틴어 문장과 영어 문장을 참조하여 다음 영어 단어를 반복적으로 선택한다. 도 3b는 생성적 화학 트랜스포머가 입력된 화학식을 재현하는 방법을 학습하는 과정을 나타내며, 생성적 화학 트랜스포머는 분자 구조와 logP, tPSA 및 QED의 세 가지 다른 물성을 학습한다. 학습 과정에서 잠재 공간의 분자 구조 및 물성 정보를 나타낸다. 도 3c는 생성적 화학 트랜스포머가 새로운 분자를 추론하여 생성하는 과정을 나타내며, 훈련된 생성적 화학 트랜스포머는 잠재 공간과 주어진 전체 조건에서 샘플링된 분자 정보를 디코딩하여 목표 물성을 만족하는 새로운 분자를 생성한다.

[0045] 도 2 및 도 3을 참조하면, 생성적 화학 트랜스포머(Generative Chemical Transformer, GCT)는 화학 언어에 대한 깊은 이해를 기반으로 목표 물성을 가진 분자를 생성하는 조건부 변화 오토인코더(Conditional Variational Autoencoder, cVAE)에 트랜스포머를 내재한 아키텍처이다. 상기 트랜스포머는 주로 NMT(Neural Machine Translator)로 사용된다.

[0046] 생성적 화학 트랜스포머는 인코더와 디코더로 구성된다. 인코더는 번역할 문장을 수신하고, 셀프 어텐션을 통해 수신된 문장을 이해한다. 문장 이해에서 처리된 정보는 디코더로 전달된다. 디코더는 인코더에서 수신한 정보와 이전 단계까지 번역된 문장을 참조하여 이 지점까지 번역된 문장 다음에 올 다음 토큰을 반복적으로 선택한다. 번역 시작 부분에 번역된 문장이 없으면 특수 토큰 '문장 시작 < sos >'가 사용된다. 디코더는 입력 정보를 사용하여 이전 단계까지 번역된 문장 다음에 올 다음 토큰을 반복적으로 선택한다. 마지막으로 디코더가 특수 토큰 '문장 종료 < eos >'를 선택하면 번역이 종료된다.

[0047] 생성적 화학 트랜스포머는 Pre-LN(Pre-Layer Normalization) 트랜스포머의 인코더와 디코더 사이에 저차원 조건부 가우시안 잠재 공간을 삽입하는 구조를 갖는다. Pre-LN 트랜스포머는 Post-LN (Post-Layer Normalization) 트랜스포머의 수정된 버전이다. 언어 모델과 변화 오토인코더의 조합은 사후 붕괴에 취약하다.

[0048] 사후 붕괴는 KL(Kullback-Leibler divergence) 어닐링에 의해 완화될 수 있다. KL 어닐링(KLA)이 그라디언트 크기를 제어하므로 Post-LN 트랜스포머의 그라디언트 흐름을 안정화하도록 설계된 Pre-LN 트랜스포머를 채택하면 KL 어닐링 조작을 용이하게 할 수 있다. 생성적 화학 트랜스포머의 손실 함수는 식 3과 같다.

[0049] [식 3]

$$L(\emptyset, \theta; x_{enc}, x_{< t}, x_t, c) =$$

$$k_w D_{KL}(q_\emptyset(z|x_{enc}, c) \parallel p(z|c)) - E_{q_\emptyset(z|x_{enc}, c)}[\log g_\theta(x_t|z, x_{< t}, c)]$$

[0052] 식 3에서, $D_{KL}(\cdot)$ 는 KL 발산(Kullback-Leibler divergence)을 나타내고, $E[\cdot]$ 은 $\cdot$ 의 기대값을 나타낸다. $q_\emptyset$ 는 매개변수화된 인코더 함수를 나타내고, g_θ 는 매개변수화된 디코더 함수(생성기)를 나타내며, $p(\cdot|c)$ 는 조건부 가우시안 사전분포를 나타낸다. 그리고, $\emptyset, \theta, x_{enc}, x_{< t}, z, x_t, c, k_w$ 는 각각 인코더의 매개변수, 디코더의 매개변수, 인코더의 입력, 디코더의 입력, 잠재 변수, 복원 목표, 목표 물성 조건, KL 어닐링 계수를 나타낸다.

[0053] 인코더와 디코더는 각각 6개의 Pre-LN 트랜스포머 블록으로 구성된다. 각 블록은 512와 8-헤드 어텐션의 차원을 가지며, 피드-포워드 블록의 차원은 2,048이다. 가우시안 잠재 공간은 128차원으로 설계된다.

[0054] 인코더의 셀프 어텐션 블록은 SMILES 스트링과 옥탄올-물 분배 계수(logP), 토폴로지 극 표면적(tPSA) 및 약물 유사성의 정량적 추정(QED)의 세 가지 물성의 연결 배열(concatenated array)을 획득한다. 디코더의 인코더-디코더 어텐션 블록은 조건(세 가지 물성)과 잠재 코드의 연결 배열(concatenated array)을 획득하고, 디코더의 셀프 어텐션 블록은 현재 단계까지 생성된 SMILES 스트링을 획득한다. 훈련 단계에서 생성적 화학 트랜스포머는

분자 물성에 대한 주어진 힌트를 참조하여 SMILES 스트링을 복원하는 작업을 수행한다. 이 과정에서 저차원의 잠재 공간이 모델의 병목 역할을 하여 병목을 통과한 제한된 정보를 최대한 활용하여 디코더에 복원할 수 있는 저차원의 의미 있는 정보를 최대한 많이 찾는다. 그런 다음 분자 구조 및 물성에 대한 의미있는 잠재 변수를 저차원 연속 잠재 공간에 표시한다. 추론 단계에서는 샘플링된 잠재 코드와 목표 물성이 학습된 디코더에 입력되고 디코더는 4-빔 서치(beam search)을 통해 반복적으로 다음 토큰을 선택한다. 이것은 일종의 트리 검색 방법이다.

[0055] 생성적 화학 트랜스포머의 드랍아웃(dropout)은 0.3이 적용된다. 학습은 아담 옵티마이저(Adam optimizer)에 의해 수행된다. 초기 학습률은 10^{-4} 이다. 모멘텀의 확장률은 0.9이고 적응형의 확장률은 0.98이다. 학습률을 예약하기 위해 두 가지 방법이 적용된다(표 1의 GCT-WarmUP 및 GCT-SGDR).

[0056] 하나는 식 4의 워밍업 스케줄러(warm-up scheduler)를 사용하는 것이다.

[0057] [식 4]

$$\eta = 3 d_{model}^{-0.5} \cdot \min(s_{current}^{-0.5}, s_{warmup}^{-1.5})$$

[0059] 식 4에서, η 는 학습률을 나타내고, $s_{current}$ 는 현재 훈련 단계를 나타내며, s_{warmup} 은 워밍업 단계를 나타낸다. s_{warmup} 은 100,000으로 설정된다.

[0060] 다른 하나는 식 5의 원 에포크 사이클(one epoch cycle)의 재시작(SGDR)과 함께 확률적 경사 하강법을 사용하는 것이다.

[0061] [식 5]

$$\eta = \eta_{min} + 0.5(\eta_{max} - \eta_{min}) \left(1 + \cos \left(\frac{s_{current}}{s_{cycle}} \pi \right) \right)$$

[0063] 식 5에서, η_{min} 은 최소 학습률을 나타내고, η_{max} 은 최대 학습률을 나타내며, s_{cycle} 은 학습률 스케줄링 스텝 사이클을 나타낸다. η_{min} , η_{max} , 및 s_{cycle} 은 각각 0, 0.0001, 및 1-에포크 단계로 설정된다. KL 어닐링은 총 25 에포크에 대하여 에포크당 0.02간격으로 k_w 가 0.02에서 0.5까지 증가하도록 적용된다.

[0065] 생성적 화학 트랜스포머 모델은 MOSES 벤치마킹 플랫폼 데이터베이스를 사용하여 훈련되고 벤치마킹되었다. MOSES 데이터베이스는 훈련 샘플(1.7M), 테스트 샘플(176k) 및 스캐폴드 테스트 샘플(176k)로 구성되며, 스캐폴드 테스트 샘플은 훈련 샘플에는 나타나지 않는 스캐폴드를 갖는다. 또, 데이터세트 간의 분포를 밀접하게 일치 시키도록 설계되었다. RDKit에서 계산된 세 가지 추가 물성(logP, tPSA 및 QED)이 생성적 화학 트랜스포머 학습에 사용된다.

[0066] 일반적으로 네트워크 훈련의 품질은 모델의 예측과 실제 라벨이 얼마나 다른지를 측정하여 평가할 수 있다. 그러나 분자 생성 모델의 경우 평균 손실 측정에서 관찰되지 않은 생성 분자의 인공물이 화학적 원자가 규칙에 맞지 않거나 분자를 비현실적으로 만들 수 있기 때문에 작은 평균 손실이 생성 모델의 성능을 보장하지 않는다. 이러한 이유로 생성된 분자의 품질을 다음 기준에 따라 확인해야 한다.

[0067] 1) 생성된 분자는 얼마나 그럴듯한가?

[0068] 2) 생성된 분자가 목표 물성을 만족하는가?

[0069] 3) 단일 전제조건 세트에 대하여 다중 후보를 생성할 수 있는가?

[0070] 4) 새로운 분자를 짧은 시간에 생성할 수 있는가?

[0072] 총 30,000개의 SMILES 스트링이 훈련된 생성적 화학 트랜스포머 모델에 의해 생성되고 MOSES 벤치마킹 스코어

메트릭에 의해 평가된다(표 1 참조). MOSES 벤치마킹 플랫폼은 유효성, 고유성, 내부 다양성, 필터 및 신규성과 같은 비교적 간단한 점수 외에도 SNN(Similarity to a Nearest Neighbor), FCD(Fr chet ChemNet Distance), Frag(Fragment similarity), 및 Scaf(Scaffold similarity)와 같은 레퍼런스 분자와의 유사성을 측정할 수 있는 메트릭을 제공한다.

[0073] SNN 점수는 식 6을 이용하여 계산된다.

[0074] [식 6]

$$\text{SNN}(G, R) = \frac{1}{|G|} \sum_{m_G \in G} \max_{m_R \in R} T(m_G, m_R)$$

[0075]

[0076] 상기 식 6에서, G 및 R은 각각 생성된 분자 및 레퍼런스 분자를 나타낸다. m은 모건 지문(Morgan fingerprints)을 나타내고, T(A,B)는 A와 B 사이의 타니모토 유사성(Tanimoto similarity)을 나타낸다.

[0077] FCD는 ChemNet에서 끝에서 두 번째 레이어의 활성화를 사용하며 바이오 활성을 예측하도록 설계되었다. 식 7을 이용하여 G와 R 사이의 분포 차이를 계산한다.

[0078] [식 7]

$$\text{FCD}(G, R) = \|\mu_G - \mu_R\|^2 + \text{Tr} \left(\sum_G + \sum_R - 2 \left(\sum_G \sum_R \right)^{1/2} \right)$$

[0079]

[0080] 상기 식 7에서, μ 는 평균을 나타내고, $\sum$ 은 공분산을 나타내며, $\text{Tr}(\cdot)$ 은 추적 오퍼레이터를 나타낸다.

[0081] Frag 점수는 식 8을 이용하여 계산된다.

[0082] [식 8]

$$\text{Frag}(G, R) = \frac{\sum_{f \in F} (c_f(G) \cdot c_f(R))}{\sqrt{\sum_{f \in F} c_f^2(G)} \sqrt{\sum_{f \in F} c_f^2(R)}}$$

[0083]

[0084] 상기 식 8에서, F는 58,315개의 BRICS 프래그먼트(fragments)를 나타내고, $c_f(A)$ 는 A세트의 분자에서 프래그먼트 $f \in F$ 가 나타날 빈도를 나타낸다.

[0085] Scaf 점수는 식 9를 이용하여 계산된다.

[0086] [식 9]

$$\text{Scaf}(G, R) = \frac{\sum_{s \in S} (c_s(G) \cdot c_s(R))}{\sqrt{\sum_{s \in S} c_s^2(G)} \sqrt{\sum_{s \in S} c_s^2(R)}}$$

[0087]

[0088] 상기 식 9에서, S는 448,854개의 BM(Bemis-Murcko) 스캐폴드를 나타내고, $c_s(A)$ 는 A세트의 분자에서 스캐폴드 $s \in S$ 가 나타날 빈도를 나타낸다.

[0090] 샘플링 조건에서 고려되는 물성은 logP, tPSA 및 QED이다. 훈련 데이터의 최대값과 최소값 사이에서 각 물성을 1,000개의 동일한 섹션으로 나눈 후 3차원 히스토그램을 도출하였다. 그리고, 각 셀에 데이터 샘플이 존재할 확률에 따라 셀을 샘플링하였다. 여기서 확률은 전체 샘플 중 해당 셀에 있는 샘플의 수이다. 30,000개의 분자가 생성될 조건 세트를 생성하기 위해 셀의 중심 값에 균일한 노이즈가 추가되었다. logP-축, tPSA-축, QED-축에 대한 균일 노이즈의 크기는 각 축 방향에서 셀 측면의 크기를 초과하지 않도록 적용된다.

[0091] 잠재 변수의 차원은 128로 설정된다. 그러나 SMILES 스트링을 구성하는 토큰의 수는 분자마다 다르기 때문에 잠재 변수의 시퀀스 길이는 매번 다르게 적용된다. 시퀀스 길이는 SMILES 스트링을 구성하는 토큰의 개수를 의미

한다. 각 분자 생성에 사용된 시퀀스 길이는 정규 분포에서 샘플링되었다. 정규 분포의 평균과 분산은 MOSES 훈련 데이터셋의 SMILES 스트링을 구성하는 토큰의 수에서 유도되었다. 시퀀스 길이가 결정된 후 잠재 변수의 값은 표준 정규 분포에서 샘플링된다.

[표 1]

			GCT		MOSES Reference Models				
			GCT-WarmUp	GCT-SGDR	VAE	AAE	Char RNN	Latent GAN	JTN-VAE
Validity	↑		0.9853	0.9916	0.9767 ±0.0012	0.9368± 0.0341	0.9748 ±0.0264	0.8966 ±0.0029	1.0±0.0
Unique@1k	↑		1.0	0.998	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0
Unique@10k	↑		0.9981	0.9797	0.9984 ±0.0005	0.9973 ±0.002	0.9994 ±0.0003	0.9968 ±0.0002	0.9996 ±0.0003
IntDiv	↑		0.8531	0.8458	0.8558 ±0.0004	0.8557 ±0.0031	0.8562 ±0.0005	0.8565 ±0.0007	0.8551 ±0.0034
Filters	↑		0.9956	0.9982	0.9970 ±0.0002	0.9960 ±0.0006	0.9943 ±0.0034	0.9735 ±0.0006	0.9760 ±0.0016
Novelty	↑		0.8144	0.6756	0.6949 ±0.0069	0.7931 ±0.0285	0.8419 ±0.0509	0.9498 ±0.0006	0.9143 ±0.0058
SNN	↑	Test	0.6179	0.6513	0.6257 ±0.0005	0.6081 ±0.0043	0.6015 ±0.0206	0.5371 ±0.0004	0.5477 ±0.0076
		TestSF	0.5771	0.5990	0.5783 ±0.0008	0.5677 ±0.0045	0.5649 ±0.0142	0.5132 ±0.0002	0.5194 ±0.007
FCD	↓	Test	0.4017	0.7980	0.0990 ±0.0125	0.5555 ±0.2033	0.0732 ±0.0247	0.2968 ±0.0087	0.3954 ±0.0234
		TestSF	0.8031	0.9949	0.5670 ±0.0338	1.0572 ±0.2375	0.5204 ±0.0379	0.8281 ±0.0117	0.9382 ±0.0531
Frag	↑	Test	0.9973	0.9922	0.9994 ±0.0001	0.9910 ±0.0051	0.9998 ±0.0002	0.9986 ±0.0004	0.9965 ±0.0003
		TestSF	0.9952	0.8562	0.9984 ±0.0003	0.9905 ±0.0039	0.9983 ±0.0003	0.9972 ±0.0007	0.9947 ±0.0002
Scaf	↑	Test	0.8905	0.8562	0.9386 ±0.0021	0.9022 ±0.0375	0.9242 ±0.0058	0.8867 ±0.0009	0.8964 ±0.0039
		TestSF	0.0921	0.0551	0.0588 ±0.0095	0.0789 ±0.009	0.1101 ±0.0081	0.1072 ±0.0098	0.1009 ±0.0105

표 1은 MOSES 벤치마킹 결과를 비교하여 나타낸다. MOSES 벤치마킹은 GCT에서 생성된 30,000 SMILES 스트링에 대해 수행되었다. 생성적 화학 트랜스포머(GCT)에서 얻은 점수를 MOSES 벤치마킹 플랫폼에서 제공하는 다른 생성 모델(VAE, AAE, CharRNN, LatentGAN 및 JTN-VAE)의 점수와 비교하였다. 10개의 지표를 사용하여 생성된 분자의 품질을 평가하였다. 위쪽 화살표는 값이 높을수록 좋고 아래쪽 화살표는 값이 낮을수록 좋다.

유효성(Validity)은 유효한 SMILES 스트링의 비율을 나타낸다. Unique@1k는 1,000개의 생성된 분자 중 고유한 샘플의 비율을 나타낸다. Unique@10k는 생성된 분자 10,000개 중 고유한 샘플의 비율을 나타낸다. IntDiv는 생성된 분자의 내부 다양성을 나타낸다. 필터(Filters)는 독성 물질 필터를 통과하는 비율을 나타낸다. 신규성(Novelty)은 훈련 데이터에 존재하지 않는 생성된 분자의 비율을 나타낸다. SNN은 생성된 모든 분자에 대한 평균 타니모토 유사성을 나타낸다. 생성된 분자와 레퍼런스 데이터 중 가장 유사한 분자 간의 유사성을 계산한다. FCD는 ChemNet 바이오 활성 예측 모델의 끝에서 두 번째 레이어의 활성화를 사용하여 생성된 분자와 레퍼런스 분자 사이의 거리 측정값을 나타낸다. Frag는 BRICS가 레퍼런스 데이터에 나타나는 빈도와 생성된 분자에 나타나는 빈도 사이의 코사인 유사성을 나타낸다. Scaf는 레퍼런스 데이터에 특정 스캐폴드가 나타나는 빈도와 생성된 분자에 나타나는 빈도 사이의 코사인 유사성을 나타낸다.

생성된 SMILES 스트링이 그럴듯한 분자 구조를 나타내는지 여부를 평가하려면 두 가지 관점에서 분석이 필요하다.

첫 번째 분석은 생성된 SMILES 스트링이 유효한 분자 그래프를 생성할 수 있는지 여부, 즉 생성된 SMILES 스트링이 화학 원자가 규칙과 SMILES 언어의 문법을 모두 충족하는지 여부이다. 벤치마킹 결과 생성된 SMILES 스트링의 98.5% 이상이 유효한 것으로 나타났다. GCT-WarmUp은 98.5%의 유효성을 보여주었고 GCT-SGDR은 99.2%의 유효성을 보여준다. 언어 기반 모델 중 가장 높은 값이다. 이 능력은 생성 기계가 SMILES 스트링을 통해 분자의 기하학적 구조를 얼마나 잘 이해할 수 있는지에 달려 있다. SMILES 언어의 문법과 화학 원자가 규칙을 충족하는

주어진 화학 스트링 다음에 오는 특징(원자, 결합 또는 가지에 해당)을 결정하기 위해 스트링에 존재하는 분자의 기하학적 구조(각 원자와 가지의 연결)를 이해해야 한다. 생성적 화학 트랜스포머(GCT)에 적용된 어텐션 메커니즘은 SMILES 언어의 의미론적 불연속성을 넘어 화학 언어의 문법을 신경망이 이해하는 데 도움이 된다.

[0100] 도 4는 인코더 블록에서 어텐션 헤드의 셀프 어텐션 점수에 대한 시각화된 예를 보여준다. 도 4는 SMILES 스트링 내의 특징에 어텐션을 기울이는 방법에 대한 두 가지 극단적인 예에 대한 결과를 보여준다. 도 4a는 4-(2-아미노프로필)-2-메톡시페놀에 대한 두 번째 인코더 블록에서 네 번째 헤드의 시각화 결과를 나타내고, 도 4b는 5,6-비스(p-메톡시페닐)-3-메틸-1,2,4-트리아진에 대한 두 번째 인코더 블록에서 세 번째 헤드의 시각화 결과를 나타낸다. 각 분자 그래프에서 원자 2에 인접한 원자가 각 SMILES 스트링에서 멀리 떨어져 있다. 도 4a의 원자 13 및 도 4b의 원자 23은 각 SMILES 스트링에서 멀리 떨어져 있다. 도면 하단의 검은색 상자에 있는 노란색 선은 어떤 토큰이 어텐션에 의해 쿼리 벡터로부터 높은 어텐션 점수가 주어지는지를 보여준다. 어텐션 점수가 높을수록 노란색 선이 불투명해진다. SMILES 스트링과 분자 그래프의 각 색상은 괄호로 구분된 SMILES 스트링의 각 가지를 나타낸다.

[0101] 도 4a를 참조하면, 원자 1, 원자 2 및 원자 13은 분자 그래프에서 서로 가깝게 위치하지만 SMILES 스트링 내에서는 서로 멀리 떨어져 있다. SMILES 스트링만 생성적 화학 트랜스포머에 제공되지만 원자 1, 원자 2 및 원자 13은 서로 관련이 있음(♠)으로 인식된다. 도 4b를 참조하면, GCT-SGDR이 원자 2와 원자 22, 23(♠) 사이의 관계를 인식한다는 것을 보여준다. 또, 특정 가지에 해당하는 원자(♣)와 고리형 가지(♥)를 인식한다. 각 어텐션-헤드는 다른 관점에 따라 분자 구조를 인식한다. 즉, 생성적 화학 트랜스포머에 적용된 어텐션 메커니즘은 생성적 화학 트랜스포머가 1차원 텍스트에 숨겨진 분자 구조를 인식하는 데 도움이 된다.

[0102] 두 번째 분석은 유효한 SMILES 스트링에서 구문 분석된 분자 구조가 실제 분자와 얼마나 유사한가이다. SNN, FCD, Frag 및 Scaf의 네 가지 메트릭이 이를 측정할 수 있다. MOSES에 있는 두 가지 테스트 세트(테스트 세트 및 TestSF 세트)에 대해 네 가지 메트릭을 평가하였다. 테스트 세트를 구성하는 분자는 트레인 세트를 구성하는 동일한 분자 스캐폴드를 가지고 있다. 반면, TestSF는 트레인 세트에 포함되지 않은 스캐폴드를 포함하는 분자로 구성된다. GCT-SGDR의 SNN(Test)과 SNN(TestSF), GCT-WarmUp의 Scaf(TestSF)를 제외하고는 VAE가 GCT보다 좋은 점수를 보였다. 그러나 다른 레퍼런스 모델의 점수와 비교하면 GCT의 점수가 나쁘지 않다. GCT-WarmUp의 SNN(Test) 및 SNN(TestSF)은 VAE를 제외한 다른 모든 참조 모델의 점수보다 높다. GCT-WarmUp의 FCD(Test)는 AAE보다 우수하고 GCT-WarmUp의 FCD(TestSF)는 AAE, LatentGAN 및 JTN-VAE의 FCD(TestSF)보다 우수하다. GCT-WarmUp의 Frag(Test) 및 Frag(TestSF)는 AAE 및 JTN-VAE의 Frag(Test) 및 Frag(TestSF)보다 높다. GCT-WarmUp의 Scaf(Test)는 LatentGAN보다 높고 GCT-WarmUp의 Scaf(TestSF)는 VAE 및 AAE보다 높다.

[0104] 도 5는 목표 물성(GCT에 주어진 전제 조건)과 30,000개의 생성된 분자의 물성 비교를 나타낸다. 도 5a~c에서 x축은 목표 물성을 나타내고, y축은 생성된 분자의 물성을 나타낸다. 도 5d에서 x축은 잠재 코드의 길이를 나타내고, y축은 생성된 SMILES 스트링을 구성하는 토큰의 수를 나타낸다. 도 5e~g는 MOSES 데이터 분포(적색선) 및 생성된 분자의 물성(청색 점)에서 무작위로 샘플링된 10개의 사전조건 세트를 나타낸다. 도 5h는 MOSES 데이터 분포(적색 마커)와 생성된 분자를 구성하는 토큰 수(청색 점)에서 무작위로 샘플링된 잠재 코드의 길이를 나타낸다.

[0105] 생성적 화학 트랜스포머는 다중 목표 물성을 충족하는 분자 구조를 생성한다. 도 5a~c는 생성적 화학 트랜스포머에서 생성된 30,000개 분자의 물성(RDKit에서 계산됨)과 목표 물성(GCT에서 주어진 전제 조건)을 비교한 결과이다. logP와 tPSA는 분자 구조와 직접적인 관련이 있는 물성이므로 분자 구조에 대한 이해를 바탕으로 목표 물성에 해당하는 분자 구조를 생성할 수 있다. 그러나 QED는 8가지 다른 물성의 기하학적 평균을 통해 약물의 유사성을 정량적으로 판단하기 위해 고안된 인공적인 지수이므로 QED의 경우 상대적으로 어렵다. 이 현상은 cRNN에서도 발견된다. 각 물성에 대한 목표 조건과 생성된 분자의 물성 사이의 절대 평균 오차는 0.177(logP), 2.923(tPSA), 0.035(QED)이다.

[0106] 생성된 SMILES 스트링의 길이는 생성적 화학 트랜스포머가 오토인코더 구조를 가지고 있기 때문에 잠재 코드의 길이에 따라 다르다. 인코더에 입력된 정보를 복원하도록 훈련된다. 훈련 단계에서 잠재 코드의 길이는 인코더에 입력된 스트링의 길이 및 디코더에서 출력되는 스트링의 길이와 동일하게 나타난다. 이것은 <eos> 및 <eos> 토큰이 입력 및 출력 디자인에 사용되는지 여부에 따라 약간 다를 수 있다. 그러나 스트링의 본성은 다르지 않다.

[0107] 추론 단계에서 입력 잠재 코드의 길이와 생성된 SMILES 스트링의 길이가 완벽하게 일치하지 않으며 GCT는 시퀀스 길이의 분포를 학습하지 않는다(도 5d 참조). 그러나 분자의 크기와 관련되는 생성된 SMILES 스트링의 길이는 잠재 코드의 길이를 조절함으로써 어느 정도 조작될 수 있다.

[0108] 주어진 다중 목표 물성이 동시에 충족되는지 확인하기 위해 생성된 분자의 물성을 훈련 데이터 분포에서 샘플링된 10개 사전조건 세트와 비교하였다(도 5e~g). 생성적 화학 트랜스포머의 골격인 조건부 모델은 다중 목표 물성을 동시에 잘 만족시키는 분자를 생성한다. 또, 생성적 화학 트랜스포머의 변화 생성기는 동일한 전체조건 세트에서 다양한 분자를 생성할 수 있도록 한다.

[0110] 생성 모델이 새로운 분자를 생성할 수 있는지 여부는 물질 발견에 대한 적용 가능성을 결정하는 중요한 기준이다. 신규성 점수는 훈련 데이터에 존재하지 않는 새로운 분자를 생성할 확률을 나타낸다. 이상하고 비현실적인 분자가 신규성 점수를 높일 수 있기 때문에 높은 신규성 점수만으로는 좋은 생성기임을 보장하지는 않는다. 따라서 신규성 점수는 생성된 분자가 현실적인지 여부를 평가하기 위해 지표와 함께 사용해야 한다.

[0111] 도 5는 각 모델의 신규성 점수와 유효성 점수의 산점도를 보여준다. 점선은 레퍼런스 모델 점수(VAE, AAE, CharRNN, GAN, JTN-VAE)의 선형 회귀이다. 흥미롭게도 모든 레퍼런스 모델에서 SNN 점수가 증가함에 따라 신규성 점수가 감소하는 것으로 관찰되었다. 반대로 신규성 점수가 높을수록 SNN 점수는 낮다. 이것은 실제 분자와 유사한 새로운 분자를 만드는 것이 쉽지 않다는 것을 의미한다. 그러나 GCT(GCT-WarmUp, GCT-SGDR, GCT-Exp1~3)가 레퍼런스 모델보다 새롭고 유사한 분자를 더 잘 생성한다는 것을 확인할 수 있다. 여기서 GCT는 다른 하이퍼파라미터 또는 다른 학습률 스케줄을 사용하여 훈련된 모델을 의미한다. 각 GCT에 대한 세부 조건 및 점수는 표 2에 요약되어 있다.

[0112] [표 2]

	GCT-Exp1 (baseline)	GCT -WarmUp	GCT -Exp2	GCT -Exp3	GCT -Exp4	GCT -SGDR
Training Conditions						
Epochs	25	25	25	25	50	25
KLA weight (k_w) [start:step:end]	[0:0.02:0.5]	[0:0.02:0.5]	[0:0.02:0.5]	[0:0.01:0.25]	[0:0.01:0.5]	[0:0.02:0.5]
Learning rate (lr)	Baseline lr	Baseline lr × 3 (= eq. 3)	Baseline lr × 5	Baseline lr	Baseline lr	eq. 4
Learning rate scheduler	Warm-Up	Warm-Up	Warm-Up	Warm-Up	Warm-Up	SGDR
MOSES Benchmarks						
Validity	0.9757	0.9853	0.9808	0.9692	0.9813	0.9916
Unique@1k	1.0	1.0	1.0	1.0	1.0	0.998
Unique@10k	0.9977	0.9981	0.9977	0.9983	0.9986	0.9797
IntDiv	0.8465	0.8531	0.8531	0.8541	0.8491	0.8458
Filters	0.9955	0.9956	0.9950	0.9944	0.9943	0.9982
Novelty	0.8019	0.8144	0.8043	0.8114	0.8572	0.6756
SNN (Test)	0.6204	0.6179	0.6206	0.6081	0.6131	0.6513
SNN (TestSF)	0.5784	0.5771	0.5791	0.5693	0.5744	0.5990
FCD (Test)	0.4181	0.4017	0.3813	0.3883	0.7180	0.7980
FCD (TestSF)	0.8560	0.8031	0.8411	0.7855	1.221	0.9949
Frag (Test)	0.9977	0.9973	0.9980	0.9979	0.9944	0.9922
Frag (TestSF)	0.9957	0.9952	0.9956	0.9960	0.9910	0.8562
Scaf (Test)	0.9021	0.8905	0.8935	0.8644	0.8893	0.8562
Scaf (TestSF)	0.0883	0.0921	0.1039	0.1049	0.0969	0.0551

[0113]

[0115] SGDR(GCT-SGDR)을 사용하는 모델은 워밍업 스케줄러(GCT-WarmUp)를 사용하는 모델보다 신규성이 낮고 유효성이 높다. 주기적으로 학습률을 감소시키는 스케줄러는 VAE 손실 함수의 KL 발산항을 줄이는데 손실이 있지만, 복원 오차항을 줄이는 이점이 있다. 일종의 사이클 어닐링 스케줄러인 SGDR 스케줄러는 GCT-SGCR이 높은 유효성과 낮은 신규성을 갖도록 한다. 분자 생성당 걸리는 시간은 8C/16T CPU 및 NVIDIA GTX 1080 Ti 환경에서 507ms, 12C/24T 및 NVIDIA Tesla T4 환경에서 440ms였다.

[0117] 본 발명의 실시예들에 따른 생성적 화학 트랜스포머 아키텍처는 어텐션 메커니즘을 사용하는 NLP 분야에서 언어 모델인 트랜스포머를 조건부 변화 생성기에 내재한다. 훈련된 생성적 화학 트랜스포머는 화학 언어에 대한 깊은

이해를 기반으로 원하는 조건을 충족하는 SMILES 스트링을 생성할 수 있다.

[0118]

생성적 화학 트랜스포머의 어텐션 메카니즘은 화학식에 스파스 어텐션(sparse attention)을 가짐으로써 논해밀턴(non-Hamiltonian) 분자 그래프를 1차원 스트링으로 변환하여 발생하는 화학 언어의 의미론적 불연속성의 한계를 넘어 분자의 기하학적 구조를 깊이 이해하는 데 도움이 된다. 화학 언어에 대한 깊은 이해는 생성된 SMILES 스트링이 SMILES 언어의 문법을 충족하고 화학 규칙을 충족하며 현실적이게 만든다. 생성적 화학 트랜스포머의 조건부 변화 생성기는 생성된 분자가 다중 목표 물성을 동시에 만족하고 다양하게 만든다. 생성적 화학 트랜스포머의 오토인코더 구조는 분자 크기를 제어할 수 있게 한다. 생성적 화학 트랜스포머는 훈련 과정에서 볼 수 없었던 새로운 분자를 생성하고 수백 밀리초 만에 분자를 생성한다. 잘 훈련된 생성적 화학 트랜스포머(GCT-WarmUp)는 98.5% 확률로 유효한 SMILES 스트링을 생성한다. 생성된 SMILES 스트링의 84.1%는 학습된 적이 없는 새로운 분자였으며 실제 분자와의 유사도는 0.681(SNN 점수)이었다. 기존 실제 분자의 패턴과 유사하지만 이전에 본 적이 없는 새로운 분자를 만드는 것은 어렵다. 비교 모델 중 생성적 화학 트랜스포머는 실제 분자와 유사한 새로운 분자를 만드는 데 최고의 성능을 보였다. 또, 생성된 분자는 다중 목표 물성을 동시에 충족했으며 세 가지 다른 물성에 대한 평균 절대 오차는 0.177(logP), 2.923(tPSA) 및 0.035(QED)이었다. 또, 생성적 화학 트랜스포머가 분자 크기를 제어할 수 있음을 확인하였다. 주어진 잠재 코드 길이와 비교하여 생성된 SMILES 토큰 수의 평균 차이는 0.332이다. 분자 생성은 개인용 컴퓨터에서 분자당 507ms가 소요되었다. 또, 생성적 화학 트랜스포머에 적용 가능한 조건을 필요에 따라 다르게 구성할 수 있으며, 생성적 화학 트랜스포머를 BERT, GPT, T5와 같은 트랜스포머 기반 아키텍처로 확장할 수 있다.

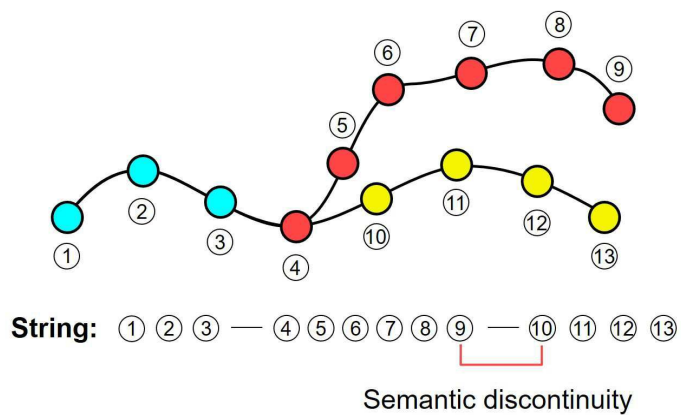
[0120]

이제까지 본 발명에 대한 구체적인 실시예들을 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

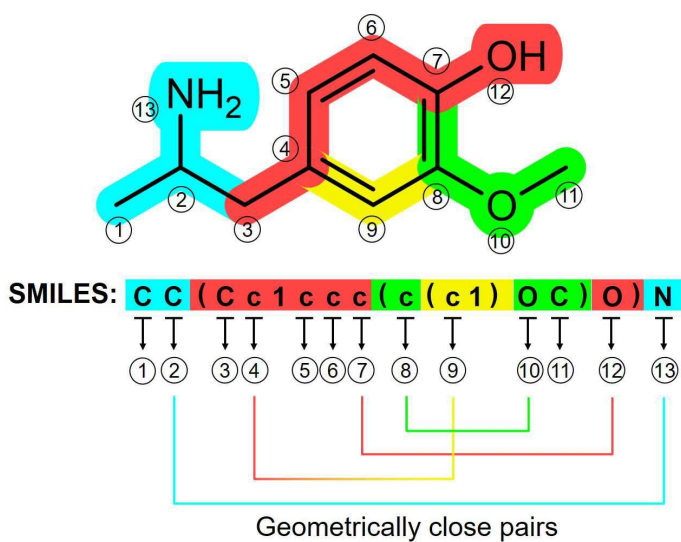
도면

도면1

a



b

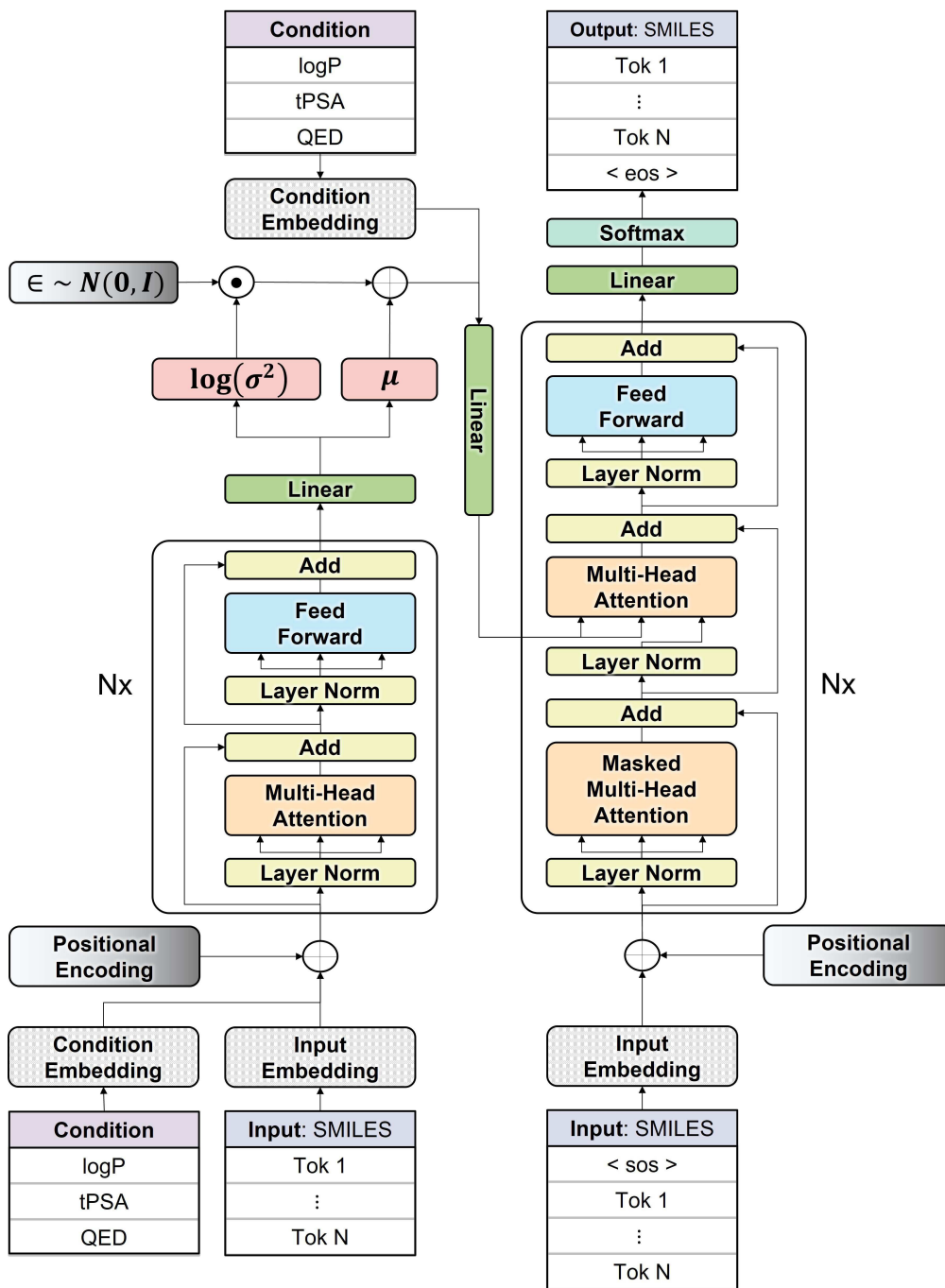


c

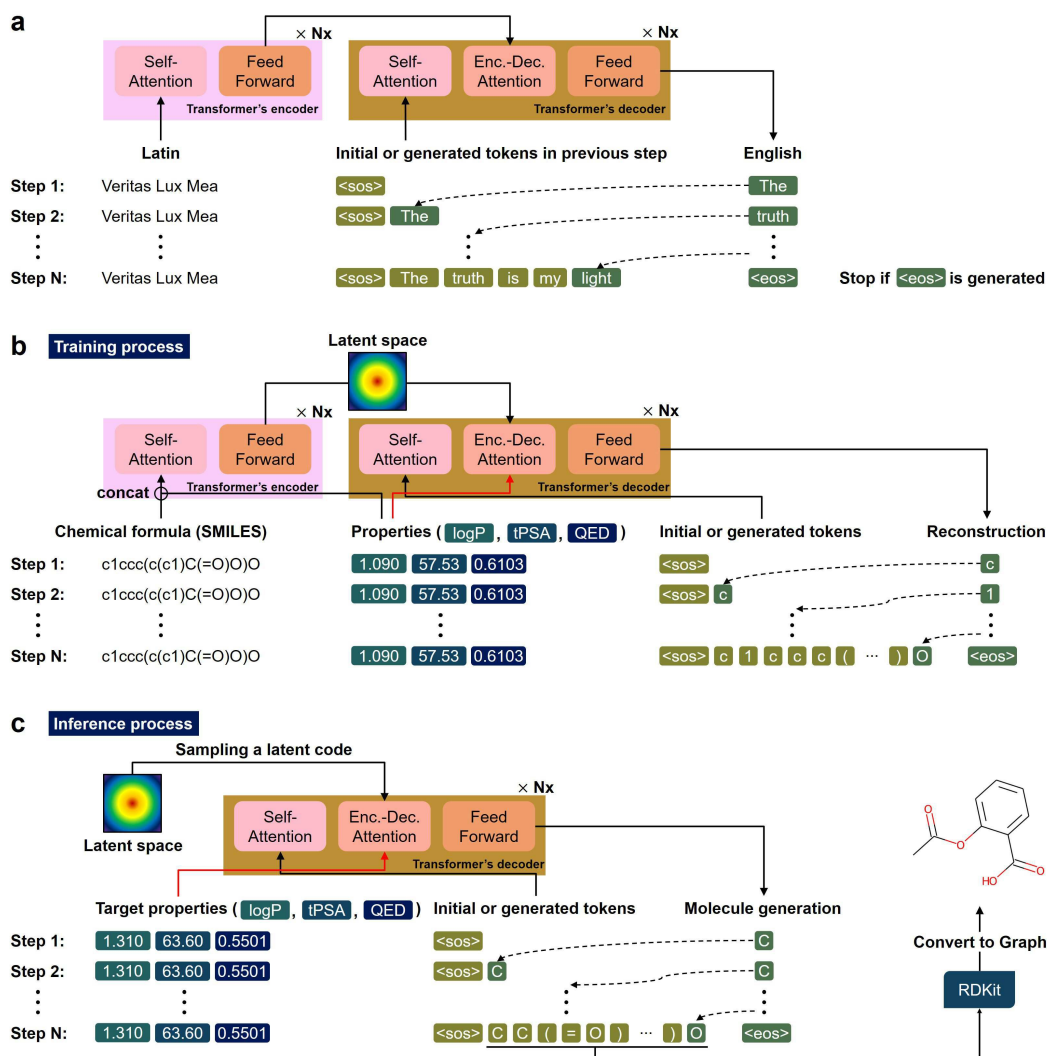
Yesterday I lost my dog on the way to a walk,
and my mom found him in the park.

Semantically close pair

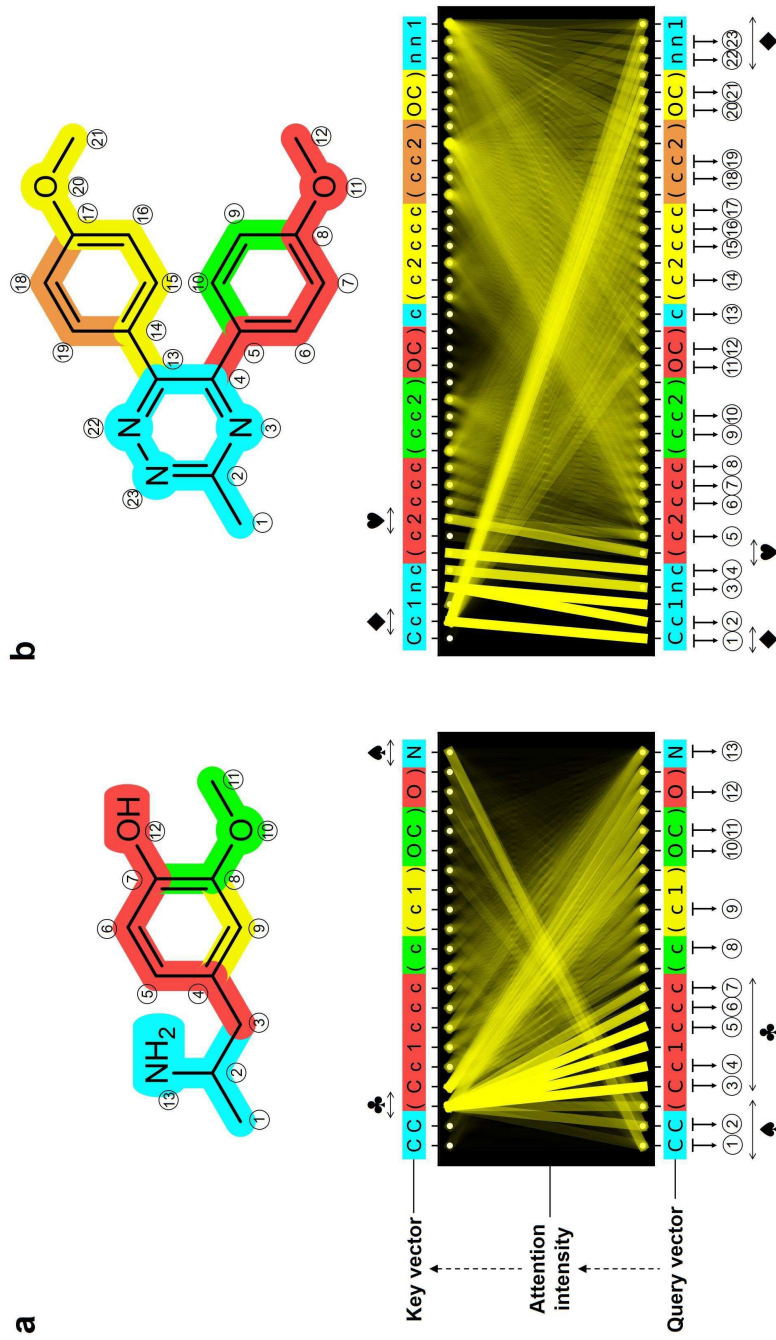
도면2



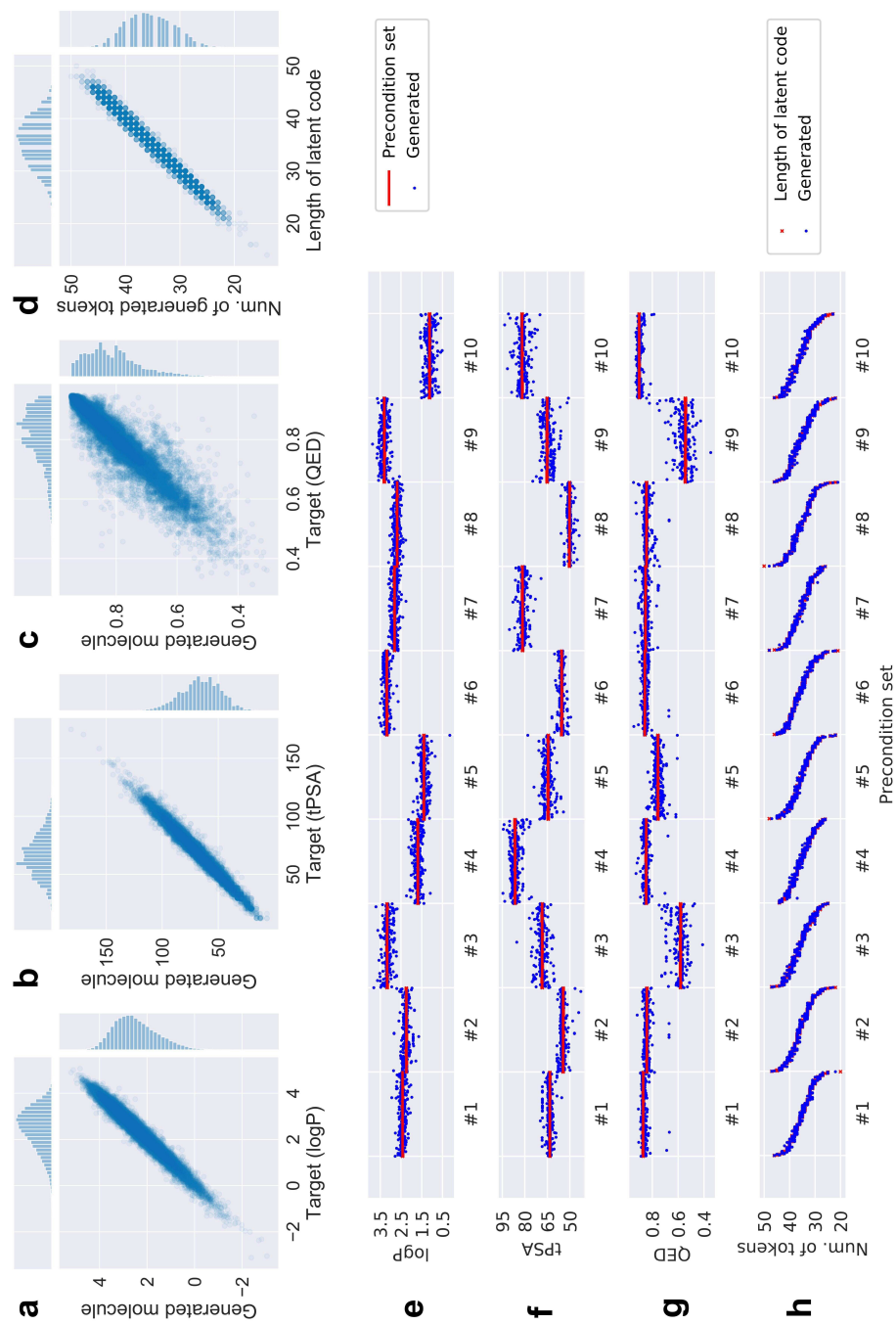
도면3



도면4



도면5



도면6

