PCT22-0037

PCT 출원서

(전자적 형태가 원본)

0	**수리관청 전용**	
0-1	국제출원번호	
0-2	국제출원일자	
0-3	수리관청 명칭 및 "PCT 국제출원"	

0-4	**서식 PCT/RO/101 - PCT 출원서**	
0-4-1	우측에 기재된 바와 같이 작성되었다.	**ePCT-Filing** **Version 4.10.010 MT/FOP 20221109/1.1**
0-5	**신청**	
	아래 서명인은 본 국제 출원서가 특허협력조약에 의해 처리될 것을 청구합니다.	
0-6	**출원인이 지정한 수리관청**	**대한민국 특허청 (RO/KR)**
0-7	**출원인 또는 대리인의 서류참조기호**	**PCT22-0037**
I	**발명의 명칭**	**딥러닝 기반 분자 설계 방법, 이를 수행하는 장치 및 컴퓨터 프로그램**
II	**출원인**	
II-1	이 사람은	**오직 출원인 (applicant only)**
II-2	우측 지정국에 관한 출원인	**모든 지정국 (all designated States)**
II-4ko	성명	**연세대학교 산학협력단**
II-4en	Name:	**UIF (UNIVERSITY INDUSTRY FOUNDATION), YONSEI UNIVERSITY**
II-5ko	주소	**대한민국** **03722** **서울특별시 서대문구 연세로 50**
II-5en	Address:	**50, Yonsei-ro,** **Seodaemun-gu, Seoul 03722** **Republic of Korea**
II-6	국적	**대한민국 KR**
II-7	거주국	**대한민국 KR**
II-8	전화번호	**+82-2-2123-5138**
II-10	이메일 주소	**patent@yonsei.ac.kr**
II-10(a)	이메일 사용동의 수리관청, 국제조사기관, 국제사무국, 국제예비심사기관이 필요 시 이 이메일 주소를 사용하여 이 국제 출원과 관련하여 발행된 통지서를 송부할 것에 동의한다.	**오직 전자적 형태의 통지서만 송부 (서면 통지서는 미발송)**
II-11	출원인 코드	**2-2005-009509-9**
III-1	**출원인 및/또는 발명자**	
III-1-1	이 사람은	**오직 발명자 (inventor only)**
III-1-3	우측 지정국에 관한 발명자	**모든 지정국 (all designated States)**
III-1-4ko	성명	**박상현**
III-1-4en	Name (LAST, First):	**PARK, Sang Hyun**
III-1-5ko	주소	**대한민국** **08004** **서울특별시 양천구 오목로 300, 204동 3701호**
III-1-5en	Address:	**204-3701, 300 Omok-ro,** **Yangcheon-gu, Seoul 08004** **Republic of Korea**

PCT 출원서

(전자적 형태가 원본)

III-2	**출원인 및/또는 발명자**	
III-2-1	이 사람은	**오직 발명자 (inventor only)**
III-2-3	우측 지정국에 관한 발명자	**모든 지정국 (all designated States)**
III-2-4ko	성명	**최종환**
III-2-4en	Name (LAST, First):	**CHOI, Jong Hwan**
III-2-5ko	주소	**대한민국 21090 인천광역시 계양구 봉오대로691번길 4, 103동 409호**
III-2-5en	Address:	**103-409, 4 Bongo-daero 691beon-gil, Gyeyang-gu, Incheon 21090 Republic of Korea**
III-3	**출원인 및/또는 발명자**	
III-3-1	이 사람은	**오직 발명자 (inventor only)**
III-3-3	우측 지정국에 관한 발명자	**모든 지정국 (all designated States)**
III-3-4ko	성명	**서상민**
III-3-4en	Name (LAST, First):	**SEO, Sang Min**
III-3-5ko	주소	**대한민국 03672 서울특별시 서대문구 명지대3길 24-12, 303호**
III-3-5en	Address:	**303, 24-12 Myongjidae 3-gil, Seodaemun-gu, Seoul 03672 Republic of Korea**
III-4	**출원인 및/또는 발명자**	
III-4-1	이 사람은	**오직 발명자 (inventor only)**
III-4-3	우측 지정국에 관한 발명자	**모든 지정국 (all designated States)**
III-4-4ko	성명	**박진욱**
III-4-4en	Name (LAST, First):	**PARK, Jin Uk**
III-4-5ko	주소	**대한민국 01707 서울특별시 노원구 덕릉로 613, 301동 307호**
III-4-5en	Address:	**301-307, 613 Deongneung-ro, Nowon-gu, Seoul 01707 Republic of Korea**

PCT22-0037

PCT 출원서

(전자적 형태가 원본)

IV-1	**대리인 또는 대표자** 아래에 기재된 자는 관할 국제기관에 대하여 우측에 표시된 자격으로 출원인을 대리하는 것으로 선임되었다.	**대리인**
IV-1-1ko	성명	**특허법인 우인**
IV-1-1en	Name:	**WOOIN PATENT & LAW FIRM**
IV-1-2ko	주소	**대한민국 06246 서울특별시 강남구 역삼로 157, 2층 (역삼동,중평빌딩)**
IV-1-2en	Address:	**(Yeoksam-dong, Jungpyeong Bldg.) 2Fl., 157 Yeoksamro, Gangnam-gu, Seoul 06246 Republic of Korea**
IV-1-3	전화번호	**+82-2-541-9841**
IV-1-4	팩스번호	**+82-2-541-9842**
IV-1-5	이메일 주소	**patent@wooinlaw.com**
IV-1-5(a)	이메일 사용동의 수리관청, 국제조사기관, 국제사무국, 국제예비심사기관이 필요 시 이 이메일 주소를 사용하여 이 국제 출원과 관련하여 발행된 통지서를 송부할 것에 동의한다.	**오직 전자적 형태의 통지서만 송부 (서면 통지서는 미발송)**
IV-1-6	대리인 코드	**9-2006-100082-1**
V	**지정국**	
V-1	**본 출원서의 제출로, 규칙 4.9(a)에 따라, 부여될 수 있는 모든 종류의 권리 보호를 위하여, 그리고 해당하는 경우 지역특허 및 국내특허 모두를 위하여 당해 국제출원일에 PCT에 기속되는 모든 체약국이 지정된다.**	
V-2	**V-2란은 출원서 제출시 또는 규칙 26의 2.1에 의해 그 이후 출원서 제6기재란에 위 특정 관련 체약국의 국내 선출원에 대한 우선권주장이 포함되어 있을 경우 당해 체약국의 국내법에 의해 해당 국내 선출원의 효력이 상실되는 것을 방지하기 위한 목적으로 당해 체약국의 지정을 제외하는 데에만 사용될 수 있다 (지정 제외시 이의 취소 불가능).**	**KR**
VI-1	**선국내출원에 대한 우선권 주장**	
VI-1-1	출원일	**2022년 03월 08일 (08.03.2022)**
VI-1-2	출원번호	**10-2022-0029395**
VI-1-3	파리협약 당사국명 또는 WTO 회원국명	**KR**
VI-2	**우선권서류 신청** 수리관청에 대하여 위에 명시된 선출원의 인증등본을 준비하여 국제사무국에 송부하여 줄 것을 신청한다.	**VI-1**
VI-3	**인용에 의한 보완** 조약 제11조(1)(iii)(d) 또는 (e)에서 규정하는 국제출원의 요소, 또는 규칙 20.5(a)에서 규정하는 발명의 설명, 청구범위 또는 도면의 일부, 또는 규칙 20.5의2(a)에서 규정하는 발명의 설명, 청구범위 또는 도면의 요소 또는 일부가 이 국제출원에는 포함되어 있지 않지만 조약 제11조(1)(iii) 규정의 요소 중 하나 이상이 수리관청에 최초로 접수된 날에 우선권주장의 기초가 된 선출원에 완전히 포함되어 있는 경우, 그 요소 또는 부분은 규칙 20.6에 따른 확인을 조건으로, 규칙 20.6과 관련하여 이 국제출원에 있어서 인용에 의해 보완된다.	
VII-1	**국제조사기관(ISA) 선택**	**대한민국 특허청 (ISA/KR)**

VIII	**선언서**	선언서 개수	
VIII-1	발명자의 신원에 관한 선언	**-**	
VIII-2	국제출원일에 특허출원 및 특허를 받을 수 있는 출원인의 자격에 관한 선언	**-**	
VIII-3	국제출원일에 선출원의 우선권을 주장할 수 있는 출원인의 자격에 관한 선언	**-**	
VIII-4	발명자 선언(미국에 대한 지정의 경우에 한함)	**-**	
VIII-5	신규성을 해치지 아니하는 개시 또는 신규성 상실의 예외에 관한 선언	**2**	

PCT 출원서

(전자적 형태가 원본)

VIII-5-1	**선언서:신규성을 해치지 아니하는 개시 또는 신규성 상실의 예외** 신규성을 해치지 아니하는 개시 또는 신규성 상실의 예외에 관한 선언서 (규칙 제4.17조(v) 및 제51조2.1(a)(v)): 성명	**본 국제출원 에 관해** **연세대학교 산학협력단 - UIF (UNIVERSITY INDUSTRY FOUNDATION), YONSEI UNIVERSITY** **국제 출원에서 청구되는 대상이 다음과 같이 개시되었음을 선언한다:**
VIII-5-1 (i)	개시의 종류:	**기타: 공개(학회 발표)**
VIII-5-1 (ii)	개시일:	**2021년 12월 12일 (12.12.2021)**
VIII-5-1 (iii)	개시의 명칭:	**MolBit: De novo Drug Design via Binary Representations of SMILES for avoiding the Posterior Collapse Problem**
VIII-5-1 (iv)	개시의 장소:	
VIII-5-1 (i)	개시의 종류:	**기타: 공개(논문 발표)**
VIII-5-1 (ii)	개시일:	**2022년 01월 14일 (14.01.2022)**
VIII-5-1 (iii)	개시의 명칭:	**MolBit: De novo Drug Design via Binary Representations of SMILES for avoiding the Posterior Collapse Problem**
VIII-5-1 (iv)	개시의 장소:	

PCT22-0037

PCT 출원서

(전자적 형태가 원본)

IX	**체크 리스트**	용지 수	전자적 파일 첨부
IX-1	출원서(선언서 포함)	**5**	✓
IX-2	발명의 설명	**13**	✓
IX-3	청구범위	**4**	✓
IX-4	요약서	**1**	✓
IX-5	도면	**7**	✓
IX-6a	발명의 설명의 서열목록 부분	**-**	**-**
IX-7	용지매수 소계	**30**	
	첨부 항목	서면 첨부	전자적 파일 첨부
IX-8	수수료 계산 용지	**-**	✓
IX-9	개별위임장 원본	**-**	✓
IX-20	**요약서에 수반되어야 할 도면 번호**	**2**	
IX-21	**국제출원의 출원 언어**	**한국어**	
X-1	**출원인, 대리인 또는 대표자의 서명 또는 날인**	**/최성우/**	
X-1-1	성명	**특허법인 우인**	
X-1-2	서명인의 성명	**최성우**	
X-1-3	권한 (출원서를 통해 서명자의 자격이 명백하지 않은 경우에는 그 자격도 표시)	**대표변리사**	

수리관청 전용

10-1	**국제출원으로 제출된 서류의 실제 접수일**	
10-2	**도면**	
10-2-1	접수	
10-2-2	미접수	
10-3	**국제출원으로 제출된 서류를 완성하는 서류 또는 도면의 추후 기간내 제출에 따른 정정된 실제 접수일**	
10-4	**PCT 제11조(2)에 따라 제출이 요구된 보완서로서 기간내 제출된 보완서의 접수일**	
10-5	**국제조사기관(ISA)**	**ISA/KR**
10-6	**조사료 납부시까지 지연된 조사용 사본의 송부**	

국제 사무국 전용

11-1	**국제 사무국의 기록원본 접수일**	

PCT 위임장

(전자적 형태가 원본)

0-1	**PCT 위임장 (특허 협력 조약에 의거하여 제출된 국제 출원) (PCT 규칙 제90.4조)**	
0-1-1	우측에 기재된 바와 같이 작성되었다.	**ePCT-Filing** **Version 4.10.010 MT/FOP 20221109/1.1**

1	**아래에 서명한 출원인**	**연세대학교 산학협력단**
1-1-1	**우측에 기재된 사람을 아래의 자격으로 선임한다.**	**특허법인 우인** **WOOIN PATENT & LAW FIRM** **대한민국** **06246** **서울특별시 강남구 역삼로 157, 2층 (역삼동,중평빌딩)** **(Yeoksam-dong, Jungpyeong Bldg.) 2Fl., 157 Yeoksamro, Gangnam-gu, Seoul 06246** **Republic of Korea**
1-2	**자격**	**대리인**
1-3	**우측 기관에 대하여**	**모든 관할 국제 기관**
1-4	**아래의 국제 출원에 관한 서명의 출원인을 대리함**	
1-4-1	**발명의 명칭**	**딥러닝 기반 분자 설계 방법, 이를 수행하는 장치 및 컴퓨터 프로그램**
1-4-2	**출원인 또는 대리인의 서류참조기호**	**PCT22-0037**
1-4-3	**국제출원번호(이용 가능한 경우)**	
1-4-4	**수리관청**	**대한민국 특허청 (RO/KR)**
1-5	**그리고 아래 서명인을 대신하여 지불하거나 지불받았다.**	

2-1	**출원인 서명**	**/김지현/**
2-1-1	성명	**연세대학교 산학협력단**
2-1-2	서명인의 성명	**김지현**
2-1-3	권한 (출원서를 통해 서명자의 자격이 명백하지 않은 경우에는 그 자격도 표시)	**대표**
3	**일자**	**2022년 12월 12일 (12.12.2022)**

PCT(부속문서 - 수수료 계산용지)

(전자적 형태가 원본)
이 페이지는 국제 출원서의 일부가 아니며 페이지수에 포함되지 않는다

0	**수리관청 전용**	
0-1	국제출원번호	
0-2	수리관청의 우편 소인 일자	

0-4	**Form PCT/RO/101 (부속문서) PCT 수수료 계산 용지**				
0-4-1	우측에 기재된 바와 같이 작성되었다.		**ePCT-Filing Version 4.10.010 MT/FOP 20221109/1.1**		
0-9	**출원인 또는 대리인의 서류참조기호**		**PCT22-0037**		
2	**출원인**		**연세대학교 산학협력단**		
12	**규정 수수료 계산**		수수료 금액/계수	총 금액 (CHF)	총 금액 (KRW)
12-1	송달료	**T**	➪		**45000**
12-2-1	조사료	**S**	➪		**450000**
12-2-2	국제조사기관		**KR**		
12-3	국제 출원 수수료 최초 30장	**i1**	**1330**		
12-4	최초 30장 초과 장수		**0**		
12-5	최초 30장 초과 1장당 추가 수수료	**(X)**	**0**		
12-6	총 추가금액	**i2**	**0**		
12-7	i1 + i2 =	**i**	**1330**		
12-12	XML 전자출원 감면	**R**	**-300**		
12-13	총 국제출원 수수료(i-R)	**I**	➪	**1030**	
12-19	**총 금액(T+S+I+P)**		➪	**1030**	**495000**
12-21	**결제 방법**		**기타 : 현재 지불금액이 없습니다.**		

명세서

발명의 명칭: 딥러닝 기반 분자 설계 방법, 이를 수행하는 장치 및 컴퓨터 프로그램

기술분야

[1] 본 발명은 딥러닝 기반 분자 설계 방법, 이를 수행하는 장치 및 컴퓨터 프로그램에 관한 것으로서, 더욱 상세하게는 딥러닝(deep learning)을 이용하여 분자를 설계하는, 방법, 장치 및 컴퓨터 프로그램에 관한 것이다.

배경기술

[2] 변분 오토인코더(variational autoencoder, VAE)는 성공적인 생성 모델이지만, 순환 신경망(recurrent neural network, RNN) 등과 같은 자기회귀 모델(autoregressive model)과 쌍을 이룰 때, 후방 붕괴(posterior collapse) 문제가 발생된다. 이 문제는 잠재 화학 공간(latent chemical space)의 학습된 확률 분포에서 샘플링된 데이터를 무시하고, 시퀀스의 이전 시간 단계의 정보로만 시퀀스 생성을 진행할 때 관찰된다. RNN-VAE 모델의 목적은 잠재 공간의 다양한 화합물에 대한 충분한 정보를 학습하는 것이기 때문에 붕괴는 바람직하지 않다.

[3] 이와 같은 후방 붕괴 문제를 해결하기 위한 종래의 방법들인, KL-annealing 방식은 후방 붕괴에 대한 간접적인 접근 방식이며, 보조 전략 방식은 속성을 정확하게 예측하지 못하기 때문에 효과적이지 않을 수 있다.

발명의 상세한 설명

기술적 과제

[4] 본 발명이 이루고자 하는 목적은, 검벨-소프트맥스(Gumbel-Softmax) 함수를 활용하여 변분 오토인코더(variational autoencoder, VAE)를 사전 학습하고, 사전 학습된 변분 오토인코더(VAE)와 유전 알고리즘(genetic algorithm, GA)을 이용하여 사용자의 요청 분자 특성에 대응되는 분자를 설계하는, 딥러닝 기반 분자 설계 방법, 이를 수행하는 장치 및 컴퓨터 프로그램을 제공하는 데 있다.

[5] 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

과제 해결 수단

[6] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법은, 사용자에 의해 분자 특성을 입력받는 단계; 및 사전 학습된 변분 오토인코더(variational autoencoder, VAE) 및 유전 알고리즘(genetic algorithm, GA)을 기반으로, 상기 분자 특성에 대응되는 목표 분자를 설계하는 단계;를 포함하며, 상기 변분 오토인코더(VAE)는, 입력되는 분자 구조의 SMILES(simplified molecular-input line-entry system) 표현을 검벨-소프트맥스(Gumbel-Softmax)를 이용하여 분자 구조의 이진 잠재

벡터(binary latent vector) 표현으로 변환하는 검벨 인코더(Gumbel encoder) 및 분자 구조의 이진 잠재 벡터 표현을 분자 구조의 SMILES 표현으로 변환하는 스토캐스틱 디코더(stochastic decoder)를 포함한다.

[7] 여기서, SMILES로 표현된 학습 데이터 세트를 기반으로, 상기 검벨-소프트맥스를 이용한 상기 변분 오토인코더(VAE)를 사전 학습하는 단계;를 더 포함할 수 있다.

[8] 여기서, 상기 분자 설계 단계는, 상기 변분 오토인코더(VAE) 및 상기 유전 알고리즘(GA)을 이용하여, 상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더의 도메인에서 상기 분자 특성을 가지는 분자에 대응되는 이진 잠재 벡터 표현을 탐색하고, 탐색된 이진 잠재 벡터 표현을 기반으로 상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더를 통해 상기 입력된 분자 특성에 대응되는 상기 목표 분자를 설계하는 것으로 이루어질 수 있다.

[9] 여기서, 상기 변분 오토인코더(VAE)의 상기 검벨 인코더 q(zlx)는, 게이트 순환 유닛(gated recurrent unit, GRU) 및 검벨-소프트맥스를 사용하여 입력 SMILES 스트링(string) x를 이진 잠재 벡터 z에 매핑할 수 있다.

[10] 여기서, 상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더 p(xlz)는, 상기 스토캐스틱 디코더에서 게이트 순환 유닛(GRU) 모델의 초기 히든 상태(hidden state)로 인코딩된 이진 잠재 벡터 z를 사용하여 입력 SMILES 스트링 x를 복원할 수 있다.

[11] 여기서, 상기 변분 오토인코더(VAE) 사전 학습 단계는, 상기 학습 데이터 세트를 기반으로, 재구성 에러(reconstruction error) 항 및 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD) 항을 포함하는 ELBO(evidence lower bound) 손실 함수(loss function)를 최소화하는 것을 통해, 상기 변분 오토인코더(VAE)를 학습하는 것으로 이루어질 수 있다.

[12] 여기서, 상기 ELBO 손실 함수 L(x)는, [식 1]을 나타내고, 상기 [식 1]은,

[13] $$L(x) = -\log(p_\theta(x)) + KL[q_\phi(z|x)||p(z)]$$ 이며,

상기 θ는, 상기 스토캐스틱 디코더의 학습 파라미터(trainable parameter)를 나타내고, 상기 Φ는, 상기 검벨 인코더의 학습 파라미터를 나타내며, 상기 $-\log(p_\theta(x))$ 는, 상기 재구성 에러 항을 나타내고, 입력 SMILES 스트링 x가 얼마나 완벽하게 복원되는지를 평가하는 것을 나타내는 $p_\theta(x)$ 의 음의 로그-우도(log-likelihood)를 나타내며, 상기 $KL[q_\phi(z|x)||p(z)]$ 은, 상기 쿨백-라이블러 발산(KLD) 항을 나타내고, 카테고리형 분포(categorical distribution) p(z)와 상기 검벨 인코더 $q_\phi(z|x)$ 의 출력 간의 차이를 측정할 수 있다.

[14] 여기서, 상기 쿨백-라이블러 발산(KLD) 항은, 이산 균일 분포(discrete uniform distribution)의 확률 질량 함수(probability mass function)를 상기 카테고리형 분포 p(z)에 대입하여, [식 2]를 통해 계산되며, 상기 [식 2]는,

[15]

$$KL[q_\phi(z|x)||p(z)] = q_\phi(z|x)\log(q_\phi(z|x)*K)$$

[16] 이고, 상기 K는, 카테고리의 개수를 나타내며, 잠재 공간(latent space)의 차원(dimension)과 동일할 수 있다.

[17] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 컴퓨터 프로그램은 컴퓨터 판독 가능한 저장 매체에 저장되어 상기한 딥러닝 기반 분자 설계 방법 중 어느 하나를 컴퓨터에서 실행시킨다.

[18] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 장치는, 딥러닝을 이용하여 분자를 설계하는 장치로서, 딥러닝을 이용하여 분자를 설계하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및 상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 딥러닝을 이용하여 분자를 설계하기 위한 동작을 수행하는 하나 이상의 프로세서;를 포함하며, 상기 프로세서는, 사용자에 의해 분자 특성을 입력받고, 사전 학습된 변분 오토인코더(variational autoencoder, VAE) 및 유전 알고리즘(genetic algorithm, GA)을 기반으로, 상기 분자 특성에 대응되는 목표 분자를 설계하며, 상기 변분 오토인코더(VAE)는, 입력되는 분자 구조의 SMILES(simplified molecular-input line-entry system) 표현을 검벨-소프트맥스(Gumbel-Softmax)를 이용하여 분자 구조의 이진 잠재 벡터(binary latent vector) 표현으로 변환하는 검벨 인코더(Gumbel encoder) 및 분자 구조의 이진 잠재 벡터 표현을 분자 구조의 SMILES 표현으로 변환하는 스토캐스틱 디코더(stochastic decoder)를 포함한다.

[19] 여기서, 상기 프로세서는, SMILES로 표현된 학습 데이터 세트를 기반으로, 상기 검벨-소프트맥스를 이용한 상기 변분 오토인코더(VAE)를 사전 학습할 수 있다.

[20] 여기서, 상기 프로세서는, 상기 변분 오토인코더(VAE) 및 상기 유전 알고리즘(GA)을 이용하여, 상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더의 도메인에서 상기 분자 특성을 가지는 분자에 대응되는 이진 잠재 벡터 표현을 탐색하고, 탐색된 이진 잠재 벡터 표현을 기반으로 상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더를 통해 상기 입력된 분자 특성에 대응되는 상기 목표 분자를 설계할 수 있다.

[21] 여기서, 상기 변분 오토인코더(VAE)의 상기 검벨 인코더 q(zlx)는, 게이트 순환 유닛(gated recurrent unit, GRU) 및 검벨-소프트맥스를 사용하여 입력 SMILES 스트링(string) x를 이진 잠재 벡터 z에 매핑하며, 상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더 p(xlz)는, 상기 스토캐스틱 디코더에서 게이트 순환 유닛(GRU) 모델의 초기 히든 상태(hidden state)로 인코딩된 이진 잠재 벡터 z를

사용하여 입력 SMILES 스트링 x를 복원하고, 상기 프로세서는, 상기 학습 데이터 세트를 기반으로, 재구성 에러(reconstruction error) 항 및 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD) 항을 포함하는 ELBO(evidence lower bound) 손실 함수(loss function)를 최소화하는 것을 통해, 상기 변분 오토인코더(VAE)를 학습할 수 있다.

발명의 효과

[22] 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법, 이를 수행하는 장치 및 컴퓨터 프로그램에 의하면, 검벨-소프트맥스(Gumbel-Softmax) 함수를 활용하여 변분 오토인코더(variational autoencoder, VAE)를 사전 학습하고, 사전 학습된 변분 오토인코더(VAE)와 유전 알고리즘(genetic algorithm, GA)을 이용하여 사용자의 요청 분자 특성에 대응되는 분자를 설계함으로써, 분자 구조의 SMILES 표현을 이진 벡터 표현으로 변환할 수 있어 잠재 공간의 탐색 범위를 유한한 공간으로 제약할 수 있다.

[23] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[24] 도 1은 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 장치를 설명하기 위한 블록도이다.

[25] 도 2는 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법을 설명하기 위한 흐름도이다.

[26] 도 3은 본 발명의 바람직한 실시예에 따른 변분 오토인코더를 설명하기 위한 도면이다.

[27] 도 4는 본 발명의 바람직한 실시예에 따른 SMILES 사전 학습을 위한 GumbelVAE 및 특성 최적화를 위한 유전 알고리즘을 활용하는 MolBit 파이프라인을 설명하기 위한 도면이다.

[28] 도 5는 본 발명의 바람직한 실시예에 따른 이진 표현에 의해 SMILES 스트링을 인코딩 및 디코딩하는 MolBit 생성기를 설명하기 위한 도면이다.

[29] 도 6은 본 발명의 바람직한 실시예에 따른 분자 특성에 대한 최적의 이진 표현을 찾기 위해 이진 잠재 공간을 탐색하는 MolBit 최적기를 설명하기 위한 도면이다.

[30] 도 7은 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법의 성능을 설명하기 위한 도면으로, 분자 특성의 후방 붕괴에 대한 평가 결과를 나타내며, (a) 종래 기술(GaussianVAE)에 대한 표준 가우시안 분포에서 샘플링된 10개의 랜덤 잠재 벡터이고, (b) 내지 (d)는 10개의 랜덤 가우시안 잠재 벡터로부터 추정된 QED(quantitative estimate of drug-likeness), LogP(partition coefficient) 및 SAscore(synthetic accessibility score) 각각의 확률 분포이며, (e)는 본

발명(GumbleVAE)에 대한 베르누이 분포에서 샘플링된 10개의 랜덤 이진 벡터이고, (f) 내지 (h)는 10개의 랜덤 이진 잠재 벡터로부터 추정된 QED, LogP 및 SAscore 각각의 확률 분포이다.

[31] 도 8은 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법의 성능을 설명하기 위한 도면으로, penalized logP 최적화를 가지는 본 발명(MolBit)에서의 유전 알고리즘 결과를 나타내며, (a)는 세대(generation)에 따른 적합도(fitness) 점수이고, (b)는 유전 알고리즘을 적용하지 않은 본 발명(MolBit before GA) 및 유전 알고리즘에 의해 penalized logP를 최대화하도록 최적화된 본 발명(MolBit after GA) 각각에 의해 생성된 30,000개의 분자의 확률 분포 비교이다.

발명의 실시를 위한 형태

[32] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세히 설명한다. 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 게시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 게시가 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. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.

[33] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[34] 본 명세서에서 "제1", "제2" 등의 용어는 하나의 구성 요소를 다른 구성 요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예컨대, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.

[35] 본 명세서에서 각 단계들에 있어 식별부호(예컨대, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[36] 본 명세서에서, "가진다", "가질 수 있다", "포함한다" 또는 "포함할 수 있다" 등의 표현은 해당 특징(예컨대, 수치, 기능, 동작, 또는 부품 등의 구성 요소)의 존재를 가리키며, 추가적인 특징의 존재를 배제하지 않는다.

[37] 이하에서 첨부한 도면을 참조하여 본 발명에 따른 딥러닝 기반 분자 설계 방법, 이를 수행하는 장치 및 컴퓨터 프로그램의 바람직한 실시예에 대해 상세하게 설명한다.

[38] 먼저, 도 1을 참조하여 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 장치에 대하여 설명한다.

[39] 도 1은 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 장치를 설명하기 위한 블록도이다.

[40] 도 1을 참조하면, 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 장치(100)는 검벨-소프트맥스(Gumbel-Softmax) 함수를 활용하여 변분 오토인코더(variational autoencoder, VAE)를 사전 학습하고, 사전 학습된 변분 오토인코더(VAE)와 유전 알고리즘(genetic algorithm, GA)을 이용하여 사용자의 요청 분자 특성에 대응되는 분자를 설계할 수 있다.

[41] 이를 위해, 딥러닝 기반 분자 설계 장치(100)는 하나 이상의 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130) 및 통신 버스(150)를 포함할 수 있다.

[42] 프로세서(110)는 딥러닝 기반 분자 설계 장치(100)가 동작하도록 제어할 수 있다. 예컨대, 프로세서(110)는 컴퓨터 판독 가능한 저장 매체(130)에 저장된 하나 이상의 프로그램(131)을 실행할 수 있다. 하나 이상의 프로그램(131)은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 컴퓨터 실행 가능 명령어는 프로세서(110)에 의해 실행되는 경우 딥러닝 기반 분자 설계 장치(100)로 하여금 딥러닝을 이용하여 분자를 설계하기 위한 동작을 수행하도록 구성될 수 있다.

[43] 컴퓨터 판독 가능한 저장 매체(130)는 딥러닝을 이용하여 분자를 설계하기 위한 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능한 저장 매체(130)에 저장된 프로그램(131)은 프로세서(110)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능한 저장 매체(130)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 딥러닝 기반 분자 설계 장치(100)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[44] 통신 버스(150)는 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130)를 포함하여 딥러닝 기반 분자 설계 장치(100)의 다른 다양한 컴포넌트들을 상호 연결한다.

[45] 딥러닝 기반 분자 설계 장치(100)는 또한 하나 이상의 입출력 장치를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(170) 및 하나 이상의 통신 인터페이스(190)를 포함할 수 있다. 입출력 인터페이스(170) 및 통신 인터페이스(190)는 통신 버스(150)에 연결된다. 입출력 장치(도시하지 않음)는

입출력 인터페이스(170)를 통해 딥러닝 기반 분자 설계 장치(100)의 다른 컴포넌트들에 연결될 수 있다.

[46] 그러면, 도 2 및 도 3을 참조하여 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법에 대하여 설명한다.

[47] 도 2는 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법을 설명하기 위한 흐름도이고, 도 3은 본 발명의 바람직한 실시예에 따른 변분 오토인코더를 설명하기 위한 도면이다.

[48] 도 2를 참조하면, 딥러닝 기반 분자 설계 장치(100)의 프로세서(110)는 SMILES(simplified molecular-input line-entry system)로 표현된 학습 데이터를 기반으로, 검벨-소프트맥스를 이용한 변분 오토인코더(VAE)를 사전 학습할 수 있다(S110).

[49] 여기서, 변분 오토인코더(VAE)는 도 3에 도시된 바와 같이, 검벨 인코더(Gumbel encoder) 및 스토캐스틱 디코더(stochastic decoder)를 포함할 수 있다. 검벨 인코더는 입력되는 분자 구조의 SMILES 표현을 검벨-소프트맥스를 이용하여 분자 구조의 이진 잠재 벡터(binary latent vector) 표현으로 변환할 수 있다. 스토캐스틱 디코더는 분자 구조의 이진 잠재 벡터 표현을 분자 구조의 SMILES 표현으로 변환할 수 있다.

[50] 이때, 프로세서(110)는 학습 데이터 세트를 기반으로, 재구성 에러(reconstruction error) 항 및 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD) 항을 포함하는 ELBO(evidence lower bound) 손실 함수(loss function)를 최소화하는 것을 통해, 변분 오토인코더(VAE)를 학습할 수 있다.

[51] 여기서, 프로세서(110)는 순환 어닐링 스케쥴(cyclical annealing schedule) 알고리즘을 이용하여 쿨백-라이블러 발산(KLD) 항을 주기적으로 온/오프할 수 있다.

[52] 이후, 프로세서(110)는 사용자에 의해 분자 특성을 입력받을 수 있다(S120).

[53] 그런 다음, 프로세서(110)는 사전 학습된 변분 오토인코더(VAE) 및 유전 알고리즘(GA)을 기반으로, 분자 특성에 대응되는 목표 분자를 설계할 수 있다(S130).

[54] 즉, 프로세서(110)는 변분 오토인코더(VAE) 및 유전 알고리즘(GA)을 이용하여, 변분 오토인코더(VAE)의 스토캐스틱 디코더의 도메인에서 분자 특성을 가지는 분자에 대응되는 이진 잠재 벡터 표현을 탐색할 수 있다.

[55] 여기서, 유전 알고리즘(GA)에 따른 염색체는 베르누이 분포(Bernoulli distribution)에서 샘플링된 이진 잠재 벡터를 나타낼 수 있다. 유전 알고리즘(GA)에 따른 적합도(fitness) 점수는 각 염색체(즉, 각 이진 잠재 벡터)에서 생성된 분자의 화학적 특성 값의 평균을 나타낼 수 있다. 유전 알고리즘(GA)에 따른 교차(crossover) 작업은 두개의 부모 염색체(즉, 두개의 이진 잠재 벡터)가 두개의 파트로 나누어 지고, 나누어진 파트가 교환되는 것을 나타낼 수 있다. 유전 알고리즘(GA)에 따른 돌연변이(mutation) 작업은

염색체(즉, 이진 잠재 벡터)의 랜덤 비트를 온하거나 오프하는 것을 나타낼 수 있다.

[56] 그리고, 프로세서(110)는 탐색된 이진 잠재 벡터 표현을 기반으로 변분 오토인코더(VAE)의 스토캐스틱 디코더를 통해 입력된 분자 특성에 대응되는 목표 분자를 설계할 수 있다.

[57] 그러면, 도 4 내지 도 6을 참조하여 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법에 대하여 보다 자세하게 설명한다.

[58] 도 4는 본 발명의 바람직한 실시예에 따른 SMILES 사전 학습을 위한 GumbelVAE 및 특성 최적화를 위한 유전 알고리즘을 활용하는 MolBit 파이프라인을 설명하기 위한 도면이고, 도 5는 본 발명의 바람직한 실시예에 따른 이진 표현에 의해 SMILES 스트링을 인코딩 및 디코딩하는 MolBit 생성기를 설명하기 위한 도면이며, 도 6은 본 발명의 바람직한 실시예에 따른 분자 특성에 대한 최적의 이진 표현을 찾기 위해 이진 잠재 공간을 탐색하는 MolBit 최적기를 설명하기 위한 도면이다.

[59] SMILES 생성 및 분자 특성 최적화를 위한 본 발명에 따른 MolBit는 도 4에 도시된 바와 같이, 검벨-소프트맥스를 활용하여 입력 SMILES 스트링(string)을 변분 오토인코더(VAE)의 이진 잠재 표현에 매핑하고, 유전 알고리즘(GA)을 활용하여 원하는 분자 특성을 가진 분자 생성을 위한 이진 잠재 공간(binary latent space)을 탐색할 수 있다.

[60] A. 검벨-소프트맥스

[61] 검벨-소프트맥스는 심플렉스(simplex)에 대한 연속 분포인 카테고리형 분포(categorical distribution)의 샘플을 근사하는 실수 벡터에 대한 미분가능 함수(differentiable function)일 수 있다. 클래스 확률 $\pi_1, \pi_2, ..., \pi_k$인 k-차원 원-핫 벡터(one-hot vector)의 경우, 검벨-소프트맥스는 아래의 [수학식 1]과 같이 정의된 근사 k-차원 실수 벡터 $y=(y_1, y_2, ...,y_k)$를 계산할 수 있다.

[62] [수식1]

$$y_i = \frac{\exp((\log \pi_i + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log \pi_j + g_i)/\tau)} \quad for \;\; i = 1, \ldots, k$$

[63] 여기서, $g_1, ..., g_k$는 표준 검벨 분포로부터 추출된 독립적인 동일 분포 샘플(independent identically distributed samples)을 나타낼 수 있다. τ는 0보다 크며, 신뢰 제어(confidence control)를 위한 온도 파라미터(temperature parameter)를 나타낼 수 있다. 검벨-소프트맥스 y의 출력은 $\tau \rightarrow 0$과 같이 원-핫 벡터가 되며, 온도가 증가할수록 균일(uniform)에 가까워 질 수 있다. 모델 훈련 단계에서, 온도는 높은 양의 값으로 초기화되고, 0이 아닌 작은 값으로 단조롭게 감소할 수

있다.

[64] B. 본 발명에 따른 GumvelVAE

[65] 본 발명에 따른 MolBit의 생성 모델인 GumvelVAE는 2개의 파트인 검벨 인코더와 스토캐스틱 디코더로 구성될 수 있다. 검벨 인코더 q(z|x)는 게이트 순환 유닛(gated recurrent unit, GRU) 및 검벨-소프트맥스를 사용하여 입력 SMILES 스트링 x를 이진 잠재 벡터 z에 매핑할 수 있다. 스토캐스틱 디코더 p(x|z)는 도 5에 도시된 바와 같이, 스토캐스틱 디코더에서 게이트 순환 유닛(GRU) 모델의 초기 히든 상태(hidden state)로 인코딩된 이진 잠재 벡터 z를 사용하여 입력 SMILES 스트링 x를 복원할 수 있다.

[66] 본 발명에 따른 GumvelVAE는 재구성 에러 항 및 쿨백-라이블러 발산(KLD) 항을 포함하는 ELBO 손실 함수를 최소화하는 것에 의해 학습될 수 있다. ELBO 손실 함수 L(x)는 아래의 [수학식 2]일 수 있다.

[67] [수식2]

$$L(x) = -\log(p_\theta(x)) + KL[q_\phi(z|x)||p(z)]$$

[68] 여기서, θ는 스토캐스틱 디코더의 학습 파라미터(trainable parameter)를 나타낼 수 있다. Φ는 검벨 인코더의 학습 파라미터를 나타낼 수 있다.

[69] $-\log(p_0(x))$는 재구성 에러 항을 나타내고, 입력 SMILES 스트링 x가 얼마나 완벽하게 복원되는지를 평가하는 것을 나타내는 $p_0(x)$의 음의 로그-우도(log-likelihood)를 나타낼 수 있다. $KL[q_\Phi(z|x)\|p(z)]$은 쿨백-라이블러 발산(KLD) 항을 나타내고, 카테고리형 분포 p(z)와 검벨 인코더 $q_\Phi(z|x)$ 의 출력 간의 차이를 측정할 수 있다. 쿨백-라이블러 발산(KLD) 항은 이산 균일 분포(discrete uniform distribution)의 확률 질량 함수(probability mass function)를 카테고리형 분포 p(z)에 대입하여, 아래의 [수학식 3]을 통해 계산될 수 있다.

[70] [수식3]

$$KL[q_\phi(z|x)||p(z)] = q_\phi(z|x)\log(q_\phi(z|x)*K)$$

[71] 여기서, K는 카테고리의 개수를 나타내며, 잠재 공간의 차원(dimension)과 동일할 수 있다.

[72] 정리하면, 본 발명에 따른 GumbelVAE의 학습 과정은 아래의 [표 1]과 같은 알고리즘 1로 표현될 수 있다.

[73] [표1]

알고리즘 1: GumbelVAE의 학습 과정

Input: Tokenized SMILES sequence $x = (x_1, x_2, \cdots, x_T)$, dimension of latent space K, a temperature τ, a learning rate η

Output: None

```
c1 ← 0                                   // initialize a hidden state
L ← 0                                    // initialize a loss
for t=1 to T do
    ot, ct+1 ← GRU(xt, ct; Θ)            // encoder GRU
end for
π ← Softmax(cT+1)
z ← GumbelSoftmax(π)                     // latent vector
h1 ← Dense(z)
for t=1 to T-1 do
    yt, ht+1 ← GRU(xt, ht; ϕ)            // decoder GRU
    pt ← Softmax(yt)
    L ← L + NLL(pt, xt+1)                // negative log-likelihood
end for
L ← L+π*log(π*K)                         // KL divergence
ΔΘ, Δϕ ← Adam(L)                         // gradient descent
Θ ← Θ-η*ΔΘ                               // encoder update
ϕ ← ϕ-η*Δϕ                               // decoder update
```

[74] C. KL-어닐링(KL-annealing) 스케쥴러

[75] KL-베니싱(KL-vanishing) 문제는 자동회귀(autoregressive) 디코더를 가지는 변분 오토인코더(VAE)에 대한 학습 문제일 수 있다. 이 베니싱(vanishing) 문제는 디코더가 인코더의 잠재 벡터 없이도 출력 시퀀스를 잘 생성할 수 있을 때 발생할 수 있다. KL-베니싱(KL-vanishing)에 대한 간단한 솔루션은 재구성 에러 항과 쿨백-라이블러 발산(KLD) 항의 가중치 균형을 맞추는 KL-어닐링(KL-annealing) 기법일 수 있다. KL-베니싱(KL-vanishing) 문제를 완화하기 위해, 본 발명은 주기적으로 쿨백-라이블러 발산(KLD) 항을 온/오프하는 순환 어닐링 스케쥴(cyclical annealing schedule) 알고리즘을 활용하여, 잠재 벡터에 대한 디코더의 종속성을 효과적으로 개선할 수 있다.

[76] D. 유전 알고리즘(GA)

[77] 유전 알고리즘(GA)은 잘-알려진 메타휴리스틱(metaheuristic) 최적화 알고리즘일 수 있다. 유전 알고리즘(GA)은 랜덤 염색체-유사 데이터로 0세대부터 시작하여, 사용자에 의해 정의된 적합도(fitness) 점수를 사용하여 염색체의 순위를 매기고, 점수가 높은 부모 염색체를 선택하고, 다음 세대를 위한 부모 염색체에 교차(crossover)와 돌연변이(mutation)와 같은 재조합 연산자를 적용할 수 있다. 이러한 평가 및 재조합 과정을 반복적으로 수행한 후, 마지막 염색체가 최적의 염색체로 결정될 수 있다.

[78] 본 발명에 따른 MolBit에서는, 베르누이 분포(Bernoulli distribution)에서 샘플링된 이진 잠재 벡터가 염색체 역할을 하며, 적합도(fitness) 점수는 각

염색체에서 생성된 분자의 화학적 특성 값의 평균으로 정의될 수 있다. 적합도(fitness) 점수는 QED(quantitative estimate of drug-likeness), LogP(partition coefficient) 등을 이용할 수 있다. 교차(crossover) 작업에서는, 두개의 부모 염색체가 두개의 파트로 나누어 지고, 그 파트가 교환될 수 있다. 돌연변이(mutation) 작업은 염색체의 랜덤 비트를 온하거나 오프할 수 있다. 유전 알고리즘(GA)을 통한 본 발명에 따른 MolBit의 분자 최적화 프로세스는 도 6에 도시된 바와 같이 이루어질 수 있다.

[79] 그러면, 도 7 및 도 8을 참조하여 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법의 성능에 대하여 설명한다.

[80] 도 7은 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법의 성능을 설명하기 위한 도면으로, 분자 특성의 후방 붕괴에 대한 평가 결과를 나타내며, (a)는 종래 기술(GaussianVAE)에 대한 표준 가우시안 분포에서 샘플링된 10개의 랜덤 잠재 벡터이고, (b) 내지 (d)는 10개의 랜덤 가우시안 잠재 벡터로부터 추정된 QED(quantitative estimate of drug-likeness), LogP(partition coefficient) 및 SAscore(synthetic accessibility score) 각각의 확률 분포이며, (e)는 본 발명(GumbleVAE)에 대한 베르누이 분포에서 샘플링된 10개의 랜덤 이진 벡터이고, (f) 내지 (h)는 10개의 랜덤 이진 잠재 벡터로부터 추정된 QED, LogP 및 SAscore 각각의 확률 분포이다.

[81] 도 7을 참조하면, 종래 기술(GaussianVAE)은 잠재 벡터 사이의 분자 특성에 대해 구별할 수 없는 분포를 생성한다. 이에 반면, 본 발명(GumbleVAE)은 다양한 잠재 벡터에 대해 다양한 분포를 생성하여, 후방 붕괴를 피할 수 있음을 확인할 수 있다.

[82] - GaussianVAE : 이 베이스라인 모델은 SMILES 스트링 생성을 위한 표준 RNN-VAE 구조를 가지고 있다. GaussianVAE는 잠재 공간이 표준 가우시안 분포에 근접한다는 점을 제외하고, 본 발명에 따른 MolBit의 GumbelVAE와 동일하다.

[83] 도 8은 본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법의 성능을 설명하기 위한 도면으로, penalized logP 최적화를 가지는 본 발명(MolBit)에서의 유전 알고리즘 결과를 나타내며, (a)는 세대(generation)에 따른 적합도(fitness) 점수이고, (b)는 유전 알고리즘을 적용하지 않은 본 발명(MolBit before GA) 및 유전 알고리즘에 의해 penalized logP를 최대화하도록 최적화된 본 발명(MolBit after GA) 각각에 의해 생성된 30,000개의 분자의 확률 분포 비교이다.

[84] 도 8을 참조하면, penalized logP 점수가 500 세대 후에 수렴됨을 확인할 수 있고, 유전 알고리즘(GA)을 적용하지 않은 본 발명(MolBit before GA) 및 유전 알고리즘(GA)에 의해 최적화된 본 발명(MolBit after GA) 각각으로부터 샘플링된 30,000개 분자에 대한 penalized logP 분포의 상대적 이동을 확인할 수 있다.

[85] < 본 발명의 특징 >

[86] - 검벨-소프트맥스 함수를 이용한 SMILES 표현의 이진 벡터 표현 방법

[87] 종래 기술에도 SMILES 표현을 이진 벡터로 변환하는 내용이 있으나, 이는 SMILES를 구성하는 ASCII 코드를 각각의 이진수로 변환하는 것을 통해 이진 벡터 표현을 구하고 있다. 즉, 종래 기술에서는 사용자가 지정한 알고리즘에 따라 이진 벡터 표현을 얻고 있다. 이에 반면, 본 발명은 딥러닝을 통해 이진 벡터 변환 알고리즘을 학습하고 있기 때문에 종래 기술과 차별점이 있다.

[88] 또한, 기존의 연구들[논문 01 ~ 논문 04]은 SMILES 표현 생성을 위해 변분 오토인코더(VAE) 활용시 검벨-소프트맥스 함수를 사용하고 있지 않아, 본 발명과 같이 이진 벡터 표현을 학습할 수 없다. 이에 따라, 본 발명은 이진 벡터 표현 변환을 통해 최적의 분자 구조를 탐색할 수 있는 점에서 기존의 연구들과 차별점이 있다.

[89] [논문 01] Gomez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276. 참조

[90] [논문 02] Winter, Robin, et al. "Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations." Chemical science 10.6 (2019): 1692-1701. 참조

[91] [논문 03] Griffiths, Ryan-Rhys, and Jose Miguel Hernandez-Lobato. "Constrained Bayesian optimization for automatic chemical design using variational autoencoders." Chemical science 11.2 (2020): 577-586. 참조

[92] [논문 04] Yan, Chaochao, et al. "Re-balancing variational autoencoder loss for molecule sequence generation." Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 2020. 참조

[93] - 유전 알고리즘(GA)을 활용한 최적 분자 구조 탐색 방법

[94] 검벨-소프트맥스 함수를 사용하지 않는 기존의 연구들[논문 01 ~ 논문 04]은 변분 오토인코더(VAE)의 디코더 도메인에서 최적 분자 구조 표현에 대응되는 실수값의 잠재 벡터를 탐색하기 위해, 모두 베이지안 최적화(bayesian optimization) 알고리즘을 사용하고 있다. 이러한 실수 공간 도메인은 이진 벡터 공간 대비 몹시 광범위하여, 임의의 잠재 벡터에 대하여 SMILES 표현을 생성해야 하는 변분 오토 인코더(VAE)의 디코더 학습이 쉽지 않고, 베이지안 최적화 알고리즘을 사용하더라도 최적의 잠재 벡터 탐색을 어렵게 한다. 이에 반면, SMILES 표현을 이진 벡터 표현으로 변환할 수 있는 본 발명은 잠재 공간의 탐색 범위를 유한한 공간으로 제약하기에, 효과적인 디코더 학습 및 최적 잠재 벡터 탐색을 가능케 한다. 나아가, 효율적인 탐색 방법으로 유전 알고리즘(GA)을 사용하는 점에서 종래 기술과 차별점이 있다.

[95] 본 실시예들에 따른 동작은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능한 저장 매체에 기록될 수

있다. 컴퓨터 판독 가능한 저장 매체는 실행을 위해 프로세서에 명령어를 제공하는데 참여한 임의의 매체를 나타낸다. 컴퓨터 판독 가능한 저장 매체는 프로그램 명령, 데이터 파일, 데이터 구조 또는 이들의 조합을 포함할 수 있다. 예컨대, 자기 매체, 광기록 매체, 메모리 등이 있을 수 있다. 컴퓨터 프로그램은 네트워크로 연결된 컴퓨터 시스템 상에 분산되어 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수도 있다. 본 실시예를 구현하기 위한 기능적인(Functional) 프로그램, 코드, 및 코드 세그먼트들은 본 실시예가 속하는 기술 분야의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.

[96] 본 실시예들은 본 실시예의 기술 사상을 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

[97] [부호의 설명]

[98] 100 : 딥러닝 기반 분자 설계 장치,

[99] 110 : 프로세서,

[100] 130 : 컴퓨터 판독 가능한 저장 매체,

[101] 131 : 프로그램,

[102] 150 : 통신 버스,

[103] 170 : 입출력 인터페이스,

[104] 190 : 통신 인터페이스

청구범위

[청구항 1] 사용자에 의해 분자 특성을 입력받는 단계; 및
사전 학습된 변분 오토인코더(variational autoencoder, VAE) 및 유전 알고리즘(genetic algorithm, GA)을 기반으로, 상기 분자 특성에 대응되는 목표 분자를 설계하는 단계;
를 포함하며,
상기 변분 오토인코더(VAE)는,
입력되는 분자 구조의 SMILES(simplified molecular-input line-entry system) 표현을 검벨-소프트맥스(Gumbel-Softmax)를 이용하여 분자 구조의 이진 잠재 벡터(binary latent vector) 표현으로 변환하는 검벨 인코더(Gumbel encoder) 및 분자 구조의 이진 잠재 벡터 표현을 분자 구조의 SMILES 표현으로 변환하는 스토캐스틱 디코더(stochastic decoder)를 포함하는,
딥러닝 기반 분자 설계 방법.

[청구항 2] 제1항에서,
SMILES로 표현된 학습 데이터 세트를 기반으로, 상기 검벨-소프트맥스를 이용한 상기 변분 오토인코더(VAE)를 사전 학습하는 단계;
를 더 포함하는 딥러닝 기반 분자 설계 방법.

[청구항 3] 제2항에서,
상기 분자 설계 단계는,
상기 변분 오토인코더(VAE) 및 상기 유전 알고리즘(GA)을 이용하여, 상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더의 도메인에서 상기 분자 특성을 가지는 분자에 대응되는 이진 잠재 벡터 표현을 탐색하고, 탐색된 이진 잠재 벡터 표현을 기반으로 상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더를 통해 상기 입력된 분자 특성에 대응되는 상기 목표 분자를 설계하는 것으로 이루어지는,
딥러닝 기반 분자 설계 방법.

[청구항 4] 제3항에서,
상기 변분 오토인코더(VAE)의 상기 검벨 인코더 q(z|x)는,
게이트 순환 유닛(gated recurrent unit, GRU) 및 검벨-소프트맥스를 사용하여 입력 SMILES 스트링(string) x를 이진 잠재 벡터 z에 매핑하는,
딥러닝 기반 분자 설계 방법.

[청구항 5] 제4항에서,
상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더 p(x|z)는,
상기 스토캐스틱 디코더에서 게이트 순환 유닛(GRU) 모델의 초기 히든 상태(hidden state)로 인코딩된 이진 잠재 벡터 z를 사용하여 입력 SMILES 스트링 x를 복원하는,
딥러닝 기반 분자 설계 방법.

[청구항 6] 제5항에서,
상기 변분 오토인코더(VAE) 사전 학습 단계는,
상기 학습 데이터 세트를 기반으로, 재구성 에러(reconstruction error) 항 및 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD) 항을 포함하는 ELBO(evidence lower bound) 손실 함수(loss function)를 최소화하는 것을 통해, 상기 변분 오토인코더(VAE)를 학습하는 것으로 이루어지는,
딥러닝 기반 분자 설계 방법.

[청구항 7] 제6항에서,
상기 ELBO 손실 함수 L(x)는,
[식 1]을 나타내고,
상기 [식 1]은,

$$L(x) = -\log(p_\theta(x)) + KL[q_\phi(z|x)||p(z)]$$

이며,
상기 θ는, 상기 스토캐스틱 디코더의 학습 파라미터(trainable parameter)를 나타내고, 상기 Φ는, 상기 검벨 인코더의 학습 파라미터를 나타내며,
상기 $-\log(p_\theta(x))$ 는, 상기 재구성 에러 항을 나타내고, 입력 SMILES 스트링 x가 얼마나 완벽하게 복원되는지를 평가하는 것을 나타내는 $p_\theta(x)$ 의 음의 로그-우도(log-likelihood)를 나타내며,
상기 $KL[q_\phi(z|x)||p(z)]$ 은, 상기 쿨백-라이블러 발산(KLD) 항을 나타내고, 카테고리형 분포(categorical distribution) p(z)와 상기 검벨 인코더 $q_\phi(z|x)$ 의 출력 간의 차이를 측정하는,
딥러닝 기반 분자 설계 방법.

[청구항 8] 제7항에서,
상기 쿨백-라이블러 발산(KLD) 항은,
이산 균일 분포(discrete uniform distribution)의 확률 질량 함수(probability mass function)를 상기 카테고리형 분포 p(z)에 대입하여, [식 2]를 통해 계산되며,
상기 [식 2]는,

$$KL[q_\phi(z|x)||p(z)] = q_\phi(z|x)\log(q_\phi(z|x)*K)$$

이고, 상기 K는, 카테고리의 개수를 나타내며, 잠재 공간(latent space)의 차원(dimension)과 동일한, 딥러닝 기반 분자 설계 방법.

[청구항 9] 제1항 내지 제8항 중 어느 한 항에 기재된 딥러닝 기반 분자 설계 방법을

컴퓨터에서 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램.

[청구항 10] 딥러닝을 이용하여 분자를 설계하는 장치로서,
딥러닝을 이용하여 분자를 설계하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및
상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 딥러닝을 이용하여 분자를 설계하기 위한 동작을 수행하는 하나 이상의 프로세서;
를 포함하며,
상기 프로세서는,
사용자에 의해 분자 특성을 입력받고,
사전 학습된 변분 오토인코더(variational autoencoder, VAE) 및 유전 알고리즘(genetic algorithm, GA)을 기반으로, 상기 분자 특성에 대응되는 목표 분자를 설계하며,
상기 변분 오토인코더(VAE)는,
입력되는 분자 구조의 SMILES(simplified molecular-input line-entry system) 표현을 검벨-소프트맥스(Gumbel-Softmax)를 이용하여 분자 구조의 이진 잠재 벡터(binary latent vector) 표현으로 변환하는 검벨 인코더(Gumbel encoder) 및 분자 구조의 이진 잠재 벡터 표현을 분자 구조의 SMILES 표현으로 변환하는 스토캐스틱 디코더(stochastic decoder)를 포함하는,
딥러닝 기반 분자 설계 장치.

[청구항 11] 제10항에서,
상기 프로세서는,
SMILES로 표현된 학습 데이터 세트를 기반으로, 상기 검벨-소프트맥스를 이용한 상기 변분 오토인코더(VAE)를 사전 학습하는,
딥러닝 기반 분자 설계 장치.

[청구항 12] 제11항에서,
상기 프로세서는,
상기 변분 오토인코더(VAE) 및 상기 유전 알고리즘(GA)을 이용하여,
상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더의 도메인에서 상기 분자 특성을 가지는 분자에 대응되는 이진 잠재 벡터 표현을 탐색하고, 탐색된 이진 잠재 벡터 표현을 기반으로 상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더를 통해 상기 입력된 분자 특성에 대응되는 상기 목표 분자를 설계하는,
딥러닝 기반 분자 설계 장치.

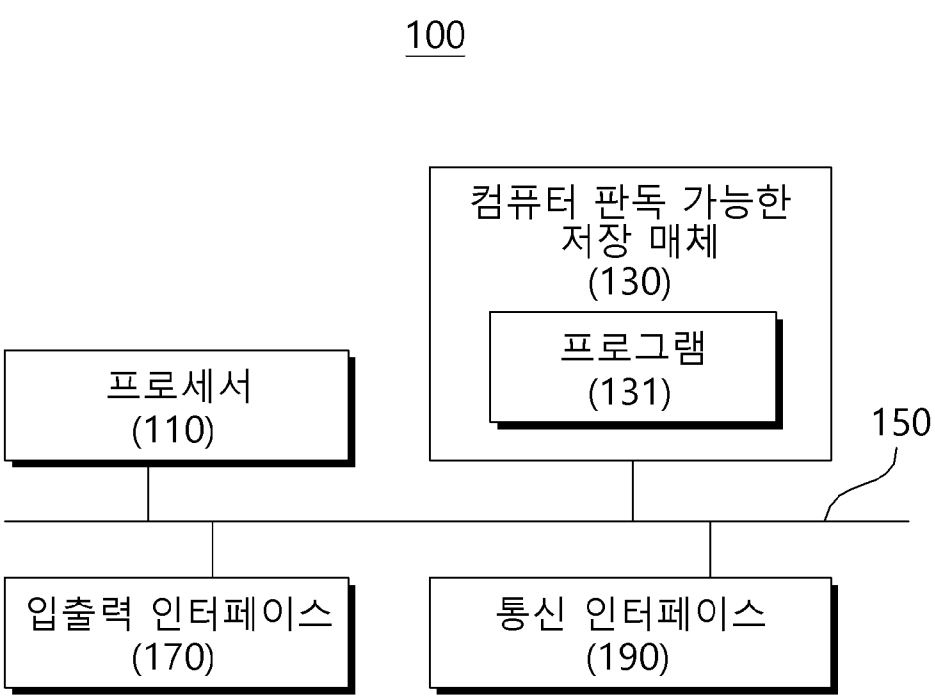
[청구항 13] 제12항에서,
상기 변분 오토인코더(VAE)의 상기 검벨 인코더 q(z|x)는,
게이트 순환 유닛(gated recurrent unit, GRU) 및 검벨-소프트맥스를 사용하여 입력 SMILES 스트링(string) x를 이진 잠재 벡터 z에 매핑하며,

상기 변분 오토인코더(VAE)의 상기 스토캐스틱 디코더 p(x|z)는,
상기 스토캐스틱 디코더에서 게이트 순환 유닛(GRU) 모델의 초기 히든 상태(hidden state)로 인코딩된 이진 잠재 벡터 z를 사용하여 입력 SMILES 스트링 x를 복원하고,
상기 프로세서는,
상기 학습 데이터 세트를 기반으로, 재구성 에러(reconstruction error) 항 및 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD) 항을 포함하는 ELBO(evidence lower bound) 손실 함수(loss function)를 최소화하는 것을 통해, 상기 변분 오토인코더(VAE)를 학습하는,
딥러닝 기반 분자 설계 장치.

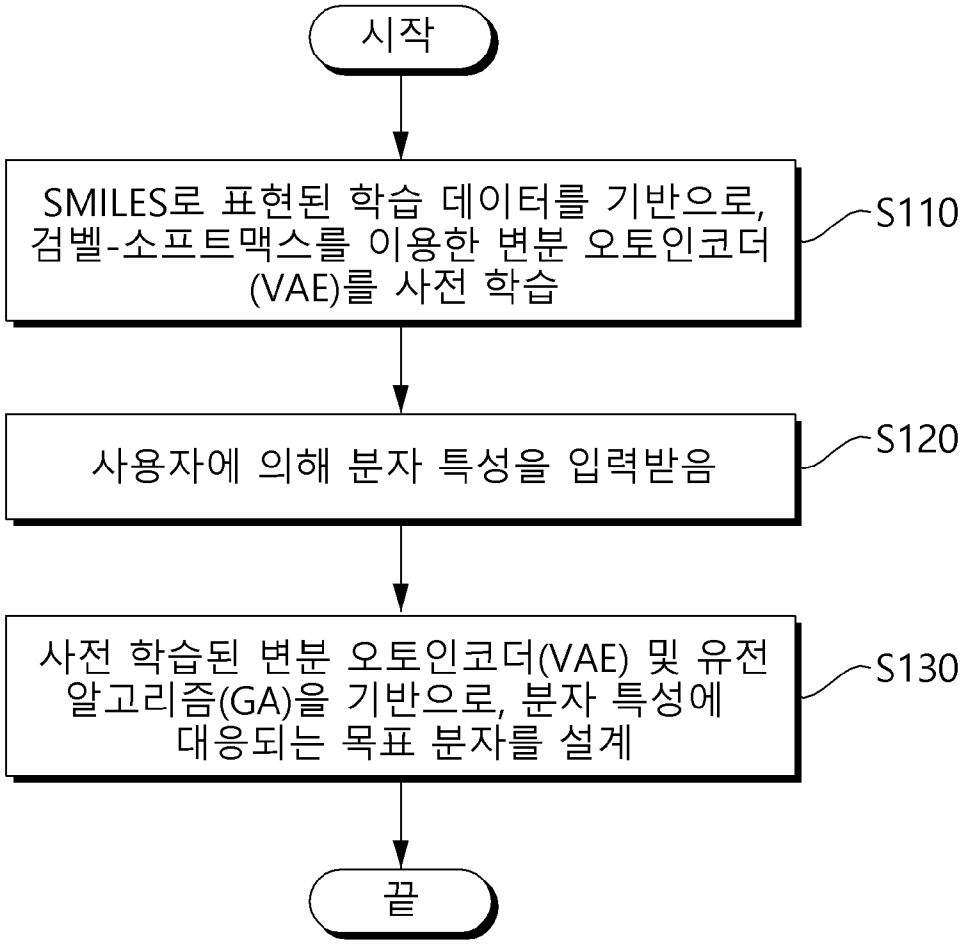
요약서

본 발명의 바람직한 실시예에 따른 딥러닝 기반 분자 설계 방법, 이를 수행하는 장치 및 컴퓨터 프로그램은, 검벨-소프트맥스(Gumbel-Softmax) 함수를 활용하여 변분 오토인코더(variational autoencoder, VAE)를 사전 학습하고, 사전 학습된 변분 오토인코더(VAE)와 유전 알고리즘(genetic algorithm, GA)을 이용하여 사용자의 요청 분자 특성에 대응되는 분자를 설계함으로써, 분자 구조의 SMILES 표현을 이진 벡터 표현으로 변환할 수 있어 잠재 공간의 탐색 범위를 유한한 공간으로 제약할 수 있다.

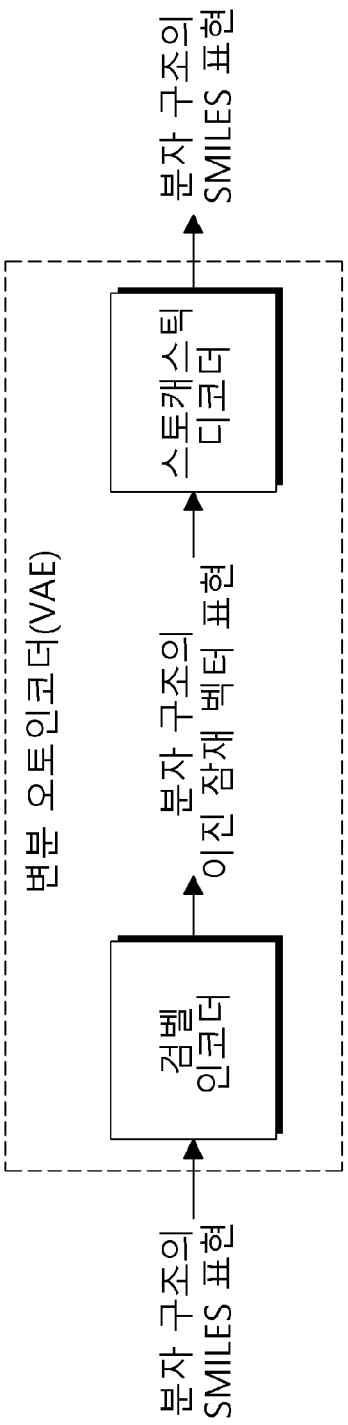
[도1]


100
컴퓨터 판독 가능한
저장 매체
(130)
프로그램
(131)
프로세서
(110)
150
입출력 인터페이스
(170)
통신 인터페이스
(190)

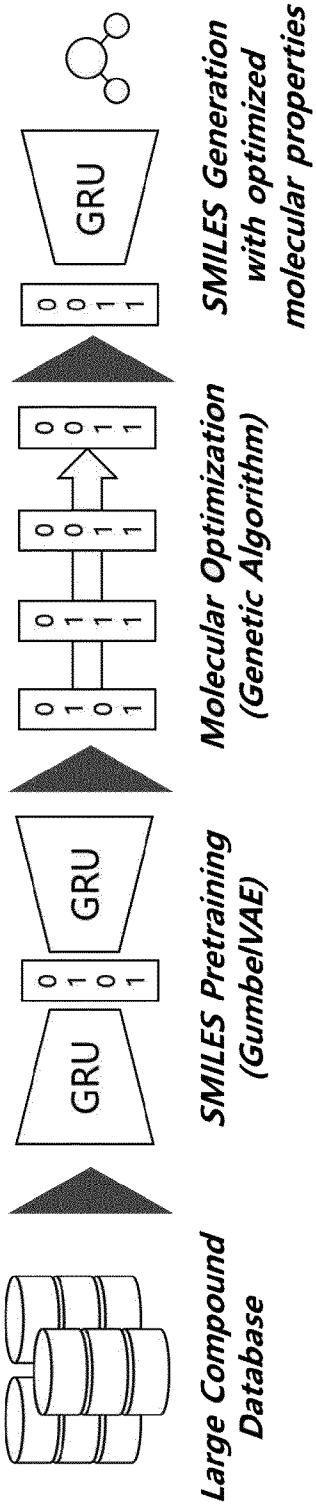
[도2]


시작
SMILES로 표현된 학습 데이터를 기반으로, 검벨-소프트맥스를 이용한 변분 오토인코더 (VAE)를 사전 학습
S110
사용자에 의해 분자 특성을 입력받음
S120
사전 학습된 변분 오토인코더(VAE) 및 유전 알고리즘(GA)을 기반으로, 분자 특성에 대응되는 목표 분자를 설계
S130
끝

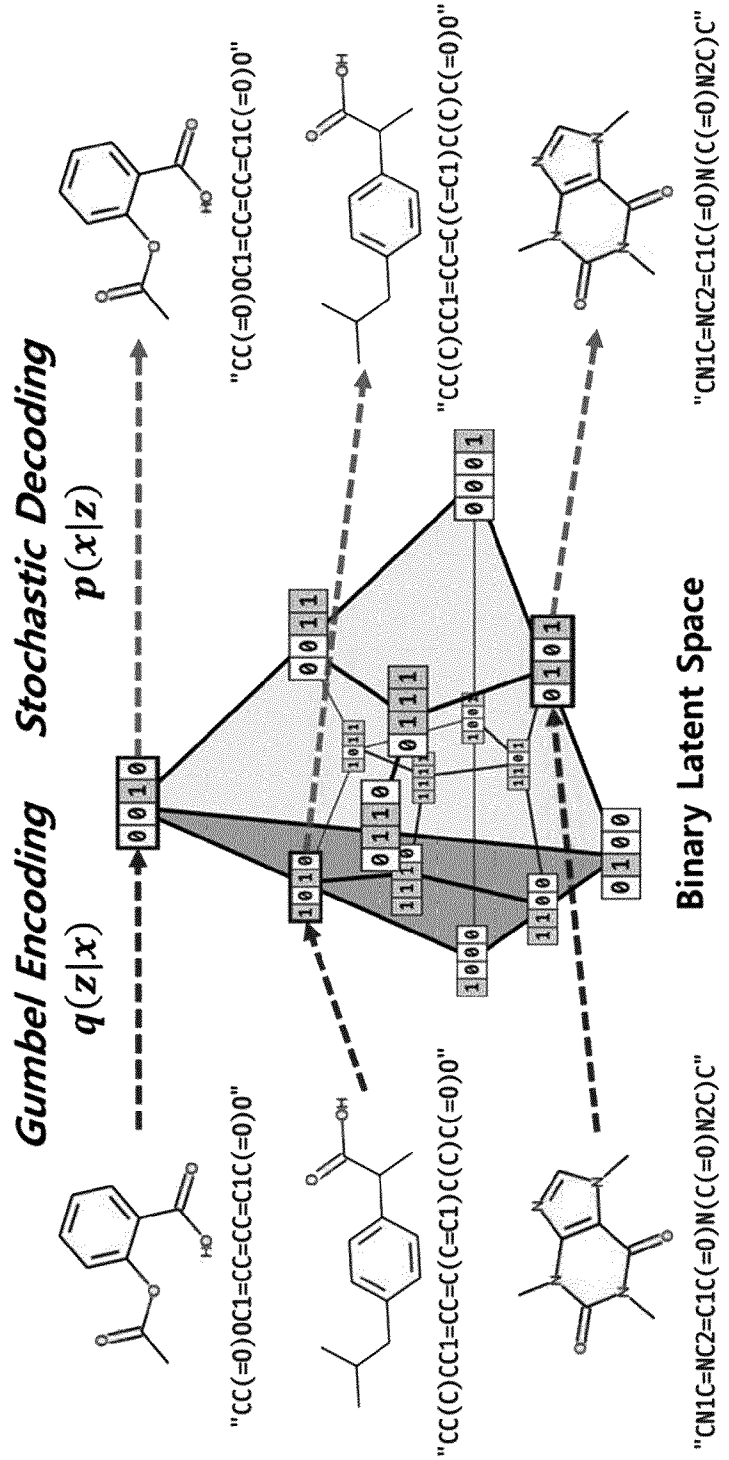
[도3]


분자 구조의
SMILES 표현
변분 오토인코더(VAE)
검벨
인코더
분자 구조의
이진 잠재 벡터 표현
스토캐스틱
디코더
분자 구조의
SMILES 표현

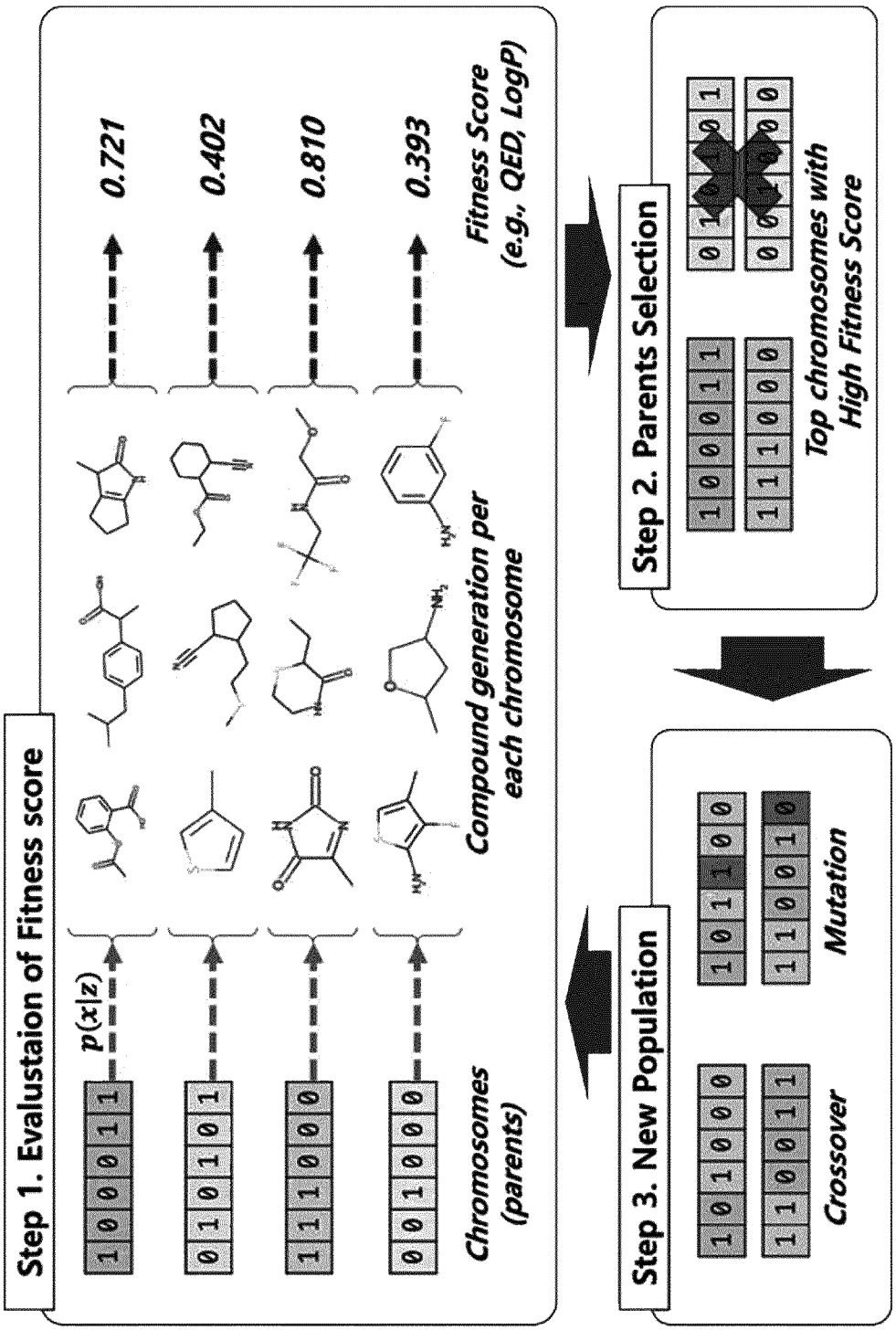
[도4]


Large Compound Database
GRU
SMILES Pretraining (GumbelVAE)
Molecular Optimization (Genetic Algorithm)
GRU
SMILES Generation with optimized molecular properties

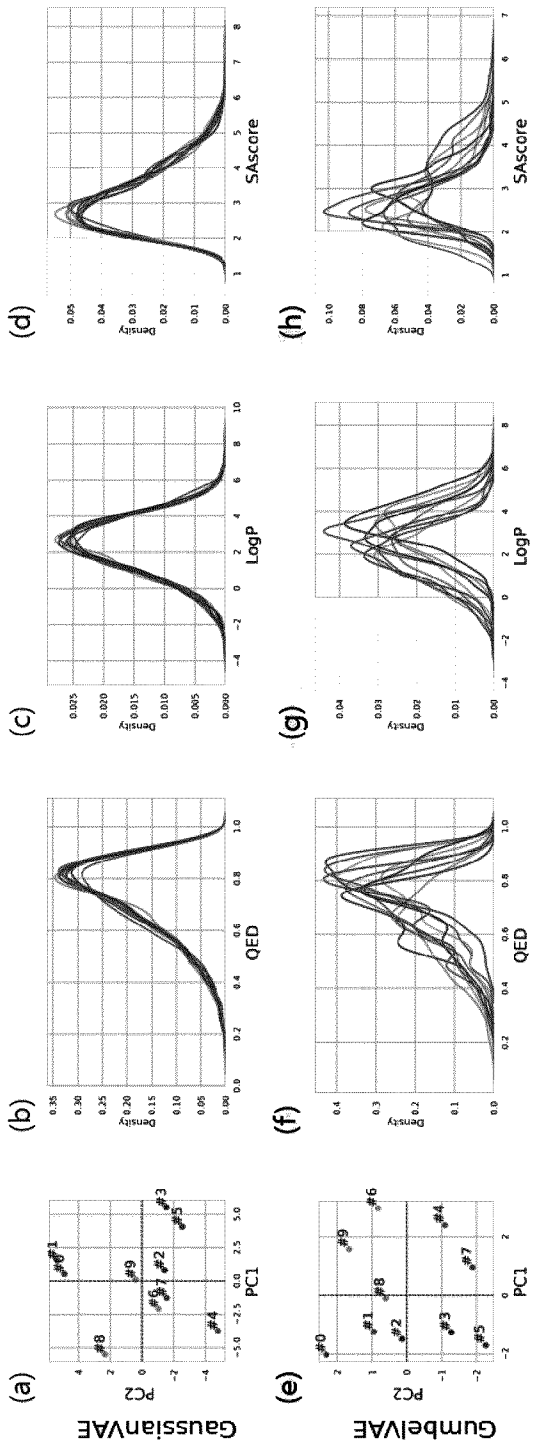
[도5]


Gumbel Encoding
$q(z|x)$
Stochastic Decoding
$p(x|z)$
Binary Latent Space
"CC(=O)OC1=CC=CC=C1C(=O)O"
"CC(=O)OC1=CC=CC=C1C(=O)O"
"CC(C)CC1=CC=C((C=C1)C(C)C(=O)O"
"CC(C)CC1=CC=C((C=C1)C(C)C(=O)O"
"CN1C=NC2=C1C(=O)N(C(=O)N2C)C"
"CN1C=NC2=C1C(=O)N(C(=O)N2C)C"

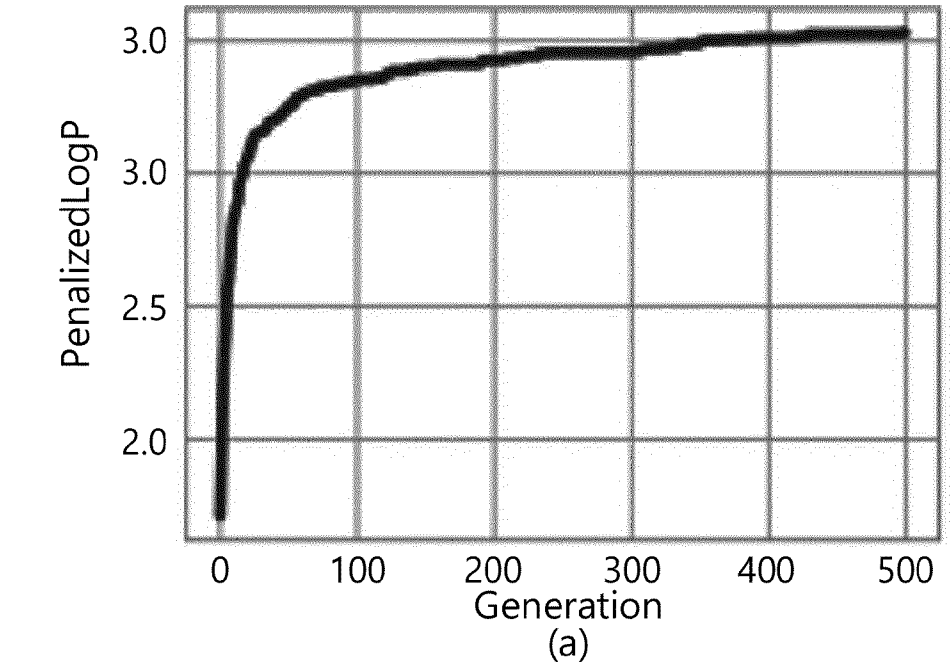
[도6]


Step 1. Evalustaion of Fitness score
p(x|z)
0.721
0.402
0.810
0.393
Fitness Score
(e.g., QED, LogP)
Compound generation per
each chromosome
Chromosomes
(parents)
Step 2. Parents Selection
Top chromosomes with
High Fitness Score
Step 3. New Population
Crossover
Mutation

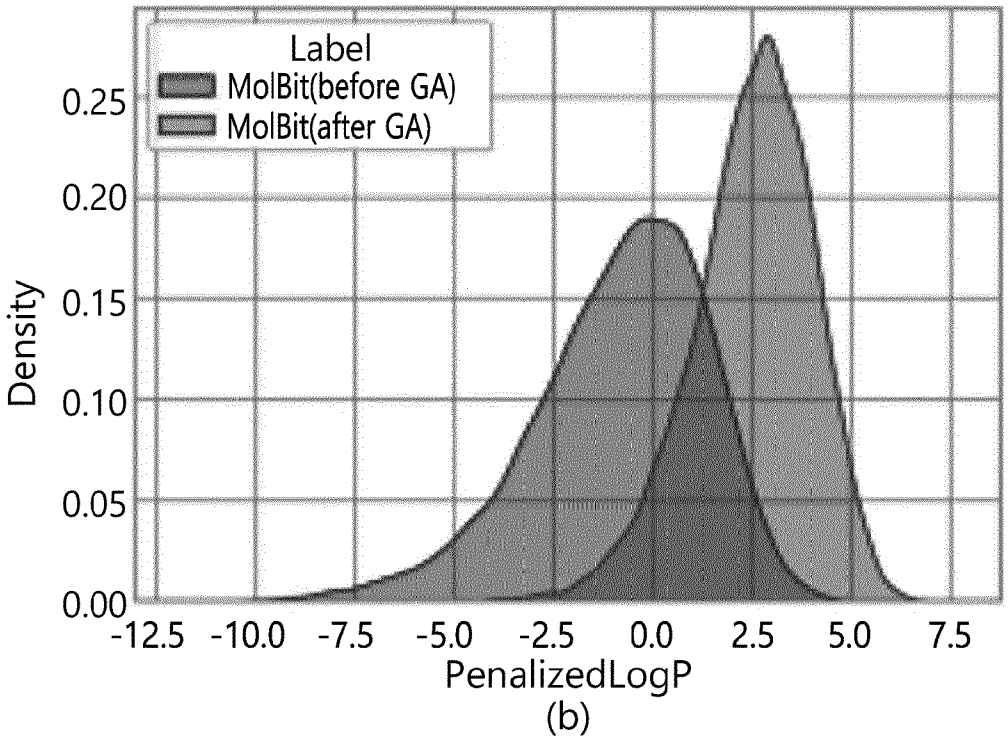
[도7]


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
GaussianVAE
GumbelVAE
PC1
PC2
QED
LogP
SAscore
Density

[도8]


PenalizedLogP
3.0
3.0
2.5
2.0
0
100
200
300
400
500
Generation

(a)


Label
MolBit(before GA)
MolBit(after GA)
Density
0.25
0.20
0.15
0.10
0.05
0.00
-12.5
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
PenalizedLogP

(b)