【서지사항】

【서류명】 특허출원서

【출원구분】 특허출원

【출원인】

【명칭】 연세대학교 산학협력단

【특허고객번호】 2-2005-009509-9

【대리인】

【성명】 권성현

【대리인번호】 9-2012-000114-4

【포괄위임등록번호】 2020-085395-8

【대리인】

【성명】 강일신

【대리인번호】 9-2013-000145-7

【포괄위임등록번호】 2020-085394-1

【대리인】

【성명】 김정연

【대리인번호】 9-2010-001352-0

【포괄위임등록번호】 2020-085398-0

【대리인】

【성명】 백두진

【대리인번호】 9-2010-000842-1

【포괄위임등록번호】 2020-085396-5

【대리인】

【성명】 유광철

【대리인번호】 9-2013-000581-3

【포괄위임등록번호】 2020-085397-2

【발명의 국문명칭】 문자열 교정을 이용한 유사 구조의 분자 생성 모델 학습 방법

【발명의 영문명칭】 LEARNING METHOD FOR MODELS TO GENERATE MOLECULES WITH SIMILAR STRUCTURES

【발명자】

【성명】 박상현

【성명의 영문표기】 SANGHYUN PARK

【주민등록번호】 670101-1XXXXXX

【우편번호】 08004

【주소】 서울특별시 양천구 오목로 300, 204동 3701호

【발명자】

【성명】 최종환

【성명의 영문표기】 JONGHWAN CHOI

【주민등록번호】 910226-1XXXXXX

【우편번호】 21090

【주소】 인천광역시 계양구 봉오대로691번길 4, 103동 409호

【발명자】

【성명】 박성민

【성명의 영문표기】 SHENGMIN PIAO

【주소】 서울특별시 서대문구 연희로35길 31, 201호

【주소의 영문표기】 201-ho, 31, Yeonhui-ro 35-gil, Seodaemun-gu, Seoul

【발명자】

【성명】 서상민

【성명의 영문표기】 SANGMIN SEO

【주민등록번호】 930507-1XXXXXX

【우편번호】 03724

【주소】 서울특별시 서대문구 연희로14길 29

【출원언어】 국어

【심사청구】 청구

【공지예외적용대상증명서류의 내용】

【공개형태】 논문 게재

【공개일자】 2022.11.14

【이 발명을 지원한 국가연구개발사업】

【과제고유번호】 1711130121

【과제번호】 2019R1A2C3005212

【부처명】 과학기술정보통신부

【과제관리(전문)기관명】 한국연구재단

【연구사업명】 개인기초연구(과기정통부)(R&D)

【연구과제명】 딥러닝을 이용한 간암 표적항암제 내성기전 규명 및 이를 극복할 새로운 표적항암제 탐색

【기여율】 1/1

【과제수행기관명】 인천대학교

【연구기간】 2021.03.01 ~ 2022.02.28

【취지】 위와 같이 특허청장에게 제출합니다.

대리인 권성현 (서명 또는 인)

대리인 강일신 (서명 또는 인)

대리인 김정연 (서명 또는 인)

대리인 백두진 (서명 또는 인)

대리인 유광철 (서명 또는 인)

【수수료】

【출원료】 0 면 46,000 원

【가산출원료】 25 면 0 원

【우선권주장료】 0 건 0 원

【심사청구료】 9 항 539,000 원

【합계】 585,000원

【감면사유】 전담조직(50%감면)[1]

【감면후 수수료】 292,500 원

【첨부서류】 1.공지예외적용대상(신규성상실의예외, 출원시의특례)규정을 적용받기 위한 증명서류_1통

1 : 공지예외적용대상(신규성상실의예외,_출원시의특례)규정을_적용받기_위한_증명서류

PDF 파일 첨부

【발명의 설명】

【발명의 명칭】

문자열 교정을 이용한 유사 구조의 분자 생성 모델 학습 방법{LEARNING METHOD FOR MODELS TO GENERATE MOLECULES WITH SIMILAR STRUCTURES}

【기술분야】

【0001】 본 개시는 시드(seed) 분자의 문자열을 교정하여 개선된 분자 특성을 가진 새로운 분자를 생성하기 위한 모델링 방법 및 장치에 관한 것이다.

【발명의 배경이 되는 기술】

【0003】 신약 개발은 인간과 질병 사이의 오랜 투쟁을 극복하기 위한 도전적인 과제이다. 신약후보물질을 설계하는 과정에 많은 시간과 비용이 소요된다는 것은 잘 알려진 사실이다. 조각 기반 스크리닝 또는 합성과 같은 전통적인 접근 방식은 대규모 화학 라이브러리를 탐색하기 위한 전문 지식과 경험을 요구할 뿐만 아니라 탐색 가능한 분자의 수가 방대하기 때문에 비효율적이라는 문제가 있다.

【발명의 내용】

【해결하고자 하는 과제】

【0005】 본 개시에서는 상술한 문제를 해결하도록 시드(seed) 분자의 문자열을 교정하여 개선된 분자 특성을 가진 새로운 분자를 생성하기 위한 모델링 방법

및 장치가 개시된다.

【과제의 해결 수단】

【0007】 본 개시의 일 실시예에 따르면, 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법은, 학습 데이터셋에 포함된 소스 분자 및 학습 데이터셋에서 소스 분자와 사전 페어링(paring)된 목표 분자의 문자열 각각에 대해 토큰화를 수행하여 소스 문자열 조각 세트 및 목표 문자열 조각 세트를 획득하는 단계, 소스 문자열 조각 세트 및 목표 문자열 조각 세트를 제1 학습 모델에 입력하여 소스 문자열 조각 세트가 임베딩된 소스 임베딩 조각 세트 및 목표 문자열 조각 세트가 임베딩된 목표 임베딩 조각 세트를 획득하는 단계 - 제1 학습 모델은 입력된 임의의 문자열 조각 세트에 포함된 하나의 문자열 조각을 반복적으로 인코딩하여 획득한 두 개의 임베딩 조각 사이의 거리는 짧아지고, 하나의 문자열 조각 및 임의의 문자열 조각 세트에 포함된 다른 하나의 문자열 조각 각각을 인코딩하여 획득한 두 개의 임베딩 조각 사이의 거리는 멀어지게 하는 목표 함수를 기초로 임베딩을 수행함 - 및 소스 임베딩 조각 세트 및 목표 임베딩 조각 세트를 제2 학습 모델에 입력하여, 소스 임베딩 조각 세트에서 하나 이상의 임베딩 조각이 수정된 최종 임베딩 조각 세트를 출력하도록 제2 학습 모델을 학습시키는 단계 - 제2 학습 모델은 입력된 임의의 임베딩 조각 세트와 목표 임베딩 조각 세트를, 임베딩 조각 단위로 비교하여 산출된 유사도가 증가하게 하는 손실 함수를 기초로, 수정을 수행함-를 포함할

수 있다.

【0008】일 실시예에 따르면, 소스 분자 및 목표 분자는 공통되는 약학적 속성을 갖도록 구성될 수 있다.

【0009】일 실시예에 따르면, 약학적 속성은 도파민 수용체 D2 (Dopamine Receptor D2; DRD2)의 활성을 억제하는 확률을 나타내고, 소스 분자의 확률은 제1 값 미만이고, 목표 분자의 확률은 제1 값보다 큰 제2 값을 초과하도록 구성될 수 있다.

【0010】일 실시예에 따르면, 페어링은 소스 분자 및 목표 분자 각각과 미리 결정된 복수의 분자 구조 각각을 대조한 결과를 기초로 결정될 수 있다.

【0011】일 실시예에 따르면, 손실 함수는 삭제 손실 함수 및 삽입 손실 함수를 포함하고, 소스 임베딩 조각 세트 및 목표 임베딩 조각 세트를 제2 학습 모델에 입력하여, 소스 임베딩 조각 세트 중 하나 이상의 임베딩 조각이 수정된 최종 임베딩 조각 세트를 출력하도록 제2 학습 모델을 학습시키는 단계는, 소스 임베딩 조각 세트 중 삭제 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각을 삭제하여 제1 중간 임베딩 조각 세트를 획득하는 단계 및 제1 중간 임베딩 조각 세트에 삽입 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각을 삽입하여 최종 임베딩 조각 세트를 출력하는 단계를 포함하고, 삭제 손실 함수에서 임의의 임베딩 조각 세트는 소스 임베딩 조각 세트로 구성되고, 삽입 손실 함수에서 임의의 임베딩 조각 세트는 제1 중간 임베딩 조각 세트로 구성될 수 있다.

【0012】일 실시예에 따르면, 삽입 손실 함수는 제1 삽입 손실 함수 및 제2 삽입 손실 함수를 포함하고, 제1 중간 임베딩 조각 세트 중 삽입 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각을 삽입하여 최종 임베딩 조각 세트를 출력하는 단계는, 제1 중간 임베딩 조각 세트에서 제1 삽입 손실 함수를 감소하게 하는 플레이스홀더의 삽입 위치 및 플레이스홀더의 개수를 결정하고, 제1 중간 임베딩 조각 세트에 개수 만큼의 플레이스홀더를 삽입 위치에 삽입하여 제2 중간 임베딩 조각 세트를 획득하는 단계 및 제2 중간 임베딩 조각 세트에서 플레이스홀더를 제2 삽입 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각으로 대체하여 최종 임베딩 조각 세트를 출력하는 단계를 포함할 수 있다.

【0013】일 실시예에 따르면, 토큰화에 의해 임의의 문자열은 상기 임의의 문자열에 포함된 하나 이상의 가지(branch)에 기초하여 제1 방향을 따라 제1 문자열 조각 세트로 분리된 후, 상기 제1 문자열 조각 세트에 포함된 하나 이상의 고리(ring)에 기초하여 상기 제1 방향과 반대되는 제2 방향을 따라 제2 문자열 조각 세트로 결합될 수 있다.

【0014】일 실시예에 따르면, 토큰화에 따라, 소스 문자열 조각 세트 및 목표 문자열 조각 세트 중 적어도 하나의 문자열 조각 세트는 제1 방향을 따라 하나 이상의 가지 조각이 분리된 후, 제1 방향과 반대되는 제2 방향을 따라 하나 이상의 가지 조각이 재결합하여 하나 이상의 고리 조각을 형성할 수 있다.

【0015】본 개시의 다른 실시예에 따르면, 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법을 실행시키도록 컴퓨터로 판독 가능한 기록매체에 기록된

컴퓨터 프로그램이 제공될 수 있다.

【발명의 효과】

【0017】 본 개시의 일부 실시예에 따르면 모델의 컴퓨팅 효율 및 분자 특성이 개선된 유사 구조의 분자 생성 모델을 제공할 수 있다.

【도면의 간단한 설명】

【0019】 도 1은 본 개시의 일 실시예에 따른 유사 구조의 분자 생성 시스템의 모식도이다.

도 2는 본 개시의 일 실시예에 따른 유사 구조의 분자 생성 모델 학습 방법의 흐름도이다.

도 3은 본 개시의 일 실시예에 따라 학습 데이터셋에 포함된 분자로부터 문자열 조각 세트를 획득하는 단계의 예시를 나타낸다.

도 4는 본 개시의 일 실시예에 따라 문자열 조각 세트를 이용하여 임베딩 조각 세트를 출력하는 제1 학습 모델을 학습시키는 단계의 일부를 나타내는 모식도이다.

도 5는 본 개시의 일 실시예에 따라 임베딩 조각 세트를 이용하여 제2 학습 모델을 학습시키는 단계의 일부를 나타내는 모식도이다.

【발명을 실시하기 위한 구체적인 내용】

【0020】이하, 본 개시의 실시를 위한 구체적인 내용을 첨부된 도면을 참조하여 상세히 설명한다. 다만, 이하의 설명에서는 본 개시의 요지를 불필요하게 흐릴 우려가 있는 경우, 널리 알려진 기능이나 구성에 관한 구체적 설명은 생략하기로 한다.

【0021】첨부된 도면에서, 동일하거나 대응하는 구성요소에는 동일한 참조부호가 부여되어 있다. 또한, 이하의 실시예들의 설명에 있어서, 동일하거나 대응되는 구성요소를 중복하여 기술하는 것이 생략될 수 있다. 그러나 구성요소에 관한 기술이 생략되어도, 그러한 구성요소가 어떤 실시예에 포함되지 않는 것으로 의도되지는 않는다.

【0022】개시된 실시예의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 개시는 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 개시가 완전하도록 하고, 본 개시가 통상의 기술자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것일 뿐이다.

【0023】본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 개시된 실시예에 대해 구체적으로 설명하기로 한다. 본 명세서에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 관련 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있

으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

【0024】본 명세서에서의 단수의 표현은 문맥상 명백하게 단수인 것으로 특정하지 않는 한, 복수의 표현을 포함한다. 또한, 복수의 표현은 문맥상 명백하게 복수인 것으로 특정하지 않는 한, 단수의 표현을 포함한다. 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.

【0025】한편, 본 명세서에서 '모델' 또는 '시스템'에 입력되거나, '모델' 또는 '시스템'으로부터 출력되거나, '모델' 또는 '시스템'에서 처리되는 'A'는 'A에 대한 정보를 포함하는 데이터'를 지칭하는 것으로 한다. 예를 들어, '모델'에 '분자'가 입력되는 경우, '분자'는 분자에 대한 정보를 포함하는 데이터로 해석될 수 있다. 다른 예를 들어, '모델'에서 '문자열'이 분할되는 경우, '문자열'은 연속적 문자로 구성되는 텍스트 데이터를 지칭할 수 있다.

【0026】도 1은 본 개시의 일 실시예에 따른 유사 구조의 분자 생성 시스템(100) 모식도이다. 도시된 바와 같이, 유사 구조의 분자 생성 시스템(100)(이하, '시스템'이라 한다.) 유사 구조 분자 생성 모델(110), 입력 분자(120) 및 출력 분자(130)를 포함할 수 있다. 본 개시의 유사 구조 분자 생성 모델(110)은 입력 분자(120)로부터 입력 분자(120)와 구조적으로 동일/유사하나 속성이 개선된 출력 분

자(130)를 획득하는 것을 목적으로 학습될 수 있다. 여기서, '속성'은 분자의 화학적 속성을 지칭할 수 있다. 예를 들어, '속성'은 약물의 기능, 효과 등으로 해석될 수 있다. 한편, 이하에서는 상술한 목적을 달성하기 위하여 유사 구조 분자 생성 모델(110)을 학습시키는 방법이 개시된다. 이 경우, 방법에 포함된 모든 동작은 각각 컴퓨팅 장치의 적어도 하나의 프로세서에 의해 수행될 수 있다.

【0027】도 2는 본 개시의 일 실시예에 따른 유사 구조의 분자 생성 모델 학습 방법(200)의 흐름도이다. 도시된 바와 같이, 방법(200)은 학습 데이터셋에 포함된 소스 분자 및 학습 데이터셋에서 소스 분자와 사전 페어링(paring)된 목표 분자의 문자열 각각에 대해 토큰화를 수행하여 소스 문자열 조각 세트 및 목표 문자열 조각 세트를 획득하는 단계(S210), 소스 문자열 조각 세트 및 목표 문자열 조각 세트를 제1 학습 모델에 입력하여 소스 문자열 조각 세트가 임베딩된 소스 임베딩 조각 세트 및 목표 문자열 조각 세트가 임베딩된 목표 임베딩 조각 세트를 획득하는 단계(S220) 및 소스 임베딩 조각 세트 및 목표 임베딩 조각 세트를 제2 학습 모델에 입력하여, 소스 임베딩 조각 세트에서 하나 이상의 임베딩 조각이 수정된 최종 임베딩 조각 세트를 출력하도록 제2 학습 모델을 학습하는 단계(S230) 중 적어도 하나를 포함할 수 있다. 한편, 도 2에서는 제2 단계(S220) 이후에 제3 단계(S230)이 실시되는 것으로 도시되었으나 이에 한정되지 않는다. 예를 들어, 제1 단계(S210), 제2 단계(S220) 및 제3 단계(S230) 중 적어도 일부의 단계는 병렬적으로 수행될 수도 있다.

【0028】도 3은 본 개시의 일 실시예에 따라 학습 데이터셋에 포함된 분자(310)로부터 문자열 조각 세트를 획득하는 단계(예: 제1 단계(S210))의 예시를 나타낸다. 도시된 바와 같이, 프로세서는 분자(310)를 문자열(320)로 변환하고, 변환된 문자열(320)에 대해 토큰화를 수행하여 제1 문자열 조각 세트(330) 및/또는 제2 문자열 조각 세트(340)를 획득할 수 있다. 예를 들어, 프로세서는 문자열(320)을 문자열(320)에 포함된 하나 이상의 가지(branch)에 기초하여 제1 방향(좌에서 우)을 따라 제1 문자열 조각 세트(320)로 분리할 수 있다. 추가적으로, 프로세서는 제1 문자열 조각 세트(320)를 문자열(320)에 포함된 하나 이상의 고리(ring)에 기초하여 제1 방향과 반대되는 제2 방향(우에서 좌)을 따라 제2 문자열 조각 세트(320)로 결합할 수도 있다.

【0029】도 4는 본 개시의 일 실시예에 따라 문자열 조각 세트를 이용하여 임베딩 조각 세트를 출력하는 제1 학습 모델을 학습시키는 단계(예: 제2 단계(S220))의 일부를 나타내는 모식도이다. 프로세서는 문자열 조각 세트를 제1 학습 모델(400)에 입력하여 문자열 조각 세트가 임베딩된 임베딩 조각 세트를 획득할 수 있다. 이 경우, 임베딩 조각 세트는 문자열 조각 세트에 포함된 복수의 문자열 조각에 대응하는 복수의 임베딩 조각을 포함할 수 있다. 예를 들어, 문자열 조각 세트가 3개의 문자 조각 vi, vj 및 vk를 포함하는 경우, 문자열 조각 세트가 임베딩된 임베딩 조각 세트는 vi에 대응하는 임베딩 조각 hi, vj에 대응하는 임베딩 조각 hj 및 vk에 대응하는 임베딩 조각 hk를 포함할 수 있다.

【0030】 한편, 제1 학습 모델은 입력된 임의의 문자열 조각 세트에 포함된 하나의 문자열 조각을 반복적으로 인코딩하여 획득한 두 개의 임베딩 조각 사이의 거리는 짧아지게 하는 목표 함수를 기초로 임베딩을 수행할 수 있다. 즉, 하나의 문자열 조각을 두 번 인코딩하여 획득한 서로 다른 두 개의 임베딩 조각을 상기 제1 학습 모델의 비지도 대조 학습에 이용되는 긍정 사례(positive instance)로 그룹핑할 수 있다. 이 경우, 서로 다른 두 개의 임베딩 조각은 dropout mask를 포함하는 인코더를 이용하여 획득될 수 있다.

【0031】 추가적으로 또는 대안적으로, 제1 학습 모델은 상기 하나의 문자열 조각 및 상기 임의의 문자열 조각 세트에 포함된 다른 하나의 문자열 조각 각각을 인코딩하여 획득한 두 개의 임베딩 조각 사이의 거리는 멀어지게 하는 목표 함수를 기초로 임베딩을 수행할 수 있다. 즉, 서로 다른 두 개의 문자열 조각 각각을 인코딩하여 획득한 서로 다른 두 개의 임베딩 조각을 제1 학습 모델의 비지도 대조 학습에 이용되는 부정 사례(negative instance)로 그룹핑할 수 있다.

【0032】 본 개시의 앞선 설명은 통상의 기술자들이 본 개시를 행하거나 이용하는 것을 가능하게 하기 위해 제공된다. 본 개시의 다양한 수정예들이 통상의 기술자들에게 쉽게 자명할 것이고, 본원에 정의된 일반적인 원리들은 본 개시의 취지 또는 범위를 벗어나지 않으면서 다양한 변형예들에 적용될 수도 있다. 따라서, 본 개시는 본원에 설명된 예들에 제한되도록 의도된 것이 아니고, 본원에 개시된 원리들 및 신규한 특징들과 일관되는 최광의의 범위가 부여되도록 의도된다.

【0033】도 5는 본 개시의 일 실시예에 따라 임베딩 조각 세트를 이용하여 제2 학습 모델을 학습시키는 단계(예: 제3 단계(S230))의 일부를 나타내는 모식도이다. 여기서, 제2 학습 모델은 소스 임베딩 조각 세트(540)로부터 제2 학습 모델은 소스 임베딩 조각 세트(540)로부터 임베딩 조각의 삭제 및 삽입 과정을 통해 최종 임베딩 조각 세트(570)를 출력하도록 학습되는 모델을 지칭할 수 있다. 구체적으로, 프로세서는 소스 임베딩 조각 세트(540) 및 목표 임베딩 조각 세트(580)를 제2 학습 모델에 입력하여, 소스 임베딩 조각 세트(540) 중 하나 이상의 임베딩 조각이 수정된 최종 임베딩 조각 세트(570)를 출력하도록 제2 학습 모델을 학습시킬 수 있다. 예를 들어, 프로세서는 소스 임베딩 조각 세트(540) 중 삭제 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각을 삭제하여 제1 중간 임베딩 조각 세트(550)를 획득할 수 있다. 그리고 나서, 프로세서는 제1 중간 임베딩 조각 세트(550)에 삽입 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각을 삽입하여 제2 학습 모델로부터 최종 임베딩 조각 세트(570)를 출력할 수 있다. 여기서, 삭제 손실 함수, 및 삽입 손실 함수는 제2 학습 모델의 손실 함수에 포함되는 것으로, 삭제 손실 함수는 제1 중간 임베딩 조각 세트(550)와 목표 임베딩 조각 세트(580)를 임베딩 조각 단위로 비교하여 산출된 유사도가 증가하도록 제2 학습 모델을 학습시킬 수 있다. 추가적으로 또는 대안적으로, 삽입 손실 함수는 제2 중간 임베딩 조각 세트(560) 및 최종 임베딩 조각 세트(570)와 목표 임베딩 조각 세트(580)를 임베딩 조각 단위로 비교하여 산출된 유사도가 증가하도록 제2 학습 모델을 학습시킬 수 있다.

【0034】 한편, 프로세서가 제1 중간 임베딩 조각 세트(550) 중 삽입 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각을 삽입하여 최종 임베딩 조각 세트를 출력하는 과정은 제1 중간 임베딩 조각 세트(550)에서 제1 삽입 손실 함수를 감소하게 하는 플레이스홀더의 삽입 위치 및 플레이스홀더의 개수를 결정하고, 제1 중간 임베딩 조각 세트(550)에 개수 만큼의 플레이스홀더를 삽입 위치에 삽입하여 제2 중간 임베딩 조각 세트(560)를 획득한 후, 제2 중간 임베딩 조각 세트(560)에서 플레이스홀더를 제2 삽입 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각으로 대체하여 최종 임베딩 조각 세트를 출력하는 과정을 포함할 수 있다.

【0035】 이하에서는, 상술한 방법이 수행된 구체적인 실험례가 후술된다.

【0036】 하드웨어: Nvidia GeForce RTX 3090 24GB GPU, Intel i9-9900K 3.60GHz CPU 및 32GB RAM. 소프트웨어: Ubuntu 18.04.6 LTS, PyTorch 1.12.1, Python 3.7.12. SELFragment 임베딩 및 SELF-EdiT(본 개시의 분자 생성 모델에 대응됨)의 하이퍼파라미터 설정을 위해 원본 SimCSE(본 개시의 제1 학습 모델에 대응됨) 및 LevT(본 개시의 제2 학습 모델에 대응됨)에서 사용된 기본값을 사용했다. 이 연구에 사용된 모든 소스 코드와 데이터 세트는 https://github.com/sungmin630/SELF-EdiT 에서 사용할 수 있다.

【0037】 본 개시의 방법에 따라 학습된 모델을 다음과 비교한다.

【0038】 *MMPA: 데이터 세트에서 여러 규칙을 추출하는 규칙 기반 분자 변환 방법이다. 추론 절차 중에 종자 분자는 서로 다른 일치 변환 규칙을 사용하여 여러

번 번역되었다.

【0039】*Junction Tree VAE(JT-VAE): 분자 그래프를 주기가 없고 생성하기 쉬운 접합 트리로 나타내는 베이지안 최적화 기반 모델이다. 인코더는 분자 그래프와 접합 트리를 모두 잠재 변수로 매핑한다. 그 후 디코더는 먼저 특정 분자 그래프로 재구성되는 청사진으로 접합 트리를 생성한다.

【0040】*GCPN: 원자와 결합을 반복적으로 추가하거나 삭제하여 분자를 수정하는 강화 학습 기반 모델. 제안된 모델은 또한 최적화된 분자의 자연성을 향상시키기 위해 적대적 학습을 채택한다.

【0041】*VSeq2Seq: SMILES 기반 시퀀스를 VAE로 최적화하는 베이지안 최적화 기반 모델. 인코더와 디코더 모두 GRU를 신경 아키텍처로 사용하며 다른 분자 생성 작업에 성공적으로 적용되었다.

【0042】*UGMMT: 이중 학습을 활용하여 분자를 최적화하는 방법이다. 임베딩 공간 간의 양방향 변환을 구현하기 위해 각 번역 네트워크는 단방향 변환에 대해 개별적으로 훈련된다.

【0043】*VJTNN(+GAN): 분자 최적화를 그래프 간 변환 작업으로 처리하는 JT-VAE 기반의 개선된 방법이다. 제안하는 방법은 학습을 위한 접합 트리 인코더-디코더를 유지하면서 베이지안 최적화 대신 적대적 학습을 사용한다.

【0044】*STONED: 소스 분자와 대상 분자를 교체하여 SELFIES를 간단히 편집하는 규칙 기반 방법. STONED는 광전지와 유사한 분자 구조 설계를 위한 가상 스크

리닝에 적용하고 소스에서 대상 분자까지의 화학적 경로를 도출하여 해석 가능성을 입증함으로써 그 우수성을 입증했다.

【0045】 **【표 1】**

Method	Molecular Optimization Performance on QED dataset				
	Success	Property	Similarity	Novelty	Total Score**
MMPA*	32.9%	-	-	99.9%	-
GCPN*	9.4%	-	-	**100%**	-
VSeq2Seq*	58.5%	-	-	99.6%	-
JT-VAE	5.6%	0.789	**0.473**	**100%**	0.380
UGMMT	27.8%	0.633	0.257	**100%**	0.461
VJTNN	58.3%	0.902	0.307	94.7%	0.625
VJTNN+GAN	59.5%	0.902	0.307	94.7%	0.628
STONED	55.3%	**0.936**	0.300	**100%**	0.627
Ours	**59.6%**	0.833	0.354	99.8%	**0.647**

* Results of MMPA, GCPN and VSeq2Seq are from Barshatski, G. and Radinsky, K. 2021, August. Unpaired Generative Molecule-to-Molecule Translation for Lead Optimization. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Minin (pp. 2554-2564).

** Total score is the geometric mean of success, property, similarity, and novelty.

【0046】 【표 2】

Method	Molecular Optimization Performance on DRD2 dataset				
	Success	Property	Similarity	Novelty	Total Score**
MMPA*	46.4%	-	-	99.9%	-
GCPN*	4.4%	-	-	**100%**	-
VSeq2Seq*	75.9%	-	-	79.7%	-
JT-VAE	3.3%	0.091	**0.465**	**100%**	0.193
UGMMT	21.3%	0.506	0.182	88.6%	0.363
VJTNN	74.8%	0.803	0.343	62.3%	0.599
VJTNN+GAN	73.3%	0.787	0.324	62.4%	0.584
STONED	51.3%	**0.958**	0.296	**100%**	0.618
Ours	**82.2%**	0.544	0.296	99.9%	**0.642**

* Results of MMPA, GCPN and VSeq2Seq are from Barshatski, G. and Radinsky, K. 2021, August. Unpaired Generative Molecule-to-Molecule Translation for Lead Optimization. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Minin (pp. 2554-2564).

** Total score is the geometric mean of success, property, similarity, and novelty.

【0047】본 명세서에서는 본 개시가 일부 실시예들과 관련하여 설명되었지만, 본 발명이 속하는 기술분야의 통상의 기술자가 이해할 수 있는 본 개시의 범위를 벗어나지 않는 범위에서 다양한 변형 및 변경이 이루어질 수 있다는 점을 알아야 할 것이다. 또한, 그러한 변형 및 변경은 본 명세서에서 첨부된 특허 청구의 범위 내에 속하는 것으로 생각되어야 한다.

【부호의 설명】

【0049】100: 유사 구조 분자 생성 시스템

110: 유사 구조 분자 생성 모델

120: 입력 분자

130: 출력 분자

【청구범위】

【청구항 1】

학습 데이터셋에 포함된 소스 분자 및 상기 학습 데이터셋에서 상기 소스 분자와 사전 페어링(paring)된 목표 분자의 문자열 각각에 대해 토큰화를 수행하여 소스 문자열 조각 세트 및 목표 문자열 조각 세트를 획득하는 단계;

상기 소스 문자열 조각 세트 및 상기 목표 문자열 조각 세트를 제1 학습 모델에 입력하여 상기 소스 문자열 조각 세트가 임베딩된 소스 임베딩 조각 세트 및 상기 목표 문자열 조각 세트가 임베딩된 목표 임베딩 조각 세트를 획득하는 단계 - 상기 제1 학습 모델은 입력된 임의의 문자열 조각 세트에 포함된 하나의 문자열 조각을 반복적으로 인코딩하여 획득한 두 개의 임베딩 조각 사이의 거리는 짧아지고, 상기 하나의 문자열 조각 및 상기 임의의 문자열 조각 세트에 포함된 다른 하나의 문자열 조각 각각을 인코딩하여 획득한 두 개의 임베딩 조각 사이의 거리는 멀어지게 하는 목표 함수를 기초로 임베딩을 수행함-; 및

상기 소스 임베딩 조각 세트 및 상기 목표 임베딩 조각 세트를 제2 학습 모델에 입력하여, 상기 소스 임베딩 조각 세트에서 하나 이상의 임베딩 조각이 수정된 최종 임베딩 조각 세트를 출력하도록 상기 제2 학습 모델을 학습시키는 단계 - 상기 제2 학습 모델은 입력된 임의의 임베딩 조각 세트와 상기 목표 임베딩 조각 세트를, 임베딩 조각 단위로 비교하여 산출된 유사도가 증가하게 하는 손실 함수를 기초로, 수정을 수행함-

를 포함하는, 적어도 하나의 프로세서에 의해 수행되는 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법.

【청구항 2】

제1항에 있어서,

상기 소스 분자 및 상기 목표 분자는 공통되는 약학적 속성을 갖도록 구성되는, 적어도 하나의 프로세서에 의해 수행되는 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법.

【청구항 3】

제2항에 있어서,

상기 약학적 속성은 DRD2의 활성을 억제하는 확률을 나타내고,

상기 소스 분자의 상기 확률은 제1 값 미만이고, 상기 목표 분자의 상기 확률은 상기 제1 값보다 큰 제2 값을 초과하도록 구성되는, 적어도 하나의 프로세서에 의해 수행되는 문자열 교정 기반의 유사 구조 분자 생성 모델 방법.

【청구항 4】

제1항에 있어서,

상기 페어링은 상기 소스 분자 및 상기 목표 분자 각각과 미리 결정된 복수

의 분자 구조 각각을 대조한 결과를 기초로 결정되는, 적어도 하나의 프로세서에 의해 수행되는 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법.

【청구항 5】

제1항에 있어서,

상기 손실 함수는 삭제 손실 함수 및 삽입 손실 함수를 포함하고,

상기 소스 임베딩 조각 세트 및 상기 목표 임베딩 조각 세트를 제2 학습 모델에 입력하여, 상기 소스 임베딩 조각 세트 중 하나 이상의 임베딩 조각이 수정된 최종 임베딩 조각 세트를 출력하도록 상기 제2 학습 모델을 학습시키는 단계는,

상기 소스 임베딩 조각 세트 중 상기 삭제 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각을 삭제하여 제1 중간 임베딩 조각 세트를 획득하는 단계; 및

상기 제1 중간 임베딩 조각 세트에 상기 삽입 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각을 삽입하여 상기 최종 임베딩 조각 세트를 출력하는 단계

를 포함하고,

상기 삭제 손실 함수에서 상기 임의의 임베딩 조각 세트는 상기 소스 임베딩 조각 세트로 구성되고, 상기 삽입 손실 함수에서 상기 임의의 임베딩 조각 세트는 상기 제1 중간 임베딩 조각 세트로 구성되는, 적어도 하나의 프로세서에 의해 수행되는 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법.

【청구항 6】

제5항에 있어서,

상기 삽입 손실 함수는 제1 삽입 손실 함수 및 제2 삽입 손실 함수를 포함하고,

상기 제1 중간 임베딩 조각 세트 중 상기 삽입 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각을 삽입하여 상기 최종 임베딩 조각 세트를 출력하는 단계는,

상기 제1 중간 임베딩 조각 세트에서 상기 제1 삽입 손실 함수를 감소하게 하는 플레이스홀더의 삽입 위치 및 상기 플레이스홀더의 개수를 결정하고, 상기 제1 중간 임베딩 조각 세트에 상기 개수 만큼의 플레이스홀더를 상기 삽입 위치에 삽입하여 제2 중간 임베딩 조각 세트를 획득하는 단계; 및

상기 제2 중간 임베딩 조각 세트에서 상기 플레이스홀더를 상기 제2 삽입 손실 함수를 감소하게 하는 하나 이상의 임베딩 조각으로 대체하여 상기 최종 임베딩 조각 세트를 출력하는 단계

를 포함하는, 적어도 하나의 프로세서에 의해 수행되는 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법.

【청구항 7】

제1항에 있어서,

상기 토큰화에 의해 상기 소스 문자열 조각 세트 및 상기 목표 문자열 조각 세트 중 적어도 하나의 문자열 조각 세트는 가지(branch) 조각 및 고리(ring) 조각 중 적어도 하나를 포함하는, 적어도 하나의 프로세서에 의해 수행되는 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법.

【청구항 8】

제1항에 있어서,

상기 토큰화에 의해 임의의 문자열은 상기 임의의 문자열에 포함된 하나 이상의 가지(branch)에 기초하여 제1 방향을 따라 제1 문자열 조각 세트로 분리된 후, 상기 제1 문자열 조각 세트에 포함된 하나 이상의 고리(ring)에 기초하여 상기 제1 방향과 반대되는 제2 방향을 따라 제2 문자열 조각 세트로 결합되는, 적어도 하나의 프로세서에 의해 수행되는 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법.

【청구항 9】

제8항에 따른 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법을 실행시키도록 컴퓨터로 판독 가능한 기록매체에 기록된 컴퓨터 프로그램.

【요약서】

【요약】

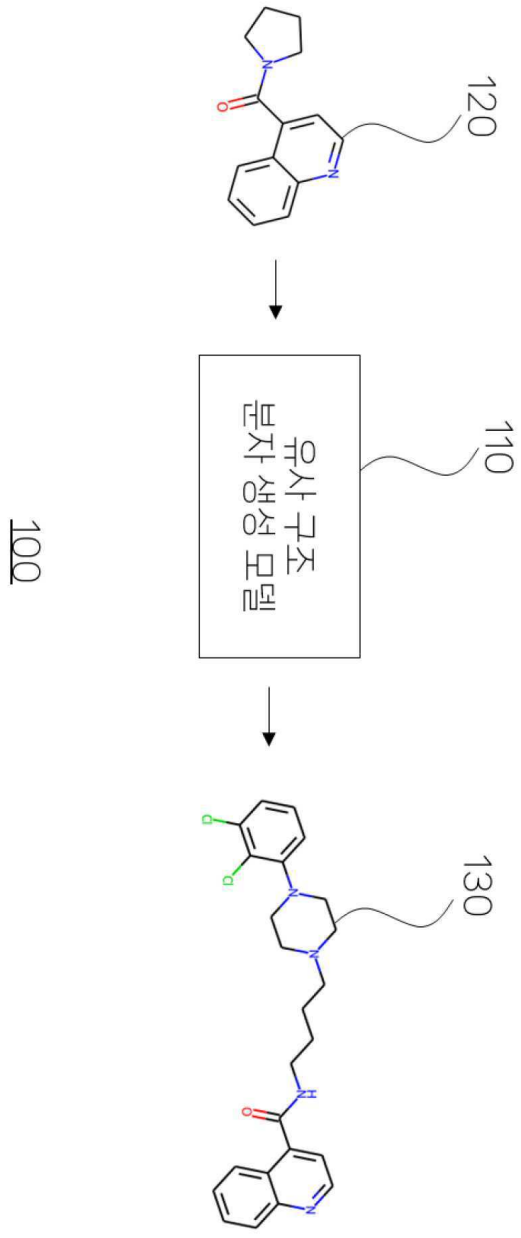
본 개시의 일 실시예에 따르면, 문자열 교정 기반의 유사 구조 분자 생성 모델 학습 방법은, 학습 데이터셋에 포함된 소스 분자 및 학습 데이터셋에서 소스 분자와 사전 페어링(paring)된 목표 분자의 문자열 각각에 대해 토큰화를 수행하여 소스 문자열 조각 세트 및 목표 문자열 조각 세트를 획득하는 단계, 소스 문자열 조각 세트 및 목표 문자열 조각 세트를 제1 학습 모델에 입력하여 소스 문자열 조각 세트가 임베딩된 소스 임베딩 조각 세트 및 목표 문자열 조각 세트가 임베딩된 목표 임베딩 조각 세트를 획득하는 및 소스 임베딩 조각 세트 및 목표 임베딩 조각 세트를 제2 학습 모델에 입력하여, 소스 임베딩 조각 세트에서 하나 이상의 임베딩 조각이 수정된 최종 임베딩 조각 세트를 출력하도록 제2 학습 모델을 학습시키는 단계를 포함할 수 있다.

【대표도】

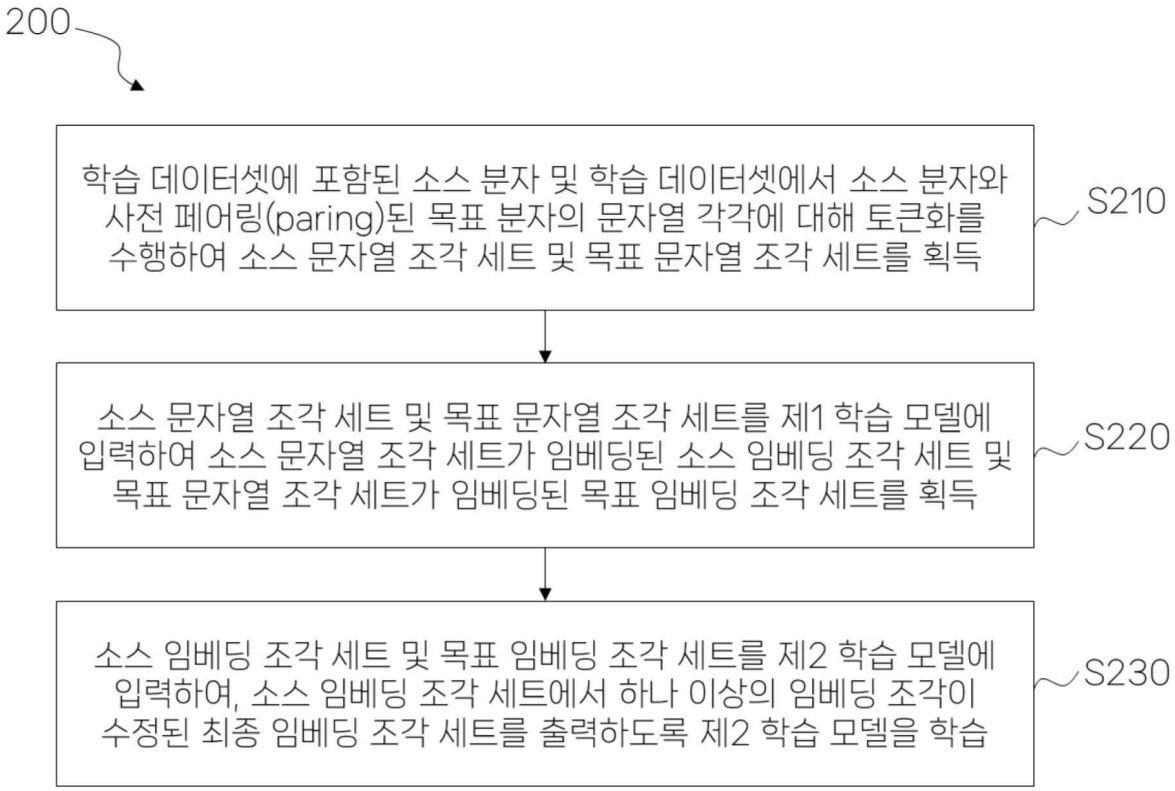
도 2

【도면】

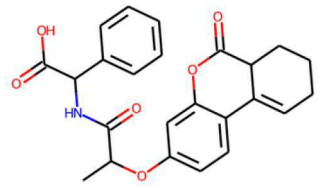
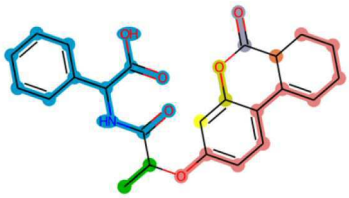
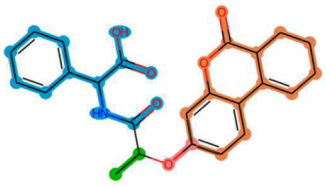
【도 1】



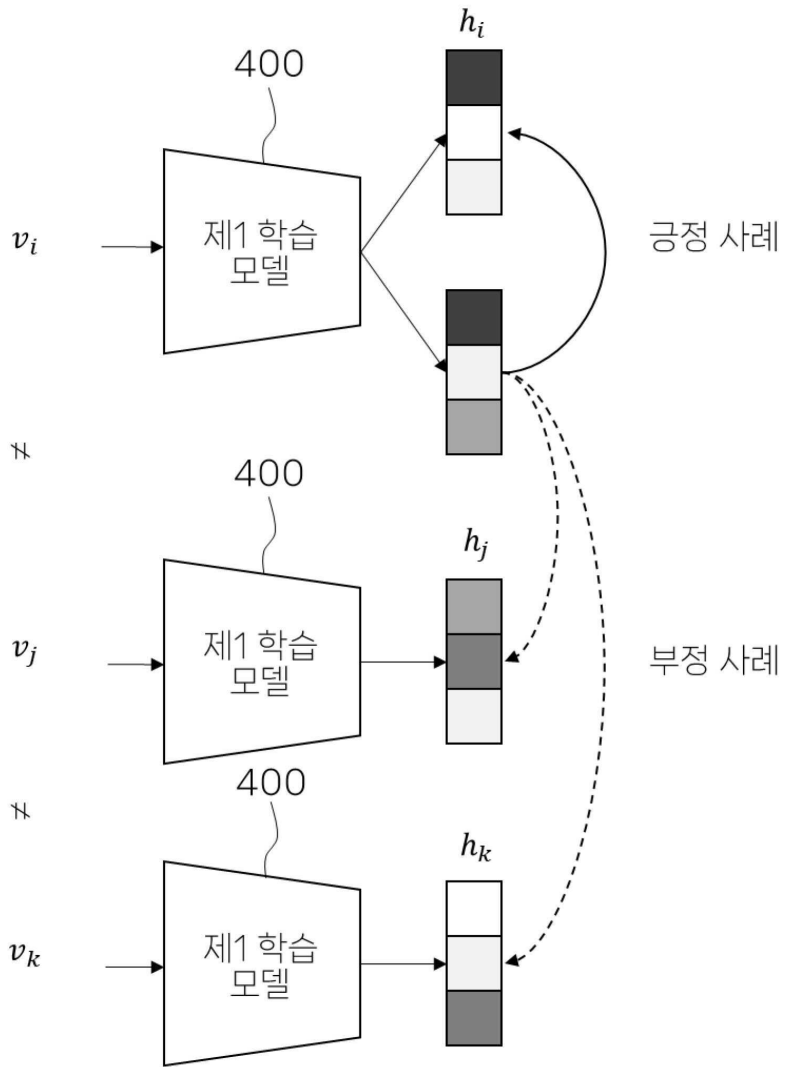
【도 2】


200
학습 데이터셋에 포함된 소스 분자 및 학습 데이터셋에서 소스 분자와 사전 페어링(paring)된 목표 분자의 문자열 각각에 대해 토큰화를 수행하여 소스 문자열 조각 세트 및 목표 문자열 조각 세트를 획득
S210
소스 문자열 조각 세트 및 목표 문자열 조각 세트를 제1 학습 모델에 입력하여 소스 문자열 조각 세트가 임베딩된 소스 임베딩 조각 세트 및 목표 문자열 조각 세트가 임베딩된 목표 임베딩 조각 세트를 획득
S220
소스 임베딩 조각 세트 및 목표 임베딩 조각 세트를 제2 학습 모델에 입력하여, 소스 임베딩 조각 세트에서 하나 이상의 임베딩 조각이 수정된 최종 임베딩 조각 세트를 출력하도록 제2 학습 모델을 학습
S230

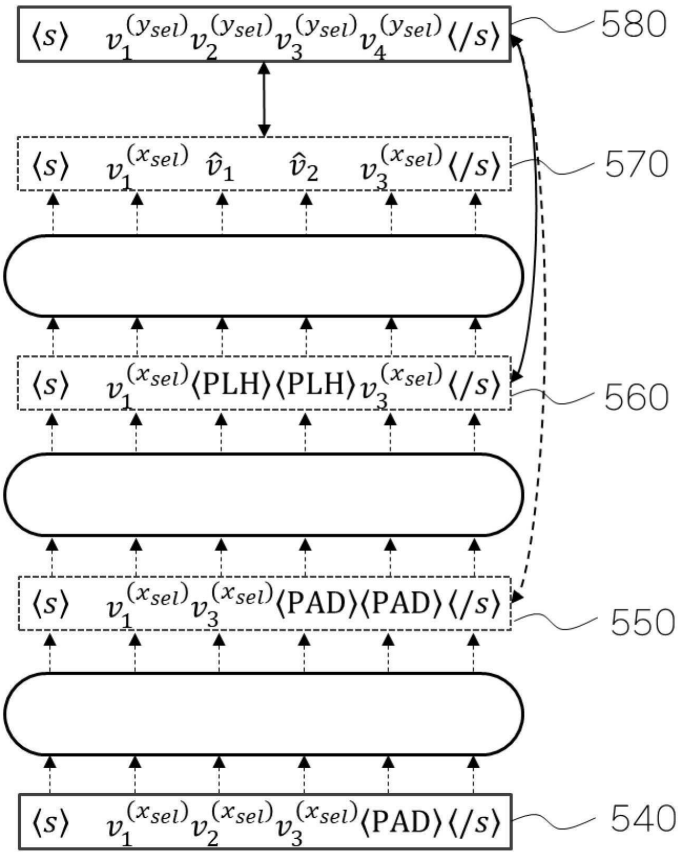
【도 3】

분자	문자열
 310	'[C][C][Branch2][Ring1][=Branch1][C][=Branch1][C][=O][N][C][Branch1][=Branch1][C][Branch1][C][O][=O][C][=C][C][=C][C][=C][Ring1][=Branch1][O][C][=C][C][=C][C][=C][C][C][C][C][Ring1][=Branch1][C][=Branch1][C][=O][O][C][Ring1][O][=C][Ring1][#C]' 320
	['[C][C]', '[Branch2][Ring1][=Branch1][C][=Branch1][C][=O][N][C][Branch1][=Branch1][C][Branch1][C][O][=O][C][=C][C][=C][C][=C][Ring1][=Branch1]', '[O][C][=C][C][=C][C][=C][C][C][C][C][Ring1][=Branch1]', '[C]', '[=Branch1][C][=O]', '[O][C][Ring1][O]', '[=C][Ring1][#C]'] 330
	['[C][C]', '[Branch2][Ring1][=Branch1][C][=Branch1][C][=O][N][C][Branch1][=Branch1][C][Branch1][C][O][=O][C][=C][C][=C][C][=C][Ring1][=Branch1]', '[O][C][=C][C][=C][C][=C][C][C][C][C][Ring1][=Branch1][C][=Branch1][C][=O][O][C][Ring1][O][=C][Ring1][#C]'] 340

【도 4】



【도 5】


$\langle s\rangle \quad v_1^{(y_{sel})} v_2^{(y_{sel})} v_3^{(y_{sel})} v_4^{(y_{sel})} \langle /s\rangle$
580
$\langle s\rangle \quad v_1^{(x_{sel})} \quad \hat{v}_1 \quad \hat{v}_2 \quad v_3^{(x_{sel})} \langle /s\rangle$
570
$\langle s\rangle \quad v_1^{(x_{sel})} \langle \mathrm{PLH}\rangle \langle \mathrm{PLH}\rangle v_3^{(x_{sel})} \langle /s\rangle$
560
$\langle s\rangle \quad v_1^{(x_{sel})} v_3^{(x_{sel})} \langle \mathrm{PAD}\rangle \langle \mathrm{PAD}\rangle \langle /s\rangle$
550
$\langle s\rangle \quad v_1^{(x_{sel})} v_2^{(x_{sel})} v_3^{(x_{sel})} \langle \mathrm{PAD}\rangle \langle /s\rangle$
540