【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】 분자 생성 모델을 위한 학습 방법 및 장치

【발명의 영문명칭】 TRAINING METHOD AND DEVICE FOR MOLECULAR GENERATION MODEL

【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】 서상민

【성명의 영문표기】 SANGMIN SEO

【주민등록번호】 930507-1XXXXXX

【우편번호】 03724

【주소】 서울특별시 서대문구 연희로14길 29

【출원언어】 국어

【심사청구】 청구

【공지예외적용대상증명서류의 내용】

【공개형태】 논문 공개

【공개일자】 2023.01.19

【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】	60	면	0 원
【우선권주장료】	0	건	0 원
【심사청구료】	10	항	583,000 원
【합계】	629,000원		
【감면사유】	전담조직(50%감면)[1]		
【감면후 수수료】	314,500 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】

분자 생성 모델을 위한 학습 방법 및 장치{TRAINING METHOD AND DEVICE FOR MOLECULAR GENERATION MODEL}

【기술분야】

【0001】본 개시는 분자 생성 모델을 위한 학습 방법 및 장치에 관한 것이다.

【발명의 배경이 되는 기술】

【0003】구조 제약 분자 생성은 목표 지향 분자 최적화 연구에서 어려운 문제에 해당한다. 구조 제약 분자 생성의 목표는 소스 약물의 분자 구조와 유사하면서도 향상된 표적 화학적 특성을 가진 새로운 분자를 생성하는 것이다. 유기 화학의 전통적인 접근 방식은 특정 질병에 대한 잠재적 활성을 가진 새로운 약물 후보를 식별하기 위해 먼저 소스 약물의 구조 중에서 표적 생물학적 실체와 결합하는 주요 영역을 조사하고, 해당 부분을 제외한 나머지 부분에 대해 치환 가능한 분자 모티프 조합을 고려하여 분자 구조의 식별을 포함한다. 그러나 이러한 brute-force-like 접근법은 10^{30}-10^{60} 범위로 추정되는 약물과 같은 화학 공간의 크기가 크기 때문에 상당한 전문 지식과 막대한 비용이 필요하다.

【0004】 이러한 비효율성 문제를 해결하기 위해 다양한 컴퓨터 지원 약물 설계 방법, 특히 인공 지능(AI) 기술 기반 응용 프로그램이 제안되었다. 그러나, 종래의 다양한 인공 지능 기술 기반 분자 생성 모델은 특정한 화학적 속성을 만족하는 분자를 생성하는 데에는 효과적이나, 동시에 소스 분자와 유사한 구조를 갖는 분자를 생성하는 것을 동시에 달성하기에는 아직 개선의 여지가 남아있다.

【발명의 내용】

【해결하고자 하는 과제】

【0006】 본 개시에서는 상술한 문제를 해결하기 위하여 화학적 속성 및 구조적 유사도를 모두 만족하는 분자 생성 모델이 제공된다.

【과제의 해결 수단】

【0008】 본 개시의 일 실시예에 따르면, 분자 생성 모델을 위한 학습 방법은 소스 분자 모델, 소스 분자 모델과의 구조적 유사도가 제1 임계치 초과인 타겟 분자 모델, 소스 분자 모델 또는 타겟 분자 모델 중 하나 이상의 모델과의 구조적 유사도가 제1 임계치 이하인 네거티브 분자 모델을 포함하는 학습 데이터셋을 획득하는 단계, 학습 데이터셋 및 제1 손실 함수를 기초로, 소스 분자 모델과 타겟 분자 모델 사이의 거리를 조정하도록 분자 생성 모델을 학습시키는 단계 및 학습 데이터셋 및 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 소스 분자 모델과 네거티

브 분자 모델 사이의 거리 및 타겟 분자 모델과 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리를 조정하도록 분자 생성 모델을 학습시키는 단계를 포함할 수 있다.

【0009】 일 실시예에 따르면, 학습 데이터셋 및 제1 손실 함수를 기초로, 소스 분자 모델과 타겟 분자 모델 사이의 거리를 조정하도록 분자 생성 모델을 학습시키는 단계는 학습 데이터셋 및 제1 손실 함수를 기초로, 소스 분자 모델과 타겟 분자 모델 사이의 거리가 가까워지도록 분자 생성 모델을 학습시키는 단계를 포함할 수 있다.

【0010】 일 실시예에 따르면, 학습 데이터셋 및 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 소스 분자 모델과 네거티브 분자 모델 사이의 거리 및 타겟 분자 모델과 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리를 조정하도록 분자 생성 모델을 학습시키는 단계는 학습 데이터셋 및 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 소스 분자 모델과 네거티브 분자 모델 사이의 거리 및 타겟 분자 모델과 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리가 멀어지도록 분자 생성 모델을 학습시키는 단계를 포함할 수 있다.

【0011】 일 실시예에 따르면, 학습 데이터셋 및 보상 함수를 기초로, 소스 분자 모델로부터 소스 분자 모델과의 구조적 유사도가 제2 임계치 초과인 분자 모델이 출력되도록, 분자 생성 모델을 학습시키는 단계를 더 포함할 수 있다.

【0012】일 실시예에 따르면, 학습 데이터셋 및 보상 함수를 기초로, 소스 분자 모델로부터 소스 분자 모델과의 구조적 유사도가 제2 임계치 초과인 분자 모델이 출력되도록, 분자 생성 모델을 학습시키는 단계는 분자 생성 모델에 소스 분자 모델을 입력하여 출력 분자 모델을 획득하는 단계 및 출력 분자 모델과 소스 분자 모델을 비교한 결과 출력 분자 모델과 소스 분자 모델 사이의 구조적 유사도가 제2 임계치를 초과하는지 여부를 기초로, 출력 분자 모델과 연관된 포지티브 가중치 또는 네거티브 가중치를 산출하여 분자 생성 모델에 부여하는 단계를 포함할 수 있다.

【0013】일 실시예에 따르면, 출력 분자 모델과 소스 분자 모델을 비교한 결과 출력 분자 모델과 소스 분자 모델 사이의 구조적 유사도가 제2 임계치를 초과하는지 여부를 기초로, 출력 분자 모델과 연관된 포지티브 가중치 또는 네거티브 가중치를 산출하여 분자 생성 모델에 부여하는 단계는 출력 분자 모델과 소스 분자 모델을 비교한 결과 출력 분자 모델과 소스 분자 모델 사이의 구조적 유사도가 제2 임계치를 초과하는지 여부 및 출력 분자 모델의 화학적 속성 스코어가 소스 분자 모델의 화학적 속성 스코어를 초과하는지 여부를 기초로, 출력 분자 모델과 연관된 포지티브 가중치 또는 네거티브 가중치를 산출하여 분자 생성 모델에 부여하는 단계를 포함할 수 있다.

【0014】일 실시예에 따르면, 학습 데이터셋 및 보상 함수를 기초로, 소스 분자 모델로부터 소스 분자 모델과의 구조적 유사도가 제2 임계치 초과인 분자 모델이 출력되도록, 분자 생성 모델을 학습시키는 단계는 학습 데이터셋 및 보상 함

수를 기초로, 소스 분자 모델로부터, 소스 분자 모델과의 구조적 유사도가 제2 임계치 초과이고, 소스 분자 모델의 화학적 속성 스코어보다 큰 화학적 속성 스코어를 갖는 분자 모델이 출력되도록, 분자 생성 모델을 학습시키는 단계를 포함할 수 있다.

【0015】 일 실시예에 따르면, 타겟 분자 모델의 화학적 속성 스코어는 소스 분자 모델의 화학적 속성 스코어보다 크게 구성될 수 있다.

【0016】 본 개시의 다른 실시예에 따르면, 분자 생성 모델을 위한 학습 방법을 실행시키도록 컴퓨터로 판독 가능한 기록매체에 기록된 컴퓨터 프로그램이 제공될 수 있다.

【0017】 본 개시의 또 다른 실시예에 따르면, 분자 생성 모델을 위한 학습 장치는 분자 생성 모델과 연관된 데이터를 저장하는 메모리 및 메모리와 연결되어 분자 생성 모델을 학습시키는 적어도 하나의 프로세서를 포함하고, 적어도 하나의 프로세서는 소스 분자 모델, 소스 분자 모델과의 구조적 유사도가 제1 임계치 초과인 타겟 분자 모델, 소스 분자 모델 또는 타겟 분자 모델 중 하나 이상의 모델과의 구조적 유사도가 제1 임계치 이하인 네거티브 분자 모델을 포함하는 학습 데이터셋을 획득하는 것 학습 데이터셋 및 제1 손실 함수를 기초로, 소스 분자 모델과 타겟 분자 모델 사이의 거리를 조정하도록 분자 생성 모델을 학습시키는 것과 학습 데이터셋 및 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 소스 분자 모델과 네거티브 분자 모델 사이의 거리 및 타겟 분자 모델과 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리를 조정하도록 분자 생성 모델을 학습시키는 것을 실행하도

록 구성된 명령어들을 포함할 수 있다.

【도면의 간단한 설명】

【0019】 도 1은 본 개시의 일 실시예에 따른 분자 생성 모델의 동작을 나타내는 아키텍처이다.

도 2는 본 개시의 일 실시예에 따른 학습 데이터셋을 이용하여 분자 생성 모델을 학습시키는 과정의 일부를 나타내는 아키텍처이다.

도 3은 본 개시의 일 실시예에 따른 입력된 임의의 분자 모델을 이용하여 분자 생성 모델을 학습시키는 과정의 다른 일부를 나타내는 아키텍처이다.

도 4는 본 개시의 일 실시예에 따른 컴퓨팅 장치(410)의 내부 구성을 나타내는 블록도이다.

도 5 내지 7은 본 개시의 일 실시예에 따라 0.40 내지 0.70 범위의 여러 구조적 유사도 임계값에 대한 성공률을 평가한 결과를 나타낸다.

도 8은 본 개시의 일 실시예에 따른 학습 방법의 이점을 입증하기 위하여 DRD2 벤치마크 데이터셋에 대한 절제 실험을 수행한 결과를 나타낸다.

도 9는 본 개시의 일 실시예에 따른 제1 손실 함수(Contractive loss) 및 제2 손실 함수(Margin loss)가 있거나 없는 각 훈련된 모델에 대해 속성, 개선도 및 유사도라는 세 가지 지표를 평가한 결과를 나타낸다.

도 10은 본 개시의 일 실시예에 따른 손실 함수들(contractive & margin)의

평균 구조적 유사도를 평가한 결과를 나타낸다.

도 11은 본 개시의 일 실시예에 따른 분자 생성 모델의 선형 프로젝션 분석 수행 결과를 나타낸다.

도 12는 본 개시의 일 실시예에 따라 소라페닙을 소스 분자 모델로 사용하여 10,000개의 분자 모델을 학습하고 생성한 후 성공률을 비교한 결과를 나타낸다.

도 13은 본 개시의 일 실시예에 따라 분자 생성 모델로부터 육안으로도 소라페닙과 유사한 구조의 분자 모델이 생성된 결과를 나타낸다.

도 14는 본 개시의 일 실시예에 따른 실험례에서 히트 후보가 소라페닙보다 ABCG2에 대한 결합 에너지가 높은 지를 확인하기 위해 AutoDock Vina를 사용하여 결합 에너지를 비교한 결과를 나타낸다.

도 15는 본 개시의 일 실시예에 따른 실험례에서 히트 후보가 소라페닙만큼 BRAF에 대한 결합 친화력이 강한 지를 확인하기 위해 Chimera를 사용하여 수용체-리간드 복합체의 3D 구조에 대한 그래픽을 도시한 결과 및 LigPlot Plus를 사용하여 수용체-리간드 복합체의 2D 구조에 대한 그래픽을 도시한 결과를 나타낸다.

도 16은 본 개시의 일 실시예에 따른 분자 생성 모델을 위한 학습 방법의 흐름도이다.

【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 흐

릴 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 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지

는 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】 또한, 명세서에서 사용되는 '부'라는 용어는 소프트웨어 또는 하드웨어 구성요소를 의미하며, '부'는 어떤 역할들을 수행한다. 그렇지만 '부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. '부'는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 '부'는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 또는 변수들 중 적어도 하나를 포함할 수 있다. 구성요소들과 '부'들은 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '부'들로 결합되거나 추가적인 구성요소들과 '부'들로 더 분리될 수 있다.

【0026】 본 개시의 일 실시예에 따르면 '부'는 프로세서 및 메모리로 구현될 수 있다. '프로세서'는 범용 프로세서, 중앙 처리 장치(CPU), 마이크로프로세서,

디지털 신호 프로세서(DSP), 제어기, 마이크로제어기, 상태 머신 등을 포함하도록 넓게 해석되어야 한다. 몇몇 환경에서는, '프로세서'는 주문형 반도체(ASIC), 프로그램 가능 로직 디바이스(PLD), 필드 프로그램가능 게이트 어레이(FPGA) 등을 지칭할 수도 있다. '프로세서'는, 예를 들어, DSP와 마이크로프로세서의 조합, 복수의 마이크로프로세서들의 조합, DSP 코어와 결합한 하나 이상의 마이크로프로세서들의 조합, 또는 임의의 다른 그러한 구성들의 조합과 같은 처리 디바이스들의 조합을 지칭할 수도 있다. 또한, '메모리'는 전자 정보를 저장 가능한 임의의 전자 컴포넌트를 포함하도록 넓게 해석되어야 한다. '메모리'는 임의 액세스 메모리(RAM), 판독-전용 메모리(ROM), 비-휘발성 임의 액세스 메모리(NVRAM), 프로그램가능 판독-전용 메모리(PROM), 소거-프로그램가능 판독 전용 메모리(EPROM), 전기적으로 소거가능 PROM(EEPROM), 플래쉬 메모리, 자기 또는 광학 데이터 저장장치, 레지스터들 등과 같은 프로세서-판독가능 매체의 다양한 유형들을 지칭할 수도 있다. 프로세서가 메모리로부터 정보를 판독하고/하거나 메모리에 정보를 기록할 수 있다면 메모리는 프로세서와 전자 통신 상태에 있다고 불린다. 프로세서에 집적된 메모리는 프로세서와 전자 통신 상태에 있다.

【0027】 도 1은 본 개시의 일 실시예에 따른 분자 생성 모델(100)의 동작을 나타내는 아키텍처이다. 도시된 바와 같이, 분자 생성 모델(100)은 입력되는 소스 분자 모델(110)로부터 최종 분자 모델(120)을 획득할 수 있다. 여기서, 최종 분자 모델(120)은 소스 분자 모델(110)에서 물리적 및/또는 화학적 속성이 변형된 분자 모델을 지칭할 수 있다. 한편, 분자 생성 모델(100)과 연관된 동작은 컴퓨팅 장치

의 적어도 하나의 프로세서에 의해 수행되며, 컴퓨팅 장치의 하드웨어적 구성은 도 4에서 상세히 후술된다.

【0028】 소스 분자 모델(110)은 최종 분자 모델(120)과 구조적으로 유사할 수 있다. 즉, 분자 생성 모델(100)은 소스 분자 모델(110)과 최종 분자 모델(120) 사이의 구조적 유사도가 사전 결정된 임계치를 초과하도록 구성될 수 있다. 이 경우, 구조적 유사도는 타니모토(Tanimoto) 유사도 측정 방식에 따라 측정될 수 있다.

【0029】 소스 분자 모델(110)은 최종 분자 모델(120)과 화학적 속성 스코어는 유사하지 않을 수 있다. 구체적으로, 최종 분자 모델(120)의 화학적 속성 스코어는 소스 분자 모델(110)의 화학적 속성 스코어보다 크게 구성될 수 있다. 이는 최종 분자 모델(120)의 화학적 속성이 소스 분자 모델(110)과 비교하여 보다 개선된 것을 의미할 수 있다. 또한, 화학적 속성 스코어는 소스 분자 모델(110) 및/또는 최종 분자 모델(120)과 같은 분자 모델의 특정한 화학적 반응 정도를 의미할 수 있다. 예를 들어, 분자 모델이 약물인 경우, 화학적 속성 스코어는 인체 내에서 분자 모델이 특정 세포와 반응하는 정도를 의미할 수 있다.

【0030】 프로세서는 소스 분자 모델(110)로부터 화학적 속성이 향상되고 분자 구조가 유사한 최종 분자 모델(120)이 출력되도록 분자 생성 모델(100)을 학습시킬 수 있다. 이를 위해, 프로세서는 복수의 분자 구조 모델 샘플들이 포함된 학습 데이터셋을 이용하여 분자 생성 모델(100)이 입력되는 분자 모델(여기서, 소스 분자 모델(110))의 화학적 속성 및 분자 구조를 학습하도록 분자 생성 모델(100)을

학습시킬 수 있다. 도 2 및 3에서는 이와 같이 프로세서가 학습 데이터셋을 구성하고, 해당 학습 데이터셋을 이용하여 분자 생성 모델(110)을 학습시키는 과정이 상세히 후술된다.

【0031】 도 2는 본 개시의 일 실시예에 따른 학습 데이터셋(110, 210, 220)을 이용하여 분자 생성 모델(100)을 학습시키는 과정의 일부를 나타내는 아키텍처이다. 도시된 바와 같이, 학습 데이터셋은 소스 분자 모델(110), 타겟 분자 모델(210) 및 네거티브 분자 모델(220) 중 적어도 하나를 포함할 수 있다. 이 경우, 타겟 분자 모델(210)은 소스 분자 모델(110)과 구조적 유사도가 제1 임계치를 초과하도록 구성될 수 있다. 또한, 네거티브 분자 모델(220)은 소스 분자 모델(110)과의 구조적 유사도가 제1 임계치 이하 및 타겟 분자 모델(210)과의 구조적 유사도가 제1 임계치 이하가 되도록 구성될 수 있다. 추가적으로 또는 대안적으로, 네거티브 분자 모델(220)은 소스 분자 모델(110)과의 구조적 유사도가 제1 임계치보다 작은 제2 임계치 이하이고, 타겟 분자 모델(210)과의 구조적 유사도가 제1 임계치보다 작은 제2 임계치 이하가 되도록 구성될 수 있다.

【0032】 프로세서는 학습 데이터셋(110, 210, 220)을 획득할 수 있다. 그리고 나서, 프로세서는 학습 데이터셋(110, 210, 220) 및 제1 손실 함수를 기초로, 소스 분자 모델(110)과 타겟 분자 모델(210) 사이의 거리를 조정하도록 분자 생성 모델(100)을 학습시킬 수 있다. 구체적으로, 프로세서는 학습 데이터셋(110, 210, 220) 및 제1 손실 함수를 기초로, 소스 분자 모델(110)과 타겟 분자 모델(210) 사이의 거리가 짧아지도록 분자 생성 모델(100)을 학습시킬 수 있다. 예를 들어, 프

로세서는 소스 분자 모델(110)을 분자 생성 모델(100)에 입력시켜, 제1 벡터(232)를 획득할 수 있다. 또한, 프로세서는 타겟 분자 모델(210)을 분자 생성 모델(100)에 입력하여, 제2 벡터(234)를 획득할 수 있다. 그리고 나서, 프로세서는 제1 벡터(232)와 제2 벡터(234) 사이의 거리를 좁힐 수 있다. 즉, 프로세서는 제1 손실 함수를 기초로 소스 분자 모델(110)과 타겟 분자 모델(210) 사이의 구조적 유사도를 학습할 수 있다.

【0033】프로세서는 학습 데이터셋(110, 210, 220) 및 제1 손실 함수와는 상이한 제2 손실 함수를 기초로, 소스 분자 모델(110)과 네거티브 분자 모델(220) 사이의 거리를 조정하도록 분자 생성 모델(100)을 학습시킬 수 있다. 구체적으로, 프로세서는 학습 데이터셋(110, 210, 220) 및 제2 손실 함수를 기초로, 소스 분자 모델(110)과 네거티브 분자 모델(220) 사이의 거리가 짧아지도록 분자 생성 모델(100)을 학습시킬 수 있다. 예를 들어, 프로세서는 네거티브 분자 모델(220)을 분자 생성 모델(100)에 입력하여, 제3 벡터(236)를 획득할 수 있다. 그리고 나서, 프로세서는 제1 벡터(232) 와 제3 벡터(236) 사이의 거리를 넓힐 수 있다. 이에 따라, 제1 벡터(232)와 제3 벡터(236) 사이의 거리는 제1 벡터(232)와 제2 벡터(234) 사이의 거리보다 멀게 구성될 수 있다. 즉, 프로세서는 제2 손실 함수를 기초로 소스 분자 모델(110)과 네거티브 분자 모델(220) 사이의 구조적 비유사도를 학습할 수 있다.

【0034】프로세서는 학습 데이터셋(110, 210, 220) 및 제2 손실 함수를 기초로, 타겟 분자 모델(210)과 네거티브 분자 모델(220) 사이의 거리를 조정하도록 분

자 생성 모델(100)을 학습시킬 수 있다. 구체적으로, 프로세서는 학습 데이터셋(110, 210, 220) 및 제2 손실 함수를 기초로, 타겟 분자 모델(210)과 네거티브 분자 모델(220) 사이의 거리를 조정하도록 분자 생성 모델(100)을 학습시킬 수 있다. 예를 들어, 제2 벡터(234)와 제3 벡터(236) 사이의 거리를 넓힐 수 있다. 이에 따라, 제2 벡터(234)와 제3 벡터(236) 사이의 거리는 제1 벡터(232)와 제2 벡터(234) 사이의 거리보다 멀게 구성될 수 있다. 즉, 프로세서는 제2 손실 함수를 기초로 타겟 분자 모델(210)과 네거티브 분자 모델(220) 사이의 구조적 비유사도를 학습할 수 있다.

【0035】 한편, 도시된 바와 같이 제1 벡터(232), 제2 벡터(234) 및 제3 벡터(236)는 분자 생성 모델(100)의 제1 모듈(240)에 각각 소스 분자 모델(110), 타겟 분자 모델(210) 및 네거티브 분자 모델(220)을 입력함으로써 획득될 수 있다. 이 경우, 제1 모듈(240)은 적어도 하나의 인코더를 포함할 수 있다.

【0036】 도 3은 본 개시의 일 실시예에 따른 입력된 임의의 분자 모델(여기서, 소스 분자 모델(110))을 이용하여 분자 생성 모델(100)을 학습시키는 과정의 다른 일부를 나타내는 아키텍처이다. 도 3에서 분자 생성 모델(100)이 학습되는 과정은 도 2에서 분자 생성 모델(100)이 학습되는 과정 이후에 수행될 수 있다. 추가적으로 또는 대안적으로, 분자 생성 모델(100)이 학습되는 과정은 도 2의 분자 생성 모델(100)이 학습되는 과정과 적어도 일부 병렬적으로 수행될 수도 있다.

【0037】 프로세서는 학습 데이터셋(예: 학습 데이터셋(110, 210, 220)) 및 보상 함수를 기초로 입력된 임의의 분자 모델로부터 임의의 분자 모델과의 구조적

유사도가 미리 결정된 임계치를 초과하는 분자 모델이 출력되도록 분자 생성 모델(100)을 학습시킬 수 있다. 즉, 보상 함수는 분자 생성 모델(100)로부터 출력된 분자 모델(여기서, 출력 분자 모델(320))의 물리적 구조 및/또는 화학적 속성을 기초로 분자 생성 모델(100)에 포지티브 가중치를 부여하여 출력된 분자 모델의 구조적 유사도 및 화학적 속성에 대한 학습을 강화할 수 있다. 또한, 보상 함수는 분자 생성 모델(100)로부터 출력된 분자 모델의 물리적 구조 및/또는 화학적 속성을 기초로 분자 생성 모델(100)에 네거티브 가중치를 부여하여 출력된 분자 모델의 구조적 비유사도 및 화학적 속성에 대한 학습을 강화할 수 있다. 이하에서는, 소스 분자 모델(110)이 입력된 경우를 예시로 분자 생성 모델(100)이 학습되는 과정이 상세히 후술된다.

【0038】 먼저, 프로세서는 학습 데이터셋에 포함된 임의의 분자 모델을 분자 생성 모델(100)에 입력하여 분자 모델에 대응하는 벡터를 획득할 수 있다. 예를 들어, 프로세서는 학습 데이터셋에 포함된 소스 분자 모델(110)을 분자 생성 모델(100)(구체적으로는, 제1 모듈(240))에 입력하여 제1 벡터(232)를 획득할 수 있다. 그리고 나서, 획득된 제1 벡터(232)를 제2 모듈(310)에 입력하여 출력 분자 모델(320)을 획득할 수 있다.

【0039】 프로세서는 획득된 출력 분자 모델(320)을 소스 분자 모델(110)과 비교할 수 있다. 구체적으로, 프로세서는 출력 분자 모델(320)과 소스 분자 모델(110) 사이의 구조적 유사도를 산출할 수 있다. 추가적으로 또는 대안적으로, 프로세서는 출력 분자 모델(320)의 화학적 속성 스코어를 산출할 수 있다.

【0040】프로세서는 출력 분자 모델(320)과 소스 분자 모델(110) 사이의 구조적 유사도가 미리 결정된 임계치를 초과하는 경우, 출력 분자 모델(320)과 연관된 포지티브 가중치를 산출할 수 있다. 이 경우, 출력 분자 모델(320)과 연관된 포지티브 가중치는 출력 분자 모델(320)의 구조적 유사도 및/또는 화학적 속성 스코어를 기초로 결정될 수 있다. 그리고 나서, 프로세서는 산출된 포지티브 가중치를 분자 생성 모델(100)에 부여할 수 있다. 반면, 프로세서는 출력 분자 모델(320)과 소스 분자 모델(110) 사이의 구조적 유사도가 미리 결정된 임계치 이하인 경우, 출력 분자 모델(320)과 연관된 네거티브 가중치를 산출할 수 있다. 이 경우, 출력 분자 모델(320)과 연관된 네거티브 가중치는 출력 분자 모델(320)의 구조적 유사도 및/또는 화학적 속성 스코어를 기초로 결정될 수 있다. 그리고 나서, 프로세서는 산출된 네거티브 가중치를 분자 생성 모델(100)에 부여할 수 있다. 즉, 프로세서는 출력 분자 모델(320)이 입력된 소스 분자 모델(110)과 구조적으로 유사한 경우에는 분자 생성 모델(100)에 리워드를 부여하고, 출력 분자 모델(320)이 입력된 소스 분자 모델(110)과 구조적으로 비유사한 경우에는 분자 생성 모델(100)에 페널티를 부여함으로써 소스 분자 모델(110)의 구조에 대한 학습을 강화할 수 있다.

【0041】유사하게, 프로세서는 출력 분자 모델(320)의 화학적 속성 스코어가 미리 결정된 임계치를 초과하는 경우, 출력 분자 모델(320)과 연관된 포지티브 가중치를 산출할 수 있다. 이 경우에도, 출력 분자 모델(320)과 연관된 포지티브 가중치는 출력 분자 모델(320)의 화학적 속성 스코어를 기초로 결정될 수 있다. 그

리고 나서, 프로세서는 산출된 포지티브 가중치를 분자 생성 모델(100)에 부여할 수 있다. 반면, 프로세서는 출력 분자 모델(320)의 화학적 속성 스코어가 미리 결정된 임계치 이하인 경우, 출력 분자 모델(320)과 연관된 네거티브 가중치를 산출할 수 있다. 이 경우 또한, 출력 분자 모델(320)과 연관된 네거티브 가중치는 출력 분자 모델(320)의 화학적 속성 스코어를 기초로 결정될 수 있다. 즉, 프로세서는 출력 분자 모델(320)이 입력된 소스 분자 모델(110)과 화학적 반응 레벨이 유사하거나 및/또는 개선된 경우에는 분자 생성 모델(100)에 리워드를 부여하고, 출력 분자 모델(320)이 입력된 소스 분자 모델(110)보다 화학적 반응 레벨이 낮은 경우에는 분자 생성 모델(100)에 페널티를 부여함으로써, 분자 생성 모델(100)의 화학적 속성에 대한 학습을 강화할 수 있다.

【0042】 한편, 프로세서는 포지티브 가중치 및/또는 네거티브 가중치를 산출함에 있어서, 구조적 유사도 및 화학적 속성 스코어를 동시에 고려할 수도 있다. 예를 들어, 출력 분자 모델(320)의 화학적 속성 스코어가 미리 결정된 임계치를 초과하고, 출력 분자 모델(320)과 소스 분자 모델(110) 사이의 구조적 유사도가 미리 결정된 임계치를 초과하는 경우, 프로세서는 출력 분자 모델(320)의 구조적 유사도 및/또는 화학적 속성 스코어를 기초로 포지티브 가중치를 산출하여 분자 생성 모델(100)에 부여할 수 있다. 다른 예를 들어, 출력 분자 모델(320)의 화학적 속성 스코어가 미리 결정된 임계치 미만이거나, 출력 분자 모델(320)과 소스 분자 모델(110) 사이의 구조적 유사도가 미리 결정된 임계치 이하인 경우, 프로세서는 출력 분자 모델(320)의 구조적 유사도 및/또는 화학적 속성 스코어를 기초로 네거티브

가중치를 산출하여 분자 생성 모델(100)에 부여할 수 있다. 마찬가지로, 화학적 속성 스코어와 구조적 유사도가 각각 모두 해당 임계치의 미만인 경우에도 프로세서는 네거티브 가중치를 산출하여 분자 생성 모델(100)에 부여할 수 있다.

【0043】 도 4는 본 개시의 일 실시예에 따른 컴퓨팅 장치(410)의 내부 구성을 나타내는 블록도이다. 여기서, 컴퓨팅 장치(410)는 본 개시에 따른 분자 생성 모델을 위한 학습 방법을 실행하기 위한 장치를 지칭할 수 있다. 컴퓨팅 장치(410)는 머신러닝을 위한 프로그램 등이 실행 가능하고 유/무선 통신이 가능한 임의의 컴퓨팅 장치로서, 데스크탑, 스마트폰, 태블릿, 노트북 등을 포함할 수 있다.

【0044】 컴퓨팅 장치(410)는 메모리(412), 프로세서(414), 통신 모듈(416) 및 입출력 인터페이스(418)를 포함할 수 있다. 도 4는 도시되지 않았으나, 컴퓨팅 장치(410)는 통신 모듈(416)을 이용하여 네트워크를 통해 정보 및/또는 데이터를 통신할 수 있도록 구성될 수 있다. 또한, 입출력 장치(420)는 입출력 인터페이스(418)를 통해 컴퓨팅 장치(410)에 정보 및/또는 데이터를 입력하거나 컴퓨팅 장치(410)로부터 생성된 정보 및/또는 데이터를 출력하도록 구성될 수 있다.

【0045】 메모리(412)는 비-일시적인 임의의 컴퓨터 판독 가능한 기록매체를 포함할 수 있다. 일 실시예에 따르면, 메모리(412)는 RAM(random access memory), ROM(read only memory), 디스크 드라이브, SSD(solid state drive), 플래시 메모리(flash memory) 등과 같은 비소멸성 대용량 저장 장치(permanent mass storage device)를 포함할 수 있다. 다른 예로서, ROM, SSD, 플래시 메모리, 디스크 드라이브 등과 같은 비소멸성 대용량 저장 장치는 메모리와는 구분되는 별도의 영구 저

장 장치로서 컴퓨팅 장치(410)에 포함될 수 있다. 또한, 메모리(412)에는 운영체제와 적어도 하나의 프로그램 코드(예를 들어, 컴퓨팅 장치(410)에 설치되어 구동되는 머신러닝을 위한 프로그램을 실행하기 위한 코드)가 저장될 수 있다.

【0046】 이러한 소프트웨어 구성요소들은 메모리(412)와는 별도의 컴퓨터에서 판독 가능한 기록매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록매체는 이러한 컴퓨팅 장치(410) 및 외부 서버에 직접 연결가능한 기록매체를 포함할 수 있는데, 예를 들어, 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록매체를 포함할 수 있다. 다른 예로서, 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록매체가 아닌 통신 모듈(416)을 통해 메모리(412)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 어플리케이션의 설치 파일을 배포하는 파일 배포 시스템이 네트워크를 통해 제공하는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여 메모리(412)에 로딩될 수 있다.

【0047】 프로세서(414)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(412) 또는 통신 모듈(416)에 의해 프로세서(414)로 제공될 수 있다. 예를 들어, 프로세서(414)는 메모리(412)와 같은 기록 장치에 저장된 프로그램 코드에 따라 수신되는 명령을 실행하도록 구성될 수 있다.

【0048】 통신 모듈(416)은 네트워크를 통해 컴퓨팅 장치(410)와 외부 서버가 서로 통신하기 위한 구성 또는 기능을 제공할 수 있으며, 컴퓨팅 장치(410) 및/또

는 외부 서버가 다른 사용자 단말 또는 다른 시스템(일례로 별도의 클라우드 시스템 등)과 통신하기 위한 구성 또는 기능을 제공할 수 있다. 일례로, 컴퓨팅 장치(410)의 프로세서(414)가 메모리(412) 등과 같은 기록 장치에 저장된 프로그램 코드에 따라 생성한 요청 또는 데이터는 통신 모듈(416)의 제어에 따라 네트워크를 통해 외부 서버로 전달될 수 있다. 역으로, 외부 서버의 프로세서의 제어에 따라 제공되는 제어 신호나 명령이 네트워크를 거쳐 컴퓨팅 장치(410)의 통신 모듈(416)을 통해 컴퓨팅 장치(410)에 수신될 수 있다.

【0049】컴퓨팅 장치(410)의 입출력 인터페이스(418)는 입출력 장치(420)와의 상호 작용을 위한 수단일 수 있다. 구체적으로, 입출력 인터페이스(418)는 터치스크린 등과 같이 입력과 출력을 수행하기 위한 구성 또는 기능이 하나로 통합된 장치와의 인터페이스를 위한 수단일 수 있다. 이 경우, 입출력 장치(420)는 오디오 센서 및/또는 이미지 센서를 포함한 카메라, 키보드, 마이크로폰, 마우스 등의 입력 장치를 포함할 수 있다. 또한, 입출력 장치(420)는 디스플레이, 스피커, 햅틱 피드백 디바이스(haptic feedback device) 등과 같은 출력 장치를 포함할 수 있다.

【0050】도 4에서는 입출력 장치(420)가 컴퓨팅 장치(410)에 포함되지 않도록 도시되어 있으나, 이에 한정되지 않으며, 컴퓨팅 장치(410)와 하나의 장치로 구성될 수 있다. 또한, 도 4에서는 입출력 인터페이스(418)가 프로세서(414)와 별도로 구성된 요소로서 도시되었으나, 이에 한정되지 않으며, 입출력 인터페이스(418)가 프로세서(414)에 포함되도록 구성될 수 있다.

【0051】컴퓨팅 장치(410)는 도 4의 구성요소들보다 더 많은 구성요소들을 포함할 수 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 일 실시예에 따르면, 컴퓨팅 장치(410)는 상술된 입출력 장치(420) 중 적어도 일부를 포함하도록 구현될 수 있다. 또한, 컴퓨팅 장치(410)는 트랜시버(transceiver), GPS(Global Positioning system) 모듈, 카메라, 각종 센서, 데이터베이스 등과 같은 다른 구성요소들을 더 포함할 수 있다. 예를 들어, 컴퓨팅 장치(410)가 스마트폰인 경우, 일반적으로 스마트폰이 포함하고 있는 구성요소를 포함할 수 있으며, 예를 들어, 가속도 센서, 자이로 센서, 마이크 모듈, 카메라 모듈, 각종 물리적인 버튼, 터치패널을 이용한 버튼, 입출력 포트, 진동을 위한 진동기 등의 다양한 구성요소들이 컴퓨팅 장치(410)에 더 포함되도록 구현될 수 있다.

【0052】컴퓨팅 장치(410) 및 외부 서버 각각에서 머신러닝을 위한 프로그램이 동작하는 동안에, 프로세서(414)는 각각 입출력 인터페이스(418)와 연결된 입출력 장치(420)를 통해 입력되거나 선택된 수치 데이터, 텍스트, 이미지, 영상 등을 수신할 수 있으며, 수신된 수치 데이터, 텍스트, 이미지 및/또는 영상 등을 메모리(412)에 저장하거나 통신 모듈(416) 및 네트워크를 통해 서로에게 제공할 수 있다.

【0053】외부 서버의 프로세서는 복수의 사용자 단말 및/또는 복수의 외부 시스템으로부터 수신된 정보 및/또는 데이터를 관리, 처리 및/또는 저장하도록 구성될 수 있다. 일 실시예에 따르면, 프로세서는 컴퓨팅 장치(410)로부터 수신된 사용자 입력 및 해당 사용자 입력에 따른 로그 데이터를 관리, 처리 및/또는 저장

할 수 있다. 추가적으로 또는 대안적으로, 프로세서는 네트워크와 연결된 별도의 클라우드 시스템, 데이터베이스 등으로부터 컴퓨팅 장치(410)의 머신러닝에 이용되는 알고리즘을 실행하기 위한 프로그램 등을 저장 및/또는 업데이트하도록 구성될 수 있다.

【0054】이하에서는, 본 개시의 분자 생성 모델(도 5 내지 15에서 “COMA”로 표시됨)을 위한 학습 방법을 구현하기 위하여 수행된 실험례가 상세히 후술된다.

【0055】분자 생성 모델의 개요

【0056】본 개시의 분자 생성 모델은 SMILES 문자열을 인코딩 및 디코딩하기 위한 GRU(Gated Recurrent Unit) 기반의 VAE(Variational Autoencoder)로, ASCII 코드를 사용하여 분자 구조를 나타낼 수 있다. 여기서, 각 ASCII 코드는 분자 구조에 포함된 원자, 원자 간 결합 유형 및/또는 결합 구조(예: 가지 구조, 고리 구조 등)를 나타낼 수 있다.

【0057】분자 생성 모델의 인코더(즉, 제1 모듈)는 제1 손실 함수를 기초로 유사한 구조를 가진 두 개의 분자를 잠재 공간에서 서로 가까운 지점에 삽입하는 반면, 제2 손실 함수를 기초로 서로 다른 구조를 가진 두 개의 분자를 잠재 공간에서 가능한 멀리 배치할 수 있다. 즉, 분자 생성 모델의 디코더는 인코더로부터 출력된 잠재 벡터(예를 들어, 도 2의 제1 벡터(232))에서 유효한 SMILES 문자열을 생성하도록 학습된다. 또한, 입력된 분자 모델(예를 들어, 도 3의 소스 분자 모델(110))보다 개선된 화학적 속성을 가진 SMILES 문자열을 선택적으로 생성하기 위하

여 보상 함수를 이용한 강화 학습이 적용될 수 있다.

【0058】 성능 평가

【0059】 분자 생성 모델의 성능을 평가하기 위하여 네 개의 벤치마크 데이터셋(DRD2, QED, pLogP04 및 pLogP06)이 이용되었다. DRD2를 이용한 학습 목적은 Tanimoto 유사도가 0.4 이상이라는 조건 하에서 소스 분자 모델보다 도파민 수용체 D2에 대하여 더 활성인 새로운 분자를 생성하는 것이고, QED를 이용한 학습 목적은 Tanimoto 유사도가 0.4 이상이라는 조건 하에서 소스 분자 모델보다 더 (기존의) 약물과 유사한 새로운 분자를 생성하는 것이다. QED 스코어의 경우 범위는 [0,1]이며 값이 클수록 약물과의 유사도가 더 높음을 나타낸다. 마지막으로, pLogP04 및 pLogP06 작업의 목표는 각각 0.4 및 0.6을 소스 분자 모델과의 구조적 유사도 임계치로 하여 페널티 logP의 스코어를 향상시키는 것이다. 여기서, 페널티 logP의 스코어는 logP의 스코어에서 분자 구조 내 고리(ring)의 크기 값과 합성 접근성 점수를 차감한 것을 나타낸다.

【0060】 비교 모델

【0061】 본 개시의 분자 구조 비교 모델은 JTVAE, VJTNN, VJTNN+GAN, CORE, HierG2G, HierG2G+BT 및 UGMMT의 7가지 최신 모델과 비교되었다. JTVAE는 베이지안 최적화 방법을 사용하여 분자 특성을 최적화하는 그래프 기반 분자 생성 모델을 나타낸다. VJTNN은 뉴럴 어텐션 기능이 추가된 JTVAE의 최신 버전 모델이고, VJTNN+GAN은 적대적 훈련이 있는 보다 최신 버전의 모델에 해당한다. CORE는 copy-and-refine 전략을 사용하여 분자를 생성하는 VJTNN+GAN의 개선된 버전이다.

HierG2G는 계층적 인코딩 방식을 사용하는 그래프 기반 생성 모델이다. HierG2G+BT는 데이터 증대를 위한 back-translation 단계를 추가한 HierG2G의 개선된 버전이다. UGMMT는 비지도 학습 방식을 사용하여 훈련되는 SMILES 기반 생성 모델이다.

【0062】평가 지표

【0063】구조에 제약이 있는 분자 생성의 다양한 평가 지표를 사용하여 분자 생성 모델과 다른 모델들을 평가했다. 먼저 각 벤치마크 작업의 학습 데이터셋으로 모든 모델을 교육하고 테스트 데이터셋의 각 소스 분자에 대해 분자를 20번 생성한 다음 생성된 분자를 7개의 지표로 평가했다.

【0064】*유효성(Validity): 테스트 데이터에서 생성된 유효한 SMILES 문자열의 비율

【0065】*Novelty: 훈련 데이터에 없는 유효한 SMILES 문자열의 비율

【0066】*속성(Property): 유효한 SMILES 문자열의 속성 스코어 평균

【0067】*개선(Improvement): 생성된 SMILES 문자열과 소스 SMILES 문자열 간의 속성 스코어 차이의 평균

【0068】*유사도(Similarity): 생성된 SMILES 문자열과 소스 SMILES 문자열 간의 Tanimoto 구조적 유사도의 평균

【0069】*다양성(Diversity): 생성된 SMILES 문자열 간의 Tanimoto 쌍 별 비유사도의 평균

【0070】*성공률(Success rate): 화학적 속성(여기서, 약물적 속성) 개선과 구조적 유사도 기준을 모두 만족하는 유효하고 새로운 SMILES 문자열의 비율

【0071】성공률 비교

【0072】도 5 내지 7은 본 개시의 일 실시예에 따라 0.40 내지 0.70 범위의 여러 구조적 유사도 임계값에 대한 성공률을 평가한 결과를 나타낸다. 성공률은 모델이 세 가지 제약 조건인 Novelty, 화학적 속성의 개선 및 구조적 유사도라는 조건을 동시에 충족하는 유효 분자를 얼마나 많이 생성하는지 측정하는 데 가장 중요한 척도이기 때문이다.

【0073】도 5를 참고하면, 본 개시의 분자 생성 모델은 여러 임계값 조건에서 기본 모델과 동등하거나 더 나은 성능을 보였고, 분자 생성 모델이 기본 모델보다 구조적 유사도 제약 조건(0.55 내지 0.70)에서 다른 조건들을 만족하는 분자를 보다 우수하게 생성할 수 있음을 확인했다.

【0074】도 6을 참고하면, 정량적 비교를 위해 구조적 유사도의 임계값에 대한 평균 성공률을 계산하고, DRD2, QED, pLogP04 및 pLogP06에 대해 본 개시의 분자 생성 모델의 평균 스코어가 각각 0.180, 0.301, 0.154 및 0.213임을 확인했다. 기본 모델과 비교할 때 분자 생성 모델의 스코어는 최신 모델과 비교하여 0.002 내지 0.240으로 더 높아 구조에 제약이 있는 분자 생성에 더 적합한 모델임을 확인했다.

【0075】전반적인 성능

【0076】도 7은 나머지 6개의 지표는 본 개시의 분자 생성 모델 및 다른 모델들의 특성을 나타낸다. 분자 생성 모델의 유효성과 Novelty는 모든 벤치마크 데이터셋에서 기본 모델을 능가했다. JTVAE의 경우, DRD2 및 QED 작업에서 분자 생성 모델보다 우수한 구조적 유사도를 보였지만 화학적 속성 스코어를 동시에 향상시키지 못하여 성공률이 낮았다. 전반적인 평가를 위해 각 모델의 총 유효성, 속성, 개선도, 유사도, Novelty 및 다양성 스코어를 계산했다. QED를 제외하고 분자 생성 모델이 가장 높은 스코어를 나타내 상술한 성공률 분석과 일치함을 확인하였다. 이러한 실험 결과는 제1 손실 함수, 제2 손실 함수 및 보상 함수를 이용한 본 개시의 학습 방법이 구조적 제약이 있는 분자 생성에 효과적임을 입증한다.

【0077】제1 손실 함수 및 제2 손실 함수에 대한 애블레이션(ablation) 연구

【0078】도 8은 본 개시의 일 실시예에 따른 학습 방법의 이점을 입증하기 위하여 DRD2 벤치마크 데이터셋에 대한 절제 실험을 수행한 결과를 나타낸다. 도시된 바와 같이, 제1 손실 함수(여기서, Contractive loss) 및 제2 손실 함수(여기서, Margin loss)를 모두 이용한 학습을 통해 높은 구조적 유사도를 달성할 수 있었다. Kruskal-Wallis H 검정을 사용하여 제1 손실 함수 및 제2 손실 함수를 모두 이용한 학습이 구조적 유사도에 대하여 통계적으로 유의하게 개선됨을 확인하였다.

【0079】성능 비교

【0080】도 9는 본 개시의 일 실시예에 따른 제1 손실 함수(여기서, Contractive loss) 및 제2 손실 함수(여기서, Margin loss)가 있거나 없는 각 훈련

된 모델에 대해 속성, 개선도 및 유사도라는 세 가지 지표를 평가한 결과를 나타낸다. 유사도에는 눈에 띄는 차이가 없었지만 제1 손실 함수 및 제2 손실 함수를 모두 사용한 경우에만 높은 속성 및 개선 스코어가 관찰되었다. 이러한 결과는 제1 손실 함수 및 제2 손실 함수가 구조적 제약이 있는 분자 생성에서 중요한 역할을 한다는 것을 나타낸다.

【0081】도 10은 본 개시의 일 실시예에 따른 손실 함수들(여기서, contractive & margin)의 평균 구조적 유사도를 평가한 결과를 나타낸다. DRD2 데이터셋의 유사한 분자 생성 작업에서 손실 함수들의 조합이 종래의 triplet loss와 contrastive loss를 능가한다는 것을 확인했다. 분자 생성 모델은 소스 분자 모델에 대해 평균 유사도가 0.423인 표적 분자를 생성한 반면, 종래의 triplet loss 및 contrastive loss를 사용한 경우, 평균 유사도가 각각 0.269 및 0.262인 표적 분자를 생성한 것으로 나타났다.

【0082】잠재 공간 분석

【0083】도 11은 본 개시의 일 실시예에 따른 분자 생성 모델의 선형 프로젝션 분석 수행 결과를 나타낸다. 분자 생성 모델의 장점은 제1 손실 함수 및 제2 손실 함수를 이용하여 구조적 유사도 측면의 성능을 높인다는 것이다. 이를 위해, 분자 생성 모델의 잠재 공간에서 구조가 유사한 분자는 서로 가깝게 만들고 구조적으로 다른 분자는 서로 멀리 떨어지도록 설계되었다. 선형 프로젝션 분석은 상술한 통계 분석에서 사용된 데이터와 동일한 데이터가 사용되었다. 잠재 공간에서 한 점 S6을 선택 하고 그 점에서 시작하여 임의의 방향으로 화살표를 그리고 화살

표 위에 있는 6개의 점에 해당하는 분자 구조를 비교한 결과, 시작점에 인접한 지점은 Tanimoto 유사도가 높고 먼 지점은 유사도 스코어가 낮음을 확인했다. 따라서 제안한 방법이 의도한 만큼 효과적이라고 판단하였다.

【0084】구체적인 실험례: 소라페닙(sorafenib) 내성에 대한 약물 발견

【0085】구조적 제약이 있는 분자 생성은 기존 약물과 유사한 새로운 분자를 생성하여 약물을 이용한 화학적 치료법에 내성이 있는 환자를 위한 약물 후보를 발견하는 데 사용할 수 있다. 약물 후보는 기존 약물의 pharmacophore 특성을 손실하지 않고 약물 내성과 관련된 화학적 특성을 감소시켜 얻을 수 있다. 본 실험례에서 소라페닙 내성 간암 환자에서 화학적 치료법의 치료 효과를 향상시키기 위해 간세포 암종(HCC)에 대한 표적 항암제인 소라페닙에 분자 생성 모델을 적용하였다.

【0086】소라페닙 내성과 ABC 수송체 사이의 연관성

【0087】소라페닙은 Raf/Mek/Erk 경로에서 종양 세포에서 세포 증식 및 혈관 신생을 억제하는 단백질 키나아제의 억제제이다. 소라페닙의 중간 정도의 치료 효과와 은폐된 약물 내성으로 인해, 소라페닙의 대안으로 사용될 수 있는 신약 후보 물질의 발견은 중요한 연구 과제에 해당한다. 소라페닙 내성과 관련된 의심되는 메커니즘 중 하나는 세포에서 약물을 끌어내는 ATP 결합 카세트(ABC) 운반체이다. 소라페닙을 포함한 다중 표적 티로신 키나아제 억제제(TKI)는 ABC 수송체 기질로 작용하기 때문에, ABC 수송체는 소라페닙이 치료 표적 단백질에 결합하기 전에 HCC 종양 세포에서 소라페닙을 빼내는 것으로 분석된다. 따라서, 소라페닙의 치료 표적 단백질에 대한 친화력의 손실 없이 ABC 수송체 단백질에 대한 소라페닙의 결합

친화도를 감소시킨다면 간세포 암종 환자에서 소라페닙 내성을 완화함과 동시에 화학적 치료법의 효과를 높일 수 있다.

【0088】ABCG2에 대한 결합 선호도 최적화

【0089】소라페닙과 유사한 적중 발견을 위한 분자 생성 모델의 proof-of-concept를 수행하기 위해, 본 실험례에서 소라페닙의 표적 키나아제인 세린/트레오닌-단백질 키나아제 B-raf (BRAF)에 대한 친화력 손실 없이 ABCG2(ABC subfamily G member 2)의 단백질에 대한 결합 친화도 스코어를 줄이면서 소라페닙의 하위 구조를 보존하는 것이 목표로 채택되었다. 이를 위해, ChEMBL 데이터베이스에서 16,000여개의 SMILES 문자열을 선별 하고 분자 생성 모델 및 UGMMT에 대한 학습 데이터셋이 구성되었다. 여기서, UGMMT는 최신 SMILES 기반 모델이기 때문에 선택되었다.

【0090】도 12는 본 개시의 일 실시예에 따라 소라페닙을 소스 분자 모델로 사용하여 10,000개의 분자 모델을 학습하고 생성한 후 성공률을 비교한 결과를 나타낸다. 성공률은 표적 분자 조건(Tanimoto 유사도 > 0.4 및 ABCG2에 대한 친화도 스코어 < 4.7)을 만족하는 신규 분자 모델의 비율로 정의되었다. 분자 생성 모델(여기서, COMA)은 높은 성공률(0.174)을 보인 반면 UGMMT는 낮은 성공률(0.001)을 보였다. UGMMT가 분자 생성 모델보다 ABCG2에 대한 결합 친화도를 더 많이 감소시켰지만 UGMMT가 소라페닙과 유사한 분자를 생성하지 못했기 때문에 UGMMT가 낮은 성공률을 보였다.

【0091】도 13은 본 개시의 일 실시예에 따라 분자 생성 모델로부터 육안으로도 소라페닙과 유사한 구조의 분자 모델이 생성된 결과를 나타낸다. 분자 생성 모델에 의해 생성된 분자들은 모두 표적 분자 조건(Tanimoto 유사도 > 0.4 및 ABCG2에 대한 친화도 스코어 < 4.7)을 만족함을 확인했다.

【0092】도 14는 본 개시의 일 실시예에 따른 실험례에서 히트 후보가 소라페닙보다 ABCG2에 대한 결합 에너지가 높은 지를 확인하기 위해 AutoDock Vina를 사용하여 결합 에너지를 비교한 결과를 나타낸다. 도킹된 포즈는 AutoDock Vina 1.2.3을 사용하여 평가되었고 Chimera 1.16 및 LigPlot Plus 2.2.5를 사용하여 시각화되었다. 리간드(ligand)를 준비하기 위해 먼저 분자 생성 모델에 의해 생성된 10,000개의 분자 중 19개의 고유한 분자를 중복 분자 모델 제거를 통해 추출했고, Open Babel 3.1.1을 사용하여 분자의 3차원 좌표를 생성했다. 그리고 나서, pH 7.4에서 분자를 양성자화하고 AutoDock 용 Python 라이브러리인 meeko 0.3.0을 사용하여 pdbqt 파일을 만들었다. ABCG2 및 BRAF 수용체를 준비하기 위해 ABCG2 및 BRAF에 대한 PDB 데이터베이스에서 각각 6VXH 및 1UWH를 포함한 3D 구조 파일을 다운로드하고 ADFR 소프트웨어 1.0을 사용하여 수소를 확인했다. 박스 중심과 크기를 정의하기 위해 Chimera를 활용하고 수용체 리간드 쌍당 20개의 포즈를 생성하기 위해 AutoDock Vina를 실행했다. 그리고 나서, 수용체-리간드 쌍 당 결합 에너지 스코어가 가장 높은 최상의 포즈를 선택하여 소라페닙과 비교했다. 도 14를 참고하면, 15개의 분자가 소라페닙보다 ABCG2에 대해 더 높은 결합 에너지를 가짐을 확인할 수 있다. 따라서 이러한 분자는 소라페닙의 대안으로 히트 후보 분자가 될 수

있다.

【0093】 도 15는 본 개시의 일 실시예에 따른 실험례에서 히트 후보가 소라페닙만큼 BRAF에 대한 결합 친화력이 강한 지를 확인하기 위해 Chimera를 사용하여 수용체-리간드 복합체의 3D 구조에 대한 그래픽을 도시한 결과 및 LigPlot Plus를 사용하여 수용체-리간드 복합체의 2D 구조에 대한 그래픽을 도시한 결과를 나타낸다. 도 15를 참고하면, 히트 후보 분자들이 BRAF에서 소라페닙에 대한 결합 포켓과 잘 맞는다는 것을 확인할 수 있다. 또한 생성된 히트 후보 분자와 소라페닙이 구조 분석 도구와 기본 매개변수를 사용하여 반 데르 발스 반경을 기반으로 공통 원자간 접촉을 가짐을 확인했다. 또한, LigPlot Plus에 의해 그려진 2D 플롯을 통해 분자가 Glu500(A) 및 Cys531(A)를 포함한 아미노산 잔기와 수소 결합을 갖고 BRAF에서 소라페닙과 상호 작용한다는 것을 확인할 수 있다.

【0094】 합성 접근성 평가

【0095】 Scifinder-n의 역합성 분석을 이용하여 분자 생성 모델에 의해 생성된 분자의 합성 가능성을 평가하였다. 대부분의 분자는 두 단계로 합성될 수 있다. 이는 생성된 분자가 기존 약물 소라페닙과 유사하기 때문에 좋은 합성성을 보장할 수 있었기 때문이며, 분자 생성 모델과 같은 구조적 제약이 있는 분자 생성 모델이 목표 지향적 약물 발견을 위한 실제 작업에 효과적인 도구가 될 것임을 시사한다. 즉, in silico 분석 결과는 분자 생성 모델에 의해 생성된 소라페닙 유도체가 약물 내성이 높은 환자에서 소라페닙의 대체 약물 후보가 될 수 있음을 나타낸다.

【0096】결론

【0097】구조적 제약이 있는 분자 생성을 위한 AI 기반 생성 모델은 효과적인 약물 발견을 위한 솔루션일 뿐만 아니라 화학자 및 약리학자를 위한 강력하고 설명 가능한 도구가 될 수 있다. 기존의 구조 제약 분자 생성 모델은 화학적 속성 개선, 신규성, 소스 분자와의 높은 유사도를 동시에 만족시키는 분자를 생산하는데 한계가 있다. 본 개시의 분자 생성 모델은 두 가지 훈련 단계를 통해 높은 속성 개선과 높은 구조적 유사도를 모두 달성했다. 또한, 유사도 제약 및 속성 개선에 있어 분자 생성 모델은 DRD2, QED, plogP04 및 plogP06의 4가지 벤치마크 데이터셋에서 다양한 최신 모델을 능가하는 성과를 나타냈다.

【0098】구현 세부 정보

【0099】본 개시의 분자 생성 모델은 Python 3.6과 PyTorch 1.10.1 및 RDKit 2021.03.5를 비롯한 여러 오픈 소스 도구를 사용하여 구현되었다. SMILES kekulization, SMILES 유효성 검사, Tanimoto 유사도 계산, QED 추정 에는 화학정보 학용 오픈 소스 도구 인 RDKit이 사용되었다. 오픈 소스 기계 학습 프레임워크인 PyTorch는 분자 생성 모델의 신경망을 구성하고 훈련하는 데 사용되었다. 모든 실험은 64GB 메모리와 GeForce RTX 3090이 장착된 Ubuntu 18.04.6 LTS에서 수행되었다.

【0100】타니모토(Tanimoto) 유사도

【0101】범위가 0에서 1인 Tanimoto 유사도는 Morgan 지문으로 표시되는 원자 쌍 및 위상 비틀림과 같은 분자 구조를 비교한다. 본 개시의 실험례에서 Morgan 지문은 반경이 2이고 2048비트인 RDKit을 사용하여 생성된 이진 벡터이다. 해당 지문 벡터 FP(x) = (p_1, p_2, ..., p_{2048}) 및 FP(y) = (q_1, q_2, ..., q_{2048})이 있는 두 개의 SMILES 문자열 x 및 y에 대해 Tanimoto 유사도 스코어는 수학식 1에 따라 산출되었다.

【0102】**【수학식 1】**

$$\mathcal{T}(x,y) = \frac{\sum_{i=1}^{2048} p_i q_i}{\sum_{j=1}^{2048} (p_j + q_j - p_j q_j)}$$

【0103】결합 친화도 예측

【0104】ABCG2 및 BRAF에 대한 결합 친화도 스코어를 예측하는 것은 소라페닙 내성에 대한 COMA 적용에 매우 중요하다. 본 개시의 실험례에서 가상 스크리닝을 위한 PyTorch 기반 라이브러리인 DeepPurpose는 460만 쌍 이상의 분자에 대한 정확하고 높은 처리량 친화도 예측에 사용되었다. 또한, UGMMT 및 COMA에 대한 훈련 데이터셋을 생성하고 COMA에서 강화 학습의 보상을 계산하기 위해 측정된 결합 친화도의 공개 데이터베이스인 BindingDB에서 사전 훈련된 메시지 전달 및 컨볼루션 신경망으로 예측 모델을 활용했다.

【0105】 벤치마크 데이터셋

【0106】 이 연구에서는 표 1에 제시된 종래에 제공된 4개의 벤치마크 데이터셋과 소라페닙 내성에 대한 원본 데이터셋을 사용했다.

【0107】 【표 1】

		DRD2	QED	pLogP04	pLogP06	Sorafenib
Number of Unique Items	Triplets (Src,Tar,Neg)	688040	1766120	1973800	1495400	4612380
	Pairs (Src,Tar)	34402	88306	98690	74770	230619
	Src	18490	38723	57856	67718	13840
	Tar	3141	13202	44759	69762	2340
	Neg	21632	51923	99066	132397	16180
Range of Tanimoto Similarity	(Src,Tar)	0.40 – 0.83	0.40 – 0.80	0.40 – 1.00	0.60 – 1.00	0.40 – 1.00
	(Src,Neg)	0.00 – 0.30	0.00 – 0.30	0.00 – 0.30	0.00 – 0.49	0.03 – 0.30
	(Tar,Neg)	0.00 – 0.30	0.00 – 0.30	0.00 – 0.30	0.00 – 0.49	0.03 – 0.30
Range of Property	Src	0.00 – 0.05	0.70 – 0.80	-62.52 – 1.66	-32.33 – 3.89	4.90 – 8.37
	Tar	0.50 – 1.00	0.90 – 0.95	-42.76 – 4.17	-30.63 – 5.48	3.39 – 4.70
	Difference (Tar - Src)	0.45 – 1.00	0.10 – 0.25	1.00 – 64.36	1.00 – 23.79	N/A

【0108】DRD2 데이터셋에는 ZINC 데이터베이스에서 파생된 DRD2 활동 스코어와 함께 34,000여개의 분자 쌍(소스 및 타겟)이 포함되어 있다. DRD2 활동 스코어의 범위는 0에서 1까지이며 종래의 서포트 벡터 머신(Support Vector Machine) 회귀 모델을 사용하여 평가되었다. DRD2 데이터셋의 각 쌍에 대해 SMILES 문자열 쌍은 Tanimoto 유사도가 0.4 이상이고 소스 및 타겟 SMILES 문자열의 DRD2 스코어가 각각 0.05 미만 및 0.5보다 큰 속성 제약 조건을 충족했다. QED 데이터셋에는 QED 스코어가 있는 ZINC 데이터베이스에서 파생된 88,000여개의 분자 쌍이 포함되어 있

다. QED 스코어의 범위는 0에서 1까지이며 RDKit을 사용하여 계산되었다. QED 데이터셋의 각 쌍에 대해 두 SMILES 문자열 간의 Tanimoto 유사도는 0.4 이상이었고 소스 및 타겟의 QED 스코어는 각각 [0.7, 0.8] 및 [0.9, 1.0] 범위에 있었다. 페널티 logP04 및 페널티 logP06 데이터셋에는 각각 페널티 logP 스코어와 함께 ZINC 데이터베이스에서 파생된 98,000여개 및 74,000여개의 분자 쌍이 포함되어 있다. 페널티 logP 스코어 범위는 -63.0에서 5.5이다. 페널티가 적용된 logP04 데이터셋의 각 쌍에 대해 두 SMILES 문자열 간의 Tanimoto 유사도는 0.4 이상이었다. 페널티가 부여된 logP06의 경우 유사도 임계값이 0.6으로 설정되었다.

【0109】COMA 응용 사례를 소개하기 위해 소라페닙 유사 분자 생성을 위한 데이터셋을 구성했다. ABCG2의 활성이 간세포 암종에서 소라페닙 내성과 관련된다는 관찰에 기초하여, 이 적용은 ABCG2에 대해 더 낮은 결합 친화도를 갖는 소라페닙 유사 분자를 생성하는 동시에 표적 키나아제 BRAF에 대한 친화도 수준을 보존하는 것을 목표로 했다. 이 데이터셋 에는 ChEMBL 데이터베이스에서 파생된 23만여개의 분자 쌍이 ABCG2 및 BRAF에 대한 결합 친화도 스코어와 함께 포함되어 있다. DeepPurpose를 사용하여 평가한 결합 친화도 스코어는 pKd였다. ABCG2 데이터셋의 각 쌍에 대해 두 분자 간의 Tanimoto 유사도는 0.4 이상이었고 소스와 타겟의 ABCG2 친화도 값은 각각 [4.9, 8.4] 및 [3.3, 4.7] 범위였다. BRAF의 경우 소스와 대상 모두 결합 친화도가 6.0보다 컸다.

【0110】도 16은 본 개시의 일 실시예에 따른 분자 생성 모델을 위한 학습 방법(1600)의 흐름도이다. 방법(1600)은 컴퓨팅 장치의 적어도 하나의 프로세서

(예: 프로세서(414))에 의해 수행될 수 있다. 한편 도시된 바와 같이, 방법(1600)은 소스 분자 모델, 소스 분자 모델과의 구조적 유사도가 제1 임계치 초과인 타겟 분자 모델, 소스 분자 모델 또는 타겟 분자 모델 중 하나 이상의 모델과의 구조적 유사도가 제1 임계치 이하인 네거티브 분자 모델을 포함하는 학습 데이터셋을 획득하는 단계(S1610)로 개시될 수 있다.

【0111】 프로세서는 학습 데이터셋 및 제1 손실 함수를 기초로, 소스 분자 모델과 타겟 분자 모델 사이의 거리를 조정하도록 분자 생성 모델을 학습시킬 수 있다(S1620). 예를 들어, 프로세서는 학습 데이터셋 및 제1 손실 함수를 기초로, 소스 분자 모델과 타겟 분자 모델 사이의 거리가 가까워지도록 분자 생성 모델을 학습시킬 수 있다.

【0112】 프로세서는 학습 데이터셋 및 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 소스 분자 모델과 네거티브 분자 모델 사이의 거리 및 타겟 분자 모델과 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리를 조정하도록 분자 생성 모델을 학습시킬 수 있다(S1630). 예를 들어, 프로세서는 학습 데이터셋 및 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 소스 분자 모델과 네거티브 분자 모델 사이의 거리 및 타겟 분자 모델과 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리가 멀어지도록 분자 생성 모델을 학습시킬 수 있다.

【0113】 추가적으로, 프로세서는 학습 데이터셋 및 보상 함수를 기초로, 소스 분자 모델로부터 소스 분자 모델과의 구조적 유사도가 제2 임계치 초과인 분자 모델이 출력되도록, 분자 생성 모델을 학습시킬 수도 있다. 예를 들어, 프로세서

는 분자 생성 모델에 소스 분자 모델을 입력하여 출력 분자 모델을 획득하는 단계 및 출력 분자 모델과 소스 분자 모델을 비교한 결과 출력 분자 모델과 소스 분자 모델 사이의 구조적 유사도가 제2 임계치를 초과하는지 여부를 기초로, 출력 분자 모델과 연관된 포지티브 가중치 또는 네거티브 가중치를 산출하여 분자 생성 모델에 부여할 수 있다. 다른 예를 들어, 프로세서는 출력 분자 모델과 소스 분자 모델을 비교한 결과 출력 분자 모델과 소스 분자 모델 사이의 구조적 유사도가 제2 임계치를 초과하는지 여부 및 출력 분자 모델의 화학적 속성 스코어가 소스 분자 모델의 화학적 속성 스코어를 초과하는지 여부를 기초로, 출력 분자 모델과 연관된 포지티브 가중치 또는 네거티브 가중치를 산출하여 분자 생성 모델에 부여할 수도 있다. 또 다른 예를 들어, 프로세서는 학습 데이터셋 및 보상 함수를 기초로, 소스 분자 모델로부터, 소스 분자 모델과의 구조적 유사도가 제2 임계치 초과이고, 소스 분자 모델의 화학적 속성 스코어보다 큰 화학적 속성 스코어를 갖는 분자 모델이 출력되도록, 분자 생성 모델을 학습시킬 수 있다.

【0114】 한편, 분자 생성 모델은 타겟 분자 모델의 화학적 속성 스코어가 소스 분자 모델의 화학적 속성 스코어를 초과하도록 구성될 수 있다.

【0115】 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.

【0116】비록 예시적인 구현예들이 하나 이상의 독립형 컴퓨터 시스템의 맥락에서 현재 개시된 주제의 양태들을 활용하는 것을 언급할 수도 있으나, 본 주제는 그렇게 제한되지 않고, 오히려 네트워크나 분산 컴퓨팅 환경과 같은 임의의 컴퓨팅 환경과 연계하여 구현될 수도 있다. 또 나아가, 현재 개시된 주제의 양상들은 복수의 프로세싱 칩들이나 디바이스들에서 또는 그들에 걸쳐 구현될 수도 있고, 스토리지는 복수의 디바이스들에 걸쳐 유사하게 영향을 받게 될 수도 있다. 이러한 디바이스들은 PC들, 네트워크 서버들, 및 핸드헬드 디바이스들을 포함할 수도 있다.

【0117】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 특허청구의 범위 내에 속하는 것으로 생각되어야 한다.

【부호의 설명】

【0119】100: 분자 생성 모델

110: 소스 분자 모델

120: 출력 분자 모델

【청구범위】

【청구항 1】

소스 분자 모델, 상기 소스 분자 모델과의 구조적 유사도가 제1 임계치 초과인 타겟 분자 모델, 상기 소스 분자 모델 또는 상기 타겟 분자 모델 중 하나 이상의 모델과의 구조적 유사도가 상기 제1 임계치 이하인 네거티브 분자 모델을 포함하는 학습 데이터셋을 획득하는 단계;

상기 학습 데이터셋 및 제1 손실 함수를 기초로, 상기 소스 분자 모델과 상기 타겟 분자 모델 사이의 거리를 조정하도록 분자 생성 모델을 학습시키는 단계; 및

상기 학습 데이터셋 및 상기 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 상기 소스 분자 모델과 상기 네거티브 분자 모델 사이의 거리 및 상기 타겟 분자 모델과 상기 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리를 조정하도록 상기 분자 생성 모델을 학습시키는 단계

를 포함하는, 적어도 하나의 프로세서에 의해 수행되는 분자 생성 모델을 위한 학습 방법.

【청구항 2】

제1항에 있어서,

상기 학습 데이터셋 및 제1 손실 함수를 기초로, 상기 소스 분자 모델과 상기 타겟 분자 모델 사이의 거리를 조정하도록 분자 생성 모델을 학습시키는

단계는,

상기 학습 데이터셋 및 제1 손실 함수를 기초로, 상기 소스 분자 모델과 상기 타겟 분자 모델 사이의 거리가 가까워지도록 상기 분자 생성 모델을 학습시키는 단계

를 포함하는, 적어도 하나의 프로세서에 의해 수행되는 분자 생성 모델을 위한 학습 방법.

【청구항 3】

제1항에 있어서,

상기 학습 데이터셋 및 상기 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 상기 소스 분자 모델과 상기 네거티브 분자 모델 사이의 거리 및 상기 타겟 분자 모델과 상기 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리를 조정하도록 상기 분자 생성 모델을 학습시키는 단계는,

상기 학습 데이터셋 및 상기 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 상기 소스 분자 모델과 상기 네거티브 분자 모델 사이의 거리 및 상기 타겟 분자 모델과 상기 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리가 멀어지도록 상기 분자 생성 모델을 학습시키는 단계

를 포함하는, 적어도 하나의 프로세서에 의해 수행되는 분자 생성 모델을 위한 학습 방법.

【청구항 4】

제1항에 있어서,

상기 학습 데이터셋 및 보상 함수를 기초로, 상기 소스 분자 모델로부터 상기 소스 분자 모델과의 구조적 유사도가 제2 임계치 초과인 분자 모델이 출력되도록, 상기 분자 생성 모델을 학습시키는 단계

를 더 포함하는, 적어도 하나의 프로세서에 의해 수행되는 분자 생성 모델을 위한 학습 방법.

【청구항 5】

제4항에 있어서,

상기 학습 데이터셋 및 보상 함수를 기초로, 상기 소스 분자 모델로부터 상기 소스 분자 모델과의 구조적 유사도가 제2 임계치 초과인 분자 모델이 출력되도록, 상기 분자 생성 모델을 학습시키는 단계는,

상기 분자 생성 모델에 상기 소스 분자 모델을 입력하여 출력 분자 모델을 획득하는 단계; 및

상기 출력 분자 모델과 상기 소스 분자 모델을 비교한 결과 상기 출력 분자 모델과 상기 소스 분자 모델 사이의 구조적 유사도가 상기 제2 임계치를 초과하는지 여부를 기초로, 상기 출력 분자 모델과 연관된 포지티브 가중치 또는 네거티브

가중치를 산출하여 상기 분자 생성 모델에 부여하는 단계

를 포함하는, 적어도 하나의 프로세서에 의해 수행되는 분자 생성 모델을 위한 학습 방법.

【청구항 6】

제5항에 있어서,

상기 출력 분자 모델과 상기 소스 분자 모델을 비교한 결과 상기 출력 분자 모델과 상기 소스 분자 모델 사이의 구조적 유사도가 상기 제2 임계치를 초과하는지 여부를 기초로, 상기 출력 분자 모델과 연관된 포지티브 가중치 또는 네거티브 가중치를 산출하여 상기 분자 생성 모델에 부여하는 단계는,

상기 출력 분자 모델과 상기 소스 분자 모델을 비교한 결과 상기 출력 분자 모델과 상기 소스 분자 모델 사이의 구조적 유사도가 상기 제2 임계치를 초과하는지 여부 및 상기 출력 분자 모델의 화학적 속성 스코어가 상기 소스 분자 모델의 화학적 속성 스코어를 초과하는지 여부를 기초로, 상기 출력 분자 모델과 연관된 포지티브 가중치 또는 네거티브 가중치를 산출하여 상기 분자 생성 모델에 부여하는 단계

를 포함하는, 적어도 하나의 프로세서에 의해 수행되는 분자 생성 모델을 위한 학습 방법.

【청구항 7】

제4항에 있어서,

상기 학습 데이터셋 및 보상 함수를 기초로, 상기 소스 분자 모델로부터 상기 소스 분자 모델과의 구조적 유사도가 제2 임계치 초과인 분자 모델이 출력되도록, 상기 분자 생성 모델을 학습시키는 단계는,

상기 학습 데이터셋 및 상기 보상 함수를 기초로, 상기 소스 분자 모델로부터, 상기 소스 분자 모델과의 구조적 유사도가 상기 제2 임계치 초과이고, 상기 소스 분자 모델의 화학적 속성 스코어보다 큰 화학적 속성 스코어를 갖는 분자 모델이 출력되도록, 상기 분자 생성 모델을 학습시키는 단계

를 포함하는, 적어도 하나의 프로세서에 의해 수행되는 분자 생성 모델을 위한 학습 방법.

【청구항 8】

제1항에 있어서,

상기 타겟 분자 모델의 화학적 속성 스코어는 상기 소스 분자 모델의 화학적 속성 스코어보다 큰, 적어도 하나의 프로세서에 의해 수행되는 분자 생성 모델을 위한 학습 방법.

【청구항 9】

제1항 내지 8항 중 어느 한 항에 따른 분자 생성 모델을 위한 학습 방법을 실행시키도록 컴퓨터로 판독 가능한 기록매체에 기록된 컴퓨터 프로그램.

【청구항 10】

분자 생성 모델과 연관된 데이터를 저장하는 메모리; 및

상기 메모리와 연결되어 상기 분자 생성 모델을 학습시키는 적어도 하나의 프로세서

를 포함하고, 상기 적어도 하나의 프로세서는

소스 분자 모델, 상기 소스 분자 모델과의 구조적 유사도가 제1 임계치 초과인 타겟 분자 모델, 상기 소스 분자 모델 또는 상기 타겟 분자 모델 중 하나 이상의 모델과의 구조적 유사도가 상기 제1 임계치 이하인 네거티브 분자 모델을 포함하는 학습 데이터셋을 획득하는 것,

상기 학습 데이터셋 및 제1 손실 함수를 기초로, 상기 소스 분자 모델과 상기 타겟 분자 모델 사이의 거리를 조정하도록 분자 생성 모델을 학습시키는 것과

상기 학습 데이터셋 및 상기 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 상기 소스 분자 모델과 상기 네거티브 분자 모델 사이의 거리 및 상기 타겟 분자 모델과 상기 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리를 조정하도록 상기 분자 생성 모델을 학습시키는 것을 실행하도록 구성된 명령어들을 포함

하는, 분자 생성 모델을 위한 학습 장치.

【요약서】

【요약】

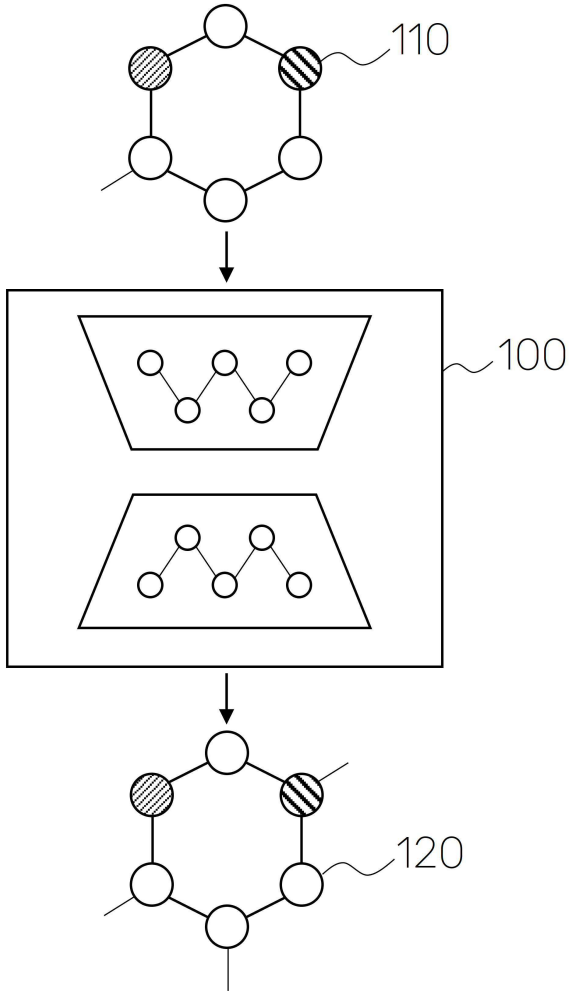
본 개시의 일 실시예에 따르면, 분자 생성 모델을 위한 학습 방법은 소스 분자 모델, 소스 분자 모델과의 구조적 유사도가 제1 임계치 초과인 타겟 분자 모델, 소스 분자 모델 또는 타겟 분자 모델 중 하나 이상의 모델과의 구조적 유사도가 제1 임계치 이하인 네거티브 분자 모델을 포함하는 학습 데이터셋을 획득하는 단계, 학습 데이터셋 및 제1 손실 함수를 기초로, 소스 분자 모델과 타겟 분자 모델 사이의 거리를 조정하도록 분자 생성 모델을 학습시키는 단계 및 학습 데이터셋 및 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 소스 분자 모델과 네거티브 분자 모델 사이의 거리 및 타겟 분자 모델과 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리를 조정하도록 분자 생성 모델을 학습시키는 단계를 포함할 수 있다.

【대표도】

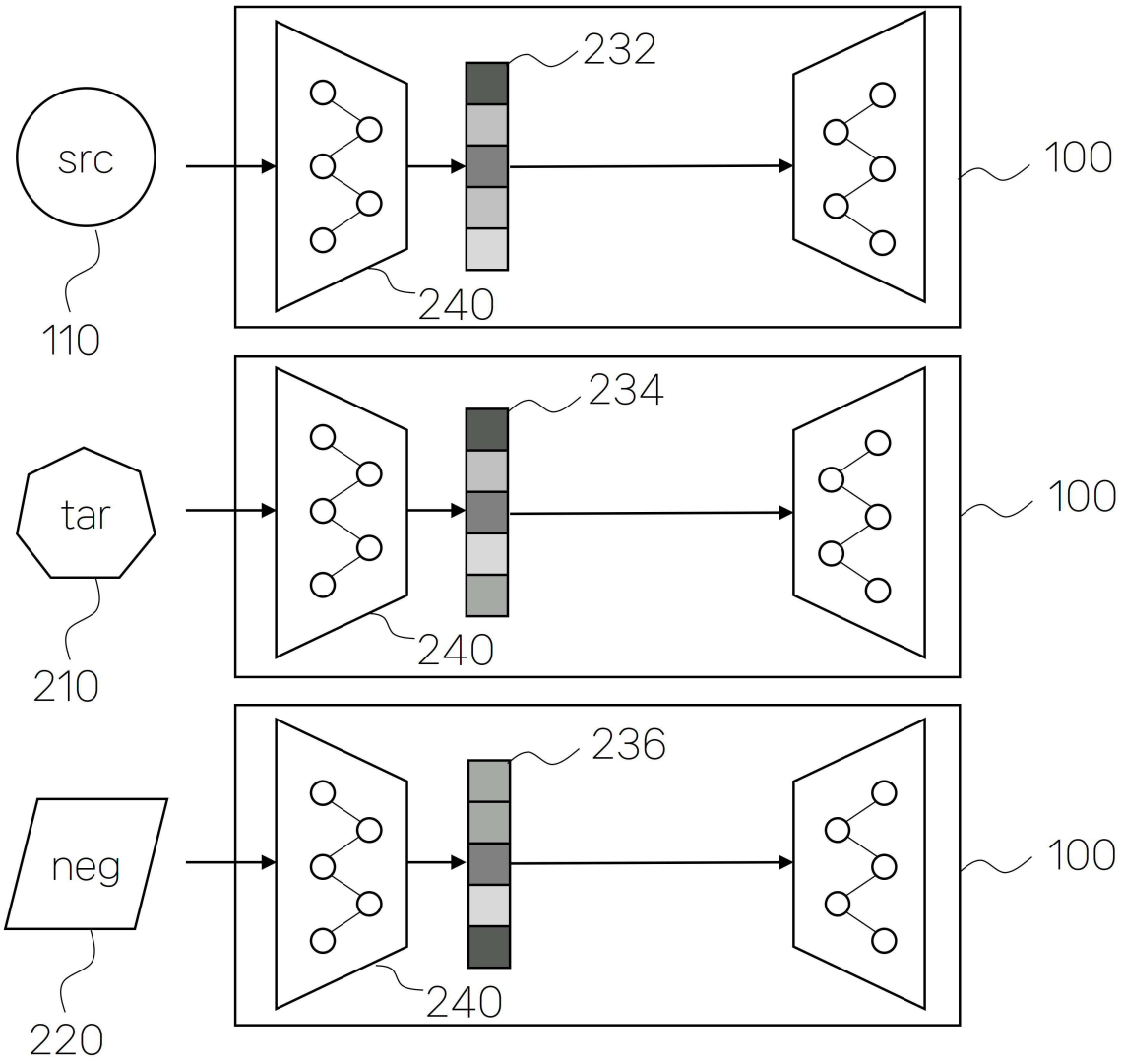
도 1

【도면】

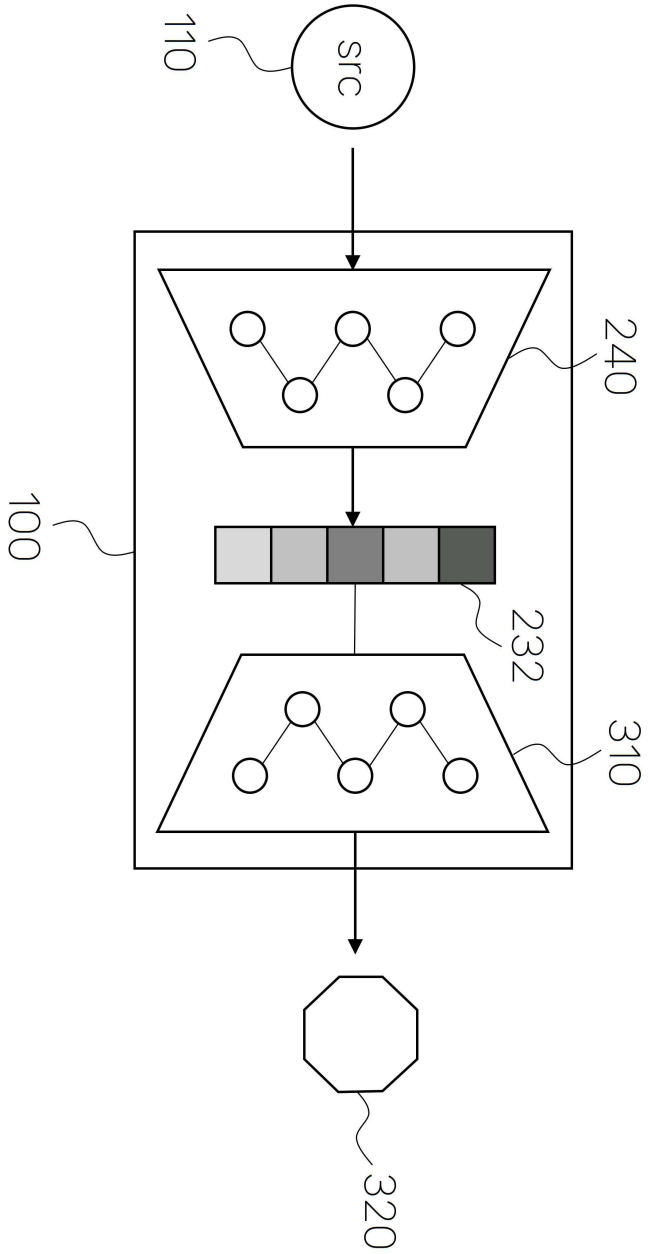
【도 1】



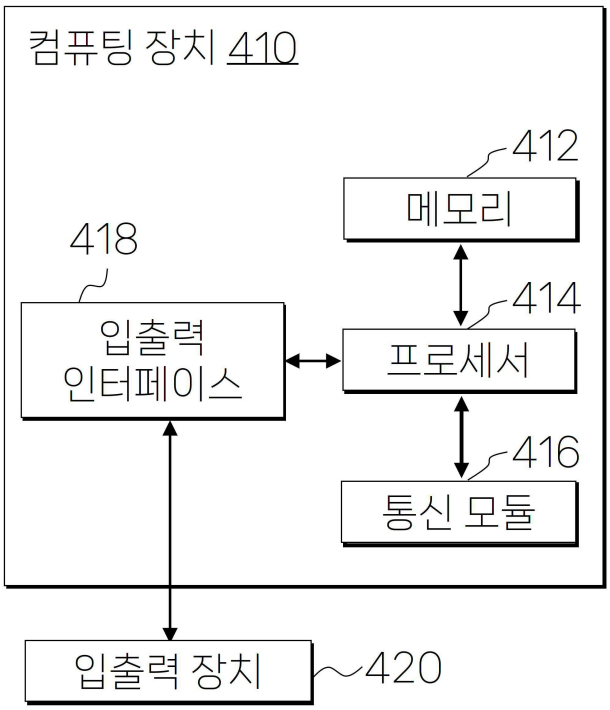
【도 2】



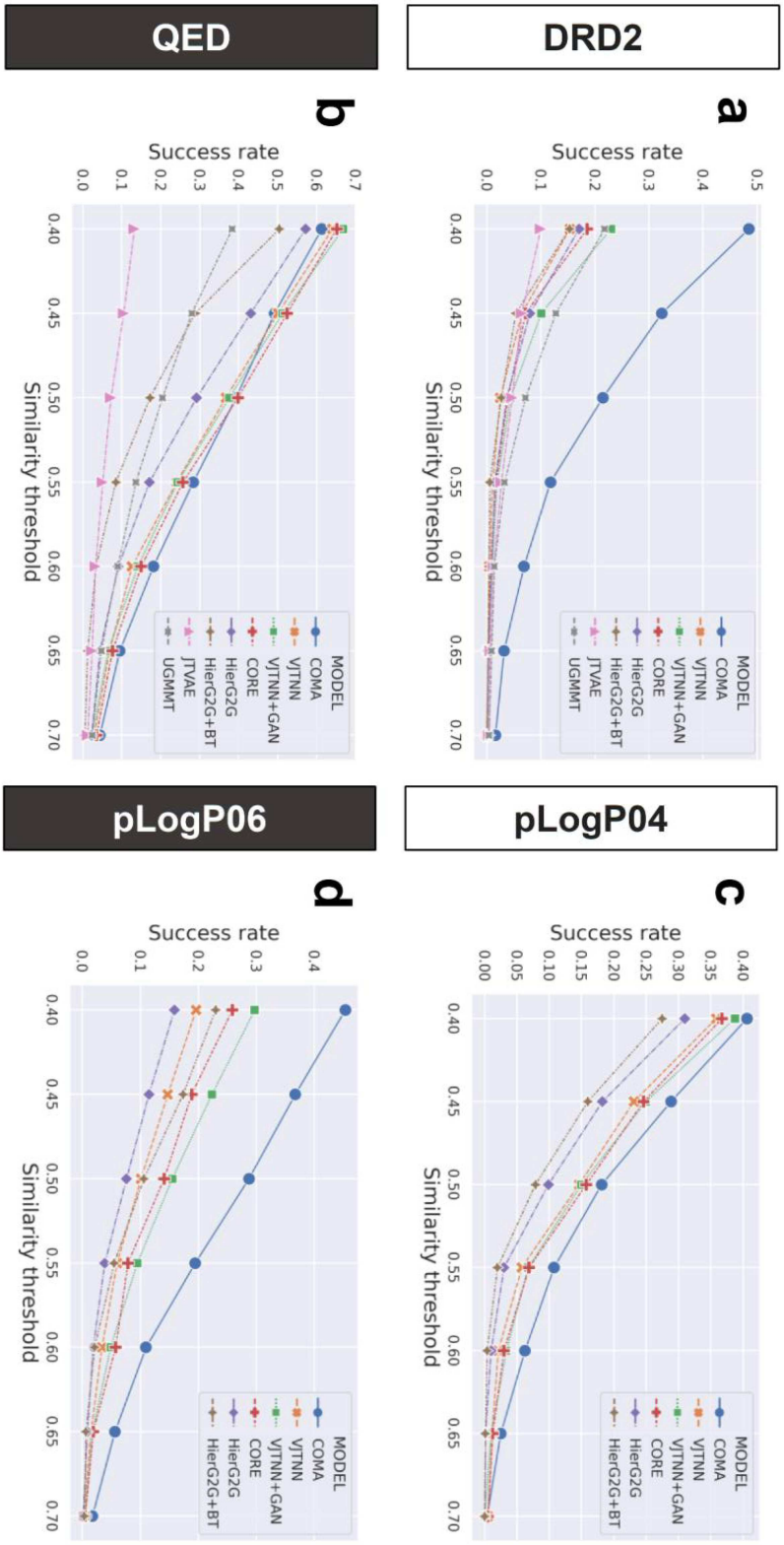
【도 3】



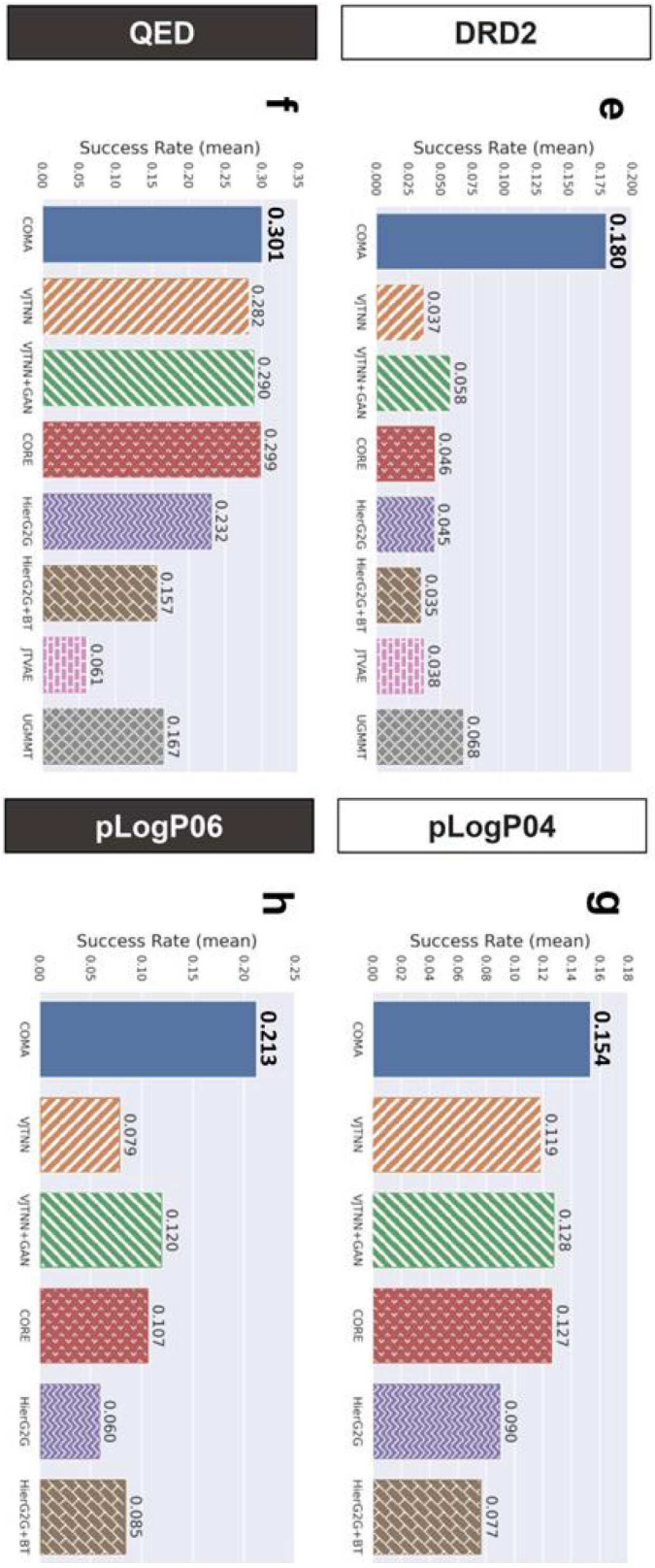
【도 4】



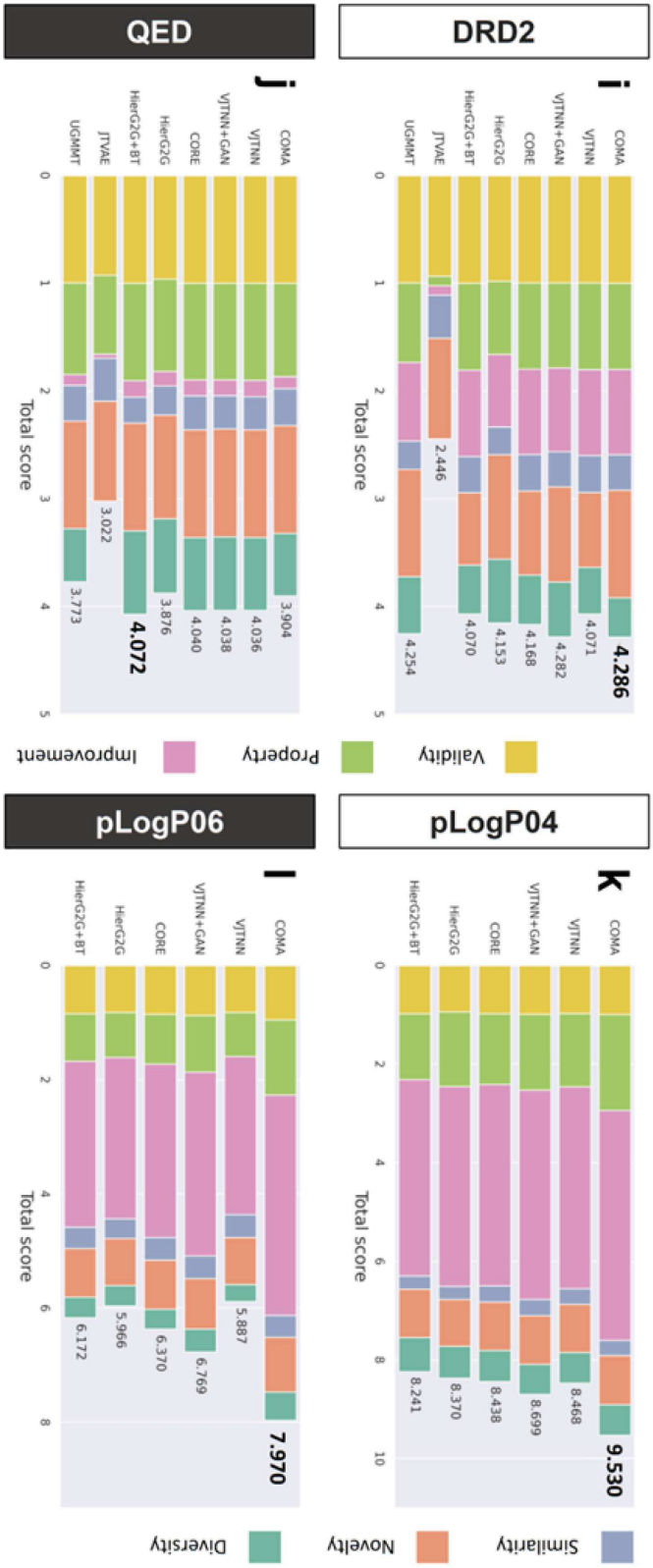
【도 5】


DRD2
QED
pLogP04
pLogP06
a
b
c
d
Success rate
Similarity threshold
MODEL
COMA
VJTNN
VJTNN+GAN
CORE
HierG2G
HierG2G+BT
JTVAE
UGMMT

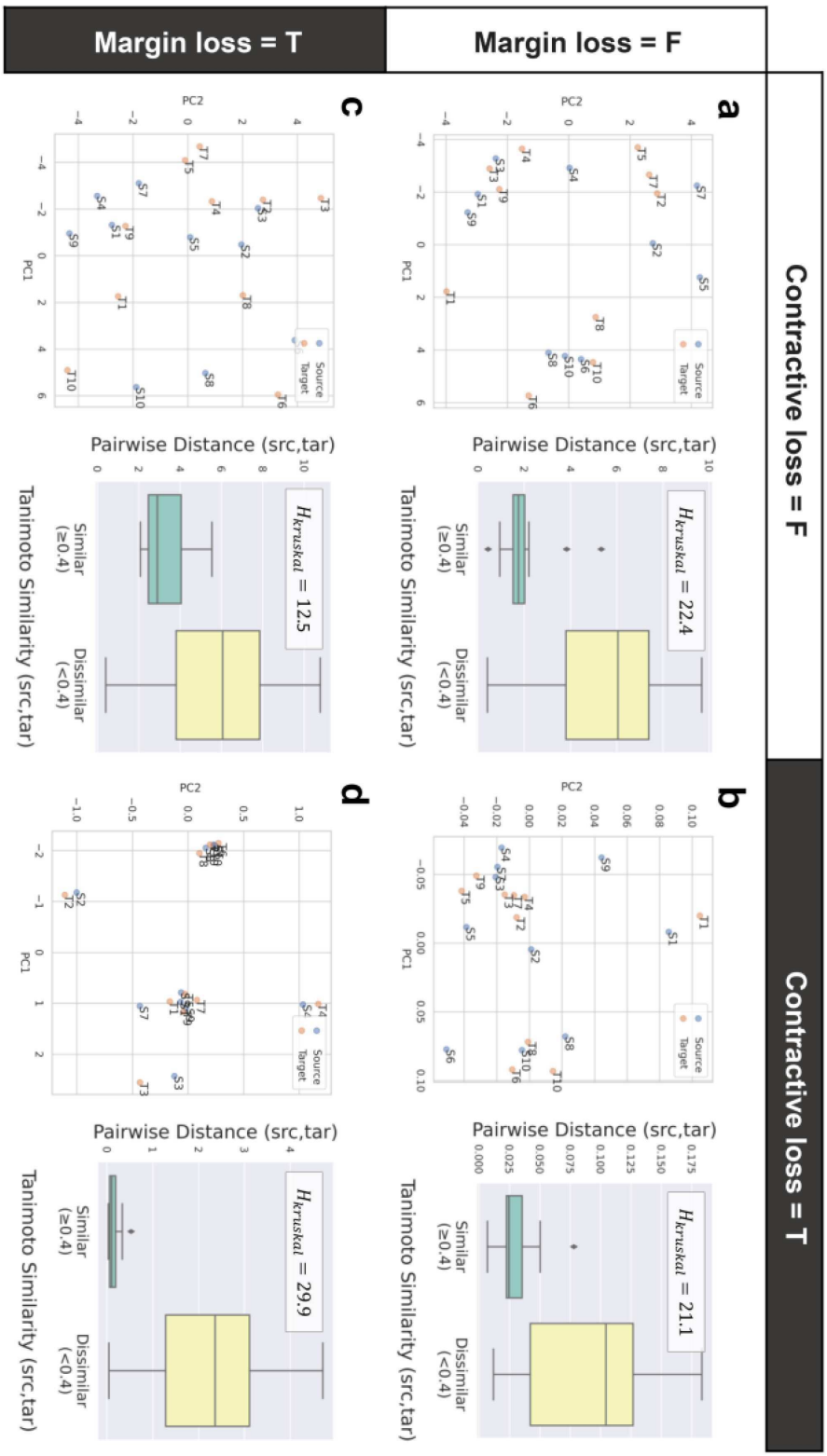
【도 6】


DRD2
QED
pLogP04
pLogP06
e
f
g
h
Success Rate (mean)
COMA
VJTNN
VJTNN+GAN
CORE
HierG2G
HierG2G+BT
JTVAE
UGMMT
0.180
0.037
0.058
0.046
0.045
0.035
0.038
0.068
0.301
0.282
0.290
0.299
0.232
0.157
0.061
0.167
0.154
0.119
0.128
0.127
0.090
0.077
0.213
0.079
0.120
0.107
0.060
0.085

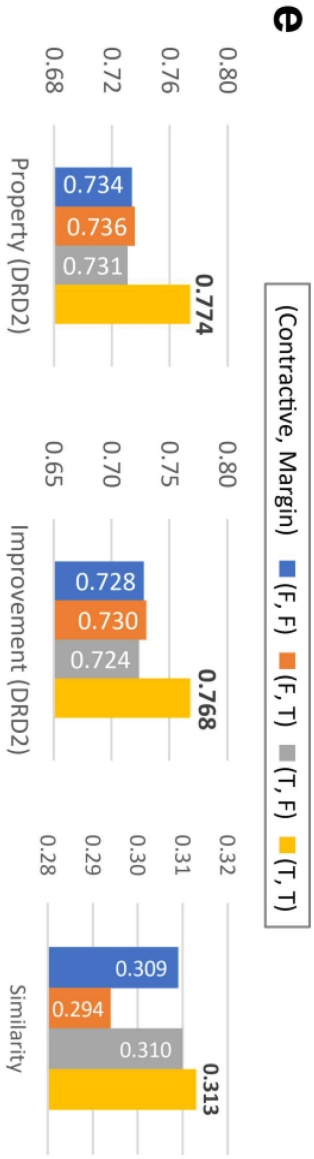
【도 7】


DRD2
QED
pLogP04
pLogP06
i
j
k
l
COMA
VJTNN
VJTNN+GAN
CORE
HierG2G
HierG2G+BT
JTVAE
UGMMT
Total score
4.286
4.071
4.282
4.168
4.153
4.070
2.446
4.254
3.904
4.036
4.038
4.040
3.876
4.072
3.022
3.773
9.530
8.468
8.699
8.438
8.370
8.241
7.970
5.887
6.769
6.370
5.966
6.172
Validity
Property
Improvement
Similarity
Novelty
Diversity

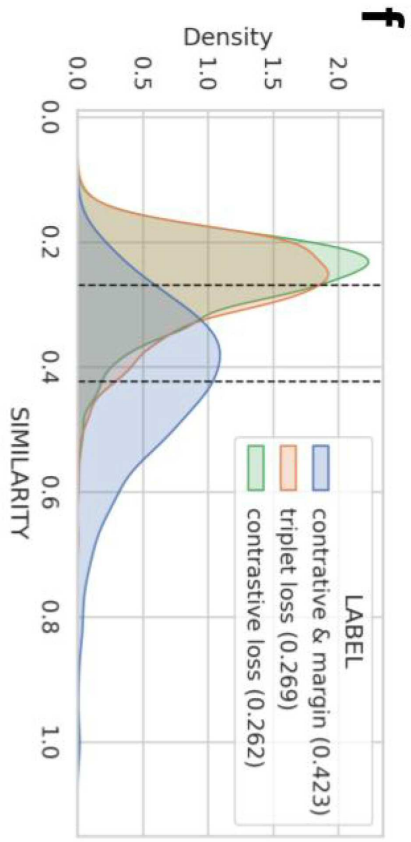
【도 8】


Margin loss = F
Margin loss = T
Contractive loss = F
Contractive loss = T
a
b
c
d
PC1
PC2
Source
Target
Pairwise Distance (src,tar)
Tanimoto Similarity (src,tar)
Similar (≥0.4)
Dissimilar (<0.4)
$H_{kruskal}$ = 22.4
$H_{kruskal}$ = 21.1
$H_{kruskal}$ = 12.5
$H_{kruskal}$ = 29.9

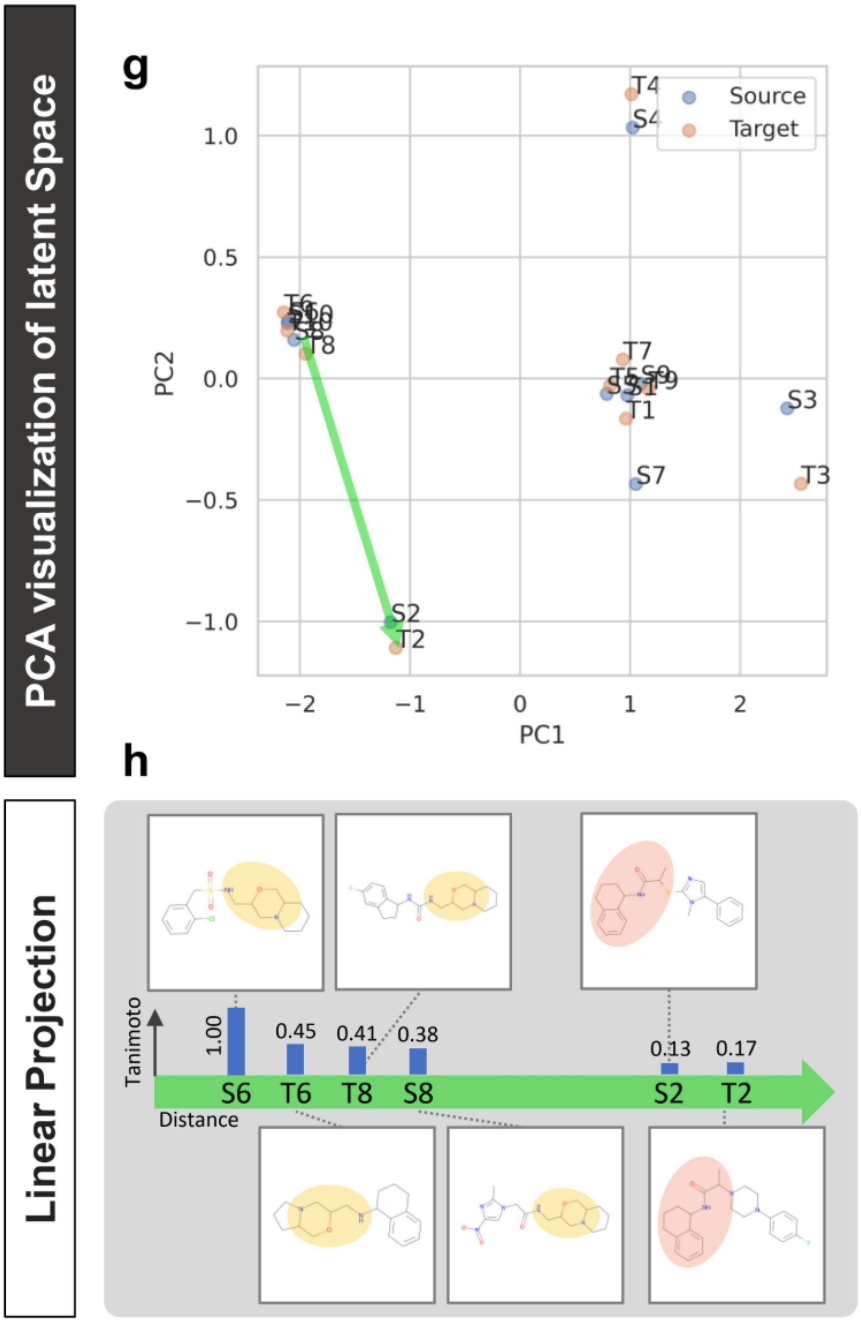
【도 9】


e
(Contractive, Margin)
(F, F)
(F, T)
(T, F)
(T, T)
Property (DRD2)
0.68
0.72
0.76
0.80
0.734
0.736
0.731
0.774
Improvement (DRD2)
0.65
0.70
0.75
0.80
0.728
0.730
0.724
0.768
Similarity
0.28
0.29
0.30
0.31
0.32
0.309
0.294
0.310
0.313

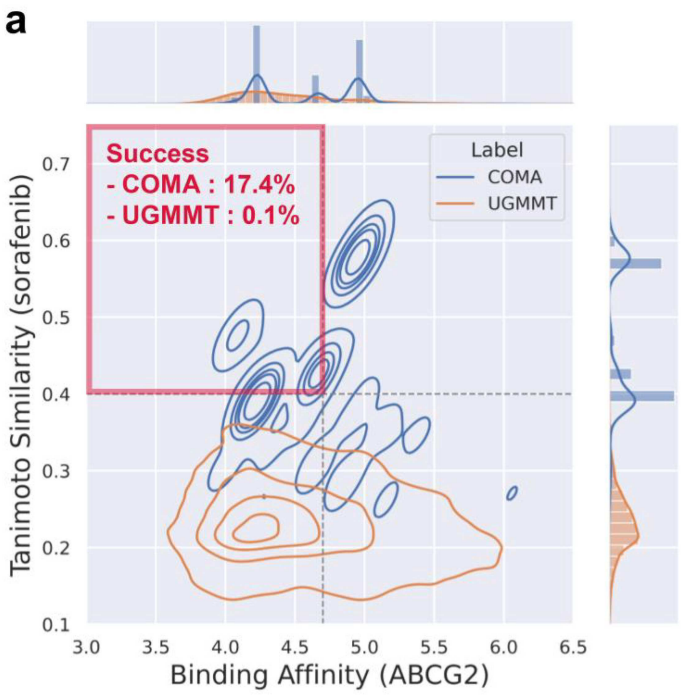
【도 10】


Density
0.0
0.5
1.0
1.5
2.0
SIMILARITY
0.0
0.2
0.4
0.6
0.8
1.0
LABEL
contrative & margin (0.423)
triplet loss (0.269)
contrastive loss (0.262)

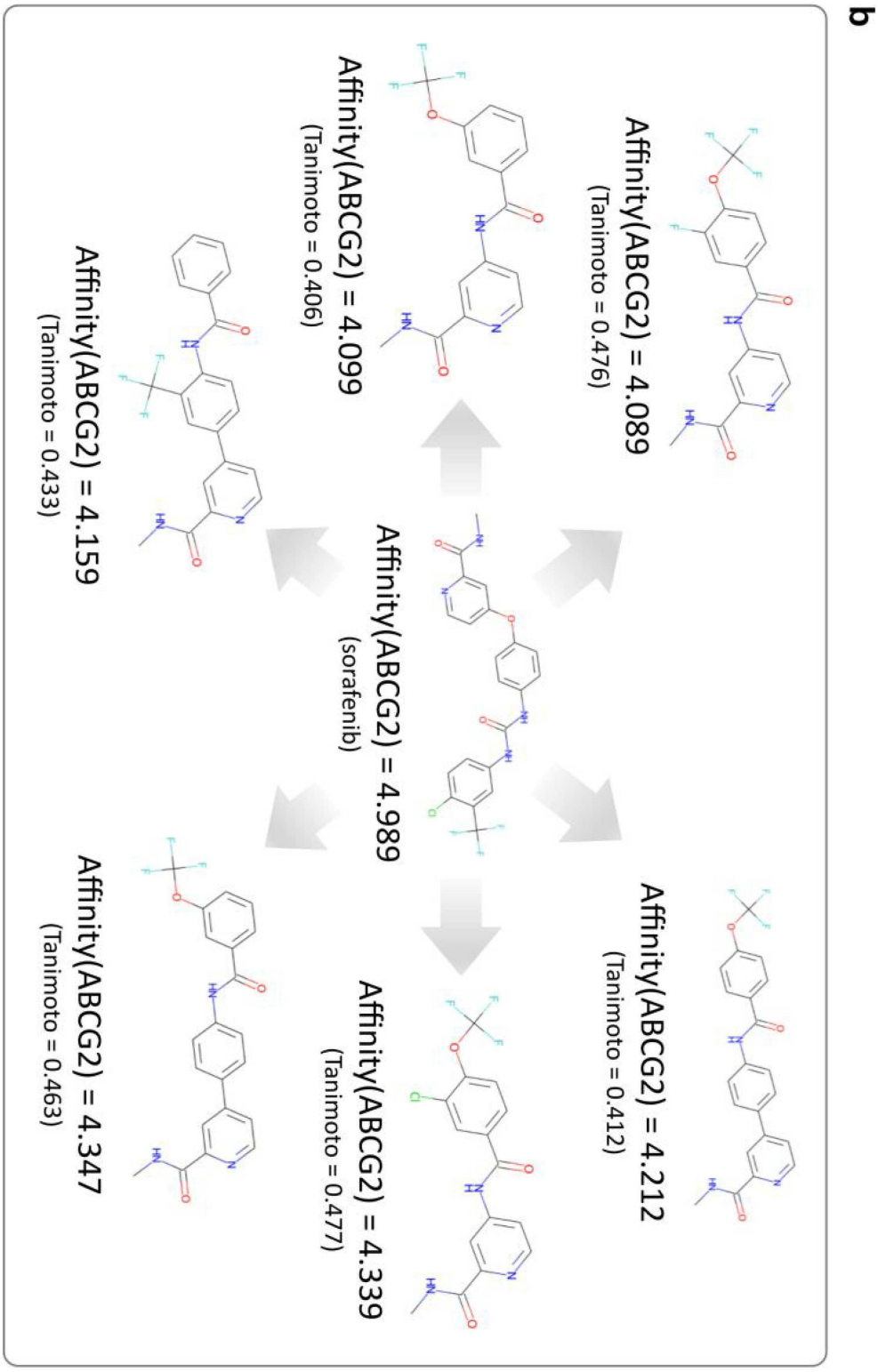
【도 11】


PCA visualization of latent Space
g
Source
Target
PC1
PC2
h
Linear Projection
Tanimoto
Distance
1.00
0.45
0.41
0.38
0.13
0.17
S6
T6
T8
S8
S2
T2

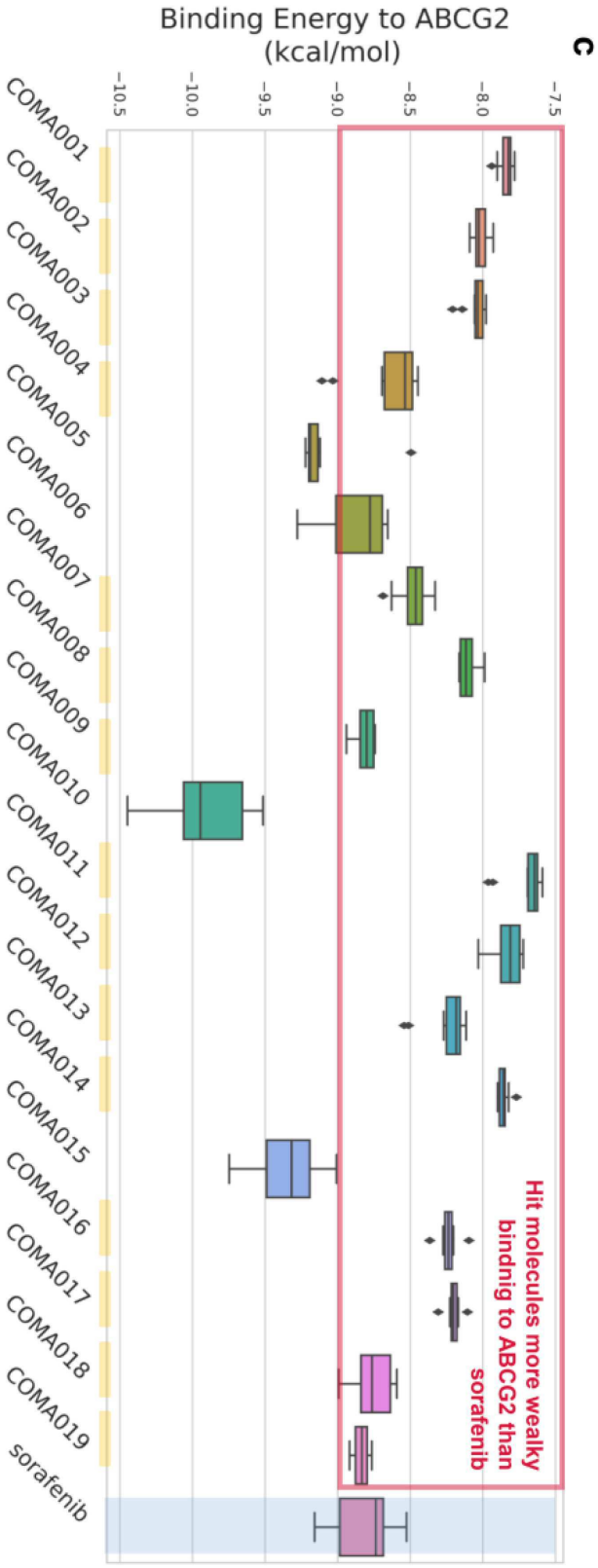
【도 12】


a
Success
- COMA : 17.4%
- UGMMT : 0.1%
Label
COMA
UGMMT
Tanimoto Similarity (sorafenib)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
Binding Affinity (ABCG2)

【도 13】

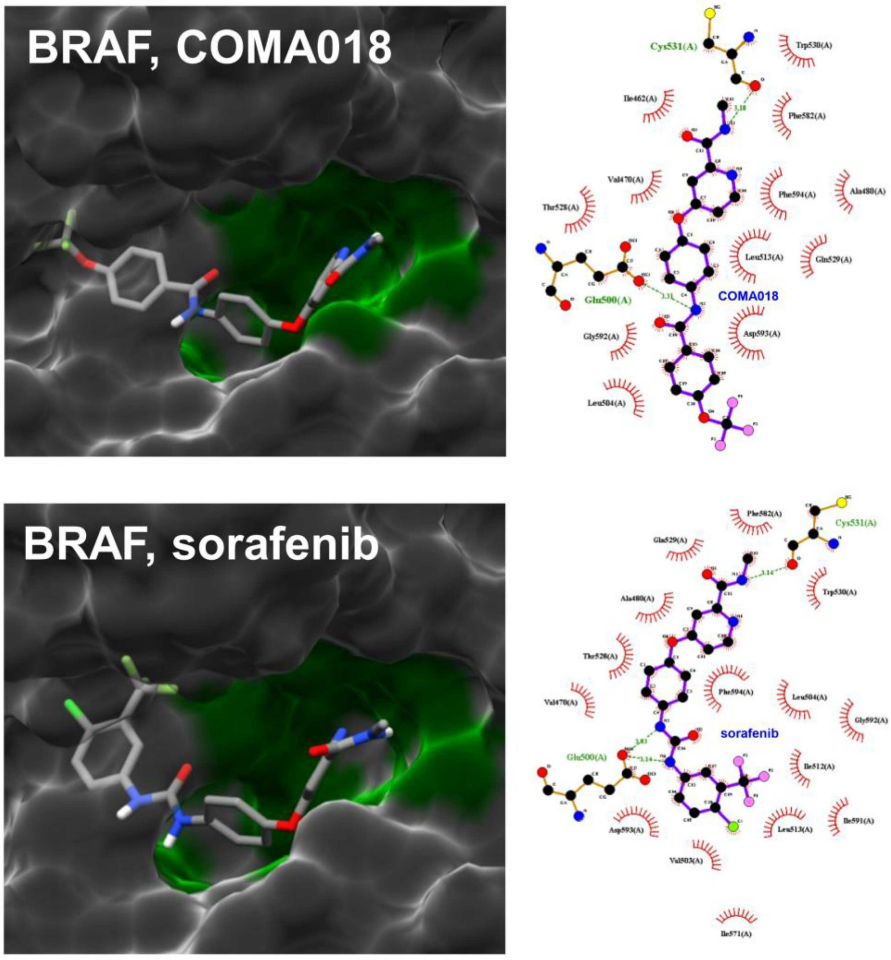

b
Affinity(ABCG2) = 4.089
(Tanimoto = 0.476)
Affinity(ABCG2) = 4.099
(Tanimoto = 0.406)
Affinity(ABCG2) = 4.989
(sorafenib)
Affinity(ABCG2) = 4.212
(Tanimoto = 0.412)
Affinity(ABCG2) = 4.339
(Tanimoto = 0.477)
Affinity(ABCG2) = 4.159
(Tanimoto = 0.433)
Affinity(ABCG2) = 4.347
(Tanimoto = 0.463)

【도 14】

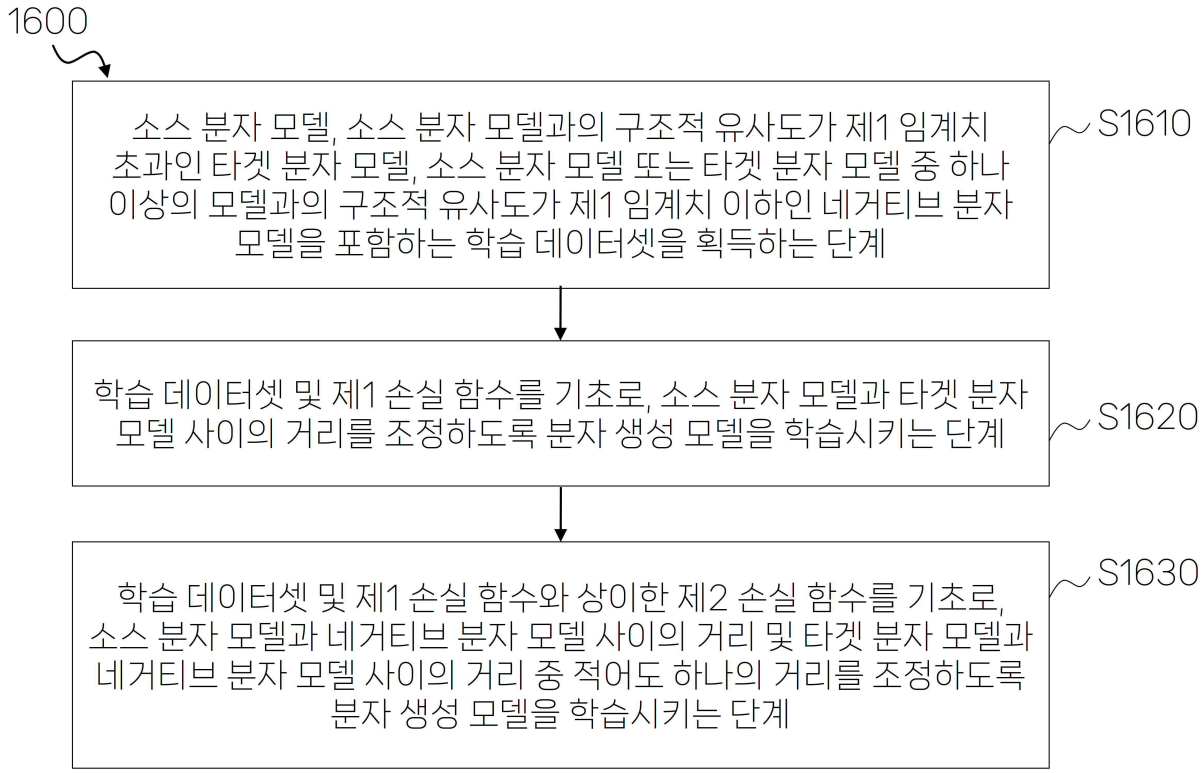

Binding Energy to ABCG2
(kcal/mol)
C
−10.5
−10.0
−9.5
−9.0
−8.5
−8.0
−7.5
COMA001
COMA002
COMA003
COMA004
COMA005
COMA006
COMA007
COMA008
COMA009
COMA010
COMA011
COMA012
COMA013
COMA014
COMA015
COMA016
COMA017
COMA018
COMA019
sorafenib
Hit molecules more weakly
bindnig to ABCG2 than
sorafenib

【도 15】

d



【도 16】


1600
소스 분자 모델, 소스 분자 모델과의 구조적 유사도가 제1 임계치 초과인 타겟 분자 모델, 소스 분자 모델 또는 타겟 분자 모델 중 하나 이상의 모델과의 구조적 유사도가 제1 임계치 이하인 네거티브 분자 모델을 포함하는 학습 데이터셋을 획득하는 단계
S1610
학습 데이터셋 및 제1 손실 함수를 기초로, 소스 분자 모델과 타겟 분자 모델 사이의 거리를 조정하도록 분자 생성 모델을 학습시키는 단계
S1620
학습 데이터셋 및 제1 손실 함수와 상이한 제2 손실 함수를 기초로, 소스 분자 모델과 네거티브 분자 모델 사이의 거리 및 타겟 분자 모델과 네거티브 분자 모델 사이의 거리 중 적어도 하나의 거리를 조정하도록 분자 생성 모델을 학습시키는 단계
S1630