(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0046481
(43) 공개일자 2024년04월09일

(51) 국제특허분류(Int. Cl.)
G16B 5/00 (2019.01) *G16B 15/30* (2019.01)
G16B 25/10 (2019.01) *G16B 40/20* (2019.01)
(52) CPC특허분류
G16B 5/00 (2019.02)
G16B 15/30 (2019.02)
(21) 출원번호 10-2024-7000495
(22) 출원일자(국제) 2022년06월15일
심사청구일자 없음
(85) 번역문제출일자 2024년01월05일
(86) 국제출원번호 PCT/US2022/033685
(87) 국제공개번호 WO 2022/266259
국제공개일자 2022년12월22일
(30) 우선권주장
63/210,930 2021년06월15일 미국(US)

(71) 출원인
플래그쉽 파이어니어링 이노베이션스 브이아이, 엘엘씨
미국 02142 매사추세츠주 캠브리지 55 캠브리지 파크웨이 8 플로어
(72) 발명자
울프, 파비안 알렉산더
미국 02142 매사추세츠 케임브리지 케임브리지 파크웨이 55 8 플로어 플래그쉽 파이어니어링 이노베이션스 브이아이 엘엘씨
하다드, 래기
미국 02142 매사추세츠 케임브리지 케임브리지 파크웨이 55 8 플로어 플래그쉽 파이어니어링 이노베이션스 브이아이 엘엘씨
플러기스, 니콜라스 매카트니
미국 02142 매사추세츠 케임브리지 케임브리지 파크웨이 55 8 플로어 플래그쉽 파이어니어링 이노베이션스 브이아이 엘엘씨
(74) 대리인
특허법인(유)남아이피그룹

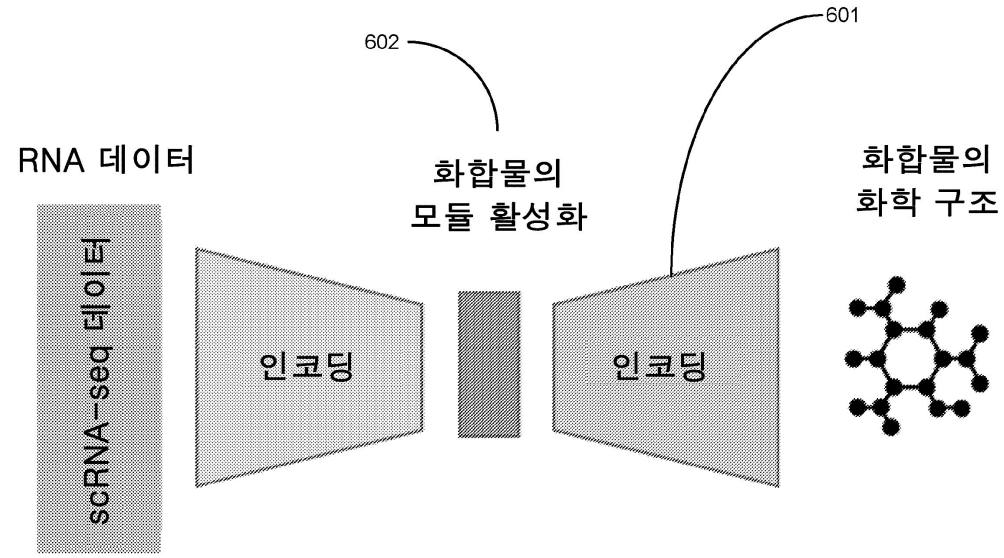
전체 청구항 수 : 총 159 항

(54) 발명의 명칭 **지문 분석을 이용하여 화합물을 생리학적 조건과 연관시키는 시스템 및 방법**

(57) 요 약

화합물을 생리학적 조건과 연관시키기 위한 시스템 및 방법이 제공된다. 화합물 화학 구조의 지문을 얻고, 하나 이상의 계산된 활성화 점수를 출력하는 모델에 입력한다. 각각의 활성화 점수는 모듈의 세트의 세포 구성성분 모듈을 나타내고, 여기서 각각의 모듈은 세포 구성성분의 서브세트를 포함하고, 모듈의 세트의 제1 모듈은 생리학적 조건과 연관된다. 제1 모듈에 대한 활성화 점수가 임계치 기준을 충족하는 경우에, 화합물은 생리학적 조건과 연관된 것으로 식별된다. 일부 양태에서, 각각의 활성화 점수는 생리학적 조건과 연관된 교란 시그니처를 나타내고, 화합물은 제1 교란 시그니처에 대한 활성화 점수가 임계치 기준을 충족하는 경우에 식별된다. 화합물을 생리학적 조건과 연관시키는 모델을 훈련시키기 위한 시스템 및 방법이 또한 제공된다.

대 표 도



(52) CPC특허분류
G16B 25/10 (2019.02)
G16B 40/20 (2019.02)

명 세 서

청구범위

청구항 1

테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법으로서, 상기 방법은,

(A) 상기 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계;

(B) 세포 구성성분 모듈의 세트에 액세스하는 단계로서,

상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 각각의 독립적인 서브세트를 포함하고,

상기 복수의 세포 구성성분의 각각의 개별 독립적인 서브세트에 대한 상응하는 복수의 세포-기반 검정 풍부도 값은 상기 생리학적 조건과 연관된 복수의 상이한 상태에 걸쳐 개별적으로 상관되고,

상기 세포 구성성분 모듈의 세트의 제1 세포 구성성분 모듈은 상기 관심 생리학적 조건과 연관되는, 상기 액세스하는 단계;

(C) 모델 내로 상기 화학 구조의 지문을 입력하는 것에 응답하여(상기 모델은 100개 이상의 파라미터를 포함함), 상기 모델로부터의 출력으로서, 상기 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈에 대한 각각의 활성화 점수를 검색하는 단계; 및

(D) 상기 제1 세포 구성성분 모듈에 대한 상기 활성화 점수가 제1 임계치 기준을 충족하는 경우에, 상기 테스트 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 단계를 포함하는, 방법.

청구항 2

제1항에 있어서, 상기 세포-기반 검정 풍부도 값은 기관(organ)의 세포의 값인, 방법.

청구항 3

제2항에 있어서, 상기 기관은 심장, 간, 폐, 근육, 뇌, 췌장, 비장, 신장, 소장, 자궁 또는 방광인, 방법.

청구항 4

제1항에 있어서, 상기 세포-기반 검정 풍부도 값은 조직의 세포의 값인, 방법.

청구항 5

제4항에 있어서, 상기 조직은 골, 연골, 관절, 기관(tracheae), 척수, 각막, 눈, 피부 또는 혈관인, 방법.

청구항 6

제1항에 있어서, 상기 세포-기반 검정 풍부도 값은 복수의 줄기 세포의 세포의 값인, 방법.

청구항 7

제6항에 있어서, 상기 복수의 줄기 세포는 복수의 배아 줄기 세포, 복수의 성체 줄기 세포, 또는 복수의 유도된 만능 줄기 세포(iPSC)인, 방법.

청구항 8

제1항에 있어서, 상기 세포-기반 검정 풍부도 값은 복수의 1차 인간 세포의 세포의 값인, 방법.

청구항 9

제8항에 있어서, 상기 복수의 1차 인간 세포는 복수의 CD34+ 세포, 복수의 CD34+ 조혈 줄기, 복수의 전구 세포(HSPC), 복수의 T-세포, 복수의 중간엽 줄기 세포(MSC), 복수의 기도 기저 줄기 세포 또는 복수의 유도된 만능

줄기 세포인, 방법.

청구항 10

제1항에 있어서, 상기 세포-기반 검정 풍부도 값은 제대혈, 말초혈 또는 골수 내의 세포의 값인, 방법.

청구항 11

제1항에 있어서, 상기 세포-기반 검정 풍부도 값은 고형 조직 내 세포의 값인, 방법.

청구항 12

제11항에 있어서, 상기 고형 조직은 태반, 간, 심장, 뇌, 신장 또는 위장관인, 방법.

청구항 13

제1항에 있어서, 상기 세포-기반 검정 풍부도 값은 복수의 분화된 세포의 값인, 방법.

청구항 14

제13항에 있어서, 상기 복수의 분화된 세포는 복수의 거핵구, 복수의 골모세포, 복수의 연골세포, 복수의 지방세포, 복수의 간세포, 복수의 간 중피 세포, 복수의 담관 상피 세포, 복수의 간 성상 세포, 복수의 간 시누소이드 내피 세포, 복수의 쿠퍼 세포, 복수의 피트 세포, 복수의 혈관 내피 세포, 복수의 췌장관 상피 세포, 복수의 췌장관 세포, 복수의 중심성 세포, 복수의 선방 세포, 복수의 랑게르한스섬, 복수의 심장 근육 세포, 복수의 섬유모세포, 복수의 각질세포, 복수의 평활근 세포, 복수의 제I형 폐포 상피 세포, 복수의 제II형 폐포 상피 세포, 복수의 클라라 세포, 복수의 섬모 상피 세포, 복수의 기저 세포, 복수의 배상 세포, 복수의 신경내분비 세포, 복수의 쿨치츠키(kultschitzky) 세포, 복수의 신세관 상피 세포, 복수의 요로상피 세포, 복수의 원주 상피 세포, 복수의 사구체 상피 세포, 복수의 사구체 내피 세포, 복수의 발세포, 복수의 사구체간 세포, 복수의 신경 세포, 복수의 교세포, 복수의 소교세포, 또는 복수의 핍지교세포인, 방법.

청구항 15

제1항 내지 제14항 중 어느 한 항에 있어서, 상기 상응하는 복수의 세포-기반 검정 풍부도 값은 복수의 세포의 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq) 데이터인, 방법.

청구항 16

제15항에 있어서, 상기 생리학적 조건과 연관된 상기 복수의 상이한 상태는 세포의 분취물이 상기 생리학적 조건에 영향을 미치는 것으로 공지된 화합물에 대한 노출에 대해 자유롭지 않은 대조군 상태에 더하여, 상이한 세포 분취물을 상기 생리학적 조건에 영향을 미치는 것으로 공지된 1종 이상의 참조 화합물에 노출시킴으로써 유도되는, 방법.

청구항 17

제1항 내지 제14항 중 어느 한 항에 있어서, 상기 상응하는 복수의 세포-기반 검정 풍부도 값은 벌크 RNA 시퀀스로부터의 것인, 방법.

청구항 18

제1항 내지 제14항 중 어느 한 항에 있어서, 상기 상응하는 복수의 세포-기반 검정 풍부도 값은 단일 세포 RNA 시퀀싱으로부터의 것인, 방법.

청구항 19

제1항 내지 제18항 중 어느 한 항에 있어서, 상기 세포 구성성분 모듈의 세트는 상기 제1 세포 구성성분 모듈로 이루어지는, 방법.

청구항 20

제1항 내지 제18항 중 어느 한 항에 있어서, 상기 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈을

포함하고, 상기 모델은 복수의 컴포넌트 모델을 포함하는 앙상블 모델이고, 상기 복수의 컴포넌트 모델의 각각의 컴포넌트 모델이 상기 복수의 컴포넌트 모델의 각각의 컴포넌트 모델에 상기 화학 구조의 지문을 입력하는 것에 응답하여 상기 세포 구성성분 모듈의 세트에서의 상이한 세포 구성성분 모듈에 대한 활성화 점수를 제공하는, 방법.

청구항 21

제1항 내지 제20항 중 어느 한 항에 있어서, 상기 방법은 상기 테스트 화학적 화합물의 단순화된 분자-입력 라인-엔트리 시스템(SMILES) 스트링 표현으로부터 상기 지문을 계산하는 단계를 더 포함하는, 방법.

청구항 22

제20항 또는 제21항에 있어서, 상기 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 상응하는 신경망인, 방법.

청구항 23

제22항에 있어서, 상기 상응하는 신경망이 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합인, 방법.

청구항 24

제20항 또는 제21항에 있어서, 상기 복수의 컴포넌트 모델의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델인, 방법.

청구항 25

제22항에 있어서,

상기 상응하는 신경망은 상응하는 완전 연결 신경망과 상응하는 메시지 전달 신경망의 조합이고,

상기 상응하는 완전 연결 신경망의 제1 출력 및 상기 상응하는 메시지 전달 신경망의 제2 출력은 조합되어, 상기 화학 구조의 지문을 상기 상응하는 완전 연결 신경망 및 상기 상응하는 메시지 전달 신경망에 입력하는 것에 응답하여, 상기 세포 구성성분 모듈의 세트의 상기 상응하는 세포 구성성분 모듈에 대한 하나 이상의 계산된 활성화 점수의 활성화 점수를 결정하는, 방법.

청구항 26

제1항에 있어서,

상기 세포 구성성분 모듈의 세트가 복수의 세포 구성성분 모듈이고,

상기 제1 세포 구성성분 모듈을 포함하는 상기 복수의 세포 구성성분 모듈의 제1 서브세트는 상기 관심 생리학적 조건과 연관되고,

상기 복수의 세포 구성성분 모듈의 제2 서브세트는 상기 관심 생리학적 조건과 연관되지 않고,

상기 제1 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 상기 제1 임계치 기준을 충족하고, 상기 복수의 세포 구성성분 모듈의 제2 서브세트에서의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 상기 제1 임계치 기준 이외의 제2 임계치 기준을 충족하는 경우에, 상기 테스트 화학적 화합물이 상기 관심 생리학적 조건으로 식별되는, 방법.

청구항 27

제1항 내지 제26항 중 어느 한 항에 있어서, 과정에 의해 상기 제1 세포 구성성분 모듈을 식별하는 단계를 더 포함하고, 상기 과정은,

전자 형태로 하나 이상의 제1 데이터세트를 획득하는 단계로서, 상기 하나 이상의 제1 데이터세트는,

제1 복수의 세포의 각각의 개별 세포에 대해(상기 제1 복수의 세포는 20개 이상의 세포를 포함하고, 집합적으로 복수의 주석화된 세포 상태를 나타냄),

상기 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해(상기 복수의 세포 구성성분은 10개 이상의 세

포 구성성분을 포함함),

상기 각각의 세포에서 상기 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하고,

이에 의해, 복수의 벡터에 액세스하거나 복수의 벡터를 형성하며, 상기 복수의 벡터의 각각의 개별 벡터는 (i) 상기 복수의 구성성분의 각각의 세포 구성성분에 상응하고, (ii) 상응하는 복수의 요소를 포함하고, 상기 상응하는 복수의 요소의 각각의 개별 요소는 상기 제1 복수의 세포의 상기 각각의 세포의 상기 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 가지는, 상기 획득하는 단계;

상기 복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계로서, 상기 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 상기 복수의 세포 구성성분의 서브세트를 포함하고, 상기 복수의 세포 구성성분 모듈은 (i) 상기 복수의 후보 세포 구성성분 모듈 및 (ii) 상기 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현으로 배열되고, 상기 복수의 세포 구성성분 모듈이 10개를 초과하는 세포 구성성분 모듈을 포함하는, 상기 식별하는 단계;

전자 형태로 하나 이상의 제2 데이터세트를 획득하는 단계로서, 상기 하나 이상의 제2 데이터세트는,

제2 복수의 세포의 각각의 개별 세포에 대해(상기 제2 복수의 세포는 20개 이상의 세포를 포함하고, 상기 관심 생리학적 조건을 알리는 복수의 공변량을 집합적으로 나타냄),

상기 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해,

상기 각각의 세포에서 상기 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하고,

이에 의해, (i) 상기 제2 복수의 세포 및 (ii) 상기 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 세포 구성성분 카운트 데이터 구조를 획득하는, 상기 획득하는 단계;

상기 복수의 세포 구성성분 또는 그 표현을 공통 차원으로서 사용하여 상기 세포 구성성분 카운트 데이터 구조 및 상기 잠재 표현을 조합함으로써 활성화 데이터 구조를 형성하는 단계로서, 상기 활성화 데이터 구조는 상기 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈에 대해,

상기 제2 복수의 세포의 각각의 세포에 대해, 각각의 활성화 가중치를 포함하는, 상기 형성하는 단계; 및

상기 복수의 공변량에서의 각각의 개별 공변량에 대해, (i) 후보 세포 구성성분 모델로의 상기 공변량의 지문의 입력시 상기 후보 세포 구성성분 모델에 의해 표현되는 각각의 세포 구성성분 모듈에 대한 계산된 활성화와 (ii) 상기 후보 세포 구성성분 모델에 의해 표현되는 각각의 세포 구성성분 모듈에 대한 실제 활성화 사이의 차이를 사용하여 상기 후보 세포 구성성분 모델을 훈련시키는 단계로서, 상기 훈련은 상기 차이에 응답하여 상기 후보 세포 구성성분 모델과 연관된 복수의 공변량 파라미터를 조정하는, 상기 훈련시키는 단계를 포함하는, 방법.

청구항 28

제27항에 있어서, 상기 복수의 공변량 파라미터는,

상기 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대해,

각각의 개별 공변량에 대해,

상기 각각의 공변량이 상기 제2 복수의 세포에 걸쳐 상기 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 상응하는 파라미터를 포함하고; 상기 방법은,

상기 후보 세포 구성성분 모델의 훈련시 상기 복수의 공변량 파라미터를 사용하여, 상기 복수의 후보 세포 구성성분 모듈에서 상기 제1 세포 구성성분 모듈을 식별하는 단계를 더 포함하는, 방법.

청구항 29

제27항 또는 제28항에 있어서, 상기 복수의 주석화된 세포 상태의 주석화된 세포 상태는 노출 조건 하에 화합물에 대한 상기 제1 복수의 세포의 세포의 노출인, 방법.

청구항 30

제29항에 있어서, 상기 노출 조건은 노출 지속기간, 화합물의 농도, 또는 노출 지속기간과 상기 화합물의 농도

의 조합인, 방법.

청구항 31

제1항 내지 제30항 중 어느 한 항에 있어서, 상기 복수의 세포 구성성분의 각각의 세포 구성성분이 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합인, 방법.

청구항 32

제27항 내지 제30항 중 어느 한 항에 있어서,

상기 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합이고,

상기 제1 또는 제2 복수의 세포에서의 상기 각각의 세포에서의 상기 각각의 세포 구성성분의 상기 상응하는 풍부도가 비색 측정치, 형광 측정치, 발광 측정치 또는 공명 에너지 전달(FRET) 측정치에 의해 결정되는, 방법

청구항 33

제11항에 있어서,

상기 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합이고,

상기 제1 또는 제2 복수의 세포에서의 상기 각각의 세포에서의 상기 각각의 세포 구성성분의 상응하는 풍부도는 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq), scTag-seq, 시퀀싱을 사용한 전위효소-접근 가능 염색질에 대한 단일-세포 검정(scATAC-seq), CyTOF/SCoP, E-MS/Abseq, miRNA-seq, CITE-seq, 또는 그 임의의 조합에 의해 결정되는, 방법.

청구항 34

제1항 내지 제30항, 제32항 또는 제33항 중 어느 한 항에 있어서, 상기 복수의 벡터를 사용하여 상기 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계는 상기 복수의 벡터의 각각의 벡터의 각각의 상응하는 복수의 요소를 사용하여 상관 모델을 상기 복수의 벡터에 적용하는 것을 포함하는, 방법.

청구항 35

제34항에 있어서, 상기 상관 모델은 그래프 클러스터링을 포함하는, 방법.

청구항 36

제34항에 있어서, 상기 그래프 클러스터링은 피어슨-상관관계-기반 거리 메트릭에 대한 라이덴 클러스터링인, 방법.

청구항 37

제34항에 있어서, 상기 그래프 클러스터링은 루뱅 클러스터링인, 방법.

청구항 38

제27항 내지 제37항 중 어느 한 항에 있어서, 상기 복수의 세포 구성성분 모듈은 10개 내지 2000개의 세포 구성성분 모듈로 이루어진 것인, 방법.

청구항 39

제27항 내지 제37항 중 어느 한 항에 있어서, 상기 복수의 세포 구성성분은 100개 내지 8,000개의 세포 구성성분으로 이루어지는, 방법.

청구항 40

제27항 내지 제37항 중 어느 한 항에 있어서, 상기 복수의 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 2개 내지 300개의 세포 구성성분으로 이루어지는, 방법.

청구항 41

제1항 내지 제40항 중 어느 한 항에 있어서, 상기 관심 생리학적 조건은 질환인, 방법.

청구항 42

제27항에 있어서, 상기 관심 생리학적 조건은 질환이고, 상기 제1 복수의 세포는 상기 복수의 주석화된 세포 상태에 의해 표시되는 바와 같은, 상기 질환을 대표하는 세포 및 상기 질환을 대표하지 않는 세포를 포함하는, 방법.

청구항 43

제27항에 있어서, 상기 복수의 공변량은 세포 배치(batch), 세포 공여자, 세포 유형, 질환 상태, 화학적 화합물에 대한 노출, 또는 그 임의의 조합을 포함하는, 방법.

청구항 44

제27항에 있어서, 상기 후보 세포 구성성분 모델을 훈련시키는 단계는 멀티-태스크 공식화에서 범주형 교차-엔트로피 손실을 사용하여 수행되며, 상기 복수의 공변량에서의 각각의 공변량은 복수의 비용 함수에서의 비용 함수에 상응하고, 상기 복수의 비용 함수의 각각의 개별 비용 함수는 공통 가중 인자를 갖는, 방법.

청구항 45

제1항 내지 제44항 중 어느 한 항에 있어서, 상기 테스트 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물인, 방법.

청구항 46

제45항에 있어서, 상기 테스트 화학적 화합물은 리핀스키 5 준칙 각각을 충족시키는 유기 화합물인, 방법.

청구항 47

제45항에 있어서, 상기 테스트 화학적 화합물은 상기 리핀스키 5 준칙의 적어도 3가지 기준을 충족시키는 유기 화합물인, 방법.

청구항 48

제1항 내지 제19항 중 어느 한 항에 있어서, 상기 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함하는, 방법.

청구항 49

제1항 내지 제48항 중 어느 한 항에 있어서, 상기 방법은 데이라이트(Daylight), BCI, ECFP4, EcFC, MDL, APFP, TTFP, UNITY 2D 지문, RNNS2S 또는 GraphConv를 사용하여 상기 테스트 화학적 화합물의 화학 구조로부터 상기 지문을 생성하는 단계를 더 포함하는, 방법.

청구항 50

제1항 내지 제18항 및 제20항 내지 제49항 중 어느 한 항에 있어서, 상기 세포 구성성분 모듈의 세트는 5개 이상의 세포 구성성분 모듈을 포함하는, 방법.

청구항 51

제1항 내지 제18항 및 제20항 내지 제50항 중 어느 한 항에 있어서, 상기 세포 구성성분 모듈의 세트는 10개 이상의 세포 구성성분 모듈을 포함하는, 방법.

청구항 52

제1항 내지 제18항 및 제20항 내지 제50항 중 어느 한 항에 있어서, 상기 세포 구성성분 모듈의 세트는 100개

이상의 세포 구성성분 모듈을 포함하는, 방법.

청구항 53

제1항 내지 제52항 중 어느 한 항에 있어서, 상기 각각의 세포 구성성분 모듈 내의 상기 복수의 세포 구성성분의 상기 독립적인 서브세트는 5개 이상의 세포 구성성분을 포함하는, 방법.

청구항 54

제1항 내지 제52항 중 어느 한 항에 있어서, 상기 각각의 세포 구성성분 모듈 내의 상기 복수의 세포 구성성분의 상기 독립적인 서브세트는 상기 관심 생리학적 조건과 연관된 분자 경로 내의 2개 내지 20개의 세포 구성성분으로 이루어지는, 방법.

청구항 55

제1항 내지 제54항 중 어느 한 항에 있어서, 상기 제1 임계치 기준은 상기 제1 세포 구성성분 모듈이 임계 활성화 점수를 가져야 한다는 요건인, 방법.

청구항 56

하나 이상의 프로세서 및 메모리를 포함하는 컴퓨터 시스템으로서, 상기 메모리는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 수행하기 위한 명령어를 저장하고, 상기 방법은,

(A) 상기 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계;

(B) 세포 구성성분 모듈의 세트에 액세스하는 단계로서,

상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 각각의 독립적인 서브세트를 포함하고,

상기 복수의 세포 구성성분의 각각의 개별 독립적인 서브세트에 대한 상응하는 복수의 세포-기반 검정 풍부도 값은 상기 생리학적 조건과 연관된 복수의 상이한 상태에 걸쳐 개별적으로 상관되고,

상기 세포 구성성분 모듈의 세트의 제1 세포 구성성분 모듈은 상기 관심 생리학적 조건과 연관되는, 액세스하는 단계;

(C) 모델 내로 상기 화학 구조의 지문을 입력하는 것에 응답하여(상기 모델은 100개 이상의 파라미터를 포함함), 상기 모델로부터의 출력으로서, 상기 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈에 대한 각각의 활성화 점수를 검색하는 단계; 및

(D) 상기 제1 세포 구성성분 모듈에 대한 상기 활성화 점수가 제1 임계치 기준을 충족하는 경우에, 상기 테스트 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 단계를 포함하는, 컴퓨터 시스템.

청구항 57

테스트 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한, 컴퓨터에 의해 실행가능한 하나 이상의 컴퓨터 프로그램을 저장하는 비-일시적 컴퓨터-판독가능 매체로서, 상기 컴퓨터는 하나 이상의 프로세서 및 메모리를 포함하고, 상기 하나 이상의 컴퓨터 프로그램은 집합적으로 방법을 수행하는 컴퓨터 실행 가능 명령어를 인코딩하고, 상기 방법은,

(A) 상기 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계;

(B) 세포 구성성분 모듈의 세트에 액세스하는 단계로서,

상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 각각의 독립적인 서브세트를 포함하고,

상기 복수의 세포 구성성분의 각각의 개별 독립적인 서브세트에 대한 상응하는 복수의 세포-기반 검정 풍부도 값은 상기 생리학적 조건과 연관된 복수의 상이한 상태에 걸쳐 개별적으로 상관되고,

상기 세포 구성성분 모듈의 세트의 제1 세포 구성성분 모듈은 상기 관심 생리학적 조건과 연관되는, 상기 액세스하는 단계;

(C) 모델 내로 상기 화학 구조의 지문을 입력하는 것에 응답하여(상기 모델은 100개 이상의 파라미터를 포함함), 상기 모델로부터의 출력으로서, 상기 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈에 대한 각각의 활성화 점수를 검색하는 단계; 및

(D) 상기 제1 세포 구성성분 모듈에 대한 상기 활성화 점수가 제1 임계치 기준을 충족할 때, 상기 테스트 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 단계를 포함하는, 비-일시적 컴퓨터-판독가능 매체.

청구항 58

테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법으로서, 상기 방법은,

(A) 상기 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계;

(B) 교란 시그니처의 세트에 액세스하는 단계로서,

상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 복수의 세포 구성성분의 각각의 독립적인 서브세트를 포함하고,

상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별을 포함하고, 상기 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 상기 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하며, 상기 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 상기 각각의 제1 세포 상태 및 상기 제2 세포 상태 중 다른 하나는 상기 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태인, 단계;

(C) 상기 지문을 모델에 입력하는 단계로서,

상기 모델은 100개 이상의 파라미터를 포함하고,

상기 모델은 상기 모델로의 상기 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력하고,

상기 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 상기 교란 시그니처의 세트에서의 상응하는 교란 시그니처를 나타내는, 상기 액세스하는 단계; 및

(D) 상기 교란 시그니처의 세트의 제1 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족할 때 상기 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 단계를 포함하는, 방법.

청구항 59

제58항에 있어서, 상기 방법은 상기 테스트 화학적 화합물의 단순화된 분자-입력 라인-엔트리 시스템(SMILES) 스트링 표현으로부터 상기 지문을 계산하는 단계를 더 포함하는, 방법.

청구항 60

제58항 또는 제59항에 있어서, 상기 모델은 신경망을 포함하는, 방법.

청구항 61

제60항에 있어서, 상기 신경망은 완전 연결 신경망, 메시지 전달 신경망, 또는 그의 조합인, 방법.

청구항 62

제58항 내지 제61항 중 어느 한 항에 있어서, 상기 모델은 복수의 컴포넌트 모델을 포함하는 앙상블 모델이고, 상기 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 상기 복수의 컴포넌트 모델 세트에서의 각각의 컴포넌트 모델에 상기 화학 구조의 지문을 입력하는 것에 응답하여 상기 교란 시그니처의 세트에서의 상이한 교란 시그니처에 대한 활성화 점수를 제공하는, 방법.

청구항 63

제62항에 있어서, 상기 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로

지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함하는, 방법.

청구항 64

제62항 또는 제63항에 있어서, 상기 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 상응하는 신경망인, 방법.

청구항 65

제64항에 있어서, 상기 상응하는 신경망은 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합인, 방법.

청구항 66

제63항 또는 제64항에 있어서, 상기 복수의 컴포넌트 모델의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델인, 방법.

청구항 67

제65항에 있어서,

상기 상응하는 신경망은 완전 연결 신경망과 메시지 전달 신경망의 조합이고,

상기 제1 신경망의 제1 출력 및 상기 제2 신경망의 제2 출력은, 상기 화학 구조의 지문을 상기 완전 연결 신경망 및 상기 메시지 전달 신경망에 입력하는 것에 응답하여 조합되어, 상기 교란 시그니처의 세트에서의 제1 교란 시그니처에 대한 상기 하나 이상의 계산된 활성화 점수에서의 활성화 점수를 결정하는, 방법.

청구항 68

제58항에 있어서,

상기 교란 시그니처의 세트는 복수의 교란 시그니처이고,

상기 제1 교란 시그니처를 포함하는 상기 복수의 교란 시그니처의 제1 서브세트는 상기 관심 생리학적 조건과 연관되고,

상기 복수의 교란 시그니처의 제2 서브세트는 상기 관심 생리학적 조건과 연관되지 않고,

상기 제1 교란 시그니처에 대한 상기 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족하고, 상기 복수의 교란 시그니처의 상기 제2 서브세트에서의 교란 시그니처에 대한 상기 각각의 계산된 활성화 점수가 상기 제1 임계치 기준 이외의 제2 임계치 기준을 충족하는 경우에, 상기 테스트 화학적 화합물은 상기 관심 생리학적 조건으로 식별되는, 방법.

청구항 69

제58항 내지 제68항 중 어느 한 항에 있어서, 상기 관심 생리학적 조건은 질환인, 방법.

청구항 70

제58항에 있어서, 상기 테스트 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물인, 방법.

청구항 71

제70항에 있어서, 상기 테스트 화학적 화합물은 상기 리핀스키 5 준칙 각각을 충족시키는 유기 화합물인, 방법.

청구항 72

제70항에 있어서, 상기 테스트 화학적 화합물은 상기 리핀스키 5 준칙의 적어도 3가지 기준을 충족시키는 유기 화합물인, 방법.

청구항 73

제58항에 있어서, 상기 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델,

선형 모델, 또는 선형 회귀 모델을 포함하는, 방법.

청구항 74

제58항 내지 제73항 중 어느 한 항에 있어서, 상기 방법은 데이라이트(Daylight), BCI, ECFP4, EcFC, MDL, APFP, TTFP, UNITY 2D 지문, RNNS2S 또는 GraphConv를 사용하여 상기 테스트 화학적 화합물의 상기 화학 구조로부터 상기 지문을 생성하는 단계를 더 포함하는, 방법.

청구항 75

제58항 내지 제74항 중 어느 한 항에 있어서, 상기 교란 시그니처의 세트는 상기 제1 교란 시그니처로 이루어지는, 방법.

청구항 76

제58항 내지 제74항 중 어느 한 항에 있어서, 상기 교란 시그니처의 세트는 5개 이상의 교란 시그니처를 포함하는, 방법.

청구항 77

제58항 내지 제74항 중 어느 한 항에 있어서, 상기 교란 시그니처의 세트는 10개 이상의 교란 시그니처를 포함하는, 방법.

청구항 78

제58항 내지 제74항 중 어느 한 항에 있어서, 상기 교란 시그니처의 세트는 100개 이상의 교란 시그니처를 포함하는, 방법.

청구항 79

제58항 내지 제74항 중 어느 한 항에 있어서, 상기 제1 임계치 기준은 상기 제1 교란 시그니처가 임계 활성화 점수를 가져야 한다는 요건인, 방법.

청구항 80

하나 이상의 프로세서 및 메모리를 포함하는 컴퓨터 시스템으로서, 상기 메모리는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 수행하기 위한 명령어를 저장하고, 상기 방법은,

(A) 상기 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계;

(B) 교란 시그니처의 세트에 액세스하는 단계로서,

상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 복수의 세포 구성성분의 각각의 독립적인 서브세트를 포함하고,

상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별을 포함하고, 상기 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하며, 상기 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 상기 각각의 제1 세포 상태 및 상기 제2 세포 상태 중 다른 하나는 상기 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태인, 상기 액세스하는 단계;

(C) 상기 지문을 모델에 입력하는 단계로서,

상기 모델은 100개 이상의 파라미터를 포함하고,

상기 모델은 상기 모델로의 상기 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력하고,

상기 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 상기 교란 시그니처의 세트에서의 상응하는 교란 시그니처를 나타내는, 상기 입력하는 단계; 및

(D) 상기 교란 시그니처의 세트의 제1 교란 시그니처에 대한 상기 각각의 계산된 활성화 점수가 제1 임계치 기

준을 충족할 때 상기 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 단계를 포함하는, 방법.

청구항 81

테스트 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한, 컴퓨터에 의해 실행가능한 하나 이상의 컴퓨터 프로그램을 저장하는 비-일시적 컴퓨터-판독가능 매체로서, 상기 컴퓨터는 하나 이상의 프로세서 및 메모리를 포함하고, 상기 하나 이상의 컴퓨터 프로그램은 집합적으로 방법을 수행하는 컴퓨터 실행 가능 명령어를 인코딩하고, 상기 방법은,

(A) 상기 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계;

(B) 교란 시그니처의 세트에 액세스하는 단계로서,

상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 복수의 세포 구성성분의 각각의 독립적인 서브세트를 포함하고,

상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별을 포함하고, 상기 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 상기 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하며, 상기 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 상기 각각의 제1 세포 상태 및 상기 제2 세포 상태 중 다른 하나는 상기 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태인, 상기 액세스하는 단계;

(C) 상기 지문을 모델에 입력하는 단계로서,

상기 모델은 100개 이상의 파라미터를 포함하고,

상기 모델은 상기 모델로의 상기 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력하고,

상기 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 상기 교란 시그니처의 세트에서의 상응하는 교란 시그니처를 나타내는, 상기 입력하는 단계; 및

(D) 상기 교란 시그니처의 세트의 제1 교란 시그니처에 대한 상기 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족할 때 상기 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 단계를 포함하는, 비-일시적 컴퓨터-판독가능 매체.

청구항 82

화학적 화합물을 관심 생리학적 조건과 연관시키는 방법으로서, 상기 방법은,

메모리 및 하나 이상의 프로세서를 포함하는 컴퓨터 시스템에서,

(A) 전자 형태로, 복수의 화합물의 각각의 개별 화합물의 상응하는 화학 구조의 각각의 지문을 획득하여, 복수의 지문을 획득하는 단계;

(B) 전자 형태로, 상기 복수의 화합물의 각각의 화합물에 대한 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수를 획득하는 단계로서, 상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함하는, 상기 획득하는 단계;

(C) 훈련되지 않은 모델을

상기 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해,

상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해,

(i) 상기 각각의 화합물의 상기 화학 구조의 지문을 상기 훈련되지 않은 모델로 입력시 상기 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수 및 (ii) 상기 세포 구성성분 모듈의 세트의 상기 각각의 화합물에 대한 상기 각각의 세포 구성성분 모듈의 상기 각각의 수치 활성화 점수 사이의 차이를 사용하여 훈련시키는 단계로서, 상기 훈련시키는 단계 (C)는 상기 차이에 응답하여 상기 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고, 상기 복수의 파라미터는 100개 이상의 파라미터를 포함하며, 이에 의해 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득하는, 상기 훈련시키는 단계를 포함하는, 방법.

청구항 83

제82항에 있어서, 상기 세포 구성성분 모듈의 세트는 단일 세포 구성성분 모듈로 이루어지는, 방법.

청구항 84

제82항에 있어서, 상기 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈을 포함하는, 방법.

청구항 85

제82항에 있어서, 상기 세포 구성성분 모듈의 세트는 2개 내지 500개의 세포 구성성분 모듈로 이루어지는, 방법.

청구항 86

제82항에 있어서, 상기 복수의 화합물은 10개 내지 1×10^6개의 화합물로 이루어지는, 방법.

청구항 87

제82항에 있어서, 상기 복수의 화합물은 100개 내지 100,000개의 화합물로 이루어지는, 방법.

청구항 88

제82항에 있어서, 상기 복수의 화합물은 1000개 내지 100,000개의 화합물로 이루어지는, 방법.

청구항 89

제82항 내지 제88항 중 어느 한 항에 있어서, 상기 훈련시키는 단계 (C)는 회귀 알고리즘에 따라 상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해 각각의 개별 화합물과 연관된 각각의 차이에 응답하여 상기 훈련되지 않은 모델과 연관된 상기 복수의 파라미터를 조정하는, 방법.

청구항 90

제89항에 있어서, 상기 회귀 알고리즘은 상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대한 각각의 개별 화합물과 연관된 각각의 차이의 최소 제곱 오차를 최적화하는, 방법.

청구항 91

제82항 내지 제90항 중 어느 한 항에 있어서, 상기 훈련된 모델은 신경망을 포함하는, 방법.

청구항 92

제91항에 있어서, 상기 신경망은 완전 연결 신경망, 메시지 전달 신경망, 또는 그의 조합인, 방법.

청구항 93

제82항 내지 제90항 중 어느 한 항에 있어서, 상기 훈련된 모델은 복수의 컴포넌트 모델의 앙상블 모델이고, 상기 복수의 컴포넌트 모델의 각각의 개별 컴포넌트 모델은 상기 복수의 세포 구성성분 모듈의 상이한 세포 구성성분 모듈에 대한 계산된 활성화 점수를 출력하는, 방법.

청구항 94

제93항에 있어서, 상기 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함하는, 방법.

청구항 95

제93항에 있어서, 상기 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 상응하는 신경망인, 방법.

청구항 96

제95항에 있어서, 상기 상응하는 신경망은 완전 연결 신경망, 메시지 전달 신경망, 또는 그의 조합인, 방법.

청구항 97

제82항 내지 제96항 중 어느 한 항에 있어서,

상기 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈이고,

상기 복수의 세포 구성성분 모듈의 제1 서브세트는 상기 관심 생리학적 조건과 연관되고,

상기 복수의 세포 구성성분 모듈의 제2 서브세트는 상기 관심 생리학적 조건과 연관되지 않는, 방법.

청구항 98

제82항 내지 제97항 중 어느 한 항에 있어서, 상기 방법은 과정에 의해 상기 복수의 세포 구성성분 모듈의 세포 구성성분 모듈을 식별하는 단계를 더 포함하고, 상기 과정은,

전자 형태로 하나 이상의 제1 데이터세트를 획득하는 단계로서, 상기 하나 이상의 제1 데이터세트는,

제1 복수의 세포의 각각의 개별 세포에 대해(상기 제1 복수의 세포는 20개 이상의 세포를 포함하고, 집합적으로 복수의 주석화된 세포 상태를 나타냄),

상기 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해(상기 복수의 세포 구성성분은 10개 이상의 세포 구성성분을 포함함),

상기 각각의 세포에서 상기 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하고,

이에 의해 복수의 벡터에 액세스하거나 복수의 벡터를 형성하며, 상기 복수의 벡터의 각각의 개별 벡터는 (i) 상기 복수의 구성성분의 각각의 세포 구성성분에 상응하고 (ii) 상응하는 복수의 요소를 포함하고, 상기 상응하는 복수의 요소의 각각의 개별 요소는 상기 제1 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 갖는, 상기 획득하는 단계;

상기 복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계로서, 상기 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 상기 복수의 세포 구성성분의 서브세트를 포함하고, 상기 복수의 세포 구성성분 모듈은 (i) 상기 복수의 후보 세포 구성성분 모듈 및 (ii) 상기 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현으로 배열되고, 상기 복수의 세포 구성성분 모듈은 10개를 초과하는 세포 구성성분 모듈을 포함하는, 상기 식별하는 단계;

전자 형태로 하나 이상의 제2 데이터세트를 획득하는 단계로서, 상기 하나 이상의 제2 데이터세트는,

제2 복수의 세포의 각각의 개별 세포에 대해(상기 제2 복수의 세포는 20개 이상의 세포를 포함하고, 상기 관심 생리학적 조건을 알리는 복수의 공변량을 집합적으로 나타냄),

상기 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해,

상기 각각의 세포에서 상기 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하고,

이에 의해, (i) 상기 제2 복수의 세포 및 (ii) 상기 복수의 세포 구성성분 또는 그 표현으로 차원화된 세포 구성성분 카운트 데이터 구조를 획득하는, 상기 획득하는 단계;

상기 복수의 세포 구성성분 또는 그 표현을 공통 차원으로서 사용하여 상기 세포 구성성분 카운트 데이터 구조 및 상기 잠재 표현을 조합함으로써 활성화 데이터 구조를 형성하는 단계로서, 상기 활성화 데이터 구조는 상기 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈에 대해,

상기 제2 복수의 세포의 각각의 세포에 대해, 각각의 활성화 가중치를 포함하고, 상기 형성하는 단계;

(i) 상기 활성화 데이터 구조를 상기 후보 모델로 입력시 상기 활성화 데이터 구조 내에 표현된 각각의 세포 구성성분 모듈 내의 복수의 공변량 중 각각의 공변량의 부재 또는 존재의 예측과 (ii) 각각의 세포 구성성분 모듈의 각각의 공변량의 실제 부재 또는 존재 사이의 차이를 사용하여 후보 세포 구성성분 모델을 훈련시키는 단계로서, 상기 훈련시키는 단계는 상기 차이에 응답하여 상기 후보 세포 구성성분 모델과 연관된 복수의 공변량 파라미터를 조정하는, 상기 훈련시키는 단계를 포함하는, 방법.

청구항 99

제98항에 있어서, 상기 복수의 공변량 파라미터는,

상기 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대해,

각각의 개별 공변량에 대해,

상기 각각의 공변량이 상기 제2 복수의 세포에 걸쳐 상기 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 상응하는 파라미터를 포함하고;

상기 후보 세포 구성성분 모델의 훈련시 상기 복수의 공변량 파라미터를 사용하여, 상기 복수의 후보 세포 구성성분 모듈에서 상기 세포 구성성분 모듈을 식별하는 단계를 포함하는, 방법.

청구항 100

제99항에 있어서, 상기 복수의 주석화된 세포 상태의 주석화된 세포 상태는 노출 조건 하에 화합물에 대한 상기 제1 복수의 세포의 세포의 노출인, 방법.

청구항 101

제99항에 있어서, 상기 노출 조건은 노출 지속기간, 화합물의 농도, 또는 노출 지속기간과 상기 화합물의 농도의 조합인, 방법.

청구항 102

제82항 내지 제101항 중 어느 한 항에 있어서, 상기 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합인, 방법.

청구항 103

제98항에 있어서, 상기 제1 또는 제2 복수의 세포의 상기 각각의 세포의 상기 각각의 세포 구성성분의 상응하는 풍부도는 비색 측정치, 형광 측정치, 발광 측정치, 또는 공명 에너지 전달(FRET) 측정치에 의해 결정되는, 방법.

청구항 104

제98항에 있어서, 상기 제1 또는 제2 복수의 세포에서의 상기 각각의 세포에서의 상기 각각의 세포 구성성분의 상기 상응하는 풍부도는 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq), scTag-seq, 시퀀싱을 사용한 전위효소-접근 가능 염색질에 대한 단일-세포 검정(scATAC-seq), CyTOF/SCoP, E-MS/Abseq, miRNA-seq, CITE-seq, 또는 그 임의의 조합에 의해 결정되는, 방법.

청구항 105

제98항에 있어서, 상기 복수의 벡터를 사용하여 상기 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계는 상기 복수의 벡터의 각각의 벡터의 각각의 상응하는 복수의 요소를 사용하여 상관 모델을 상기 복수의 벡터에 적용하는 것을 포함하는, 방법.

청구항 106

제105항에 있어서, 상기 상관 모델은 그래프 클러스터링을 포함하는, 방법.

청구항 107

제106항에 있어서, 상기 그래프 클러스터링 방법은 피어슨-상관관계-기반 거리 메트릭에 대한 라이덴 클러스터링이거나 또는 루뱅 클러스터링인, 방법.

청구항 108

제82항 내지 제107항 중 어느 한 항에 있어서, 상기 복수의 세포 구성성분은 100개 내지 8,000개의 세포 구성성분으로 이루어진 것인, 방법.

청구항 109

제98항에 있어서, 상기 복수의 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 2개 내지 300개의 세포 구성성분으로 이루어지는, 방법.

청구항 110

제82항 내지 제109항 중 어느 한 항에 있어서, 상기 관심 생리학적 조건은 질환인, 방법.

청구항 111

제98항에 있어서, 상기 생리학적 조건은 질환이고, 상기 제1 복수의 세포는 상기 복수의 주석화된 세포 상태에 의해 표시되는 바와 같은, 상기 질환을 대표하는 세포 및 상기 질환을 대표하지 않는 세포를 포함하는, 방법.

청구항 112

제98항에 있어서, 상기 복수의 공변량은 세포 배치, 세포 공여자, 세포 유형, 질환 상태, 또는 화학적 화합물에 대한 노출을 포함하는, 방법.

청구항 113

제98항에 있어서, 상기 후보 세포 구성성분 모델을 훈련시키는 단계는 멀티-태스크 공식화에서 범주형 교차-엔트로피 손실을 사용하여 수행되며, 상기 복수의 공변량에서의 각각의 공변량은 복수의 비용 함수에서의 비용 함수에 상응하고, 상기 복수의 비용 함수의 각각의 개별 비용 함수는 공통 가중 인자를 갖는, 방법.

청구항 114

제82항 내지 제113항 중 어느 한 항에 있어서, 상기 복수의 화학적 화합물의 각각의 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물인, 방법.

청구항 115

제82항 내지 제113항 중 어느 한 항에 있어서, 상기 복수의 화학적 화합물의 각각의 화학적 화합물은 상기 리핀스키 5 준칙 각각을 충족시키는, 방법.

청구항 116

제82항 내지 제113항 중 어느 한 항에 있어서, 상기 복수의 화학적 화합물의 각각의 화학적 화합물은 상기 리핀스키 5 준칙의 적어도 3가지 기준을 충족시키는, 방법.

청구항 117

제82항 내지 제116항 중 어느 한 항에 있어서, 상기 훈련된 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함하는, 방법.

청구항 118

제82항 내지 제117항 중 어느 한 항에 있어서, 상기 방법은 데이라이트(Daylight), BCI, ECFP4, EcFC, MDL, APFP, TTFP, 유니티(UNITY) 2D 지문, RNNS2S 또는 GraphConv를 사용하여 상응하는 화학 구조로부터 각각의 개별 지문을 생성하는 단계를 더 포함하는, 방법.

청구항 119

제82항에 있어서, 상기 세포 구성성분 모듈의 세트는 5개 이상의 세포 구성성분 모듈을 포함하는, 방법.

청구항 120

제82항에 있어서, 상기 세포 구성성분 모듈의 세트는 10개 이상의 세포 구성성분 모듈을 포함하는, 방법.

청구항 121

제82항에 있어서, 상기 세포 구성성분 모듈의 세트는 100개 이상의 세포 구성성분 모듈을 포함하는, 방법.

청구항 122

하나 이상의 프로세서 및 메모리를 포함하는 컴퓨터 시스템으로서, 상기 메모리는 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 수행하기 위한 명령어를 저장하고, 상기 방법은,

(A) 전자 형태로, 복수의 화합물의 각각의 개별 화합물의 상응하는 화학 구조의 각각의 지문을 획득하여, 복수의 지문을 획득하는 단계;

(B) 전자 형태로, 상기 복수의 화합물의 각각의 화합물에 대한 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수를 획득하는 단계로서, 상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함하는, 상기 획득하는 단계;

(C) 훈련되지 않은 모델을

상기 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해,

상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해,

(i) 상기 각각의 화합물의 화학 구조의 지문을 상기 훈련되지 않은 모델로 입력시 상기 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수 및 (ii) 상기 세포 구성성분 모듈의 세트의 상기 각각의 화합물에 대한 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련시키는 단계로서, 상기 훈련시키는 단계 (C)는 상기 차이에 응답하여 상기 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고, 상기 복수의 파라미터는 100개 이상의 파라미터를 포함하며, 이에 의해 상기 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득하는, 상기 훈련시키는 단계를 포함하는, 방법.

청구항 123

화학적 화합물을 관심 생리학적 조건과 연관시키기 위한, 컴퓨터에 의해 실행가능한 하나 이상의 컴퓨터 프로그램을 저장하는 비-일시적 컴퓨터-판독가능 매체로서, 상기 컴퓨터는 하나 이상의 프로세서 및 메모리를 포함하고, 상기 하나 이상의 컴퓨터 프로그램은 집합적으로 방법을 수행하는 컴퓨터 실행 가능 명령어를 인코딩하고, 상기 방법은,

(A) 전자 형태로, 복수의 화합물의 각각의 개별 화합물의 상응하는 화학 구조의 각각의 지문을 획득하여, 복수의 지문을 획득하는 단계;

(B) 전자 형태로, 상기 복수의 화합물의 각각의 화합물에 대한 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수를 획득하는 단계로서, 상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함하는, 상기 획득하는 단계;

(C) 훈련되지 않은 모델을

상기 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해,

상기 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해,

(i) 상기 각각의 화합물의 상기 화학 구조의 상기 지문을 상기 훈련되지 않은 모델로 입력시 상기 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수 및 (ii) 상기 세포 구성성분 모듈의 세트의 상기 각각의 화합물에 대한 상기 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련시키는 단계로서, 상기 훈련시키는 단계 (C)는 상기 차이에 응답하여 상기 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고, 상기 복수의 파라미터는 100개 이상의 파라미터를 포함하며, 이에 의해 상기 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득하는, 상기 훈련시키는 단계를 포함하는, 비-일시적 컴퓨터-판독가능 매체.

청구항 124

화학적 화합물을 관심 생리학적 조건과 연관시키는 방법으로서, 상기 방법은,

메모리 및 하나 이상의 프로세서를 포함하는 컴퓨터 시스템에서,

(A) 전자 형태로, 복수의 화합물의 각각의 개별 화합물의 상응하는 화학 구조의 각각의 지문을 획득하여, 복수의 지문을 획득하는 단계;

(B) 전자 형태로, 상기 복수의 화합물의 각각의 상응하는 화합물에 대한 교란 시그니처의 세트의 각각의 개별 교란 시그니처의 각각의 수치 활성화 점수를 획득하는 단계로서, 상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별, 및 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하고, 상기 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 상기 각각의 제1 세포 상태 및 상기 제2 세포 상태 중 다른 하나는 상기 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태인, 상기 획득하는 단계;

(C) 훈련되지 않은 모델을

상기 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해,

상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해:

(i) 상기 각각의 화합물의 상기 화학 구조의 상기 지문을 상기 훈련되지 않은 모델로 입력시의 각각의 교란 시그니처에 대한 각각의 계산된 활성화 점수 및 (ii) 상기 교란 시그니처의 세트의 상기 상응하는 화합물에 대한 상기 각각의 교란 시그니처의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련시키는 단계로서, 상기 훈련시키는 단계 (C)는 상기 차이에 응답하여 상기 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고, 상기 복수의 파라미터는 100개 이상의 파라미터를 포함하며, 이에 의해 상기 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득하는, 상기 훈련시키는 단계를 포함하는, 방법.

청구항 125

제124항에 있어서, 상기 교란 시그니처의 세트는 단일 교란 시그니처로 이루어지는, 방법.

청구항 126

제124항에 있어서, 상기 교란 시그니처의 세트는 2개 내지 500개의 교란 시그니처로 이루어지는, 방법.

청구항 127

제124항 내지 제126항 중 어느 한 항에 있어서, 상기 복수의 화합물은 10개 내지 1×10^6개의 화합물로 이루어지는, 방법.

청구항 128

제124항 내지 제126항 중 어느 한 항에 있어서, 상기 복수의 화합물은 100개 내지 100,000개의 화합물로 이루어지는, 방법.

청구항 129

제124항 내지 제126항 중 어느 한 항에 있어서, 상기 복수의 화합물은 1000개 내지 100,000개의 화합물로 이루어지는, 방법.

청구항 130

제124항 내지 제129항 중 어느 한 항에 있어서, 상기 훈련시키는 단계(C)는 회귀 알고리즘에 따라 상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해 각각의 상응하는 화합물과 연관된 각각의 차이에 응답하여 상기 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하는, 방법.

청구항 131

제130항에 있어서, 상기 회귀 알고리즘은 상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해 각각의 상응하는 화합물과 연관된 각각의 차이의 최소 제곱 오차를 최적화하는, 방법.

청구항 132

제124항 내지 제131항 중 어느 한 항에 있어서, 상기 훈련된 모델은 신경망을 포함하는, 방법.

청구항 133

제132항에 있어서, 상기 신경망은 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합인, 방법.

청구항 134

제124항에 있어서, 상기 훈련된 모델은 복수의 컴포넌트 모델의 앙상블 모델이고, 상기 복수의 컴포넌트 모델의 각각의 개별 컴포넌트 모델은 상기 복수의 컴포넌트 모델 세트에서의 각각의 컴포넌트 모델에 각각의 화학 구조의 지문을 입력하는 것에 응답하여 상기 복수의 교란 시그니처의 세트에서의 상이한 교란 시그니처의 세트에 대한 계산된 활성화 점수를 출력하는, 방법.

청구항 135

제134항에 있어서, 상기 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함하는, 방법.

청구항 136

제134항에 있어서, 상기 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 상응하는 신경망인, 방법.

청구항 137

제136항에 있어서, 상기 상응하는 신경망은 완전 연결 신경망, 메시지 전달 신경망, 또는 그의 조합인, 방법.

청구항 138

제124항 내지 제137항 중 어느 한 항에 있어서,

상기 교란 시그니처의 세트는 복수의 교란 시그니처를 포함하고,

상기 복수의 교란 시그니처의 제1 서브세트는 상기 관심 생리학적 조건과 연관되고,

상기 복수의 교란 시그니처의 제2 서브세트는 상기 관심 생리학적 조건과 연관되지 않는, 방법.

청구항 139

제124항 내지 제138항 중 어느 한 항에 있어서, 상기 관심 생리학적 조건은 질환인, 방법.

청구항 140

제124항 내지 제139항 중 어느 한 항에 있어서, 상기 복수의 화학적 화합물의 각각의 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물인, 방법.

청구항 141

제124항 내지 제140항 중 어느 한 항에 있어서, 상기 복수의 화학적 화합물의 각각의 화학적 화합물은 리핀스키 5 준칙 각각을 충족시키는, 방법.

청구항 142

제124항 내지 제140항 중 어느 한 항에 있어서, 상기 복수의 화학적 화합물의 각각의 화학적 화합물은 상기 리핀스키 5 준칙의 적어도 3가지 기준을 충족시키는, 방법.

청구항 143

제124항에 있어서, 상기 훈련된 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모

델, 선형 모델, 또는 선형 회귀 모델을 포함하는, 방법.

청구항 144

제124항 내지 제143항 중 어느 한 항에 있어서, 상기 방법은 데이라이트(Daylight), BCI, ECFP4, EcFC, MDL, APFP, TTFP, 유니티(UNITY) 2D 지문, RNNS2S 또는 GraphConv를 사용하여 상기 상응하는 화학 구조로부터 각각의 개별 지문을 생성하는 단계를 더 포함하는, 방법.

청구항 145

제124항에 있어서, 상기 교란 시그니처의 세트는 5개 이상의 교란 시그니처를 포함하는, 방법.

청구항 146

제124항에 있어서, 상기 교란 시그니처의 세트는 10개 이상의 교란 시그니처를 포함하는, 방법.

청구항 147

제124항에 있어서, 상기 교란 시그니처의 세트는 100개 이상의 교란 시그니처를 포함하는, 방법.

청구항 148

제124항에 있어서, 상기 방법은 절차에 의해 상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처의 각각의 수치 활성화 점수를 획득하는 단계를 더 포함하며, 상기 절차는,

전자 형태로, 변경되지 않은 세포 상태와 변경된 세포 상태 사이의 차등 세포 구성성분 풍부도의 척도를 나타내는 단일-세포 전이 시그니처에 액세스하는 단계로서,

상기 변경된 세포 상태는 상기 변경되지 않은 세포 상태로부터 변경된 세포 상태로의 세포 전이를 통해 발생하고,

(i) 상기 변경되지 않은 세포 상태, (ii) 상기 변경된 세포 상태, 및 (iii) 상기 변경되지 않은 세포 상태로부터 상기 변경된 세포 상태로의 전이 중 적어도 하나는 상기 관심 생리학적 조건과 연관되고,

상기 단일-세포 전이 시그니처는 기준 복수의 세포 구성성분의 식별, 및 복수의 기준 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 상기 각각의 세포 구성성분의 풍부도의 변화와 상기 변경되지 않은 세포 상태와 상기 변경된 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 제1 유의성 점수를 포함하는, 상기 액세스하는 단계; 및

상기 단일-세포 전이 시그니처 및 상기 각각의 교란 시그니처를 비교함으로써 상기 각각의 교란 시그니처의 상기 각각의 수치 활성화 점수를 결정하는 단계를 포함하는, 방법.

청구항 149

제148항에 있어서, 상기 단일-세포 전이 시그니처 및 상기 교란 시그니처를 비교하여 각각의 교란 시그니처의 각각의 수치 활성화 점수를 결정하는 단계는, 상기 단일-세포 전이 시그니처의 기준 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해,

상기 각각의 세포 구성성분의 상기 제1 유의성 점수와 상기 각각의 교란 시그니처의 상기 상응하는 세포 구성성분의 상기 상응하는 유의성 점수를 비교하는 단계를 포함하는, 방법.

청구항 150

제148항 또는 제149항에 있어서, 상기 각각의 교란 시그니처의 상기 활성화 점수는 상기 교란 시그니처의 세트의 다른 교란 시그니처에 대한 각각의 교란 시그니처의 단일-세포 전이 시그니처에 대한 관련성의 상대 순위인, 방법.

청구항 151

제150항에 있어서, 상기 상대 순위는 윌콕슨 순위-합계 검정, t-검정, 로지스틱 회귀 또는 일반화된 선형 모델에 의해 결정되는, 방법.

청구항 152

제148항 내지 제151항 중 어느 한 항에 있어서, 상기 단일-세포 전이 시그니처의 상기 변경되지 않은 세포 상태는 상기 각각의 교란 시그니처의 상기 제1 세포 상태 또는 상기 제2 세포 상태와 동일한, 방법.

청구항 153

제148항 내지 제151항 중 어느 한 항에 있어서, 상기 단일-세포 전이 시그니처의 상기 변경되지 않은 세포 상태는 상기 각각의 교란 시그니처의 상기 제1 세포 상태 및 상기 제2 세포 상태 둘 모두와 상이한, 방법.

청구항 154

제148항 내지 제153항 중 어느 한 항에 있어서,

상기 단일-세포 전이 시그니처의 상기 기준 복수의 세포 구성성분 및 상기 각각의 교란 시그니처의 상기 각각의 복수의 세포 구성성분을 프루닝하여 전사 인자와의 상기 비교를 제한하는 단계를 더 포함하는, 방법.

청구항 155

제124항 내지 제154항 중 어느 한 항에 있어서, 상기 복수의 교란 시그니처에서의 상기 각각의 교란 시그니처의 상기 교란된 세포 상태는 상기 복수의 화합물 중의 화합물에 노출되지 않은 대조군 세포에 의해 나타나는, 방법.

청구항 156

제124항 내지 제154항 중 어느 한 항에 있어서, 상기 복수의 교란 시그니처의 각각의 교란 시그니처의 상기 교란된 세포 상태가 상기 각각의 교란 시그니처와 연관된 상기 화합물 이외의 상기 복수의 화학적 화합물의 화학적 화합물에 노출된 관련되지 않은 교란된 세포에 걸친 평균으로 나타내어지는, 방법.

청구항 157

하나 이상의 프로세서 및 메모리를 포함하는 컴퓨터 시스템으로서, 상기 메모리는 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한 명령어를 저장하고, 방법은,

(A) 전자 형태로, 복수의 화합물의 각각의 개별 화합물의 상응하는 화학 구조의 각각의 지문을 획득하여, 복수의 지문을 획득하는 단계;

(B) 전자 형태로, 상기 복수의 화합물의 각각의 상응하는 화합물에 대한 교란 시그니처의 세트의 각각의 개별 교란 시그니처의 각각의 수치 활성화 점수를 획득하는 단계로서, 상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별, 및 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하고, 상기 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 상기 각각의 제1 세포 상태 및 상기 제2 세포 상태 중 다른 하나는 상기 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태인, 상기 획득하는 단계;

(C) 훈련되지 않은 모델을

상기 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해,

상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해,

(i) 상기 각각의 화합물의 상기 화학 구조의 상기 지문을 훈련되지 않은 모델로 입력시의 각각의 교란 시그니처에 대한 각각의 계산된 활성화 점수 및 (ii) 상기 교란 시그니처의 세트의 상기 상응하는 화합물에 대한 상기 각각의 교란 시그니처의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련시키는 단계로서, 상기 훈련시키는 단계 (C)는 상기 차이에 응답하여 상기 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고, 상기 복수의 파라미터는 100개 이상의 파라미터를 포함하며, 이에 의해 상기 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득하는, 상기 훈련시키는 단계를 포함하는, 컴퓨터 시스템.

청구항 158

화학적 화합물을 관심 생리학적 조건과 연관시키기 위한, 컴퓨터에 의해 실행가능한 하나 이상의 컴퓨터 프로그램을 저장하는 비-일시적 컴퓨터-판독가능 매체로서, 상기 컴퓨터는 하나 이상의 프로세서 및 메모리를 포함하고, 상기 하나 이상의 컴퓨터 프로그램은 집합적으로 방법을 수행하는 컴퓨터 실행 가능 명령어를 인코딩하고, 상기 방법은,

(A) 전자 형태로, 복수의 화합물의 각각의 개별 화합물의 상응하는 화학 구조의 각각의 지문을 획득하여, 복수의 지문을 획득하는 단계;

(B) 전자 형태로, 상기 복수의 화합물의 각각의 상응하는 화합물에 대한 교란 시그니처의 세트의 각각의 개별 교란 시그니처의 각각의 수치 활성화 점수를 획득하는 단계로서, 상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별, 및 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하고, 상기 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 상기 각각의 제1 세포 상태 및 상기 제2 세포 상태 중 다른 하나는 상기 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태인, 상기 획득하는 단계;

(C) 훈련되지 않은 모델을

상기 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해,

상기 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해,

(i) 상기 각각의 화합물의 상기 화학 구조의 상기 지문을 훈련되지 않은 모델로 입력시의 각각의 교란 시그니처에 대한 각각의 계산된 활성화 점수 및 (ii) 상기 교란 시그니처의 세트의 상기 상응하는 화합물에 대한 상기 각각의 교란 시그니처의 상기 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련시키는 단계로서, 상기 훈련시키는 단계 (C)는 상기 차이에 응답하여 상기 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고, 상기 복수의 파라미터는 100개 이상의 파라미터를 포함하며, 이에 의해 상기 화학적 화합물을 상기 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득하는, 상기 훈련시키는 단계를 포함하는, 비-일시적 컴퓨터-판독가능 매체.

청구항 159

제1항 내지 제158항 중 어느 한 항에 있어서, 모델은 리그레서(regressor)인, 방법.

발명의 설명

기 술 분 야

[0001] **관련 출원에 대한 상호 참조**

[0002] 본 출원은 2021년 6월 15일에 출원된 발명의 명칭이 "SYSTEMS AND METHODS FOR ASSOCIATING COMPOUNDS WITH PHYSIOLOGICAL CONDITIONS USING FINGERPRINT ANALYSIS"인 미국 특허 가출원 제63/210,930호; 및 2021년 6월 15일에 출원된 발명의 명칭이 "COMPUTATIONAL MODELING PLATFORM"인 미국 특허 가출원 제63/210,679호에 대한 우선권을 주장하며, 이들 각각은 그 전문이 본원에 참조로 포함된다.

[0003] **기술 분야**

[0004] 본 발명은 일반적으로 화합물을 생리학적 조건과 연관시키기 위한 시스템 및 방법에 관한 것이다.

배 경 기 술

[0005] 세포 메커니즘의 연구는 질환을 이해하는데 중요하다.

[0006] 생물학적 조직은 동적이고 고도로 네트워크화된 다세포 시스템이다. 특정 세포에서의 세포하 네트워크에서의 기능장애는 세포 거동의 전체 랜드스케이프를 이동시키고, 질환 상태로 이어진다. 기존의 약물 발견 노력은 세포가 건강한 상태에서 질환 상태로 전이되게 하는 분자 메커니즘을 특징화하고, 이들 전이를 역전시키거나 억제

하는 약리학적 접근법을 식별하고자 한다. 과거의 노력은 또한 이들 전이를 특징짓는 분자 시그니처를 식별하고, 이들 시그니처를 역전시키는 약리학적 접근법을 식별하기를 추구하여 왔다.

[0007] 조직 또는 표면 마커에 의해 풍부화된 세포에서, 세포의 벌크 수집물에 대한 분자 데이터는 집단 내의 개별 세포의 표현형 및 분자 다양성을 차폐한다. 세포의 이들 벌크 수집물에서 세포의 이질성은 질환-유도 메커니즘을 규명하는 것을 목표로 하는 현재의 노력의 결과를 잘못 유도하거나 심지어 완전히 부정확하게 한다. 새로운 접근법, 예컨대 단일-세포 RNA 시퀀싱은 분자 수준에서 개별 세포를 특징화할 수 있다. 이들 데이터는 보다 높은 분해능에서 다양한 세포 상태를 이해하기 위한 기질을 제공하고, 세포가 보유하는 상태의 풍부하고 현저한 다양성을 밝혀낸다.

[0008] 단일 세포 데이터를 해석할 때, 상당한 과제, 즉 이들 분자 측정의 정확도의 불확실성과 함께, 이들 데이터의 희소성, 세포에 존재하는 분자의 존재의 간과 및 노이즈라는 과제가 존재한다. 따라서, 개별 세포 상태를 제어하기 위한 약리학적 접근법에 대한 통찰을 유도하고, 상응하게 질환을 해결하기 위한 새로운 접근법이 요구된다.

[0009] 또한, 복합 질환은 종종 단일 또는 소수의 분자 표적으로 분해될 수 없다. 시험관내 질환 모델에 대한 고처리량 영상화 기술 및 고처리량 스크리닝에서의 최근의 진보에도 불구하고, 시험관내-기반 스크리닝 접근법으로부터 생성된 후보 표적을 유효한 약물로 번역하는 것은 종종 비교적 느리고 비효율적인 분자 표적-기반 약물 발견 접근법으로의 복귀를 수반하는 상당한 과제이다.

[0010] 상기 배경을 고려하여, 본 기술 분야에는 약물 발견을 위한 후보 화합물의 식별을 위한 시스템 및 방법이 필요하다.

발명의 내용

해결하려는 과제

[0011] 본 개시는 전술한 단점을 해결한다. 본 개시는, 적어도 부분적으로, 관심 생리학적 조건(예를 들어, 관심 표현형, 질환, 세포 상태 및/또는 세포 과정)에 상응하는 세포 구성성분 데이터(예를 들어, 유전자의 풍부도 및/또는 교란 시그니처)로, 그리고, 잠재 표현 및 기계 학습을 사용하여 세포 구성성분의 모듈(예를 들어, 서브세트)과 관심 생리학적 조건 사이의 연관성(예를 들어, 가중치 및/또는 상관)을 결정함으로써, 이들 단점을 해결한다. 특히, 본 개시는 다양한 생리학적 조건, 예컨대 질환의 근간이 되는 분자 메커니즘을 규명하기 위한 시스템 및 방법을 제공한다.

과제의 해결 수단

[0012] 본 개시의 한 양태는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 제공한다. 방법은 (A) 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계를 포함한다.

[0013] 방법은 (B) 세포 구성성분 모듈의 세트에 액세스하는 단계를 더 포함한다. 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 각각의 독립적인 서브세트를 포함한다. 복수의 세포 구성성분의 각각의 개별 독립적인 서브세트에 대한 상응하는 복수의 세포-기반 검정 풍부도 값이 생리학적 조건과 연관된 복수의 상이한 상태에 걸쳐 개별적으로 상관된다. 세포 구성성분 모듈의 세트의 제1 세포 구성성분 모듈은 관심 생리학적 조건과 연관된다.

[0014] 방법은 (C) 화학 구조의 지문을 모델 내로 입력하는 것에 응답하여, 모델로부터의 출력으로서, 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈에 대한 각각의 활성화 점수를 검색하는 단계를 더 포함한다. 일부 실시예에서, 모델은 50개 이상의 파라미터, 100개 이상의 파라미터, 1000개 이상의 파라미터, 또는 10,000개 이상의 파라미터를 포함한다.

[0015] 방법은 (D) 제1 세포 구성성분 모듈에 대한 활성화 점수가 제1 임계치 기준을 충족하는 경우에, 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 단계를 더 포함한다.

[0016] 일부 실시예에서, 세포-기반 검정 풍부도 값은 기관(organ)의 세포의 값이다. 일부 이러한 실시예에서, 기관은 심장, 간, 폐, 근육, 뇌, 췌장, 비장, 신장, 소장, 자궁 또는 방광이다.

[0017] 일부 실시예에서, 세포-기반 검정 풍부도 값은 조직의 세포의 값이다. 일부 실시예에서, 조직은 골, 연골, 관절, 기관(tracheae), 척수, 각막, 눈, 피부 또는 혈관이다.

[0018] 일부 실시예에서, 세포-기반 검정 풍부도 값은 복수의 줄기 세포의 세포의 값이다. 일부 실시예에서, 복수의 줄기 세포는 복수의 배아 줄기 세포, 복수의 성체 줄기 세포, 또는 복수의 유도된 만능 줄기 세포(iPSC)이다.

[0019] 일부 실시예에서, 세포-기반 검정 풍부도 값은 복수의 1차 인간 세포의 세포의 값이다. 일부 이러한 실시예에서, 복수의 1차 인간 세포는 복수의 CD34+ 세포, 복수의 CD34+ 조혈 줄기, 복수의 전구 세포(HSPC), 복수의 T-세포, 복수의 중간엽 줄기 세포(MSC), 복수의 기도 기저 줄기 세포 또는 복수의 유도된 만능 줄기 세포이다.

[0020] 일부 실시예에서, 세포-기반 검정 풍부도 값은 제대혈, 말초혈 또는 골수 내의 세포의 값이다.

[0021] 일부 실시예에서, 세포-기반 검정 풍부도 값은 고형 조직 내의 세포의 값이다. 일부 이러한 실시예에서, 고형 조직은 태반, 간, 심장, 뇌, 신장 또는 위장관이다.

[0022] 일부 실시예에서, 세포-기반 검정 풍부도 값은 복수의 분화된 세포의 값이다. 일부 이러한 실시예에서, 복수의 분화된 세포는 복수의 거핵구, 복수의 골모세포, 복수의 연골세포, 복수의 지방세포, 복수의 간세포, 복수의 간 중피 세포, 복수의 담관 상피 세포, 복수의 간 성상 세포, 복수의 간 시누소이드 내피 세포, 복수의 쿠퍼 세포, 복수의 피트 세포, 복수의 혈관 내피 세포, 복수의 췌장관 상피 세포, 복수의 췌장관 세포, 복수의 중심성 세포, 복수의 선방 세포, 복수의 랑게르한스섬, 복수의 심장 근육 세포, 복수의 섬유모세포, 복수의 각질세포, 복수의 평활근 세포, 복수의 제I형 폐포 상피 세포, 복수의 제II형 폐포 상피 세포, 복수의 클라라 세포, 복수의 섬모 상피 세포, 복수의 기저 세포, 복수의 배상 세포, 복수의 신경내분비 세포, 복수의 쿨치츠키(kultschitzky) 세포, 복수의 신세관 상피 세포, 복수의 요로상피 세포, 복수의 원주 상피 세포, 복수의 사구체 상피 세포, 복수의 사구체 내피 세포, 복수의 발세포, 복수의 사구체간 세포, 복수의 신경 세포, 복수의 교세포, 복수의 소교세포, 또는 복수의 핍지교세포이다.

[0023] 일부 실시예에서, 상응하는 복수의 세포-기반 검정 풍부도 값은 복수의 세포의 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq) 데이터이다. 일부 이러한 실시예에서, 생리학적 조건과 연관된 복수의 상이한 상태는 세포의 분취물이 해당 생리학적 조건에 영향을 미치는 것으로 공지된 화합물에 대한 노출에 대해 자유롭지 않은 대조군 상태에 더하여, 상이한 세포 분취물을 해당 생리학적 조건에 영향을 미치는 것으로 공지된 1종 이상의 참조 화합물에 노출시킴으로써 유도된다.

[0024] 일부 실시예에서, 상응하는 복수의 세포-기반 검정 풍부도 값은 벌크 RNA 시퀀스로부터의 값이다.

[0025] 일부 실시예에서, 상응하는 복수의 세포-기반 검정 풍부도 값은 단일 세포 RNA 시퀀싱으로부터의 값이다.

[0026] 일부 실시예에서, 세포 구성성분 모듈의 세트는 제1 세포 구성성분 모듈로 이루어진다.

[0027] 일부 실시예에서, 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈을 포함하고, 모델은 복수의 컴포넌트 모델을 포함하는 앙상블 모델이다. 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 화학 구조의 지문을 복수의 컴포넌트 모델의 각각의 컴포넌트 모델에 입력하는 것에 응답하여 세포 구성성분 모듈의 세트의 상이한 세포 구성성분 모듈에 대한 활성화 점수를 제공한다.

[0028] 일부 실시예에서, 방법은 테스트 화학적 화합물의 단순화된 분자-입력 라인-엔트리 시스템(SMILES) 스트링 표현으로부터 지문을 계산하는 단계를 더 포함한다.

[0029] 일부 실시예에서, 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 상응하는 신경망(예를 들어, 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합)이다. 일부 실시예에서, 상응하는 신경망은 상응하는 완전 연결 신경망 및 상응하는 메시지 전달 신경망의 조합이고, 상응하는 완전 연결 신경망의 제1 출력 및 상응하는 메시지 전달 신경망의 제2 출력은 조합되어, 화학 구조의 지문을 상응하는 완전 연결 신경망 및 상응하는 메시지 전달 신경망에 입력하는 것에 응답하여, 세포 구성성분 모듈의 세트에서 상응하는 세포 구성성분 모듈에 대한 하나 이상의 계산된 활성화 점수의 활성화 점수를 결정한다.

[0030] 일부 이러한 실시예에서, 복수의 컴포넌트 모델의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델이다.

[0031] 일부 실시예에서, 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈이고, 제1 세포 구성성분 모듈을 포함하는 복수의 세포 구성성분 모듈의 제1 서브세트는 관심 생리학적 조건과 연관되고, 복수의 세포 구성성분 모듈의 제2 서브세트는 관심 생리학적 조건과 연관되지 않고, 제1 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족하고, 복수의 세포 구성성분 모듈의 제2 서브세트의 세포 구성성분 모듈에 대

한 각각의 계산된 활성화 점수가 제1 임계치 기준 이외의 제2 임계치 기준을 충족하는 경우, 테스트 화학적 화합물이 관심 생리학적 조건으로 식별된다.

[0032] 일부 실시예에서, 방법은 제1 세포 구성성분 모듈을 식별하는 단계를 더 포함하고, 이 식별 단계는 전자 형태로 하나 이상의 제1 데이터세트를 획득하는 단계를 포함하는 과정에 의해 이루어지며, 상기 하나 이상의 제1 데이터세트는 제1 복수의 세포의 각각의 개별 세포에 대해(제1 복수의 세포는 20개 이상의 세포를 포함하고, 집합적으로 복수의 주석화된 세포 상태를 나타냄): 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해(예를 들어, 적어도 10, 20, 30, 100, 또는 1000 또는 그 이상의 세포 구성성분): 각각의 세포에서 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하고, 이에 의해 복수의 벡터에 액세스하거나 복수의 벡터를 형성하며, 복수의 벡터의 각각의 개별 벡터는 (i) 복수의 구성성분의 각각의 세포 구성성분에 상응하고 (ii) 상응하는 복수의 요소를 포함하고, 상응하는 복수의 요소의 각각의 개별 요소는 제1 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 갖는다. 방법은 복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계를 더 포함하고, 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈이 복수의 세포 구성성분의 서브세트를 포함하고, 이때 복수의 세포 구성성분 모듈은 (i) 복수의 후보 세포 구성성분 모듈 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현으로 배열되고, 복수의 세포 구성성분 모듈은 10개를 초과하는 세포 구성성분 모듈을 포함한다. 방법은 전자 형태로 하나 이상의 제2 데이터세트를 획득하는 단계를 더 포함하고, 하나 이상의 제2 데이터세트는 제2 복수의 세포의 각각의 개별 세포에 대해(제2 복수의 세포는 20개 이상의 세포를 포함하고 집합적으로 관심 생리학적 조건을 알리는 복수의 공변량을 나타냄): 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하며, 이에 의해, (i) 제2 복수의 세포 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 세포 구성성분 카운트 데이터 구조를 획득한다. 방법은 복수의 세포 구성성분 또는 그 표현을 공통 차원으로서 사용하여 세포 구성성분 카운트 데이터 구조 및 잠재 표현을 조합함으로써 활성화 데이터 구조를 형성하는 단계로서, 활성화 데이터 구조는, 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈에 대해, 제2 복수의 세포의 각각의 세포에 대해, 각각의 활성화 가중치를 포함하는, 상기 형성하는 단계; 및 복수의 공변량에서의 각각의 개별 공변량에 대해, (i) 후보 세포 구성성분 모델로의 공변량의 지문 입력 시 후보 세포 구성성분 모델에 의해 표현되는 각각의 세포 구성성분 모듈에 대해 계산된 활성화와 (ii) 후보 세포 구성성분 모델에 의해 표현되는 각각의 세포 구성성분 모듈에 대한 실제 활성화 사이의 차이를 사용하여 후보 세포 구성성분 모델을 훈련시키는 단계로서, 훈련시키는 단계는 차이에 응답하여 후보 세포 구성성분 모델과 연관된 복수의 공변량 파라미터를 조정하는, 상기 훈련시키는 단계를 더 포함한다. 일부 이러한 실시예에서, 복수의 공변량 파라미터는 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대해: 각각의 개별 공변량에 대해: 각각의 공변량이 제2 복수의 세포에 걸쳐 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 상응하는 파라미터를 포함하고; 방법은 후보 세포 구성성분 모델의 훈련시 복수의 공변량 파라미터를 사용하여, 복수의 후보 세포 구성성분 모듈 내의 제1 세포 구성성분 모듈을 식별하는 단계를 더 포함한다. 일부 이러한 실시예에서 방법은 복수의 주석화된 세포 상태의 주석화된 세포 상태가 노출 조건 하에 제1 복수의 세포의 세포의 화합물에 대한 노출(예를 들어, 노출 지속기간, 화합물의 농도, 또는 노출 지속기간과 화합물의 농도의 조합)인 것을 더 포함한다.

[0033] 일부 실시예에서, 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합이다.

[0034] 일부 실시예에서, 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합이고, 제1 또는 제2 복수의 세포의 각각의 세포 중의 각각의 세포 구성성분의 상응하는 풍부도는 비색 측정치, 형광 측정치, 발광 측정치 또는 공명 에너지 전달 (FRET) 측정치에 의해 결정된다.

[0035] 일부 실시예에서, 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합이고, 제1 또는 제2 복수의 세포의 각각의 세포 중의 각각의 세포 구성성분의 상응하는 풍부도는 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq), scTag-seq, 시퀀싱을 사용한 전위효소-접근 가능 염색질에 대한 단일-세포 검정(scATAC-seq), CyTOF/SCoP, E-MS/Abseq, miRNA-seq, CITE-seq 또는 그 임의의 조합에 의해 결정된다.

[0036] 일부 실시예에서, 복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계는 복수의 벡터의 각각의 벡터의 각각의 상응하는 복수의 요소를 사용하여 상관 모델을 복수의

벡터에 적용하는 것을 포함한다. 일부 이러한 실시예에서, 상관 모델은 그래프 클러스터링(예를 들어, 피어슨-상관관계-기반 거리 메트릭에 대한 라이덴 클러스터링, 루뱅 클러스터링 등)을 포함한다.

[0037] 일부 실시예에서, 복수의 세포 구성성분 모듈은 10개 내지 2000개의 세포 구성성분 모듈, 또는 100개 내지 8,000개의 세포 구성성분으로 이루어진다. 일부 실시예에서, 복수의 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 2개 내지 300개의 세포 구성성분으로 이루어진다.

[0038] 일부 실시예에서, 관심 생리학적 조건은 질환이다.

[0039] 일부 실시예에서, 관심 생리학적 조건은 질환이고, 제1 복수의 세포는 복수의 주석화된 세포 상태에 의해 표시되는 바와 같은, 질환을 대표하는 세포 및 질환을 대표하지 않는 세포를 포함한다.

[0040] 일부 실시예에서, 복수의 공변량은 세포 배치(batch), 세포 공여자, 세포 유형, 질환 상태, 화학적 화합물에 대한 노출, 또는 그 임의의 조합을 포함한다.

[0041] 일부 실시예에서, 후보 세포 구성성분 모델을 훈련시키는 것은 멀티-태스크 공식화에서 범주형 교차-엔트로피 손실을 사용하여 수행되며, 여기서 복수의 공변량에서의 각각의 공변량은 복수의 비용 함수에서의 비용 함수에 상응하고, 복수의 비용 함수에서의 각각의 개별 비용 함수는 공통 가중 인자를 갖는다.

[0042] 일부 실시예에서, 테스트 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물이다. 일부 이러한 실시예에서, 테스트 화학적 화합물은 리핀스키 5 준칙 각각을 충족시키는 유기 화합물이다. 일부 실시예에서, 테스트 화학적 화합물은 리핀스키 5 준칙의 적어도 3가지 기준을 충족하는 유기 화합물이다. 일부 실시예에서, 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0043] 일부 실시예에서, 방법은 데이라이트(Daylight), BCI, ECFP4, EcFC, MDL, APFP, TTFP, UNITY 2D 지문, RNNS2S 또는 GraphConv를 사용하여 테스트 화학적 화합물의 화학 구조로부터 지문을 생성하는 것을 더 포함한다.

[0044] 일부 실시예에서, 세포 구성성분 모듈의 세트는 5개 이상의 세포 구성성분 모듈, 10개 이상의 세포 구성성분 모듈, 또는 100개 이상의 세포 구성성분 모듈을 포함한다.

[0045] 일부 실시예에서, 각각의 세포 구성성분 모듈 내의 복수의 세포 구성성분의 독립적인 서브세트는 5개 이상의 세포 구성성분을 포함한다.

[0046] 일부 실시예에서, 각각의 세포 구성성분 모듈 내의 복수의 세포 구성성분의 독립적인 서브세트는 관심 생리학적 조건과 연관된 분자 경로 내의 2개 내지 20개의 세포 구성성분으로 이루어진다.

[0047] 일부 실시예에서, 제1 임계치 기준은 제1 세포 구성성분 모듈이 임계 활성화 점수를 가져야 한다는 요건이다.

[0048] 본 개시의 또 다른 양태는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 제공한다.

[0049] 방법은 (A) 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계를 포함한다.

[0050] 방법은 (B) 교란 시그니처의 세트에 액세스하는 단계를 더 포함하고, 여기서 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 복수의 세포 구성성분의 각각의 독립적인 서브세트를 포함하고, 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별, 및 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하고, 여기서 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태이다.

[0051] 방법은 (C) 지문을 모델에 입력하는 단계를 더 포함하며, 여기서 모델은 50, 100, 500, 1000, 또는 10,000 또는 그 이상의 파라미터를 포함하고, 모델은 모델로의 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력하고, 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 교란 시그니처의 세트의 상응하는 교란 시그니처를 나타낸다.

[0052] 방법은 (D) 교란 시그니처의 세트의 제1 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족할 때 화학적 화합물을 관심 생리학적 조건과 연관시키는 단계를 더 포함한다.

[0053] 일부 실시예에서, 방법은 테스트 화학적 화합물의 단순화된 분자-입력 라인-엔트리 시스템(SMILES) 스트링 표현

으로부터 지문을 계산하는 단계를 더 포함한다.

[0054] 일부 실시예에서, 모델은 신경망을 포함한다. 일부 이러한 실시예에서, 신경망은 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합이다.

[0055] 일부 실시예에서, 모델은 복수의 컴포넌트 모델을 포함하는 앙상블 모델이고, 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 화학 구조의 지문을 복수의 컴포넌트 모델 세트에서의 각각의 컴포넌트 모델에 입력하는 것에 응답하여 교란 시그니처의 세트에서의 상이한 교란 시그니처에 대한 활성화 점수를 제공한다.

[0056] 일부 실시예에서, 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0057] 일부 실시예에서, 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 상응하는 신경망이다(예를 들어, 상응하는 신경망은 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합임).

[0058] 일부 실시예에서, 복수의 컴포넌트 모델의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델이다.

[0059] 일부 실시예에서, 상응하는 신경망은 완전 연결 신경망 및 메시지 전달 신경망의 조합이고, 화학 구조의 지문을 완전 연결 신경망 및 메시지 전달 신경망에 입력하는 것에 응답하여, 제1 신경망의 제1 출력 및 제2 신경망의 제2 출력이 조합되어, 교란 시그니처의 세트의 제1 교란 시그니처에 대한 하나 이상의 계산된 활성화 점수의 활성화 점수를 결정한다.

[0060] 일부 실시예에서, 교란 시그니처의 세트는 복수의 교란 시그니처이고, 제1 교란 시그니처를 비롯한 복수의 교란 시그니처의 제1 서브세트는 관심 생리학적 조건과 연관되고, 복수의 교란 시그니처의 제2 서브세트는 관심 생리학적 조건과 연관되지 않고, 제1 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족하고, 복수의 교란 시그니처의 제2 서브세트에서의 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준 이외의 제2 임계치 기준을 충족하는 경우에, 테스트 화학적 화합물은 관심 생리학적 조건으로 식별된다.

[0061] 일부 실시예에서, 관심 생리학적 조건은 질환이다.

[0062] 일부 실시예에서, 테스트 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물이다.

[0063] 일부 실시예에서, 테스트 화학적 화합물은 리핀스키 5 준칙 각각을 충족시키는 유기 화합물이다. 일부 이러한 실시예에서, 테스트 화학적 화합물은 리핀스키 5 준칙의 적어도 3가지 기준을 충족하는 유기 화합물이다.

[0064] 일부 실시예에서, 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0065] 일부 실시예에서, 방법은 데이라이트(Daylight), BCI, ECFP4, EcFC, MDL, APFP, TTFP, UNITY 2D 지문, RNNS2S 또는 GraphConv를 사용하여 테스트 화학적 화합물의 화학 구조로부터 지문을 생성하는 것을 더 포함한다.

[0066] 일부 실시예에서, 교란 시그니처의 세트는 제1 교란 시그니처로 이루어진다.

[0067] 일부 실시예에서, 교란 시그니처의 세트는 5개 이상의 교란 시그니처, 10개 이상의 교란 시그니처 또는 100개 이상의 교란 시그니처를 포함한다.

[0068] 일부 실시예에서, 제1 임계치 기준은 제1 교란 시그니처가 임계 활성화 점수를 가져야 한다는 요건이다.

[0069] 본 개시의 또 다른 양태는 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 제공한다.

[0070] 방법은 메모리 및 하나 이상의 프로세서를 포함하는 컴퓨터 시스템에서: (A) 전자 형태로, 복수의 화합물의 각각의 개별 화합물의 상응하는 화학 구조의 각각의 지문을 획득함으로써, 복수의 지문을 획득하는 것을 포함한다.

[0071] 방법은 (B) 전자 형태로, 복수의 화합물의 각각의 화합물에 대한 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수를 획득하는 단계를 더 포함하고, 여기서 세포 구성성분 모듈의 세트의

각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함한다.

[0072] 방법은 (C) 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해, (i) 각각의 화합물의 화학 구조의 지문을 훈련되지 않은 모델로 입력시 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수 및 (ii) 세포 구성성분 모듈의 세트의 각각의 화합물에 대한 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련되지 않은 모델을 훈련시키는 단계를 더 포함하고, 여기서 훈련 단계 (C)는 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정함으로써(복수의 파라미터는 50, 100, 200, 500, 1000, 또는 10,000 또는 그 이상의 파라미터를 포함함), 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득한다.

[0073] 일부 실시예에서, 세포 구성성분 모듈의 세트는 단일 세포 구성성분 모듈로 이루어진다.

[0074] 일부 실시예에서, 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈을 포함한다.

[0075] 일부 실시예에서, 세포 구성성분 모듈의 세트는 2개 내지 500개의 세포 구성성분 모듈로 이루어진다.

[0076] 일부 실시예에서, 복수의 화합물은 10개 내지 1×10^6개의 화합물로 이루어진다.

[0077] 일부 실시예에서, 복수의 화합물은 100개 내지 100,000개의 화합물로 이루어진다.

[0078] 일부 실시예에서, 복수의 화합물은 1000개 내지 100,000개의 화합물로 이루어진다.

[0079] 일부 실시예에서, 훈련 단계 (C)는 회귀 알고리즘에 따라 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대한 각각의 개별 화합물과 연관된 각각의 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정한다. 일부 이러한 실시예에서, 회귀 알고리즘은 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대한 각각의 개별 화합물과 연관된 각각의 차이의 최소 제곱 오차를 최적화한다.

[0080] 일부 실시예에서, 훈련된 모델은 신경망(예를 들어, 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합)을 포함한다.

[0081] 일부 실시예에서, 훈련된 모델은 복수의 컴포넌트 모델의 앙상블 모델이고, 복수의 컴포넌트 모델의 각각의 개별 컴포넌트 모델은 복수의 세포 구성성분 모듈의 상이한 세포 구성성분 모듈에 대한 계산된 활성화 점수를 출력한다. 일부 이러한 실시예에서, 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0082] 일부 실시예에서, 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 상응하는 신경망이다. 일부 이러한 실시예에서, 상응하는 신경망은 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합이다.

[0083] 일부 실시예에서, 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈이고, 복수의 세포 구성성분 모듈의 제1 서브세트는 관심 생리학적 조건과 연관되고, 복수의 세포 구성성분 모듈의 제2 서브세트는 관심 생리학적 조건과 연관되지 않는다.

[0084] 일부 실시예에서, 방법은 전자 형태로 하나 이상의 제1 데이터세트를 획득하는 단계를 포함하는 과정에 의해 복수의 세포 구성성분 모듈의 세포 구성성분 모듈을 식별하는 단계를 더 포함하며, 하나 이상의 제1 데이터세트는 제1 복수의 세포의 각각의 개별 세포에 대해(여기서, 제1 복수의 세포는 20개 이상의 세포를 포함하고, 복수의 주석화된 세포 상태를 집합적으로 나타냄), 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해(여기서, 복수의 세포 구성성분은 5, 10, 15, 20, 25, 50, 또는 100개 또는 그 이상의 세포 구성성분을 포함함), 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 또는 집합적으로 포함하고, 이에 의해, 복수의 벡터에 액세스하거나 복수의 벡터를 형성한다. 복수의 벡터의 각각의 개별 벡터는 (i) 복수의 구성성분의 각각의 세포 구성성분에 상응하고, (ii) 상응하는 복수의 요소를 포함한다. 상응하는 복수의 요소의 각각의 개별 요소는 제1 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 갖는다. 복수의 벡터는 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는데 사용되고, 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 복수의 세포 구성성분의 서브세트를 포함한다. 복수의 세포 구성성분 모듈은 (i) 복수의 후보 세포 구성성분 모듈 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현으로 배열되고, 복수의 세포 구성성분 모듈은 3, 5, 10, 15, 20, 또는 100개를 초과하는 세포 구성성분 모듈을 포함한다. 하나 이상의 제2 데이터세트가 전자 형태로 획득되고, 이러한 하나 이상의 제2 데이터세트는 제2 복수의 세포의 각각의 개별 세포에 대해(제2 복수의 세포는 20

개 이상의 세포를 포함하고 집합적으로 관심 생리학적 조건을 알리는 복수의 공변량을 나타냄), 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하며, 이에 의해, (i) 제2 복수의 세포 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 세포 구성성분 카운트 데이터 구조를 획득한다. 활성화 데이터 구조는 복수의 세포 구성성분 또는 그 표현을 공통 차원으로서 사용하여 세포 구성성분 카운트 데이터 구조 및 잠재 표현을 조합함으로써 형성되고, 여기서 활성화 데이터 구조는 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈에 대해: 제2 복수의 세포의 각각의 세포에 대해, 각각의 활성화 가중치를 포함한다. 후보 세포 구성성분 모델은 (i) 후보 모델로의 활성화 데이터 구조의 입력 시 활성화 데이터 구조에 표현된 각각의 세포 구성성분 모듈에서의 복수의 공변량 중 각각의 공변량의 부재 또는 존재의 예측과 (ii) 각각의 세포 구성성분 모듈에서의 각각의 공변량의 실제 부재 또는 존재 사이의 차이를 사용하여 훈련된다. 이러한 훈련은 차이에 반응하는 후보 세포 구성성분 모델과 연관된 복수의 공변량 파라미터를 조정한다.

[0085] 일부 실시예에서, 복수의 공변량 파라미터는 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대해: 각각의 개별 공변량에 대해: 각각의 공변량이 제2 복수의 세포에 걸쳐 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 상응하는 파라미터를 포함하고, 후보 세포 구성성분 모델의 훈련시 복수의 공변량 파라미터를 사용하여 복수의 후보 세포 구성성분 모듈 중 세포 구성성분 모듈이 식별된다.

[0086] 일부 실시예에서, 복수의 주석화된 세포 상태의 주석화된 세포 상태는 노출 조건 하에 화합물에 대한 제1 복수의 세포의 세포의 노출이다.

[0087] 일부 실시예에서, 노출 조건은 노출 지속기간, 화합물의 농도, 또는 노출 지속기간과 화합물의 농도의 조합이다.

[0088] 일부 실시예에서, 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합이다.

[0089] 일부 실시예에서, 제1 또는 제2 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도는 비색 측정치, 형광 측정치, 발광 측정치, 또는 공명 에너지 전달(FRET) 측정치에 의해 결정된다.

[0090] 일부 실시예에서, 제1 또는 제2 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도는 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq), scTag-seq, 시퀀싱을 사용하는 전위효소-접근 가능 염색질에 대한 단일-세포 검정(scATAC-seq), CyTOF/SCoP, E-MS/Abseq, miRNA-seq, CITE-seq, 또는 그 임의의 조합에 의해 결정된다.

[0091] 일부 실시예에서, 복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계는 복수의 벡터의 각각의 벡터의 각각의 상응하는 복수의 요소를 사용하여 상관 모델을 복수의 벡터에 적용하는 것을 포함한다. 일부 이러한 실시예에서, 상관 모델은 그래프 클러스터링(예를 들어, 피어슨-상관관계-기반 거리 메트릭에 대한 라이덴 클러스터링 또는 루뱅 클러스터링)을 포함한다.

[0092] 일부 실시예에서, 복수의 세포 구성성분은 100개 내지 8,000개의 세포 구성성분으로 이루어진다.

[0093] 일부 실시예에서, 복수의 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 2개 내지 300개의 세포 구성성분으로 이루어진다.

[0094] 일부 실시예에서, 관심 생리학적 조건은 질환이다.

[0095] 일부 실시예에서, 생리학적 조건은 질환이고, 제1 복수의 세포는 복수의 주석화된 세포 상태에 의해 표시되는 바와 같은, 질환을 대표하는 세포 및 질환을 대표하지 않는 세포를 포함한다.

[0096] 일부 실시예에서, 복수의 공변량은 세포 배치, 세포 공여자, 세포 유형, 질환 상태 또는 화학적 화합물에 대한 노출을 포함한다.

[0097] 일부 실시예에서, 후보 세포 구성성분 모델을 훈련시키는 것은 멀티-태스크 공식화에서 범주형 교차-엔트로피 손실을 사용하여 수행되며, 여기서 복수의 공변량에서의 각각의 공변량은 복수의 비용 함수에서의 비용 함수에 상응하고, 복수의 비용 함수에서의 각각의 개별 비용 함수는 공통 가중 인자를 갖는다.

[0098] 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물이다.

[0099] 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 리핀스키 5 준칙 각각을 충족시킨다. 일부 이러한 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 리핀스키 5 준칙의 적어도 3가지 기준을 충족시킨다.

[0100] 일부 실시예에서, 훈련된 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0101] 일부 실시예에서, 방법은 데이라이트, BCI, ECFP4, EcFC, MDL, APFP, TTFP, UNITY 2D 지문, RNNS2S 또는 GraphConv를 사용하여 상응하는 화학 구조로부터 각각의 개별 지문을 생성하는 것을 더 포함한다.

[0102] 일부 실시예에서, 세포 구성성분 모듈의 세트는 5개 이상의 세포 구성성분 모듈, 10개 이상의 세포 구성성분 모듈, 또는 100개 이상의 세포 구성성분 모듈을 포함한다.

[0103] 본 개시의 또 다른 양태는 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 제공한다. 방법은 예를 들어 메모리 및 하나 이상의 프로세서를 포함하는 컴퓨터 시스템에서 수행될 수 있다.

[0104] 방법은 (A) 전자 형태로, 복수의 화합물의 각각의 개별 화합물의 상응하는 화학 구조의 각각의 지문을 획득함으로써, 복수의 지문을 획득하는 것을 포함한다.

[0105] 방법은 (B) 전자 형태로, 복수의 화합물의 각각의 상응하는 화합물에 대한 교란 시그니처의 세트의 각각의 개별 교란 시그니처의 각각의 수치 활성화 점수를 획득하는 단계를 더 포함하고, 여기서 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별, 및 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함한다. 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 유발된 각각의 교란된 세포 상태이다.

[0106] 방법은 (C) 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해, (i) 각각의 화합물의 화학 구조의 지문을 훈련되지 않은 모델로 입력 시 각각의 교란 시그니처에 대한 각각의 계산된 활성화 점수 및 (ii) 교란 시그니처의 세트에서 상응하는 화합물에 대한 각각의 교란 시그니처의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련되지 않은 모델을 훈련시키는 것을 더 포함한다. 훈련 단계 (C)는 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하여, 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득한다. 일부 실시예에서, 복수의 파라미터는 50, 100, 200, 500, 1000, 10,000, 또는 1×10^6 또는 그 이상의 파라미터를 포함한다.

[0107] 일부 실시예에서, 교란 시그니처의 세트는 단일 교란 시그니처로 이루어진다.

[0108] 일부 실시예에서, 교란 시그니처의 세트는 2개 내지 500개의 교란 시그니처로 이루어진다.

[0109] 일부 실시예에서, 복수의 화합물은 10개 내지 1×10^6개의 화합물로 이루어진다. 일부 실시예에서, 복수의 화합물은 100개 내지 100,000개의 화합물로 이루어진다. 일부 실시예에서, 복수의 화합물은 1000개 내지 100,000개의 화합물로 이루어진다.

[0110] 일부 실시예에서, 훈련 단계 (C)는 회귀 알고리즘에 따라 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해 각각의 상응하는 화합물과 연관된 각각의 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정한다. 일부 이러한 실시예에서, 회귀 알고리즘은 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해 각각의 상응하는 화합물과 연관된 각각의 차이의 최소 제곱 오차를 최적화한다.

[0111] 일부 실시예에서, 훈련된 모델은 신경망(예를 들어, 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합)을 포함한다.

[0112] 일부 실시예에서, 훈련된 모델은 복수의 컴포넌트 모델의 앙상블 모델이고, 복수의 컴포넌트 모델의 각각의 개별 컴포넌트 모델은 각각의 화학 구조의 지문을 복수의 컴포넌트 모델의 세트에서의 각각의 컴포넌트 모델에 입력하는 것에 응답하여 복수의 교란 시그니처의 세트에서의 상이한 교란 시그니처의 세트에 대한 계산된 활성화 점수를 출력한다. 일부 이러한 실시예에서, 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정

트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0113] 일부 실시예에서, 복수의 컴포넌트 모델의 각각의 컴포넌트 모델은 상응하는 신경망(예를 들어, 완전 연결 신경망, 메시지 전달 신경망, 또는 이들의 조합)이다.

[0114] 일부 실시예에서, 교란 시그니처의 세트는 복수의 교란 시그니처를 포함하고, 복수의 교란 시그니처의 제1 서브세트는 관심 생리학적 조건과 연관되고, 복수의 교란 시그니처의 제2 서브세트는 관심 생리학적 조건과 연관되지 않는다.

[0115] 일부 실시예에서, 관심 생리학적 조건은 질환이다.

[0116] 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물이다.

[0117] 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 리핀스키 5 준칙 각각을 충족시킨다.

[0118] 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 리핀스키 5 준칙의 적어도 3가지 기준을 충족시킨다.

[0119] 일부 실시예에서, 훈련된 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0120] 일부 실시예에서, 방법은 데이라이트, BCI, ECFP4, EcFC, MDL, APFP, TTFP, UNITY 2D 지문, RNNS2S 또는 GraphConv를 사용하여 상응하는 화학 구조로부터 각각의 개별 지문을 생성하는 것을 더 포함한다.

[0121] 일부 실시예에서, 교란 시그니처의 세트는 5개 이상의 교란 시그니처, 10개 이상의 교란 시그니처, 또는 100개 이상의 교란 시그니처를 포함한다.

[0122] 일부 실시예에서, 방법은 변경되지 않은 세포 상태와 변경된 세포 상태 사이의 차등 세포 구성성분 풍부도의 척도를 나타내는 단일-세포 전이 시그니처에 전자 형태로 액세스하는 것을 포함하는 절차에 의해 교란 시그니처의 세트의 각각의 교란 시그니처의 각각의 수치 활성화 점수를 획득하는 단계를 더 포함하고, 여기서 변경된 세포 상태는 변경되지 않은 세포 상태로부터 변경된 세포 상태로의 세포 전이를 통해 발생하고, (i) 변경되지 않은 세포 상태, (ii) 변경된 세포 상태, 및 (iii) 변경되지 않은 세포 상태로부터 변경된 세포 상태로의 전이 중 적어도 하나는 관심 생리학적 조건과 연관되고, 단일-세포 전이 시그니처는 기준 복수의 세포 구성성분의 식별을 포함하고, 복수의 기준 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 변경되지 않은 세포 상태와 변경된 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 제1 유의성 점수를 포함한다. 추가로, 단일-세포 전이 시그니처 및 각각의 교란 시그니처를 비교하고 그에 의해 각각의 교란 시그니처의 각각의 수치 활성화 점수를 결정한다.

[0123] 일부 실시예에서, 단일-세포 전이 시그니처 및 교란 시그니처를 비교하여 각각의 교란 시그니처의 각각의 수치 활성화 점수를 결정하는 것은, 단일-세포 전이 시그니처의 복수의 기준 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 제1 유의성 점수를 각각의 교란 시그니처의 상응하는 세포 구성성분의 상응하는 유의성 점수와 비교하는 것을 포함한다.

[0124] 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 교란 시그니처의 세트의 다른 교란 시그니처에 대한 각각의 교란 시그니처의 단일-세포 전이 시그니처에 대한 관련성의 상대 순위이다.

[0125] 일부 실시예에서, 상대 순위는 윌콕슨 순위-합계 검정, t-검정, 로지스틱 회귀 또는 일반화된 선형 모델에 의해 결정된다.

[0126] 일부 실시예에서, 단일-세포 전이 시그니처의 변경되지 않은 세포 상태는 각각의 교란 시그니처의 제1 세포 상태 또는 제2 세포 상태와 동일하다.

[0127] 일부 실시예에서, 단일-세포 전이 시그니처의 변경되지 않은 세포 상태는 각각의 교란 시그니처의 제1 세포 상태 및 제2 세포 상태 둘 다와 상이하다.

[0128] 일부 실시예에서, 방법은 단일-세포 전이 시그니처의 기준 복수의 세포 구성성분 및 각각의 교란 시그니처의 각각의 복수의 세포 구성성분을 프루닝(pruning)하여 전사 인자에 대한 비교를 제한하는 것을 더 포함한다.

[0129] 일부 실시예에서, 복수의 교란 시그니처에서의 각각의 교란 시그니처의 교란된 세포 상태는 복수의 화합물 중의

화합물에 노출되지 않은 대조군 세포에 의해 나타내어진다.

[0130] 일부 실시예에서, 복수의 교란 시그니처의 각각의 교란 시그니처의 교란된 세포 상태는 각각의 교란 시그니처와 과 연관된 화합물 이외의 복수의 화학적 화합물의 화학적 화합물에 노출된 관련되지 않은 교란된 세포에 걸친 평균에 의해 나타내어진다.

[0131] 개시된 실시예 중 일부에서, 모델은 리그레서(regressor)이다.

[0132] 본 개시의 또 다른 양태는 하나 이상의 프로세서, 및 하나 이상의 프로세서에 의한 실행을 위한 하나 이상의 프로그램을 저장하는 메모리를 갖는 컴퓨터 시스템을 제공하고, 하나 이상의 프로그램은 본원에 개시된 임의의 방법 및/또는 실시예를 수행하기 위한 명령어를 포함한다.

[0133] 본 개시의 또 다른 양태는 컴퓨터에 의해 실행되도록 구성된 하나 이상의 프로그램을 저장하는 비-일시적 컴퓨터 판독가능 저장 매체를 제공하고, 하나 이상의 프로그램은 본원에 개시된 방법 및/또는 실시예 중 임의의 것을 수행하기 위한 명령어를 포함한다.

[0134] 본 개시의 추가의 양태 및 이점은 하기 상세한 설명으로부터 관련 기술분야의 통상의 기술자에게 용이하게 명백해질 것이며, 여기서 단지 본 개시의 예시적 실시예가 제시되고 설명된다. 실현되는 바와 같이, 본 개시는 다른 및 상이한 실시예가 가능하며, 그의 여러 세부사항은 모두 개시로부터 벗어나지 않으면서 다양한 명백한 양태에서 변형이 가능하다. 따라서, 도면 및 설명은 제한적인 것이 아니라, 사실상 예시적인 것으로 간주되어야 한다.

도면의 간단한 설명

[0135] 본원에 개시된 실시예는 첨부 도면의 도면에서 제한이 아닌 예로서 예시된다. 유사한 참조 번호는 도면 전반에 걸쳐 상응하는 부분을 지칭한다.

도 1은 본 개시의 한 실시예에 따른 예시적인 시스템 및 컴퓨팅 장치의 블록도를 예시한다.

도 2a 및 2b는 본 개시의 다양한 실시예에 따른, 복수의 세포 구성성분을 관심 생리학적 조건과 연관시키기 위한 예시적인 방법의 과정 및 특징의 흐름도를 집합적으로 제공한다.

도 3a, 3b, 3c, 3d, 및 3e는 본 개시의 다양한 실시예에 따른, 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한 예시적인 방법의 과정 및 특징의 흐름도를 제공하며, 여기서 파선 박스는 선택적인 요소를 나타낸다.

도 4는 본 개시의 일부 실시예에 따른, 세포 구성성분 모듈의 복수의 벡터의 예 및 세포 구성성분의 잠재 표현의 예를 예시한다.

도 5는 본 개시의 일부 실시예에 따른, 세포 구성성분 카운트 데이터 구조의 예 및 예시적인 활성화 데이터 구조를 예시한다.

도 6은 본 개시의 일부 실시예에 따른, 복수의 화합물 가중치를 조정하기 위해 모델을 훈련시키는 방법의 예를 예시한다.

도 7은 본 개시의 일부 실시예에 따른, 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한 예시적인 방법의 과정 및 특징의 흐름도를 제공하며, 여기서 파선 박스는 선택적인 요소를 나타낸다.

도 8은 본 개시의 한 실시예에 따른, 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한 예시적인 방법의 과정 및 특징의 흐름도를 제공하며, 여기서 파선 박스는 선택적인 요소를 나타낸다.

도 9는 본 개시의 한 실시예에 따른, 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한 예시적인 방법의 과정 및 특징의 흐름도를 제공하며, 여기서 파선 박스는 선택적인 요소를 나타낸다.

도 10a, 10b, 10c, 10d 및 10e는 본 개시의 실시예에 따른, 지방산-관련 세포 프로그램의 활성화에 대한 화학 구조를 예측하기 위한 예시적인 방법의 성능 및 4-중첩 검증을 예시한다. 도 10a는 화학 구조를 예측하기 위한 모델 아키텍처의 개략도를 예시한다. 도 10b는 1,200개의 무작위로 선택된 화합물의 테스트 세트에 대한 성능을 예시한다. 도 10c는 훈련 세트와 상이한 스캐폴드를 갖는 1,200개의 화합물의 테스트 세트에 대한 성능을 예시한다. 도 10d는 시험관내 지방전구세포 검정에서 전사 활성화에 기초한 갈색지방-관련 모듈에 대한 검증을 예시한다. 도 10e는 표적 모듈에 대한 5백만개 화합물의 데이터베이스로부터 도출된 예측 화합물의 최적화를

예시한다.

도 11은 본 개시의 한 실시예에 따른, 태아 적혈구생성 및 T-세포 소진과 관련된 세포 거동의 활성화에 대한 화학 구조를 예측하기 위한 예시적인 방법의 검증을 예시한다.

도 12는 본 개시의 실시예에 따라 단일 세포 RNA 시퀀싱(scRNA-seq)을 이용하여 인간 지방전구세포 유전자 모듈 활성화에 대한 공지된 피페리딘-함유 화합물("KPCC") 및 6개의 새로 합성된 히트 "합성 히트"의 영향을 평가하기 위한 예시적인 방법의 개략도를 예시한다.

도 13은 본 개시의 실시예에 따른, 원하는 전사 변화의 활성화에 대한 KPCC 및 6개의 합성 히트의 효과를 예시한다.

도 14a, 14b, 14c 및 14d는 선택적인 요소가 파선 박스로 표시된 세포 구성성분 모듈을 식별하는 흐름도를 제공한다.

발명을 실시하기 위한 구체적인 내용

[0136] *서론.*

[0137] 상기 배경을 고려하여, 본 개시는 질환에 중요한 세포 과정 및 프로그램을 표적화하는 약물 발견에 대한 접근법을 설명한다. 이러한 접근법은, 일부 양태에서, 생리학적 조건(예를 들어, 세포 프로그램, 세포 과정, 및/또는 세포 상태) 및 화합물의 화학 구조의 계산적 조작 표현을 사용하여 화학 구조-관련 양식 및 그의 특성을 예측함으로써 실현된다. 인코딩된 화학 구조는 이어서 세포 프로그램 및/또는 세포 상태의 표현 상에 맵핑되고, 따라서 화합물을 생리학적 조건과 연관시킬 수 있다.

[0138] 예를 들어, 일부 양태에서, 본 개시는 분자 프로파일(예를 들어, 유전자 모듈)과 관심 생물학적 과정(예를 들어, 세포 프로그램 및/또는 세포 상태) 및 화합물의 화학 구조 사이의 연관성을 획득하기 위한 시스템 및 방법을 제공한다. 이들 연관은 약물 발견을 위해 새로운 화학 구조, 예컨대 유사한 기능적 또는 구조적 특성을 갖는 것을 예측하는데 사용될 수 있다.

[0139] 일부 실시예에서, 예측 능력을 갖는 컴퓨터 모델링 아키텍처가 하나 이상의 도메인 및/또는 데이터 유형에 걸쳐 생리학상 관련된 화학 구조의 잠재 표현의 생성을 통해 이들 연관성을 발견하는데 사용된다. 연관은, 예를 들어 세포 거동의 프로파일, 예컨대 1종 이상의 화합물에 대한 세포의 노출에 반응한 차등 유전자 발현 또는 세포 상태 전이를 제공하는 교란 데이터로부터 유도될 수 있다. 일부 구현에서, 방법은 다양한 도메인(예를 들어, 분자, 세포, 임상, 생체내, 시험관내, 지식-기반 등) 및/또는 다양한 데이터 유형(잠재 표현 및 기계 학습을 사용하여 전사, 유전적, 후생적, 공변량 등) 사이의 상관을 조합하고 결정하여 생리학상 관련된 화학 구조를 예측한다.

[0140] 예시적인 실시예에서, 본 개시는 화합물에 대한 잠재 표현을 사용하는 모델링 접근법을 제공한다. 복수의 화합물의 각각의 개별 화합물에 대해, 방법은 각각의 화합물이 복수의 생리학적 조건에서 각각의 생리학적 조건을 유도할 가능성을 나타내는 벡터를 저장하는 잠재 표현을 생성하는 것을 포함한다. 생리학적 조건은 특정한 표현형, 세포 과정 및/또는 질환과 연관된 세포 상태 전이 및/또는 세포 구성성분 모듈(예를 들어, 유전자 모듈)을 포함할 수 있다. 따라서, 방법은 화합물 및 생리학적 조건(예를 들어, 세포 상태 및/또는 유전자 모듈)에 의해 차원화되고, 예를 들어 *n_compounds x n_cell_states* 또는 *n_compounds x n_gene_modules*로 표시되는, 모델에 대한 멀티-태스크 훈련 표지로서의 역할을 하는 행렬 표현을 생성한다.

[0141] 화합물을 생리학적 조건과 연관시키기 위한 기계 학습 모델에 대한 입력은 화합물의 화학 구조를 인코딩하는 각각의 화합물의 정규 이성질체 SMILES 표현 및/또는 그래프-기반 표현을 포함하고, 모델을 훈련시키는데 추가로 사용된다. 훈련 표지는 각각의 화합물을 각각의 생리학적 조건과 연관시키는 수치 활성화 점수로서 제공된다. 예를 들어, 각각의 화합물에 대한 벡터는 복수의 연관된 가중치를 포함할 수 있으며, 여기서 각각의 가중치는 화합물이 각각의 생리학적 조건, 예컨대 각각의 세포 상태, 세포 상태 전이, 교란 시그니처 및/또는 각각의 유전자 모듈의 활성화를 유도할 가능성을 나타낸다.

[0142] 입력으로서 행렬 표현을 수신할 때, 모델은 회귀 문제를 해결함으로써 화학 구조로부터 세포 상태(예를 들어, 교란 시그니처) 및/또는 유전자 모듈 활성화를 학습하도록 훈련된다. 회귀 문제를 해결하기 위해 2가지 예시적인 모델 아키텍처가 사용된다. 제1 모델은 SMILES 스트링의 표준 지문에 대해 완전히 연결된 네트워크를 이용하며, 여기서 네트워크 아키텍처는 ReLU 활성화를 갖는 3-계층 네트워크이다. 제2 모델은 DGL 라이브러리로부

터의 MPNN 네트워크를 포함한다. 이들 모델 각각은 회귀 예측의 최소 제곱 오차를 최적화함으로써 서로 독립적으로 훈련된다. 테스트 시간에, 이들 모델의 예측이 평균화되며, 따라서, 제1 및 제2 모델을 포함하는 앙상블 모델을 형성한다. 이어서, 앙상블 모델은 화합물과 생리학적 조건 사이의 연관성을 결정하는데 사용될 수 있으며, 이는 화학 구조로부터의 가능한 생리학적 활성화의 예측 및/또는 특정한 생리학적 조건을 유도할 가능성이 있는 화학 구조의 예측을 획득하는데 추가로 적용될 수 있다.

[0143] 유리하게는, 본원에 개시된 시스템 및 방법은 약물 발견을 위한 체계적인 확장가능한 접근법을 제공함으로써 설명된 단점을 해결한다. 예를 들어, 약물 발견과 관련된 통상적인 기계 학습 접근법은 딥 러닝 방법 및 고성능 컴퓨팅과 쌍형성된 3D 단백질 및 화학 구조 표현을 사용하는 인 실리코(in silico) 표적 스크리닝 능력을 이용하여 표적의 라이브러리를 향한 후보 화합물의 작용 방법을 계산한다. 그러나, 이들 접근법은 표적-집중 스크리닝 패러다임에 속하며, 이는 생물학적 과정의 근간이 되는 동적 및 고도로 네트워크화된 다세포 시스템의 복잡성을 적절하게 해결하지 못한다. 약물 발견을 위한 다른 통상적인 방법은 단일 세포 및 세포주가 전사체 데이터 또는 영상화 데이터에 기초하여 어떻게 교란에 반응하는지를 모델링하기 위해 기계 학습 접근법을 사용한다. 이러한 방법에서, 고처리량 데이터세트는 질환의 표현형 표현 및 세포 시험관내 시스템의 화합물 교란을 학습하는데 사용된다. 이들은 표현형 질환 반응을 유도하거나 상쇄시킬 화합물을 예측하는데 사용된다. 그러나, 전통적인 고처리량 데이터 모델링 접근법은 그럼에도 불구하고, 다수의 후보 표적의 식별에 대한 잠재력 및 큐레이션의 결여로 인해 불리하다. 고처리량 스크리닝으로부터 획득된 각각의 잠재적 후보의 검증은 종종 시험관내 스크리닝을 위해 수백개 또는 심지어 수천개의 화합물 또는 화합물의 분자 표적-기반 최적화 또는 합성을 필요로 하는 힘든 과정이다.

[0144] 이들 접근법과 대조적으로, 본 개시는 유리하게는 표현 화학 구조 데이터(예를 들어, 화합물 치료에 대한 세포 반응)를 획득하기 위한 시스템 및 방법을 제공하며, 이는 이어서 생물학적 과정과 연관된 세포 상태, 교란 시그니처 및/또는 세포 구성성분(예를 들어, 관심 생리학적 조건에 수반되는 유전자 모듈 또는 교란 시그니처)의 표현에 걸쳐 맵핑된다. 그럼에도 불구하고, 이러한 표적-애그노스틱 접근법은 후보 표적의 체계적인 큐레이션 및 최적화를 허용하고, 따라서 표적 발견과 시스템에 걸친 예측 번역 사이의 상당한 갭을 가교시킨다.

[0145] 예를 들어, 하기 예에 예시된 바와 같이, 지방산 대사에 관여하는 후보 약물작용발생단은 본원에 개시된 시스템 및 방법의 실시예를 사용하여 식별되었다. 예 4에 추가로 예시된 바와 같이, 후보 약물작용발생단에 기초한 예측 번역은 6개의 새로운 화학 물질을 생성하였으며, 이들 모두는 인간 지방세포 상에서 테스트시 지방산-관련 세포 과정에 관여하는 유전자 모듈을 활성화시키는 것으로 밝혀졌다. 후보 약물작용발생단의 식별 및 6개의 새로운 화학 물질의 설계는 고처리량 스크리닝, 단백질 표적에 대한 식별 또는 최적화, 또는 수백 또는 수천개의 새로운 화합물의 합성을 필요로 하지 않고 수행되었다. 따라서, 본원에 제공된 시스템 및 방법은 표적 발견에서 예측 번역 및 검증까지, 통상적인 분자 표적-기반 또는 표현형-기반 접근법에 비해 약물 발견 및 개발 과정의 용이성 및 효율을 개선시킨다.

[0146] 유리하게는, 본 개시는 화합물과 생리학적 조건 사이의 연관성(예를 들어, 가중치 및/또는 상관)의 표적화된 결정을 위한 모델의 훈련 및 사용을 개선시킴으로써, 화합물과 생리학적 조건의 연관성을 개선시키는 다양한 시스템 및 방법을 추가로 제공한다. 기계 학습 모델의 복잡성은 시간 복잡성(실행 시간, 또는 주어진 입력 크기 n에 대한 알고리즘의 속도의 척도), 공간 복잡성(공간 요건, 또는 주어진 입력 크기 n에 대한 알고리즘을 실행하는데 필요한 컴퓨팅 전력 또는 메모리의 양), 또는 둘 다를 포함한다. 복잡성(및 후속 계산 부담)은 주어진 모델에 의한 훈련 및 예측 둘 다에 적용된다.

[0147] 일부 경우에, 계산 복잡성은 구현, 추가의 알고리즘 또는 교차-검증 방법의 통합, 및/또는 하나 이상의 파라미터(예를 들어, 가중치 및/또는 하이퍼파라미터)에 의해 영향을 받는다. 일부 경우에, 계산 복잡성은 입력 크기 n의 함수로서 표현되고, 여기서 입력 데이터는 인스턴스의 수(예를 들어, 훈련 샘플의 수), 차원 p(예를 들어, 특징의 수), 트리의 수 n_{trees}(예를 들어, 트리에 기초한 방법의 경우), 서포트 벡터의 수 n_{sv}(예를 들어, 서포트 벡터에 기초한 방법의 경우), 이웃의 수 k(예를 들어, k 최근접 이웃 모델의 경우), 클래스의 수 c, 및/또는 계층 i에서의 뉴런의 수 n_i(예를 들어, 신경망의 경우)이다. 입력 크기 n과 관련하여, 그 후, 계산 복잡성의 근사치(예를 들어, 빅 O 표기법에서)는 입력 크기가 증가함에 따라 실행 시간 및/또는 공간 요건이 어떻게 증가하는지를 나타낸다. 함수는 입력 크기의 증가에 비해 더 느리거나 더 빠른 속도로 복잡성이 증가할 수 있다. 계산 복잡성의 다양한 근사치는 상수(예를 들어, $O(1)$), 대수(예를 들어, $O(\log n)$), 선형(예를 들어, $O(n)$), 로그선형(예를 들어, $O(n \log n)$), 2차(예를 들어, $O(n^2)$), 다항식(예를 들어, $O(n^c)$), 지수(예를 들어, $O(c^n)$), 및/또는 펙토리얼(예를 들어,$O(n!)$)를 포함하나 이에 제한되지는 않는다. 일부 경우에, 상수 함수의 경우에서

와 같이, 더 단순한 함수는 입력 크기가 증가함에 따라 더 낮은 수준의 계산 복잡성을 동반하는 반면, 퀵토리얼 함수와 같은 더 복잡한 함수는 입력 크기의 약간의 증가에 응답하여 복잡성의 상당한 증가를 나타낼 수 있다.

[0148] 기계 학습 모델의 계산 복잡도는 함수에 의해(예를 들어, 빅오(Big O) 표기법으로) 유사하게 표현될 수 있고, 복잡도는 모델의 유형, 하나 이상의 입력 또는 치수의 크기, 사용량(예를 들어, 훈련 및/또는 예측), 및/또는 시간 또는 공간 복잡도가 평가되는지 여부에 따라 달라질 수 있다. 예를 들어, 의사결정 트리 모델에서의 복잡도는 훈련에 대해 $O(n^2p)$ 및 예측에 대해 $O(p)$로서 근사화되는 반면, 선형 회귀 모델에서의 복잡도는 훈련에 대해 $O(p^2n + p^3)$ 및 예측에 대해 $O(p)$로서 근사화된다. 랜덤 포레스트 모델의 경우, 훈련 복잡도는 $O(n^2pn_{trees})$로 근사화되는 반면, 예측 복잡도는 $O(pn_{trees})$로 근사화된다. 구배 부스팅 모델의 경우, 복잡도는 훈련에 대해 $O(npn_{trees})$ 및 예측에 대해 $O(pn_{trees})$로서 근사화된다. 커널 서포트 벡터 머신의 경우, 복잡도는 훈련에 대해 $O(n^2p + n^3)$ 및 예측에 대해 $O(n_{sv}p)$로서 근사화된다. 나이브 베이즈 모델의 경우, 복잡도는 훈련에 대해 $O(np)$ 및 예측에 대해 $O(p)$로서 표현되고, 신경망의 경우, 복잡도는 예측에 대해 $O(pn_1 + n_1n_2 + \cdots)$로서 근사화된다. K 최근접 이웃 모델에서의 복잡도는 시간에 대해 $O(knp)$ 및 공간에 대해 $O(np)$로서 근사화된다. 로지스틱 회귀 모델의 경우, 복잡도는 시간에 대해 $O(np)$ 및 공간에 대해 $O(p)$로서 근사화된다. 로지스틱 회귀 모델의 경우, 복잡도는 시간에 대해 $O(np)$ 및 공간에 대해 $O(p)$로서 근사화된다.

[0149] 앞서 설명된 바와 같이, 기계 학습 모델의 경우, 계산 복잡도는 입력, 특징 및/또는 클래스 크기를 증가시키기 위한, 뿐만 아니라 모델 아키텍처에서의 변형을 위한 모델(예를 들어, 리그레서)의 확장성 및 따라서 전체 유효성 및 유용성을 결정한다. 대규모 데이터세트와 관련하여, 적어도 10개, 적어도 100개, 적어도 1000개 또는 그 이상의 세포에 대해 획득된 적어도 10개, 적어도 100개, 적어도 1000개 또는 그 이상의 유전자의 풍부도를 포함하는 유전자 발현 데이터세트의 경우에서와 같이, 이러한 대규모 데이터세트에 대해 수행된 함수의 계산 복잡도는 많은 기존 시스템의 능력에 부담을 줄 수 있다. 또한, 입력 특징의 수(예를 들어, 세포 구성성분(예를 들어, 유전자)의 수 및/또는 화합물의 수) 및/또는 경우의 수(예를 들어, 세포의 수, 세포 상태 주석, 교란 시그니처, 모듈, 및/또는 공변량)가 기술적 진보, 주석의 이용가능성의 증가, 및 하류 적용 및 가능성의 확장과 함께 증가함에 따라, 임의의 주어진 분류 모델의 계산 복잡도는 각각의 시스템의 사양에 의해 제공되는 시간 및 공간 용량을 빠르게 압도할 수 있다.

[0150] 따라서, 화합물을 생리학적 조건과 연관시키는 것에 대한 최소 입력 크기(예를 들어, 적어도 10개, 적어도 100개, 적어도 1000개 또는 그 이상의 화합물; 각각의 세포 구성성분 모듈에 대한 적어도 10개, 적어도 50개, 적어도 100개 또는 그 이상의 세포 구성성분; 적어도 5개, 적어도 10개, 적어도 100개 또는 그 이상의 교란 시그니처; 및/또는 적어도 5개, 적어도 10개, 적어도 100개 또는 그 이상의 세포 구성성분 모듈) 및/또는 상응하는 최소 수의 파라미터(예를 들어, 기계 학습 모델에 입력된 모든 특징의 모든 가능한 쌍형성에 상응하는 적어도 50개, 적어도 100개 또는 적어도 1000개의 파라미터 및/또는 파라미터)를 갖는 기계 학습 모델을 사용함으로써, 계산 복잡도가 비례적으로 증가되며, 따라서, 머리로는 수행할 수 없고, 본 방법은 계산 문제를 해결한다. 예를 들어, 본 개시의 한 실시예에서, 복수의 적어도 10개의 세포 구성성분 모듈 및 복수의 적어도 50개의 화합물에 의해 차원화된 활성화 점수 행렬을 획득하는 단계는 적어도 500개의 파라미터(예를 들어, 가중치)를 획득하는 것을 포함한다. 본 개시의 또 다른 실시예에서, 복수의 적어도 10개의 교란 시그니처에서 각각의 교란 시그니처에 대해, 복수의 적어도 50개의 화합물 내의 각각의 화합물에 대한 각각의 활성화 가중치를 획득하는 것은 적어도 500개의 활성화 가중치를 획득하는 것을 포함한다. 세포 상태 전이, 세포 구성성분, 세포, 화합물, 공변량, 샘플, 시점, 복제물 및/또는 배치의 수를 포함하나 이에 제한되지 않는 추가의 입력 특징 및/또는 경우에 대해 유사한 최소값을 부과하는 것은 방법의 계산 복잡성에 유사하게 영향을 미칠 것이다.

[0151] 기계 학습 모델에서의 계산 복잡도에 대한 추가의 세부사항은 thekerneltrip.com/machine/learning/computational-complexity-learning-algorithms에서 온라인으로 이용가능한 문헌 ["Computational complexity of machine learning algorithms," published April 16, 2018]; 문헌 [Hastie, 2001, *The Elements of Statistical Learning*, Springer, New York]; 및 문헌 [Arora and Barak, 2009, *Computational Complexity: A Modern Approach*, Cambridge University Press, New York]에 제공되어 있으며; 이들 각각은 그 전문이 본 출원에 참조로 포함된다.

[0152] 이제, 실시예를 상세하게 언급할 것이고, 그의 예를 첨부 도면에 예시한다. 하기 상세한 설명에서, 본 개시의 철저한 이해를 제공하기 위해 수많은 구체적 세부사항이 제시된다. 그러나, 본 개시가 이들 구체적 세부사항

없이 실시될 수 있다는 것이 관련 기술분야의 통상의 기술자에게 명백할 것이다. 다른 경우에, 널리 공지된 방법, 프로시저, 컴포넌트, 회로 및 네트워크는 실시예의 양태를 불필요하게 모호하게 하지 않도록 상세하게 기재되지 않았다.

[0153] 복수의 사례가 단일 예로서 본원에 설명되는 컴포넌트, 동작 또는 구조에 대해 제공될 수 있다. 마지막으로, 다양한 컴포넌트, 동작, 및 데이터 저장소 사이의 경계는 다소 임의적이고, 특정한 동작은 특정한 예시적인 구성에 관련하여 예시된다. 다른 형태의 기능이 구상되고 구현(들)의 범주 내에 속할 수 있다. 일반적으로, 예시적인 구성에서 별개의 컴포넌트로서 제시된 구조 및 기능은 조합된 구조 또는 컴포넌트로서 구현될 수 있다. 유사하게, 단일 컴포넌트로서 제시된 구조 및 기능성은 별개의 컴포넌트로서 구현될 수 있다. 이들 및 다른 변형, 변형, 추가 및 개선은 구현(들)의 범주 내에 속한다.

[0154] 또한, 용어 "제1", "제2" 등이 다양한 요소를 설명하기 위해 본원에 사용될 수 있지만, 이들 요소가 이들 용어에 의해 제한되지 않아야 함이 이해될 것이다. 이들 용어는 단지 하나의 요소를 또 다른 요소와 구별하기 위해 사용될 뿐이다. 예를 들어, 본 발명의 범주로부터 벗어나지 않으면서, 제1 데이터세트는 제2 데이터세트로 명명될 수 있고, 유사하게, 제2 데이터세트는 제1 데이터세트로 명명될 수 있다. 제1 데이터세트 및 제2 데이터세트는 둘 다 데이터세트이지만, 이들은 동일한 데이터세트는 아니다.

[0155] 본원에 사용된 용어는 단지 특정한 구현을 설명하기 위한 것이며, 청구범위를 제한하는 것으로 의도되지 않는다. 구현 및 첨부된 청구범위의 설명에 사용된 바와 같이, 단수 형태는 문맥이 달리 명백하게 나타내지 않는 한 복수 형태를 또한 포함하는 것을 의도한다. 또한, 본원에 사용된 용어 "및/또는"은 연관된 열거된 항목 중 하나 이상의 임의의 및 모든 가능한 조합을 지칭하고 포괄하는 것으로 이해될 것이다. 추가로, 용어 "포함하다" 및/또는 "포함하는"은, 본 명세서에서 사용되는 경우에, 언급된 특징, 정수, 단계, 동작, 요소 및/또는 컴포넌트의 존재를 명시하지만, 하나 이상의 다른 특징, 정수, 단계, 동작, 요소, 컴포넌트 및/또는 그의 그룹의 존재 또는 추가를 배제하지 않는 것으로 이해될 것이다.

[0156] 본원에 사용된 용어 "~인 경우"는 문맥에 따라 언급된 조건 선례가 참일 "때" 또는 "시" 또는 "~이라고 결정하는 것에 응답하여" 또는 "~이라는 결정에 따라" 또는 "~이라고 검출하는 것에 응답하여"를 의미하는 것으로 해석될 수 있다. 유사하게, 어구 "(언급된 조건 선례가 참인 것으로) 결정된 경우" 또는 "(언급된 조건 선례가 참인) 경우" 또는 "(언급된 조건 선례가 참일) 때"는 문맥에 따라, 언급된 조건 선례가 참인 것으로 "결정시" 또는 "결정하는 것에 응답하여" 또는 "결정에 따라" 또는 "검출시" 또는 "검출하는 것에 응답하여"를 의미하는 것으로 해석될 수 있다.

[0157] 또한, 참조 번호가 "i번째" 표기법으로 주어지는 경우에, 참조 번호는 일반적 컴포넌트, 세트 또는 실시예를 지칭한다. 예를 들어, "세포-컴포넌트 *i*"로 명명된 세포-컴포넌트는 복수의 세포-컴포넌트에서의 i번째 세포-컴포넌트를 지칭한다.

[0158] 전술한 설명은 예시적인 구현을 구현하는 예시적인 시스템, 방법, 기술, 명령어 시퀀스, 및 컴퓨팅 머신 프로그램 제품을 포함하였다. 설명을 위해, 본 발명의 대상의 다양한 구현의 이해를 제공하기 위해 수많은 구체적 세부사항이 제시된다. 그러나, 본 발명의 대상의 구현이 이들 구체적 세부사항 없이 실시될 수 있음이 관련 기술분야의 통상의 기술자에게 명백할 것이다. 일반적으로, 널리 공지된 명령어 인스턴스, 프로토콜, 구조 및 기술은 상세하게 제시되지 않았다.

[0159] 전술한 설명은, 설명의 목적을 위해, 구체적 구현을 참조하여 기재되었다. 그러나, 하기 예시적인 논의는 모든 것을 설명하거나 개시된 정확한 형태로 구현을 제한하려는 의도가 아니다. 상기 교시를 고려하여 많은 수정 및 변형이 가능하다. 구현은 원리 및 그의 실제 응용을 가장 잘 설명하기 위해 선택 및 설명되고, 이렇게 함으로써, 관련 기술분야의 다른 통상의 기술자가 고려되는 특정 용도에 적합한 다양한 수정을 가지는 구현들 및 다양한 구현을 가장 잘 이용할 수 있게 한다.

[0160] 명확성을 위해, 본원에 설명된 구현의 모든 일상적인 특징이 도시되고 설명되지는 않는다. 임의의 이러한 실제 구현의 개발에서, 수많은 구현-특정적 결정이 사용 사례- 및 비즈니스-관련 제약의 준수와 같은 설계자의 특정 목표를 달성하기 위해 이루어지고, 이들 특정 목표는 구현마다 및 설계자마다 달라질 것임이 이해될 것이다. 또한, 이러한 설계 노력은 복잡하고 시간-소모적일 수 있지만, 그럼에도 불구하고 본 개시의 이익을 갖는 본 기술 분야의 숙련자에게는 일상적인 엔지니어링 수행이라는 것이 이해될 것이다.

[0161] 본 설명의 일부 부분은 정보에 대한 동작의 알고리즘 및 기호 표현의 관점에서 본 발명의 실시예를 설명한다. 이러한 알고리즘적 설명 및 표현은 데이터 처리 분야의 통상의 기술자가 그의 연구의 본질을 관련 기술분야의

다른 통상의 기술자에게 효과적으로 전달하기 위해 통상적으로 사용된다. 이러한 동작은, 기능적으로, 계산적으로, 또는 논리적으로 설명되지만, 컴퓨터 프로그램 또는 등가의 전기 회로, 마이크로코드 등에 의해 구현되는 것으로 이해된다.

[0162] 본 명세서에서 사용된 언어는 주로 가독성 및 교육 목적을 위해 선택되었고, 본 발명의 주제를 서술하거나 제한하기 위해 선택된 것은 아닐 수 있다. 따라서, 본 발명의 범위는 이 상세한 설명에 의해서가 아니라, 오히려 본원에 기초한 출원에 대해 허여된 임의의 청구범위에 의해 제한되는 것으로 의도된다. 따라서, 본 발명의 실시예의 개시는 본 발명의 범위를 제한하는 것이 아니라 예시하는 것으로 의도된다.

[0163] 일반적으로, 청구범위 및 명세서에 사용된 용어는 관련 기술분야의 통상의 기술자에 의해 이해되는 명확한 의미를 갖는 것으로 해석되도록 의도된다. 특정 용어는 추가적인 명확성을 제공하기 위해 하기에 정의된다. 명확한 의미와 제공된 정의 사이에서 상충되는 경우에, 제공된 정의가 사용되어야 한다.

[0164] 본원에 직접 정의되지 않은 임의의 용어는 본 발명의 관련 기술분야 내에서 이해되는 바와 같은 그와 통상적으로 연관된 의미를 갖는 것으로 이해될 것이다. 본 발명의 양태의 조성, 디바이스, 방법 등, 및 그의 제조 또는 사용 방법을 설명함에 있어 실시자에게 추가의 지침을 제공하기 위해 본원에서 특정 용어가 논의된다. 동일한 것이 둘 이상의 방식으로 언급될 수 있음을 이해할 수 있을 것이다. 결과적으로, 대안적 언어 및 동의어가 본원에 논의된 용어 중 임의의 하나 이상에 사용될 수 있다. 용어가 본원에서 상술되거나 논의되었는지 또는 그렇지 않은지는 중요하지 않다. 일부 동의어 또는 대체가능한 방법, 물질 등이 제공된다. 하나 또는 몇몇의 동의어 또는 등가물에 대한 언급은 명시적으로 언급되지 않는 한 다른 동의어 또는 등가물의 사용을 배제하지 않는다. 용어의 예를 비롯한 예의 사용은 단지 예시적 목적을 위한 것이며, 본원의 본 발명의 양태의 범위 및 의미를 제한하지는 않는다.

[0165] *정의.*

[0166] 본원에 사용된 용어 "약" 또는 "대략"은 관련 기술분야의 통상의 기술자에 의해 결정된 바와 같은 특정한 값에 대한 허용되는 오차 범위 내를 의미하며, 이는 부분적으로 값이 측정 또는 결정되는 방법, 예를 들어 측정 시스템의 한계에 따라 달라진다. 예를 들어, 일부 실시예에서 "약"은 본 기술 분야의 관례에 따라 1 또는 1 초과의 표준 편차 이내를 의미한다. 일부 실시예에서, "약"은 주어진 값의 ±20%, ±10%, ±5% 또는 ±1%의 범위를 의미한다. 일부 실시예에서, 용어 "약" 또는 "대략"은 값의 한 자릿수 이내, 5배 이내, 또는 2배 이내를 의미한다. 특정한 값이 본 출원 및 청구범위에 설명된 경우에, 달리 언급되지 않는 한, 용어 "약"은 특정한 값에 대해 허용되는 오차 범위 내를 의미하는 것으로 가정될 수 있다. 본원의 상세한 설명 내의 모든 수치값은 "약" 표시된 값에 의해 수식되고, 관련 기술분야의 통상의 기술자에 의해 예상되는 실험 오차 및 변동을 고려한다. 용어 "약"은 관련 기술분야의 통상의 기술자에 의해 통상적으로 이해되는 바와 같은 의미를 가질 수 있다. 일부 실시예에서, 용어 "약"은 ±10%를 지칭한다. 일부 실시예에서, 용어 "약"은 ±5%를 지칭한다.

[0167] 본원에 사용된 용어 "풍부도", "풍부도 수준" 또는 "발현 수준"은 하나 이상의 세포에 존재하는 세포 구성성분(예를 들어, 유전자 산물, 예컨대 RNA 종, 예를 들어 mRNA 또는 miRNA, 또는 단백질 분자)의 양, 또는 다중 세포에 걸쳐 존재하는 세포 구성성분의 평균 양을 지칭한다. mRNA 또는 단백질 발현을 언급할 때, 일반적으로 이러한 용어는 특정 게놈 유전자좌, 예를 들어, 특정 유전자에 상응하는 임의의 RNA 또는 단백질 종의 양을 지칭한다. 그러나, 일부 실시예에서, 풍부도는 다중 mRNA 또는 단백질 이소형을 생성하는 특정한 유전자에 상응하는 mRNA 또는 단백질의 특정한 이소형의 양을 지칭할 수 있다. 게놈 유전자좌는 유전자 명칭, 염색체 위치, 또는 임의의 다른 유전자 맵핑 메트릭을 사용하여 식별될 수 있다.

[0168] 본원에서 상호교환가능하게 사용되는 "세포 상태" 또는 "생물학적 상태"는 세포 또는 세포 집단의 상태 또는 표현형을 의미한다. 예를 들어, 세포 상태는 건강하거나 질환에 걸려 있을 수 있다. 세포 상태는 복수의 질환 중 하나일 수 있다. 세포 상태는 화합물 치료 및/또는 분화된 세포 계통에 대한 반응일 수 있다. 세포 상태는 하나 이상의 유전자, 하나 이상의 단백질, 및/또는 하나 이상의 생물학적 경로를 포함하나 이에 제한되지 않는 하나 이상의 세포 구성성분의 척도(예를 들어, 활성화, 발현, 및/또는 풍부도의 척도)를 특징으로 할 수 있다.

[0169] 본원에 사용된 "세포 상태 전이" 또는 "세포 전이"는 제1 세포 상태로부터 제2 세포 상태로의 세포 상태의 전이를 지칭한다. 일부 실시예에서, 제2 세포 상태는 변경된 세포 상태(예를 들어, 건강한 세포 상태 내지 이환된 세포 상태)이다. 일부 실시예에서, 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 소정 조건에 대한 세포의 노출에 의해 유발된 교란된 상태이다. 교란된 상태는 화합물에 대한 세포의 노출에 의해 유발될 수 있다. 세포 상태 전이는 세

포 내의 세포 구성성분 풍부도의 변화, 및 따라서 세포에 의해 생산된 세포 구성성분(예를 들어, mRNA, 전사 인자)의 아이덴티티 및 양(예를 들어, 교란 시그니처)에 의해 표시될 수 있다.

[0170] 세포 또는 복수의 세포에 대한 세포 구성성분 풍부도 측정치와 관련하여 본원에 사용된 용어 "데이터세트"는 일부 문맥에서 단일 세포로부터 수집된 데이터의 고차원 세트(예를 들어, 단일-세포 세포 구성성분 풍부도 데이터세트)를 지칭할 수 있다. 다른 문맥에서, 용어 "데이터세트"는 단일 세포로부터 수집된 데이터의 복수의 고차원 세트(예를 들어, 복수의 단일-세포 세포 구성성분 풍부도 데이터세트)를 지칭할 수 있으며, 복수의 데이터의 각각의 세트는 복수의 세포 중 하나의 세포로부터 수집된다.

[0171] 본원에 사용된 용어 "차등 풍부도" 또는 "차등 발현"은 제2 엔티티(예를 들어, 제2 세포, 복수의 세포, 및/또는 샘플)와 비교하여 제1 엔티티(예를 들어, 제1 세포, 복수의 세포, 및/또는 샘플)에 존재하는 세포 구성성분의 양 및/또는 빈도에서의 차이를 지칭한다. 일부 실시예에서, 제1 엔티티는 제1 세포 상태(예를 들어, 질환 표현형)를 특징으로 하는 샘플이고, 제2 엔티티는 제2 세포 상태(예를 들어, 정상 또는 건강한 표현형)를 특징으로 하는 샘플이다. 예를 들어, 세포 구성성분은 제2 세포 상태를 특징으로 하는 엔티티에 비교하여 제1 세포 상태를 특징으로 하는 엔티티에서 상승된 수준 또는 감소된 수준으로 존재하는 폴리뉴클레오티드(예를 들어, mRNA 전사체)일 수 있다. 일부 실시예에서, 세포 구성성분은 제2 세포 상태를 특징으로 하는 엔티티에 비해 제1 세포 상태를 특징으로 하는 엔티티에서 보다 높은 빈도 또는 보다 낮은 빈도로 검출되는 폴리뉴클레오티드일 수 있다. 세포 구성성분은 양, 빈도 또는 둘 다의 관점에서 차등적으로 풍부할 수 있다. 일부 경우에, 세포 구성성분은 한 엔티티 내의 세포 구성성분의 양이 다른 엔티티 내의 세포 구성성분의 양과 통계적으로 유의하게 상이한 경우에 2개의 엔티티 사이에 차등적으로 풍부하다. 예를 들어, 세포 구성성분은 한 엔티티에서 다른 엔티티에 존재하는 것보다 적어도 약 120%, 적어도 약 130%, 적어도 약 150%, 적어도 약 180%, 적어도 약 200%, 적어도 약 300%, 적어도 약 500%, 적어도 약 700%, 적어도 약 900%, 또는 적어도 약 1000% 더 많이 존재하는 경우에, 또는 한 엔티티에서는 검출가능하고 다른 엔티티에서는 검출가능하지 않은 경우에, 2개의 엔티티에 차등적으로 풍부하다. 일부 경우에, 세포 구성성분은 엔티티의 제1 서브세트(예를 들어, 주석화된 세포 상태의 제1 서브세트를 나타내는 세포)에서 세포 구성성분을 검출하는 빈도가 엔티티의 제2 서브세트(예를 들어, 주석화된 세포 상태의 제2 서브세트를 나타내는 세포)에서보다 통계적으로 유의하게 더 높거나 더 낮은 경우에 엔티티의 2개의 세트에서 차등 발현된다. 예를 들어, 세포 구성성분은 한 세트의 엔티티에서 다른 세트의 엔티티보다 적어도 약 120%, 적어도 약 130%, 적어도 약 150%, 적어도 약 180%, 적어도 약 200%, 적어도 약 300%, 적어도 약 500%, 적어도 약 700%, 적어도 약 900% 또는 적어도 약 1000% 더 빈번하게 또는 덜 빈번하게 관찰되는 것으로 검출되는 경우에 2개의 세트의 엔티티에서 차등 발현된다.

[0172] 본원에 사용된 용어 "건강한"은 건강한 상태(예를 들어, 우수한 건강을 갖는 대상으로부터 획득됨)를 특징으로 하는 샘플을 지칭한다. 건강한 대상은 임의의 악성 또는 비-악성 질환의 부재를 입증할 수 있다. "건강한" 개체는, 통상적으로 "건강한" 것으로 간주될 수 없는, 검정되는 상태와 무관한 다른 질환 또는 상태를 가질 수 있다.

[0173] 세포와 관련하여 본원에 사용된 용어 "교란"(예를 들어, 세포의 교란 또는 세포 교란)은 하나 이상의 조건에 대한 세포의 임의의 노출, 예컨대 1종 이상의 화합물에 의한 치료를 지칭한다. 이들 화합물은 "교란원(perturbagens)"으로 지칭될 수 있다. 일부 실시예에서, 교란원은 예를 들어 소분자, 생물제제, 치료제, 단백질, 소분자와 조합된 단백질, ADC, 핵산, 예컨대 siRNA 또는 간섭 RNA, 야생형 및/또는 돌연변이체 shRNA를 과다발현하는 cDNA, 야생형 및/또는 돌연변이체 가이드 RNA를 과다발현하는 cDNA(예를 들어, Cas9 시스템 또는 다른 유전자 편집 시스템), 또는 임의의 전술한 것의 임의의 조합을 포함할 수 있다. 교란은 세포의 표현형의 변화 및/또는 세포 내의 하나 이상의 세포 구성성분의 발현 또는 풍부도 수준의 변화(예를 들어, 교란 시그니처)를 유도하거나 또는 이를 특징으로 할 수 있다. 예를 들어, 교란은 세포의 전사 프로파일의 변화를 특징으로 할 수 있다.

[0174] 본원에 사용된 용어 "샘플", "생물학적 샘플" 또는 "환자 샘플"은 대상과 연관된 생물학적 상태를 반영할 수 있는, 대상으로부터 취한 임의의 샘플을 지칭한다. 샘플의 예는 대상의 혈액, 전혈, 혈장, 혈청, 소변, 뇌척수액, 분변, 타액, 땀, 눈물, 흉막액, 심막액 또는 복막액을 포함하나, 이에 제한되지 않는다. 샘플은 살아있거나 죽은 대상으로부터 유도된 임의의 조직 또는 물질을 포함할 수 있다. 샘플은 무세포 샘플일 수 있다. 샘플은 하나 이상의 세포 구성성분을 포함할 수 있다. 예를 들어, 샘플은 핵산(예를 들어, DNA 또는 RNA) 또는 그의 단편, 또는 단백질을 포함할 수 있다. 용어 "핵산"은 데옥시리보핵산(DNA), 리보핵산(RNA) 또는 그의 임의의 하이브리드 또는 단편을 지칭할 수 있다. 샘플 내의 핵산은 무세포 핵산일 수 있다. 샘플은 액체 샘플 또는 고체 샘플(예를 들어, 세포 또는 조직 샘플)일 수 있다. 샘플은 체액일 수 있다. 샘플은 대변 샘플일 수

있다. 샘플은 조직 또는 세포 구조를 물리적으로 파괴하도록 처리(예를 들어, 원심분리 및/또는 세포 용해)될 수 있고, 따라서 분석을 위한 샘플을 제조하는데 사용될 수 있는 효소, 완충제, 염, 세제 등을 추가로 함유할 수 있는 용액 내로 세포내 컴포넌트를 방출시킨다.

[0175] 본원에 사용된 바와 같이, 화합물의 지문에서와 같은 용어 "지문"은 화합물의 디지털 다이제스트이다. 이러한 디지털 다이제스트의 비제한적 예는 데이라이트 지문, BCI 지문, ECFC4 지문, ECFP4 지문, EcFC 지문, MDL 지문, 원자 쌍 지문(APFP 지문), 위상적 비틀림 지문(TTFP) 지문, UNITY 2D 지문, RNNS2S 지문, 또는 GraphConv 지문을 포함한다. 문헌 [Franco, 2014, "The Use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation," J. Cheminform 6, p. 5], 및 문헌 [Rensi and Altman, 2017, "Flexible Analog Search with Kernel PCA Embedded Molecule Vectors," Computational and Structural Biotechnology Journal, doi:10.1016/j.csbj.2017.03.003]을 참조하며, 이들 각각은 본원에 참조로 포함된다. 또한, 문헌 [Raymond and Willett, 2002, "Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases," Journal of Computer-Aided Molecular Design 16, 59-71] 및 문헌 [Franco 등의 2014, "The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation" Journal of chemoinformatics 6(5)]을 참조하며, 이들 각각은 본원에 참조로 포함된다.

[0176] 본원에 사용된 용어 "분류"는 엔티티(예를 들어, 세포, 샘플, 세포 구성성분, 세포 구성성분 모듈 등)의 특정한 특성(예를 들어, 세포 과정, 공변량, 세포 상태 주석 등)과 연관된 임의의 수(들) 또는 다른 문자(들)를 지칭할 수 있다. 예를 들어, "+" 기호(또는 단어 "양성")는 특정 엔티티가 특정 특성에 대해 양성으로 분류된다는 것을 의미할 수 있다(예를 들어, 세포 구성성분 모듈은 관심 세포 과정과 양성으로 연관된다). 또 다른 예에서, 용어 "분류"는 엔티티와 특정한 특성 사이의 상관(예를 들어, 각각의 공변량과의 세포 구성성분 모듈 사이의 상관)의 결정을 지칭할 수 있다. 일부 실시예에서, 분류는 상관 계수 및/또는 가중치이다. 분류는 이진적일(예를 들어, 양성 또는 음성) 수 있거나 또는 보다 많은 수준의 분류(예를 들어, 1 내지 10 또는 0 내지 1의 척도)를 가질 수 있다. 용어 "컷오프" 및 "임계치"는 동작에 사용되는 미리 결정된 수를 지칭할 수 있다. 예를 들어, 컷오프 값은 그를 초과하면 엔티티가 배제되는 값을 지칭할 수 있다. 임계치 값은 특정 분류가 적용되는 값을 초과하거나 또는 그 미만일 수 있다. 이들 용어 중 어느 하나가 이들 문맥 중 어느 하나에서 사용될 수 있다.

[0177] 본원에서 상호교환가능하게 사용되는 용어 "분류자", "모델", "알고리즘", "리그레서" 및/또는 "분류자"는 기계 학습 모델 또는 알고리즘을 지칭한다. 일부 실시예에서, 모델은 비지도 학습 알고리즘이다. 비지도 학습 알고리즘의 한 예는 클러스터 분석이다.

[0178] 일부 실시예에서, 모델은 지도 기계 학습이다. 지도 학습 알고리즘의 비제한적 예는 로지스틱 회귀, 신경망, 서포트 벡터 머신, 나이브 베이즈 알고리즘, 최근접 이웃 알고리즘, 랜덤 포레스트 알고리즘, 의사결정 트리 알고리즘, 부스팅 트리 알고리즘, 다항 로지스틱 회귀 알고리즘, 선형 모델, 선형 회귀, 구배부스팅, 혼합 모델, 은닉 마르코프 모델, 가우시안 NB 알고리즘, 선형 판별 분석, 또는 그 임의의 조합을 포함하나, 이에 제한되지 않는다. 일부 실시예에서, 모델은 다항식 분류자 알고리즘이다. 일부 실시예에서, 모델은 2-단계 확률적 구배 하강(SGD) 모델이다. 일부 실시예에서, 모델은 심층 신경망(예를 들어, 심층-광범위 샘플-수준 모델)이다. 일부 실시예에서, 본 개시의 분류자 또는 모델은 25개 이상, 100개 이상, 1000개 이상, 10,000개 이상, 100,000개 이상 또는 1×10^6개 이상의 파라미터를 가지며, 따라서 모델의 계산은 머리로는 수행할 수 없다.

[0179] 또한, 본원에 사용된 용어 "파라미터"는 알고리즘, 모델, 리그레서 및/또는 분류자에서 하나 이상의 입력, 출력 및/또는 기능에 영향을 미칠 수 있는(예를 들어, 이를 수정, 맞춤화 및/또는 조정할 수 있는) 알고리즘, 모델, 리그레서 및/또는 분류자에서 임의의 계수, 또는 유사하게 내부 또는 외부 요소의 임의의 값(예를 들어, 가중치 및/또는 하이퍼파라미터)을 지칭한다. 예를 들어, 일부 실시예에서, 파라미터는 알고리즘, 모델, 리그레서 및/또는 분류자의 거동, 학습 및/또는 성능을 제어, 수정, 맞춤화 및/또는 조정하는데 사용될 수 있는 임의의 계수, 가중치 및/또는 하이퍼파라미터를 지칭한다. 일부 경우에, 파라미터는 알고리즘, 모델, 리그레서 및/또는 분류자에 대한 입력(예를 들어, 특징)의 영향을 증가 또는 감소시키는데 사용된다. 비제한적 예로서, 일부 실시예에서, 파라미터는 노드(예를 들어, 신경망)의 영향을 증가 또는 감소시키는데 사용되며, 여기서 노드는 하나 이상의 활성화 함수를 포함한다. 특정 입력, 출력, 및/또는 함수에 대한 파라미터의 할당은 주어진 알고리즘, 모델, 리그레서, 및/또는 분류자에 대한 임의의 하나의 패러다임으로 제한되지 않고, 목적하는 성능을 위한 임의의 적합한 알고리즘, 모델, 리그레서, 및/또는 분류자 아키텍처에서 사용될 수 있다. 일부 실시예에서,

파라미터는 고정된 값을 갖는다. 일부 실시예에서, 파라미터의 값은 수동으로 및/또는 자동으로 조정가능하다. 일부 실시예에서, 파라미터의 값은 알고리즘, 모델, 리그레서 및/또는 분류자에 대한 검증 및/또는 훈련 과정에 의해(예를 들어, 오차 최소화 및/또는 역전파 방법에 의해) 수정된다. 일부 실시예에서, 본 개시의 알고리즘, 모델, 리그레서 및/또는 분류자는 복수의 파라미터를 포함한다. 일부 실시예에서, 복수의 파라미터는 n개의 파라미터이며, 여기서 $n \geq 2$; $n \geq 5$; $n \geq 10$; $n \geq 25$; $n \geq 40$; $n \geq 50$; $n \geq 75$; $n \geq 100$; $n \geq 125$; $n \geq 150$; $n \geq 200$; $n \geq 225$; $n \geq 250$; $n \geq 350$; $n \geq 500$; $n \geq 600$; $n \geq 750$; $n \geq 1,000$; $n \geq 2,000$; $n \geq 4,000$; $n \geq 5,000$; $n \geq 7,500$; $n \geq 10,000$; $n \geq 20,000$; $n \geq 40,000$; $n \geq 75,000$; $n \geq 100,000$; $n \geq 200,000$; $n \geq 500,000$, $n \geq 1 \times 10^6$, $n \geq 5 \times 10^6$, 또는 $n \geq 1 \times 10^7$이다. 이와 같이, 본 개시의 알고리즘, 모델, 리그레서 및/또는 분류자는 머리로는 수행할 수 없다. 일부 실시예에서, n은 10,000 내지 1×10^7, 100,000 내지 5×10^6, 또는 500,000 내지 1×10^6이다. 일부 실시예에서, 본 개시의 알고리즘, 모델, 리그레서 및/또는 분류자는 k-차원 공간에서 작동하며, 여기서 k는 5 이상의 양의 정수(예를 들어, 5, 6, 7, 8, 9, 10 등)이다. 이와 같이, 본 개시의 알고리즘, 모델, 리그레서 및/또는 분류자는 머리로는 수행할 수 없다.

[0180] *신경망.* 일부 실시예에서, 모델은 신경망(예를 들어, 컨볼루션 신경망 및/또는 잔차 신경망)이다. 인공 신경망(ANN)이라고도 공지된 신경망 모델은 컨볼루션 및/또는 잔차 신경망 모델(딥 러닝 모델)을 포함한다. 신경망은 입력 데이터 세트를 출력 데이터 세트에 맵핑하도록 훈련될 수 있는 기계 학습 모델들일 수 있으며, 여기서 신경망은 노드의 다수의 계층으로 조직화된 노드의 상호접속된 그룹을 포함한다. 예를 들어, 신경망 아키텍처는 적어도 입력 계층, 하나 이상의 은닉 계층, 및 출력 계층을 포함할 수 있다. 신경망은 임의의 총수의 계층, 및 임의의 수의 은닉 계층을 포함할 수 있으며, 여기서 은닉 계층은 입력 데이터 세트를 출력 값 또는 출력 값 세트에 맵핑하는 것을 가능하게 하는 훈련가능한 특징 추출기로서 기능한다. 본원에 사용될 때, 딥 러닝 모델(DNN)은 복수의 은닉 계층, 예를 들어, 2개 이상의 은닉 계층을 포함하는 신경망일 수 있다. 신경망의 각각의 계층은 다수의 노드(또는 "뉴런")를 포함할 수 있다. 노드는 입력 데이터 또는 이전 계층에서의 노드의 출력 중 어느 하나로부터 직접 오는 입력을 수신하고, 특정 동작, 예를 들어, 합산 동작을 수행할 수 있다. 일부 실시예에서, 입력으로부터 노드로의 연결은 파라미터(예를 들어, 가중치 및/또는 가중 인자)와 연관된다. 일부 실시예에서, 노드는 모든 입력 쌍, x_i 및 그의 관련 파라미터의 곱을 합산할 수 있다. 일부 실시예에서, 가중 합계는 바이어스 b로 오프셋된다. 일부 실시예에서, 노드 또는 뉴런의 출력은 선형 또는 비선형 함수일 수 있는 임계치 또는 활성화 함수 f를 사용하여 게이팅될 수 있다. 활성화 함수는, 예를 들어, 정류 선형 단위(ReLU) 활성화 함수, 리키(Leaky) ReLU 활성화 함수, 또는 포화 쌍곡선 탄젠트, 아이덴티티, 이진 스텝, 로지스틱, arcTan, 소프트사인, 파라미터 정류 선형 단위, 지수 선형 단위, softPlus, 벤트 아이덴티티, softExponential, 시누소이드, 사인, 가우시안, 또는 시그모이드 함수, 또는 그 임의의 조합과 같은 다른 함수일 수 있다.

[0181] 신경망의 가중 인자, 바이어스 값, 및 임계값, 또는 다른 계산 파라미터는 하나 이상의 훈련 데이터 세트를 사용하여 훈련 단계에서 "교시"되거나 "학습"될 수 있다. 예를 들어, ANN이 계산하는 출력 값(들)이 훈련 데이터 세트에 포함된 예와 일치하도록 훈련 데이터 세트로부터의 입력 데이터 및 구배 하강 또는 역방향 전파 방법을 사용하여 파라미터가 훈련될 수 있다. 파라미터는 역전파 신경망 훈련 과정으로부터 획득될 수 있다.

[0182] 임의의 다양한 신경망이 대상의 영상을 분석하는데 사용하기에 적합할 수 있다. 예는 피드포워드 신경망, 방사형 기저 기능 네트워크, 회귀 신경망, 잔차 신경망, 컨볼루션 신경망, 잔차 컨볼루션 신경망 등, 또는 그 임의의 조합을 포함할 수 있으나, 이에 제한되지는 않는다. 일부 실시예에서, 기계 학습은 사전-훈련된 및/또는 전이 학습된 ANN 또는 딥 러닝 아키텍처를 사용한다. 컨볼루션 및/또는 잔차 신경망이 본 개시에 따라 대상의 영상을 분석하는데 사용될 수 있다.

[0183] 예를 들어, 심층 신경망 모델은 입력 계층, 복수의 개별적으로 파라미터화된(예를 들어, 가중된) 컨볼루션 계층, 및 출력 스코어러를 포함한다. 컨볼루션 계층뿐만 아니라 입력 계층 각각의 파라미터(예를 들어, 가중치)는 심층 신경망 모델과 연관된 복수의 파라미터(예를 들어, 가중치)에 기여한다. 일부 실시예에서, 적어도 100개의 파라미터, 적어도 1000개의 파라미터, 적어도 2000개의 파라미터 또는 적어도 5000개의 파라미터가 심층 신경망 모델과 연관된다. 이와 같이, 심층 신경망 모델은 이들이 머리로는 해석할 수 없기 때문에 컴퓨터를 사용할 필요가 있다. 즉, 모델에 대한 입력이 주어지면, 모델 출력은 이러한 실시예에서 머리가 아닌 컴퓨터를 사용하여 결정될 필요가 있다. 예를 들어, 문헌 [Krizhevsky 등의 2012, "Imagenet classification with deep convolutional neural networks, "*Advances in Neural Information Processing Systems* 2, Pereira, Burges, Bottou, Weinberger, eds., pp. 1097-1105, Curran Associates, Inc.]; 문헌 [Zeiler, 2012

"ADADELTA: an adaptive learning rate method,"' CoRR, vol. abs/1212.5701] 및 문헌 [Rumelhart 등의 1988, "Neurocomputing: Foundations of research," ch. Learning Representations by Back-propagating Errors, pp. 696-699, Cambridge, MA, USA: MIT Press]을 참조하며, 이들 각각은 본원에 참조로 포함된다.

[0184] 모델로서 사용하기에 적합한, 컨볼루션 신경망 모델을 포함한 신경망 모델은, 예를 들어 문헌 [Vincent 등의 2010, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," J Mach Learn Res 11, pp. 3371-3408]; 문헌 [Larochelle 등의 2009, "Exploring strategies for training deep neural networks," J Mach Learn Res 10, pp. 1-40]; 및 문헌 [Hassoun, 1995, Fundamentals of Artificial Neural Networks, Massachusetts Institute of Technology]에 개시되어 있으며, 이들 각각은 본 출원에 참조로 포함된다. 모델로서 사용하기에 적합한 추가의 예시적인 신경망은 문헌 [Duda 등의 2001, *Pattern Classification*, Second Edition, John Wiley & Sons, Inc., New York]; 및 문헌 [Hastie 등의 2001, *The Elements of Statistical Learning*, Springer-Verlag, New York]에 개시되어 있고, 이들 각각은 그 전문이 본원에 참조로 포함된다. 모델로서 사용하기에 적합한 추가의 예시적인 신경망은 또한 문헌 [Draghici, 2003, *Data Analysis Tools for DNA Microarrays*, Chapman & Hall/CRC]; 및 문헌 [Mount, 2001, *Bioinformatics: sequence and genome analysis*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York]에 설명되어 있고, 이들 각각은 그 전문이 본원에 참조로 포함된다.

[0185] *서포트 벡터 머신.* 일부 실시예에서, 모델은 서포트 벡터 머신(SVM)이다. 모델로서 사용하기에 적합한 SVM 모델은 예를 들어, 문헌 [Cristianini and Shawe-Taylor, 2000, "An Introduction to Support Vector Machines," Cambridge University Press, Cambridge]; 문헌 [Boser 등의 1992, "A training algorithm for optimal margin models," in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, ACM Press, Pittsburgh, Pa., pp. 142-152]; 문헌 [Vapnik, 1998, Statistical Learning Theory, Wiley, New York; Mount, 2001, Bioinformatics: sequence and genome analysis, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.]; 문헌 [Duda, Pattern Classification, Second Edition, 2001, John Wiley & Sons, Inc., pp. 259, 262-265]; 및 문헌 [Hastie, 2001, The Elements of Statistical Learning, Springer, New York; and Furey 등의 2000, Bioinformatics 16, 906-914]에 설명되어 있으며, 이들 각각은 그 전문이 본원에 참조로 포함된다. 분류에 사용될 때, SVM은 표지된 데이터로부터 최대로 떨어진 초평면으로 주어진 세트의 이진 표지된 데이터를 분리한다. 선형 분리가 가능하지 않은 경우에, SVM은 특징 공간에 대한 비선형 맵핑을 자동으로 실현하는 '커널'의 기술과 조합하여 작동할 수 있다. 특징 공간에서 SVM에 의해 발견된 초평면은 입력 공간에서의 비선형 결정 경계에 상응할 수 있다. 일부 실시예에서, SVM과 연관된 복수의 파라미터(예를 들어, 가중치)는 초평면을 정의한다. 일부 실시예에서, 초평면은 적어도 10개, 적어도 20개, 적어도 50개, 또는 적어도 100개의 파라미터에 의해 정의되고, SVM 모델은 머리로는 해석할 수 없기 때문에 계산을 위해 컴퓨터가 필요하다.

[0186] *나이브 베이즈 모델.* 일부 실시예에서, 모델은 나이브 베이즈 모델이다. 모델로서 사용하기에 적합한 나이브 베이즈 모델은, 예를 들어 본원에 참조로 포함된 문헌 [Ng 등의 2002, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," Advances in Neural Information Processing Systems, 14]에 개시되어 있다. 나이브 베이즈 분류자는 특징 사이에 강한 (나이브) 독립 가정을 갖는 베이즈의 이론을 적용하는 것에 기초한 "확률적 분류자"의 패밀리에서의 임의의 분류자이다. 일부 실시예에서, 이들은 커널 밀도 추정과 결합된다. 예를 들어, 본원에 참조로 포함된 문헌 [Hastie 등의 2001, *The elements of statistical learning : data mining, inference, and prediction*, eds. Tibshirani and Friedman, Springer, New York]을 참조한다.

[0187] *최근접 이웃 모델.* 일부 실시예에서, 모델은 최근접 이웃 모델이다. 최근접 이웃 모델은 메모리-기반일 수 있고, 피팅될 모델을 포함하지 않는다. 최근접 이웃에 대해, 질의 지점 x_0(테스트 대상)가 주어지면, x_0까지의 거리가 가장 가까운 k개의 훈련 지점 $x_{(r)}$, r, … , k(여기서는 훈련 대상)을 식별한 다음, k개의 최근접 이웃을 사용하여 지점 x_0를 분류한다. 여기서, 이들 이웃까지의 거리는 식별 유전자 세트의 풍부도 값의 함수이다. 일부 실시예에서, 특징 공간에서의 유클리드 거리가 사용되어 다음과 같이 거리를 결정한다: $d_{(i)} = \|x_{(i)} - x_{(o)}\|$. 전형적으로, 최근접 이웃 모델이 사용될 때, 선형 판별자를 계산하는데 사용된 풍부도 데이터는 평균 0 및 분산 1을 갖도록 표준화된다. 최근접 이웃 규칙은 비균등 클래스 프라이어, 차등 오분류 비용, 및 특징 선택의 문제를 해결하도록 정밀화될 수 있다. 이들 정밀화 중 다수는 이웃에 대한 일부 형태의 가중 투표를 수반한다. 최근접 이웃 분석에 대한 보다 많은 정보에 대해서는, 문헌 [Duda, *Pattern*

Classification, Second Edition, 2001, John Wiley & Sons, Inc] 및 문헌 [Hastie, 2001, The Elements of Statistical Learning, Springer, New York]을 참조하며, 이들 각각은 본 출원에 참조로 포함되어 있다.

[0188] k-최근접 이웃 모델은 입력이 특징 공간 내의 k개의 가장 가까운 훈련 예로 이루어지는 비-파라미터 기계 학습 방법이다. 출력은 클래스 멤버십이다. 오브젝트는 그의 이웃의 복수의 투표에 의해 분류되며, 오브젝트는 그의 k개의 최근접 이웃 중에서 가장 흔한 클래스에 할당된다(k는 양의 정수, 전형적으로 작음). k = 1이면, 오브젝트는 단순히 그 단일 최근접 이웃의 클래스에 할당된다. 본원에 참조로 포함된 문헌 [Duda 등의 2001, *Pattern Classification*, Second Edition, John Wiley & Sons]을 참조한다. 일부 실시예에서, k-최근접 이웃 모델을 풀기 위해 필요한 거리 계산의 수는 머리로는 수행할 수 없기 때문에, 주어진 입력에 대한 모델을 풀기 위해 컴퓨터를 사용해야 한다.

[0189] *랜덤 포레스트, 의사결정 트리, 및 부스팅된 트리 모델.* 일부 실시예에서, 모델은 의사결정 트리이다. 모델로서 사용하기에 적합한 의사결정 트리는 일반적으로 문헌 [Duda, 2001, Pattern Classification, John Wiley & Sons, Inc., New York, pp. 395-396]에 설명되어 있으며, 이는 본원에 참조로 포함된다. 트리-기반 방법은 특징 공간을 직사각형의 세트로 분할하고, 이어서 각각의 것에서(상수와 유사) 모델을 피팅한다. 일부 실시예에서, 의사결정 트리는 랜덤 포레스트 회귀이다. 사용될 수 있는 한 특정 모델은 분류 및 회귀 트리(CART)이다. 다른 구체적인 의사결정 트리 모델은 ID3, C4.5, MART 및 랜덤 포레스트를 포함하나, 이에 제한되지는 않는다. CART, ID3 및 C4.5는 문헌 [Duda, 2001, Pattern Classification, John Wiley & Sons, Inc., New York, pp. 396-408 및 pp. 411-412]에 기재되어 있으며, 이는 본원에 참조로 포함된다. CART, MART, 및 C4.5는 문헌 [Hastie 등의 2001, The Elements of Statistical Learning, Springer-Verlag, New York, Chapter 9]에 기재되어 있으며, 이는 그 전문이 본원에 참조로 포함된다. 랜덤 포레스트는 그 전문이 본원에 참조로 포함되는 문헌 [Breiman, 1999, "Random Forests--Random Features," Technical Report 567, Statistics Department, U.C. Berkeley, September 1999]에 기재되어 있다. 일부 실시예에서, 의사결정 트리 모델은 적어도 10개, 적어도 20개, 적어도 50개, 또는 적어도 100개의 파라미터(예를 들어, 가중치 및/또는 결정)를 포함하고, 머리로는 해석할 수 없기 때문에 계산을 위해 컴퓨터가 필요하다.

[0190] *회귀.* 일부 실시예에서, 모델은 회귀를 사용한다. 회귀 알고리즘은 임의의 유형의 회귀일 수 있다. 예를 들어, 일부 실시예에서, 회귀는 로지스틱 회귀이다. 일부 실시예에서, 회귀는 라쏘, L2 또는 엘라스틱 넷 정규화를 사용한 로지스틱 회귀이다. 일부 실시예에서, 임계치 값을 충족시키지 못하는 상응하는 회귀 계수를 갖는 이들 추출된 특징은 프루닝된(제거된) 고려사항이다. 일부 실시예에서, 멀티카테고리 반응을 취급하는 로지스틱 회귀 모델의 일반화가 모델로서 사용된다. 로지스틱 회귀는 문헌 [Agresti, *An Introduction to Categorical Data Analysis*, 1996, Chapter 5, pp. 103-144, John Wiley & Son, New York]에 개시되어 있으며, 이는 본원에 참조로 포함된다. 일부 실시예에서, 모델은 문헌 [Hastie 등의 2001, *The Elements of Statistical Learning*, Springer-Verlag, New York]에 개시된 회귀 모델을 이용한다. 일부 실시예에서, 로지스틱 회귀 모델은 적어도 10개, 적어도 20개, 적어도 50개, 적어도 100개, 또는 적어도 1000개의 파라미터(예를 들어, 가중치)를 포함하고, 머리로는 해석할 수 없기 때문에 계산을 위해 컴퓨터가 필요하다.

[0191] *선형 판별 분석.* 선형 판별 분석(LDA), 정상 판별 분석(NDA), 또는 판별 함수 분석은 2개 이상의 클래스의 오브젝트 또는 이벤트를 특성화하거나 분리하는 특징의 선형 조합을 발견하기 위해 통계, 패턴 인식, 및 기계 학습에서 사용되는 방법인 피셔 선형 판별의 일반화일 수 있다. 생성된 조합은 본 개시의 일부 실시예에서 모델(선형 모델)로서 사용될 수 있다.

[0192] *혼합 모델 및 은닉 마르코프 모델.* 일부 실시예에서, 모델은 혼합 모델, 예컨대 문헌 [McLachlan 등의 Bioinformatics 18(3):413-422, 2002]에 설명된 것이다. 일부 실시예에서, 특히 시간 컴포넌트를 포함하는 실시예에서, 모델은 문헌 [Schliep 등의 2003, Bioinformatics 19(1):i255-i263]에 설명된 바와 같은 은닉 마르코프 모델이다.

[0193] *클러스터링.* 일부 실시예에서, 모델은 비지도 클러스터링 모델이다. 일부 실시예에서, 모델은 지도 클러스터링 모델이다. 모델로서 사용하기에 적합한 클러스터링은, 예를 들어 그 전문이 본원에 참조로 포함되는 문헌 [Duda and Hart, Pattern Classification and Scene Analysis, 1973, John Wiley & Sons, Inc., New York](이하 "Duda 1973")의 페이지 211-256에 기재되어 있다. 클러스터링 문제는 데이터세트에서 자연적 그룹화를 찾는 것 중 하나로서 설명될 수 있다. 자연적 그룹화를 식별하기 위해, 2가지 문제가 다루어질 수 있다. 먼저, 2개의 샘플 사이의 유사성(또는 비유사성)을 측정하는 방식을 결정할 수 있다. 이 메트릭(예를 들어, 유사성 척도)은 하나의 클러스터 내의 샘플이 다른 클러스터 내의 샘플에 대한 것보다 서로 더 유사함을 보장하기 위해

사용될 수 있다. 둘째, 유사성 척도를 사용하여 데이터를 클러스터로 분할하기 위한 메커니즘이 결정될 수 있다. 클러스터링 조사를 시작하는 한 방식은 거리 함수를 정의하고 훈련 세트의 모든 샘플 쌍 사이의 거리의 행렬을 계산하는 것일 수 있다. 거리가 유사성의 우수한 척도이면, 이때, 동일한 클러스터 내의 참조 엔티티 사이의 거리는 상이한 클러스터 내의 참조 엔티티 사이의 거리보다 유의하게 더 작을 수 있다. 그러나, 클러스터링은 거리 메트릭을 사용하지 않을 수 있다. 예를 들어, 비메트릭 유사성 함수 s(x, x')를 사용하여 2개의 벡터 x 및 x'를 비교할 수 있다. s(x, x')는 x 및 x'가 어떻게든 "유사"할 때 그 값이 큰 대칭 함수일 수 있다. 데이터세트에서 지점 사이의 "유사성" 또는 "비유사성"을 측정하는 방법이 선택되면, 클러스터링은 데이터의 임의의 분할의 클러스터링 품질을 측정하는 기준 함수를 사용할 수 있다. 기준 함수를 극단화하는 데이터 세트의 파티션이 데이터를 클러스터링하는데 사용될 수 있다. 본 개시에서 사용될 수 있는 특정한 예시적인 클러스터링 기술은 계층적 클러스터링(최근접-이웃 알고리즘, 최원접-이웃 알고리즘, 평균 연관 알고리즘, 중심 알고리즘, 또는 제곱합 알고리즘을 사용하는 응집 클러스터링), k-평균 클러스터링, 퍼지 k-평균 클러스터링, 및 자비스-패트릭 클러스터링을 포함할 수 있으나 이에 제한되지는 않는다. 일부 실시예에서, 클러스터링은 (예를 들어, 클러스터의 수를 미리 인지하지 않는 및/또는 클러스터 할당을 사전결정하지 않는) 비지도 클러스터링을 포함한다.

[0194] *모델 및 부스팅의 앙상블.* 일부 실시예에서, 모델의 앙상블(2개 이상)이 사용된다. 일부 실시예에서, 부스팅 기술, 예컨대 아다부스트(AdaBoost)가 모델의 성능을 개선시키기 위해 많은 다른 유형의 학습 알고리즘과 함께 사용된다. 이러한 접근법에서, 본원에 개시된 임의의 모델 또는 그의 등가물의 출력은 부스팅된 모델의 최종 출력을 나타내는 가중 합계로 조합된다. 일부 실시예에서, 모델로부터의 복수의 출력은 평균, 중앙값, 모드, 가중 평균, 가중 중앙값, 가중 모드 등을 포함하나 이에 제한되지는 않는 관련 기술분야에 공지된 중심 집중 경향의 임의의 척도를 사용하여 조합된다. 일부 실시예에서, 복수의 출력은 투표 방법을 사용하여 조합된다. 일부 실시예에서, 모델의 앙상블에서의 각각의 모델은 가중되거나 또는 가중되지 않는다.

[0195] 본원에 사용된 용어 "훈련되지 않은 모델"(예를 들어, "훈련되지 않은 리그레서" 및/또는 "훈련되지 않은 분류자")은 훈련 데이터세트에 대해 훈련되지 않은 기계 학습 모델, 예컨대 리그레서 또는 분류자를 지칭한다. 본원에 사용된 용어 "모델을 훈련시키는"은 훈련되지 않은 또는 부분적으로 훈련된 모델을 훈련시키는 과정을 지칭한다. 예를 들어, 일부 실시예에서, 모델을 훈련시키는 것은 잠재 표현으로 배열된 복수의 세포 구성성분 모듈 및 하기에서 논의되는 세포 구성성분 카운트 데이터 구조를 획득하는 것을 포함한다. 잠재 표현으로 배열된 복수의 세포 구성성분 모듈 및 세포 구성성분 카운트 데이터 구조는 조합되어 활성화 데이터 구조를 형성하고, 활성화 데이터 구조는, 활성화 데이터 구조(이하 "1차 훈련 데이터세트")에서 복수의 세포 구성성분 모듈에 대한 복수의 공변량에서의 각각의 공변량의 존재의 실제 부재와 함께, 훈련되지 않은 또는 부분적으로 훈련된 모델에 대한 집합적 입력으로서 적용되어, 공변량-모듈 상관에 대해 훈련되지 않은 또는 부분적으로 훈련된 모델을 훈련시킴으로써, 훈련된 모델을 획득한다. 또한, 용어 "훈련되지 않은 모델"은 전이 학습 기술이 훈련되지 않은 모델의 이러한 훈련에 사용되는 가능성을 배제하지 않는다는 것이 이해될 것이다. 예를 들어, 본원에 참조로 포함되는 문헌 [Fernandes 등의 2017, "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening," Pattern Recognition and Image Analysis: 8th Iberian Conference Proceedings, 243-250]은 이러한 전이 학습의 비제한적 예를 제공한다. 전이 학습이 사용되는 경우에, 앞서 설명된 훈련되지 않은 모델에는 1차 훈련 데이터세트의 데이터를 초과한, 그 이외의 추가 데이터가 제공된다. 즉, 전이 학습 실시예의 비제한적 예에서, 훈련되지 않은 모델은 (i) 1차 훈련 데이터세트 및 (ii) 추가 데이터를 수신한다. 전형적으로, 이러한 추가 데이터는 또 다른 보조 훈련 데이터세트로부터 학습된 계수(예를 들어, 회귀 계수)의 형태이다. 또한, 단일 보조 훈련 데이터세트의 설명이 개시되었지만, 본 개시에서 훈련되지 않은 모델을 훈련하는데 있어서 1차 훈련 데이터세트를 보완하기 위해 사용될 수 있는 보조 훈련 데이터세트의 수에 대한 제한은 없음을 이해할 것이다. 예를 들어, 일부 실시예에서, 2개 이상의 보조 훈련 데이터세트, 3개 이상의 보조 훈련 데이터세트, 4개 이상의 보조 훈련 데이터세트 또는 5개 이상의 보조 훈련 데이터세트가 전이 학습을 통해 1차 훈련 데이터세트를 보완하는데 사용되며, 각각의 이러한 보조 데이터세트는 1차 훈련 데이터세트와 상이하다. 임의의 방식의 전이 학습이 이러한 실시예에서 사용될 수 있다. 예를 들어, 1차 훈련 데이터세트에 더하여 제1 보조 훈련 데이터세트 및 제2 보조 훈련 데이터세트가 있는 경우를 고려한다. (회귀와 같은 모델을 제1 보조 훈련 데이터세트에 적용함으로써) 제1 보조 훈련 데이터세트로부터 학습된 계수는 전이 학습 기술(예를 들어, 2차원 행렬 곱셈)을 사용하여 제2 보조 훈련 데이터세트에 적용될 수 있으며, 이는 다시 그 계수가 그 후 1차 훈련 데이터세트에 적용되는 훈련된 중간 모델을 초래할 수 있고, 이는 1차 훈련 데이터세트 자체와 함께 훈련되지 않은 모델에 적용된다. 대안적으로, (회귀와 같은 모델을 제1 보조 훈련 데이터세트에 적용함으로써) 제1 보조 훈련 데이터세트로부터 학습된 계수의 제1 세트 및 (회귀와 같은 모델을 제2 보조 훈련 데이터세트에 적용

함으로써) 제2 보조 훈련 데이터세트로부터 학습된 계수의 제2 세트가 (예를 들어, 개별 독립 행렬 곱셈에 의해) 각각 개별적으로 1차 훈련 데이터세트의 개별 인스턴스에 적용될 수 있고, 1차 훈련 데이터세트 자체(또는 1차 훈련 세트로부터 학습된 주 컴포넌트 또는 회귀 계수와 같은 1차 훈련 데이터세트의 일부 감소된 형태)와 함께 1차 훈련 데이터세트의 개별 인스턴스에 대한 계수의 이러한 적용들 모두는 그 후 훈련되지 않은 모델을 훈련하기 위해 훈련되지 않은 모델에 적용될 수 있다. 어느 예에서나, 제1 및 제2 보조 훈련 데이터세트로부터 유도된 공변량-모듈 상관(예를 들어, 추가의 세포 상태 주석, 추가의 공변량, 및/또는 그의 세포 구성성분 풍부도 등)에 관한 지식이 훈련되지 않은 모델을 훈련시키기 위해 공변량-표지된 1차 훈련 데이터세트와 함께 사용된다.

[0196] 본원에서 상호교환가능하게 사용되는 용어 "뉴런", "노드", "단위", "은닉 뉴런", "은닉 단위" 등은 입력을 수용하고 활성화 함수 및 하나 이상의 파라미터(예를 들어, 계수 및/또는 가중치)를 통해 출력을 제공하는 신경망의 단위를 지칭한다. 예를 들어, 은닉 뉴런은 이전 계층으로부터의 하나 이상의 입력을 수용하고, 후속 계층에 대한 입력으로서 기능하는 출력을 제공할 수 있다. 일부 실시예에서, 신경망은 단지 1개의 출력 뉴런을 포함한다. 일부 실시예에서, 신경망은 복수의 출력 뉴런을 포함한다. 일반적으로, 출력은 예측 값, 예컨대 확률 또는 가능성, 이진 결정(예를 들어, 존재 또는 부재, 양성 또는 음성 결과), 및/또는 관심 조건, 예컨대 공변량, 세포 상태 주석, 또는 관심 세포 과정의 표지(예를 들어, 분류 및/또는 상관 계수)이다. 단일-클래스 분류 모델의 경우, 출력은 입력 특징(예를 들어, 하나 이상의 세포 구성성분 모듈)이 조건(예를 들어, 공변량, 세포 상태 주석, 및/또는 관심 세포 과정)을 가질 가능성(예를 들어, 상관 계수 및/또는 가중치)일 수 있다. 다중 클래스 분류 모델의 경우, 다수의 예측 값이 생성될 수 있고, 각각의 예측 값은 각각의 관심 조건에 대한 입력 특징의 가능성을 나타낸다.

[0197] 본원에 사용될 때, 용어 "파라미터"는 모델, 분류자, 또는 알고리즘 내의 하나 이상의 입력, 출력, 및/또는 기능에 영향을 미칠 수 있는(예를 들어, 수정, 맞춤화, 및/또는 조정할 수 있는) 모델, 분류자, 또는 알고리즘 내의 임의의 계수, 또는 유사하게 내부 또는 외부 요소의 임의의 값(예를 들어, 가중치 및/또는 하이퍼파라미터)을 의미한다. 일부 실시예에서, 파라미터는 모델에서 하나 이상의 입력, 출력 또는 함수를 조정하는 계수(예를 들어, 가중치)이다. 예를 들어, 파라미터의 값은 모델에 대한 입력(예를 들어, 특징)의 영향을 상향가중 또는 하향가중하는데 사용될 수 있다. 특징은 로지스틱 회귀, SVM, 또는 나이브 베이즈 모델에서와 같이 파라미터와 연관될 수 있다. 파라미터의 값은, 대안적으로 또는 추가적으로, 신경망 내의 노드(예를 들어, 여기서 노드는 입력의 출력으로의 변환을 정의하는 하나 이상의 활성화 함수를 포함함), 클래스, 또는 (예를 들어, 복수의 세포에서의 세포의) 인스턴스의 영향을 상향가중 또는 하향가중하는데 사용될 수 있다. 특정 입력, 출력, 기능, 또는 특징에 대한 파라미터의 할당은 주어진 모델에 대한 임의의 하나의 패러다임으로 제한되지 않고, 최적 성능을 위해 임의의 적합한 모델 아키텍처에서 사용될 수 있다. 일부 경우에, 모델의 입력, 출력, 함수, 또는 특징과 연관된 파라미터(예를 들어, 계수)에 대한 참조는, 예컨대 기계 학습 모델의 계산 복잡성에 관련하여, 그의 수, 성능, 또는 최적화의 지표로서 유사하게 사용될 수 있다. 일부 실시예에서, 파라미터는 고정된 값을 갖는다. 일부 실시예에서, 파라미터의 값은 (예를 들어, 하이퍼파라미터 최적화 방법을 사용하여) 수동으로 및/또는 자동으로 조정가능하다. 일부 실시예에서, 파라미터의 값은 모델 검증 및/또는 훈련 과정에 의해(예를 들어, 본원의 다른 곳에 설명된 바와 같은 오차 최소화 및/또는 역전파 방법에 의해) 수정된다.

[0198] 본원에 사용된 용어 "벡터"는 요소의 열거된 목록, 예컨대 요소의 어레이이며, 여기서 각각의 요소는 할당된 의미를 갖는다. 이와 같이, 본 개시에 사용된 용어 "벡터"는 용어 "텐서"와 상호교환가능하다. 예로서, 벡터가 복수의 세포에서 각각의 세포 구성성분에 대한 풍부도 카운트를 포함하는 경우에, 복수의 세포의 각각의 세포에 대해 벡터 내에 미리 결정된 요소가 존재한다. 제시의 용이성을 위해, 일부 경우에 벡터는 1차원인 것으로 설명될 수 있다. 그러나, 본 개시는 이에 제한되지 않는다. 벡터 내의 각각의 요소가 나타내는 것에 대한 설명이 정의된다면(예를 들어, 요소 1은 복수의 세포의 세포 1의 풍부도 카운트를 나타냄 등), 임의의 차원의 벡터가 본 개시에서 사용될 수 있다.

[0199] **I. 예시적인 시스템 실시예**

[0200] 이제, 본 개시의 일부 양태 및 본 개시에 사용된 일부 정의의 개요가 제공되었고, 예시적인 시스템의 세부사항이 도 1과 함께 설명된다.

[0201] 도 1은 본 개시의 일부 실시예에 따른 시스템(100)을 예시하는 블록도를 제공한다. 시스템(100)은 관심 세포 과정과 연관된 복수의 세포 구성성분 모듈에서 하나 이상의 세포 구성성분 모듈의 결정을 제공한다. 도 1에서, 시스템(100)은 컴퓨팅 장치로서 예시되어 있다. 컴퓨터 시스템(100)의 다른 토폴로지가 가능하다. 예를 들어,

일부 실시예에서, 시스템(100)은 사실상 네트워크에서 함께 연결되거나, 또는 클라우드 컴퓨팅 환경에서 가상 기계 또는 컨테이너인 여러 컴퓨터 시스템을 구성할 수 있다. 이와 같이, 도 1에 도시된 예시적인 토폴로지는 단지 관련 기술분야의 통상의 기술자에게 쉽게 이해될 방식으로 본 개시의 실시예의 특징을 설명하기 위해 제공된다.

[0202] 도 1을 참조하면, 일부 실시예에서 컴퓨터 시스템(100)(예를 들어, 컴퓨팅 장치)은 네트워크 인터페이스(104)를 포함한다. 일부 실시예에서, 네트워크 인터페이스(104)는 시스템 내의 시스템(100) 컴퓨팅 장치를 서로간에 뿐만 아니라 선택적인 외부 시스템 및 디바이스와 하나 이상의 통신 네트워크를 통해(예를 들어, 네트워크 통신 모듈(158)을 통해) 상호접속한다. 일부 실시예에서, 네트워크 인터페이스(104)는 인터넷, 하나 이상의 근거리 통신망(LAN), 하나 이상의 광역 네트워크(WAN), 다른 유형의 네트워크, 또는 이러한 네트워크의 조합을 통해 네트워크 통신 모듈(158)을 통한 통신을 선택적으로 제공한다.

[0203] 네트워크의 예에는 월드 와이드 웹(WWW), 인트라넷 및/또는 무선 네트워크, 예컨대 휴대 전화 네트워크, 무선 근거리 통신망(LAN) 및/또는 대도시 통신망(MAN), 및 무선 통신에 의한 다른 디바이스가 포함된다. 무선 통신은 GSM(Global System for Mobile Communications), EDGE(Enhanced Data GSM Environment), HSDPA(high-speed downlink packet access), HSUPA(high-speed uplink packet access), EV-DO(Evolution, Data-Only), HSPA, HSPA+, DC-HSPDA(Dual-Cell HSPA), LTE(long term evolution), NFC(near field communication), W-CDMA(wideband code division multiple access), CDMA(code division multiple access), TDMA(time division multiple access), Bluetooth, 무선 충실도(Wi-Fi)(예를 들어, IEEE 802.11a, IEEE 802.11ac, IEEE 802.11ax, IEEE 802.11b, IEEE 802.11g 및/또는 IEEE 802.11n), VoIP(voice over Internet Protocol), Wi-MAX, e-메일용 프로토콜(예를 들어, IMAP (Internet message access protocol) 및/또는 POP(post office protocol)), 인스턴트 메시징(예를 들어, XMPP(extensible messaging and presence protocol), SIMPLE(Session Initiation Protocol for Instant Messaging and Presence Leveraging Extensions), IMPS (Instant Messaging and Presence Service)), 및/또는 SMS(Short Message Service), 또는 본 명세서의 출원일 현재 아직 개발되지 않은 통신 프로토콜을 포함하는 임의의 다른 적합한 통신 프로토콜을 포함하는, 복수의 통신 표준, 프로토콜 및 기술 중 임의의 것을 선택적으로 사용한다.

[0204] 시스템(100)은 일부 실시예에서 하나 이상의 처리 유닛(CPU(들))(102)(예를 들어, 프로세서, 처리 코어 등), 하나 이상의 네트워크 인터페이스(104), (선택적으로) 사용자에 의한 사용을 위한 디스플레이(108) 및 입력 시스템(105)(예를 들어, 입력/출력 인터페이스, 키보드, 마우스 등)을 포함하는 사용자 인터페이스(106), 메모리(예를 들어, 비-영구 메모리(107), 영구 메모리(109), 및 상기 언급된 컴포넌트를 상호접속하기 위한 하나 이상의 통신 버스(103)를 포함한다. 하나 이상의 통신 버스(103)는 시스템 컴포넌트 사이의 통신을 상호접속하고 제어하는 회로부(때때로 칩셋이라고 함)를 선택적으로 포함한다. 비-영구 메모리(107)는 전형적으로 고속 랜덤 액세스 메모리, 예컨대 DRAM, SRAM, DDR RAM, ROM, EEPROM, 플래쉬 메모리를 포함하는 반면, 영구 메모리(109)는 전형적으로 CD-ROM, 디지털 다기능 디스크(DVD) 또는 다른 광학 스토리지, 자기 카세트, 자기 테이프, 자기 디스크 저장 디바이스 또는 다른 자기 저장 디바이스, 자기 디스크 저장 디바이스, 광학 디스크 저장 디바이스, 플래쉬 메모리 디바이스, 또는 다른 비휘발성 솔리드 스테이트 저장 디바이스를 포함한다. 영구 메모리(109)는 선택적으로 CPU(들)(102)로부터 원격으로 위치된 하나 이상의 저장 디바이스를 포함한다. 영구 메모리(109), 및 비-영구 메모리(109) 내의 비휘발성 메모리 디바이스(들)는 비-일시적 컴퓨터 판독가능 저장 매체를 포함한다. 일부 실시예에서, 비-영구 메모리(107) 또는 대안적으로 비-일시적 컴퓨터 판독가능 저장 매체는 때때로 영구 메모리(109)와 함께 다음 프로그램, 모듈 및 데이터 구조, 또는 그의 서브세트를 저장한다:

[0205] · 다양한 기본 시스템 서비스를 핸들링하고 하드웨어 의존적 태스크를 수행하기 위한 프로시저를 포함하는, 선택적 운영 체제(156)(예를 들어, ANDROID, iOS, DARWIN, RTXC, LINUX, UNIX, OS X, WINDOWS, 또는 VxWorks와 같은 임베디드 운영 체제);

[0206] · 시스템(100)을 다른 디바이스 및/또는 통신 네트워크(104)와 연결하기 위한 선택적 네트워크 통신 모듈(또는 명령어)(158);

[0207] · 복수의 화합물의 각각의 화합물에 대한 각각의 화학 구조(122)(예를 들어, (122-1), …(122-R)) 또는 그 표현(예를 들어, 화학 구조의 지문)을 포함하는 화합물 구조 데이터 저장소(120);

[0208] · 세포 구성성분 모듈(132)의 세트(예를 들어, 132-1, … 132-K)를 포함하는 세포 구성성분 모듈 데이터 저장소(130)- 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분(134)(예를 들어, 134-1-1,… 134-1-Z)의 서브세트를 포함함 -;

[0209] · 교란 시그니처의 세트(142)(예를 들어, 142-1, … 142-P)를 포함하는 교란 데이터 저장소(140)- 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별, 및 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수(144)(예를 들어, 144-1-1, … 144-1-Q)를 포함함 -;

[0210] · 복수의 화합물의 각각의 개별 화합물에 대한 각각의 개별 화학 구조(152)(예를 들어, 152-1, … 152-R)에 대해 다음을 포함하는 활성화 데이터 구조(150):

[0211] o 선택적으로, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해, 각각의 수치 활성화 점수(154)(예를 들어, 154-1-1, … 154-1-K), 및/또는

[0212] o 선택적으로, 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해, 각각의 수치 활성화 점수(156)(예를 들어, 156-1-1, … 156-1-P); 및

[0213] · 복수의 파라미터(예를 들어, 100개 이상의 파라미터)를 포함하는 모델- 복수의 파라미터는 계산된 활성화 점수와 각각의 화학 구조에 대한 수치 활성화 점수 사이의 차이에 응답하여 조정됨 -.

[0214] 다양한 실시예에서, 상기 식별된 요소 중 하나 이상이 이전에 언급된 메모리 디바이스 중 하나 이상에 저장되고, 상기 설명된 기능을 수행하기 위한 명령어 세트에 상응한다. 상기 식별된 모듈, 데이터, 또는 프로그램(예를 들어, 명령어 세트)은 개별 소프트웨어 프로그램, 프로시저, 데이터세트, 또는 모듈로서 구현될 필요가 없고, 따라서 이들 모듈 및 데이터의 다양한 서브세트가 다양한 구현에서 조합되거나 그렇지 않으면 재배열될 수 있다. 일부 구현에서, 비-영구 메모리(107)는 위에서 식별된 모듈 및 데이터 구조의 서브세트를 선택적으로 저장한다. 또한, 일부 실시예에서, 메모리는 앞서 설명되지 않은 추가의 모듈 및 데이터 구조를 저장한다. 일부 실시예에서, 상기에 식별된 요소 중 하나 이상은 시스템(100)에 의해 어드레싱가능한, 시스템(100)의 컴퓨터 시스템 이외의 컴퓨터 시스템에 저장되어, 시스템(100)은 필요할 때 이러한 데이터의 전부 또는 일부를 검색할 수 있다.

[0215] 도 1은 "시스템(100)"을 도시하지만, 도면은 본원에 설명된 구현의 구조적 개략도보다는 컴퓨터 시스템에 존재할 수 있는 다양한 특징에 대한 기능적 설명을 더 의도하는 것이다. 실제로, 그리고 관련 기술분야의 통상의 기술자에 의해 인식되는 바와 같이, 개별적으로 제시된 항목은 조합될 수 있고, 일부 항목은 분리될 수 있다. 또한, 도 1은 비-영구 메모리(107) 내의 특정 데이터 및 모듈을 도시하지만, 이들 데이터 및 모듈의 일부 또는 전부는 대신에 영구 메모리(109) 또는 둘 이상의 메모리에 저장될 수 있다. 예를 들어, 일부 실시예에서, 적어도 화합물 구조 데이터 저장소(120) 및 활성화 데이터 구조(150)는 클라우드-기반 인프라구조의 일부일 수 있는 원격 저장 디바이스에 저장된다. 일부 실시예에서, 적어도 화합물 구조 데이터 저장소(120) 및 활성화 데이터 구조(150)는 클라우드-기반 인프라구조 상에 저장된다. 일부 실시예에서, 화합물 구조 데이터 저장소(120) 및 활성화 데이터 구조(150)는 또한 원격 저장 디바이스(들)에 저장될 수 있다.

[0216] 본 개시에 따른 시스템이 도 1을 참조하여 개시되었고, 본 개시에 따른 방법(200, 300, 700, 800, 900, 및 1500)을 이제 도 2, 3, 7, 8, 9, 및 14를 참조하여 상세히 설명한다.

[0217] **II. 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법**

[0218] *생리학적 조건.*

[0219] 도 3a-3e를 참조하면, 본 개시의 한 양태는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법(300)을 제공한다.

[0220] 일부 실시예에서, 관심 생리학적 조건은 질환이다.

[0221] 일부 실시예에서, 질환은 감염성 또는 기생충성 질환; 신생물; 혈액 또는 혈액-형성 기관의 질환; 면역계의 질환; 내분비, 영양 또는 대사 질환; 정신, 행동 또는 신경발달 장애; 수면-각성 장애; 신경계의 질환; 시각계의 질환; 귀 또는 유양돌기의 질환; 순환계의 질환; 호흡계의 질환; 소화계의 질환; 피부의 질환; 근골격계 또는 결합 조직의 질환; 비뇨생식기계의 질환; 성적 건강과 관련된 상태; 임신, 출산 또는 산후기와 관련된 질환; 주산기 기간 기원의 특정 상태; 및 발달 이상으로 이루어진 그룹으로부터 선택된다. 일부 실시예에서, 질환은 ICD-11 MMS 또는 국제 질환 분류의 하나 이상의 엔트리이다. ICD는 질환, 손상 및 사망 원인을 분류하는 방법을 제공한다. 세계 보건 기구(WHO)는 진단된 질환의 사례를 기록하고 추적하는 방법을 표준화하기 위해 ICD를

공개한다.

[0222] 일부 실시예에서, 관심 생리학적 조건은 질환 자극제, 예컨대 질환 전제조건 또는 동반이환이다.

[0223] 일부 실시예에서, 관심 생리학적 조건은 세포 시스템에서 발생하거나, 또는 세포 시스템과 관련하여 측정된다. 일부 실시예에서, 관심 생리학적 조건은 하나 이상의 세포에서 발생하거나 또는 그와 관련하여 측정되며, 여기서 하나 이상의 세포는 단일 세포, 세포주, 생검 샘플 세포 및/또는 배양된 1차 세포를 포함한다. 일부 실시예에서, 관심 생리학적 조건은 인간 세포에서 발생하는 생리학적 조건이다. 일부 실시예에서, 관심 생리학적 조건은 샘플, 예컨대 본원에 설명된 임의의 샘플에서 발생하는 생리학적 조건이다(예를 들어, 정의: 샘플 참조). 일부 실시예에서, 관심 생리학적 조건은 대상, 예컨대 인간 또는 동물에서 발생하는 생리학적 조건이다.

[0224] 일부 실시예에서, 관심 생리학적 조건은 관심 세포 과정이거나, 또는 그와 관련된다.

[0225] 일부 실시예에서, 관심 세포 과정은 이상 세포 과정이다. 일부 실시예에서, 관심 세포 과정은 질환과 연관된 세포 과정이다. 예를 들어, 앞서 설명된 바와 같이, 일부 실시예에서, 방법은 질환에 중요한 세포 과정 및 프로그램의 표적화 및 설명을 제공한다. 일부 실시예에서, 관심 세포 과정은 질환의 발병, 진행, 증상, 중증도 및/또는 해소를 포함하나 이에 제한되지는 않는 질환의 임의의 특징의 근간이 되는 메커니즘을 나타내거나 또는 그와 관련된다. 일부 실시예에서, 관심 세포 과정은 기능적 경로이다. 일부 실시예에서, 관심 세포 과정은 신호전달 경로이다. 일부 실시예에서, 관심 세포 과정은(예를 들어, 화합물, 소분자 및/또는 치료제의) 작용 메커니즘이다. 일부 실시예에서, 관심 세포 과정은 전사 네트워크(예를 들어, 유전자 조절 네트워크)에 의해 특징화되고/거나 조절된다. 일부 실시예에서, 관심 세포 과정은 제1 세포 상태와 제2 세포 상태 사이의 전이 동안 발생하는 세포 과정이다.

[0226] 일부 실시예에서, 관심 세포 과정은 주석, 예컨대 유전자 세트 풍부화 검정(GSEA) 주석, 유전자 온톨로지 주석, 기능적 및/또는 신호전달 경로 주석, 및/또는 세포 시그니처 주석이다. 주석은 NIH 유전자 발현 옴니버스(GEO), EBI ArrayExpress, NCBI, BLAST, EMBL-EBI, GenBank, Ensembl, KEGG 경로 데이터베이스, LINCS(Library of Integrated Network-based Cellular Signatures) L1000 데이터세트, 리액톰 경로 데이터베이스, 유전자 온톨로지 프로젝트 및/또는 임의의 질환-특정 데이터베이스를 포함하나 이에 제한되지는 않는 임의의 공공 지식 데이터베이스로부터 획득될 수 있다.

[0227] 따라서, 일부 실시예에서, 관심 생리학적 조건은 본원에 설명된 바와 같은 임의의 각각의 질환, 기능적 경로, 신호전달 경로, 작용 메커니즘, 전사 네트워크, 불일치, 및/또는 세포 또는 생물학적 과정이다.

[0228] 일부 실시예에서, 관심 생리학적 조건은 표현형이다. 예를 들어, 일부 실시예에서, 관심 생리학적 조건은 화합물, 소분자 및/또는 치료제의 생리학적 징후, 예컨대 질환의 독성 및/또는 해소이다. 일부 실시예에서, 생리학적 조건은 유동 세포측정법 판독, 영상화 및 현미경검사 주석(예를 들어, H&E 슬라이드, IHC 슬라이드, 방사선학 영상, 및/또는 다른 의료 영상화), 및/또는 세포 구성성분 데이터를 포함하나 이에 제한되지는 않는 실험 데이터를 사용하여 측정된 표현형이다.

[0229] 일부 실시예에서, 관심 생리학적 조건은 독성의 척도이다. 일부 실시예에서, 생리학적 조건은 핵 수용체의 억제 또는 활성화, 및/또는 핵 수용체의 억제의 양 또는 활성화의 양이다. 일부 실시예에서, 생리학적 조건은 생물학적 경로(예를 들어, 스트레스 반응 경로)의 억제 또는 활성화, 및/또는 억제의 양 또는 활성화의 양이다. 예시적인 핵 수용체 및 예시적인 스트레스 반응 경로, 뿐만 아니라 이들 핵 수용체에 대한 억제 또는 활성화 데이터 및 본 개시에 사용될 수 있는 예시적인 스트레스 반응 경로는 문헌 [Huang 등의 2016, "Modelling the Tox21 10 K chemical profiles for *in vivo* toxicity prediction and mechanism characterization," Nat Commun. 7, p. 10425]에 설명된 바와 같이 대략 10,000개의 화합물에 대해 기재되어 있으며, 이 문헌은 본 명세서에 참조로 포함된다.

[0230] 일부 실시예에서, 관심 생리학적 조건은 세포 구성성분의 세트(예를 들어, 세포 구성성분 모듈)의 활성화 및/또는 교란 시그니처(예를 들어, 교란에 반응한 복수의 분석물의 차등 발현 프로파일)을 특징으로 한다.

[0231] 예를 들어, 일부 실시예에서, 관심 생리학적 조건은 세포 구성성분의 세트를 포함하는 세포 구성성분 모듈이다. 임의의 유형의 분석물(예를 들어, 유전자, 전사체, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질, 또는 이들의 조합)이 각각의 세포 구성성분 모듈 내의 세포 구성성분의 세트에서의 사용에 고려된다. 일부 실시예에서, 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 세포 구성성분 모듈은 관련 기술분야에 공지된 임의의 세포 또는 생물학적 과정, 뿐만 아니라 그 임의의 이상과 연관된다. 본원에 개시된 시스템 및 방법과 함께 사용하기에 적합한 세포 구성성분 모듈은 하기 명칭 "세포 구성성분 및 세포 구성성분 모듈" 섹션에 추가로 기재

되어 있다.

[0232] 일부 실시예에서, 관심 생리학적 조건은 제1 세포 상태와 제2 세포 상태 사이의 불일치를 특징으로 하는 교란 시그니처(예를 들어, 세포 상태 전이 시그니처)이다.

[0233] 일부 이러한 실시예에서, 관심 생리학적 조건은 이환 상태(예를 들어, 이환 대상 및/또는 이환 조직으로부터 획득된 세포)와 건강한 상태(예를 들어, 건강한 또는 대조군 대상 및/또는 조직으로부터 획득된 세포) 사이의 불일치에 의해 식별된다. 예를 들어, 일부 실시예에서, 이환된 상태는 세포의 기능의 손실, 세포의 기능의 획득, 세포의 진행(예를 들어, 세포의 분화된 상태로의 전이), 세포의 정체(예를 들어, 세포의 분화된 상태의 전이 불능), 세포의 침입(예를 들어, 비정상적 위치에서의 세포의 출현), 세포의 소멸(예를 들어, 세포가 정상적으로 존재하는 위치에서의 세포의 부재), 세포의 장애(예를 들어, 세포 내 및/또는 주변의 구조적, 형태학적 및/또는 공간적 변화), 세포의 네트워크의 손실(예를 들어, 자손 세포 또는 세포 하류의 세포에서의 정상 효과를 제거하는 세포의 변화), 세포의 네트워크의 획득(예를 들어, 세포 하류의 세포의 자손 세포에서 새로운 하류 효과를 촉발하는 세포의 변화), 세포의 과잉(예를 들어, 세포의 과잉), 세포의 결핍(예를 들어, 세포의 밀도가 임계 임계치 미만임), 세포에서의 세포 구성성분 비 및/또는 양의 차이, 세포에서의 전이 속도의 차이, 또는 그 임의의 조합에 의해 식별된다.

[0234] 본원에 개시된 시스템 및 방법과 함께 사용하기에 적합한 교란 시그니처는 하기 "교란 시그니처"라는 명칭의 섹션에 추가로 기재되어 있다.

[0235] 일부 실시예에서, 관심 생리학적 조건은 복수의 생리학적 조건(예를 들어, 세포 과정, 세포 구성성분 모듈, 및/또는 교란 시그니처)을 포함한다. 일부 실시예에서, 관심 생리학적 조건은 적어도 3개, 적어도 4개, 적어도 5개, 적어도 6개, 적어도 7개, 적어도 8개, 적어도 9개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 또는 적어도 100개의 생리학적 조건을 포함한다. 일부 실시예에서, 관심 생리학적 조건은 200개 이하, 100개 이하, 90개 이하, 80개 이하, 70개 이하, 60개 이하, 50개 이하, 20개 이하, 또는 10개 이하의 생리학적 조건을 포함한다. 일부 실시예에서, 관심 생리학적 조건은 1개 내지 5개, 5개 내지 10개, 2개 내지 20개, 10개 내지 50개, 또는 20개 내지 100개의 생리학적 조건을 포함한다. 일부 실시예에서, 관심 생리학적 조건은 3가지 이상의 생리학적 조건에서 시작하여 200가지 이하의 생리학적 조건에서 끝나는 또 다른 범위 내에 속하는 복수의 생리학적 조건을 포함한다.

[0236] 일부 실시예에서, 본 개시의 화합물은 리핀스키 5 준칙을 충족시키는 화학적 화합물이다. 일부 실시예에서, 본 개시의 화합물은 다음과 같은 리핀스키 5 준칙 중 2개 이상의 규칙, 3개 이상의 규칙, 또는 모든 4개의 규칙을 충족시키는 유기 화합물이다: (i) 5개 이하의 수소 결합 공여자(예를 들어, OH 및 NH 기), (ii) 10개 이하의 수소 결합 수용자(예를 들어, N 및 O), (iii) 500 달톤 미만의 분자량, 및 (iv) 5 미만의 LogP. 4개의 기준 중 3개에 숫자 5가 관여되기 때문에 "5의 규칙"이라 지칭된다. 문헌 [Lipinski, 1997, Adv. Drug Del. Rev. 23, 3]을 참조하며, 이는 그 전문이 본원에 참조로 포함된다. 일부 실시예에서, 본 개시의 화합물은 리핀스키의 5 준칙 외에 하나 이상의 기준을 충족한다. 예를 들어, 일부 실시예에서, 본 개시의 화합물은 5개 이하의 방향족 고리, 4개 이하의 방향족 고리, 3개 이하의 방향족 고리, 또는 2개 이하의 방향족 고리를 갖는다.

[0237] 블록 302를 참조하면, 방법(300)은 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 것을 포함한다.

[0238] 예를 들어, 일부 구현에서, 기계 학습 접근법에 대한 테스트 화학적 화합물의 적용은 분자 데이터(예를 들어, 화합물의 화학 구조)를 기계 학습 모델에 의해 판독 및 조작될 수 있는 포맷으로 변환시키는 것을 포함한다.

[0239] 도 3a의 블록 304를 참조하면, 화학 구조를 기계 학습-판독가능 포맷으로 변환시키는 한 접근법은 분자를 텍스트의 스트링으로서 나타내는 단순화된 분자-입력 라인-엔트리 시스템(SMILES)을 사용하여 화학 구조의 "지문"을 결정하는 것을 포함한다. 따라서, 일부 실시예에서, 방법은 테스트 화학적 화합물의 단순화된 분자-입력 라인-엔트리 시스템(SMILES) 스트링 표현으로부터 지문을 계산하는 것을 더 포함한다. SMILES 스트링을 사용한 분자 핑거프린팅은, 예를 들어 그 전문이 본원에 참조로 포함되는 문헌 [Honda 등의 2019, "SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery," arXiv:1911.04738]에 추가로 기재되어 있다.

[0240] 화학 구조를 기계-학습 판독가능 포맷으로 변환시키는 또 다른 접근법은 그래프-기반 분자 지문을 결정하는 것을 포함한다. 그래프-기반 분자 핑거프린팅에서, 원래의 분자 구조는 그래프에 의해 나타내어지며, 여기서 노드는 개별 원자를 나타내고, 에지는 원자 사이의 결합을 나타낸다. 그래프-기반 접근법은 보다 낮은 크기 요건 및 따라서 보다 낮은 계산 부담을 갖는 다중 하위구조를 효율적으로 인코딩하는 능력, 뿐만 아니라 지문 사이의

구조적 유사성의 지표를 인코딩하는 능력을 포함한 여러 이점을 제공한다. 그래프-기반 핑거프린팅은, 예를 들어, 그 전문이 본원에 참조로 포함되는 문헌 [Duvenaud 등의 2015, "Convolutional networks on graphs for learning molecular fingerprints," NeurIPS, 2224-2232]에 추가로 기재되어 있다. 일부 실시예에서, 지문은 그래프 컨볼루션 네트워크로부터 생성된다. 일부 실시예에서, 지문은 공간 그래프 컨볼루션 네트워크, 예컨대 그래프 주의 네트워크(GAT), 그래프 동형 네트워크(GIN), 또는 그래프 하위구조 인덱스-기반 근사 그래프(SAGA)로부터 생성된다. 일부 실시예에서, 지문은 Chebyshev 다항식 필터링을 사용하는 스펙트럼 그래프 컨볼루션과 같은 스펙트럼 그래프 컨볼루션 네트워크로부터 생성된다.

[0241] 도 3a의 블록 306을 참조하면, 일부 실시예에서, 지문은 SMILES 변환기, ECFP4, RNNS2S 및/또는 GraphConv를 사용하여 화학 구조로부터 생성된다.

[0242] *모델 아키텍처.*

[0243] 도 3b의 블록 308을 참조하면, 이 방법은 지문을 모델에 입력하는 단계를 포함한다. 일부 실시예에서, 모델은 복수의(예를 들어, 100, 200, 300, 500, 1000, 10,000 또는 그 이상의) 파라미터를 포함한다.

[0244] 일부 실시예에서, 모델은 복수의 파라미터(예를 들어, 가중치 및/또는 하이퍼파라미터)를 포함한다. 일부 실시예에서, 모델에 대한 복수의 파라미터는 적어도 10, 적어도 50, 적어도 100, 적어도 500, 적어도 1000, 적어도 2000, 적어도 5000, 적어도 10,000, 적어도 20,000, 적어도 50,000, 적어도 100,000, 적어도 200,000, 적어도 500,000, 적어도 1백만, 적어도 2백만, 적어도 3백만, 적어도 4백만 또는 적어도 5백만개의 파라미터를 포함한다. 일부 실시예에서, 모델에 대한 복수의 파라미터는 8백만개 이하, 5백만개 이하, 4백만개 이하, 1백만개 이하, 500,000개 이하, 100,000개 이하, 50,000개 이하, 10,000개 이하, 5000개 이하, 1000개 이하, 또는 500개 이하의 파라미터를 포함한다. 일부 실시예에서, 모델에 대한 복수의 파라미터는 10개 내지 5000개, 500개 내지 10,000개, 10,000개 내지 500,000개, 20,000개 내지 1백만개, 또는 1백만개 내지 5백만개의 파라미터를 포함한다. 일부 실시예에서, 모델에 대한 복수의 파라미터는 10개 이상의 파라미터에서 시작하여 8백만개 이하의 파라미터에서 끝나는 또 다른 범위 내에 속한다.

[0245] 일부 실시예에서, 모델의 훈련은 하나 이상의 하이퍼파라미터(예를 들어, 훈련 동안 튜닝될 수 있는 하나 이상의 값)를 추가로 특징으로 한다. 일부 실시예에서, 하이퍼파라미터 값은 훈련 동안 튜닝(예를 들어, 조정)된다. 일부 실시예에서, 하이퍼파라미터 값은 훈련 데이터세트의 특정 요소 및/또는 하나 이상의 입력(예를 들어, 세포, 세포 구성성분 모듈, 공변량 등)에 기초하여 결정된다. 일부 실시예에서, 하이퍼파라미터 값은 실험적 최적화를 사용하여 결정된다. 일부 실시예에서, 하이퍼파라미터 값은 하이퍼파라미터 스위프를 사용하여 결정된다. 일부 실시예에서, 하이퍼파라미터 값은 이전 템플릿 또는 디폴트 값에 기초하여 할당된다.

[0246] 일부 실시예에서, 하나 이상의 하이퍼파라미터의 각각의 하이퍼파라미터는 학습율을 포함한다. 일부 실시예에서, 학습율은 적어도 0.0001, 적어도 0.0005, 적어도 0.001, 적어도 0.005, 적어도 0.01, 적어도 0.05, 적어도 0.1, 적어도 0.2, 적어도 0.3, 적어도 0.4, 적어도 0.5, 적어도 0.6, 적어도 0.7, 적어도 0.8, 적어도 0.9, 또는 적어도 1이다. 일부 실시예에서, 학습율은 1 이하, 0.9 이하, 0.8 이하, 0.7 이하, 0.6 이하, 0.5 이하, 0.4 이하, 0.3 이하, 0.2 이하, 0.1 이하, 0.05 이하, 0.01 이하, 또는 그 미만이다. 일부 실시예에서, 학습율은 0.0001 내지 0.01, 0.001 내지 0.5, 0.001 내지 0.01, 0.005 내지 0.8, 또는 0.005 내지 1이다. 일부 실시예에서, 학습율은 0.0001 이상에서 시작하여 1 이하에서 끝나는 또 다른 범위 내에 속한다. 일부 실시예에서, 하나 이상의 하이퍼파라미터는 정규화 강도(예를 들어, L2 가중치 페널티, 탈락률 등)를 더 포함한다. 예를 들어, 일부 실시예에서, 모델(예를 들어, 신경망)은 복수의 은닉 뉴런에서 각각의 은닉 뉴런의 상응하는 파라미터(예를 들어, 가중치)에 대한 정규화를 사용하여 훈련된다. 일부 실시예에서, 정규화는 L1 또는 L2 페널티를 포함한다.

[0247] 일부 실시예에서, 하나 이상의 하이퍼파라미터의 각각의 하이퍼파라미터는 손실 함수이다. 일부 실시예에서, 손실 함수는 평균 제곱 오차, 평탄화 평균 제곱 오차, 2차 손실, 평균 절대 오차, 평균 바이어스 오차, 힌지, 다중 클래스 서포트 벡터 머신, 및/또는 교차-엔트로피이다. 일부 실시예에서, 손실 함수는 구배 하강 알고리즘 및/또는 최소화 함수이다.

[0248] 일부 실시예에서, 모델은 하나 이상의 활성화 함수와 연관된다. 일부 실시예에서, 하나 이상의 활성화 함수에서의 활성화 함수는 tanh, 시그모이드, 소프트맥스, 가우시안, 볼츠만-가중 평균, 절대값, 선형, 정류 선형 단위(ReLU), 바운딩된 정류 선형, 소프트 정류 선형, 파라미터화된 정류 선형, 평균, max, min, 부호, 제곱, 제곱근, 멀티쿼드릭, 역 2차, 역 멀티쿼드릭, 폴리하모닉 스플라인, 스위시, 미시, 가우시안 오차 선형 단위(GeLU),

및/또는 박판 스플라인이다. 모델은 모델로의 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력한다.

[0249] 도 3b의 블록 310을 참조하면, 일부 실시예에서, 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 및/또는 선형 회귀 모델을 포함한다. 일부 실시예에서, 모델은 리그레서이다. 일부 실시예에서, 모델은 본원에 개시된 임의의 모델이다(예를 들어, 정의: 모델 참조).

[0250] 도 3b의 블록 312를 참조하면, 일부 실시예에서, 모델은 신경망을 포함한다.

[0251] 일부 실시예에서, 신경망은 ReLU 활성화를 갖는 완전 연결 신경망이다. 예를 들어, 일부 실시예에서, 모델은 상응하는 하나 이상의 입력을 포함하는 신경망이며, 여기서 상응하는 하나 이상의 입력에서의 각각의 입력은 테스트 화학적 화합물에 대한 화학 구조를 위한 것이고, 상응하는 제1 은닉 계층은 상응하는 복수의 은닉 뉴런을 포함하며, 여기서 상응하는 복수의 은닉 뉴런에서의 각각의 은닉 뉴런은 (i) 복수의 입력에서의 각각의 입력에 완전히 연결되고, (ii) 제1 활성화 함수 유형과 연관되고, (iii) 신경망에 대한 복수의 파라미터에서의 상응하는 파라미터(예를 들어, 가중치) 및 하나 이상의 상응하는 신경망 출력과 연관되며, 하나 이상의 상응하는 신경망 출력 중 각각의 개별 신경망 출력은 (i) 상응하는 복수의 은닉 뉴런에서의 각각의 은닉 뉴런의 출력을 입력으로서 직접 또는 간접적으로 수신하고, (ii) 제2 활성화 함수 유형과 연관된다. 일부 이러한 실시예에서, 신경망은 완전히 연결된 네트워크이다.

[0252] 일부 실시예에서, 신경망은 복수의 은닉 계층을 포함한다. 앞서 설명된 바와 같이, 은닉 계층은 입력 계층과 출력 계층 사이에 위치한다(예를 들어, 추가의 복잡성을 포착하기 위해). 일부 실시예에서, 복수의 은닉 계층이 있는 경우, 각각의 은닉 계층은 동일하거나 상이한 각각의 수의 뉴런을 가질 수 있다.

[0253] 일부 실시예에서, 각각의 은닉 뉴런(예를 들어, 신경망 내의 각각의 은닉 계층 내)은 입력 데이터에 대한 함수(예를 들어, 선형 또는 비선형 함수)를 수행하는 활성화 함수와 연관된다. 일반적으로, 활성화 함수의 목적은 신경망이 원래의 데이터의 표현에 대해 훈련되고 후속하여 새로운(예를 들어, 이전에 보지 못한) 데이터의 추가 표현을 "피팅" 또는 생성할 수 있도록 비선형성을 데이터에 도입하는 것이다. 활성화 함수(예를 들어, 제1 및/또는 제2 활성화 함수)의 선택은 신경망의 사용 사례에 의존하는데, 그 이유는 특정 활성화 함수가 데이터세트(예를 들어, tanh 및/또는 시그모이드 함수)의 극단 단부에서 포화로 이어질 수 있기 때문이다. 예를 들어, 일부 실시예에서, 활성화 함수(예를 들어, 제1 및/또는 제2 활성화 함수)는 본원에 개시된 임의의 활성화 함수를 포함하나 이에 제한되지는 않는, 관련 기술분야에 공지된 임의의 적합한 활성화 함수로부터 선택된다.

[0254] 일부 실시예에서, 각각의 은닉 뉴런은 활성화 함수에 기초하여 결정된, 신경망의 출력에 기여하는 파라미터(예를 들어, 가중치 및/또는 바이어스 값)와 추가로 연관된다. 일부 실시예에서, 은닉 뉴런은 임의의 파라미터(예를 들어, 무작위화된 가중치)로 초기화된다. 일부 대안적 실시예에서, 은닉 뉴런은 미리 결정된 파라미터 세트로 초기화된다.

[0255] 일부 실시예에서, 신경망 내의(예를 들어, 하나 이상의 은닉 계층에 걸친) 복수의 은닉 뉴런은 적어도 2개, 적어도 3개, 적어도 4개, 적어도 5개, 적어도 6개, 적어도 7개, 적어도 8개, 적어도 9개, 적어도 10개, 적어도 11개, 적어도 12개, 적어도 13개, 적어도 14개, 적어도 15개, 적어도 16개, 적어도 17개, 적어도 18개, 적어도 19개, 적어도 20개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 또는 적어도 500개의 뉴런이다. 일부 실시예에서, 복수의 은닉 뉴런은 적어도 100개, 적어도 500개, 적어도 800개, 적어도 1000개, 적어도 2000개, 적어도 3000개, 적어도 4000개, 적어도 5000개, 적어도 6000개, 적어도 7000개, 적어도 8000개, 적어도 9000개, 적어도 10,000개, 적어도 15,000개, 적어도 20,000개 또는 적어도 30,000개의 뉴런이다. 일부 실시예에서, 복수의 은닉 뉴런은 30,000개 이하, 20,000개 이하, 15,000개 이하, 10,000개 이하, 9000개 이하, 8000개 이하, 7000개 이하, 6000개 이하, 5000개 이하, 4000개 이하, 3000개 이하, 2000개 이하, 1000개 이하, 900개 이하, 800개 이하, 700개 이하, 600개 이하, 500개 이하, 400개 이하, 300개 이하, 200개 이하, 100개 이하, 또는 50개 이하의 뉴런이다. 일부 실시예에서, 복수의 은닉 뉴런은 2개 내지 20개, 2개 내지 200개, 2개 내지 1000개, 10개 내지 50개, 10개 내지 200개, 20개 내지 500개, 100개 내지 800개, 50개 내지 1000개, 500개 내지 2000개, 1000개 내지 5000개, 5000개 내지 10,000개, 10,000개 내지 15,000개, 15,000개 내지 20,000개, 또는 20,000개 내지 30,000개의 뉴런이다. 일부 실시예에서, 복수의 은닉 뉴런은 2개 이상의 뉴런에서 시작하여 30,000개 이하의 뉴런에서 끝나는 또 다른 범위 내에 속한다.

[0256] 일부 실시예에서, 신경망은 1개 내지 20개의 은닉 계층을 포함한다. 일부 실시예에서, 신경망은 1개 내지 20개의 은닉 계층을 포함한다. 일부 실시예에서, 신경망은 적어도 2개, 적어도 3개, 적어도 4개, 적어도 5개, 적어도 6개, 적어도 7개, 적어도 8개, 적어도 9개, 적어도 10개, 적어도 11개, 적어도 12개, 적어도 13개, 적어도 14개, 적어도 15개, 적어도 16개, 적어도 17개, 적어도 18개, 적어도 19개, 적어도 20개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 또는 적어도 100개의 은닉 계층을 포함한다. 일부 실시예에서, 신경망은 100개 이하, 90개 이하, 80개 이하, 70개 이하, 60개 이하, 50개 이하, 40개 이하, 30개 이하, 20개 이하, 10개 이하, 9개 이하, 8개 이하, 7개 이하, 6개 이하, 또는 5개 이하의 은닉 계층을 포함한다. 일부 실시예에서, 신경망은 1개 내지 5개, 1개 내지 10개, 1개 내지 20개, 10개 내지 50개, 2개 내지 80개, 5개 내지 100개, 10개 내지 100개, 50개 내지 100개, 또는 3개 내지 30개의 은닉 계층을 포함한다. 일부 실시예에서, 신경망은 1개 계층 이상으로 시작하여 100개 계층 이하로 끝나는 다른 범위 내에 속하는 복수의 은닉 계층을 포함한다.

[0257] 일부 실시예에서, 신경망은 얕은 신경망을 포함한다. 얕은 신경망은 적은 수의 은닉 계층을 갖는 신경망을 지칭한다. 일부 실시예에서, 이러한 신경망 아키텍처는 신경망 훈련의 효율을 개선시키고 훈련에 관여된 계층의 감소된 수로 인해 계산력을 보존한다. 일부 실시예에서, 신경망은 1개의 은닉 계층을 포함한다. 일부 실시예에서, 신경망은 2개, 3개, 4개 또는 5개의 은닉 계층을 포함한다.

[0258] 일부 실시예에서, 신경망은 메시지 전달 신경망이다. 메시지 전달 신경망은 그래프(예를 들어, 화학 구조의 그래프-기반 표현) 상에서의 지도 학습을 위한 프레임워크를 지칭하며, 여기서 노드는 원자를 나타내고, 에지는 원자 사이의 결합을 나타낸다. 일반적으로, 메시지 전달 신경망은 순방향 경로에서의 2개의 단계, 메시지 전달 단계 및 판독 단계를 포함한다. 메시지 전달 단계는 T 간격의 기간 동안 실행되고, 메시지 함수 M_t 및 버텍스 업데이트 함수 U_t에 따라 그래프 내의 각각의 노드에서 은닉 상태를 업데이트하는 것을 포함한다. 판독 단계는 판독 함수 R을 사용하여 그래프에 대한 특징 벡터를 계산한다. 일부 실시예에서, 메시지 전달 신경망은 컨볼루션 네트워크(예를 들어, 공간 그래프 컨볼루션 네트워크 및/또는 스펙트럼 그래프 컨볼루션 네트워크), 게이팅 그래프 신경망(GG-NN), 상호작용 네트워크, 분자 그래프 컨볼루션, 심층 텐서 신경망, 및/또는 라플라시안-기반 방법을 포함한다. 예를 들어, 문헌 [Gilmer 등의 2017, "Neural Message Passing for Quantum Chemistry," arXiv:1704.01212v2]을 참조하며, 이는 그 전문이 본원에 참조로 포함된다.

[0259] 도 3b의 블록 314를 참조하면, 일부 실시예에서, 모델은 복수의 컴포넌트 모델의 앙상블 모델이다. 예를 들어, 블록 316을 참조하면, 일부 실시예에서, 하나 이상의 계산된 활성화 점수에서의 각각의 계산된 활성화 점수는 복수의 컴포넌트 모델의 각각의 컴포넌트 모델의 출력의 중심 집중 경향의 척도이다.

[0260] 도 3b의 블록 318을 참조하면, 일부 실시예에서, 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 선형 회귀 모델, 또는 복수의 신경망을 포함한다.

[0261] 일부 실시예에서, 앙상블 모델은 적어도 2개, 적어도 3개, 적어도 4개, 적어도 5개, 적어도 6개, 적어도 7개, 적어도 8개, 적어도 9개, 적어도 10개, 적어도 20개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 또는 적어도 500개의 컴포넌트 모델을 포함한다. 일부 실시예에서, 앙상블 모델은 500개 이하, 400개 이하, 300개 이하, 200개 이하, 또는 100개 이하의 컴포넌트 모델을 포함한다. 일부 실시예에서, 앙상블 모델은 100개 이하, 50개 이하, 40개 이하, 30개 이하, 또는 20개 이하의 컴포넌트 모델을 포함한다. 일부 실시예에서, 앙상블 모델은 1개 내지 50개, 2개 내지 20개, 5개 내지 50개, 10개 내지 80개, 5개 내지 15개, 3개 내지 30개, 10개 내지 500개, 2개 내지 100개, 또는 50개 내지 100개의 컴포넌트 모델을 포함한다. 일부 실시예에서, 앙상블 모델은 2개 이상의 컴포넌트 모델로 시작하여 500개 이하의 컴포넌트 모델로 끝나는 또 다른 범위의 컴포넌트 모델을 포함한다.

[0262] 일부 실시예에서, 앙상블 모델은 복수의 컴포넌트 모델로부터 획득된 복수의 출력(예를 들어, 활성화 점수)을 조합함으로써 형성된다. 일부 실시예에서, 분류자로부터의 복수의 출력(예를 들어, 활성화 점수)은 평균, 중앙값, 모드, 가중 평균, 가중 중앙값, 가중 모드, 산술 평균, 중간범위, 미드힌지, 삼평균 및/또는 윈저화 평균을 포함하나 이에 제한되지는 않는 관련 기술분야에 공지된 중심 집중 경향의 임의의 척도를 사용하여 조합된다. 예를 들어, 앙상블 모델로부터의 최종 결정은 앙상블 모델에서의 모든 컴포넌트 모델에 걸친 출력의 평균에 기초하여 획득될 수 있다.

[0263] 일부 실시예에서, 복수의 출력은 투표 방법을 사용하여 조합된다. 예를 들어, 일부 실시예에서, 복수의 출력은 각각의 화학 구조와 각각의 관심 생리학적 조건 사이의 연관성을 나타내는, 앙상블 모델의 각각의 컴포넌트 모델로부터의 출력의 수(예를 들어, 활성화 점수)를 기록함으로써 조합된다. 일부 실시예에서, 컴포넌트 모델로부터의 복수의 출력(예를 들어, 활성화 점수)은 과반수 투표를 사용하여 조합된다. 일부 이러한 실시예에서, 컴포넌트 모델로부터의 복수의 출력은 연관성을 나타내는 출력의 기록(예를 들어, 임계치 기준을 초과하는 활성화 점수의 기록)이 투표 임계치보다 큰 경우에 각각의 화학 구조와 각각의 관심 생리학적 조건 사이의 연관성을 결정함으로써 조합된다. 일부 실시예에서, 투표 임계치는 앙상블 모델에서의 복수의 컴포넌트 모델로부터의 총 투표의 적어도 50%이다. 일부 실시예에서, 투표 임계치는 앙상블 모델에서의 복수의 컴포넌트 모델로부터의 총 투표의 적어도 20%, 적어도 30%, 적어도 40%, 적어도 50%, 적어도 60%, 적어도 70%, 적어도 80%, 적어도 90%, 또는 적어도 95%이다.

[0264] 일부 실시예에서, 앙상블 모델의 각각의 컴포넌트 모델은 가중되지 않는다(예를 들어, 각각의 컴포넌트 모델은 앙상블 모델에서 하나의 투표를 갖는다). 일부 실시예에서, 앙상블 모델에서의 하나 이상의 컴포넌트 모델이 추가로 가중된다(예를 들어, 앙상블 모델에서 1을 초과하는 투표를 갖는다).

[0265] 일부 실시예에서, 방법은 단일 앙상블 모델 또는 복수의 앙상블 모델을 획득하는 것을 포함한다. 관련 기술분야에 공지된 임의의 아키텍처가 앙상블 모델에 대해 고려된다. 예를 들어, 일부 실시예에서, 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 및/또는 선형 회귀 모델을 포함한다. 일부 실시예에서, 복수의 컴포넌트 모델은 복수의 신경망을 포함한다.

[0266] 도 3b의 블록 320을 참조하면, 일부 실시예에서, 모델은 복수의 신경망의 앙상블 모델이다. 도 3b의 블록 322를 참조하면, 일부 실시예에서, 모델은 복수의 신경망을 포함하는 앙상블 모델이고, 여기서 복수의 신경망 내의 제1 신경망은 ReLU 활성화를 갖는 완전 연결 신경망이며, 복수의 신경망 내의 제2 신경망은 메시지 전달 신경망이다. 일부 이러한 실시예에서, 제1 신경망은 SMILES 표현으로서 화학 구조에 대한 분자 지문을 입력으로서 수용하는 완전 연결 3-계층 신경망이다. 일부 실시예에서, 제2 신경망은 그래프-기반 표현으로서 화학 구조에 대한 분자 지문을 입력으로서 수용하는 메시지 전달 신경망(MPNN)이다.

[0267] *세포 구성성분 및 세포 구성성분 모듈.*

[0268] 앞서 설명된 바와 같이, 다시 블록 308을 참조하면, 화학 구조의 지문을 모델로 입력하는 것에 응답하여, 모델은 세포 구성성분 모듈의 세트에 대한 하나 이상의 계산된 활성화 점수를 출력한다. 도 3c의 블록 326을 참조하면, 하나 이상의 계산된 활성화 점수 중 각각의 개별 계산된 활성화 점수는 세포 구성성분 모듈의 세트의 상응하는 세포 구성성분 모듈을 나타낸다.

[0269] 도 3c의 블록 328을 참조하면, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함한다.

[0270] 일부 실시예에서, 세포 구성성분은 유전자, 유전자 산물(예를 들어, mRNA 및/또는 단백질), 탄수화물, 지질, 후생적 특징, 대사산물 및/또는 그의 조합이다. 일부 실시예에서, 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합이다.

[0271] 일부 실시예에서, 복수의 세포 구성성분은 DNA, 변형된(예를 들어, 메틸화된) DNA, 코딩(예를 들어, mRNA) 또는 비코딩 RNA(예를 들어, sncRNA)를 포함하는 RNA, 전사후 변형된 단백질(예를 들어, 인산화, 글리코실화, 미리스틸화 등의 단백질)을 포함하는 단백질, 지질, 탄수화물, 시클릭 뉴클레오티드, 예컨대 시클릭 아데노신 모노포스페이트(cAMP) 및 시클릭 구아노신 모노포스페이트(cGMP)를 포함하는 뉴클레오티드(예를 들어, 아데노신 트리포스페이트(ATP), 아데노신 디포스페이트(ADP) 및 아데노신 모노포스페이트(AMP)), 다른 소분자 세포 구성성분, 예컨대 산화 및 환원된 형태의 니코틴아미드 아데닌 디뉴클레오티드(NADP/NADPH), 및 그 임의의 조합을 포함하는 핵산을 포함한다.

[0272] 일부 실시예에서, 복수의 세포 구성성분은 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 25개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 600개, 적어도 700개, 적어도 800개, 적어도 900개, 적어도 1000개, 적어도 2000개, 적어도 3000개, 적어도 4000개, 적어도 5000개, 적어도 6000개, 적어도 7000개, 적어도 8000개, 적어도 9000개, 적어도 10,000개, 적어도 20,000개, 적어도 30,000개, 적어도 50,000

개, 또는 50,000개를 초과하는 세포 구성성분을 포함한다. 일부 실시예에서, 복수의 세포 구성성분은 70,000개 이하, 50,000개 이하, 30,000개 이하, 10,000개 이하, 5000개 이하, 1000개 이하, 500개 이하, 200개 이하, 100개 이하, 90개 이하, 80개 이하, 70개 이하, 60개 이하, 50개 이하, 또는 40개 이하의 세포 구성성분을 포함한다. 일부 실시예에서, 복수의 세포 구성성분은 20개 내지 10,000개의 세포 구성성분으로 이루어진다. 일부 실시예에서, 복수의 세포 구성성분은 100개 내지 8,000개의 세포 구성성분으로 이루어진다. 일부 실시예에서, 복수의 세포 구성성분은 5개 내지 20개, 20개 내지 50개, 50개 내지 100개, 100개 내지 200개, 200개 내지 500개, 500개 내지 1000개, 1000개 내지 5000개, 5000개 내지 10,000개, 또는 10,000개 내지 50,000개의 세포 구성성분을 포함한다. 일부 실시예에서, 복수의 세포 구성성분은 5개 이상의 세포 구성성분에서 시작하여 70,000개 이하의 세포 구성성분에서 끝나는 또 다른 범위 내에 속한다.

[0273] 예로서, 일부 실시예에서, 복수의 세포 구성성분은 선택적으로 RNA 수준에서 측정된 복수의 유전자를 포함한다. 일부 실시예에서, 복수의 유전자는 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 25개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 600개, 적어도 700개, 적어도 800개, 적어도 900개 또는 적어도 1000개의 유전자를 포함한다. 일부 실시예에서, 복수의 유전자는 적어도 1000개, 적어도 2000개, 적어도 3000개, 적어도 4000개, 적어도 5000개, 적어도 10,000개, 적어도 30,000개, 적어도 50,000개, 또는 50,000개를 초과하는 유전자를 포함한다. 일부 실시예에서, 복수의 유전자는 5개 내지 20개, 20개 내지 50개, 50개 내지 100개, 100개 내지 200개, 200개 내지 500개, 500개 내지 1000개, 1000개 내지 5000개, 5000개 내지 10,000개, 또는 10,000개 내지 50,000개의 유전자를 포함한다.

[0274] 또 다른 예로서, 일부 실시예에서, 복수의 세포 구성성분은 복수의 단백질을 포함한다. 일부 실시예에서, 복수의 단백질은 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 25개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 600개, 적어도 700개, 적어도 800개, 적어도 900개 또는 적어도 1000개의 단백질을 포함한다. 일부 실시예에서, 복수의 단백질은 적어도 1000개, 적어도 2000개, 적어도 3000개, 적어도 4000개, 적어도 5000개, 적어도 10,000개, 적어도 30,000개, 적어도 50,000개, 또는 50,000개를 초과하는 단백질을 포함한다. 일부 실시예에서, 복수의 단백질은 5개 내지 20개, 20개 내지 50개, 50개 내지 100개, 100개 내지 200개, 200개 내지 500개, 500개 내지 1000개, 1000개 내지 5000개, 5000개 내지 10,000개, 또는 10,000개 내지 50,000개의 단백질을 포함한다.

[0275] 세포 구성성분 모듈의 각각의 세포 구성성분이 고유할 필요는 없다. 예를 들어, 세포 구성성분 모듈 A가 세포 구성성분 1, 3 및 10을 함유하는 경우를 고려한다. 세포 구성성분 모듈의 세트의 다른 세포 구성성분 모듈 또한 이러한 세포 구성성분을 함유할 수 있다. 여기서, 용어 "독립적"은 특정한 세포 구성성분 모듈에서 복수의 세포 구성성분의 서브세트가 전체로서 고유함을 의미한다. 따라서, 상기 예시적인 세포 구성성분 모듈 A를 고려하면, 세포 구성성분 모듈의 세트의 또 다른 세포 구성성분 모듈은 세포 구성성분 1, 3 및 10을 함유할 수 있고, 단 세포 구성성분 모듈 A가 함유하지 않는 다른 세포 구성성분을 추가로 함유한다. 상기 예시적인 세포 구성성분 모듈 A를 추가로 고려하여, 세포 구성성분 모듈의 세트의 또 다른 세포 구성성분 모듈은 세포 구성성분 1, 3 및 10의 서브세트로 제한될 수 있고, 단 세포 구성성분 모듈 A가 함유하지 않는 다른 세포 구성성분을 추가로 함유할 필요는 없다(그러나, 이는 이러한 추가적인 세포 구성성분을 또한 가질 수 있다).

[0276] 일부 실시예에서, 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈은 복수의 세포 구성성분의 각각의 독립적인 서브세트의 동일하거나 상이한 수의 세포 구성성분을 포함한다. 일부 실시예에서, 각각의 개별 세포 구성성분 모듈에 상응하는 세포 구성성분의 각각의 개별 독립적인 서브세트는 세포 구성성분의 고유한 서브세트이다(예를 들어, 비-중첩이고, 여기서 복수의 세포 구성성분의 각각의 세포 구성성분은 하나 이하의 모듈로 그룹화된다). 일부 실시예에서, 제1 세포 구성성분 모듈은 제2 세포 구성성분 모듈에 상응하는 세포 구성성분의 제2 서브세트와 중첩되는 세포 구성성분의 제1 서브세트를 갖는다(예를 들어, 복수의 세포 구성성분 중 적어도 하나의 세포 구성성분이 2개 이상의 상이한 모듈에 공통적인 경우에 중첩됨).

[0277] 도 3c의 블록 330을 참조하면, 일부 실시예에서, 각각의 세포 구성성분 모듈 내의 복수의 세포 구성성분의 독립적인 서브세트는 5개 이상의 세포 구성성분을 포함한다. 일부 실시예에서, 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈 내의 복수의 세포 구성성분의 독립적인 서브세트는 적어도 2개, 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 25개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 600개, 적어도 700개, 적어도 800개, 적어도 900개, 적어도 1000개, 적어도 2000개, 또는 적어도 3000개의

세포 구성성분을 포함한다. 일부 실시예에서, 복수의 세포 구성성분의 독립적인 서브세트는 5000개 이하, 3000개 이하, 1000개 이하, 500개 이하, 200개 이하, 100개 이하, 90개 이하, 80개 이하, 70개 이하, 60개 이하, 또는 50개 이하의 세포 구성성분을 포함한다. 일부 실시예에서, 복수의 세포 구성성분의 독립적인 서브세트는 5개 내지 100개, 2개 내지 300개, 20개 내지 500개, 200개 내지 1000개, 또는 1000개 내지 5000개의 세포 구성성분을 포함한다. 일부 실시예에서, 복수의 세포 구성성분의 독립적인 서브세트는 2개 이상의 세포 구성성분에서 시작하여 5000개 이하의 세포 구성성분에서 끝나는 또 다른 범위 내에 속한다.

[0278] 일부 실시예에서, 각각의 세포 구성성분 모듈 내의 복수의 세포 구성성분의 독립적인 서브세트는 관심 생리학적 조건과 연관된 세포 과정(예를 들어, 분자 경로) 내의 세포 구성성분으로 이루어진다. 예를 들어, 도 3c의 블록 332를 참조하면, 일부 실시예에서, 각각의 세포 구성성분 모듈 내의 복수의 세포 구성성분의 독립적인 서브세트는 관심 생리학적 조건과 연관된 분자 경로 내의 2개 내지 20개의 세포 구성성분으로 이루어진다.

[0279] 도 3d의 블록 334를 참조하면, 세포 구성성분 모듈의 세트의 적어도 제1 세포 구성성분 모듈이 관심 생리학적 조건과 연관된다. 실제로, 수많은 세포 구성성분 모듈이 관심 생리학적 조건과 연관될 수 있다.

[0280] 도 3d의 블록 336을 참조하면, 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 세포 구성성분 모듈의 세트의 상응하는 세포 구성성분 모듈을 나타낸다.

[0281] 도 3d의 블록 338을 참조하면, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함한다.

[0282] 도 3d의 블록 340을 참조하면, 일부 실시예에서, 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈이다. 제1 세포 구성성분 모듈을 포함하는 복수의 세포 구성성분 모듈의 제1 서브세트는 관심 생리학적 조건과 연관된다. 즉, 이러한 세포 구성성분 모듈은 관심 생리학적 조건에 관여하는 세포 구성성분을 나타낸다. 예를 들어, 이러한 세포 구성성분 모듈의 이러한 세포 구성성분은 일부 기준선, 야생형 상태의 세포에 비해 관심 생리학적 조건을 나타내는 세포에서 하향조절 또는 상향조절될 수 있다. 또한, 복수의 세포 구성성분 모듈의 제2 서브세트는 관심 생리학적 조건과 연관되지 않는다. 즉, 이러한 세포 구성성분 모듈의 세포 구성성분은 관심 생리학적 조건에 수반되지 않는 세포 구성성분을 나타낸다. 예를 들어, 이러한 세포 구성성분은 일부 기준선, 야생형 상태의 세포에 비해 관심 생리학적 조건을 나타내는 세포에서 하향조절 또는 상향조절되지 않는다. 이러한 실시예에서, 화학적 화합물은 제1 세포 구성성분 모듈(세포 구성성분 모듈의 제1 서브세트에 있음)에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족하고, 복수의 세포 구성성분 모듈의 제2 서브세트에서 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 제2 임계치 기준을 충족하는 경우에 관심 생리학적 조건으로 식별된다. 예시적인 제1 임계치 기준은 도 3e의 블록 348과 관련하여 하기에 논의된다. 일반적으로, (제1 임계치를 충족하는 계산된 활성화 점수를 가짐으로써 입증된 바에 따라) 세포 구성성분 모듈의 제1 서브세트에서 세포 구성성분 모듈로 식별하지만, (제2 임계치를 충족하는 계산된 활성화 점수를 가짐으로써 입증된 바에 따라) 세포 구성성분 모듈의 제2 서브세트에서 세포 구성성분 모듈로 식별하지 않는 화학적 화합물이 추구된다. 예를 들어, 일부 실시예에서, 제1 임계치의 만족은 제1 미리 결정된 수치값을 초과하는 활성화 점수를 필요로 하는 반면, 제2 임계치의 만족은 제2 미리 결정된 수치값 미만의 활성화 점수를 필요로 하며, 여기서 정확한 제1 및 제2 미리 결정된 수치값은 응용에 따라 달라진다.

[0283] 나타낸 바와 같이, 일부 구현에서, 방법은 관심 생리학적 조건(예를 들어, 세포 과정)을 특징화하기 위해 하나 이상의 유형의 분자 데이터(예를 들어, 세포 구성성분)를 사용하는 것을 포함한다. 이러한 분자 데이터는 측정 가능한 속성(예를 들어, 풍부도 및/또는 발현 수준), 예컨대 체학 프로파일링(예를 들어, 전사체학, 단백질체학, 대사체학 등)을 갖는 임의의 분석물을 포함할 수 있다.

[0284] 일반적으로, 세포 과정과 연관되는 경우, 세포 구성성분(예를 들어, 유전자)의 세포 구성성분 모듈은 스위칭 이벤트의 시퀀스로부터 발생하는 것으로 생각될 수 있고, 이때 유사한 시간에 스위칭되는 세포 구성성분(예를 들어, 유전자)은 함께 모듈을 형성한다. 따라서, 예를 들어, 일부 실시예에서, 각각의 세포 구성성분 모듈은 복수의 세포 구성성분의 각각의 서브세트를 포함하고, 이때 세포 구성성분의 서브세트는 각각의 관심 생리학적 조건(예를 들어, 관심 세포 과정)과 연관된 거동의 유사성을 기초로 그룹화된다. 한 예에서, 각각의 관심 생리학적 조건과 연관된 세포 구성성분 모듈은 각각의 생리학적 조건을 갖는 복수의 세포 유형에 걸쳐 유사하게 거동하는(예를 들어, 유사한 발현 프로파일을 나타내는) 유전자의 서브세트를 포함할 수 있다.

[0285] 도 3d의 블록 342를 참조하면, 일부 실시예에서, 세포 구성성분 모듈의 세트는 제1 세포 구성성분 모듈로 이루어진다.

[0286] 도 3d의 블록 344를 참조하면, 일부 실시예에서, 세포 구성성분 모듈의 세트는 5개 이상의 세포 구성성분 모듈을 포함한다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 20개 이상, 30개 이상, 40개 이상, 50개 이상, 60개 이상, 70개 이상, 80개 이상, 90개 이상, 또는 100개 이상의 세포 구성성분 모듈을 포함한다.

[0287] 일부 실시예에서, 세포 구성성분 모듈의 세트는 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 25개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 600개, 적어도 700개, 적어도 800개, 적어도 900개, 적어도 1000개, 적어도 2000개, 적어도 3000개, 적어도 4000개 또는 적어도 5000개의 세포 구성성분 모듈을 포함한다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 10,000개 이하, 5000개 이하, 2000개 이하, 1000개 이하, 500개 이하, 300개 이하, 200개 이하, 100개 이하, 90개 이하, 80개 이하, 70개 이하, 60개 이하, 또는 50개 이하의 세포 구성성분 모듈을 포함한다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 10개 내지 2000개의 세포 구성성분 모듈로 이루어진다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 50개 내지 500개의 세포 구성성분 모듈로 이루어진다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 5개 내지 20개, 20개 내지 50개, 50개 내지 100개, 100개 내지 200개, 200개 내지 500개, 500개 내지 1000개, 1000개 내지 5000개, 또는 5000개 내지 10,000개의 세포 구성성분 모듈을 포함한다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 5개 이상의 세포 구성성분 모듈에서 시작하여 10,000개 이하의 세포 구성성분 모듈에서 끝나는 또 다른 범위 내에 속한다.

[0288] 일부 실시예에서, 방법은 관심 생리학적 조건과 연관된 세포 구성성분 모듈을 식별하는 것을 더 포함한다. 이러한 방법은 하기 도 14a-14d와 함께 명칭 *세포 구성성분 모듈의 식별* 섹션에서 논의된다.

[0289] *활성화 점수*.

[0290] 도 3b의 블록 308에 설명된 바와 같이, 모델은 모델로의 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력한다. 일반적으로, 훈련된 모델(블록 308의 모델)의 출력은 표지(예를 들어, 수치 활성화 점수)를 포함하는 훈련 데이터세트에 대해 학습하고 훈련된 모델의 출력이 최소 성능 수준을 만족할 때까지, 예컨대 검증 단계를 통해 복수의 파라미터를 조정하는 과정을 통해 정의된다. 훈련 모델은 "모델 훈련"이라는 명칭의 섹션에서 하기에 추가로 개시된다.

[0291] 일부 실시예에서, 하나 이상의 계산된 활성화 점수의 활성화 점수는 각각의 화합물에 상응하는 각각의 세포 구성성분 모듈에 대한 각각의 활성화 가중치이다. 예를 들어, 일부 실시예에서, 활성화 점수는 도 2a, 도 2b 및 도 14a 내지 도 14d를 참조로 "세포 구성성분 모듈의 식별"이라는 명칭의 하기 섹션에서 설명되고 도 5의 활성화 데이터 구조에 예시된 바와 같이 획득된 활성화 가중치이고, 이때 활성화 점수는 각각의 화합물로의 치료와 상관되고/되거나 이에 응답한 각각의(예를 들어, 제1) 세포 구성성분 모듈의 활성화(예를 들어, 유도 및/또는 차등 발현)를 표시한다.

[0292] 따라서, 일부 이러한 실시예에서, 훈련된 모델은 출력으로서, 테스트 화학적 화합물과 관심 생리학적 조건의 연관(예를 들어, 관심 생리학적 조건과 연관된 제1 세포 구성성분 모듈)을 나타내는 계산된 활성화 점수를 제공한다. 이어서, 도 3e의 블록 348을 참조하면, 방법은 제1 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족할 때 관심 생리학적 조건을 갖는 화학적 화합물을 식별하는 것(예를 들어, 그에 대한 연관을 결정하는 것)을 포함한다.

[0293] 도 3e의 블록 350을 참조하면, 일부 실시예에서, 제1 임계치 기준은 제1 세포 구성성분 모듈이 임계 활성화 점수를 가져야 한다는 요건이다. 일반적으로, (제1 임계치를 충족하는 계산된 활성화 점수를 가짐으로써 입증된 바와 같이) 관심 생리학적 조건을 갖는 것을 식별하는 화학적 화합물이 추구된다. 예를 들어, 일부 실시예에서, 제1 임계치의 충족은 제1의 미리 결정된 수치값을 초과하는 활성화 점수를 필요로 한다.

[0294] 예를 들어, 일부 실시예에서, 활성화 점수는 "0" 내지 "1"(또는 일부 다른 범위 "A" 내지 "B", 여기서 A 및 B는 2개의 상이한 수임)의 정규화된 연속 값으로서 표현되고, 여기서 "1"에 더 가까운 값(예를 들어, .89, .90, .91, .92 등)은 세포 구성성분 모듈(및 세포 구성성분 모듈이 나타내는 화학적 화합물)과 관심 생리학적 조건 사이의 강한 연관성을 나타낸다. "0"에 더 가까운 값(예를 들어, 0.01, 0.02, 0.03, 0.04 등)은 세포 구성성분 모듈(및 세포 구성성분이 나타내는 화학적 화합물)과 관심 생리학적 조건 사이에 연관성이 없음을 나타낸다. 이러한 경우에, 제1 임계치는 "0" 내지 "1"(또는 일부 다른 범위 "A" 내지 "B", 여기서 A 및 B는 2개의 상이한 수임)에서 선택되고, 세포 구성성분 모듈(및 이것이 나타내는 화학 구조)은 활성화 점수가 제1 임계치를 초과하는 경우에 관심 생리학적 조건과 연관된 것으로 간주되는 반면, 세포 구성성분 모듈(및 이것이 나타내는 화학

구조)은 활성화 점수가 제1 임계치 미만인 경우에 관심 생리학적 조건과 연관되지 않은 것으로 간주된다. 일부 이러한 실시예에서, 활성화 점수는 "0" 내지 "1"(또는 A 및 B가 2개의 상이한 수인 일부 다른 범위 "A" 내지 "B")의 연속 척도에서의 정규화된 값으로서 표현되고, 제1 임계치는 0 내지 1, 0.10 내지 0.90, 0.20 내지 0.80, 0.30 내지 0.70, 0.50 내지 0.99, 0.60 내지 0.99, 0.70 내지 0.99, 0.80 내지 0.99, 또는 0.90 내지 0.99의 값이다.

[0295] 또 다른 예로서, 일부 실시예에서, 활성화 점수는 "0" 내지 "1"(또는 일부 다른 범위 "A" 내지 "B", 여기서 A 및 B는 2개의 상이한 수임)의 연속 척도 상의 정규화된 값으로서 표현되고, 여기서 "1"에 더 가까운 값(예를 들어, .89, .90, .91, .92 등)은 세포 구성성분 모듈(및 세포 구성성분 모듈이 나타내는 화학적 화합물)과 관심 생리학적 조건 사이에 연관성이 없음을 나타낸다. "0"에 더 가까운 값(예를 들어, 0.01, 0.02, 0.03, 0.04 등)은 세포 구성성분 모듈(및 세포 구성성분이 나타내는 화학적 화합물)과 관심 생리학적 조건 사이의 연관성을 나타낸다. 이러한 경우에, 제1 임계치는 "0" 내지 "1"(또는 일부 다른 범위 "A" 내지 "B", 여기서 A 및 B는 2개의 상이한 수임)로 선택되고, 세포 구성성분 모듈(및 이것이 나타내는 화학 구조)은 활성화 점수가 제1 임계치 미만인 경우에 관심 생리학적 조건과 연관되는 것으로 간주되는 반면, 세포 구성성분 모듈(및 이것이 나타내는 화학 구조)은 활성화 점수가 제1 임계치를 초과하는 경우에 관심 생리학적 조건과 연관되지 않는 것으로 간주된다. 일부 이러한 실시예에서, 활성화 점수는 "0" 내지 "1"(또는 A 및 B가 2개의 상이한 수인 일부 다른 범위 "A" 내지 "B")의 연속 척도 상의 정규화된 값으로서 표현되고, 제1 임계치는 0 내지 1, 0.10 내지 0.90, 0.20 내지 0.80, 0.30 내지 0.70, 0.50 내지 0.99, 0.60 내지 0.99, 0.70 내지 0.99, 0.80 내지 0.99, 또는 0.90 내지 0.99의 값이다.

[0296] 도 3e의 블록 352를 참조하면, 일부 실시예에서, 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈(예를 들어, 2개 내지 1000개, 10개 내지 100개, 2개 내지 100개, 4개 내지 50개의 세포 구성성분 모듈)이고, 블록 348의 식별은 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족시킬 것을 요구한다. 예를 들어, 세포 구성성분 모듈의 세트가 2개의 세포 구성성분 모듈 A 및 B로 이루어진 경우를 고려한다. 도 3e의 블록 352는 세포 구성성분 모듈 A 및 B의 활성화 점수가 각각 제1 임계치 조건을 충족시킬 것을 요구한다. 예를 들어, 세포 구성성분 모듈 A가 0.25의 계산된 활성화 점수를 갖고, 세포 구성성분 모듈 B가 0.75의 계산된 활성화 점수를 가지며, 제1 임계치 조건의 만족은 각각의 활성화 점수가 0.4를 초과할 것을 요구하는 경우를 고려한다. 이러한 예에서, 세포 구성성분 모듈의 세트는 도 3e의 블록 352의 요건을 충족시키지 않는데, 이는 각각의 활성화 점수가 0.4 임계치 요건 이하이기 때문이다.

[0297] 도 3e의 블록 354를 참조하면, 일부 실시예에서, 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈(예를 들어, 2개 내지 1000개, 10개 내지 100개, 2개 내지 100개, 4개 내지 50개의 세포 구성성분 모듈)이고, 블록 348의 식별은 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 계산된 활성화 점수에 걸친 중심 집중 경향의 척도가 제1 임계치 기준을 충족시킬 것을 요구한다. 예를 들어, 세포 구성성분 모듈의 세트가 2개의 세포 구성성분 모듈 A 및 B로 이루어진 경우를 고려한다. 도 3e의 블록 354는 세포 구성성분 모듈 A 및 B의 활성화 점수의 소정의 중심 집중 경향의 척도가 제1 임계치 조건을 충족시킬 것을 요구한다. 예를 들어, 중심 집중 경향의 척도가 평균화인 경우를 고려하면, 세포 구성성분 모듈 A는 0.25의 계산된 활성화 점수를 갖고, 세포 구성성분 모듈 B는 0.75의 계산된 활성화 점수를 갖고, 제1 임계치 조건의 충족은 평균 활성화 점수가 0.4를 초과할 것을 요구한다. 이러한 경우에, 세포 구성성분 모듈의 세트는 도 3e의 블록 354의 요건을 충족시키는데, 이는 이들이 0.4 임계치 요건보다 큰 0.25 + 0.75 / 2 또는 0.5의 평균 활성화 점수를 갖기 때문이다. 일부 실시예에서, 중심 집중 경향의 척도는 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 개별 계산된 활성화 점수의 산술 평균, 가중 평균, 중간범위, 미드힌지, 삼평균, 윈저화 평균, 평균, 또는 모드이다.

[0298] *화합물.*

[0299] 일부 실시예에서, 테스트 화학적 화합물은 소분자, 생물제제, 단백질, 소분자와 조합된 단백질, ADC, 핵산, 예컨대 siRNA 또는 간섭 RNA, 야생형 및/또는 돌연변이체 shRNA를 과다발현하는 cDNA, 야생형 및/또는 돌연변이체 가이드 RNA를 과다발현하는 cDNA(예를 들어, Cas9 시스템 또는 다른 세포-컴포넌트 편집 시스템), 및/또는 임의의 전술한 것의 임의의 조합이다.

[0300] 일부 실시예에서, 테스트 화학적 화합물은 무기 또는 유기이다.

[0301] 예를 들어, 도 3e의 블록 356과 관련하여, 일부 실시예에서, 테스트 화학적 화합물은 2000 달톤(Da) 미만의 분자량을 갖는 유기 화합물이다. 일부 실시예에서, 테스트 화학적 화합물은 적어도 10 Da, 적어도 20 Da, 적어도

50 Da, 적어도 100 Da, 적어도 200 Da, 적어도 500 Da, 적어도 1 kDa, 적어도 2 kDa, 적어도 3 kDa, 적어도 5 kDa, 적어도 10 kDa, 적어도 20 kDa, 적어도 30 kDa, 적어도 50 kDa, 적어도 100 kDa 또는 적어도 500 kDa의 분자량을 갖는다. 일부 실시예에서, 테스트 화학적 화합물은 1000 kDa 이하, 500 kDa 이하, 100 kDa 이하, 50 kDa 이하, 10 kDa 이하, 5 kDa 이하, 2 kDa 이하, 1 kDa 이하, 500 Da 이하, 300 Da 이하, 100 Da 이하, 또는 50 Da 이하의 분자량을 갖는다. 일부 실시예에서, 테스트 화학적 화합물은 10 Da 내지 900 Da, 50 Da 내지 1000 Da, 100 Da 내지 2000 Da, 1 kDa 내지 10 kDa, 5 kDa 내지 500 kDa, 또는 100 kDa 내지 1000 kDa의 분자량을 갖는다. 일부 실시예에서, 테스트 화학적 화합물은 10 달톤 이상에서 시작하여 1000 kDa 이하에서 끝나는 또 다른 범위 내에 속하는 분자량을 갖는다.

[0302] 도 3e의 블록 358을 참조하면, 일부 실시예에서, 테스트 화학적 화합물은 리핀스키 5 준칙 각각을 충족시키는 유기 화합물이다. 리핀스키 5 준칙(예를 들어, RO5)은 약물유사성을 평가하는데, 예컨대 각각의 약리학적 또는 생물학적 활성을 갖는 각각의 화합물이 인간에 대한 투여에 적합한 상응하는 화학적 또는 물리적 특성을 갖는지 여부를 결정하는데 사용되는 일련의 가이드라인이다. 리핀스키 5 준칙은 화합물의 약물유사성을 결정하기 위한 다음 기준을 포함한다: (i) 500 Da 미만의 분자 질량, (ii) 5개 이하의 수소 결합 공여자, (iii) 10개 이하의 수소 결합 수용자, 및 (iv) 5 이하의 옥탄올-물 분배 계수 log *P*.

[0303] 도 3e의 블록 360과 관련하여, 일부 실시예에서, 테스트 화학적 화합물은 리핀스키 5 준칙의 적어도 2, 3 또는 4가지 기준을 충족시키는 유기 화합물이다. 일부 실시예에서, 테스트 화학적 화합물은 리핀스키 5 준칙의 0, 1, 2, 3가지 또는 모든 4가지 기준을 충족시키는 유기 화합물이다.

[0304] 일부 실시예에서, 테스트 화학적 화합물은 데이터베이스로부터 선택된다. 약물 스크린, 주석 및/또는 일반적 정보, 예컨대 화합물 표적 및 화합물의 화학적 특성으로부터의 결과를 제공하는 적합한 화합물 데이터베이스의 예는 암에서의 약물 감수성의 유전체학, 암 치료 반응 포털, 연결성 맵, PharmacoDB, BoBER(Base of Bioisosterically Exchangeable Replacements) 및/또는 DrugBank를 포함하나, 이에 제한되지는 않는다. 일부 실시예에서, 테스트 화학적 화합물은 유전자 및 유전자 산물, 교란-유도된 세포 구성성분 시그니처 및/또는 경로 주석에 대한 정보를 제공하는 데이터베이스로부터 선택된다. 적합한 데이터베이스의 예는 NIH 유전자 발현 옴니버스(GEO), EBI ArrayExpress, NCBI, BLAST, EMBL-EBI, GenBank, Ensembl, KEGG 경로 데이터베이스, LINCS(Library of Integrated Network-based Cellular Signatures) L1000 데이터세트, 리액톰 경로 데이터베이스, 및/또는 유전자 온톨로지 프로젝트를 포함하나, 이에 제한되지는 않는다.

[0305] *실제 응용에서의 방법(300)의 결과의 사용.*

[0306] 일부 실시예에서, 도 3과 관련하여 앞서 설명된 방법(300)을 사용하여 관심 생리학적 조건에 대해 복수의 테스트 화합물을 평가한다. 이러한 실시예에서, 복수의 테스트 화합물의 각각의 테스트 화합물은 도 3의 방법(300)을 통해 실행된다. 따라서, 100개의 테스트 화합물 및 1개의 관심 생리학적 조건이 존재하는 경우에, 이러한 실시예에서, 방법(300)은 100회 실행되며, 여기서 100회 중 각각의 인스턴스는 테스트 화합물 중 상이한 것들에 대한 것이다.

[0307] 또한, 일부 실시예에서, 도 3과 관련하여 앞서 설명된 방법(300)을 사용하여 복수의 관심 생리학적 조건에 대해 복수의 화합물을 평가한다. 이러한 실시예에서, 각각의 관심 생리학적 조건에 대해, 복수의 테스트 화합물의 각각의 개별 테스트 화합물은 도 3의 방법(300)을 통해 실행된다. 따라서, 100개의 테스트 화합물 및 2개의 관심 생리학적 조건이 존재하는 경우에, 이러한 실시예에서, 방법(300)은 200회 실행되며, 200회의 각각의 인스턴스는 제1 또는 제2 관심 생리학적 조건 중 어느 하나에 대한 테스트 화합물 중 상이한 것들에 대한 것이다.

[0308] 일부 실시예에서, 복수의 테스트 화합물은 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 800개, 적어도 1000개, 적어도 2000개, 적어도 3000개, 적어도 4000개, 적어도 5000개, 적어도 8000개, 적어도 10,000개, 적어도 20,000개, 적어도 30,000개, 적어도 50,000개, 적어도 80,000개, 적어도 100,000개, 적어도 200,000개, 적어도 500,000개, 적어도 800,000개, 적어도 1백만개 또는 적어도 2백만개의 테스트 화합물을 포함하고, 관심 단일 생리학적 조건이 존재한다. 일부 이러한 실시예에서, 방법(300)은 각각의 테스트 화합물에 대해 하나씩 적어도 5회, 적어도 10회, 적어도 15회, 적어도 20회, 적어도 30회, 적어도 40회, 적어도 50회, 적어도 100회, 적어도 200회, 적어도 300회, 적어도 400회, 적어도 500회, 적어도 800회, 적어도 1000회, 적어도 2000회, 적어도 3000회, 적어도 4000회, 적어도 5000회, 적어도 8000회, 적어도 10,000회, 적어도 20,000회, 적어도 30,000회, 적어도 50,000회, 적어도 80,000회, 적어도 100,000회, 적어도 200,000회, 적어도 500,000회, 적어도 800,000회, 적어도 1백만회, 또는 적어도 2백만회 실행되어, 적어도 5회, 적어도 10회, 적어

도 15회, 적어도 20회, 적어도 30회, 적어도 40회, 적어도 50회, 적어도 100회, 적어도 200회, 적어도 300회, 적어도 400회, 적어도 500회, 적어도 800회, 적어도 1000회, 적어도 2000회, 적어도 3000회, 적어도 4000회, 적어도 5000회, 적어도 8000회, 적어도 10,000회, 적어도 20,000회, 적어도 30,000회, 적어도 50,000회, 적어도 80,000회, 적어도 100,000회, 적어도 200,000회, 적어도 500,000회, 적어도 800,000회, 적어도 1백만회, 또는 적어도 2백만회의 활성화 점수를 실현한다.

[0309] 일부 실시예에서, 복수의 화합물은 1천만개 이하, 5백만개 이하, 1백만개 이하, 500,000개 이하, 100,000개 이하, 50,000개 이하, 10,000개 이하, 8000개 이하, 5000개 이하, 2000개 이하, 1000개 이하, 800개 이하, 500개 이하, 200개 이하, 또는 100개 이하의 테스트 화합물을 포함한다. 일부 실시예에서, 복수의 화합물은 10개 내지 500개, 100개 내지 10,000개, 5000개 내지 200,000개, 또는 10,000개 내지 1백만개의 테스트 화합물로 이루어진다.

[0310] 일부 실시예에서, 복수의 테스트 화합물은 10개 내지 1×10^6개의 테스트 화합물이다. 일부 실시예에서, 복수의 테스트 화합물은 100개 내지 100,000개의 테스트 화합물이다. 일부 실시예에서, 복수의 테스트 화합물은 1000개 내지 100,000개의 테스트 화합물이다.

[0311] 따라서, 방법(300)을 사용하여 다수의 테스트 화합물에 대한 활성화 점수를 획득할 수 있다. 이들 활성화 점수에 대한 제1 임계치의 적용은 관심 생리학적 조건과 연관된 테스트된 많은 테스트 화합물 중에서 테스트 화합물을 식별하는데 사용될 수 있다. 전형적인 실시예에서, 테스트 화합물의 선택된 수는 이들이 관심 생리학적 조건과 연관되지만 반면에 다른 것은 연관되지 않음을 나타내는 활성화 점수를 갖는다. 선택된 수의 테스트 화합물의 분석은 관심 생리학적 조건과 연관을 생성하는 테스트 화합물에 대한 분자 특성을 결정하는데 사용될 수 있다. 예를 들어, 관심 생리학적 조건과 연관됨을 나타내는 활성화 점수를 갖는 선택된 수의 테스트 화합물의 화학 구조를, 관심 생리학적 조건과 연관되지 않는 테스트 화합물과 구별되는 그의 구조의 유사성에 대해 시각적으로 검사할 수 있다. 이러한 분자 특성은 이어서 모델(601)에 의해 평가된 원래 테스트 분자에 포함되지 않았고 모델(601)을 훈련시키는데 사용되지 않았던 새로운 테스트 분자에 포함될 수 있다.

[0312] 또한, 보다 형식적인 접근법이 테스트 화합물(방법(300)에 의해 부과된 제1 임계치를 충족하는 것과 충족하지 않는 것 둘 다)을 분석하는데 사용될 수 있다. 예를 들어, 하위구조 마이닝이 이러한 화합물이 관심 생리학적 조건과 연관되게 하는 테스트 화합물 내의 하위구조를 식별하는데 사용될 수 있다. 하위구조 마이닝의 예는 MOSS(문헌 [Borgetl and Meinl, 2006, "Full Perfect Extension Pruning for Frequent Graph Mining," Proc. Workshop on Mining Complex Data (MCD 2006 at ICDM 2006, Hong Kong, China, IEEE Press, Piscataway, NJ, USA](본원에 참조로 포함됨)), 및 MOFA(문헌 [Meinl and Worlein, 2006 "Mining Molecular Datasets on Symmetric Processor Systems," International conference on Systems, man and Cybernetics 2, pp. 1269-1274](본원에 참조로 포함됨))를 포함하나, 이에 제한되지는 않는다.

[0313] 또한, 최대 공통 하위구조(MCS) 분석을 사용하여, 이러한 화합물이 관심 생리학적 조건과 연관되게 하는 테스트 화합물 내의 하위구조를 식별할 수 있다. MCS 분석의 예는 LIBMCS(Chemaxon, Library MCS, 2008), MCSS(OEChem TK version 2.0.0, OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com), 및 CncMCS(http://www.chemnavigator.com/cnc/products/downloads.asp)를 포함하나, 이에 제한되지는 않는다.

[0314] 또한, SMARTS는 이러한 화합물이 관심 생리학적 조건과 연관되게 하는 테스트 화합물 내의 하위구조를 식별하는데 사용될 수 있다. SMART 분석의 예는 CDK 디스크립터 GUI이다.

[0315] 또한, 고빈도 서브그래프 마이닝을 사용하여 테스트 화합물 내의 하위구조를 식별할 수 있으며, 이는 이러한 화합물이 관심 생리학적 조건과 연관되게 한다. 고빈도 서브그래프 마이닝의 예는 ParMol(Uni Erlangen)이다.

[0316] 또한, 그래프 및 화학적 마이닝을 사용하여 테스트 화합물 내의 하위구조를 식별할 수 있으며, 이는 이러한 화합물이 관심 생리학적 조건과 연관되게 한다. 그래프 및 화학적 마이닝의 예는 PAFI/AFGen(Karypis Lab UMN)이다.

[0317] *교란 시그니처.*

[0318] 앞서 설명된 바와 같이, 일부 실시예에서, 관심 생리학적 조건은 교란 시그니처(예를 들어, 교란에 반응한 제1 세포 상태와 제2 세포 상태 사이의 불일치를 특징으로 함)이다. 따라서, 본 개시의 또 다른 양태는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법(700)을 제공한다. 일부 실시예에서, 관심 생리학적 조건은 질환이다.

[0319] 블록 702를 참조하면, 방법은 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 것을 포함한다. "생리학적 조건" 및 "화합물"이라는 명칭의 상기 섹션에 개시된 바와 같은 생리학적 조건, 화합물, 지문, 및/또는 지문을 획득하는 방법의 임의의 적합한 실시예가 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 그의 임의의 치환, 변형, 추가, 결실, 및/또는 조합을 포함하여 고려된다.

[0320] 예를 들어, 일부 실시예에서, 테스트 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물이다. 일부 실시예에서, 테스트 화학적 화합물은 리핀스키 5 준칙 각각을 충족시키는 유기 화합물이다. 일부 실시예에서, 테스트 화학적 화합물은 리핀스키 5 준칙의 적어도 3가지 기준을 충족시키는 유기 화합물이다. 일부 실시예에서, 방법은 테스트 화학적 화합물의 단순화된 분자-입력 라인-엔트리 시스템(SMILES) 스트링 표현으로부터 지문을 계산하는 것을 더 포함한다. 일부 실시예에서, 지문은 SMILES 변환기, ECFP4, RNNS2S 또는 GraphConv를 사용하여 화학 구조로부터 생성된다.

[0321] 블록 704를 참조하면, 이 방법은 지문을 모델에 입력하는 단계를 더 포함하고, 여기서 모델은 100개 이상의 파라미터를 포함하고, 모델은 지문을 모델에 입력하는 것에 응답하여 하나 이상의 계산된 활성화 점수를 출력하고, 하나 이상의 계산된 활성화 점수 중 각각의 개별 계산된 활성화 점수는 한 세트의 교란 시그니처에서의 상응하는 교란 시그니처를 나타낸다.

[0322] 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 모델의 임의의 적합한 실시예, 예컨대 "모델 아키텍처"라는 명칭의 상기 섹션에 개시된 것, 및 그의 임의의 치환, 변형, 추가, 결실 및/또는 조합이 고려된다. 예를 들어, 일부 실시예에서, 모델은 신경망을 포함한다. 일부 이러한 실시예에서, 신경망은 ReLU 활성화를 갖는 완전 연결 신경망이다. 일부 실시예에서, 신경망은 메시지 전달 신경망이다.

[0323] 일부 실시예에서, 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0324] 일부 실시예에서, 모델은 복수의 컴포넌트 모델의 앙상블 모델이고, 하나 이상의 계산된 활성화 점수에서의 각각의 계산된 활성화 점수는 복수의 컴포넌트 모델의 각각의 컴포넌트 모델의 출력의 중심 집중 경향의 척도이다.

[0325] 일부 실시예에서, 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0326] 일부 실시예에서, 복수의 컴포넌트 모델은 복수의 신경망을 포함한다. 일부 이러한 실시예에서, 복수의 신경망 내의 제1 신경망은 ReLU 활성화를 갖는 완전 연결 신경망이고, 복수의 신경망 내의 제2 신경망은 메시지 전달 신경망이다.

[0327] 앞서 정의된 바와 같이, 교란은 하나 이상의 조건에 대한 세포의 임의의 노출, 예컨대 1종 이상의 화합물에 의한 치료를 지칭한다. 일부 실시예에서, 교란 시그니처는 교란에 의해 유도된 세포 내의 하나 이상의 세포 구성성분의 발현 또는 풍부도 수준에서의 변화이다.

[0328] 예시적인 교란은 유전자 녹다운, 자극에 대한 세포 반응, 조직 성장 및 재생, 및/또는 화합물을 사용한 치료 또는 화합물에 대한 노출을 포함하나, 이에 제한되지는 않는다. 예시적인 교란원은 소분자, 생물제제, 치료제, 단백질, 소분자와 조합된 단백질, ADC, 핵산, 예컨대 siRNA 또는 간섭 RNA, 야생형 및/또는 돌연변이체 shRNA를 과다발현하는 cDNA, 야생형 및/또는 돌연변이체 가이드 RNA를 과다발현하는 cDNA(예를 들어, Cas9 시스템 또는 다른 유전자 편집 시스템), 또는 임의의 전술한 것의 임의의 조합을 포함하나, 이에 제한되지는 않는다.

[0329] 일부 실시예에서, 시스템 수준(예를 들어, 결합 또는 도킹 활성)에서 및/또는 하류 효과 및 기관-수준 표현형과 관련하여 교란이 특징화된다. 일부 실시예에서, 교란은(예를 들어, 교란 전 또는 후에 바이오마커, 세포 생존율 및/또는 약물-단백질 상호작용을 식별 또는 측정함으로써) 분자, 세포 및/또는 조직 수준에서 교란원에 대한 반응을 유도하거나 또는 그 근간이 되는 메커니즘의 함수로서 특징화된다. 예를 들어, 교란의 측정은 표현형 측정(예를 들어, IC50 값) 및/또는 세포 구성성분 시그니처(예를 들어, 체학 프로파일링)를 포함할 수 있다.

[0330] 일부 실시예에서, 각각의 교란 및/또는 상응하는 교란 시그니처는 공개적으로 이용가능한 데이터베이스, 예컨대 암에서의 약물 감수성의 유전체학, 암 치료 반응 포털, 연결성 맵, PharmacoDB, BoBER(Base of Bioisosterically Exchangeable Replacements), DrugBank, 인간 세포 아틀라스, MSigDB(Molecular Signatures

Database), 및/또는 Enrichr로부터 획득된다. 교란 데이터가 획득될 수 있는 다른 적합한 데이터베이스는 NIH 유전자 발현 옴니버스(GEO), EBI ArrayExpress, NCBI, BLAST, EMBL-EBI, GenBank, Ensembl, KEGG 경로 데이터베이스, LINCS(Library of Integrated Network-based Cellular Signatures) L1000 데이터세트, 리액톰 경로 데이터베이스, 및/또는 유전자 온톨로지 프로젝트를 포함한다.

[0331] 교란 데이터를 획득하는 방법은, 예를 들어, 교란-seq, CRISP-seq, CROP-seq, CRISPRi, TAP-seq, CRISPRa, 교란-CITE-seq, sci-Plex, 멀티플렉스화, MIX-seq, CyTOF, 및/또는 scRNA-seq를 사용하는 세포 구성성분 데이터의 측정을 포함한다. 교란 데이터를 획득하는 방법은 질량 분광측정법(예를 들어, LCMS, GCMS), 유동 세포측정법, 정량적 폴리머라제 연쇄 반응(qPCR), 겔 전기영동, 유전자-칩 분석, 마이크로어레이, 세포형광측정 분석, 형광 현미경검사, 공초점 레이저 스캐닝 현미경검사, 레이저 스캐닝 세포측정법, 친화성 크로마토그래피, 수동 배치 모드 분리, 전기장 현탁액, 시퀀싱, 및/또는 그 임의의 조합을 포함한, 체학 데이터를 획득하는 임의의 방법을 더 포함한다. 일부 실시예에서, 본원에 개시된 세포 구성성분 풍부도 값을 획득하기 위한 임의의 방법은 (예를 들어, 교란 시그니처에 대한) 교란 데이터를 획득하는데 사용하기 위해 고려된다.

[0332] 일부 실시예에서, 교란 시그니처의 세트는 제1 교란 시그니처로 이루어진다. 일부 실시예에서, 교란 시그니처의 세트는 5개 이상의 교란 시그니처를 포함한다. 일부 실시예에서, 교란 시그니처의 세트는 10개 이상의 교란 시그니처를 포함한다. 일부 실시예에서, 교란 시그니처의 세트는 100개 이상의 교란 시그니처를 포함한다.

[0333] 일부 실시예에서, 교란 시그니처의 세트는 적어도 2개, 적어도 3개, 적어도 4개, 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 800개, 적어도 1000개, 적어도 2000개, 또는 적어도 5000개의 교란 시그니처를 포함한다. 일부 실시예에서, 교란 시그니처의 세트는 10,000개 이하, 5000개 이하, 1000개 이하, 800개 이하, 500개 이하, 200개 이하, 100개 이하, 50개 이하, 또는 20개 이하의 교란 시그니처를 포함한다. 일부 실시예에서, 교란 시그니처의 세트는 5개 내지 50개, 2개 내지 100개, 20개 내지 500개, 10개 내지 1000개, 800개 내지 5000개, 또는 50개 내지 2000개의 교란 시그니처를 포함한다. 일부 실시예에서, 교란 시그니처의 세트는 2개 이상의 교란 시그니처에서 시작하여 10,000개 이하의 교란 시그니처에서 끝나는 또 다른 범위에 속한다.

[0334] 블록 706을 참조하면, 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별, 및 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하고, 여기서 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태이다.

[0335] 일부 실시예에서, 복수의 교란 시그니처에서 각각의 교란 시그니처의 교란된 상태는 복수의 화합물 중의 화합물에 노출되지 않은 대조군 세포에 의해 나타난다. 일부 실시예에서, 복수의 교란 시그니처에서 각각의 교란 시그니처의 교란된 상태는 각각의 교란 시그니처와 연관된 화합물 이외의 복수의 화학적 화합물의 화학적 화합물에 노출된 관련되지 않은 교란된 세포에 걸친 평균에 의해 나타내어진다.

[0336] 일부 실시예에서, 세포 상태에서의 변화는 변경되지 않은 세포 상태와 변경된 세포 상태 사이의 변화를 지칭하며, 여기서 변경된 세포 상태는 변경되지 않은 세포 상태로부터 변경된 세포 상태로의 세포 전이를 통해 발생한다. 또한, (i) 변경되지 않은 세포 상태, (ii) 변경된 세포 상태, 및 (iii) 변경되지 않은 세포 상태로부터 변경된 세포 상태로의 전이 중 적어도 하나는 관심 생리학적 조건과 연관된다.

[0337] 일부 실시예에서, 교란 시그니처의 세트의 각각의 교란 시그니처는 비제한적인 예로서 본원에서 참조로 포함된, 2019년 7월 15일 출원된 발명의 명칭이 "세포 분석 방법"인 미국 특허 출원 번호 16/511,691에 개시된 임의의 방법을 사용하여 결정될 수 있다.

[0338] 특정 실시예에서, 교란의 공변량(예를 들어, 세포의 특정한 화학적 조성물에 대한 노출)이 존재할 수 있다. 예를 들어, 화학적 조성물의 공변량은 화학적 조성물의 특정 용량, 화학적 조성물에 노출된 세포를 측정하여 세포 구성성분을 정량화하는 시간, 및/또는 화학적 조성물에 노출된 세포의 아이덴티티(예를 들어, 세포주)를 포함할 수 있다. 일부 실시예에서, 교란(예를 들어, 세포의 특정한 화학적 조성물에 대한 노출)은 그의 공변량의 임계치 양이 또한 특정한 세포 전이에 영향을 미칠 것으로 예측되는 경우에만 특정한 세포 전이에 영향을 미칠 것으로 예측된다. 다시 말해서, 일부 실시예에서, 특정한 교란 시그니처의 계산된 활성화 점수는 적어도 부분적으

로, 특정한 교란 시그니처의 화학적 조성의 공변량이 또한 관심 생리학적 조건과 연관된 특정한 세포 전이에 영향을 미칠 것으로 예측되는지 여부에 의해 결정된다.

[0339] 일반적으로, 앞서 설명된 바와 같이, 훈련된 모델의 출력은 표지(예를 들어, 수치 활성화 점수)를 포함하는 훈련 데이터세트에 대해 학습하고, 훈련된 모델의 출력이 최소 성능 수준을 충족시킬 때까지, 예컨대 검증 단계를 통해 복수의 파라미터를 조정하는 과정을 통해 정의된다. 훈련 모델은 "모델 훈련"이라는 명칭의 섹션에서 하기에 추가로 개시된다. 따라서, 일부 이러한 실시예에서, 훈련된 모델은 출력으로서, 테스트 화학적 화합물과 관심 생리학적 조건의 연관을 나타내는 제1 교란 시그니처에 대한 계산된 활성화 점수를 제공한다(예를 들어, 여기서 제1 교란 시그니처는 관심 생리학적 조건과 연관된 세포 상태 전이와 연관됨).

[0340] 이어서, 블록 708을 참조하면, 방법은 교란 시그니처의 세트의 제1 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족할 때 관심 생리학적 조건을 갖는 화학적 화합물을 식별하는 것을 포함한다.

[0341] "활성화 점수"라는 명칭의 섹션에 개시된 바와 같은 활성화 점수의 임의의 적합한 실시예는 하나 이상의 계산된 활성화 점수를 획득하기 위해 고려되며, 여기서 각각의 활성화 점수는 관련 기술분야의 통상의 기술자에게 명백할 바와 같이 임의의 치환, 변형, 추가, 결실 및/또는 그의 조합을 포함한 교란 시그니처의 세트에서 상응하는 교란 시그니처를 나타낸다.

[0342] 일반적으로, (제1 임계치 기준을 충족하는 계산된 활성화 점수를 가짐으로써 입증된 바와 같이) 관심 생리학적 조건으로 식별되는 화학적 화합물이 추구된다. 예를 들어, 일부 실시예에서, 제1 임계치의 충족은 제1의 미리 결정된 수치값을 초과하는 활성화 점수를 필요로 한다.

[0343] 예를 들어, 일부 실시예에서, 활성화 점수는 "0" 내지 "1"(또는 A 및 B가 2개의 상이한 수인 일부 다른 범위 "A" 내지 "B")의 연속 척도 상의 정규화된 값으로서 표현되고, 여기서 "1"에 더 가까운 값(예를 들어, .89, .90, .91, .92 등)은 교란 시그니처(및 교란 시그니처가 나타내는 화학적 화합물)과 관심 생리학적 조건 사이의 강한 연관성을 나타낸다. "0"에 더 가까운 값(예를 들어, 0.01, 0.02, 0.03, 0.04 등)은 교란 시그니처(및 교란 시그니처가 나타내는 화학적 화합물)과 관심 생리학적 조건 사이에 연관성이 없음을 나타낸다. 이러한 경우에, 제1 임계치는 "0" 내지 "1"(또는 일부 다른 범위 "A" 내지 "B", 여기서 A 및 B는 2개의 상이한 수임)로 선택되고, 교란 시그니처(및 그것이 나타내는 화학 구조)는 활성화 점수가 제1 임계치를 초과하는 경우에 관심 생리학적 조건과 연관된 것으로 간주되는 반면, 교란 시그니처(및 그것이 나타내는 화학 구조)는 활성화 점수가 제1 임계치 미만인 경우에 관심 생리학적 조건과 연관되지 않은 것으로 간주된다. 일부 이러한 실시예에서, 활성화 점수는 "0" 내지 "1"(또는 A 및 B가 2개의 상이한 수인 일부 다른 범위 "A" 내지 "B")의 연속 척도 상의 정규화된 값으로서 표현되고, 제1 임계치는 0 내지 1, 0.10 내지 0.90, 0.20 내지 0.80, 0.30 내지 0.70, 0.50 내지 0.99, 0.60 내지 0.99, 0.70 내지 0.99, 0.80 내지 0.99, 또는 0.90 내지 0.99의 값이다.

[0344] 또 다른 예로서, 일부 실시예에서, 활성화 점수는 "0" 내지 "1"(또는 A 및 B가 2개의 상이한 수인 일부 다른 범위 "A" 내지 "B")의 연속 척도 상의 정규화된 값으로서 표현되고, 여기서 "1"에 더 가까운 값(예를 들어, 0.89, 0.90, 0.91, 0.92 등)은 교란 시그니처(및 교란 시그니처가 나타내는 화학적 화합물)과 관심 생리학적 조건 사이에 연관성이 없음을 나타낸다. "0"에 더 가까운 값(예를 들어, 0.01, 0.02, 0.03, 0.04 등)은 교란 시그니처(및 교란 시그니처가 나타내는 화학적 화합물)와 관심 생리학적 조건 사이에 연관성이 없음을 나타낸다. 이러한 경우에, 제1 임계치는 "0" 내지 "1"(또는 일부 다른 범위 "A" 내지 "B", 여기서 A 및 B는 2개의 상이한 수임)로 선택되고, 교란 시그니처(및 그것이 나타내는 화학 구조)는 활성화 점수가 제1 임계치 미만인 경우에 관심 생리학적 조건과 연관된 것으로 간주되는 반면, 교란 시그니처(및 그것이 나타내는 화학 구조)는 활성화 점수가 제1 임계치를 초과하는 경우에 관심 생리학적 조건과 연관되지 않은 것으로 간주된다. 일부 이러한 실시예에서, 활성화 점수는 "0" 내지 "1"(또는 A 및 B가 2개의 상이한 수인 일부 다른 범위 "A" 내지 "B")의 연속 척도 상의 정규화된 값으로서 표현되고, 제1 임계치는 0 내지 1, 0.10 내지 0.90, 0.20 내지 0.80, 0.30 내지 0.70, 0.50 내지 0.99, 0.60 내지 0.99, 0.70 내지 0.99, 0.80 내지 0.99, 또는 0.90 내지 0.99의 값이다.

[0345] 일부 실시예에서, 제1 임계치 기준은 제1 교란 시그니처가 임계 활성화 점수를 가져야 한다는 요건이다.

[0346] 일부 실시예에서, 제1 임계치 기준은 교란 시그니처의 세트에서 제1 교란 시그니처가 적어도 임계 순위를 가져야 한다는 요건이며, 여기서 교란 시그니처의 세트는 교란 시그니처의 세트의 각각의 교란 시그니처와 참조 시그니처(예를 들어, 단일-세포 전이 시그니처)의 비교에 기초하여 순위화된다. 화학적 화합물을 생리학적 조건과 연관시키는데 사용하기에 적합한 기준 시그니처(예를 들어, 단일-세포 전이 시그니처)에 대한 교란 시그니처의 비교 방법은 "교란 시그니처에 대한 수치 활성화 점수"라는 명칭의 섹션에서 하기에 추가로 상세하게 설명된

다.

[0347] 일부 실시예에서, 식별은 교란 시그니처의 세트의 각각의 교란 시그니처의 각각의 계산된 활성화 점수가 임계치 기준을 충족시킬 것을 요구한다. 일부 실시예에서, 식별은 교란 시그니처의 세트의 각각의 교란 시그니처의 각각의 계산된 활성화 점수에 걸친 중심 집중 경향의 척도가 임계치 기준을 충족시킬 것을 요구한다. 일부 실시예에서, 중심 집중 경향의 척도는 산술 평균, 가중 평균, 중간범위, 미드힌지, 삼평균, 윈저화 평균, 평균, 또는 교란 시그니처의 세트에서의 각각의 교란 시그니처의 각각의 개별 계산된 활성화 점수의 모드이다.

[0348] 일부 실시예에서, 교란 시그니처의 세트는 2개 내지 100개의 교란 시그니처이고, 식별은 교란 시그니처의 세트의 각각의 교란 시그니처의 각각의 계산된 활성화 점수가 임계치 기준을 충족시킬 것을 요구한다. 일부 실시예에서, 교란 시그니처의 세트는 2개 내지 100개의 교란 시그니처이고, 식별은 교란 시그니처의 세트의 각각의 교란 시그니처의 각각의 계산된 활성화 점수에 걸친 중심 집중 경향의 척도를 필요로 하고, 임계치 기준을 충족시킨다. 일부 실시예에서, 중심 집중 경향의 척도는 산술 평균, 가중 평균, 중간범위, 미드힌지, 삼평균, 윈저화 평균, 평균, 또는 교란 시그니처의 세트에서의 각각의 교란 시그니처의 각각의 개별 계산된 활성화 점수의 모드이다.

[0349] 일부 실시예에서, 교란 시그니처의 세트는 복수의 교란 시그니처이고, 제1 교란 시그니처를 비롯한 복수의 교란 시그니처의 제1 서브세트는 관심 생리학적 조건과 연관되고, 복수의 교란 시그니처의 제2 서브세트는 관심 생리학적 조건과 연관되지 않고, 제1 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족하고, 복수의 교란 시그니처의 제2 서브세트에서의 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제2 임계치 기준을 충족할 때 테스트 화학적 화합물은 관심 생리학적 조건으로 식별된다.

[0350] 일부 실시예에서, 제2 임계치 기준은 복수의 교란 시그니처의 제2 서브세트에서의 교란 시그니처에 대한 각각의 계산된 활성화 점수가 임계 활성화 점수를 가져야 한다는 요건이다.

[0351] 일부 실시예에서, 제2 임계치 기준은 복수의 교란 시그니처의 제2 서브세트의 교란 시그니처에 대한 각각의 계산된 활성화 점수가 교란 시그니처의 세트 내에서 적어도 임계 순위를 가져야 한다는 요건이고, 이때 교란 시그니처의 세트는 교란 시그니처의 세트의 각각의 교란 시그니처와 기준 시그니처(예를 들어, 단일-세포 전이 시그니처)의 비교를 기초로 순위화된다.

[0352] 일부 실시예에서, 식별은 교란 시그니처의 제2 서브세트에서의 각각의 교란 시그니처의 각각의 계산된 활성화 점수가 제2 임계치 기준을 충족시킬 것을 요구한다. 일부 실시예에서, 식별은 교란 시그니처의 제2 서브세트의 각각의 교란 시그니처의 각각의 계산된 활성화 점수에 걸친 중심 집중 경향의 척도가 제2 임계치 기준을 충족시킬 것을 요구한다. 일부 실시예에서, 중심 집중 경향의 척도는 산술 평균, 가중 평균, 중간범위, 미드힌지, 삼평균, 윈저화 평균, 평균, 또는 교란 시그니처의 세트에서의 각각의 교란 시그니처의 각각의 개별 계산된 활성화 점수의 모드이다.

[0353] **III. 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법**

[0354] *모델 훈련.*

[0355] 본 개시의 또 다른 양태는 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법(800)을 제공한다. 일부 실시예에서, 관심 생리학적 조건은 질환이다.

[0356] 블록 802를 참조하면, 방법은, 전자 형태로, 복수의 화합물에서의 각각의 화합물의 화학 구조의 각각의 지문을 획득함으로써 복수의 지문을 획득하는 것을 포함한다. "생리학적 조건" 및 "화합물"이라는 명칭의 상기 섹션에 개시된 바와 같은 생리학적 조건, 화합물, 지문, 및/또는 지문을 획득하는 방법의 임의의 적합한 실시예가 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 그의 임의의 치환, 변형, 추가, 결실, 및/또는 조합을 포함하여 고려된다.

[0357] 예를 들어, 일부 실시예에서, 복수의 화합물은 10개 내지 1×10^6개의 화합물이다. 일부 실시예에서, 복수의 화합물은 100개 내지 100,000개의 화합물이다. 일부 실시예에서, 복수의 화합물은 1000개 내지 100,000개의 화합물이다.

[0358] 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물이다. 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 리핀스키 5 준칙 각각을 충족시킨다. 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 리핀스키 5 준칙의 적어도 3가지 기준

을 충족시킨다. 일부 실시예에서, 각각의 개별 지문은 SMILES 변환기, ECFP4, RNNS2S 또는 GraphConv를 사용하여 화학 구조로부터 생성된다.

[0359] 블록 804를 참조하면, 방법은 복수의 화합물의 각각의 화합물에 대한 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수를 전자 형태로 획득하는 것을 포함하며, 여기서 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함한다. 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 그 임의의 치환, 변형, 추가, 결실, 및/또는 조합을 포함하여, "세포 구성성분 및 세포 구성성분 모듈" 및 하기 "세포 구성성분 모듈의 식별"이라는 명칭의 상기 섹션에 개시된 바와 같은 세포 구성성분, 세포 구성성분 모듈, 및/또는 세포 구성성분 모듈 식별 방법의 임의의 적합한 실시예가 고려된다.

[0360] 예를 들어, 일부 실시예에서, 세포 구성성분 모듈의 세트는 단일 세포 구성성분 모듈이다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈이다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 2개 내지 500개의 세포 구성성분 모듈이다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 단일 세포 구성성분 모듈로 이루어진다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 5개 이상의 세포 구성성분 모듈을 포함한다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 10개 이상의 세포 구성성분 모듈을 포함한다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 100개 이상의 세포 구성성분 모듈을 포함한다. 일부 실시예에서, 세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈이고, 복수의 세포 구성성분 모듈의 제1 서브세트는 관심 생리학적 조건과 연관되고, 복수의 세포 구성성분 모듈의 제2 서브세트는 관심 생리학적 조건과 연관되지 않는다.

[0361] 일부 실시예에서, 도 2a 및 도 2b의 예시적인 작업흐름에 의해 예시된 바와 같이, 방법은 전자 형태로 하나 이상의 제1 데이터세트를 획득하는 것을 포함하는 과정에 의해 복수의 세포 구성성분 모듈 내의 세포 구성성분 모듈을 식별하는 단계를 더 포함하고, 하나 이상의 제1 데이터세트는 제1 복수의 세포의 각각의 개별 세포에 대해(제1 복수의 세포는 20개 이상의 세포를 포함하고 복수의 주석화된 세포 상태를 집합적으로 나타냄), 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해(복수의 세포 구성성분은 10개 이상의 세포 구성성분을 포함함), 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함한다. 따라서, 방법은 복수의 벡터에 액세스하거나 또는 복수의 벡터를 형성하고, 복수의 벡터의 각각의 개별 벡터는 (i) 복수의 구성성분의 각각의 세포 구성성분에 상응하고, (ii) 상응하는 복수의 요소를 포함하며, 상응하는 복수의 요소의 각각의 개별 요소는 제1 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 갖는다. 복수의 벡터는 복수의 후보 세포 구성성분 모듈에서 각각의 후보 세포 구성성분 모듈을 식별하는데 사용된다. 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 복수의 세포 구성성분의 서브세트를 포함하고, 이때 복수의 세포 구성성분 모듈은 (i) 복수의 후보 세포 구성성분 모듈 및 (ii) 복수의 세포 구성성분 또는 그 표현으로 차원화된 잠재 표현으로 배열되고, 복수의 세포 구성성분 모듈은 10개를 초과하는 세포 구성성분 모듈을 포함한다.

[0362] 하나 이상의 제2 데이터세트가 전자 형태로 획득되고, 하나 이상의 제2 데이터세트는 제2 복수의 세포의 각각의 개별 세포에 대해- 제2 복수의 세포는 20개 이상의 세포를 포함하고, 관심 생리학적 조건을 알리는 복수의 공변량을 집합적으로 나타냄 -, 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함한다. 따라서, (i) 제2 복수의 세포 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 세포 구성성분 카운트 데이터 구조가 획득된다. 활성화 데이터 구조는 복수의 세포 구성성분 또는 그 표현을 공통 차원으로서 사용하여 세포 구성성분 카운트 데이터 구조 및 잠재 표현을 조합함으로써 형성되고, 여기서 활성화 데이터 구조는 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈에 대해, 제2 복수의 세포의 각각의 세포에 대해 각각의 활성화 가중치를 포함한다.

[0363] 후보 세포 구성성분 모델은 (i) 후보 모델로의 활성화 데이터 구조의 입력 시 활성화 데이터 구조에 표현된 각각의 세포 구성성분 모듈 내의 복수의 공변량 중 각각의 공변량의 부재 또는 존재의 예측과 (ii) 각각의 세포 구성성분 모듈의 각각의 공변량의 실제 부재 또는 존재 사이의 차이를 사용하여 훈련되고, 여기서 훈련은 차이에 응답하여 후보 세포 구성성분 모델과 연관된 복수의 공변량 가중치를 조정하고, 복수의 공변량 가중치는, 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대해, 각각의 개별 공변량에 대해, 각각의 공변량이 활성화 데이터 구조에 걸쳐 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 상응하는 가중치를 포함한다. 후보 세포 구성성분 모델의 훈련 시, 복수의 공변량 가중치를 사용하여 복수의 후보 세포 구성성분 모듈(예를 들어, 관심 생리학적 조건과 연관됨)에서 세포 구성성분 모듈을 식별한다.

[0364] 일부 실시예에서, 관심 생리학적 조건은 질환이고, 제1 복수의 세포는 복수의 주석화된 세포 상태에 의해 입증

된 바와 같이 질환을 대표하는 세포 및 질환을 대표하지 않는 세포를 포함한다. 일부 실시예에서, 복수의 주석화된 세포 상태의 주석화된 세포 상태는 노출 조건 하에 화합물에 대한 제1 복수의 세포의 세포의 노출이다. 일부 실시예에서, 노출 조건은 노출 지속기간, 화합물의 농도, 또는 노출 지속기간과 화합물의 농도의 조합이다.

[0365] 일부 실시예에서, 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합이다. 일부 실시예에서, 제1 또는 제2 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도는 비색 측정치, 형광 측정치, 발광 측정치, 또는 공명 에너지 전달(FRET) 측정치에 의해 결정된다. 일부 실시예에서, 제1 또는 제2 복수의 세포에서의 각각의 세포에서의 각각의 세포 구성성분의 상응하는 풍부도는 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq), scTag-seq, 시퀀싱을 사용한 전위효소-접근 가능 염색질에 대한 단일-세포 검정(scATAC-seq), CyTOF/SCoP, E-MS/Abseq, miRNA-seq, CITE-seq, 및 그 임의의 조합에 의해 결정된다. 일부 실시예에서, 복수의 세포 구성성분은 100개 내지 8,000개의 세포 구성성분으로 이루어진다.

[0366] 일부 실시예에서, 복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계는 복수의 벡터의 각각의 벡터의 각각의 상응하는 복수의 요소를 사용하여 상관 모델을 복수의 벡터에 적용하는 것을 포함한다. 일부 실시예에서, 상관 모델은 그래프 클러스터링을 포함한다. 일부 실시예에서, 그래프 클러스터링 방법은 피어슨(Pearson) 상관-기반 거리 메트릭에 대한 라이덴(Leiden) 클러스터링이거나 또는 루뱅(Louvain) 클러스터링이다.

[0367] 일부 실시예에서, 복수의 세포 구성성분 모듈은 10개 내지 2000개의 세포 구성성분 모듈로 이루어진다. 일부 실시예에서, 복수의 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 2개 내지 300개의 세포 구성성분으로 이루어진다.

[0368] 일부 실시예에서, 복수의 공변량은 세포 배치, 세포 공여자, 세포 유형, 질환 상태 또는 화학적 화합물에 대한 노출을 포함한다.

[0369] 일부 실시예에서, 후보 세포 구성성분 모델을 훈련시키는 것은 멀티-태스크 공식화에서 범주형 교차-엔트로피 손실을 사용하여 수행되며, 여기서 복수의 공변량에서의 각각의 공변량은 복수의 비용 함수에서의 비용 함수에 상응하고, 복수의 비용 함수에서의 각각의 개별 비용 함수는 공통 가중 인자를 갖는다.

[0370] 블록 806을 참조하면, 방법은 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해, (i) 훈련되지 않은 모델로의 각각의 화합물의 화학 구조의 지문의 입력 시 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수와 (ii) 세포 구성성분 모듈의 세트의 각각의 화합물에 대한 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련되지 않은 모델을 훈련시키는 것을 더 포함한다.

[0371] 일부 실시예에서, 하나 이상의 계산된 활성화 점수의 활성화 점수는 각각의 화합물에 상응하는 각각의 세포 구성성분 모듈에 대한 각각의 활성화 가중치이다. 예를 들어, 일부 실시예에서, 활성화 점수는 도 2a 및 도 2b에 설명되고 도 5의 활성화 데이터 구조에 예시된 바와 같이 획득된 활성화 가중치이고, 이때 활성화 점수는 각각의 화합물로의 치료와 상관되고/되거나 이에 반응한 각각의(예를 들어, 제1) 세포 구성성분 모듈의 활성화(예를 들어, 유도 및/또는 차등 발현)를 표시한다.

[0372] 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 모델의 임의의 적합한 실시예, 예컨대 "모델 아키텍처"라는 명칭의 상기 섹션에 개시된 것, 및 그의 임의의 치환, 변형, 추가, 결실, 및/또는 조합이 고려된다. 예를 들어, 일부 실시예에서, 훈련된 모델은 신경망을 포함한다. 일부 실시예에서, 신경망은 ReLU 활성화를 갖는 완전 연결 신경망이다. 일부 실시예에서, 신경망은 메시지 전달 신경망이다. 일부 실시예에서, 훈련된 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0373] 일부 실시예에서, 훈련된 모델은 복수의 컴포넌트 모델의 앙상블 모델이고, 각각의 계산된 활성화 점수는 복수의 컴포넌트 모델에서 각각의 컴포넌트 모델의 출력의 중심 집중 경향의 척도이다. 일부 실시예에서, 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다. 일부 실시예에서, 복수의 컴포넌트 모델은 복수의 신경망을 포함한다. 일부

실시예에서, 복수의 신경망 내의 제1 신경망은 ReLU 활성화를 갖는 완전 연결 신경망이고, 복수의 신경망 내의 제2 신경망은 메시지 전달 신경망이다.

[0374] 블록 808을 참조하면, 훈련은 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고(복수의 파라미터는 100개 이상의 파라미터를 포함함), 이에 의해 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득한다.

[0375] 일부 실시예에서, 모델에 대한 입력은 복수의 활성화 점수를 포함하고, 각각의 개별 활성화 점수는 복수의 화합물의 각각의 화합물에 대한 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈에 상응한다. 각각의 개별 화합물에 대한 각각의 개별 세포 구성성분 모듈에 상응하는 활성화 점수는 모듈과 화합물 사이의 연관성(예를 들어, 가중치 및/또는 상관)을 식별하기 위해 멀티-태스크 모델을 훈련시키기 위한 표지(예를 들어, 모듈과 화합물 사이의 연관성의 실제 존재 또는 부재를 나타내는 수치 활성화 점수)로서 기능한다. 예를 들어, 앞서 설명된 바와 같이, 일부 실시예에서, 복수의 세포 구성성분 모듈의 제1 서브세트는 관심 생리학적 조건과 연관되고, 복수의 세포 구성성분 모듈의 제2 서브세트는 관심 생리학적 조건과 연관되지 않는다. 따라서, 일부 이러한 실시예에서, 연관의 실제 존재는 복수의 세포 구성성분 모듈의 제1 서브세트를 표지로서 사용하여 훈련 데이터세트에 포함될 수 있고, 연관의 실제 부재는 복수의 세포 구성성분 모듈의 제2 서브세트를 표지로서 사용하여 훈련 데이터세트에 포함될 수 있다.

[0376] 일부 실시예에서, 복수의 화합물은 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 800개, 적어도 1000개, 적어도 2000개, 적어도 3000개, 적어도 4000개, 적어도 5000개, 적어도 8000개, 적어도 10,000개, 적어도 20,000개, 적어도 30,000개, 적어도 50,000개, 적어도 80,000개, 적어도 100,000개, 적어도 200,000개, 적어도 500,000개, 적어도 800,000개, 적어도 1백만개, 또는 적어도 2백만개의 화합물을 포함하고, 여기서 복수의 화합물의 각각의 화합물에 대해, 모델에 대한 입력은 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대한 각각의 활성화 점수를 포함한다.

[0377] 일부 실시예에서, 복수의 화합물은 1천만개 이하, 5백만개 이하, 1백만개 이하, 500,000개 이하, 100,000개 이하, 50,000개 이하, 10,000개 이하, 8000개 이하, 5000개 이하, 2000개 이하, 1000개 이하, 800개 이하, 500개 이하, 200개 이하, 또는 100개 이하의 화합물을 포함하고, 여기서 복수의 화합물의 각각의 화합물에 대해, 모델에 대한 입력은 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대한 각각의 활성화 점수를 포함한다. 일부 실시예에서, 복수의 화합물은 10개 내지 500개, 100개 내지 10,000개, 5000개 내지 200,000개, 또는 10,000개 내지 1백만개의 화합물로 이루어지고, 여기서 복수의 화합물의 각각의 화합물에 대해, 모델에 대한 입력은 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대한 각각의 활성화 점수를 포함한다.

[0378] 일부 실시예에서, 앞서 설명된 바와 같이, 복수의 수치 활성화 점수 내의 각각의 수치 활성화 점수는 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대한, 복수의 화합물의 각각의 화합물에 대한 활성화 가중치이다(예를 들어, 도 5의 활성화 데이터 구조에 예시됨).

[0379] 앞서 설명된 바와 같이, 일부 실시예에서, 모델의 출력은 복수의 화합물의 각각의 화합물(예를 들어, 테스트 화학적 화합물)이 복수의 세포 구성성분 모듈의 각각의 하나 이상의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 하나 이상의 계산된 활성화 점수를 포함한다.

[0380] 일반적으로, 모델(예를 들어, 신경망)을 훈련시키는 것은 역전파(예를 들어, 구배 하강)를 통해 각각의 모델에 대한 복수의 파라미터(예를 들어, 가중치)를 업데이트하는 것을 포함한다. 먼저, 입력 데이터(예를 들어, 복수의 화합물의 각각의 개별 화합물에 대한, 복수의 모듈의 각각의 개별 세포 구성성분 모듈에 대한 복수의 활성화 점수)가 신경망 내로 수용되는 순방향 전파가 수행되고, 출력은 선택된 활성화 함수 및 파라미터(예를 들어, 가중치 및/또는 하이퍼파라미터)의 초기 세트에 기초하여 계산된다. 일부 실시예에서, 파라미터(예를 들어, 가중치 및/또는 하이퍼파라미터)는 훈련되지 않은 또는 부분적으로 훈련된 모델에 대해 무작위로 할당된다(예를 들어, 초기화된다). 일부 실시예에서, 파라미터는 이전에 저장된 복수의 파라미터로부터 또는 사전-훈련된 모델로부터(예를 들어, 전이 학습에 의해) 전달된다.

[0381] 그 후, 각각의 계층 내의 각각의 개별 유닛에 상응하는 각각의 개별 파라미터에 대한 오차 구배를 계산함으로써 역방향 통과가 수행되고, 여기서 각각의 파라미터에 대한 오차는 네트워크 출력(예를 들어, 계산된 활성화 점수로서 각각의 화합물과 각각의 세포 구성성분 모듈 사이의 연관성의 예측된 부재 또는 존재) 및 입력 데이터(예

를 들어, 예상 값 또는 참 표지; 수치 활성화 점수로서 각각의 화합물과 각각의 세포 구성성분 모듈 사이의 연관의 실제 부재 또는 존재)에 기초하여 손실(예를 들어, 오차)을 계산함으로써 결정된다. 이어서, 계산된 손실에 기초하여 값을 조정함으로써 파라미터(예를 들어, 가중치)를 업데이트함으로써 모델을 훈련시킨다.

[0382] 예를 들어, 기계 학습의 일부 일반적인 실시예에서, 역전파는 복수의 가중치(예를 들어, 임베딩)를 포함하는 은닉 계층으로 네트워크를 훈련시키는 방법이다. 훈련되지 않은 모델의 출력(예를 들어, 계산된 활성화 점수로서의 연관성의 예측된 부재 또는 존재)은 먼저 임의적으로 선택된 초기 가중치의 세트를 사용하여 생성된다. 이어서, 오차 함수를 평가하여 오차를 계산함으로써(예를 들어, 손실 함수를 사용하여), 출력을 원래의 입력(예를 들어, 수치 활성화 점수로서의 연관성의 실제 부재 또는 존재)과 비교한다. 이어서, 오차가 최소화되도록(예를 들어, 손실 함수에 따라) 가중치를 업데이트한다. 일부 실시예에서, 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 다양한 역전파 알고리즘 및/또는 방법 중 어느 하나를 사용하여 복수의 가중치를 업데이트한다.

[0383] 일부 실시예에서, 손실 함수는 평균 제곱 오차, 2차 손실, 평균 절대 오차, 평균 바이어스 오차, 힌지, 다중 클래스 서포트 벡터 머신, 및/또는 교차-엔트로피이다. 일부 실시예에서, 훈련되지 않은 또는 부분적으로 훈련된 모델을 훈련시키는 것은 구배 하강 알고리즘 및/또는 최소화 함수에 따라 오차를 계산하는 것을 포함한다. 일부 실시예에서, 훈련되지 않은 또는 부분적으로 훈련된 모델을 훈련시키는 것은 복수의 손실 함수를 사용하여 복수의 오차를 계산하는 것을 포함한다. 일부 실시예에서, 복수의 손실 함수에서의 각각의 손실 함수는 동일하거나 상이한 가중 인자를 받는다.

[0384] 도 6은 본 개시의 일부 실시예에 따른, 모델을 훈련시키기 위한 방법의 일 예를 나타낸 도면이다. 활성화 데이터 구조(상부 패널)는 복수의 K 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈과 복수의 G 세포의 각각의 세포 사이의 연관을 표시하는 복수의 활성화 점수를 포함하는 입력을 모델에 제공하고, 여기서 각각의 세포는 복수의 화합물의 각각의 화합물을 나타낸다. 복수의 세포 구성성분 모듈(중간 패널) 중 각각의 개별 세포 구성성분 모듈에 대해, 복수의 세포(예를 들어, W 화합물)에 의해 집합적으로 나타내어지는 복수의 화합물의 각각의 개별 화합물에 대한 상응하는 가중치가 훈련 전에(예를 들어, 랜덤 가중치로) 초기화된다. 따라서, 복수의 화합물 가중치는 화합물 가중치 행렬을 포함한다(중간 패널). 복수의 화합물 가중치의 조정은 (i) 훈련되지 않은 모델(예를 들어, 예측) 내로의 각각의 화합물의 화학 구조의 지문의 입력 시 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수와 (ii) 세포 구성성분 모듈의 세트(예를 들어, 실제) 내의 각각의 화합물에 대한 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수 사이의 차이를 사용하여 수행된다(하부 패널). 일부 실시예에서, 실제 활성화는 예를 들어 도 2a, 도 2b 및 도 14a 내지 도 14d를 참조하여 "세포 구성성분 모듈의 식별"이라는 명칭의 하기 섹션에 설명된 세포 구성성분 모듈의 식별 방법을 이용하여 획득되며, 여기서 복수의 공변량은 복수의 화합물을 포함한다. 이어서, 훈련된 모델이 형성될 때까지(예를 들어, 최소 수의 조정의 완료 및/또는 최소 성능 임계치의 만족을 통해) 훈련(예를 들어, 화합물 가중치의 조정)을 수행할 수 있다.

[0385] 일부 실시예에서, 오차 함수는 하나 이상의 파라미터의 값을 계산된 손실에 비례하는 양만큼 조정함으로써 모델(예를 들어, 신경망)에서 하나 이상의 파라미터(예를 들어, 가중치)를 업데이트함으로써 모델을 훈련시키는데 사용된다. 일부 실시예에서, 파라미터가 조정되는 양은 파라미터가 업데이트되는 정도 또는 중증도를 좌우하는 학습율 하이퍼파라미터에 의해 계량된다(예를 들어, 보다 작거나 보다 큰 조정). 따라서, 일부 실시예에서, 훈련은 학습율에 기초하여 복수의 파라미터의 전부 또는 서브세트를 업데이트한다. 일부 실시예에서, 학습율은 차등 학습율이다.

[0386] 일부 실시예에서, 모델(예를 들어, 신경망)을 훈련시키는 것은 상응하는 복수의 은닉 뉴런에서 각각의 은닉 뉴런의 상응하는 파라미터에 대한 정규화를 추가로 사용한다. 예를 들어, 일부 실시예에서, 정규화는 손실 함수에 페널티를 가산함으로써 수행되며, 여기서 페널티는 신경망에서의 파라미터의 값에 비례한다. 일반적으로, 정규화는 하나 이상의 파라미터에 페널티를 가산함으로써 모델의 복잡성을 감소시켜 이들 파라미터와 연관된 각각의 은닉 뉴런의 중요성을 감소시킨다. 이러한 실시는 보다 일반화된 모델을 생성하고, 데이터의 오버피팅을 감소시킬 수 있다. 일부 실시예에서, 정규화는 L1 또는 L2 페널티를 포함한다. 예를 들어, 일부 바람직한 실시예에서, 정규화는 하위 및 상위 파라미터에 대한 L2 페널티를 포함한다. 일부 실시예에서, 정규화는 공간 정규화(예를 들어, 선험적 및/또는 실험적 지식에 기초하여 결정됨) 또는 탈락 정규화를 포함한다. 일부 실시예에서, 정규화는 독립적으로 최적화된 페널티를 포함한다.

[0387] 일부 실시예에서, 모델과 연관된 복수의 화합물 가중치를 조정하는 것(예를 들어, 예측 표지와 실제 표지 사이의 차이에 반응함)을 포함하는 훈련 과정이 복수의 훈련 사례 내의 각각의 훈련 사례에 대해 반복된다.

[0388] 일부 실시예에서, 복수의 훈련 인스턴스는 적어도 3개, 적어도 4개, 적어도 5개, 적어도 6개, 적어도 7개, 적어

도 8개, 적어도 9개, 적어도 10개, 적어도 50개, 적어도 100개, 적어도 500개, 적어도 1000개, 적어도 2000개, 적어도 3000개, 적어도 4000개, 적어도 5000개, 또는 적어도 7500개의 훈련 인스턴스를 포함한다. 일부 실시예에서, 복수의 훈련 인스턴스는 10,000개 이하, 5000개 이하, 1000개 이하, 500개 이하, 100개 이하, 또는 50개 이하의 훈련 인스턴스를 포함한다. 일부 실시예에서, 복수의 훈련 인스턴스는 3개 내지 10개, 5개 내지 100개, 100개 내지 5000개, 또는 1000개 내지 10,000개의 훈련 인스턴스를 포함한다. 일부 실시예에서, 복수의 훈련 사례는 3개 이상의 훈련 사례에서 시작하여 10,000개 이하의 훈련 사례에서 끝나는 또 다른 범위 내에 속한다.

[0389] 일부 이러한 실시예에서, 훈련은 복수의 훈련 사례에 걸쳐(예를 들어, 역전파를 통해) 모델의 파라미터의 조정을 반복하는 것을 포함하고, 따라서 각각의 화합물이 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 것에서 모델의 정확도를 증가시킨다.

[0390] 일부 실시예에서, 훈련은 전이 학습을 포함한다. 전이 학습은, 예를 들어 정의 섹션(상기 "훈련되지 않은 모델" 참조)에 추가로 기재되어 있다.

[0391] 일부 실시예에서, 훈련되지 않은 또는 부분적으로 훈련된 모델을 훈련시키는 것은 오차 함수의 제1 평가에 이어 훈련된 모델을 형성한다. 일부 이러한 실시예에서, 훈련된 모델은 오차 함수의 제1 평가에 기초하여 하나 이상의 파라미터의 제1 업데이트 후에 형성된다. 일부 대안적 실시예에서, 훈련된 모델은 오차 함수의 적어도 1회, 적어도 2회, 적어도 3회, 적어도 4회, 적어도 5회, 적어도 6회, 적어도 7회, 적어도 8회, 적어도 9회, 적어도 10회, 적어도 20회, 적어도 30회, 적어도 40회, 적어도 50회, 적어도 100회, 적어도 500회, 적어도 1000회, 적어도 10,000회, 적어도 50,000회, 적어도 100,000회, 적어도 200,000회, 적어도 500,000회, 또는 적어도 1백만회의 평가 후에 형성된다. 일부 이러한 실시예에서, 훈련된 모델은 오차 함수의 적어도 1, 적어도 2, 적어도 3, 적어도 4, 적어도 5, 적어도 6, 적어도 7, 적어도 8, 적어도 9, 적어도 10, 적어도 20, 적어도 30, 적어도 40, 적어도 50, 적어도 100, 적어도 500, 적어도 1000, 적어도 10,000, 적어도 50,000, 적어도 100,000, 적어도 200,000, 적어도 500,000, 또는 적어도 1백만회의 평가에 기초하여 하나 이상의 파라미터의 적어도 1, 적어도 2, 적어도 3, 적어도 4, 적어도 5, 적어도 6, 적어도 7, 적어도 8, 적어도 9, 적어도 10, 적어도 20, 적어도 30, 적어도 40, 적어도 50, 적어도 100, 적어도 500, 적어도 1000, 적어도 10,000, 적어도 50,000, 적어도 100,000, 적어도 200,000, 적어도 500,000, 또는 적어도 1백만회의 업데이트 후에 형성된다.

[0392] 일부 실시예에서, 훈련된 모델은 모델이 최소 성능 요건을 충족할 때 형성된다. 예를 들어, 일부 실시예에서, 훈련된 모델은 오차 함수(예를 들어, 각각의 화합물과 각각의 세포 구성성분 모듈 사이의 예측된 연관과 실제 연관 사이의 차이)의 평가 후에 훈련된 모델에 대해 계산된 오차가 오차 임계치를 충족하는 경우에 형성된다. 일부 실시예에서, 오차 함수에 의해 계산된 오차는 오차가 20% 미만, 18% 미만, 15% 미만, 10% 미만, 5% 미만, 또는 3% 미만일 때 오차 임계치를 충족시킨다.

[0393] 예시적인 실시예에서, 모델의 훈련은 복수의 공변량에서의 각각의 공변량이 복수의 비용 함수 중의 비용 함수에 상응하고 복수의 비용 함수의 각각의 개별 비용 함수가 공통 가중 인자를 갖는 멀티-태스크 공식화에서의 범주형 교차-엔트로피 손실을 사용하여 수행된다.

[0394] 일부 실시예에서, 훈련은 회귀 모델에 따라 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대한 각각의 개별 화합물과 연관된 각각의 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정한다. 일부 실시예에서, 회귀 모델은 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대한 각각의 개별 화합물과 연관된 각각의 차이의 최소 제곱 오차를 최적화한다.

[0395] 모델 훈련의 상기 논의는 화합물과 세포 구성성분 모듈 사이의 연관성을 나타내는 활성화 점수를 얻고 사용하는 것을 설명하고 있지만, 실제로, 화합물과 임의의 다른 관심 생리학적 조건 사이의 연관성을 나타내는 활성화 점수, 또는 그의 임의의 세포 과정은 화합물을 생리학적 조건과 연관시키기 위해 모델을 훈련시키고 사용하는 데 사용하기 위해 고려된다. 예를 들어, 하기 섹션에서 설명될 바와 같이, 본 개시의 또 다른 양태는 교란 시그니처를 사용하여 모델을 훈련시키는 것을 포함한다. 구체적으로, 일부 실시예에서, 훈련 표지로서 교란 시그니처에 대한 수치 활성화 점수를 사용하여 모델이 훈련된다. 이어서, 훈련된 모델을 방법(700)에 설명된 바와 같이 사용하여, 출력으로서, 모델로의 화학 구조 지문의 입력에 반응성인 하나 이상의 계산된 활성화 점수를 획득하며, 여기서 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 교란 시그니처의 세트의 상응하는 교란 시그니처를 나타낸다.

[0396] *교란 시그니처에 대한 수치 활성화 점수의 획득.*

[0397] 따라서, 본 개시의 또 다른 양태는 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법(900)을 제공한다.

일부 실시예에서, 관심 생리학적 조건은 질환이다.

[0398] 블록 902를 참조하면, 방법은, 전자 형태로, 복수의 화합물에서의 각각의 화합물의 화학 구조의 각각의 지문을 획득함으로써 복수의 지문을 획득하는 것을 포함한다. "생리학적 조건" 및 "화합물"이라는 명칭의 상기 섹션에 개시된 바와 같은 생리학적 조건, 화합물, 지문, 및/또는 지문을 획득하는 방법의 임의의 적합한 실시예가 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 그의 임의의 치환, 변형, 추가, 결실, 및/또는 조합을 포함하여 고려된다.

[0399] 예를 들어, 일부 실시예에서, 복수의 화합물은 10개 내지 1 x 10^6개의 화합물이다. 일부 실시예에서, 복수의 화합물은 100개 내지 100,000개의 화합물이다. 일부 실시예에서, 복수의 화합물은 1000개 내지 100,000개의 화합물이다. 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물이다. 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 리핀스키 5 준칙 각각을 충족시킨다. 일부 실시예에서, 복수의 화학적 화합물의 각각의 화학적 화합물은 리핀스키 5 준칙의 적어도 3가지 기준을 충족시킨다. 일부 실시예에서, 각각의 개별 지문은 SMILES 변환기, ECFP4, RNNS2S 또는 GraphConv를 사용하여 화학 구조로부터 생성된다.

[0400] 블록 904를 참조하면, 방법은 복수의 화합물의 각각의 상응하는 화합물에 대한 교란 시그니처의 세트의 각각의 개별 교란 시그니처의 각각의 수치 활성화 점수를 전자 형태로 획득하는 것을 포함한다. 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 그 임의의 치환, 변형, 추가, 결실, 및/또는 조합을 포함하는, 상기 섹션 명칭 "교란 시그니처"에 개시된 바와 같은 교란 시그니처의 임의의 적합한 실시예가 고려된다.

[0401] 예를 들어, 일부 실시예에서, 교란 시그니처의 세트는 단일 교란 시그니처이다. 일부 실시예에서, 교란 시그니처의 세트는 복수의 교란 시그니처이다. 일부 실시예에서, 교란 시그니처의 세트는 2개 내지 500개의 교란 시그니처이다. 일부 실시예에서, 교란 시그니처의 세트는 5개 이상의 교란 시그니처를 포함한다. 일부 실시예에서, 교란 시그니처의 세트는 10개 이상의 교란 시그니처를 포함한다. 일부 실시예에서, 교란 시그니처의 세트는 100개 이상의 교란 시그니처를 포함한다. 일부 실시예에서, 교란 시그니처의 세트는 복수의 교란 시그니처이고, 복수의 교란 시그니처의 제1 서브세트는 관심 생리학적 조건과 연관되고, 복수의 교란 시그니처의 제2 서브세트는 관심 생리학적 조건과 연관되지 않는다.

[0402] 블록 906을 참조하면, 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별, 및 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하고, 여기서 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태이다.

[0403] 일부 실시예에서, 교란 시그니처의 세트의 각각의 교란 시그니처의 각각의 수치 활성화 점수는 변경되지 않은 세포 상태와 변경된 세포 상태 사이의 차등 세포 구성성분 풍부도의 척도를 나타내는 단일-세포 전이 시그니처에 전자 형태로 액세스하는 것을 포함하는 절차에 의해 획득된다. 변경된 세포 상태는 변경되지 않은 세포 상태로부터 변경된 세포 상태로의 세포 전이를 통해 발생하고, 여기서 (i) 변경되지 않은 세포 상태, (ii) 변경된 세포 상태, 및 (iii) 변경되지 않은 세포 상태로부터 변경된 세포 상태로의 전이 중 적어도 하나는 관심 생리학적 조건과 연관된다. 단일-세포 전이 시그니처는 기준 복수의 세포 구성성분의 식별, 및 복수의 기준 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 변경되지 않은 세포 상태와 변경된 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 제1 유의성 점수를 포함한다. 각각의 교란 시그니처의 각각의 수치 활성화 점수를 결정하기 위해 단일-세포 전이 시그니처 및 각각의 교란 시그니처가 비교된다.

[0404] 일부 실시예에서, 단일-세포 전이 시그니처 및 교란 시그니처를 비교하여 각각의 교란 시그니처의 각각의 수치 활성화 점수를 결정하는 것은 단일-세포 전이 시그니처의 기준 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 단일-세포 전이 시그니처의 각각의 세포 구성성분의 제1 유의성 점수를 각각의 교란 시그니처의 상응하는 세포 구성성분의 상응하는 유의성 점수와 비교하는 것을 포함한다.

[0405] 일부 실시예에서, 단일-세포 전이 시그니처 및 교란 시그니처를 비교하여 각각의 교란 시그니처의 각각의 수치 활성화 점수를 결정하는 것은, 단일-세포 전이 시그니처의 기준 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 단일-세포 전이 시그니처 내의 복수의 기준 세포 구성성분의 각각의 개별 세포 구성성분의 유의성

점수를 각각의 교란 시그니처 내의 복수의 세포 구성성분의 각각의 상응하는 세포 구성성분의 상응하는 유의성 점수와 비교하는 것을 포함한다.

[0406] 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 교란 시그니처의 세트의 다른 교란 시그니처에 대한 각각의 교란 시그니처의 단일-세포 전이 시그니처에 대한 관련성의 상대 순위이다. 일부 실시예에서, 상대 순위는 윌콕슨 순위-합계 검정, t-검정, 로지스틱 회귀 또는 일반화된 선형 모델에 의해 결정된다. 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 순위에 기초하지 않는다.

[0407] 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 각각의 교란 시그니처에 대한 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대한 상응하는 유의성 점수의 중심 집중 경향의 척도이다. 일부 실시예에서, 중심 집중 경향의 척도는 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대한 산술 평균, 가중 평균, 중간범위, 미드힌지, 삼평균, 윈저화 평균, 평균, 또는 상응하는 유의성 점수의 모드이다.

[0408] 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 (i) 각각의 교란 시그니처에 대한, 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대한 상응하는 유의성 점수의 중심 집중 경향의 척도와 (ii) 단일-세포 전이 시그니처에 대한, 복수의 기준 세포 구성성분의 각각의 개별 세포 구성성분에 대한 상응하는 제1 유의성 점수의 중심 집중 경향의 척도 사이의 차이이다.

[0409] 일부 실시예에서, 단일-세포 전이 시그니처의 변경되지 않은 세포 상태는 각각의 교란 시그니처의 제1 세포 상태 또는 제2 세포 상태와 동일하다. 일부 실시예에서, 단일-세포 전이 시그니처의 변경되지 않은 세포 상태는 각각의 교란 시그니처의 제1 세포 상태 및 제2 세포 상태 둘 다와 상이하다.

[0410] 일부 실시예에서, 방법은 단일-세포 전이 시그니처의 기준 복수의 세포 구성성분 및 각각의 교란 시그니처의 각각의 복수의 세포 구성성분을 프루닝하여 전사 인자에 대한 비교를 제한하는 것을 더 포함한다. 일부 실시예에서, 방법은 또 다른 세포 구성성분 유형(예를 들어, 유전자, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질, 및/또는 그의 조합)과의 비교를 제한하기 위해 단일-세포 전이 시그니처의 기준 복수의 세포 구성성분 및 각각의 교란 시그니처의 각각의 복수의 세포 구성성분을 프루닝하는 것을 더 포함한다. 일부 실시예에서, 기준 복수의 세포 구성성분 및 각각의 복수의 세포 구성성분은 프루닝되지 않는다.

[0411] 일부 실시예에서, 복수의 교란 시그니처에서 각각의 교란 시그니처의 교란된 상태는 복수의 화합물 중의 화합물에 노출되지 않은 대조군 세포에 의해 나타난다. 일부 실시예에서, 복수의 교란 시그니처에서 각각의 교란 시그니처의 교란된 상태는 각각의 교란 시그니처와 연관된 화합물 이외의 복수의 화학적 화합물의 화학적 화합물에 노출된 관련되지 않은 교란된 세포에 걸친 평균에 의해 나타내어진다.

[0412] 위에서 설명 바와 같이, 일부 실시예에서, 교란 시그니처의 세트의 각각의 교란 시그니처는 비제한적인 예로서 본원에서 참조로 포함된, 2019년 7월 15일 출원된 발명의 명칭이 "세포 분석 방법"인 미국 특허 출원 번호 16/511,691에 개시된 임의의 방법을 사용하여 결정될 수 있다.

[0413] 각각의 교란 시그니처는 각각의 복수의 세포 구성성분의 식별, 및 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함한다. 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 다른 하나는 각각의 교란 시그니처에 상응하는 화합물에 대한 세포의 노출에 의해 유발된 각각의 교란된 세포 상태이다. 또한, 위에서 설명 바와 같이, 각각의 교란 시그니처는 수치 활성화 점수를 포함한다. 일부 실시예에서, 각각의 교란 시그니처에 대한 수치 활성화 점수는 연속 척도 상의 절대값이다. 일부 실시예에서, 각각의 교란 시그니처에 대한 수치 활성화 점수는 하기에 보다 상세히 논의된 바와 같이 상대적 순위화이다.

[0414] 일부 실시예에서, 교란 시그니처의 세트의 각각의 교란 시그니처의 각각의 수치 활성화 점수는 변경되지 않은 세포 상태와 변경된 세포 상태 사이의 차등 세포 구성성분 풍부도의 척도를 나타내는 단일-세포 전이 시그니처에 전자 형태로 액세스하는 것을 포함하는 절차에 의해 획득된다. 여기서, 변경된 세포 상태는 변경되지 않은 세포 상태로부터 변경된 세포 상태로의 세포 전이를 통해 발생한다. 또한, (i) 변경되지 않은 세포 상태, (ii) 변경된 세포 상태, 및 (iii) 변경되지 않은 세포 상태로부터 변경된 세포 상태로의 전이 중 적어도 하나는 관심 생리학적 조건과 연관된다.

[0415] 단일-세포 전이 시그니처는 기준 복수의 세포 구성성분의 식별, 및 복수의 기준 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 변경되지 않은 세포 상태와 변경된 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 제1 유의성 점수를 포함한다. 일부

실시예에서, 단일-세포 전이 시그니처는 2019년 7월 15일에 출원된 발명의 명칭 "세포 분석 방법"의 미국 특허 출원 번호 16/511,691(본원에 참조로 포함됨)에 개시된 임의의 방법을 사용하여 결정된다.

[0416] 일단 획득되면, 단일-세포 전이 시그니처를 각각의 교란 시그니처와 비교하여 각각의 교란 시그니처의 각각의 수치 활성화 점수를 결정한다. 일부 실시예에서, 단일-세포 전이 시그니처를 각각의 교란 시그니처와 비교하여, 2019년 7월 15일에 출원된 발명의 명칭이 "세포 분석 방법"인 미국 특허 출원 번호 16/511,691에 개시된 복수의 교란 시그니처에서의 다른 교란 시그니처에 대한 각각의 교란 시그니처의 상대 순위를 결정하는 방법 중 임의의 것이 사용될 수 있으며, 여기서 예를 들어 이러한 상대 순위는 이어서 각각의 교란 시그니처의 각각의 수치 활성화 점수로 간주될 것이다.

[0417] 일부 실시예에서, 각각의 교란 시그니처의 각각의 수치 활성화 점수를 결정하기 위한 단일-세포 전이 시그니처 및 교란 시그니처의 비교는 단일-세포 전이 시그니처의 기준 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 제1 유의성 점수를 각각의 교란 시그니처의 상응하는 세포 구성성분의 상응하는 유의성 점수와 비교하는 것을 포함한다. 일부 이러한 실시예에서, 각각의 교란 시그니처의 활성화 점수는 교란 시그니처의 세트의 다른 교란 시그니처에 대한 각각의 교란 시그니처의 단일-세포 전이 시그니처에 대한 관련성의 상대 순위이다. 일부 이러한 실시예에서, 상대 순위는 윌콕슨 순위-합계 검정, t-검정, 로지스틱 회귀 또는 일반화된 선형 모델에 의해 결정된다. 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 각각의 교란 시그니처의 관련성의 상대 순위가 아니라, 오히려 단일-세포 전이 시그니처에 대한 다른 교란 시그니처의 순위와 독립적으로 결정된다.

[0418] 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 순위에 기초하지 않는다. 예를 들어, 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 각각의 교란 시그니처에 대한 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대한 상응하는 유의성 점수를 포함하는 복수의 유의성 점수이다.

[0419] 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 각각의 교란 시그니처에 대한 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대한 상응하는 유의성 점수의 중심 집중 경향의 척도이다. 일부 실시예에서, 중심 집중 경향의 척도는 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대한 산술 평균, 가중 평균, 중간범위, 미드힌지, 삼평균, 윈저화 평균, 평균, 또는 상응하는 유의성 점수의 모드이다.

[0420] 일부 실시예에서, 각각의 교란 시그니처의 활성화 점수는 (i) 각각의 교란 시그니처에 대한, 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대한 상응하는 유의성 점수의 중심 집중 경향의 척도와 (ii) 단일-세포 전이 시그니처에 대한, 복수의 기준 세포 구성성분의 각각의 개별 세포 구성성분에 대한 상응하는 제1 유의성 점수의 중심 집중 경향의 척도 사이의 차이이다.

[0421] 한 실시예에서, 단일-세포 전이 시그니처와 각각의 교란 시그니처 사이의 비교를 수행하기 위해, 교란 시그니처의 세포 구성성분을 행렬로서 나타낸다. 행렬의 각각의 행은 단일 교란(예를 들어, 복수의 화합물 중 단일 화합물)과 연관된다. 행렬 상의 각각의 열은 각각의 상태 사이에 차등적 풍부도를 나타내는 세포 구성성분 중 하나와 연관된다. 행렬 내의 각각의 엔트리는 특정 교란 시그니처에 대해 식별된 세포 구성성분에 대한 유의성 점수(예를 들어, p-값, t-점수)를 포함한다. 이 행렬은 단일-세포 전이 시그니처에 있는 세포 구성만을 포함하도록 필터링된다. 이러한 필터링은 임계치 p-값, 임계 수의 세포-컴포넌트의 사용 등을 사용하여 달성될 수 있다.

[0422] 행렬 내의 각각의 유의성 점수가 이산 매칭 점수로 대체된다. 각각의 유의성 점수를 이산 매칭 점수로 대체하기 위해, 세포 전이에 대해 유의하게 상향조절된 세포 구성성분 및 세포 전이에 대해 유의하게 하향조절된 세포 구성성분을 식별한다. 단일-세포 전이 시그니처에 의해 식별된 유의하게 상향조절된 세포 구성성분 각각에 대해, 세포 구성성분이 또한 그 교란(예를 들어, 화학적 조성)에 대한 교란 시그니처에 대해 유의하게 상향조절되는 경우에, 그 세포 구성성분/교란 조합에 대한 행렬에서의 유의성 점수는 "1"의 이산 매칭 점수로 대체된다. 세포 구성성분이 단일-세포 전이 시그니처에 비해 교란 시그니처에 대해 유의하게 하향조절되는 경우에, 그 세포 구성성분/교란 조합에 대한 행렬에서의 유의성 점수는 "-2"의 이산 매칭 점수로 대체된다. 세포 구성성분이 교란 시그니처에 대해 유의하게 상향조절 또는 하향조절되지 않는 경우에, 세포 구성성분/교란 조합에 대한 행렬에서의 유의성 점수는 "0"의 이산 매칭 점수로 대체된다.

[0423] 반대로, 단일-세포 전이 시그니처에서 식별된 유의하게 하향조절된 세포 구성성분 각각에 대해, 세포-컴포넌트가 또한 교란에 대해 유의하게 하향조절되는 경우, 이러한 세포 구성성분/교란 조합에 대한 행렬에서의 유의성 점수가 "-1"의 이산 매칭 점수로 대체된다. 세포 구성성분이 교란에 대해 유의하게 상향조절되는 경우에, 그

세포 구성성분/교란 조합에 대한 행렬에서의 유의성 점수는 "2"의 이산 매칭 점수로 대체된다. 세포 구성성분이 교란에 대해 유의하게 상향조절 또는 하향조절되지 않는 경우에, 그 세포 구성성분/교란 조합에 대한 행렬에서의 유의성 점수는 "0"의 이산 매칭 점수로 대체된다. 관련 기술분야의 통상의 기술자는 이들 특정한 점수 대체가 일부 실시예에서 다른 수치값으로 대체될 수 있다는 것을 인지할 것이다. 또한, 상향조절 또는 하향조절보다는, 각각의 세포 구성성분에 대한 임계치 풍부도 값의 사용이 사용될 수 있고, 이때, 상기 언급된 클래스 표지(예를 들어, "-1", "2", "0" 등)를 행렬의 각각의 요소에 할당하는데 있어서 주어진 세포 구성성분이 임계치 풍부도 값 초과 또는 미만인지 여부의 고려가 이루어진다.

[0424] 결과는 교란의 수(복수의 화학적 조성물 중 화학적 조성의 수 및 따라서 복수의 교란 시그니처 중 교란 시그니처의 수)에 의해 주어진 행의 수 및 앞서 설명된 매칭 점수를 나타내는 행렬 요소 엔트리와 단일-세포 전이로부터의 차등 세포 구성성분에 의해 주어진 열의 수를 갖는 행렬이다.

[0425] 행렬 내의 유의성 점수를 앞서 설명된 바와 같은 이산 매칭 점수로 대체한 후에, 행렬의 각각의 행 내의 이산 매칭 점수를 합산하여 각각의 행에 대한 합산된 매칭 점수를 생성한다. 이어서, 각각 교란 시그니처에 상응하는 행렬의 행을 합산 매칭 점수가 감소하는 순서로 순위화한다. 상위 순위 행은 단일-세포 전이 시그니처의 식별된 세포 전이와 연관될 가능성이 가장 큰 교란 시그니처와 연관된다. 또한, 각각의 행의 순위는 각각의 행에 상응하는 교란 시그니처에 대한 활성화 점수로서 사용될 수 있다.

[0426] 일부 실시예에서, 행렬 내의 각각의 행의 합산된 매칭 점수에 대해, 거짓 세포-컴포넌트 발견율의 추정은 2019년 7월 15일에 출원된 발명의 명칭이 "세포 분석 방법"인 미국 특허 출원 번호 16/511,691(본원에 참조로 포함됨)에 논의된 바와 같이 추정된다.

[0427] 특정 실시예에서, 교란의 공변량(예를 들어, 세포의 특정한 화학적 조성물에 대한 노출)이 존재할 수 있다. 예를 들어, 화학적 조성물의 공변량은 화학적 조성물의 특정 용량, 화학적 조성물에 노출된 세포를 측정하여 세포 구성성분을 정량화하는 시간, 및/또는 화학적 조성물에 노출된 세포의 아이덴티티(예를 들어, 세포주)를 포함할 수 있다. 일부 실시예에서, 교란(예를 들어, 세포의 특정한 화학적 조성물에 대한 노출)은 그의 공변량의 임계치 양이 또한 특정한 세포 전이에 영향을 미칠 것으로 예측되는 경우에만 특정한 세포 전이에 영향을 미칠 것으로 예측된다. 다시 말해서, 일부 실시예에서, 특정한 교란 시그니처의 수치 활성화 점수는 적어도 부분적으로, 특정한 교란 시그니처의 화학적 조성의 공변량이 또한 단일-세포 전이 점수와 연관된 특정한 세포 전이에 영향을 미칠 것으로 예측되는지 여부에 의해 결정된다.

[0428] 각각의 교란 시그니처를 단일 세포-전이 시그니처와 비교하는 대안적 방법을 사용하여 각각의 교란 시그니처의 수치 활성화 점수를 결정할 수 있다. 예를 들어, 세포 구성성분은 웹 인터페이스(예를 들어, 예컨대 L1000CDS2. An ultra-fast LINCS L1000 Characteristic Direction Signature Search Engine, on world wide web at amp.pharm.mssm.edu/L1000CDS2/#/index)를 사용하여 데이터베이스에 매칭될 수 있다.

[0429] 일부 실시예에서, 단일-세포 전이 시그니처의 변경되지 않은 세포 상태는 각각의 교란 시그니처의 제1 세포 상태 또는 제2 세포 상태와 동일하다. 일부 실시예에서, 단일-세포 전이 시그니처의 변경되지 않은 세포 상태는 각각의 교란 시그니처의 제1 세포 상태 및 제2 세포 상태 둘 다와 상이하다.

[0430] 일부 실시예에서, 방법은 단일-세포 전이 시그니처의 기준 복수의 세포 구성성분 및 각각의 교란 시그니처의 각각의 복수의 세포 구성성분을 프루닝하여 전사 인자에 대한 비교를 제한하는 것을 더 포함한다. 일부 실시예에서, 복수의 교란 시그니처에서 각각의 교란 시그니처의 교란된 상태는 복수의 화합물 중의 화합물에 노출되지 않은 대조군 세포에 의해 나타난다.

[0431] 일부 실시예에서, 복수의 교란 시그니처에서 각각의 교란 시그니처의 교란된 상태는 각각의 교란 시그니처와 연관된 화합물 이외의 복수의 화학적 화합물의 화학적 화합물에 노출된 관련되지 않은 교란된 세포에 걸친 평균에 의해 나타내어진다.

[0432] 블록 908을 참조하면, 방법은 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해, (i) 각각의 화합물의 화학 구조의 지문을 훈련되지 않은 모델로 입력 시 각각의 교란 시그니처에 대한 각각의 계산된 활성화 점수 및 (ii) 교란 시그니처의 세트에서 상응하는 화합물에 대한 각각의 교란 시그니처의 각각의 수치 활성화 점수 사이의 각각의 차이를 이용하여 훈련되지 않은 모델을 훈련시키는 것을 더 포함한다.

[0433] 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 모델의 임의의 적합한 실시예, 예컨대 "모델 아키텍처"라는 명칭의 상기 섹션에 개시된 것, 및 그의 임의의 치환, 변형, 추가, 결실, 및/또는 조합이 고려된다. 예를

들어, 일부 실시예에서, 훈련된 모델은 신경망을 포함한다. 일부 실시예에서, 신경망은 ReLU 활성화를 갖는 완전 연결 신경망이다. 일부 실시예에서, 신경망은 메시지 전달 신경망이다. 일부 실시예에서, 훈련된 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다.

[0434] 일부 실시예에서, 훈련된 모델은 복수의 컴포넌트 모델의 앙상블 모델이고, 각각의 계산된 활성화 점수는 복수의 컴포넌트 모델에서 각각의 컴포넌트 모델의 출력의 중심 집중 경향의 척도이다. 일부 실시예에서, 복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신 모델, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 또는 선형 회귀 모델을 포함한다. 일부 실시예에서, 복수의 컴포넌트 모델은 복수의 신경망을 포함한다. 일부 실시예에서, 복수의 신경망 내의 제1 신경망은 ReLU 활성화를 갖는 완전 연결 신경망이고, 복수의 신경망 내의 제2 신경망은 메시지 전달 신경망이다.

[0435] 블록 910을 참조하면, 훈련은 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고(복수의 파라미터는 100개 이상의 파라미터를 포함함), 이에 의해 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득한다.

[0436] 관련 기술분야의 통상의 기술자가 명백히 알 수 있는 바와 같이, 훈련되지 않은 또는 부분적으로 훈련된 모델을 훈련시키기 위한 임의의 적합한 방법 및 실시예, 예컨대 "모델 훈련"이라는 명칭의 상기 섹션에 개시된 것(그 임의의 치환, 변형, 추가, 결실 및/또는 조합 포함)이 고려된다.

[0437] 일부 실시예에서, 모델에 대한 입력은 복수의 활성화 점수를 포함하며, 각각의 개별 활성화 점수는 복수의 화합물의 각각의 화합물에 대한 교란 시그니처의 세트에서의 각각의 교란 시그니처에 상응한다. 각각의 개별 화합물에 대한 각각의 개별 교란 시그니처에 상응하는 활성화 점수는 교란 시그니처와 화합물 사이의 연관성(예를 들어, 가중치 및/또는 상관)을 식별하기 위해 멀티-태스크 모델을 훈련시키기 위한 표지(예를 들어, 교란 시그니처와 화합물 사이의 연관성의 실제 존재 또는 부재를 나타내는 수치 활성화 점수)로서 기능한다. 예를 들어, 앞서 설명된 바와 같이, 일부 실시예에서, 복수의 교란 시그니처의 제1 서브세트는 관심 생리학적 조건과 연관되고, 복수의 교란 시그니처의 제2 서브세트는 관심 생리학적 조건과 연관되지 않는다. 따라서, 일부 이러한 실시예에서, 복수의 교란 시그니처의 제1 서브세트를 표지로서 사용하여 훈련 데이터세트에 연관성의 실제 존재가 포함될 수 있고, 복수의 교란 시그니처의 제2 서브세트를 표지로서 사용하여 훈련 데이터세트에 연관성의 실제 부재가 포함될 수 있다.

[0438] 일부 실시예에서, 훈련은 회귀 모델에 따라 교란 시그니처의 세트 s에서의 각각의 개별 교란 시그니처에 대해 각각의 상응하는 화합물과 연관된 각각의 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정한다. 일부 실시예에서, 회귀 모델은 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해 각각의 상응하는 화합물과 연관된 각각의 차이의 최소 제곱 오차를 최적화한다.

[0439] 일부 실시예에서, 모델은 세포 구성성분 모듈, 교란 시그니처, 또는 둘 다에 대한 활성화 점수에 기초하여 관심 생리학적 조건과 화합물을 연관시키기 위해 훈련되고/거나 사용된다. 일부 실시예에서, 모델은 복수의 도메인(예를 들어, 표지 유형, 예컨대 모듈 및/또는 교란 시그니처) 및/또는 데이터 유형(예를 들어, 분석물 및/또는 세포 구성성분, 예컨대 유전자 발현 프로파일, 대사체학, 단백질체학, 후생학 등)에 대한 활성화 점수에 기초하여 관심 생리학적 조건과 화합물을 연관시키기 위해 훈련되고/거나 사용된다. 일부 실시예에서, 모델은 임의의 하나 이상의 관심 생리학적 조건(예를 들어, 화합물의 독성, 질환 상태의 해소 등)에 대한 활성화 점수에 기초하여 화합물을 관심 생리학적 조건과 연관시키기 위해 훈련되고/거나 사용된다. 일부 실시예에서, 모델은 복수의 시스템에 걸쳐 훈련되며, 여기서 시스템은 임의의 하나 이상의 생리학적 조건, 임의의 하나 이상의 도메인, 및/또는 본원에 개시된 임의의 하나 이상의 데이터 유형, 또는 관련 기술분야의 통상의 기술자에게 명백할 임의의 치환, 변형, 추가, 결실, 및/또는 조합을 지칭한다. 예를 들어, 일부 실시예에서, 모델은 테스트 화학적 화합물, 독성의 유전자 모듈 특징의 활성화, 및 질환 해소를 나타내는 교란 시그니처 사이의 연관성을 집합적으로 결정하도록 공동으로 훈련된다.

[0440] *추가의 실시예.*

[0441] 본 개시의 또 다른 양태는 하나 이상의 프로세서 및 메모리를 포함하는 컴퓨터 시스템을 제공하며, 메모리는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 수행하기 위한 명령어를 저장한다. 이 방법은

테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계, 및 지문을 모델에 입력하는 단계를 포함하며, 여기서 모델은 100개 이상의 파라미터를 포함하고, 모델은 모델로의 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력하고, 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 세포 구성성분 모듈의 세트의 상응하는 세포 구성성분 모듈을 나타내고, 세포 구성성분 모듈의 세트에서의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함하고, 세포 구성성분 모듈의 세트의 제1 세포 구성성분 모듈은 관심 생리학적 조건과 연관된다. 방법은 제1 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족하는 경우에 관심 생리학적 조건을 갖는 화학적 화합물을 식별하는 것을 더 포함한다.

[0442] 본 개시의 또 다른 양태는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한, 컴퓨터에 의해 실행가능한 하나 이상의 컴퓨터 프로그램을 저장하는 비-일시적 컴퓨터-판독가능 매체를 제공하며, 여기서 컴퓨터는 하나 이상의 프로세서 및 메모리를 포함하고, 하나 이상의 컴퓨터 프로그램은 방법을 수행하는 컴퓨터 실행가능 명령어를 집합적으로 인코딩한다. 이 방법은 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 단계, 및 지문을 모델에 입력하는 단계를 포함하며, 여기서 모델은 100개 이상의 파라미터를 포함하고, 모델은 모델로의 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력하고, 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 세포 구성성분 모듈의 세트의 상응하는 세포 구성성분 모듈을 나타내고, 세포 구성성분 모듈의 세트에서의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함하고, 세포 구성성분 모듈의 세트의 제1 세포 구성성분 모듈은 관심 생리학적 조건과 연관된다. 방법은 제1 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족하는 경우에 관심 생리학적 조건을 갖는 화학적 화합물을 식별하는 것을 더 포함한다.

[0443] 본 개시의 또 다른 양태는 하나 이상의 프로세서 및 메모리를 포함하는 컴퓨터 시스템을 제공하며, 메모리는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 수행하기 위한 명령어를 저장한다. 방법은 테스트 화학적 화합물의 화학 구조의 지문을 얻고, 지문을 모델에 입력하는 것을 포함하며, 여기서 모델은 100개 이상의 파라미터를 포함한다. 모델은 모델로의 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력한다. 하나 이상의 계산된 활성화 점수에서 각각의 개별 계산된 활성화 점수는 교란 시그니처의 세트에서 상응하는 교란 시그니처를 나타낸다. 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별을 포함하고, 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하며, 여기서 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태이다. 방법은 교란 시그니처의 세트의 제1 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족할 때 관심 생리학적 조건을 갖는 화학적 화합물을 식별하는 것을 더 포함한다.

[0444] 본 개시의 또 다른 양태는 테스트 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한, 컴퓨터에 의해 실행가능한 하나 이상의 컴퓨터 프로그램을 저장하는 비-일시적 컴퓨터-판독가능 매체를 제공하며, 여기서 컴퓨터는 하나 이상의 프로세서 및 메모리를 포함하고, 하나 이상의 컴퓨터 프로그램은 방법을 수행하는 컴퓨터 실행가능 명령어를 집합적으로 인코딩한다. 방법은 테스트 화학적 화합물의 화학 구조의 지문을 얻고, 지문을 모델에 입력하는 것을 포함하며, 여기서 모델은 100개 이상의 파라미터를 포함한다. 모델은 모델로의 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력한다. 하나 이상의 계산된 활성화 점수에서 각각의 개별 계산된 활성화 점수는 교란 시그니처의 세트에서 상응하는 교란 시그니처를 나타낸다. 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별을 포함하고, 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하며, 여기서 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태이다. 방법은 교란 시그니처의 세트의 제1 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족할 때 관심 생리학적 조건을 갖는 화학적 화합물을 식별하는 것을 더 포함한다.

[0445] 본 개시의 또 다른 양태는 하나 이상의 프로세서 및 메모리를 포함하며, 메모리가 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법을 수행하기 위한 명령어를 저장하는 것인 컴퓨터 시스템을 제공한다. 방법은, 전자 형태로, 복수의 화합물에서의 각각의 화합물의 화학 구조의 각각의 지문을 획득함으로써 복수의 지문을 획득하

는 것을 포함한다. 방법은 복수의 화합물의 각각의 화합물에 대한 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수를 전자 형태로 획득하는 것을 포함하며, 여기서 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함한다. 방법은 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해, (i) 훈련되지 않은 모델로의 각각의 화합물의 화학 구조의 지문의 입력 시 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수와 (ii) 세포 구성성분 모듈의 세트의 각각의 화합물에 대한 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련되지 않은 모델을 훈련시키는 것을 더 포함한다. 훈련은 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고(복수의 파라미터는 100개 이상의 파라미터를 포함함), 이에 의해 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득한다.

[0446] 본 개시의 또 다른 양태는 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한, 컴퓨터에 의해 실행가능한 하나 이상의 컴퓨터 프로그램을 저장하는 비-일시적 컴퓨터-판독가능 매체를 제공하며, 컴퓨터는 하나 이상의 프로세서 및 메모리를 포함하고, 하나 이상의 컴퓨터 프로그램은 방법을 수행하는 컴퓨터 실행가능 명령어를 집합적으로 인코딩한다. 방법은, 전자 형태로, 복수의 화합물에서의 각각의 화합물의 화학 구조의 각각의 지문을 획득함으로써 복수의 지문을 획득하는 것을 포함한다. 방법은 복수의 화합물의 각각의 화합물에 대한 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수를 전자 형태로 획득하는 것을 더 포함하고, 여기서 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함한다. 방법은 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해, (i) 훈련되지 않은 모델로의 각각의 화합물의 화학 구조의 지문의 입력 시 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수와 (ii) 세포 구성성분 모듈의 세트의 각각의 화합물에 대한 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련되지 않은 모델을 훈련시키는 것을 더 포함한다. 훈련은 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고(복수의 파라미터는 100개 이상의 파라미터를 포함함), 이에 의해 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득한다.

[0447] 본 개시의 또 다른 양태는 하나 이상의 프로세서 및 메모리를 포함하는 컴퓨터 시스템을 제공하며, 메모리는 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한 명령어를 저장하고, 방법은 전자 형태로 복수의 화합물에서의 각각의 화합물의 화학 구조의 각각의 지문을 획득함으로써 복수의 지문을 획득하는 것을 포함한다. 방법은 복수의 화합물의 각각의 상응하는 화합물에 대한 교란 시그니처의 세트의 각각의 개별 교란 시그니처의 각각의 수치 활성화 점수를 전자 형태로 획득하는 것을 더 포함한다. 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별을 포함하고, 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수를 포함하며, 여기서 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태이다. 방법은 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해, (i) 각각의 화합물의 화학 구조의 지문을 훈련되지 않은 모델로 입력 시 각각의 교란 시그니처에 대한 각각의 계산된 활성화 점수 및 (ii) 교란 시그니처의 세트에서 상응하는 화합물에 대한 각각의 교란 시그니처의 각각의 수치 활성화 점수 사이의 각각의 차이를 이용하여 훈련되지 않은 모델을 훈련시키는 것을 포함한다. 훈련은 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고(복수의 파라미터는 100개 이상의 파라미터를 포함함), 이에 의해 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득한다.

[0448] 본 개시의 또 다른 양태는 화학적 화합물을 관심 생리학적 조건과 연관시키기 위한, 컴퓨터에 의해 실행가능한 하나 이상의 컴퓨터 프로그램을 저장하는 비-일시적 컴퓨터-판독가능 매체를 제공하며, 여기서 컴퓨터는 하나 이상의 프로세서 및 메모리를 포함하고, 하나 이상의 컴퓨터 프로그램은 전자 형태로, 복수의 화합물의 각각의 화합물의 화학 구조의 각각의 지문을 획득함으로써 복수의 지문을 획득하는 것을 포함하는 방법을 수행하는 컴퓨터 실행가능 명령어를 집합적으로 인코딩한다. 방법은 복수의 화합물의 각각의 상응하는 화합물에 대한 교란 시그니처의 세트의 각각의 개별 교란 시그니처의 각각의 수치 활성화 점수를 전자 형태로 획득하는 것을 더 포함한다. 교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별을 포함하고, 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하

는 유의성 점수를 포함하며, 여기서 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태이다. 방법은 복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 교란 시그니처의 세트의 각각의 개별 교란 시그니처에 대해, (i) 각각의 화합물의 화학 구조의 지문을 훈련되지 않은 모델로 입력 시 각각의 교란 시그니처에 대한 각각의 계산된 활성화 점수 및 (ii) 교란 시그니처의 세트에서 상응하는 화합물에 대한 각각의 교란 시그니처의 각각의 수치 활성화 점수 사이의 각각의 차이를 이용하여 훈련되지 않은 모델을 훈련시키는 것을 더 포함한다. 훈련은 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고(복수의 파라미터는 100개 이상의 파라미터를 포함함), 이에 의해 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득한다.

[0449] 본 개시의 또 다른 양태는 하나 이상의 프로세서, 및 하나 이상의 프로세서에 의한 실행을 위한 하나 이상의 프로그램을 저장하는 메모리를 갖는 컴퓨터 시스템을 제공하며, 하나 이상의 프로그램은 본원에 개시된 임의의 방법 및/또는 실시예를 수행하기 위한 명령어를 포함한다. 일부 실시예에서, 본원에 개시된 방법 및/또는 실시예 중 임의의 것은 하나 이상의 프로세서, 및 하나 이상의 프로세서에 의한 실행을 위한 하나 이상의 프로그램을 저장하는 메모리를 갖는 컴퓨터 시스템에서 수행된다.

[0450] 본 개시의 또 다른 양태는 컴퓨터에 의해 실행되도록 구성된 하나 이상의 프로그램을 저장하는 비-일시적 컴퓨터 판독가능 저장 매체를 제공하며, 이 하나 이상의 프로그램은 본원에 개시된 임의의 방법을 수행하기 위한 명령어를 포함한다.

[0451] **IV. 세포 구성성분 모듈의 식별**

[0452] 일부 실시예에서, 관심 생리학적 조건과 연관된 세포 구성성분 모듈(132)이 식별된다. 이러한 방법은 도 2 및 14와 함께 여기서 논의된다. 특히, 도 14a의 블록 1500을 참조하면, 일부 실시예에서, 방법은 관심 생리학적 조건과 연관된 제1 세포 구성성분 모듈(132)을 식별하는 것을 더 포함한다.

[0453] 본 개시의 일부 실시예에 따른, 세포 구성성분을 관심 생리학적 조건과 연관시키기 위한 방법(200)의 예시적인 작업흐름이 도 2a 및 도 2b를 참조로 제공된다.

[0454] 도 2a의 블록 202 및 도 14a의 블록 1502를 참조하면, 방법은 전자 형태로 하나 이상의 제1 데이터세트를 획득하는 단계를 포함한다. 도 14b의 블록 1504를 참조하면, 하나 이상의 제1 데이터세트는 제1 복수의 세포의 각각의 개별 세포에 대해, 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함한다. 이러한 방식으로, 복수의 벡터가 획득된다.

[0455] 일부 실시예에서, 관심 생리학적 조건은 질환이고, 제1 복수의 세포는 복수의 주석화된 세포 상태에 의해 입증된 바와 같이 질환을 대표하는 세포 및 질환을 대표하지 않는 세포를 포함한다.

[0456] 일부 실시예에서, 도 3a의 블록 300의 관심 생리학적 조건은 질환과 연관된 이상 세포 과정이고, 제1 복수의 세포는 복수의 주석화된 세포 상태에 의해 입증된 바와 같이 질환을 나타내는 세포 및 질환을 나타내지 않는 세포를 포함한다.

[0457] 일부 실시예에서, 도 3a의 블록 300의 관심 생리학적 조건은 질환과 연관된 이상 세포 과정이고, 제1 복수의 세포는 복수의 주석화된 세포 상태에 의해 입증된 바와 같이 질환 상태를 나타내는 세포 및 건강한 또는 대조군 상태를 나타내는 세포를 포함한다.

[0458] 일부 실시예에서, 도 3a의 블록 300의 관심 생리학적 조건은 복수의 질환과 연관된 이상 세포 과정이고, 제1 복수의 세포는 복수의 서브세트의 세포를 포함하고, 각각의 개별 서브세트의 세포는 복수의 주석화된 세포 상태에 의해 입증된 바와 같이 복수의 질환에서 각각의 질환을 나타낸다.

[0459] 도 14b의 블록 1506을 참조하면, 일부 실시예에서, 제1 복수의 세포는 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 100, 200, 또는 1000 또는 그 이상의 세포를 포함하고, 집합적으로 복수개(예를 들어, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 100, 200, 또는 1000개)의 주석화된 세포 상태를 나타낸다.

[0460] 도 14b의 블록 1508을 참조하면, 일부 실시예에서, 복수의 세포 구성성분은 2개, 3개, 4개, 5개, 6개, 7개, 8개, 9개, 10개, 15개, 20개, 25개, 30개, 35개, 50개, 100개, 500개, 1000개, 5000개, 10,000개 또는 그 이상의 세포 구성성분을 포함한다. 일부 실시예에서, 복수의 세포 구성성분은 2개 내지 10,000개의 세포 구성성분

으로 이루어진다. 일부 실시예에서, 복수의 세포 구성성분은 100개 내지 10,000개의 세포 구성성분으로 이루어진다.

[0461] 도 2a의 블록 204를 참조하면, 방법은 복수의 벡터에 액세스하거나 복수의 벡터를 형성하는 것을 포함한다. 도 14a의 블록 1510을 참조하면, 복수의 벡터의 각각의 개별 벡터는 (i) 복수의 구성성분의 각각의 세포 구성성분에 상응하고, (ii) 상응하는 복수의 요소를 포함한다. 도 14a의 블록 1512를 참조하면, 상응하는 복수의 요소의 각각의 개별 요소는 제1 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 갖는다.

[0462] 블록 206을 참조하면, 복수의 벡터가 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는데 사용된다. 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 복수의 세포 구성성분의 서브세트를 포함한다. 복수의 세포 구성성분 모듈은 (i) 복수의 후보 세포 구성성분 모듈 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현으로 배열되고, 이때 복수의 세포 구성성분 모듈은 10개를 초과하는 세포 구성성분 모듈을 포함한다.

[0463] 도 14b의 블록 1514를 참조하면, 일부 실시예에서 복수의 주석화된 세포 상태의 주석화된 세포 상태는 노출 조건(예를 들어, 노출 지속기간, 화합물의 농도, 또는 노출 지속기간과 화합물의 농도의 조합) 하에서의 제1 복수의 세포의 세포 화합물에 대한 노출이다.

[0464] 도 14b의 블록 1518을 참조하면, 일부 실시예에서, 복수의 세포 구성성분의 각각의 세포 구성성분은 특정 유전자, 유전자와 연관된 특정 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질, 또는 이들의 조합이다.

[0465] 도 14b의 블록 1520을 참조하면, 일부 실시예에서, 제1 또는 제2 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도는 비색 측정치, 형광 측정치, 발광 측정치, 또는 공명 에너지 전달(FRET) 측정치에 의해 결정된다.

[0466] 도 14b의 블록 1522를 참조하면, 일부 실시예에서, 제1 또는 제2 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도는 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq), scTag-seq, 시퀀싱을 사용한 전위효소-접근 가능 염색질에 대한 단일-세포 검정(scATAC-seq), CyTOF/SCoP, E-MS/Abseq, miRNA-seq, CITE-seq, 또는 그 임의의 조합에 의해 결정된다.

[0467] 도 14b의 블록 1524를 참조하면, 일부 실시예에서, 관심 생리학적 조건은 질환이고, 제1 복수의 세포는 복수의 주석화된 세포 상태에 의해 입증된 바와 같이 질환을 대표하는 세포 및 질환을 대표하지 않는 세포를 포함한다.

[0468] 도 14b의 블록 1526을 참조하면, 복수의 벡터가 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는데 사용되고, 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 복수의 세포 구성성분의 서브세트를 포함한다. 복수의 세포 구성성분 모듈은 (i) 복수의 후보 세포 구성성분 모듈 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현으로 배열되고, 이때 복수의 세포 구성성분 모듈은 10개를 초과하는 세포 구성성분 모듈을 포함한다.

[0469] 도 14c의 블록 1528을 참조하면, 일부 실시예에서, 복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계는 복수의 벡터의 각각의 벡터의 각각의 상응하는 복수의 요소를 사용하여 상관 모델을 복수의 벡터에 적용하는 것을 포함한다. 일부 실시예에서, 상관 모델은 그래프 클러스터링 알고리즘(예를 들어, 그래프 클러스터링 방법은 피어슨-상관관계-기반 거리 메트릭에 대한 라이덴 클러스터링이고, 그래프 클러스터링 방법은 루뱅 클러스터링 등임)을 포함한다.

[0470] 도 14c의 블록 1532를 참조하면, 일부 실시예에서, 복수의 세포 구성성분 모듈은 10개 내지 2000개, 100개 내지 10000개, 20개 내지 5000개, 2개 내지 15,000개, 80개 내지 5000개, 100개 내지 500개의 세포 구성성분 모듈로 이루어진다. 일부 실시예에서, 복수의 세포 구성성분 모듈은 2개 내지 500개의 세포 구성성분 모듈이다.

[0471] 도 14c의 블록 1534를 참조하면, 일부 실시예에서, 복수의 세포 구성성분은 10개 내지 2000개, 100개 내지 10000개, 20개 내지 5000개, 2개 내지 15,000개, 80개 내지 5000개, 100개 내지 500개의 세포 구성성분으로 이루어진다. 일부 실시예에서, 복수의 세포 구성성분은 2개 내지 500개의 세포 구성성분이다.

[0472] 도 14c의 블록 1536을 참조하면, 일부 실시예에서, 복수의 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 2개 내지 300개의 세포 구성성분으로 구성된다.

[0473] 도 2a의 블록 208 및 도 14c의 블록 1538을 참조하면, 방법은 전자 형태로 하나 이상의 제2 데이터세트를 획득

하는 단계를 포함한다. 하나 이상의 제2 데이터세트는 제2 복수의 세포의 각각의 개별 세포에 대해(제2 복수의 세포는 20개 이상의 세포를 포함하고, 관심 생리학적 조건을 알리는 복수의 공변량을 집합적으로 나타냄), 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함한다. 따라서, 세포 구성성분 카운트 데이터 구조가 획득되며, 여기서 세포 구성성분 카운트 데이터 구조는 (i) 제2 복수의 세포 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된다.

[0474] 도 14c의 블록 1540을 참조하면, 일부 실시예에서, 복수의 공변량은 세포 배치, 세포 공여자, 세포 유형, 질환 상태 또는 화학적 화합물에 대한 노출을 포함한다.

[0475] 도 2b의 블록 210 및 도 14d의 블록 1542를 참조하면, 활성화 데이터 구조는 공통 차원으로서 복수의 세포 구성성분 또는 그 표현을 사용하여 세포 구성성분 카운트 데이터 구조 및 잠재 표현을 조합함으로써 형성된다. 활성화 데이터 구조는, 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈에 대해, 제2 복수의 세포의 각각의 세포에 대해, 각각의 활성화 가중치를 포함한다.

[0476] 도 2b의 블록 212 및 도 14d의 블록 1544를 참조하면, 방법은 (i) 활성화 데이터 구조를 후보 모델로 입력 시 활성화 데이터 구조 내에 표현된 각각의 세포 구성성분 모듈 내의 복수의 공변량 중 각각의 공변량의 부재 또는 존재의 예측과 (ii) 각각의 세포 구성성분 모듈의 각각의 공변량의 실제 부재 또는 존재 사이의 차이를 사용하여 후보 세포 구성성분 모델을 훈련시키는 것을 더 포함한다. 훈련은 차이에 응답하여 후보 세포 구성성분 모델과 연관된 복수의 공변량 가중치를 조정하고, 여기서 복수의 공변량 가중치는 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대해, 각각의 개별 공변량에 대해, 각각의 공변량이 활성화 데이터 구조에 걸쳐 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 상응하는 가중치를 포함한다.

[0477] 도 14d의 블록 1546을 참조하면, 후보 세포 구성성분 모델의 훈련이 멀티-태스크 공식화에서 범주형 교차-엔트로피 손실을 사용하여 수행되고, 여기서 복수의 공변량 중 각각의 공변량은 복수의 비용 함수 내의 비용 함수에 상응하고, 복수의 비용 함수 내의 각각의 개별 비용 함수는 공통 가중 인자를 갖는다.

[0478] 따라서, 도 2c의 블록 214 및 도 14d의 블록 1548을 참조하면, 복수의 공변량 가중치는 후보 세포 구성성분 모델을 훈련시킬 때, 복수의 후보 세포 구성성분 모듈 내의 제1 세포 구성성분 모듈을 식별하기 위해 사용되고, 여기서 복수의 후보 세포 구성성분 모듈 내의 제1 세포 구성성분 모듈은 관심 생리학적 조건과 연관된다.

[0479] 일부 실시예에서, 제1 및/또는 제2 복수의 세포는 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 1000개, 적어도 2000개, 적어도 3000개, 적어도 4000개, 적어도 5000개, 적어도 10,000개, 적어도 20,000개, 적어도 30,000개, 적어도 50,000개, 적어도 80,000개, 적어도 100,000개, 적어도 500,000개 또는 적어도 1백만개의 세포를 포함한다. 일부 실시예에서, 제1 및/또는 제2 복수의 세포는 5백만개 이하, 1백만개 이하, 500,000개 이하, 100,000개 이하, 50,000개 이하, 10,000개 이하, 5000개 이하, 1000개 이하, 500개 이하, 200개 이하, 100개 이하 또는 50개 이하의 세포를 포함한다. 일부 실시예에서, 제1 및/또는 제2 복수의 세포는 5개 내지 100개, 10개 내지 50개, 20개 내지 500개, 200개 내지 10,000개, 1000개 내지 100,000개, 50,000개 내지 500,000개, 또는 10,000개 내지 1백만개의 세포를 포함한다. 일부 실시예에서, 제1 및/또는 제2 복수의 세포는 5개 이상의 세포에서 시작하여 5백만개 이하의 세포에서 끝나는 또 다른 범위 내에 속한다.

[0480] 일부 실시예에서, 제2 복수의 세포는 제1 복수의 세포에 포함된 임의의 세포를 포함하지 않는다. 일부 실시예에서, 제2 복수의 세포는 제1 복수의 세포에 포함된 세포의 일부 또는 전부를 포함한다.

[0481] 일부 실시예에서, 복수의 주석화된 세포 상태는 세포 표현형, 세포 거동, 질환 상태, 유전자 돌연변이, 유전자 또는 유전자 산물의 교란(예를 들어, 녹다운, 침묵, 과다발현 등), 및/또는 화합물에 대한 노출 중 하나 이상을 포함한다. 일부 실시예에서, 복수의 주석화된 세포 상태의 주석화된 세포 상태는 노출 조건 하에 화합물에 대한 제1 복수의 세포의 세포의 노출이다. 예를 들어, 세포의 노출은 1종 이상의 화합물을 사용한 세포의 임의의 치료를 포함한다. 일부 실시예에서, 1종 이상의 화합물은, 예를 들어 소분자, 생물제제, 치료제, 단백질, 소분자와 조합된 단백질, ADC, 핵산(예를 들어, siRNA, 간섭 RNA, 야생형 및/또는 돌연변이체 shRNA를 과다발현하는 cDNA, 야생형 및/또는 돌연변이체 가이드 RNA를 과다발현하는 cDNA(예를 들어, Cas9 시스템 또는 다른 세포-컴포넌트 편집 시스템) 등), 및/또는 임의의 전술한 것의 임의의 조합을 포함한다. 일부 실시예에서, 노출 조건은 노출 지속기간, 화합물의 농도, 또는 노출 지속기간과 화합물의 농도의 조합이다. 일부 실시예에서, 화합물은 본원에 설명된 임의의 실시예, 예컨대 상기 "화합물"이라는 명칭의 섹션이다.

[0482] 일부 실시예에서, 복수의 주석화된 세포 상태는 세포 배치, 세포 공여자, 세포 유형, 세포주, 질환 상태, 시점, 복제물 및/또는 관련 메타데이터의 하나 이상의 적응증을 포함한다. 일부 실시예에서, 복수의 주석화된 세포 상태는 실험 데이터(예를 들어, 유동 세포측정법 판독, 영상화 및 현미경검사 주석, 세포 구성성분 데이터 등)를 포함한다. 일부 실시예에서, 복수의 주석화된 세포 상태는 1종 이상의 유전자 마커(예를 들어, 카피수 변이, 단일 뉴클레오티드 변이체, 다중 뉴클레오티드 다형성, 삽입, 결실, 유전자 융합, 미소위성체 불안정성 상태, 증폭 및/또는 이소형)를 포함한다. 일부 실시예에서, 복수의 주석화된 세포 상태는 본원에 개시된 임의의 공변량 및/또는 본원에 개시된 임의의 관심 생리학적 조건, 예컨대 상기 "생리학적 조건"이라는 명칭의 섹션을 포함한다.

[0483] 본원에 개시된 임의의 세포 구성성분 및/또는 임의의 세포 구성성분 모듈, 및 그의 임의의 실시예, 치환, 변형, 추가, 결실 및/또는 조합은 상기 "세포 구성성분 및 세포 구성성분 모듈"이라는 명칭의 섹션에 설명된 바와 같이 세포 구성성분 모듈의 식별을 위해 고려된다. 예를 들어, 일부 실시예에서, 복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사산물, 단백질 또는 그의 조합이다. 일부 실시예에서, 복수의 세포 구성성분은 100개 내지 8,000개의 세포 구성성분으로 이루어진다. 일부 실시예에서, 복수의 세포 구성성분 모듈은 10개 내지 2000개의 세포 구성성분 모듈로 이루어진다. 일부 실시예에서, 복수의 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 2개 내지 300개의 세포 구성성분으로 이루어진다.

[0484] 일부 실시예에서, 각각의 세포 구성성분의 상응하는 풍부도는 상기 개시된 세포 구성성분 중 임의의 것의 풍부도를 포함한다.

[0485] 다수의 풍부도 카운팅 기술(예를 들어, 세포 구성성분 측정 기술) 중 어느 하나를 사용하여 각각의 개별 세포에서의 각각의 개별 세포 구성성분에 대한 상응하는 풍부도를 획득할 수 있다. 예를 들어, 표 1은 본 개시의 일부 실시예에 따라 단일-세포 세포 구성성분 측정을 위한 비제한적 기술을 열거한다.

[0486] 일부 실시예에서, 각각의 세포 구성성분의 상응하는 풍부도는 형광, 화학발광, 전기 신호 검출, 폴리머라제 연쇄 반응(PCR), 역전사효소 폴리머라제 연쇄 반응(RT-PCR), 디지털 액적 PCR(ddPCR), 고체-상태 나노포어 검출, RNA 스위치 활성화, 노던 블롯, 및/또는 유전자 발현의 연속 분석(SAGE)을 통한 마이크로어레이 분석을 포함하는 하나 이상의 방법을 사용하여 결정된다. 일부 실시예에서, 제1 및/또는 제2 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도는 비색 측정치, 형광 측정치, 발광 측정치, 또는 공명 에너지 전달(FRET) 측정치에 의해 결정된다.

[0487] 일부 실시예에서, 제1 및/또는 제2 복수의 세포에서의 각각의 세포에서의 유전자 발현은 세포를 시퀀싱한 후, 시퀀싱 동안 식별된 각각의 유전자 전사체의 양을 카운팅함으로써 측정될 수 있다. 일부 실시예에서, 시퀀싱 및 정량화된 유전자 전사체는 RNA, 예컨대 mRNA를 포함한다. 일부 실시예에서, 시퀀싱 및 정량화된 유전자 전사체는 mRNA의 하류 생성물, 예컨대 단백질(예를 들어, 전사 인자)을 포함한다. 일반적으로, 본원에 사용된 용어 "유전자 전사체"는 번역후 변형을 포함한 유전자 전사 또는 번역의 임의의 하류 생성물을 나타내기 위해 사용될 수 있고, "유전자 발현"은 일반적으로 유전자 전사체의 임의의 척도를 지칭하기 위해 사용될 수 있다.

[0488] 일부 실시예에서, 각각의 세포 구성성분의 상응하는 풍부도는 RNA 풍부도(예를 들어, 유전자 발현)이고, 각각의 세포 구성성분의 풍부도는 각각의 유전자에 상응하는 하나 이상의 핵산 분자의 폴리뉴클레오티드 수준을 측정함으로써 결정된다. 각각의 유전자의 전사체 수준은 제1 및/또는 제2 복수의 세포에서 각각의 세포에 존재하는 mRNA 또는 그로부터 유도된 폴리뉴클레오티드의 양으로부터 결정될 수 있다. 폴리뉴클레오티드는 마이크로어레이 분석, 폴리머라제 연쇄 반응(PCR), 역전사효소 폴리머라제 연쇄 반응(RT-PCR), 노던 블롯, 유전자 발현의 연속 분석(SAGE), RNA 스위치, RNA 핑거프린팅, 리가제 연쇄 반응, Qbeta 레플리카제, 등온 증폭 방법, 가닥 치환 증폭, 전사 기반 증폭 시스템, 뉴클레아제 보호 검정(Si 뉴클레아제 또는 RNAse 보호 검정), 및/또는 고체-상태 나노포어 검출을 포함하나 이에 제한되지는 않는 다양한 방법에 의해 검출 및 정량화될 수 있다. 예를 들어, 문헌 [Draghici, Data Analysis Tools for DNA Microarrays, Chapman and Hall/CRC, 2003]; 문헌 [Simon 등의 Design and Analysis of DNA Microarray Investigations, Springer, 2004]; 문헌 [Real-Time PCR: Current Technology and Applications, Logan, Edwards, and Saunders eds., Caister Academic Press, 2009]; 문헌 [Bustin A-Z of Quantitative PCR (IUL Biotechnology, No. 5), International University Line, 2004]; 문헌 [Velculescu 등의 (1995) Science 270: 484-487; Matsumura 등의 (2005) Cell. Microbiol. 7: 11-18]; 문헌 [Serial Analysis of Gene Expression (SAGE): Methods and Protocols (Methods in Molecular Biology), Humana Press, 2008]을 참조하며, 이들 각각은 본 출원에 그 전문이 참조로 포함되어 있다.

[0489] 일부 실시예에서, 각각의 세포 구성성분의 상응하는 풍부도는 자연 발생 핵산 분자, 뿐만 아니라 합성 핵산 분자를 포함한 제1 및/또는 제2 복수의 세포에서 각각의 세포로부터 발현된 RNA 또는 그로부터 유도된 핵산(예를 들어, RNA 폴리머라제 프로모터를 포함하는 cDNA 또는 cDNA로부터 유도된 증폭된 RNA)으로부터 획득된다. 따라서, 일부 실시예에서, 각각의 세포 구성성분의 상응하는 풍부도는 총 세포 RNA, poly(A)+ 메신저 RNA(mRNA) 또는 그의 일부, 세포질 mRNA, 또는 cDNA로부터 전사된 RNA(예를 들어, cRNA)와 같은 비제한적 공급원으로부터 획득된다. 전체 및 폴리(A)+ RNA를 제조하는 방법은 관련 기술분야에 널리 공지되어 있고, 일반적으로, 예를 들어 문헌 [Sambrook 등의 Molecular Cloning: A Laboratory Manual (3rd Edition, 2001)]에 기재되어 있다. 구아니디늄 티오시아네이트 용해에 후속한 CsCl 원심분리(예를 들어, 문헌 [Chirgwin 등의 1979, Biochemistry 18:5294-5299] 참조), 실리카 겔-기반 컬럼(예를 들어, RNeasy(퀴아젠(Qiagen), 캘리포니아주 발렌시아) 또는 스트라타프렙(StrataPrep)(스트라타진(Stratagene), 캘리포니아주 라 졸라))를 사용하거나, 또는 문헌 [Ausubel 등의 eds., 1989, Current Protocols In Molecular Biology, Vol. III, Green Publishing Associates, Inc., John Wiley & Sons, Inc., New York, at pp. 13.12.1-13.12.5]에 설명된 바와 같이 페놀 및 클로로포름을 사용하여 관심 세포로부터 RNA를 추출할 수 있다. 폴리(A)+ RNA는, 예를 들어 올리고-dT 셀룰로스를 사용한 선택에 의해, 또는 대안적으로 전체 세포 RNA의 올리고-dT 프라이밍된 역전사에 의해 선택될 수 있다. RNA는 관련 기술분야에 공지된 방법에 의해, 예를 들어 ZnCl2와의 인큐베이션에 의해 단편화되어 RNA의 단편을 생성할 수 있다.

[0490] 일부 실시예에서, 제1 및/또는 제2 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도는 시퀀싱에 의해 결정된다. 일부 실시예에서, 제1 및/또는 제2 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도는 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq), scTag-seq, 시퀀싱을 사용하는 전위효소-접근 가능 염색질에 대한 단일-세포 검정(scATAC-seq), CyTOF/SCoP, E-MS/Abseq, miRNA-seq, CITE-seq, 및 그 임의의 조합에 의해 결정된다.

[0491] 세포 구성성분 풍부도 측정 기술은 측정될 목적하는 세포 구성성분에 기초하여 선택될 수 있다. 예를 들어, scRNA-seq, scTag-seq, 및 miRNA-seq를 사용하여 RNA 발현을 측정할 수 있다. 구체적으로, scRNA-seq는 RNA 전사체의 발현을 측정하고, scTag-seq는 희귀 mRNA 종의 검출을 허용하고, miRNA-seq는 마이크로-RNA의 발현을 측정한다. CyTOF/SCoP 및 E-MS/Abseq를 사용하여 세포에서 단백질 발현을 측정할 수 있다. CITE-seq는 세포에서 유전자 발현 및 단백질 발현 둘 다를 동시에 측정하고, scATAC-seq는 세포에서 염색질 입체형태를 측정한다. 하기 표 1은 앞서 설명된 세포 구성성분 풍부도 측정 기술 각각을 수행하기 위한 예시적인 프로토콜을 제공한다.

표 1

표 1 – 예시적인 측정 프로토콜

기술	프로토콜
RNA-seq	*Olsen* 등의 (2018), "Introduction to Single-Cell RNA Sequencing," Current protocols in molecular biology 122(1), 페이지 57.
Tag-seq	*Rozenberg* 등의 (2016), "Digital gene expression analysis with sample multiplexing and PCR duplicate detection: A straightforward protocol," BioTechniques, 61(1), 페이지 26.
ATAC-seq	*Buenrostro* 등의 (2015), "ATAC-seq: a method for assaying chromatic accessibility genome-wide," Current protcols in molecular biology, 109(1), 페이지 21.
miRNA-seq	*Faridani* 등의 (2016), "Single-cell sequencing of the small-RNA transcriptome," Nature biotechnology, 34(12), 페이지 1264.
CyTOF/SCoPE-MS/Abseq	*Bandura* 등의 (2009), "Mass cytometry: technique for real time single cell multitarget immunoassay based on inductivitely coupled plasma time-of-flight mass spectrometry," Analystic chemistry, 81(16), 페이지 6813. *Budnik* 등의 (2018), "SCoPE-ME: mass scpectrometry of single mammalian cells quantifies proteome heterogenity during cell differentiation," Genome biology, 19(1), 페이지 161. *Shahi* 등의 (2017), "Abseq: Ultrahigh-throughoutput single cell protein profiling with droplep microfluidic barcoding," Scientific reports, 7, 페이지 44447.
CITE-seq	*Stoeckius* 등의 (2017), "Simultaneous epitope and transcritome measurement in single cells," Nature Methods, 14(9), 페이지 856.

[0492]

[0493] 일부 실시예에서, 복수의 세포 구성성분은 단일 시점에 측정된다. 일부 실시예에서, 복수의 세포 구성성분은 다수의 시점에 측정된다. 예를 들어, 일부 실시예에서, 복수의 세포 구성성분은 세포 상태 전이(예를 들어, 분화 과정, 화합물에 대한 노출에 대한 반응, 발달 과정 등) 전반에 걸쳐 다수의 시점에 측정된다.

[0494] 본 개시가 세포(예를 들어, 단일 세포)로부터 획득된 다른 세포 구성성분의 측정을 사용하는 유사한 방법을 포괄하기 때문에, 이는 제한이 아닌 예시로서 이해하여야 한다. 본 개시는 본 개시에 설명된 방법을 실시하는 개인 또는 기관에 의해 수행된 실험 작업으로부터 직접 획득된 측정을 사용하는 방법, 뿐만 아니라 예를 들어 다른 사람에 의해 수행되고 제3자 공개물에 보고된 데이터, 데이터베이스, 계약자에 의해 수행된 검정, 또는 개시된 방법을 실시하는데 유용한 적합한 입력 데이터의 다른 공급원을 포함한 임의의 수단 또는 메커니즘을 통해 이용가능하게 된 실험 작업의 결과의 보고로부터 간접적으로 획득된 측정을 사용하는 방법을 포괄하는 것임을 추가로 이해하여야 한다.

[0495] 일부 실시예에서, 제1 및/또는 제2 복수의 세포(예를 들어, 하나 이상의 제1 데이터세트 및/또는 하나 이상의 제2 데이터세트) 내의 복수의 세포 구성성분에 대한 상응하는 풍부도가 전처리된다. 일부 실시예에서, 전처리는 필터링, 정규화, (예를 들어, 참조 시퀀스에 대한) 맵핑, 정량화, 스케일링, 디컨볼루션, 세정, 차원 감소, 변환, 통계적 분석 및/또는 응집 중 하나 이상을 포함한다.

[0496] 예를 들어, 일부 실시예에서, 복수의 세포 구성성분이 원하는 품질, 예를 들어, 핵산 시퀀스의 크기 및/또는 품

질, 또는 각각의 세포 구성성분에 대한 최소 및/또는 최대 풍부도 값을 기초로 필터링된다. 일부 실시예에서, 필터링은 다양한 소프트웨어 도구, 예컨대 스케워(Skewer)에 의해 부분적으로 또는 전체적으로 수행된다. 문헌 [Jiang, H. 등의 BMC Bioinformatics 15(182): 1-12 (2014)]을 참조한다. 일부 실시예에서, 예를 들어, 시퀀싱 데이터 QC 소프트웨어 예컨대 AfterQC, Kraken, RNA-SeQC, FastQC, 또는 또 다른 유사한 소프트웨어 프로그램을 사용하여, 품질 제어를 위해 복수의 세포 구성성분이 필터링된다. 일부 실시예에서, 복수의 세포 구성성분은, 예를 들어 풀-다운, 증폭 및/또는 시퀀싱 바이어스(예를 들어, 맵핑가능성, GC 바이어스 등)를 설명하기 위해 정규화된다. 예를 들어, 문헌 [Schwartz 등의 PLoS ONE 6(1):e16685 (2011) and Benjamini and Speed, Nucleic Acids Research 40(10):e72 (2012)]을 참조하며, 그의 내용은 모든 목적을 위해 그 전문이 본원에 참조로 포함된다. 일부 실시예에서, 전처리는 복수의 세포 구성성분으로부터 세포 구성성분의 서브세트를 제거한다. 일부 실시예에서, 복수의 세포 구성성분에 대한 상응하는 풍부도를 전처리하는 것은 높은 신호-대-잡음 비를 개선시킨다(예를 들어, 저하시킨다).

[0497] 일부 실시예에서, 전처리는 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 기준 풍부도에 비교하는 것을 수행하는 것을 포함한다. 일부 실시예에서, 기준 풍부도는, 예를 들어, 정상 샘플, 매칭된 샘플, 기준 풍부도 값을 포함하는 기준 데이터세트, 기준 세포 구성성분 예컨대 하우스키핑 유전자, 및/또는 기준 표준으로부터 획득된다. 일부 실시예에서, 세포 구성성분 풍부도의 이러한 비교는 평균 차이 검정, 윌콕슨 순위-합계 검정(만 휘트니 U 검정), t-검정, 로지스틱 회귀 및 일반화된 선형 모델을 포함하나 이에 제한되지는 않는 임의의 차등 발현 테스트를 사용하여 수행된다. 관련 기술분야의 통상의 기술자는 세포 구성성분 풍부도의 비교 및/또는 정규화를 위해 다른 측정 기준이 또한 가능하다는 것을 인지할 것이다.

[0498] 따라서, 일부 실시예에서, 하나 이상의 제1 데이터세트 및/또는 하나 이상의 제2 데이터세트의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도는 미가공 풍부도 값, 절대 풍부도 값(예를 들어, 전사체 수), 상대 풍부도 값(예를 들어, 상대 형광 단위, 전사체 분석, 및/또는 유전자 세트 발현 분석(GSEA)), 화합물 또는 응집된 풍부도 값, 변환된 풍부도 값(예를 들어, 변환된 log2 및/또는 log10), 기준(예를 들어, 정상 샘플, 매칭된 샘플, 기준 데이터세트, 하우스키핑 유전자, 및/또는 기준 표준)에 대한 변화(예를 들어, 배수- 또는 로그-변화), 표준화된 풍부도 값, 중심 집중 경향의 척도(예를 들어, 평균, 중앙값, 모드, 가중 평균, 가중 중앙값, 및/또는 가중 모드), 분산의 척도(예를 들어, 분산, 표준 편차, 및/또는 표준 오차), 조정된 풍부도 값(예를 들어, 정규화, 스케일링, 및/또는 오차-보정), 차원-감소된 풍부도 값(예를 들어, 주 컴포넌트 벡터 및/또는 잠재성 컴포넌트), 및/또는 이들의 조합을 비제한적으로 포함하는 다양한 형태 중 어느 하나를 포함한다. 차원 감소 기술을 사용하여 세포 구성성분 풍부도를 획득하는 방법은 관련 기술분야에 공지되어 있고, 관련 기술분야의 통상의 기술자에게 명백한 바와 같이, 주 컴포넌트 분석, 인자 분석, 선형 판별 분석, 다차원 스케일링, 등척성 특징 맵핑, 국부 선형 포매, 헤시안 아이겐맵핑, 스펙트럼 포매, t-분포 확률적 이웃 포매, 및/또는 그의 임의의 치환, 추가, 결실, 변형, 및/또는 조합을 포함하나 이에 제한되지는 않는 하기에 추가로 상세히 설명된다. 예를 들어, 그 전문이 본원에 참조로 포함되는 문헌 [Sumithra 등의 2015, "A Review of Various Linear and Non Linear Dimensionality Reduction Techniques," Int J Comp Sci and Inf Tech, 6(3), 2354-2360]을 참조한다.

[0499] 일부 실시예에서, 복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하는 단계는 복수의 벡터의 각각의 벡터의 각각의 상응하는 복수의 요소를 사용하여 상관 모델을 복수의 벡터에 적용하는 것을 포함한다.

[0500] 일부 실시예에서, 상관 모델은 클러스터링 방법(예를 들어, 클러스터링 모델)을 포함한다. 일부 실시예에서, 상관 모델은 그래프 클러스터링 방법(예를 들어, 모델) 및/또는 비-그래프 클러스터링 방법을 포함한다. 일부 실시예에서, 그래프 클러스터링 방법은 피어슨-상관관계-기반 거리 메트릭에 대한 라이덴 클러스터링이다. 일부 실시예에서, 그래프 클러스터링 방법은 루뱅 클러스터링이다.

[0501] 예를 들어, 일부 구현에서, 이 방법은 상관-기반 비용 함수의 적용을 포함한다. 상관-기반 비용 함수를 최적화하는 것은 세포 구성성분(예를 들어, 유전자) 사이의 이웃 관계를 정의하는 최근접-이웃 그래프를 계산하고, 각각의 개별 세포 구성성분을 각각의 세포 내의 세포 구성성분에 대한 풍부도 카운트(예를 들어, 발현 값)를 저장함으로써 형성된 벡터에 의해 나타내는 것, 및 세포 구성성분 사이의 상관을 계산하는 것을 포함한다. 서로 간의 상관이 높은 세포 구성성분이 최근접 이웃인 것으로 결정되고, 그래프 클러스터링 방법(예를 들어, 라이덴 및/또는 루뱅)을 사용하여 그래프를 클러스터링함으로써 세포 구성성분 모듈을 형성하는데 사용된다.

[0502] 다수의 클러스터링 기술 중 어느 하나가 사용될 수 있으며, 그의 예는 계층적 클러스터링, k-평균 클러스터링 및 밀도 기반 클러스터링을 포함하나, 이에 제한되지는 않는다. 한 실시예에서, 계층적 밀도 기반 클러스터링

이 사용된다(HDBSCAN으로 지칭됨, 예를 들어, 문헌 [Campello 등의 (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans Knowl Disc Data*, *10*(1), 5] 참조). 또 다른 실시예에서, 클러스터링에 기초한 커뮤니티 검출, 예컨대 루뱅 클러스터링이 사용된다(예를 들어, 문헌 [Blondel 등의 (2008). Fast unfolding of communities in large networks. *J stat mech: theor exp*, *2008*(10), P10008] 참조). 또 다른 실시예에서, 라이덴 클러스터링이 사용된다. 라이덴 알고리즘은 개별 노드를 커뮤니티 사이에서 이동시켜 파티션을 결정하고, 파티션을 정밀화하며, 정밀화된 파티션에 기초하여 집성 네트워크를 생성함으로써 진행한다. 집성 네트워크는 과정의 이전 단계에서 결정된 정밀화되지 않은 파티션에 기초하여 추가로 파티셔닝되고, 새로운 파티션은 각각의 집성 네트워크 내의 개별 노드를 이동시킴으로써 정밀화된다. 예를 들어, 문헌 [Traag 등의 (2019), "From Louvain to Leiden: guaranteeing well-connected communities,"*Sci Rep* 9:5233, doi: 10.1038/s41598-019-41695-z]을 참조한다. 또 다른 실시예에서, 확산 경로 알고리즘이 사용된다.

[0503] 일반적으로, 클러스터링, 예컨대 루뱅(Louvain) 클러스터링 및/또는 라이덴(Leiden) 클러스터링은 하드 분할 기술을 사용하며, 여기서 각각의 요소(예를 들어, 각각의 세포 구성성분)는 중첩 없이 단일 클러스터에 고유하게 할당된다. 그러나, 어느 하나의 특정 이론에 매이지 않으면서, 세포 과정(예를 들어, 관심있는 생리학적 조건과 연관된)는 세포 내의 세포 구성성분의 네트워크 사이의 복잡한 동적 상호작용을 특징으로 할 수 있고, 여기서 단일 유전자는 예를 들어 임의의 수의 동일한 또는 상이한 과정 및 경로에서 유사하게 기능하는 임의의 수의 다른 유전자와 조합하여 세포 내의 2, 3, 4 또는 그 이상의 세포 과정에서 역할을 할 수 있다. 따라서, 세포내 활성의 복잡성과 병행하여, 세포 구성성분의 제1 모듈로의 클러스터링은 임의의 다른 모듈을 반드시 배제할 필요는 없다. 따라서, 일부 실시예에서, 세포 구성성분 모듈의 식별은 세포 구성성분의 중첩 서브세트를 갖는 모듈을 획득하는 것을 포함한다.

[0504] 상관-기반 모델을 사용하는 하드 분할 기술을 사용하는 것에 대안적으로 또는 추가로, 일부 실시예에서, 복수의 벡터를 사용하여 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈을 식별하는 것은 복수의 세포 구성성분의 표현을 복수의 차원 감소 컴포넌트로서 생산하는 사전 학습 모델을 포함한다. 일부 실시예에서, 사전 학습 모델은 L0-정규화된 오토인코더이다. 이들 모델의 이점은 이들이 모듈과 세포 구성성분 사이에 1:1 상응성을 시행하지 않지만 세포 구성성분이 동시에 여러 모듈에서 나타나게 한다는 것이다.

[0505] 예를 들어, 일부 구현에서, 방법은 예비 오토인코더 비용 함수의 적용을 포함한다. 일부 이러한 경우에, 희소한 오토인코더 비용 함수를 최적화하는 것은 파이토치 또는 텐서플로우에서 구현되는 표준 훈련을 사용하여, 1-계층 오토인코더를 그의 가중치의 L0 정규화 및 재구성 손실로 훈련시키는 것을 포함한다.

[0506] 퍼지 K-평균, 중첩 K-평균(OKM), 가중 OKM(WOKM), 중첩 분할 클러스터(OPC), 및 다중-클러스터 중첩 K-평균 연장(MCOKE), 및/또는 그의 임의의 변형 또는 조합을 포함하나 이에 제한되지는 않는 중첩 분할 알고리즘의 다른 방법이 가능하다.

[0507] 일부 실시예에서, 하나 이상의 제1 데이터세트 내에 인코딩된 잠재성 정보의 형상을 보존하면서, 고차원 데이터(예를 들어, 복수의 주석화된 세포 상태를 집합적으로 나타내는 제1 복수의 세포의 각각의 세포에 대한, 복수의 세포 구성성분 모듈에 걸친 복수의 세포 구성성분의 풍부도)를 더 낮은 차원의 공간으로 압축하기 위해 통계적 기술이 사용될 수 있다. 예를 들어, 도 4의 상부 패널에 예시된 바와 같이, 카운트 행렬은 제1 복수의 세포의 각각의 세포에 대해, 복수의 세포 구성성분의 각각의 세포 구성성분에 대해, 상응하는 카운트(예를 들어, 풍부도)를 포함한다. 카운트 행렬은 도 4의 하부 패널에 예시된 잠재 표현으로 변환될 수 있고, 여기서 데이터는 상이한 주석화된 세포 상태(예를 들어, 세포 유형, 노출 조건, 질환 등)의 조건 하에서의 그 상응하는 풍부도의 유사성을 기초로, 제1 복수의 세포에 걸친 세포 구성성분의 클러스터링을 나타내는 더 낮은 차원의 공간으로 감소된다. 따라서, 클러스터링된 세포 구성성분은 세포 구성성분 모듈로서 나타내어지고, 이는 잠재 표현에서 복수의 세포 상태에 걸친 거동의 유사성을 인코딩한다.

[0508] 도 4에 예시된 잠재 표현을 다시 참조하면, 각각의 행-열 그룹화에서의 엔트리에서의 값은 원래 입력 데이터세트에 기초한 차원수 감소에 의해 결정된다. 예를 들어, 각각의 엔트리는 각각의 행에 의해 표현되는 각각의 세포 구성성분 모듈 내에 포함된 복수의 세포 구성성분의 서브세트의 각각의 열에 의해 표현되는 각각의 개별 세포 구성성분에 대한 멤버십의 표시를 포함할 수 있다(예를 들어, $weight_{1-1}$, $weight_{1-2}$ 등). 특히, 일부 실시예에서, 각각의 엔트리는 각각의 세포 구성성분이 각각의 모듈에 포함되는지 여부를 나타내는 가중치이다. 일부 구현에서, 가중치는 멤버십의 이진 표시이다(예를 들어, 각각의 모듈에서의 존재 또는 부재는 각각 1 또는 0으로 표시됨). 일부 구현에서, 가중치는 각각의 모듈에 대한 세포 구성성분의 상대적 중요성(예를 들어, 멤버십

및/또는 상관의 가능성)을 나타내도록 스케일링된다.

[0509] 일부 실시예에서, 잠재 표현에서의 각각의 차원은 각각의 세포 구성성분의 표현에 상응한다. 세포 구성성분의 표현은, 예를 들어 세포 구성성분의 비선형 표현으로부터 발생할 수 있고, 예컨대 잠재 표현 행렬 내의 각각의 엔트리(예를 들어, 가중치)는 복수의 세포 구성성분에 상응한다. 세포 구성성분의 표현을 포함하는 다른 실시예는 주 컴포넌트 분석을 사용하여 획득된 잠재 표현을 포함하며, 여기서 각각의 주 컴포넌트는 복수의 세포 구성성분에 상응하는 데이터의 분산 및/또는 다른 변환을 나타낸다.

[0510] 일부 실시예에서, 차원수 감소 기술은 데이터의 일부 손실 압축을 초래한다. 그러나, 결과적인 잠재 표현(예를 들어, 잠재 표현(118))은 계산 저장 크기에서 더 작고, 따라서 모델 훈련과 같은 다른 하류 기술과 함께 분석하기 위해 더 적은 컴퓨팅 처리 능력을 요구한다. 따라서, 잠재 표현으로의 복수의 세포 구성성분 모듈의 배열은 현재 시대의 컴퓨팅 장치를 사용하여 본원에 개시된 방법의 컴퓨터 실행가능성을 증가시킨다.

[0511] 다양한 차원수 감소 기술이 사용될 수 있다. 일부 실시예에서, 차원 감소는 주 컴포넌트(PCA), 랜덤 투사, 독립적 컴포넌트 분석, 특징 선택, 인자 분석, Sammon 맵핑, 곡선 컴포넌트 분석, 확률적 이웃 포매(SNE), 이소맵, 최대 분산 언폴딩, 국부 선형 포매, t-SNE, 비-음성 행렬 인자화, 커널 주 컴포넌트 분석, 그래프-기반 커널 주 컴포넌트 분석, 선형 판별 분석(LDA), 일반화된 판별 분석, 균일 매니폴드 근사 및 투사(UMAP), LargeVis, 라플라시안 아이겐맵, 확산 맵, 네트워크(예를 들어, 신경망) 기술, 및/또는 피셔 선형 판별 분석이다. 예를 들어, 문헌 [Fodor, 2002, "A survey of dimension reduction techniques," Center for Applied Scientific Computing, Lawrence Livermore National, Technical Report UCRL-ID-148494]; 문헌 [Cunningham, 2007, "Dimension Reduction," University College Dublin, Technical Report UCD-CSI-2007-7], 문헌 [Zahorian 등의 2011, "Nonlinear Dimensionality Reduction Methods for Use with Automatic Speech Recognition," Speech Technologies, doi: 10.5772/16863. ISBN 978-953-307-996-7]; 및 문헌 [Lakshmi 등의 2016, "2016 IEEE 6th International Conference on Advanced Computing (IACC)," pp. 31-34. doi: 10.1109/IACC.2016.16, ISBN 978-1-4673-8286-1]를 참조하며, 이들 각각은 본 출원에 참조로 포함되어 있다. 따라서, 일부 실시예에서, 차원 감소는 주 컴포넌트 분석(PCA)이고, 각각의 개별 추출된 차원 감소 컴포넌트는 PCA에 의해 유도된 각각의 주 컴포넌트를 포함한다. 이러한 실시예에서, 복수의 주 컴포넌트 중 주 컴포넌트의 수는 PCA에 의해 계산된 주 컴포넌트의 임계 수로 제한될 수 있다. 주 컴포넌트의 임계 수는, 예를 들어 적어도 5개, 적어도 10개, 적어도 20개, 적어도 50개, 적어도 100개, 적어도 1000개, 적어도 1500개, 또는 임의의 다른 수일 수 있다. 일부 실시예에서, PCA에 의해 계산된 각각의 주 컴포넌트는 PCA에 의해 고유값을 할당받고, 제1 복수의 추출된 특징의 상응하는 서브세트는 최고 고유값을 할당받는 주 컴포넌트의 임계 수로 제한된다. 복수의 세포 구성성분 벡터 내의 각각의 개별 세포 구성성분 벡터에 대해, 복수의 차원 감소 컴포넌트를 각각의 세포 구성성분 벡터에 적용하여 복수의 차원 감소 컴포넌트 내의 각각의 개별 차원 감소 컴포넌트에 대한 차원 감소 컴포넌트 값을 포함하는 상응하는 차원 감소 벡터를 형성한다. 이는 복수의 세포 구성성분 벡터로부터 상응하는 복수의 차원 감소 벡터를 형성하여, 잠재 표현으로 배열된 복수의 세포 구성성분 모듈을 형성한다.

[0512] 일부 실시예에서, 방법은 잠재 표현으로 배열된 복수의 세포 구성성분 모듈을 사용하여 매니폴드 학습을 수행하는 것을 더 포함한다. 일반적으로, 매니폴드 학습은 데이터세트에서 최대 변동을 결정함으로써 고차원 데이터의 저차원 구조를 설명하는데 사용된다. 예는 힘 방향 레이아웃(Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129-1164)(예를 들어, Force Atlas 2), t-SNE(t-distributed stochastic neighbor embedding), 국부 선형 포매(Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323-2326), ISOMAP(local linear isometric mapping)(Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319-2323), 커널 PCA, 그래프-기반 커널 PCA, PHATE(Potential of Heat-Diffusion for Affinity Based Trajectory Embedding), GDA(generalized discriminant analysis), UMAP(Uniform Manifold Approximation and Projection), 또는 커널 판별 분석을 포함하나, 이에 제한되지는 않는다. 판별 분석은 특히 일부 정보가 각각의 세포의 특정 세포 유형에 대해 미리 공지된 경우에 사용될 수 있다. 힘-지향 레이아웃은 기저 세포 과정으로부터 발생하는 기저 데이터의 비선형 양태를 인코딩하는 새로운 보다 낮은 차원을 식별하는 그의 능력 때문에 다양한 특정한 실시예에서 유용하다. 힘 방향 레이아웃은 데이터를 가장 잘 나타내는 감소된 차원수를 결정하기 위한 메커니즘으로서 물리학-기반 모델을 사용한다. 예로서, 힘 방향 레이아웃은 본 실시예에서, 하나 이상의 제1 데이터세트의 각각의 세포가 "반발력"을 할당받고, 제1 복수의 세포에 걸쳐 계산될 때,

이들 경쟁하는 "힘" 하에서 함께 "확산"하는 데이터의 섹터를 식별하는 글로벌 "중력"이 존재하는 물리적 시뮬레이션의 형태를 사용한다. 힘 방향 레이아웃은 데이터의 구조에 대해 거의 가정하지 않으며, 노이즈-제거 접근법을 부과하지 않는다.

[0513] 매니폴드 학습은, 예를 들어, 그 전문이 본원에 참조로 포함되는 문헌 [Wang 등의 2004, "Adaptive Manifold Learning," Advances in Neural Information Processing Systems 17]에 추가로 기재되어 있다.

[0514] 일부 실시예에서, 복수의 공변량은 세포 배치, 세포 공여자, 세포 유형, 질환 상태 또는 화학적 화합물에 대한 노출을 포함한다. 일부 실시예에서, 복수의 공변량은 제2 복수의 세포에서의 하나 이상의 세포와 관련된 시점, 복제물 및/또는 관련 메타데이터의 하나 이상의 지표를 포함한다. 일부 실시예에서, 복수의 공변량은 실험 데이터(예를 들어, 유동 세포측정법 판독, 영상화 및 현미경검사 주석, 세포 구성성분 데이터 등)를 포함한다. 일부 실시예에서, 복수의 공변량은 제2 복수의 세포에서의 1종 이상의 세포에 특징적인 1종 이상의 유전자 마커(예를 들어, 카피수 변이, 단일 뉴클레오티드 변이체, 다중 뉴클레오티드 다형성, 삽입, 결실, 유전자 융합, 미소위성체 불안정성 상태, 증폭 및/또는 이소형)를 포함한다. 일부 실시예에서, 복수의 공변량은 세포 표현형, 세포 거동, 질환 상태, 유전자 돌연변이, 유전자 또는 유전자 산물의 교란(예를 들어, 녹다운, 침묵, 과다발현 등), 및/또는 제2 복수의 세포에서 하나 이상의 세포에 대한 노출 조건 중 하나 이상을 포함한다.

[0515] 예를 들어, 일부 실시예에서, 공변량은 노출 조건 하에 화합물에 대한 제2 복수의 세포 중의 세포의 노출에 대한 노출 또는 반응이다. 일부 실시예에서, 세포의 노출은 1종 이상의 화합물을 사용한 세포의 임의의 치료를 포함한다. 일부 실시예에서, 1종 이상의 화합물은, 예를 들어 소분자, 생물제제, 치료제, 단백질, 소분자와 조합된 단백질, ADC, 핵산(예를 들어, siRNA, 간섭 RNA, 야생형 및/또는 돌연변이체 shRNA를 과다발현하는 cDNA, 야생형 및/또는 돌연변이체 가이드 RNA를 과다발현하는 cDNA(예를 들어, Cas9 시스템 또는 다른 세포-컴포넌트 편집 시스템) 등), 및/또는 임의의 전술한 것의 임의의 조합을 포함한다. 일부 실시예에서, 노출 조건은 노출 지속기간, 화합물의 농도, 또는 노출 지속기간과 화합물의 농도의 조합이다.

[0516] 일부 실시예에서, 공변량은 하나 이상의 세포에서 세포 상태 전이 및/또는 교란 시그니처를 유도하는, 하나 이상의 세포에 적용되는 화합물이다(예를 들어, 교란원).

[0517] 일부 실시예에서, 공변량은 복수의 세포 구성성분 중의 세포 구성성분 또는 제2 복수의 세포 중의 세포와 연관된 지식 용어(예를 들어, 주석)이다. 예를 들어, 일부 실시예에서, 공변량은 전장 연관성 연구(GWAS; genome-wide association study) 주석, 유전자 세트 풍부화 검정(GSEA) 주석, 유전자 온톨로지 주석, 기능적 및/또는 신호전달 경로 주석, 및/또는 세포 시그니처 주석이다. 일부 실시예에서, 공변량은 NIH 유전자 발현 옴니버스(GEO), EBI ArrayExpress, NCBI, BLAST, EMBL-EBI, GenBank, Ensembl, KEGG 경로 데이터베이스, 및/또는 임의의 질환-특정 데이터베이스를 포함하나 이에 제한되지는 않는, 관련 기술분야에 공지된 임의의 공공 지식 데이터베이스로부터 획득된다. 일부 실시예에서, 공변량은 교란(예를 들어, 소분자) 유도된 유전자 발현 시그니처를 제공하는 데이터베이스, 예컨대 통합 네트워크-기반 세포 시그니처의 라이브러리(LINCS) L1000 데이터세트로부터 획득된다. 예를 들어, 그 전문이 본원에 참조로 포함되는 문헌 [Duan, 2016, "L1000CDS^2: An ultra-fast LINCS L1000 Characteristic Direction Signature Search Engine," Systems Biology and Applications 2, article 16015]을 참조한다.

[0518] 일부 실시예에서, 복수의 공변량은 적어도 3개, 적어도 5개, 적어도 10개, 적어도 15개, 적어도 20개, 적어도 30개, 적어도 40개, 적어도 50개, 적어도 60개, 적어도 70개, 적어도 80개, 적어도 90개, 적어도 100개, 적어도 200개, 적어도 300개, 적어도 400개, 적어도 500개, 적어도 600개, 적어도 700개, 적어도 800개, 적어도 900개, 적어도 1000개, 적어도 2000개 또는 적어도 3000개의 공변량을 포함한다. 일부 실시예에서, 복수의 공변량은 5000개 이하, 1000개 이하, 500개 이하, 200개 이하, 100개 이하, 50개 이하, 또는 20개 이하의 공변량을 포함한다. 일부 실시예에서, 복수의 공변량은 3개 내지 10개, 10개 내지 50개, 20개 내지 500개, 200개 내지 1000개, 또는 1000개 내지 5000개의 공변량을 포함한다. 일부 실시예에서, 복수의 공변량은 3개 이상의 공변량에서 시작하여 5000개 이하의 공변량에서 끝나는 또 다른 범위 내에 속한다.

[0519] 일부 실시예에서, 복수의 공변량에서의 각각의 공변량은 세포 상태 전이 및/또는 교란 시그니처를 유도하는 하나 이상의 세포에 적용되는 화합물이고, 복수의 공변량은 복수의 화합물이다. 일부 실시예에서, 복수의 공변량은 상기 "화합물"이라는 명칭의 섹션에 개시된 바와 같은 복수의 화합물로 이루어진다.

[0520] 도 5는 세포 구성성분 카운트 데이터 구조(예를 들어, 관심 생리학적 조건을 알리는 복수의 공변량을 집합적으로 나타내는 제2 복수의 세포를 사용하여 획득됨) 및 복수의 세포 구성성분 또는 그 표현을 공통 차원으로서 사

용하는 잠재 표현을 조합함으로써 형성된 예시적인 활성화 데이터 구조를 예시한다. 이를 달성하기 위해, 일부 실시예에서, 제2 복수의 세포에 대한 카운트 행렬(예를 들어, 도 4에 예시된 제1 복수의 세포에 대한 카운트 행렬과 구조가 유사함) 및 잠재 표현이 함께 곱해져서, 잠재 표현 행렬의 가중치가 카운트 행렬의 정규화된 카운트와 곱해진다. 일반적으로, 2개의 행렬은 공통 차원(예를 들어, 제1 행렬의 x-축 및 제2 행렬의 y-축)으로 함께 곱해질 수 있다. 제1 및 제2 행렬의 그의 공통 차원에 의한 행렬 곱셈은 제1 행렬 및/또는 제2 행렬에 대안적으로 또는 추가로, 훈련되지 않은 또는 부분적으로 훈련된 모델에 적용될 수 있는 보조 데이터의 제3 행렬을 산출한다.

[0521] 따라서, 일부 이러한 실시예에서, 카운트 행렬은 *n_cells x n_genes*의 차원을 갖고, 잠재 표현은 *n_genes x n_modules*의 차원을 갖고, 여기서 *n_cells*는 제2 복수의 세포 내의 세포의 수이고, *n_genes*는 복수의 세포 구성성분 내의 세포 구성성분(예를 들어, 유전자) 또는 그 표현의 수이고, *n_modules*은 복수의 세포 구성성분 모듈 내의 모듈의 수이다. 이는 카운트 행렬 내의 세포 구성성분의 풍부도를 각각의 세포(예를 들어, 하나 이상의 관심 공변량에 상응함)가 그의 모듈 활성화를 특징으로 하는 공간 내로 맵핑하고, 생성된 행렬 표현(예를 들어, 활성화 데이터 구조)은 *n_cells x n_modules*의 차원을 갖는다(예를 들어, *n_genes*의 공통 차원을 곱한 이후).

[0522] 예를 들어, 행렬 곱셈을 사용한 잠재 표현 및 세포 구성성분 카운트 데이터 구조의 조합, 및 행렬 형태의 생성된 활성화 데이터 구조가 도 5에 집합적으로 예시된다. 잠재 표현(도 5의 상부 패널에 예시됨)은 차원 Z x K를 갖고, 여기서 Z는 세포 구성성분 또는 그 표현의 수이고, K는 세포 구성성분 모듈의 수이다. 세포 구성성분 카운트 데이터 구조(하부 좌측 패널에 예시됨)는 차원 G x Z를 가지며, 여기서 G는 제2 복수의 세포에서의 세포의 수이고, 잠재 표현에 대해, Z는 세포 구성성분의 수 또는 그 표현이다. Z(세포 구성성분의 수 또는 그의 표현)를 공통 차원으로서 사용하는 행렬 곱셈에 의한 조합은 차원 G x K를 갖는 생성된 활성화 데이터 구조를 생성한다. 각각의 개별 행에서 각각의 개별 열에 대한 각각의 엔트리는 각각의 열에 상응하는 제2 복수의 세포의 각각의 세포의 각각의 개별 세포 구성성분 모듈의 활성화를 나타내는 활성화 가중치이다. 따라서, 도 5에 예시된 바와 같이, 모듈 1에 상응하는 카운트는 세포 1에 상응하는 활성화 $\text{weight}_{1\text{-}1}$, 세포 G에 상응하는 활성화 $\text{weight}_{1\text{-}G}$ 등을 포함한다.

[0523] 일부 실시예에서, 활성화 데이터 구조 내의 복수의 활성화 가중치는 차등 모듈 활성화를 포함한다. 일부 실시예에서, 차등 모듈 활성화(예를 들어, 활성화 데이터 구조 내의 제2 복수의 세포 내의 세포 사이의 각각의 모듈의 차등 활성화 가중치)는 함수 $(mu_1 - mu_2) / (var_1 + var_2)^{-0.5}$를 사용하여 v-점수를 계산함으로써 획득되며, 여기서 *mu_i*는 각각의 조건 *i*(예를 들어, 공변량 *i*)를 갖는 세포에 걸친 모듈 활성화의 평균을 나타내고, *var_i*는 조건 *i*에서의 모듈 활성화의 분산을 나타낸다. V-점수는 분모 내 세포의 수에 의해 정규화되지 않은 t-점수로 설명될 수 있다.

[0524] 일부 실시예에서, 활성화 데이터 구조 내의 제2 복수의 세포의 각각의 개별 세포는 각각의 공변량을 나타낸다. 일부 실시예에서, 활성화 데이터 구조 내의 제2 복수의 세포의 각각의 개별 세포는 세포 상태 전이 및/또는 교란 시그니처를 유도하는 하나 이상의 세포에 적용된 각각의 화합물을 나타낸다.

[0525] 따라서, 일부 실시예에서, 활성화 데이터 구조는 제2 복수의 세포에 의해 나타내어지는 복수의 화합물의 각각의 화합물에 대한 노출에 상응하는(예를 들어, 이에 상관되는 및/또는 이에 반응하는) 각각의 세포 구성성분 모듈의 활성화(예를 들어, 활성화 수준 또는 정도)를 나타낸다. 예를 들어, 제2 복수의 세포의 각각의 개별 세포가 각각의 교란원(예를 들어, 하나 이상의 세포가 노출되는 화합물 및/또는 세포 상태 전이 및/또는 교란 시그니처를 유도하는 화합물)을 나타내는 일부 실시예에서, 활성화 데이터 구조는 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대한 각각의 활성화 가중치를 포함하고, 이는 각각의 화합물을 사용한 치료와 상관되고/되거나 이에 반응한 각각의 세포 구성성분 모듈의 활성화(예를 들어, 유도 및/또는 차등 발현)를 나타낸다.

[0526] 일부 실시예에서, 후보 세포 구성성분 모델은 상기 "모델 아키텍처"라는 명칭의 섹션에 설명된 바와 같은, 본원에 개시된 모델 아키텍처 중 임의의 것을 포함한다.

[0527] 일부 실시예에서, 후보 세포 구성성분 모델은 오토인코더, 희소 오토인코더, 및/또는 희소 다중-판독, 지식-결합 오토인코더이다. 일부 실시예에서, 후보 세포 구성성분 모델은 반-지도 모델이다. 일부 실시예에서, 후보 세포 구성성분 모델은 1-계층 신경망(예를 들어, 소프트맥스(SoftMax) 및/또는 로지스틱 회귀 모델)이다. 일부 실시예에서, 후보 세포 구성성분 모델은 1차원 후버 아웃라이어 리그레서 모델이다.

[0528] 일부 실시예에서, 후보 세포 구성성분 모델은 복수의 계층을 포함하는 희소 다중-판독, 지식-결합 오토인코더이

고, 여기서 제1 계층은 잠재 표현을 획득하는데 사용되고, 제2 계층은 세포 구성성분 모듈 지식 구성(예를 들어, 공변량 가중치 행렬)를 획득하는데 사용된다.

[0529] 일부 실시예에서, 후보 세포 구성성분 모델을 훈련시키는 것은 멀티-태스크 공식화에서 범주형 교차-엔트로피 손실을 사용하여 수행되며, 여기서 복수의 공변량에서의 각각의 공변량은 복수의 비용 함수에서의 비용 함수에 상응하고, 복수의 비용 함수에서의 각각의 개별 비용 함수는 공통 가중 인자를 갖는다.

[0530] 일부 실시예에서, 후보 세포 구성성분 모델을 훈련시키는 것은 관심 생리학적 조건과 연관된 세포 구성성분 모듈의 세트의 제1 세포 구성성분 모듈을 식별하도록 모델을 훈련시킨다. 훈련 모델을 위한 방법은 본원에 추가로 상세하게 기재되어 있다. 본원에 개시된 방법 및/또는 실시예 중 임의의 것이 상기 "모델 훈련"이라는 명칭의 섹션에 설명된 바와 같이 후보 세포 구성성분 모델을 훈련시키는데 사용하기 위해 고려된다.

[0531] **V. 예**

[0532] 화합물을 생리학적 조건과 연관시키기 위한 모델의 예시적인 성능 척도 및 치료 응용이 본원에 제공된다.

[0533] **예 1. 지방산-관련 세포 과정의 활성화에 대한 화학 구조의 예측.**

[0534] 본 예에서, 세포 구성성분 모듈을 먼저 정의하였다. 이는 세포가 관심 생리학적 조건과 연관된 상이한 상태를 나타내는 세포에 대한 발현 데이터를 획득함으로써 수행되었다. 이는 원 출원의 제27항을 따른다. 세포 구성성분 풍부도 값을 각각의 세포로부터 측정하고, 이 데이터를 사용하여 세포 구성성분을 클러스터링한다. 세포가 나타내는 다양한 상태에 걸쳐 발현 값이 서로 상관되는 세포 구성성분은 세포 구성성분 모듈로 그룹화된다. 이는 여러 세포 구성성분 모듈을 생성하며, 이들 각각은 세포 구성성분 샘플의 상이한 서브세트를 포함한다. 일부 실시예에서, 각각의 세포 구성성분 모듈이 세포 구성성분의 상이한 서브세트를 갖지만, 하나의 세포 구성성분 모듈 내의 세포 구성성분과 또 다른 세포 구성성분 모듈 내의 세포 구성성분 사이에 중첩이 존재하는 것이 가능하다.

[0535] 추가로, 본 예에서, 추가의 훈련 데이터가 제2 훈련 세트의 형태로 획득된다. 이러한 제2 훈련 세트는 세포 구성성분에 대한 단일 세포 풍부도 데이터를 또한 포함한다. 그러나, 이 제2 훈련 세트에서, 각각의 세포는 복수의 훈련 화학적 화합물 내의 상이한 화학적 화합물에 노출되었다. 이러한 훈련 세트에서, 공지된 양은 각각의 상이한 화학적 화합물의 지문, 및 이러한 화합물에 노출된 세포의 생성된 세포 구성성분 풍부도 데이터이다. 제2 데이터세트에 대한 데이터는 세포 구성성분 아이덴티티에 대한 제1 축 및 세포 아이덴티티에 대한 제2 축을 갖는 카운트 행렬(502)(도 5에 예시됨)로서 배열될 수 있다. 따라서, 카운트 행렬(502) 내의 각각의 요소는 주어진 세포 내의 주어진 세포 구성성분의 풍부도다이다. 또한, 카운트 행렬(502) 내의 각각의 개별 열(이는 특정한 세포에 상응함)은 특정한 세포가 노출된 특정한 화합물로 표지된다. 따라서, 카운트 행렬(502)의 각각의 열은 특정한 화합물(예를 들어, 훈련 화합물)로 표지되고, 각각의 요소는 상응하는 세포(X-축)에 대한 상응하는 세포 구성성분의 카운트(Y-축)이다.

[0536] 도 5에 예시된 바와 같이, 제1 데이터세트(잠재 표현(404)) 및 제2 데이터세트(카운트 행렬(502))로부터의 데이터는 조합되어 활성화 데이터 구조(예를 들어, 도 5에 예시된 바와 같은 활성화 데이터 구조(504))를 형성한다. 예를 들어, 이를 달성하기 위한 한 가지 방식은 제1 축이 세포 구성성분 모듈을 나타내고 제2 축이 각각의 세포 구성성분을 나타내도록 세포 구성성분 모듈을 잠재 표현(404) 내의 행으로서 배열하는 것이다. 이러한 방식으로, 활성화 데이터 구조(504)를 생성하기 위해, 잠재 표현(404) 및 카운트 행렬(502)은 행렬 곱셈을 통해 그의 공통 축, 세포 구성성분 수가 곱해져서 활성화 데이터 구조(504)에 도달한다. 활성화 데이터 구조(504)는 카운트 행렬(502)로부터의 세포 아이덴티티 축 및 잠재 표현(504)으로부터의 세포 구성성분 모듈 축을 보유한다. 상이한 세포 유형에 대해 상이한 활성화 구조가 형성될 수 있다. 즉, 카운트 행렬(502)을 형성하는데 사용된 세포는 특정한 관심 질환 상태를 나타낼 수 있다. 따라서, 상이한 활성화 데이터 구조(504)가 상이한 질환 상태 또는 다른 관심 표현형에 대해 형성될 수 있다.

[0537] 도 6을 참조하면, 일부 경우에, 활성화 데이터 구조(504)의 각각의 행(도 5로부터 그리고 이제 도 6의 상단에 있음)은 상이한 모델(601)에 대한 훈련 데이터로서 역할한다. 예를 들어, 모델(601)이 화합물 1 내지 W가 각각 세포 구성성분 모듈(1)을 활성화시키는 정도를 나타내는 행(604-1)($Weight_{1-1}$ 내지 $Weight_{1-W}$)의 가중치를 포함하는 경우를 고려한다. 이 모델(601)은 활성화 데이터 구조(504)의 행(640)의 요소에 대해 훈련되고, 이는 각각의 훈련 화합물(1), …, G가 세포 구성성분 모듈(1)을 활성화시키는 정도를 제공한다. 이러한 훈련에서, 먼저 세포 1이 노출된 화합물의 지문 표현을 모델 601에 입력한다. 이러한 입력에 응답하여, 세포 구성성분 모듈 1

에 대한 모델(601)은 활성화 값(Pred로 명명됨)을 출력한다. 도 6의 명명법에서의 $Value_1$. 이 출력 활성화 값을 활성화 데이터 구조(504)의 Act_{1-1}인 실제 활성화 값과 비교한다. 이어서, 세포 2가 노출된 화합물의 지문 표현을 모델 601에 입력하였다. 이 입력에 응답하여, 모델은 활성화 값(Pred. $Value_2$)을 출력한다. 이 출력 활성화 값을 활성화 데이터 구조(504)의 Act_{1-2}인 화합물 2에 대한 실제 활성화 값과 비교한다. 이 과정은 세포 G를 통해 진행된다. 세포 G가 노출된 화합물의 지문 표현을 모델(601)에 입력하였다. 이에 응답하여, 모델은 활성화 값(Pred. $Value_G$)을 출력할 것이다. 이 출력 활성화 값은 활성화 데이터 구조(504)의 Act_{1-G}인 세포 G에 대한 실제 활성화 값과 비교된다. 본 예에서, W 및 G는 동일한 값을 갖는다. 이러한 방식으로, 세포 구성성분 모듈 1에 대해 도 5에 요약된 바와 같은 활성화 데이터 구조를 유도하는데 사용된 화합물의 훈련 세트의 각각의 화합물에 대해 결과적인 예측(Pred. Value)이 이루어진다. (활성화 값의) 앞서 설명된 계산된 예측은 각각의 이들 화합물에 대한 앞서 설명된 실제 활성화 값과 비교되고, 예측 및 실제 활성화 값 사이의 차이는 역전파 및 관련 모델 정밀화 기술을 사용하여 모델(601)을 추가로 훈련시키는데 사용된다.

[0538] 따라서, 결과는 각각의 세포 구성성분 모듈에 대해 하나씩인 일련의 훈련된 모델(601)이다. 테스트 화합물의 지문은 각각의 훈련된 모델에 입력될 수 있고, 각각의 개별 훈련된 모델(601)은 예측된 활성화 값을 출력하며, 그의 크기는 각각의 훈련된 모델에 상응하는 세포 구성성분 모듈이 테스트 화합물에 의해 활성화되는지 여부를 나타낸다. 이제, 과정의 개관을 설명하였고, 각각의 단계를 본 예에서 사용된 실험 데이터와 함께 설명한다.

[0539] 제1 세포 구성성분 모듈(도 1, 도 4 132-1)을 다음 과정에 의해 식별한다. 하나 이상의 제1 데이터세트를 전자 형태로 획득한다. 하나 이상의 제1 데이터세트는 복수의 주석화된(예를 들어, 표지된 또는 공지된) 세포 상태를 집합적으로 나타내는 제1 복수의 세포(예를 들어, 20개 이상의 세포)에 대한 데이터를 포함한다. 제1 데이터세트는, 제1 복수의 세포의 각각의 개별 세포에 대해, 복수의 세포 구성성분(예를 들어, 10개 이상의 세포 구성성분) 내의 각각의 개별 세포 구성성분에 대해, 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함한다. 예를 들어, 각각의 세포에 대한 전사 데이터이다. 이러한 방식으로, 복수의 벡터가 액세스 또는 형성된다. 복수의 벡터의 각각의 개별 벡터는 복수의 구성성분의 각각의 세포 구성성분에 상응하고, 상응하는 복수의 요소를 포함한다. 벡터의 상응하는 복수의 요소의 각각의 개별 요소는 제1 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 갖는다. 따라서, 일부 이러한 실시예에서, 복수의 세포 상태에서 각각의 세포 상태에 대한 전사 데이터가 획득된다.

[0540] 예시하기 위해, 도 4에 예시된 형태의 카운트 행렬(402)이 형성된다. 본 예에서는, 지방전구세포에서 대사적으로 활성인 과정을 유도하는 것으로 공지된 소분자 교란원을 사용하였다. 지방전구세포 세포주의 분취물을 24시간 동안 교란원에 노출시키고, 교란된 조건에서 세포주의 노출된 분취물에 대해 scRNA-seq 판독을 획득하였다. 또한 교란원에 노출되지 않은 세포주의 분취물에 대해 scRNA-seq 판독을 얻고, 이들 판독은 대조군 조건을 나타낸다. 이러한 방식으로, 도 14a의 블록 1504에 따라, 제1 복수의 세포의 각각의 개별 세포에 대해, 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하고, 그에 의해, 복수의 벡터에 액세스하거나 복수의 벡터를 형성하는 제1 데이터세트를 획득하였다. 즉, 교란원에 노출된 세포 및 교란원에 노출되지 않은 세포(대조군 세포) 모두에서 측정된 각각의 세포 구성성분(예를 들어, 유전자)의 발현 값은 도 4에 예시된 카운트 행렬(402)의 요소를 형성하였다. 도 4에 도시되고 도 14a의 블록 1510에 나타낸 바와 같이, 카운트 행렬(402)은 각각의 세포 구성성분에 대한 벡터를 포함하고, 따라서 복수의 벡터가 존재한다. 복수의 벡터의 각각의 개별 벡터는 (i) 복수의 구성성분의 각각의 세포 구성성분에 상응하고, (ii) 상응하는 복수의 요소를 포함한다.

[0541] 예를 들어, 세포 구성성분 1(예를 들어, 유전자 1)에 대해, 카운트 1-1, …, 카운트 1-N은 세포 1 내지 N에서의 유전자 1의 발현의 측정이며, 여기서 N개 세포 중 일부는 교란원에 노출되었고 일부는 노출되지 않았고, 이들 카운트는 세포 구성성분 1에 대한 벡터의 요소를 형성한다. 즉, 도 14a의 블록 1512에 따라, 세포 구성성분 1에 대한 벡터의 상응하는 복수의 요소의 각각의 개별 요소는 제1 복수의 세포의 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 갖는다. 본 예는 2가지 상태(교란원에 노출되거나 노출되지 않음)를 포함하지만, 원론적으로 임의의 수의 상태, 예컨대 교란원의 상이한 농도, 노출 시간 등이 포함될 수 있다.

[0542] 도 14a의 블록 1514에 따르면, 본 예 1에는 2가지 주석화된 상태가 있다: 대조군(교란원에 대한 노출 없음) 및 교란원의 노출. 즉, 복수의 주석화된 세포 상태 중 하나의 주석화된 세포 상태는 노출 조건(예를 들어, 노출 지속기간, 여기서 24시간) 하에 화합물(여기서, 교란원)에 대한 제1 복수의 세포의 세포의 노출이다. 본 예는

2가지 상태(교란원에 노출되거나 노출되지 않음)로 이루어지지만, 원론적으로 임의의 수의 상태, 예컨대 교란원의 상이한 농도, 노출 시간 등이 포함될 수 있다.

[0543] 카운트 행렬(402)을 필터링 및 정규화 단계를 통해 전처리하여, 높은 신호-대-잡음 비를 갖는 여러 유전자를 함유하는 전처리된 카운트 행렬을 생성하였다.

[0544] 복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별한다. 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 복수의 세포 구성성분의 서브세트를 포함한다. 복수의 세포 구성성분 모듈은 (i) 복수의 후보 세포 구성성분 모듈 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현으로 배열되고, 이때 복수의 세포 구성성분 모듈은 10개를 초과하는 세포 구성성분 모듈을 포함한다.

[0545] 일부 실시예에서, 각각의 후보 세포 구성성분 모듈은 후보 전사 지문이다.

[0546] 본 예에서, 카운트 행렬(402)을 사용하여 세포 구성성분 모듈(132)을 식별하였다. 이는 도 14b의 블록 1526에 따라 수행되었다: 복수의 벡터(도 4의 카운트 행렬(402)의 각각의 행)를 사용하여 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하고, 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈은 복수의 세포 구성성분의 서브세트를 포함한다.

[0547] 이는 (i) 복수의 후보 세포 구성성분 모듈 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현을 초래하였고, 이때 복수의 후보 세포 구성성분 모듈은 10개를 초과하는 세포 구성성분 모듈을 포함한다. 이러한 잠재 표현의 예는 도 4의 잠재 표현(404)이고, 이때, 각각의 개별 후보 세포 구성성분 모듈(132)에 대해, 어느 세포 구성성분이 각각의 후보 세포 구성성분 모듈 내에 있는지에 대한 표시가 존재한다.

[0548] 잠재 표현(404)은 도 14c의 블록 1528에 따라 형성되었다: 복수의 벡터의 각각의 벡터의 각각의 상응하는 복수의 요소를 사용하여 상관 모델을 복수의 벡터에 적용함으로써 복수의 벡터(카운트 행렬(402)의 세포 구성성분 벡터)를 사용하여(잠재 표현(404)의) 복수의 후보 세포 구성성분 모듈의 각각의 후보 세포 구성성분 모듈을 식별하였다. 특히, 상관-기반 비용 함수가 최적화되었고, 이는 세포 구성성분 벡터 사이의 이웃 관계를 정의하는 최근접-이웃 그래프를 계산하고, 카운트 행렬(402)의 세포 구성성분 벡터 사이의 상관을 계산하는 것에 해당한다. 복수의 세포에 걸쳐 서로 높은 상관을 갖는 세포 구성성분(여기서 유전자)은 최근접 이웃이 되는 것으로 귀결되었으며, 라이덴(Leiden) 또는 임의의 다른 그래프 클러스터링 방법을 사용하여 그래프를 클러스터링함으로써 잠재 표현(402) 내에 세포 구성성분 모듈을 형성하였다. 파이토치(pytorch) 또는 텐서플로우(tensorflow)에서 구현되는 표준 훈련을 사용하여, 그 가중치의 L0 정규화 및 재구성 손실로 1-계층 오토인코더를 훈련시키는 것에 해당하는 희소 오토인코더 비용 함수를 최적화한다. 본 예에서, 이는 108개의 세포 구성성분 모듈이 훈련 동안 학습되게 하였다. 즉, 도 4의 잠재 표현(404)은 108개의 세포 구성성분 모듈(132)을 가졌고, 각각은 그에 대한 발현 데이터가 카운트 행렬(402)에서 이용가능한 세포 구성성분의 독립적인 서브세트를 갖는다.

[0549] 108개의 세포 모듈 중에서, "모듈 78"로 명명된 세포 구성성분 모듈 132는 교란된 샘플 및 대조군 샘플에 걸쳐 계산된, 각각의 세포 구성성분에 대한 평균 t-점수일 때 가장 강한 활성화를 나타내었다. 다시 말해, 카운트 행렬 데이터 내의 발현 데이터는 잠재 표현(404) 내의 각각의 개별 세포 구성성분 모듈에 대해, 교란원에 노출된 세포와 교란원에 노출되지 않은 세포 사이에서 각각의 세포 구성성분 모듈의 각각의 세포 구성성분의 차등 발현에 대한 t-점수를 수행함으로써 세포 구성성분을 검증하기 위해 사용되었다. 또한, 모듈 78에는 지방산 및 지질 연관 생물학적 과정에 연루된 세포 구성성분이 풍부하다. 요약하면, 모듈 78은 대사 활성의 마커인 FABP3을 비롯한 28개의 유전자로 이루어진다.

[0550] 세포 구성성분 모듈 이외에, 훈련 화합물에 대한 세포의 노출시 세포 기반 세포 구성성분 반응 데이터가 필요하다.

[0551] 따라서, 하나 이상의 제2 데이터세트를 전자 형태로 획득하였다. 하나 이상의 제2 데이터세트는 제2 복수의 세포로부터의 데이터를 포함한다. 제2 복수의 세포는 20개 이상의 세포를 포함한다. 제2 복수의 세포는 집합적으로 관심 생리학적 조건을 알리는 복수의 공변량을 나타냈다. 예를 들어, 복수의 공변량은 일부 경우에 훈련 화합물이다. 이어서, 제2 복수의 세포의 각각의 세포에 대해, 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도가 획득되고, 이에 의해 (i) 제2 복수의 세포 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 세포 구성성분 카운트 데이터 구조가 획득된다.

[0552] 이는 제2 데이터세트가 전자 형태로 얻어지고, 제2 복수의 세포의 각각의 개별 세포에 대해(제2 복수의 세포는

20개 이상의 세포를 포함하고 관심있는 생리학적 조건을 알리는 복수의 공변량(여기서는 복수의 상이한 화학적 화합물)을 집합적으로 나타냄): 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해: 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하며, 이에 의해 (i) 제2 복수의 세포 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 세포 구성성분 카운트 데이터 구조를 획득하는 것을 설명하는 도 14c의 블록 1538에 따른다.

[0553] 이 카운트 행렬의 형태의 예시는 도 5의 카운트 행렬(502)이다. 도 5의 카운트 행렬(502)에 예시된 바와 같이, 각각의 개별 세포 구성성분(예를 들어, 유전자)에 대해, 제2 복수의 세포의 각각의 세포에 대한 발현 데이터가 존재한다. 예를 들어, 복수의 유전자 각각의 전사 활성은 제2 복수의 세포에 걸쳐 측정된다. 각각의 세포는 공변량, 여기서 훈련 화학적 화합물에 노출되었다.

[0554] 복수의 세포 구성성분 또는 그 표현을 공통 차원으로서 사용하여 세포 구성성분 카운트 데이터 구조 및 잠재 표현을 조합함으로써 활성화 데이터 구조를 형성하며, 여기서 활성화 데이터 구조는 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈에 대해, 제2 복수의 세포의 각각의 세포에 대해 각각의 활성화 가중치를 포함한다.

[0555] 카운트 행렬(502)에 잠재 표현(404)을 행렬 곱하여 도 5에 예시된 활성화 데이터 구조(504)를 획득하였다. 활성화 데이터 구조(504)는, 각각의 개별 세포 구성성분 모듈에 대해, 제2 복수의 세포의 각각의 세포에 대해, 활성화 값 Act_{K-G}를 가지며, 그의 값은 카운트 행렬(502)에 의한 잠재 표현(404)의 상응하는 행렬 곱셈에 의해 결정된다.

[0556] (i) 활성화 데이터 구조가 후보 모델 내로 입력되었을 때 활성화 데이터 구조 내에 표현된 각각의 세포 구성성분 모듈 내의 복수의 공변량 중 각각의 공변량의 부재 또는 존재의 예측과 (ii) 각각의 세포 구성성분 모듈의 각각의 공변량의 실제 부재 또는 존재 사이의 차이를 사용하여 후보 세포 구성성분 모델을 훈련시키고, 이때 훈련은 차이에 응답하여 후보 세포 구성성분 모델과 연관된 복수의 공변량 가중치를 조정한다.

[0557] 활성화 데이터 구조(502)는 도 6의 모델(601)에 대한 훈련 데이터(표지 데이터)로서 기능하였으며, 이는 그 자체가 차원 N 화합물 x M 세포 구성성분 모듈의 잠재 표현(602)이다. 본 예에서, 8000개의 상이한 화합물 및 108개의 세포 구성성분 모듈을 고려하였다. 따라서, 도 5의 명명법에서, Z는 108이고 G는 8000이었다. 활성화 데이터 구조를 2가지 방식으로 훈련 및 테스트 세트로 분할하였다. 먼저, 1200개의 화합물을 테스트 세트로, 나머지 6800개의 화합물을 훈련 세트로 그룹화하는 "랜덤 분할"을 선택하였다. 또한, "교차-스캐폴드 분할"은 테스트 세트가 훈련 세트와 상이한 스캐폴드를 갖는 화합물을 함유하는 것을 보장하는 오픈-소스 소프트웨어 패키지 RDKit 내의 기능을 사용하여 정의하였다.

[0558] 도 6에 예시된 바와 같이, 활성화 데이터 구조(504)의 각각의 개별 행은 화합물이 각각의 행에 의해 표현되는 상응하는 세포 구성성분 모듈의 세포 구성성분을 유도할 가능성을 나타내는 벡터이다. 모델(601)의 각각의 인스턴스는 활성화 데이터 구조(504)의 행에 대해 훈련되었다. 활성화 데이터 구조(504)는 6800 훈련 화합물을 사용하여 형성되었다. 주어진 모델(601)에 대해, 특정 화학적 화합물의 지문이 모델(601)에 입력되고, 이 입력에 응답하여, 상응하는 세포 구성성분 모듈에 대한 예측된 활성화 값이 계산된다. 본 예측된 활성화 값은 활성화 데이터 구조체(504) 내의 대응 요소 내의 실제 활성화 값에 직접 비교될 수 있다. 따라서, 이러한 방식으로, (i) 활성화 데이터 구조(504)를 모델(601)에 입력 시 활성화 데이터 구조(504)에 표현된 각각의 세포 구성성분 모듈에 대한 훈련 화합물 내의 각각의 화합물의 부재 또는 존재의 예측과 (ii) 각각의 세포 구성성분 모듈에 대한 각각의 화합물의 실제 부재 또는 존재 사이의 차이를 계산하고, 차이에 응답하여 후보 세포 구성성분 모델과 연관된 복수의 공변량 가중치(604)를 조정함으로써 모델(601)을 훈련시키는데 사용할 수 있다. 도 6에 예시된 바와 같이, 복수의 공변량 가중치는 복수의 세포 구성성분 모듈의 각각의 개별 세포 구성성분 모듈에 대해: 각각의 개별 공변량에 대해: 각각의 공변량이 활성화 데이터 구조에 걸쳐 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 상응하는 가중치를 포함한다. 일부 실시예에서, 각각의 세포 구성성분 모듈에 대해 상이한 모델(601)이 있었다. 다시 말해서, 도 6을 참조하면, 일부 실시예에서, 각각의 행(604)은 상이한 모델(601)에 있다. 따라서, 이러한 실시예에서, 각각의 이러한 모델(601)은 활성화 데이터 구조 내의 상응하는 행(예를 들어, 각각의 모델(601)과 동일한 세포 구성성분 모듈에 상응하는 행)을 사용하여 훈련된다.

[0559] 도 6에 예시된 바와 같이, 훈련된 모델(601)(또는 모델)은 각각의 공변량(여기서는 훈련 화학적 조성)에 대한 가중치를 제공한다. 즉, 모델(601)의 잠재 표현(602)은 얼마나 많은 각각의 공변량(화학적 조성)이 세포 구성성분 모듈의 활성화와 연관되는지를 설명하는 가중치(예를 들어, 도 6의 $weight_{1-1}$ 또는 행(604-1))를 제공한다. 이러한 가중치는 세포 구성성분 모듈의 세트의 각각의 화합물에 대한 각각의 세포 구성성분 모듈의 각각의 수치

활성화 점수로 간주된다. 상이한 모델(601)이 각각의 세포 구성성분 모듈에 대해 형성되는 실시예에서, 잠재 표현(602)은 각각의 모델(601)의 총 잠재 표현이다. 일부 실시예에서, 표현에서의 각각의 가중치는 범주형이다(예를 들어, 화합물이 세포 구성성분 모듈 "0"에 영향을 미치거나, 또는 화합물이 세포 구성성분 모듈 "1"에 영향을 미치지 않음). 다른 실시예에서, 각각의 가중치는 연속 척도 상에 있고, 이때 척도의 한쪽 끝부분은 훈련 화합물이 세포 구성성분 모듈에 크게 영향을 미친다는 것을 표시하고, 척도의 다른쪽 끝부분은 훈련 화합물이 세포 구성성분 모듈에 영향을 미치지 않는다는 것을 나타낸다. 본원에 사용된 용어 "영향"은 적용 의존적이지만, 일반적으로 화합물의 부재 또는 존재가 세포 구성성분 모듈에서 세포 구성성분의 풍부도를 변화시킴을 의미한다.

[0560] 모델(601)의 훈련을 위해, 본 예에서, 도 6의 활성화 데이터 구조(504)에 나타낸 화합물의 SMILES 표현은 ECFP4 지문 표현, 및 추가로 그래프 표현으로 변환된다. 이어서, 2개의 모델을 훈련시킨다. 즉, 모델(601)은 본 예에서 2개의 상이한 모델의 앙상블이다: A) 완전 연결 신경망 아키텍처가 ECFP4 표현 상에서 훈련하는 데 사용되고, B) 메시지 전달 신경망(MPNN)이 그래프 표현 상에서 훈련하는 데 사용된다. 오픈 소스 소프트웨어 패키지 파이토치 및 DGL을 사용하여 이 훈련을 수행하였다. 훈련되지 않은 모델(601)은 훈련 세트의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해, (i) 각각의 화합물의 화학 구조의 지문을 훈련되지 않은 모델로 입력 시 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수 및 (ii) 세포 구성성분 모듈의 세트(활성화 데이터 구조(504)로부터 획득됨) 내의 각각의 화합물에 대한 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련되고, 여기서 훈련은 차이에 응답하여 훈련되지 않은 모델(601)과 연관된 복수의 파라미터를 조정하고, 여기서 복수의 파라미터는 100개 이상의 파라미터를 포함하며, 이에 의해 훈련된 모델을 획득한다.

[0561] 상기 언급된 바와 같이, 본 예에서, 모델(601)은(i) SMILES 스트링의 표준 지문 상의 완전히 연결된 네트워크(여기서, 네트워크 아키텍처는 ReLU 활성화를 갖는 3-계층 네트워크임) 및 (ii) DGL 라이브러리로부터의 MPNN 네트워크의 앙상블이다. 화학적 구조 정보의 입력 시, 모델(601)은 훈련된 각각의 세포 구성성분 모듈(132)의 활성화 점수를 제공한다.

[0562] 실제로, 일부 실시예에서, 본 예에서 각각의 세포 구성성분 모듈에 대한 별개의 앙상블 모델(601)이 있다. 즉, 모델(601)은 화학 구조의 입력시 복수의 세포 구성성분 모듈 각각에 대한 개별 활성화 점수를 제공하는 멀티-태스크 인코더였다. 또한, 일부 실시예에서, 위에서 설명 바와 같이, 각각의 개별 세포 구성성분 모듈에 대한 별개의 모델(601)이 있다. 이러한 실시예에서, 각각의 이러한 각각의 모델(601)은 상응하는 세포 구성성분 모델과 관련하여 각각의 화합물에 대한 활성화 가중치를 포함한다.

[0563] 이제 훈련된 각각의 개별 모델(601)은, 훈련 세트의 일부이든 아니든, 임의의 화합물에 대한 그의 상응하는 세포 구성성분 모듈에 대한 활성화 점수를 제공한다. 즉, 각각의 모델(601)은 그의 상응하는 세포 구성성분 모듈이 테스트 화합물과 연관되는지 여부를 보고할 수 있다. 만약 그렇다면, 모델은 그의 상응하는 세포 구성성분 모듈이 테스트 화합물과 연관됨을 나타내는 점수를 출력한다. 일부 실시예에서, 이러한 점수는 범주형이다(예를 들어, 상응하는 세포 구성성분 모듈이 테스트 화합물과 연관되는 경우에는 "1"이고, 그렇지 않은 경우에는 "0"임). 일부 실시예에서, 상기 점수는 예를 들어, 0 내지 1의 척도 상의 확률 또는 가능성이고, 여기서 1에 가까운 수(예를 들어, 0.85)는 상응하는 세포 구성성분 모듈이 테스트 화합물과 연관될 가능성을 나타낸다. 일부 실시예에서, 이 점수는 "A" 내지 "B"의 연속 척도 상에 있으며, 여기서 A 및 B는 2개의 상이한 수이다. 각각 상이한 세포 구성성분 모듈에 상응하는 여러 모델(601)이 존재하기 때문에, 테스트 화합물을 여러 상이한 모델(601)에 대해 실행하여 어느 세포 구성성분 모듈이 화합물에 의해 활성화(연관)되는지를 결정한다. 각각의 인스턴스에서, 화학 구조는 위에서 설명 바와 같이 지문으로 전환되고, 이는 각각의 모델에 적용되는 지문이다. 생물학적 관점에서, 주어진 테스트 화합물은 임의의 수의 상이한 세포 구성성분 모듈(예를 들어, 1, 2, 3, 4, 5 또는 그 이상)을 활성화시킬 수 있는 것으로 예상될 수 있음을 유의한다. 또한, 본 개시에 설명된 접근법은 모델(601)이 훈련되지 않았지만, 어떤 세포 구성성분 모듈이 테스트 화합물에 의해 활성화되어야 하는지가 공지되어 있는 화합물을 테스트함으로써 검증될 수 있다. 이는 하기 제시된 바와 같이 본 예에서 수행되었다. 특히, 화합물을 생리학적 조건과 연관시키기 위한 훈련된 모델(601)은 본 예에서 4-중첩 검증되었다. 본 테스트는 원출원의 제1항을 따른다.

[0564] 먼저, 모델 601로부터의 모델 예측을 고처리량 스크린으로부터 앞서 설명된 1200개의 무작위로 선택된 보지 못했던 화합물에 의해, 및 6800개 화합물 훈련 세트에 대해 비-중첩 스캐폴드로 앞서 설명된 1200개 화합물에 의해 유도된 지방산 생성-관련 세포 구성성분 모듈의 활성화에 대해 획득하였다. 무작위로 선택된 화합물에 대해 획득된 각각의 모델(601) 예측(예측된 세포 구성성분 활성화 점수)을 도 10b에 예시한다. 즉, 도 10b는 2개의

상이한 모델(601)(하나는 세포 구성성분 모듈(78)에 대한 것이고- "모듈 78" -, 하나는 세포 구성성분 모듈 "90"에 대한 것임)로부터의 결과를 보여준다. 모듈 78은 세포 대사에 중요한 지방산-관련 세포 과정을 나타내고, 그의 상응하는 훈련된 모델 601은 높은 결정 계수(R^2 = 0.28)를 나타내었다. 대조적으로, 동일한 scRNA-seq 데이터세트로부터 학습된, 세포 대사와 무관한 세포 구성성분 "모듈 90"(모듈 90에서의 세포 구성성분은 지방산-관련 과정과 관련되지 않음)에 대한 훈련된 모델 601은 낮은 결정 계수(R^2 = 0.08)를 갖는다. 모든 벤치마크는 매우 유의한 상관을 산출하였다(피어슨 상관 계수 p_s = 각각 ~0.5 및 ~0.2).

[0565] 원 출원의 제1항의 언어에서, 이러한 제1 검증 접근법은 테스트 화학적 화합물(고처리량 스크린으로부터 설명된 1200개의 무작위로 선택된 보지 못했던 화합물 중 하나, 및 앞서 설명된 1200개의 화합물에 의해 6800 화합물 훈련 세트에 대해 비-중첩 스캐폴드를 가짐)을 관심 생리학적 조건(여기서, 본 예에서, 세포 대사에 중요한 지방산-관련 세포 과정)과 연관시키는 방법을 제공한다. 방법은, 메모리 및 하나 이상의 프로세서를 포함하는 컴퓨터 시스템에서, 테스트 화학적 화합물의 화학 구조의 지문을 획득하는 것을 포함한다. 따라서, 테스트 화학적 화합물의 화학 구조의 지문이 획득되고, 이는 본 예에서 도 1의 각각의 모델(601)에 입력된다. 원 출원의 제1항에 관련하여, 이 모델이 모델로 지칭된다. 이러한 모델은 앙상블 모델을 포괄하고, 앙상블 모델의 각각의 컴포넌트 모델은 도 6의 모델(601)에 대해 열거된 파라미터의 단일 행을 포함하고, 이러한 행은 컴포넌트 모델과 연관된 주어진 세포 구성성분 모듈에 대한 가중치에 대한 파라미터이다. 도 6에서 이러한 가중치가 단일 행으로 나타내어지지만, 이들은 앙상블 모델의 컴포넌트 모델에서의 행 포맷일 필요는 없으며, 그의 임의의 등가물이 본 개시의 범주 내에 있다는 것을 이해할 것이다. 또한, 도 6의 모델(601)은 회귀에 기초한 모델(601)에서 적합한, 그에 대해 훈련된 각각의 화합물에 대한 단일 가중치를 포함하지만, 일부 실시예에서 모델(601)에서의 가중치의 수와 모델에 대해 훈련된 화합물의 수 사이에 명확한 관계는 존재하지 않는다. 일부 실시예에서, 모델(601)은 100개 이상, 1000개 이상, 10,000개 이상, 또는 100,000개 이상의 파라미터를 포함한다.

[0566] 원 출원의 제1항에 따라, 테스트 화합물의 지문을 모델에 입력한다. 원 출원의 제1항에 설명된 바와 같이, 모델은 100개 이상의 파라미터를 포함한다. 즉, 테스트 화합물의 지문을 입력 시, 모델 출력의 계산은 머리로는 수행할 수 없다. 모델은 모델로의 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력한다. 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 세포 구성성분 모듈의 세트의 상응하는 세포 구성성분 모듈을 나타낸다. 본 예에서, 모델은 모델(601)의 앙상블이고, 각각은 상이한 세포 구성성분 모듈을 나타내고, 따라서 앙상블 내의 각각의 모델(601)은 세포 구성성분 모듈의 세트의 상응하는 단일 세포 구성성분 모듈을 나타내는 하나 이상의 계산된 활성화 점수 내의 계산된 활성화 점수를 출력한다. 이와 관련하여, 및 상기 언급된 바와 같이, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함한다. 또한, 세포 구성성분 모듈의 세트의 적어도 제1 세포 구성성분 모듈은 관심 생리학적 조건과 연관된다. 본 예에서, 모듈(78)은 관심 생리학적 조건과 연관된다. 도 10b에 예시된 바와 같이, 모듈 78을 정확하게 활성화시키고, 따라서, (예를 들어, 제1 임계치 기준을 충족하는 제1 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수에 의해) 모듈 78의 관심 생리학적 조건(세포 대사에 중요한 지방산-관련 세포 과정)과 연관된 화합물이 식별된다.

[0567] 청구된 접근법의 제2 검증으로서, 모듈 78 및 90에 대한 각각의 훈련된 모델을 이어서 훈련 동안 도 6의 모델 601에 도입되지 않은 또 다른 테스트 세트인 지방전구세포에 노출된 특정 소분자 "합성 히트"의 scRNA-seq 특성화에 적용하였다. 도 10d는 합성 히트에 의해 모듈 90에 대해 훈련된 모델 601에 의해 표시된 활성화가 거의 또는 전혀 없는 것과 비교하여, 합성 히트에 의해 모듈 78에 대해 훈련된 모델 601에 의해 표시된 활성화의 높은 상관 및 충실한 예측을 예시한다.

[0568] 셋째, 모듈 78에 대한 훈련된 모델 601을 사용하여, 공공 데이터베이스에서 5백만개의 화합물로부터 샘플링된 200,000개 화합물의 랜덤 서브세트에 대한 세포 구성성분 모듈 78(모듈 78)에 대한 세포 구성성분 활성화 점수를 예측하였다. 이로부터, 세포 구성성분 모듈(78)을 고도로 활성화시킬 것으로 예측되는 상위 50개의 화합물을 선택하고, 공지된 화합물(본원에서 공지된 피페리딘-함유 화합물("KPCC")로 지칭됨)의 화학 구조로부터 유도된 합성 히트 유사체 및 LINCS L1000 데이터세트로부터의 화합물을 포함하는 데이터베이스에서의 화합물 세트와 비교하였다. 이러한 비교의 분포가 도 10e에서 예시된다. 분포의 꼬리 끝에서, 세포 구성성분 모듈(78)에 대한 훈련된 모델(601)에 대해 획득된 예측은 LINCS 및 합성 히트에서 모든 화합물을 유의하게 초과하는 화합물을 식별하였다. 이러한 접근법은 특정한 목적하는 세포 과정에 대해 화학 구조를 최적화하는 방법을 강조한다.

[0569] 네번째로, 상위 50개 예측에서 식별된 화학 구조를 시각적으로 검사하였고, 공지된 지방 조직-표적화 약물작용 발생단을 나타내는 명백한 화학 구조를 함유하는 것으로 밝혀졌고, 따라서 모듈 78과 연관된 세포 구성성분 모

듈을 올바르게 활성화한다.

[0570] 본 제1 예는 또한 원 출원의 제58항을 따른다. 제1항과 제58항의 차이는 교란 시그니처 대 세포 구성성분 모듈 중 하나이다. 교란에 적용된 세포의 발현을 교란에 적용되지 않은 세포에 대해 비교함으로써 교란 시그니처가 획득된다. 따라서, 지방전구세포에서 대사적 활성 과정을 유도하는 것으로 공지된 소분자 교란원이 사용될 수 있다. 지방전구세포 세포주를 24시간 동안 교란원에 노출시키고, scRNA-seq 판독을 교란 및 대조군 조건에 대해 얻을 수 있다. 이로부터, 교란 시그니처를 얻을 수 있다. 대안적으로, 별개의 교란 시그니처는 제2 데이터세트에 사용된 화학적 공변량 중 어느 하나에 노출된 세포의 세포 발현을 비교함으로써 획득될 수 있다. 실제로, 제2 데이터세트에 사용된 각각의 화학적 공변량에 대해 이러한 방식으로 별개의 교란 시그니처가 획득될 수 있다. 각각의 그러한 교란 시그니처는 각각의 그러한 가중치가 이제 이진 척도가 아닌 연속적인 척도 상에 있는 것을 제외하고는 잠재 표현(404) 내의 행의 형태를 갖는다. 예를 들어, 일부 실시예에서 각각의 가중치는 0 내지 1(또는 일부 다른 범위 "A" 내지 "B", 여기서 A 및 B는 2개의 상이한 수, 예컨대 -100 및 100임)의 연속 척도 상의 값이다. 그로부터, 훈련 과정은 잠재 표현(404), 카운트 행렬(502), 활성화 데이터 구조, 및 컴포넌트 모델(601)의 훈련의 사용에 관하여 위에서 설명한 것과 동일하고, 여기서 각각의 그러한 모델은 이제 교란 시그니처의 세트의 상이한 교란 시그니처를 나타낸다.

[0571] **예 2. 태아 적혈구생성 프로그램을 활성화시키고 T-세포 소진을 차단하기 위한 화학 구조의 예측.**

[0572] 2개의 추가의 예에서, 태아 적혈구생성 및 T 세포 소진에 관련된 2개의 scRNA-seq 데이터세트에 대한 2개의 모델을 훈련시켰다.

[0573] 태아 적혈구생성을 위해, CD34 조혈 줄기 세포를 도구 화합물 CLT-AAA-12로 처리하였으며, 이에 대해 태아 적혈구생성의 종점 마커, 특히 유동 세포측정법에 의한 판독으로서 검정에서 F 세포의 수가 유도된다는 것이 이전에 확립되었다.

[0574] T 세포 소진을 위해, 나이브 T 세포를 소진 유도 배지로 처리하였다.

[0575] 두 세포 시스템 모두가 scRNA-seq로 특성화된다. 후속적으로, 약물 반사체 모델(본원에 참조로 포함되는, 2019년 7월 15일에 출원된 발명의 명칭 "세포 분석 방법"의 미국 특허 출원 번호 16/511,691 참조)을, 교란된 세포 대 대조군 세포에 의해 정의된 세포 상태 전이를 그의 각각의 샘플에 입력함으로써 scRNA-seq 데이터세트에 적용하였다. 약물 반사체는 약물 반사체 잠재 표현에서 각각의 8000개 화합물에 대한 세포 상태 활성화 점수를 할당한다. 이는 둘 모두의 전이(태아 헤모글로빈 및 T 세포 소진)에 대한 세포 상태 활성화 점수가 있는 2개의 벡터를 생성한다. 이들 2개의 벡터는 모델(601)에 대한 훈련 데이터로서의 역할을 한다.

[0576] 이 모델을 사용하여 T-세포 소진 및 조혈 줄기 세포에서 태아 적혈구생성을 활성화시키는 화합물을 예측하였다. 조혈 줄기 세포에서의 태아 적혈구생성은, 최근 수년간 겸상 적혈구 질환에 대한 돌발 CRISPR 요법으로 이어진 세포 과정인 반면, T-세포 소진은 암에 대한 체크포인트 억제제 요법의 보다 넓은 성공을 방지하는 주요 메커니즘이다.

[0577] 공공 데이터베이스에서 5백만개의 화합물로부터 샘플링된 2,000개의 화합물의 서브세트를 사용하여 예측을 수행하였으며, 여기서 서브세트는 무작위로 또는 스캐폴드 상에서 분할되었다. 도 11의 상부 패널은 조혈 줄기 세포에서의 태아 적혈구생성과 관련된 히트 화합물 CLT-AAA-12의 교란 시그니처와 샘플링된 화합물의 유의한 R^2 및 상관 계수 p_s를 보여주는, 무작위로 분할된 2,000개의 화합물의 테스트 세트 및 스캐폴드에 대한 본 예의 모델의 성능을 보여준다. 도 11의 하부 패널은 무작위로 및 스캐폴드 상에서 분할된 2,000개의 화합물의 테스트 세트의 성능을 보여주며, 이는 T-세포 소진과 관련된 세포 전이 시그니처를 갖는 샘플링된 화합물의 유의한 R^2 및 상관 계수 p_s를 보여준다. 따라서, 도 11은 모델(601)이 관심 교란 시그니처 및/또는 세포 전이 시그니처와 동일한 세포 거동 효과를 유도하는 새로운 스캐폴드를 예측할 수 있음을 입증한다.

[0578] **예 3. 질환-임계적 세포 거동에 기초한 특징 속성: 새로운 분자의 설계를 위한 약물작용발생단의 예측.**

[0579] 예 1에 설명된 바와 같이, 본원에 개시된 시스템 및 방법에 따라 예측된 화학 구조는 관심 생리학적 조건(예를 들어, 지방 조직-표적화)과 잠재적으로 관련된 분자 특징, 예컨대 약물작용발생단을 식별하는데 사용될 수 있다. 예 1에서와 같이, 이들 약물작용발생단은 공지된 화학 구조에 의해 검증될 수 있거나, 또는 추가의 검증을 위한 신규 구조를 제시할 수 있다. 예를 들어, 약물작용발생단에 기초한 알고리즘의 예시적인 사용 사례는 생물등입체적 교환가능 대체 염기(BoBER, Base of Bioisosterically Exchangeable Replacements) 데이터베이스

를 비롯한, 이전에 문헌에 설명된 기능적 의미를 갖는 약물작용발생단의 데이터베이스를 활용하는 것을 포함한다. 사용 사례의 또 다른 예는 교란에 대한 시스템의 복합 반응에서 식별된 약물작용발생단의 역할에 관한 직관을 얻기 위해 전문 지식을, 예컨대 의약 화학자에 의해 적용하는 것을 포함한다.

[0580] 새로운 분자의 설계를 위한 약물작용발생단을 예측하기 위한 모델을 수행하였으며, 여기서 모델은 약물작용발생단이 화학 구조에 함유되었는지 여부를 나타내는 표현을 달성하기 위해 터버스키 유사성을 사용하는 점수에 기초하여 선택된 개입 라이브러리로부터의 소분자의 특징화를 포함하였다. 이러한 표현(화학적 지문)을 예 1의 모듈 78에 대한 모델 601에 입력하였다. 예 1에서 식별된 지방-표적화 약물작용발생단을 사용하여, 예 1의 모듈 78에 대한 모델을 사용하여, 공지된 피페리딘-함유 화합물("KPCC")의 지방-표적화 약물작용발생단의 연관을 결정하였고, 단리물에서 0.04064 내지 0.04633 범위의 활성화 점수로 지방산 모듈의 전사 활성화가 관찰되었다.

[0581] *예 4. 잠재적인 세포 거동에 기초한 합성 히트 화합물의 생성.*

[0582] 테스트 사례로서, 본원에서 "6개의 합성 히트"로 지칭되는 새로 합성된 소분자 히트 중 6개를 시험관내 및 생체내 검증된 지방세포 갈색지방 화합물 및 그의 잠재적인 공간 표현에 기초하여 설계하였다. 각각의 6개의 합성 히트는 인간 지방전구세포에 대해 목적하는 세포 거동 변화를 도출하였다. 먼저, KPCC 클러스터의 약물작용발생단을 식별하였다. 이어서, 분자를 신규 생체-동배체의 혼입과 함께 이 클러스터에서 그 약물작용발생단의 풍부화에 의해 설계하여, 6개의 합성 히트의 최종 설계로 이어졌다. 이들 6개의 구조적으로 다양한 합성 히트에 대한 목표는 KPCC를 포함한 공지된 화학 물질(KCE)과 동일한 세포 거동 효과를 유도하는 것이었다. 도 13의 개략도에 의해 예시된 바와 같이, 세포 거동 효과는 인간 지방전구세포를 1 μM KPCC 및 6개의 합성 히트로 24시간 동안 처리하고, scRNA-seq를 사용하여 유전자 발현을 측정하고, 상기 예 1에 설명된 지방 대사 유전자 모듈(모듈 78)에서의 변화에 의해 표현된 세포 반응을 평가함으로써 결정하였다. 예를 들어, 지방 대사 모듈에서의 유전자는 특히 FABP3, FDPS 및 LPIN1을 포함하였다.

[0583] 지방전구세포에 대한 이들 화합물의 영향의 평가는 각각의 합성 히트가 KPCC와 동일한 지방 대사 유전자 모듈을 활성화시켰음을 밝혀내었다(도 13; 박스 1302에 강조된 모듈 78). 즉, 강조된 박스(1302)는 화합물의 지문을 도 13의 차트의 Y-축 상에 열거된 모델에 입력 시 예 1의 모듈(78)에 대한 모델에 의해 출력되는 활성화 점수를 나타낸다. 이들 결과는 목적하는 세포 거동을 예측가능하게 표적화하는 모델 플랫폼에 기초하여 합성 히트를 생성하는 능력에서 높은 신뢰도를 제공한다. 특히, 본 개시의 모델(601)(예를 들어, 예 1의 모듈 78에 대한 모델(601))은 고처리량 스크리닝, 분자 표적-기반 식별 또는 최적화, 또는 검증을 위한 수백 또는 수천개의 새로운 화합물의 합성에 대한 필요 없이 생리학적 조건과 연관된 유전자 모듈을 표적화하는 합성 히트를 예측하는데 사용될 수 있다.

[0584] **인용된 참고문헌 및 대안적 실시예**

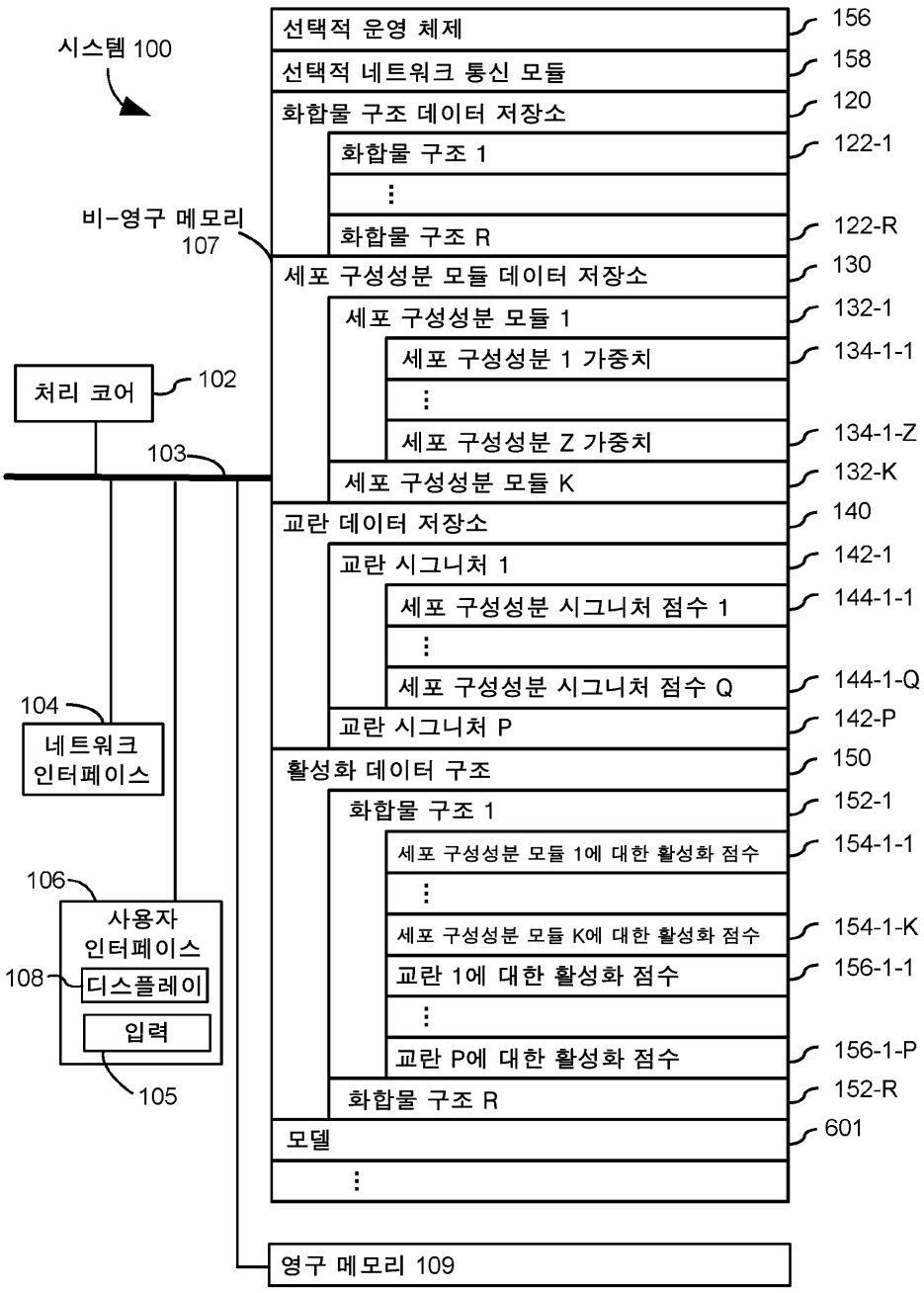
[0585] 본원에 인용된 모든 참고문헌은, 각각의 개별 공보 또는 특허 또는 특허 출원이 모든 목적을 위해 그 전문이 참조로 포함되는 것으로 구체적으로 및 개별적으로 표시된 것과 동일한 정도로 모든 목적을 위해 그 전문이 본원에 참조로 포함된다.

[0586] 본 발명은 비-일시적 컴퓨터 판독가능 저장 매체에 내장된 컴퓨터 프로그램 메커니즘을 포함하는 컴퓨터 프로그램 제품으로서 구현될 수 있다. 예를 들어, 컴퓨터 프로그램 제품은 도 1-3 및 7-9의 임의의 조합에 나타난 프로그램 모듈을 함유할 수 있다. 이들 프로그램 모듈은 CD-ROM, DVD, 자기 디스크 저장 제품, 또는 임의의 다른 비-일시적 컴퓨터 판독가능 데이터 또는 프로그램 저장 제품 상에 저장될 수 있다.

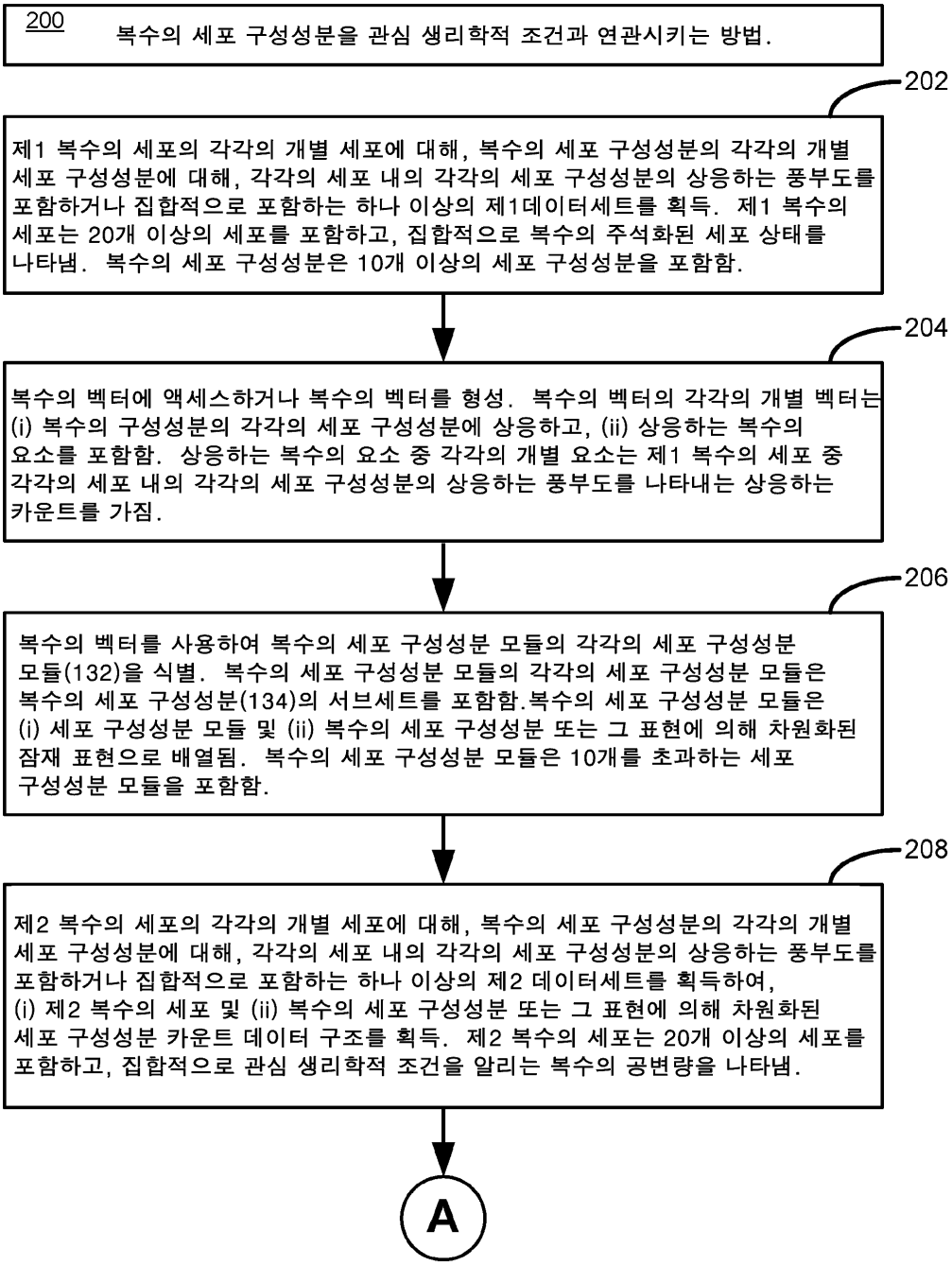
[0587] 관련 기술분야의 통상의 기술자에게 명백할 바와 같이, 본 발명의 많은 수정 및 변형이 그의 취지 및 범주로부터 벗어나지 않으면서 이루어질 수 있다. 본원에 설명된 구체적 실시예는 단지 예로서 제공된다. 실시예는 본 발명의 원리 및 그의 실제 응용을 가장 잘 설명하기 위해 선택되고 기재되었으며, 이로써 관련 기술분야의 다른 통상의 기술자가 고려된 특정한 용도에 적합한 바와 같은 다양한 수정을 갖는 본 발명 및 다양한 실시예를 가장 잘 이용할 수 있게 한다. 본 발명은 첨부된 청구범위의 자격이 있는 등가물의 전체 범주와 함께, 첨부된 청구범위의 관점에 의해서만 제한된다.

도면

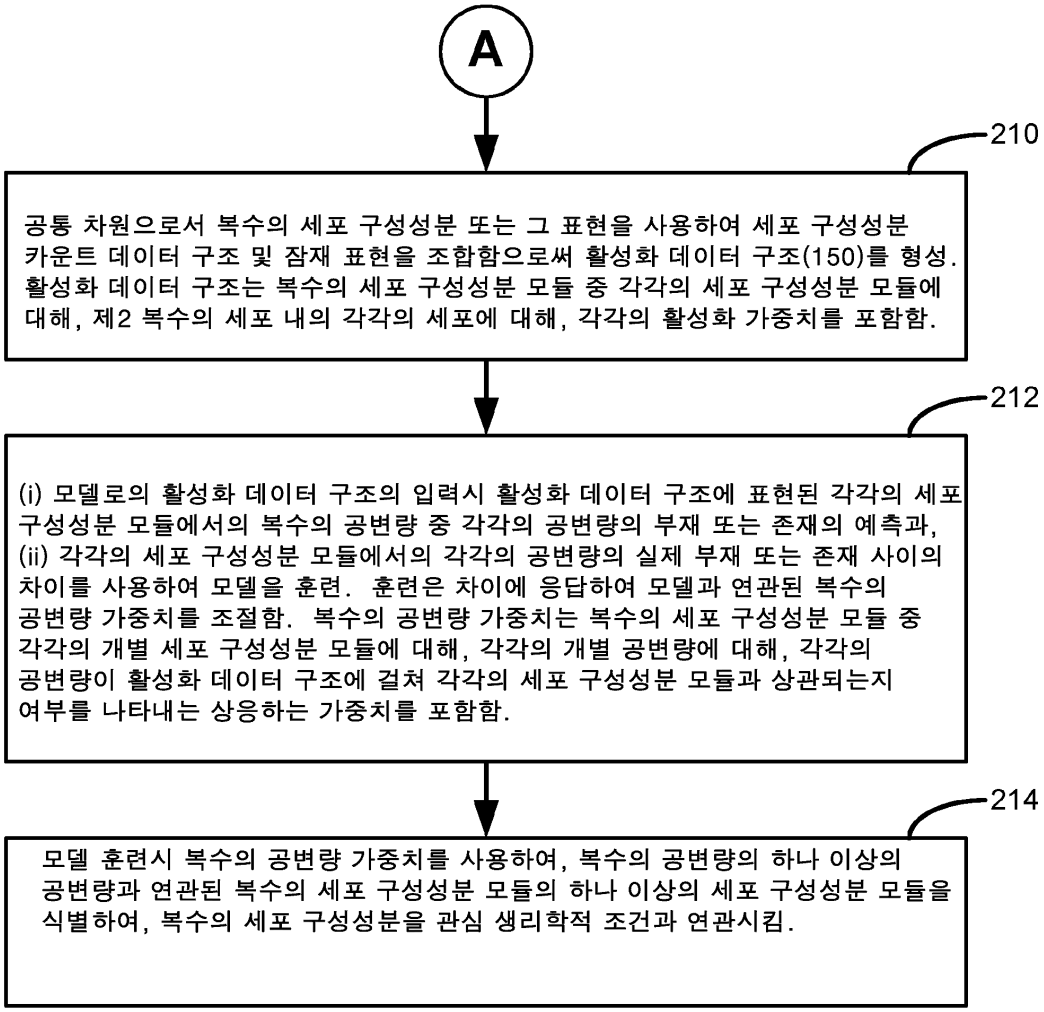
도면1


시스템 100
비-영구 메모리
107
처리 코어
102
103
104
네트워크
인터페이스
106
사용자
인터페이스
108
디스플레이
입력
105
선택적 운영 체제
156
선택적 네트워크 통신 모듈
158
화합물 구조 데이터 저장소
120
화합물 구조 1
122-1
화합물 구조 R
122-R
세포 구성성분 모듈 데이터 저장소
130
세포 구성성분 모듈 1
132-1
세포 구성성분 1 가중치
134-1-1
세포 구성성분 Z 가중치
134-1-Z
세포 구성성분 모듈 K
132-K
교란 데이터 저장소
140
교란 시그니처 1
142-1
세포 구성성분 시그니처 점수 1
144-1-1
세포 구성성분 시그니처 점수 Q
144-1-Q
교란 시그니처 P
142-P
활성화 데이터 구조
150
화합물 구조 1
152-1
세포 구성성분 모듈 1에 대한 활성화 점수
154-1-1
세포 구성성분 모듈 K에 대한 활성화 점수
154-1-K
교란 1에 대한 활성화 점수
156-1-1
교란 P에 대한 활성화 점수
156-1-P
화합물 구조 R
152-R
모델
601
영구 메모리 109

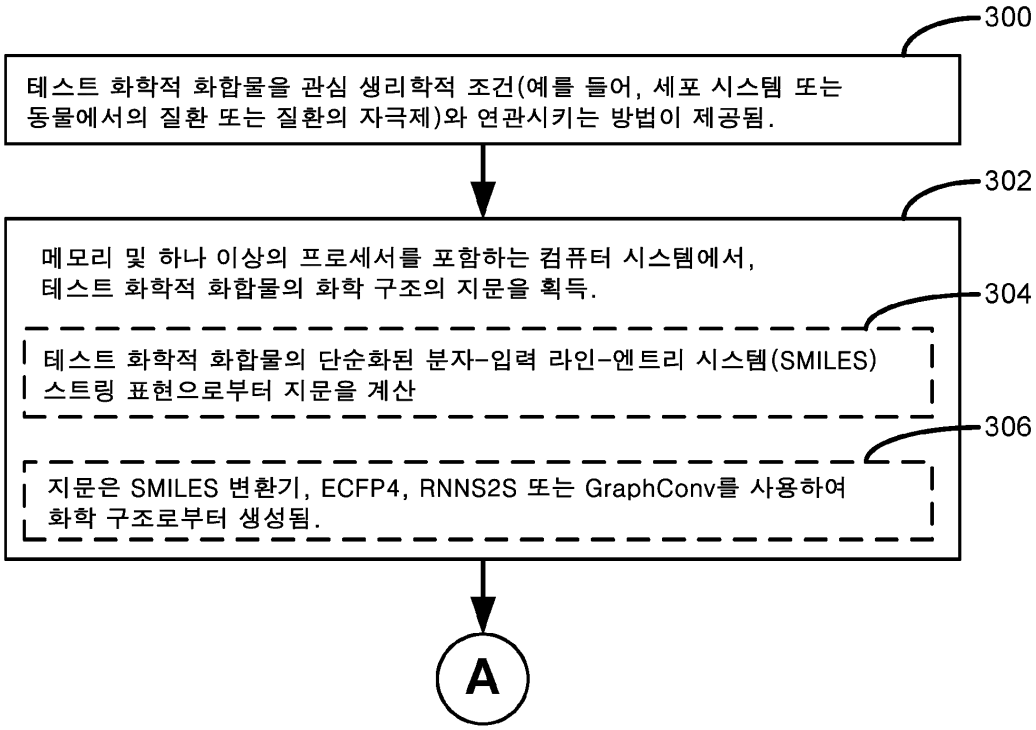
도면2a


200
복수의 세포 구성성분을 관심 생리학적 조건과 연관시키는 방법.
202
제1 복수의 세포의 각각의 개별 세포에 대해, 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 내의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하는 하나 이상의 제1데이터세트를 획득. 제1 복수의 세포는 20개 이상의 세포를 포함하고, 집합적으로 복수의 주석화된 세포 상태를 나타냄. 복수의 세포 구성성분은 10개 이상의 세포 구성성분을 포함함.
204
복수의 벡터에 액세스하거나 복수의 벡터를 형성. 복수의 벡터의 각각의 개별 벡터는 (i) 복수의 구성성분의 각각의 세포 구성성분에 상응하고, (ii) 상응하는 복수의 요소를 포함함. 상응하는 복수의 요소 중 각각의 개별 요소는 제1 복수의 세포 중 각각의 세포 내의 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 가짐.
206
복수의 벡터를 사용하여 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈(132)을 식별. 복수의 세포 구성성분 모듈의 각각의 세포 구성성분 모듈은 복수의 세포 구성성분(134)의 서브세트를 포함함.복수의 세포 구성성분 모듈은 (i) 세포 구성성분 모듈 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현으로 배열됨. 복수의 세포 구성성분 모듈은 10개를 초과하는 세포 구성성분 모듈을 포함함.
208
제2 복수의 세포의 각각의 개별 세포에 대해, 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 내의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하는 하나 이상의 제2 데이터세트를 획득하여, (i) 제2 복수의 세포 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 세포 구성성분 카운트 데이터 구조를 획득. 제2 복수의 세포는 20개 이상의 세포를 포함하고, 집합적으로 관심 생리학적 조건을 알리는 복수의 공변량을 나타냄.
A

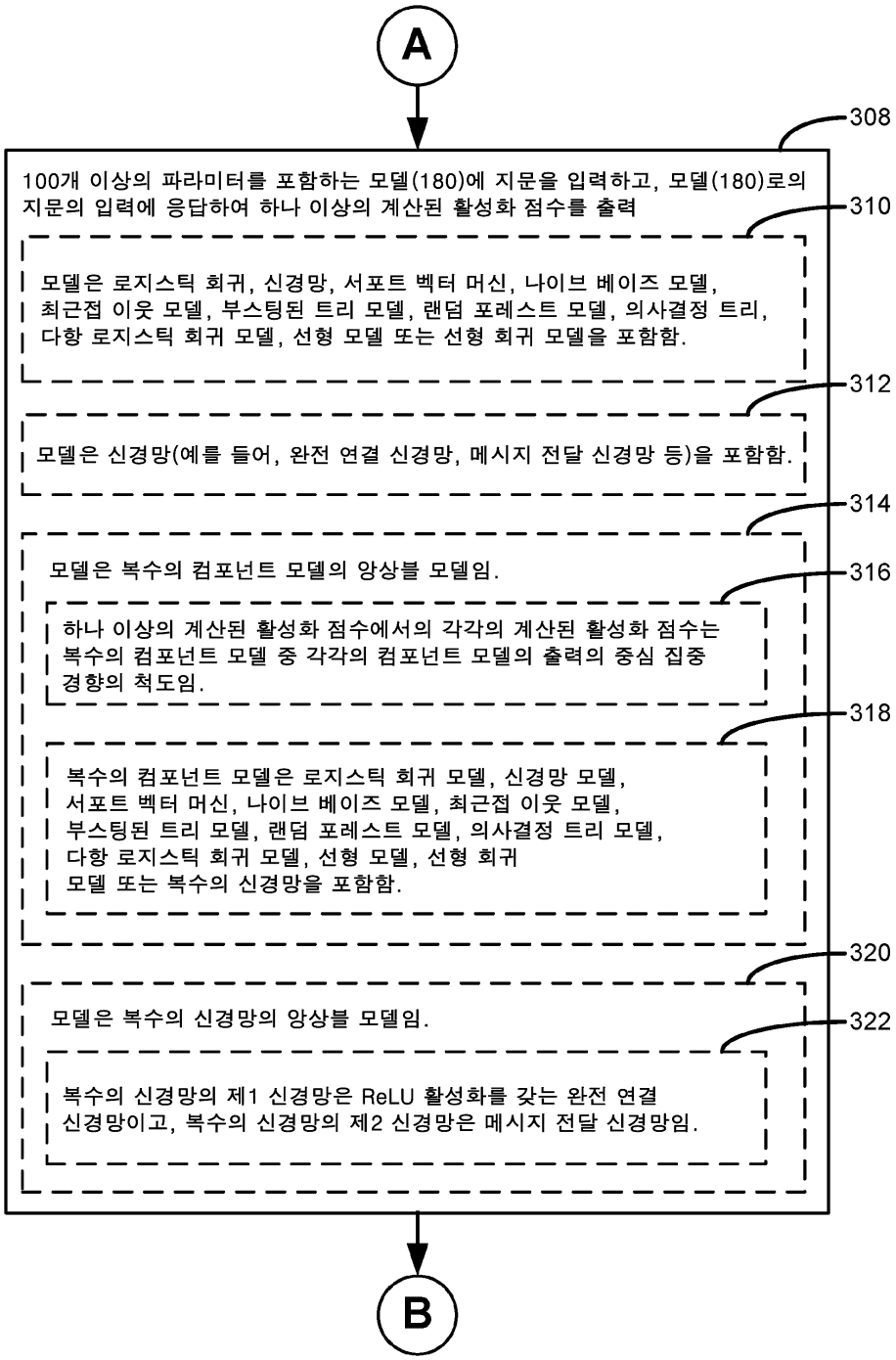
도면2b


A
210
공통 차원으로서 복수의 세포 구성성분 또는 그 표현을 사용하여 세포 구성성분 카운트 데이터 구조 및 잠재 표현을 조합함으로써 활성화 데이터 구조(150)를 형성. 활성화 데이터 구조는 복수의 세포 구성성분 모듈 중 각각의 세포 구성성분 모듈에 대해, 제2 복수의 세포 내의 각각의 세포에 대해, 각각의 활성화 가중치를 포함함.
212
(i) 모델로의 활성화 데이터 구조의 입력시 활성화 데이터 구조에 표현된 각각의 세포 구성성분 모듈에서의 복수의 공변량 중 각각의 공변량의 부재 또는 존재의 예측과, (ii) 각각의 세포 구성성분 모듈에서의 각각의 공변량의 실제 부재 또는 존재 사이의 차이를 사용하여 모델을 훈련. 훈련은 차이에 응답하여 모델과 연관된 복수의 공변량 가중치를 조절함. 복수의 공변량 가중치는 복수의 세포 구성성분 모듈 중 각각의 개별 세포 구성성분 모듈에 대해, 각각의 개별 공변량에 대해, 각각의 공변량이 활성화 데이터 구조에 걸쳐 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 상응하는 가중치를 포함함.
214
모델 훈련시 복수의 공변량 가중치를 사용하여, 복수의 공변량의 하나 이상의 공변량과 연관된 복수의 세포 구성성분 모듈의 하나 이상의 세포 구성성분 모듈을 식별하여, 복수의 세포 구성성분을 관심 생리학적 조건과 연관시킴.

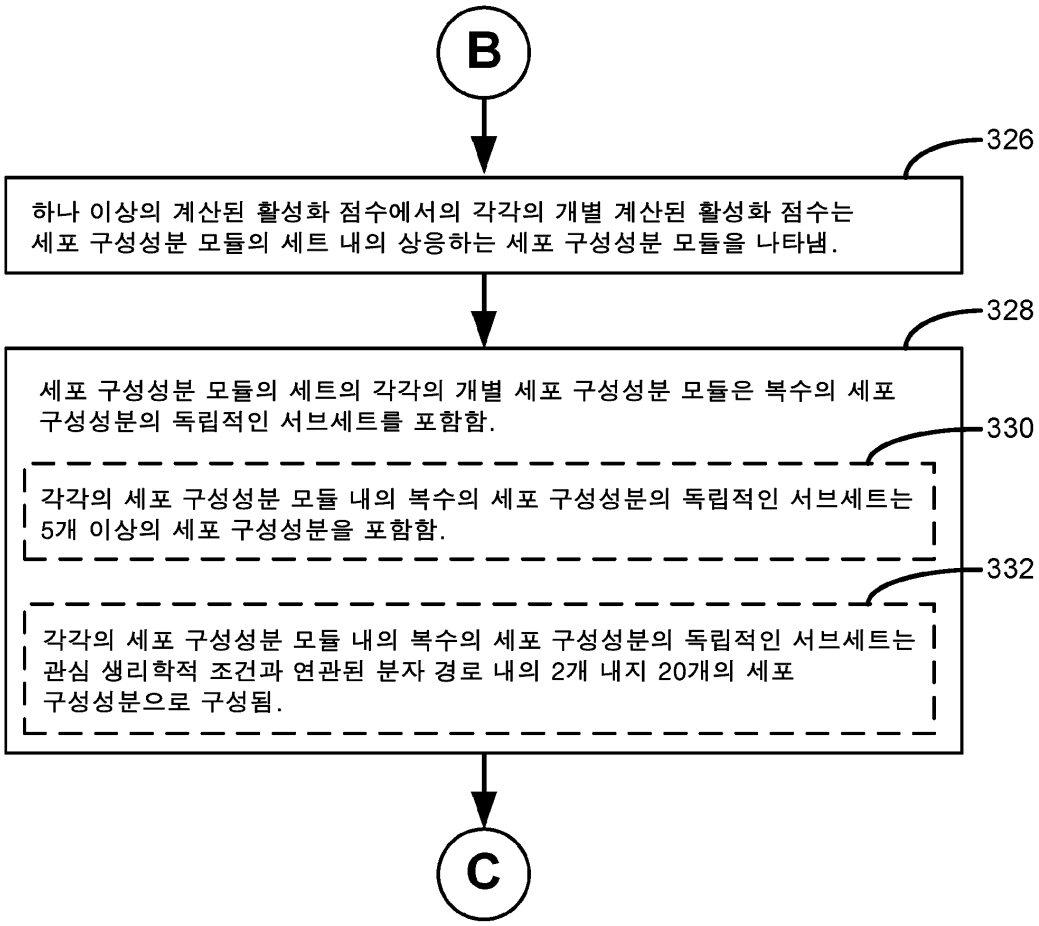
도면3a


300
테스트 화학적 화합물을 관심 생리학적 조건(예를 들어, 세포 시스템 또는 동물에서의 질환 또는 질환의 자극제)와 연관시키는 방법이 제공됨.
302
메모리 및 하나 이상의 프로세서를 포함하는 컴퓨터 시스템에서, 테스트 화학적 화합물의 화학 구조의 지문을 획득.
304
테스트 화학적 화합물의 단순화된 분자-입력 라인-엔트리 시스템(SMILES) 스트링 표현으로부터 지문을 계산
306
지문은 SMILES 변환기, ECFP4, RNNS2S 또는 GraphConv를 사용하여 화학 구조로부터 생성됨.
A

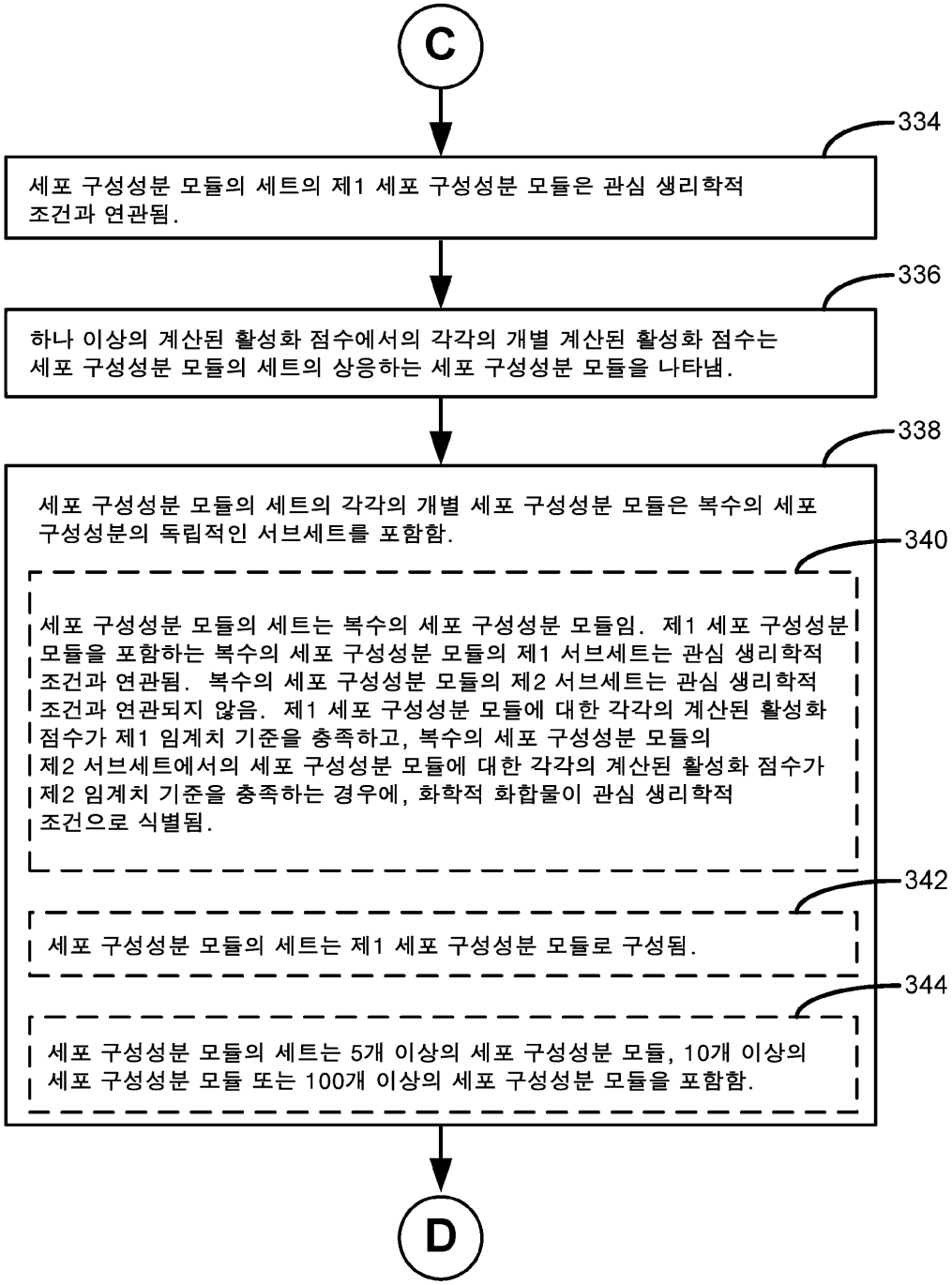
도면3b


A
308
100개 이상의 파라미터를 포함하는 모델(180)에 지문을 입력하고, 모델(180)로의 지문의 입력에 응답하여 하나 이상의 계산된 활성화 점수를 출력
310
모델은 로지스틱 회귀, 신경망, 서포트 벡터 머신, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리, 다항 로지스틱 회귀 모델, 선형 모델 또는 선형 회귀 모델을 포함함.
312
모델은 신경망(예를 들어, 완전 연결 신경망, 메시지 전달 신경망 등)을 포함함.
314
모델은 복수의 컴포넌트 모델의 앙상블 모델임.
316
하나 이상의 계산된 활성화 점수에서의 각각의 계산된 활성화 점수는 복수의 컴포넌트 모델 중 각각의 컴포넌트 모델의 출력의 중심 집중 경향의 척도임.
318
복수의 컴포넌트 모델은 로지스틱 회귀 모델, 신경망 모델, 서포트 벡터 머신, 나이브 베이즈 모델, 최근접 이웃 모델, 부스팅된 트리 모델, 랜덤 포레스트 모델, 의사결정 트리 모델, 다항 로지스틱 회귀 모델, 선형 모델, 선형 회귀 모델 또는 복수의 신경망을 포함함.
320
모델은 복수의 신경망의 앙상블 모델임.
322
복수의 신경망의 제1 신경망은 ReLU 활성화를 갖는 완전 연결 신경망이고, 복수의 신경망의 제2 신경망은 메시지 전달 신경망임.
B

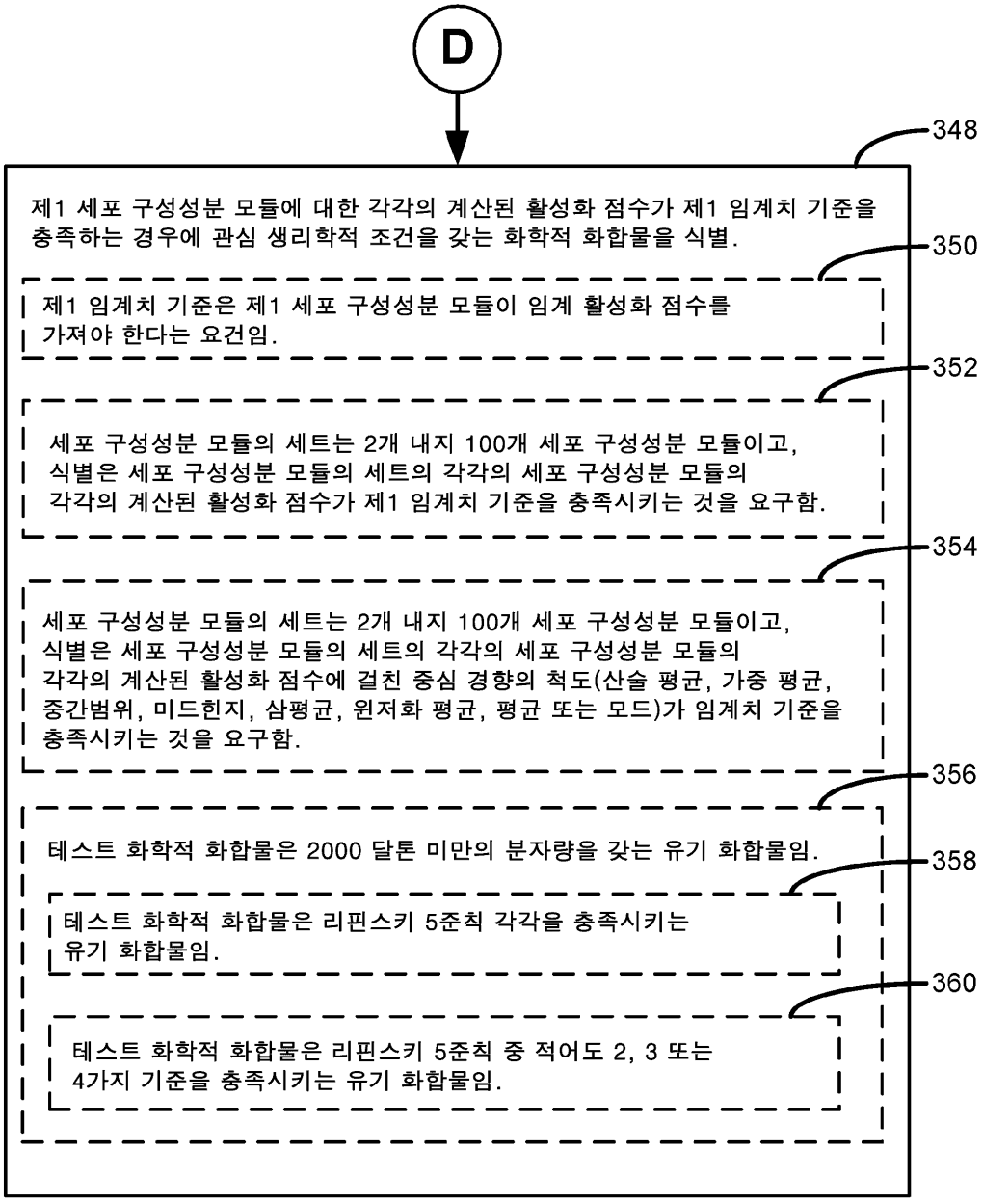
도면3c


B
326
하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 세포 구성성분 모듈의 세트 내의 상응하는 세포 구성성분 모듈을 나타냄.
328
세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함함.
330
각각의 세포 구성성분 모듈 내의 복수의 세포 구성성분의 독립적인 서브세트는 5개 이상의 세포 구성성분을 포함함.
332
각각의 세포 구성성분 모듈 내의 복수의 세포 구성성분의 독립적인 서브세트는 관심 생리학적 조건과 연관된 분자 경로 내의 2개 내지 20개의 세포 구성성분으로 구성됨.
C

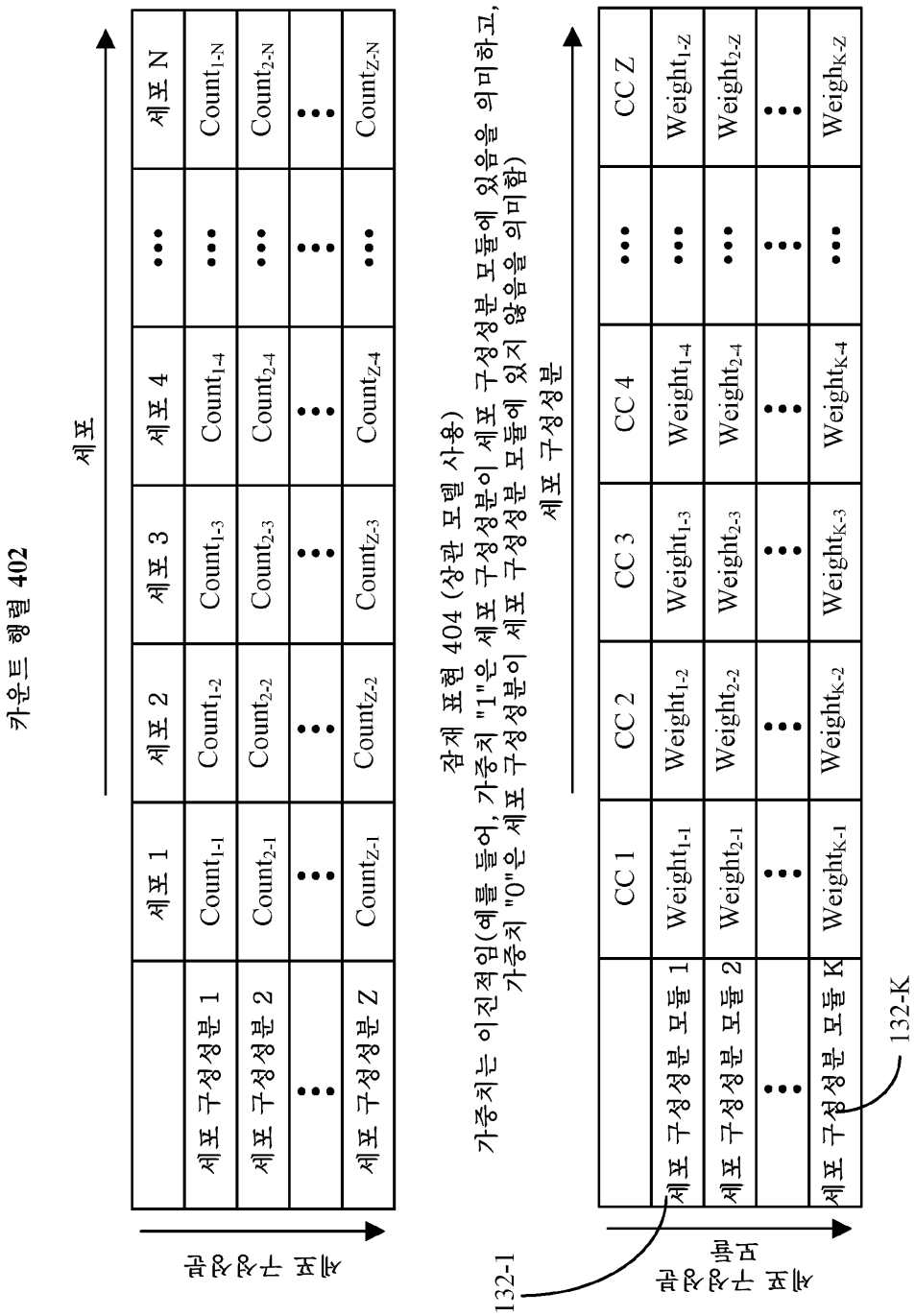
도면3d


C
334
세포 구성성분 모듈의 세트의 제1 세포 구성성분 모듈은 관심 생리학적 조건과 연관됨.
336
하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 세포 구성성분 모듈의 세트의 상응하는 세포 구성성분 모듈을 나타냄.
338
세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분의 독립적인 서브세트를 포함함.
340
세포 구성성분 모듈의 세트는 복수의 세포 구성성분 모듈임. 제1 세포 구성성분 모듈을 포함하는 복수의 세포 구성성분 모듈의 제1 서브세트는 관심 생리학적 조건과 연관됨. 복수의 세포 구성성분 모듈의 제2 서브세트는 관심 생리학적 조건과 연관되지 않음. 제1 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족하고, 복수의 세포 구성성분 모듈의 제2 서브세트에서의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 제2 임계치 기준을 충족하는 경우에, 화학적 화합물이 관심 생리학적 조건으로 식별됨.
342
세포 구성성분 모듈의 세트는 제1 세포 구성성분 모듈로 구성됨.
344
세포 구성성분 모듈의 세트는 5개 이상의 세포 구성성분 모듈, 10개 이상의 세포 구성성분 모듈 또는 100개 이상의 세포 구성성분 모듈을 포함함.
D

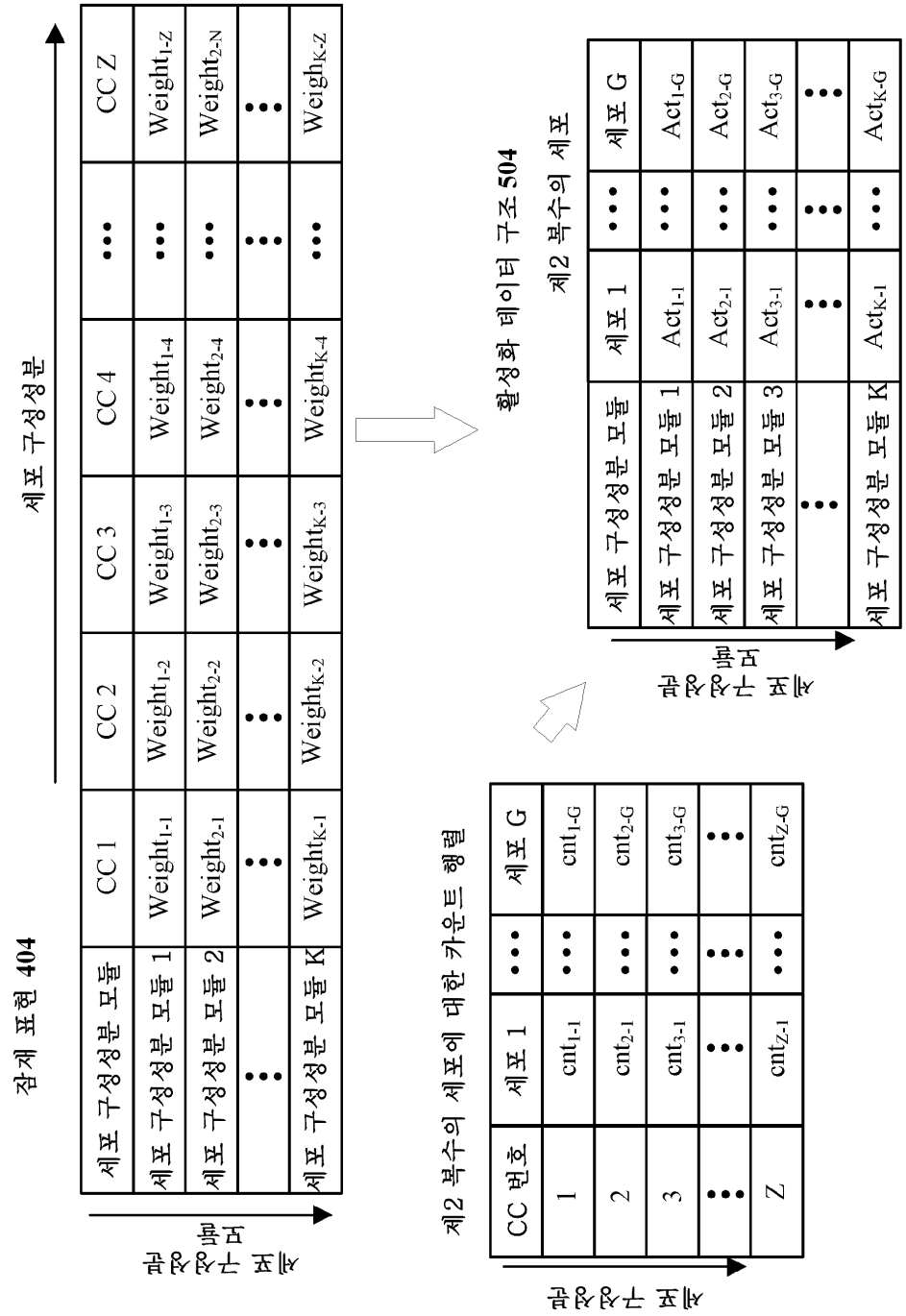
도면3e


D
348
제1 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족하는 경우에 관심 생리학적 조건을 갖는 화학적 화합물을 식별.
350
제1 임계치 기준은 제1 세포 구성성분 모듈이 임계 활성화 점수를 가져야 한다는 요건임.
352
세포 구성성분 모듈의 세트는 2개 내지 100개 세포 구성성분 모듈이고, 식별은 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족시키는 것을 요구함.
354
세포 구성성분 모듈의 세트는 2개 내지 100개 세포 구성성분 모듈이고, 식별은 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈의 각각의 계산된 활성화 점수에 걸친 중심 경향의 척도(산술 평균, 가중 평균, 중간범위, 미드힌지, 삼평균, 윈저화 평균, 평균 또는 모드)가 임계치 기준을 충족시키는 것을 요구함.
356
테스트 화학적 화합물은 2000 달톤 미만의 분자량을 갖는 유기 화합물임.
358
테스트 화학적 화합물은 리핀스키 5준칙 각각을 충족시키는 유기 화합물임.
360
테스트 화학적 화합물은 리핀스키 5준칙 중 적어도 2, 3 또는 4가지 기준을 충족시키는 유기 화합물임.

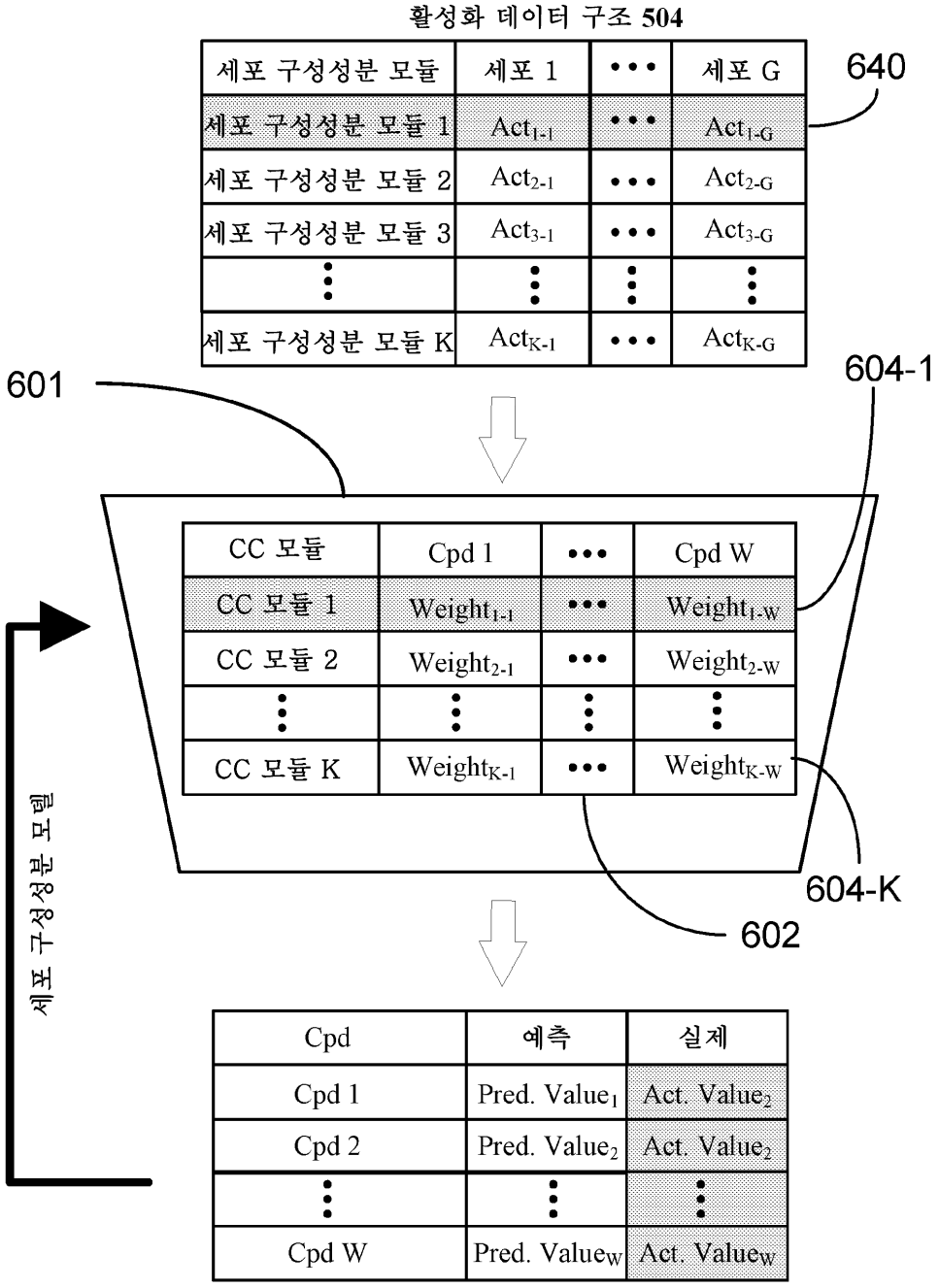
도면4


카운트 행렬 402
세포
세포 구성성분
세포 1
세포 2
세포 3
세포 4
세포 N
세포 구성성분 1
세포 구성성분 2
세포 구성성분 Z
Count1-1
Count1-2
Count1-3
Count1-4
Count1-N
Count2-1
Count2-2
Count2-3
Count2-4
Count2-N
CountZ-1
CountZ-2
CountZ-3
CountZ-4
CountZ-N
잠재 표현 404 (상관 모델 사용)
가중치는 이진적임(예를 들어, 가중치 "1"은 세포 구성성분이 세포 구성성분 모듈에 있음을 의미하고, 가중치 "0"은 세포 구성성분이 세포 구성성분 모듈에 있지 않음을 의미함)
세포 구성성분
CC 1
CC 2
CC 3
CC 4
CC Z
세포 구성성분 모듈 1
세포 구성성분 모듈 2
세포 구성성분 모듈 K
Weight1-1
Weight1-2
Weight1-3
Weight1-4
Weight1-Z
Weight2-1
Weight2-2
Weight2-3
Weight2-4
Weight2-Z
WeightK-1
WeightK-2
WeightK-3
WeightK-4
WeightK-Z
132-1
132-K
세포 구성성분 모듈

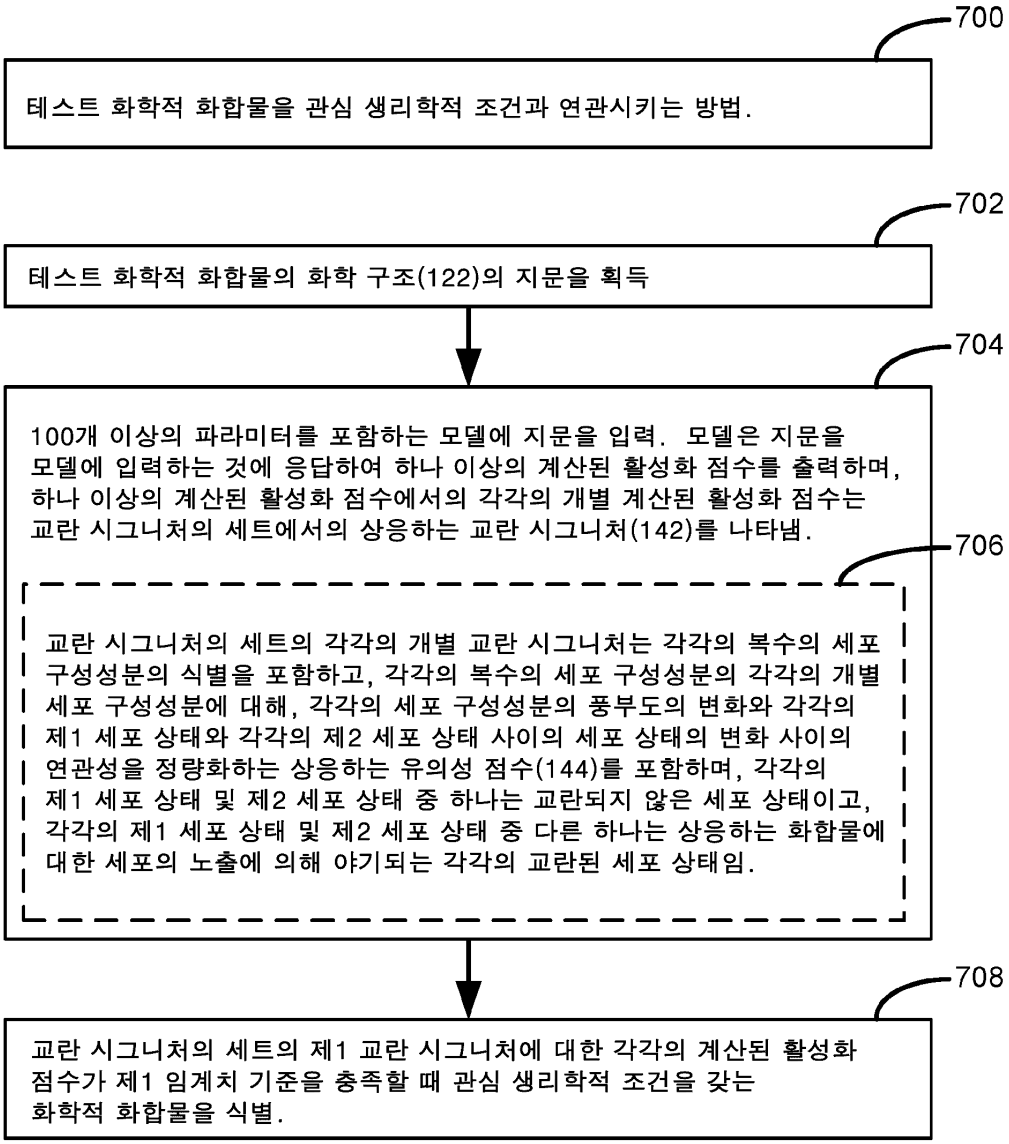
도면5


잠재 표현 404
세포 구성성분
세포 구성성분 모듈
세포 구성성분 모듈 1
세포 구성성분 모듈 2
세포 구성성분 모듈 K
CC 1
CC 2
CC 3
CC 4
CC Z
Weight1-1
Weight1-2
Weight1-3
Weight1-4
Weight1-Z
Weight2-1
Weight2-2
Weight2-3
Weight2-4
Weight2-N
WeightK-1
WeightK-2
WeightK-3
WeightK-4
WeightK-Z
제2 복수의 세포에 대한 카운트 행렬
CC 번호
세포 1
세포 G
cnt1-1
cnt1-G
cnt2-1
cnt2-G
cnt3-1
cnt3-G
cntZ-1
cntZ-G
활성화 데이터 구조 504
제2 복수의 세포
세포 구성성분 모듈
세포 구성성분 모듈 1
세포 구성성분 모듈 2
세포 구성성분 모듈 3
세포 구성성분 모듈 K
세포 1
세포 G
Act1-1
Act1-G
Act2-1
Act2-G
Act3-1
Act3-G
ActK-1
ActK-G

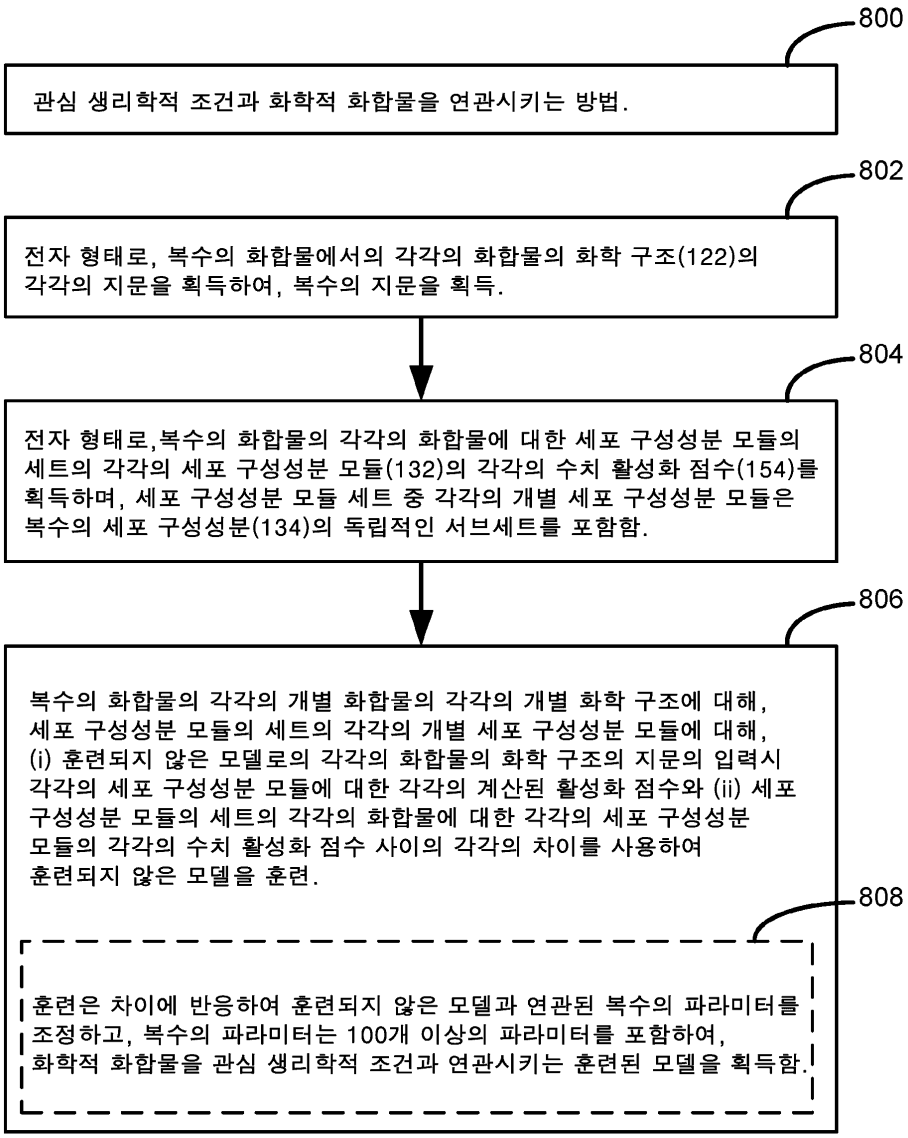
도면6


활성화 데이터 구조 504
세포 구성성분 모듈
세포 1
세포 G
640
세포 구성성분 모듈 1
Act1-1
Act1-G
세포 구성성분 모듈 2
Act2-1
Act2-G
세포 구성성분 모듈 3
Act3-1
Act3-G
세포 구성성분 모듈 K
ActK-1
ActK-G
601
604-1
CC 모듈
Cpd 1
Cpd W
CC 모듈 1
Weight1-1
Weight1-W
CC 모듈 2
Weight2-1
Weight2-W
CC 모듈 K
WeightK-1
WeightK-W
604-K
602
세포 구성성분 모델
Cpd
예측
실제
Cpd 1
Pred. Value1
Act. Value2
Cpd 2
Pred. Value2
Act. Value2
Cpd W
Pred. ValueW
Act. ValueW

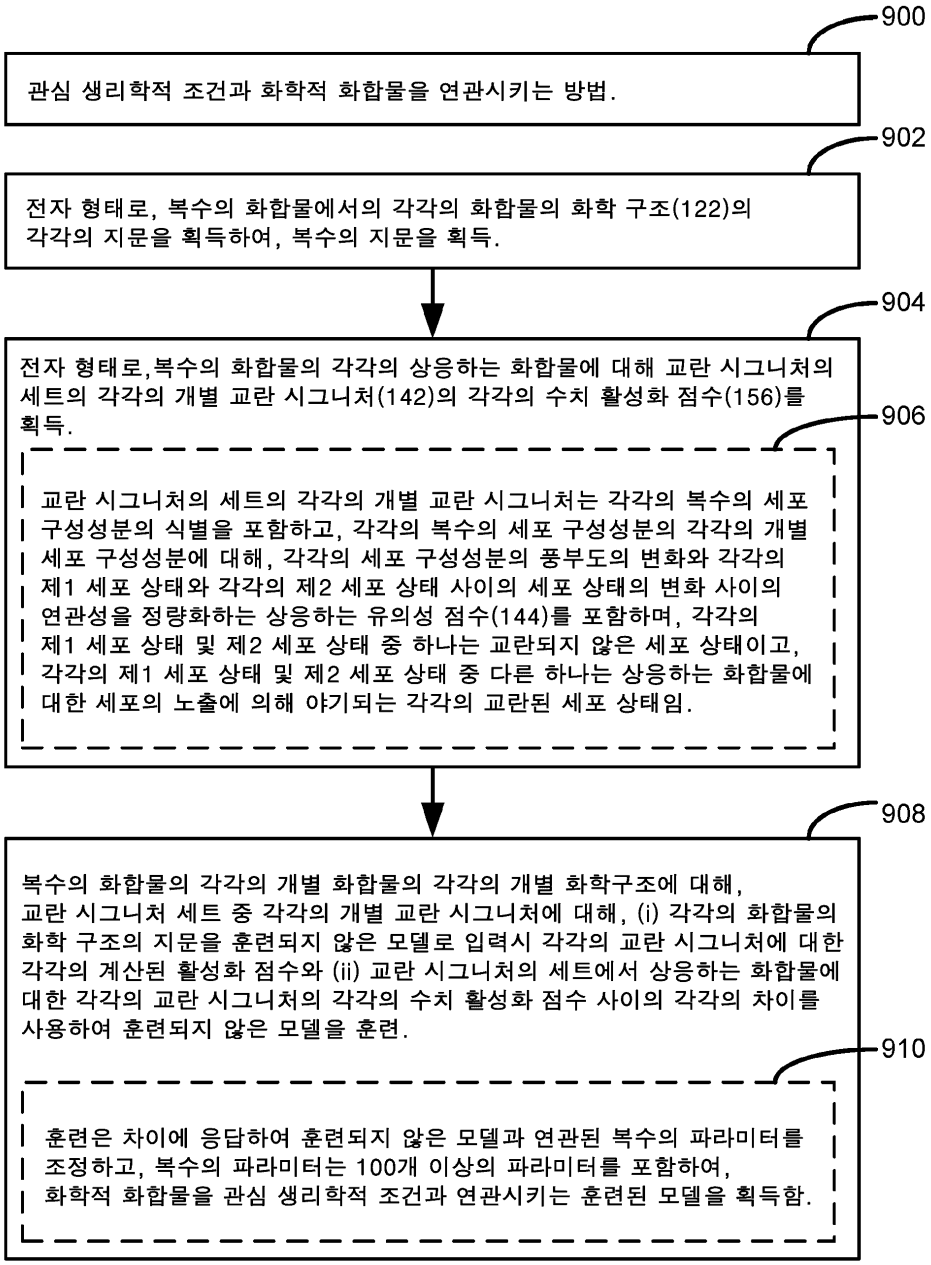
도면7


700
테스트 화학적 화합물을 관심 생리학적 조건과 연관시키는 방법.
702
테스트 화학적 화합물의 화학 구조(122)의 지문을 획득
704
100개 이상의 파라미터를 포함하는 모델에 지문을 입력. 모델은 지문을 모델에 입력하는 것에 응답하여 하나 이상의 계산된 활성화 점수를 출력하며, 하나 이상의 계산된 활성화 점수에서의 각각의 개별 계산된 활성화 점수는 교란 시그니처의 세트에서의 상응하는 교란 시그니처(142)를 나타냄.
706
교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별을 포함하고, 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수(144)를 포함하며, 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태임.
708
교란 시그니처의 세트의 제1 교란 시그니처에 대한 각각의 계산된 활성화 점수가 제1 임계치 기준을 충족할 때 관심 생리학적 조건을 갖는 화학적 화합물을 식별.

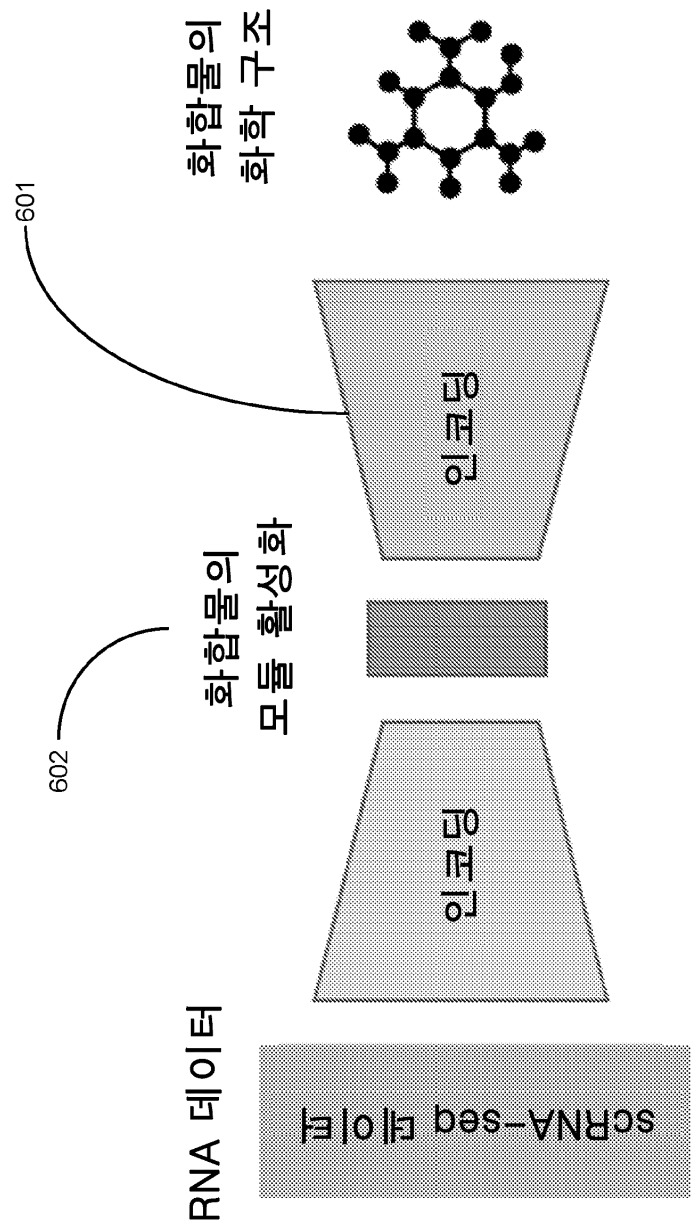
도면8


800
관심 생리학적 조건과 화학적 화합물을 연관시키는 방법.
802
전자 형태로, 복수의 화합물에서의 각각의 화합물의 화학 구조(122)의 각각의 지문을 획득하여, 복수의 지문을 획득.
804
전자 형태로,복수의 화합물의 각각의 화합물에 대한 세포 구성성분 모듈의 세트의 각각의 세포 구성성분 모듈(132)의 각각의 수치 활성화 점수(154)를 획득하며, 세포 구성성분 모듈 세트 중 각각의 개별 세포 구성성분 모듈은 복수의 세포 구성성분(134)의 독립적인 서브세트를 포함함.
806
복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학 구조에 대해, 세포 구성성분 모듈의 세트의 각각의 개별 세포 구성성분 모듈에 대해, (i) 훈련되지 않은 모델로의 각각의 화합물의 화학 구조의 지문의 입력시 각각의 세포 구성성분 모듈에 대한 각각의 계산된 활성화 점수와 (ii) 세포 구성성분 모듈의 세트의 각각의 화합물에 대한 각각의 세포 구성성분 모듈의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련되지 않은 모델을 훈련.
808
훈련은 차이에 반응하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고, 복수의 파라미터는 100개 이상의 파라미터를 포함하여, 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득함.

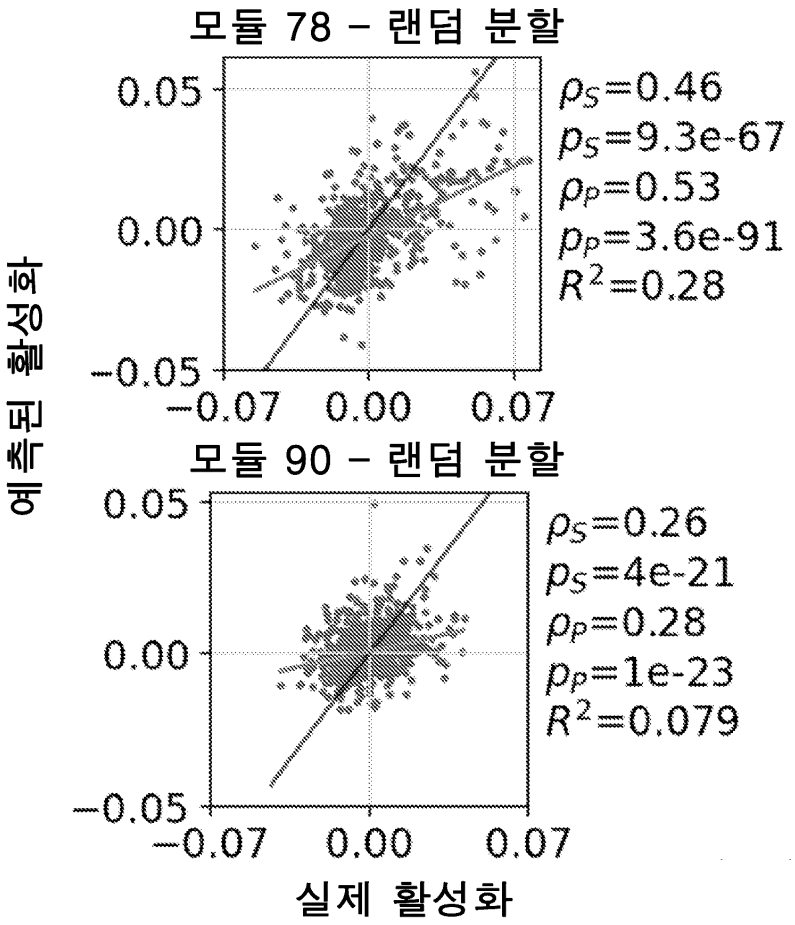
도면9


900
관심 생리학적 조건과 화학적 화합물을 연관시키는 방법.
902
전자 형태로, 복수의 화합물에서의 각각의 화합물의 화학 구조(122)의 각각의 지문을 획득하여, 복수의 지문을 획득.
904
전자 형태로,복수의 화합물의 각각의 상응하는 화합물에 대해 교란 시그니처의 세트의 각각의 개별 교란 시그니처(142)의 각각의 수치 활성화 점수(156)를 획득.
906
교란 시그니처의 세트의 각각의 개별 교란 시그니처는 각각의 복수의 세포 구성성분의 식별을 포함하고, 각각의 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 구성성분의 풍부도의 변화와 각각의 제1 세포 상태와 각각의 제2 세포 상태 사이의 세포 상태의 변화 사이의 연관성을 정량화하는 상응하는 유의성 점수(144)를 포함하며, 각각의 제1 세포 상태 및 제2 세포 상태 중 하나는 교란되지 않은 세포 상태이고, 각각의 제1 세포 상태 및 제2 세포 상태 중 다른 하나는 상응하는 화합물에 대한 세포의 노출에 의해 야기되는 각각의 교란된 세포 상태임.
908
복수의 화합물의 각각의 개별 화합물의 각각의 개별 화학구조에 대해, 교란 시그니처 세트 중 각각의 개별 교란 시그니처에 대해, (i) 각각의 화합물의 화학 구조의 지문을 훈련되지 않은 모델로 입력시 각각의 교란 시그니처에 대한 각각의 계산된 활성화 점수와 (ii) 교란 시그니처의 세트에서 상응하는 화합물에 대한 각각의 교란 시그니처의 각각의 수치 활성화 점수 사이의 각각의 차이를 사용하여 훈련되지 않은 모델을 훈련.
910
훈련은 차이에 응답하여 훈련되지 않은 모델과 연관된 복수의 파라미터를 조정하고, 복수의 파라미터는 100개 이상의 파라미터를 포함하여, 화학적 화합물을 관심 생리학적 조건과 연관시키는 훈련된 모델을 획득함.

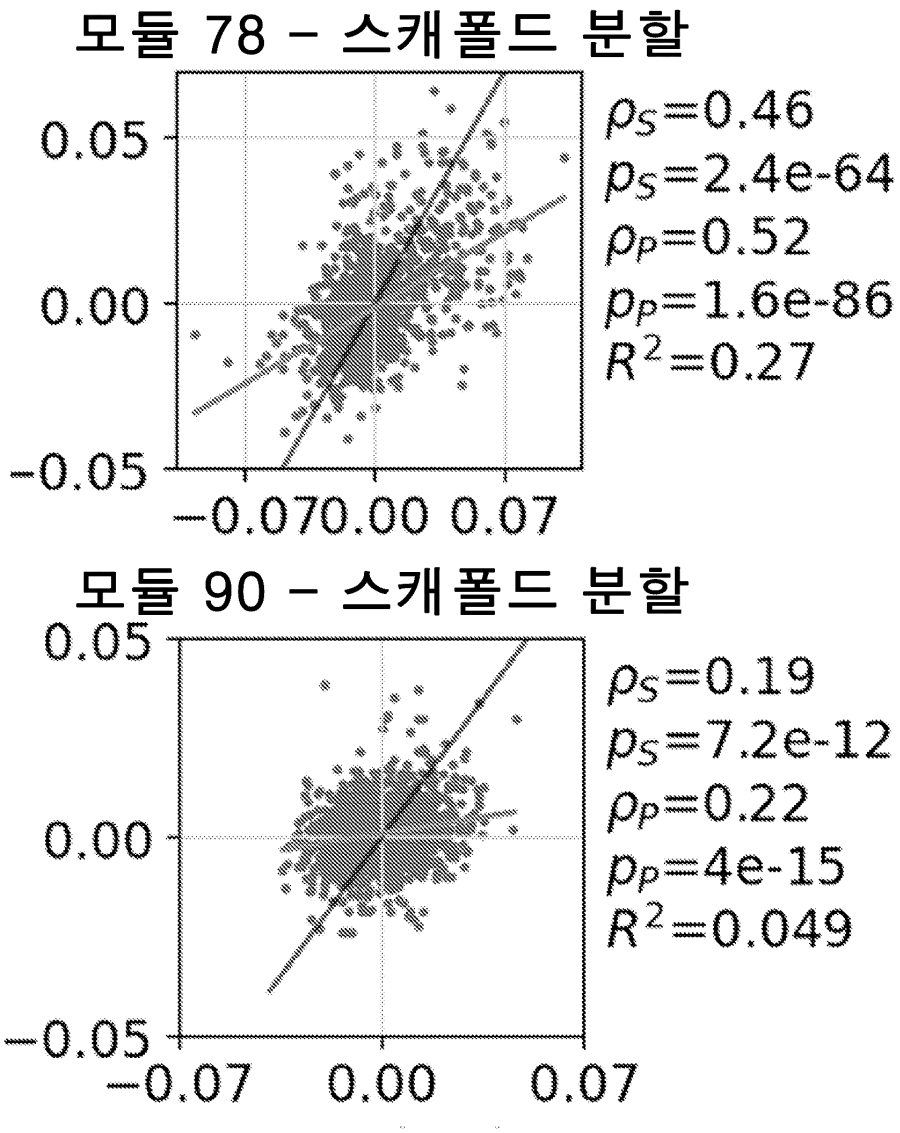
도면10a



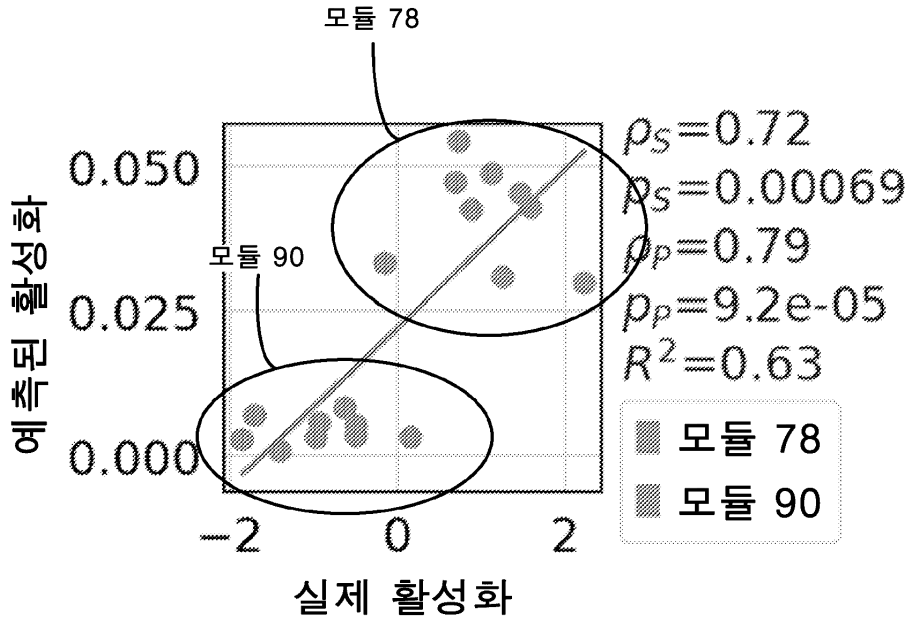
도면10b


모듈 78 – 랜덤 분할
0.05
0.00
−0.05
−0.07
0.00
0.07
ρ_S=0.46
p_S=9.3e-67
ρ_P=0.53
p_P=3.6e-91
R^2=0.28
모듈 90 – 랜덤 분할
0.05
0.00
−0.05
−0.07
0.00
0.07
ρ_S=0.26
p_S=4e-21
ρ_P=0.28
p_P=1e-23
R^2=0.079
예측된 활성화
실제 활성화

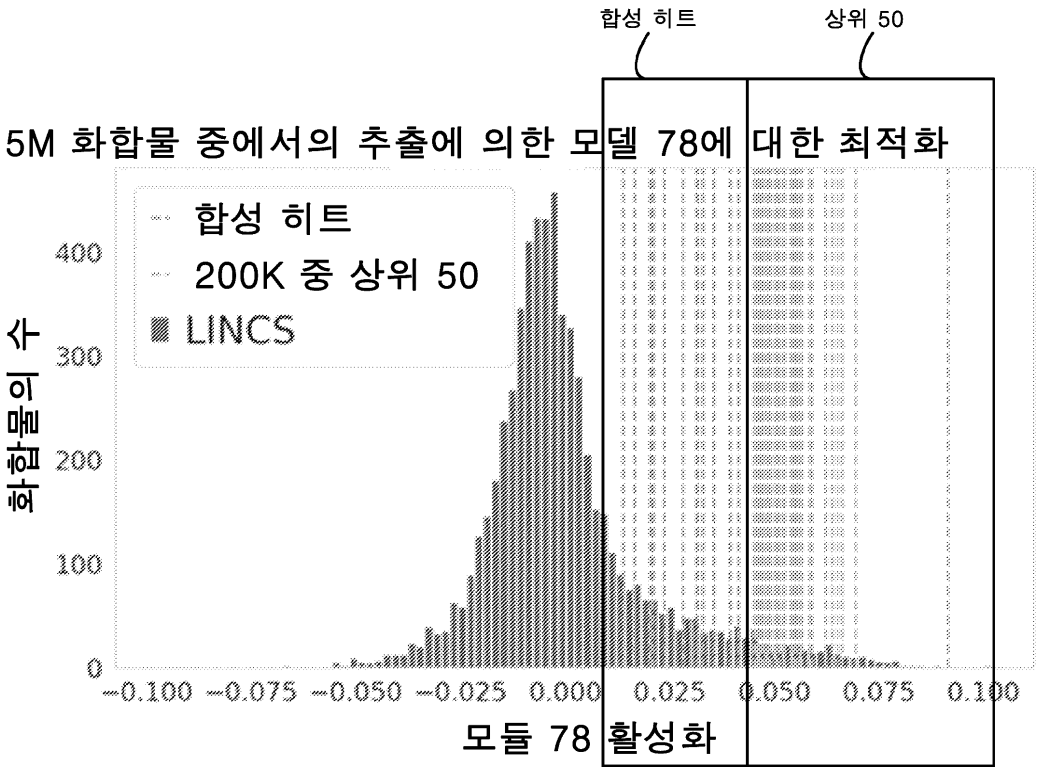
도면10c


모듈 78 – 스캐폴드 분할
0.05
0.00
−0.05
−0.07 0.00 0.07
ρ_S=0.46
p_S=2.4e-64
ρ_P=0.52
p_P=1.6e-86
R^2=0.27
모듈 90 – 스캐폴드 분할
0.05
0.00
−0.05
−0.07 0.00 0.07
ρ_S=0.19
p_S=7.2e-12
ρ_P=0.22
p_P=4e-15
R^2=0.049
실제 활성화

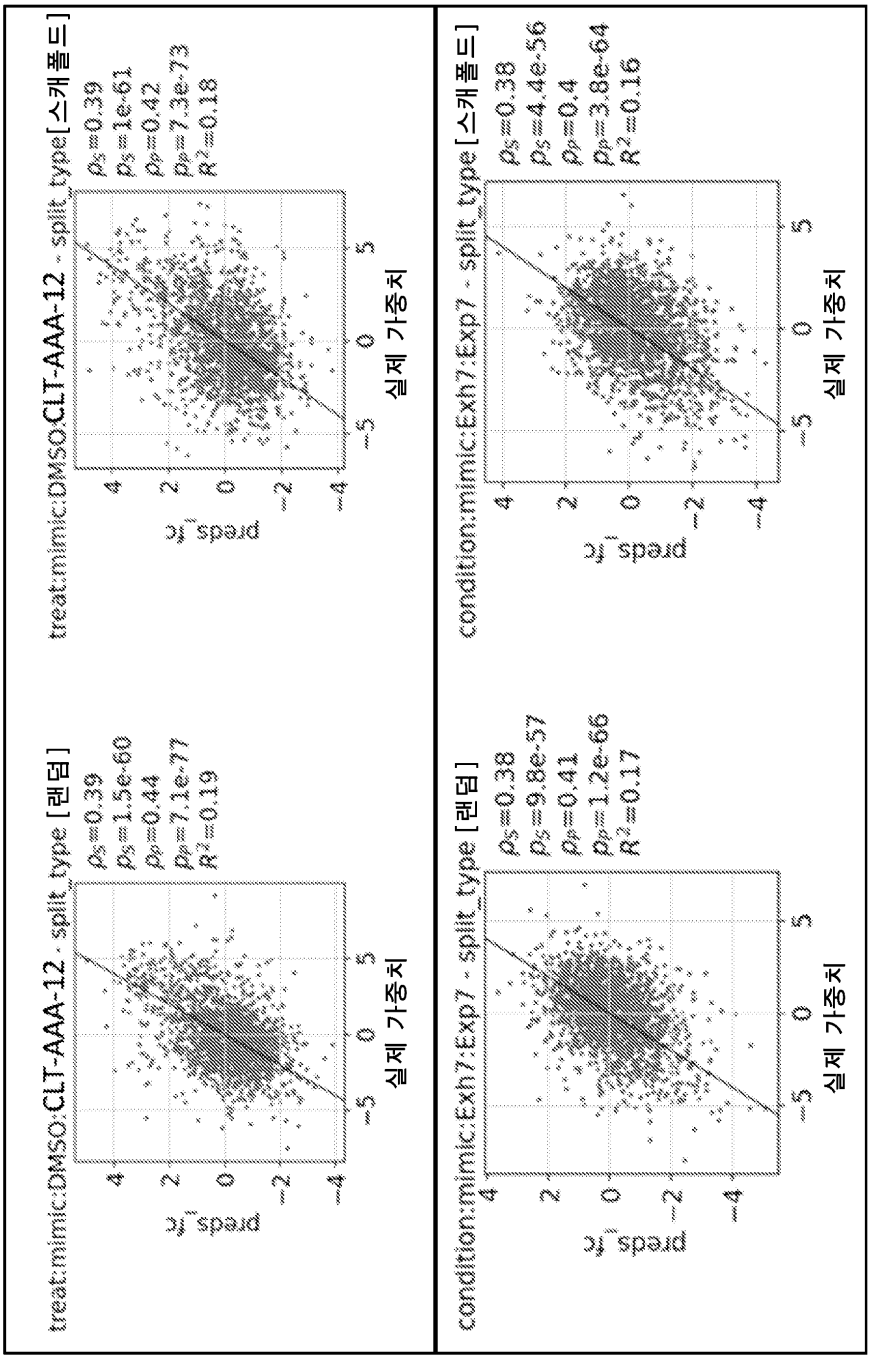
도면10d


모듈 78
모듈 90
0.050
0.025
0.000
−2
0
2
실제 활성화
예측된 활성화
ρ_S=0.72
p_S=0.00069
ρ_P=0.79
p_P=9.2e-05
R^2=0.63
모듈 78
모듈 90

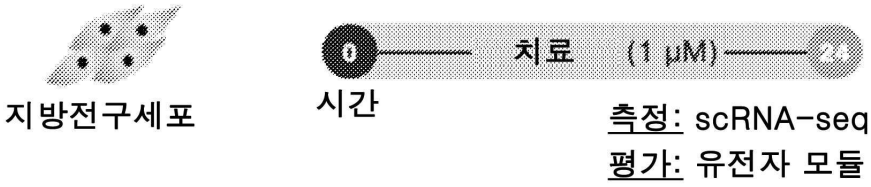
도면10e


합성 히트
상위 50
5M 화합물 중에서의 추출에 의한 모델 78에 대한 최적화
합성 히트
200K 중 상위 50
LINCS
400
300
200
100
0
수
화합물의
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
모듈 78 활성화

도면11

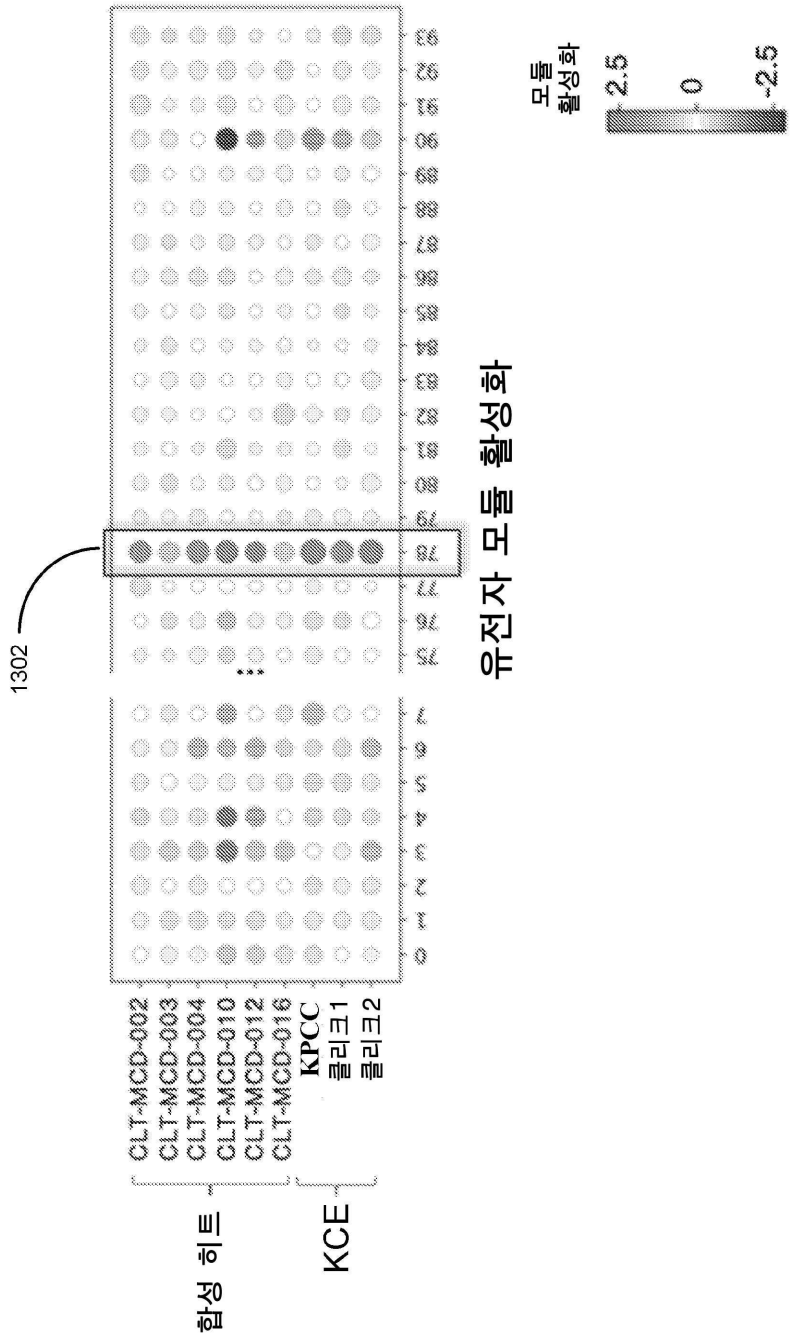

treat:mimic:DMSO:CLT-AAA-12 - split_type [랜덤]
ρ_S=0.39
p_S=1.5e-60
ρ_P=0.44
p_P=7.1e-77
R^2=0.19
treat:mimic:DMSO:CLT-AAA-12 - split_type [스캐폴드]
ρ_S=0.39
p_S=1e-61
ρ_P=0.42
p_P=7.3e-73
R^2=0.18
condition:mimic:Exh7:Exp7 - split_type [랜덤]
ρ_S=0.38
p_S=9.8e-57
ρ_P=0.41
p_P=1.2e-66
R^2=0.17
condition:mimic:Exh7:Exp7 - split_type [스캐폴드]
ρ_S=0.38
p_S=4.4e-56
ρ_P=0.4
p_P=3.8e-64
R^2=0.16
preds_fc
실제 가중치

도면12

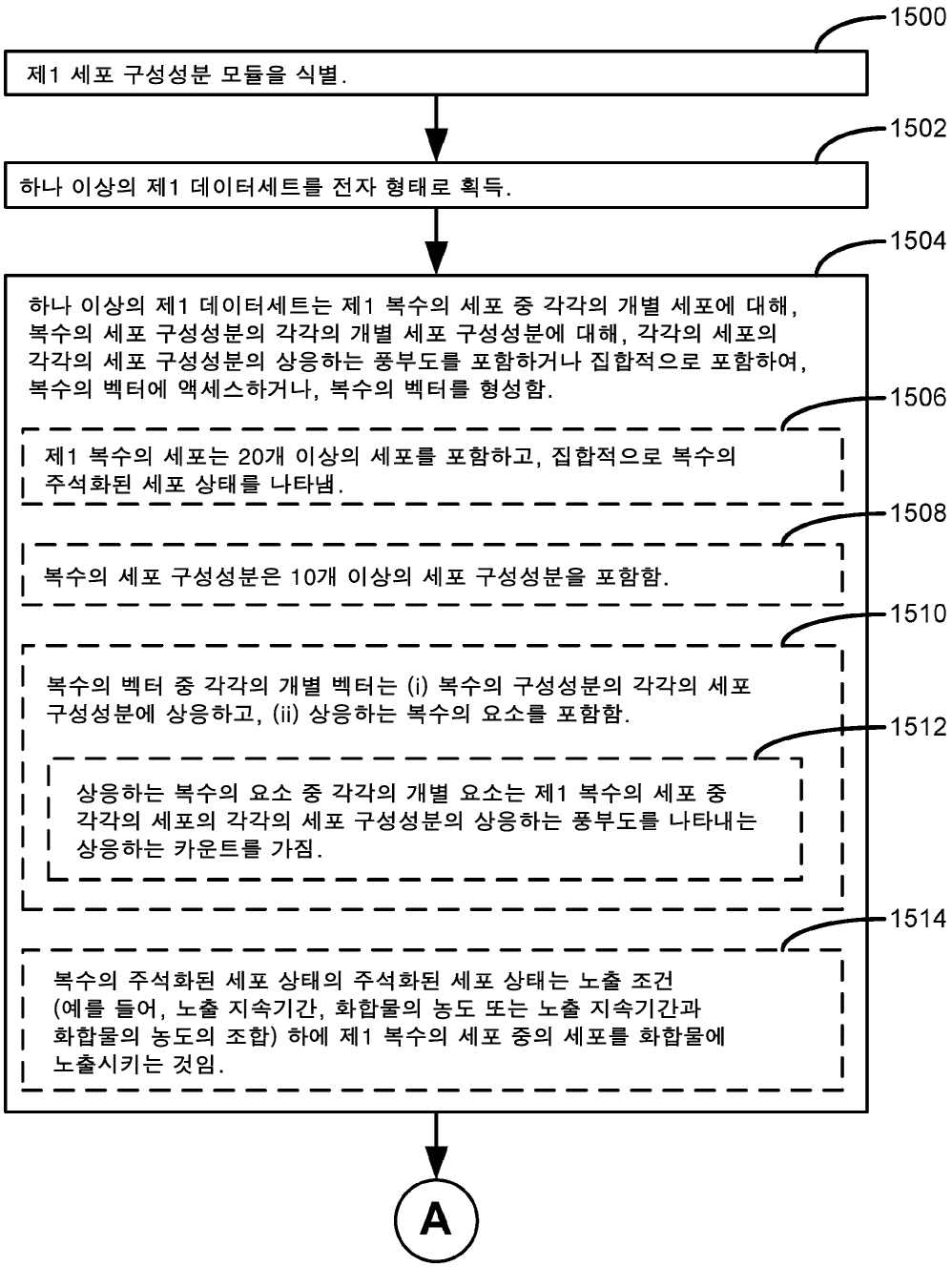


	KCE	합성 히트	
DMSO	KPCC	CLT-MCD-002	CLT-MCD-010
	클리크 1	CLT-MCD-003	CLT-MCD-012
	클리크 2	CLT-MCD-004	CLT-MCD-016

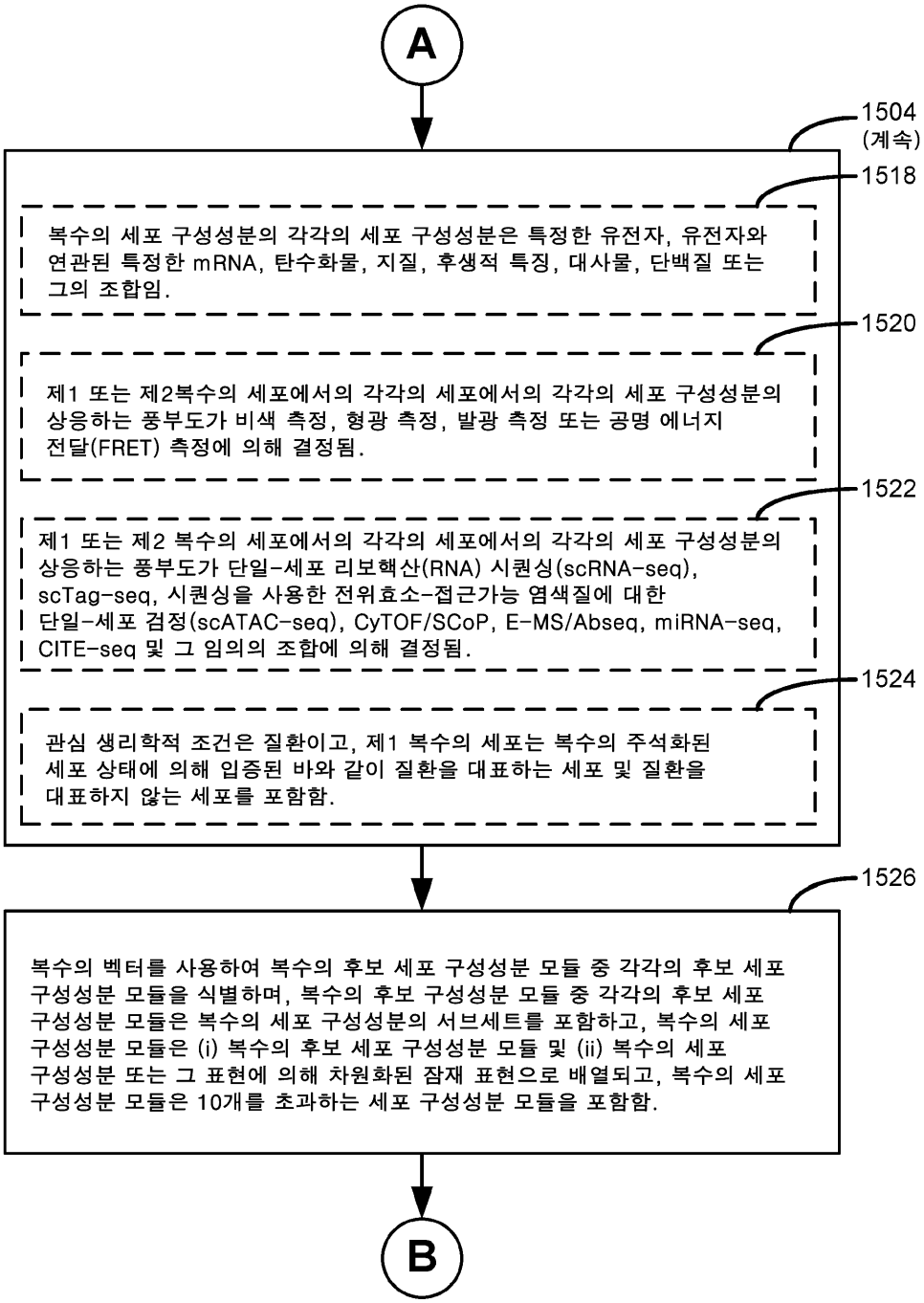
도면13


1302
유전자 모듈 활성화
모듈 활성화
2.5
0
-2.5
CLT-MCD-002
CLT-MCD-003
CLT-MCD-004
CLT-MCD-010
CLT-MCD-012
CLT-MCD-016
KPCC 클리크1
클리크2
합성 히트
KCE

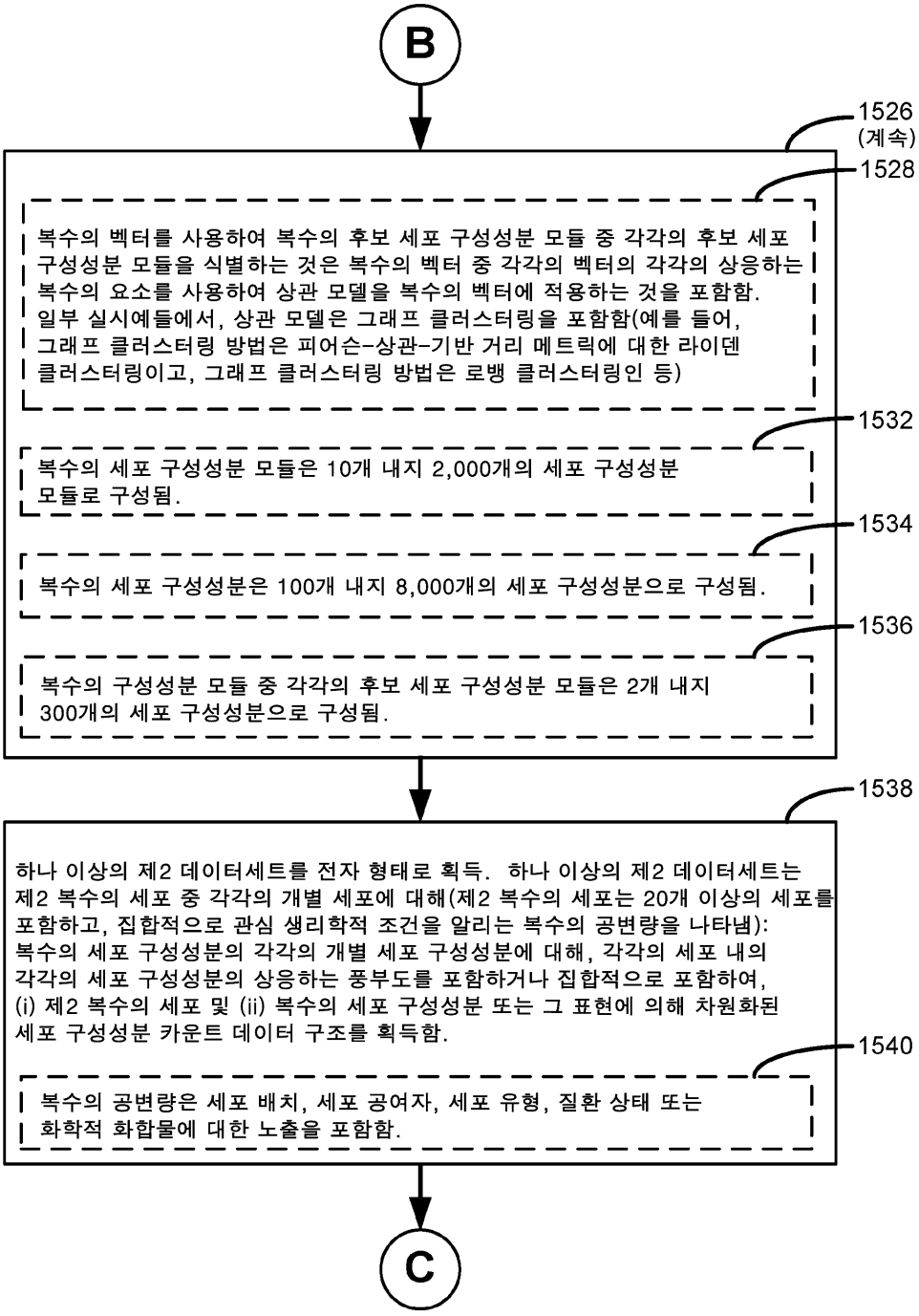
도면14a


1500
제1 세포 구성성분 모듈을 식별.
1502
하나 이상의 제1 데이터세트를 전자 형태로 획득.
1504
하나 이상의 제1 데이터세트는 제1 복수의 세포 중 각각의 개별 세포에 대해, 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하여, 복수의 벡터에 액세스하거나, 복수의 벡터를 형성함.
1506
제1 복수의 세포는 20개 이상의 세포를 포함하고, 집합적으로 복수의 주석화된 세포 상태를 나타냄.
1508
복수의 세포 구성성분은 10개 이상의 세포 구성성분을 포함함.
1510
복수의 벡터 중 각각의 개별 벡터는 (i) 복수의 구성성분의 각각의 세포 구성성분에 상응하고, (ii) 상응하는 복수의 요소를 포함함.
1512
상응하는 복수의 요소 중 각각의 개별 요소는 제1 복수의 세포 중 각각의 세포의 각각의 세포 구성성분의 상응하는 풍부도를 나타내는 상응하는 카운트를 가짐.
1514
복수의 주석화된 세포 상태의 주석화된 세포 상태는 노출 조건 (예를 들어, 노출 지속기간, 화합물의 농도 또는 노출 지속기간과 화합물의 농도의 조합) 하에 제1 복수의 세포 중의 세포를 화합물에 노출시키는 것임.
A

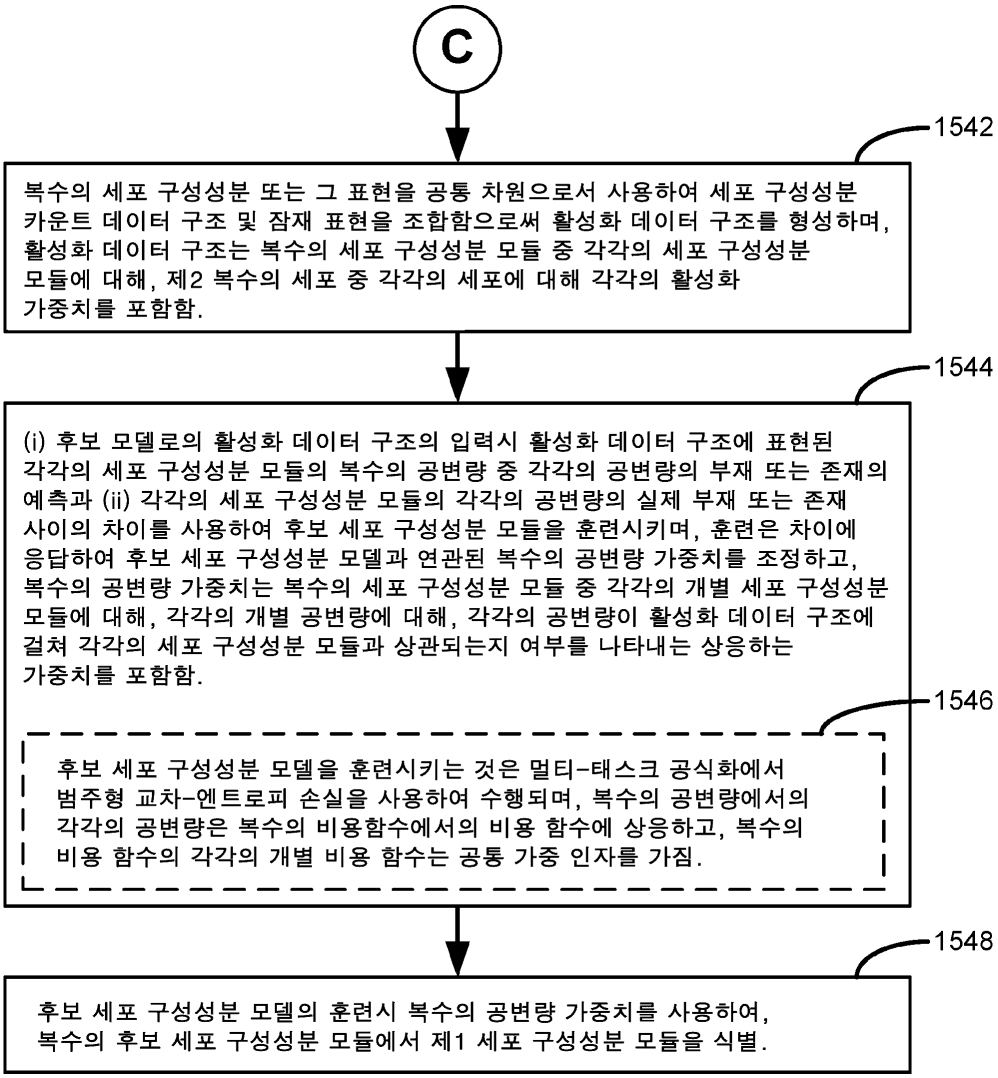
도면14b


A
1504
(계속)
1518
복수의 세포 구성성분의 각각의 세포 구성성분은 특정한 유전자, 유전자와 연관된 특정한 mRNA, 탄수화물, 지질, 후생적 특징, 대사물, 단백질 또는 그의 조합임.
1520
제1 또는 제2복수의 세포에서의 각각의 세포에서의 각각의 세포 구성성분의 상응하는 풍부도가 비색 측정, 형광 측정, 발광 측정 또는 공명 에너지 전달(FRET) 측정에 의해 결정됨.
1522
제1 또는 제2 복수의 세포에서의 각각의 세포에서의 각각의 세포 구성성분의 상응하는 풍부도가 단일-세포 리보핵산(RNA) 시퀀싱(scRNA-seq), scTag-seq, 시퀀싱을 사용한 전위효소-접근가능 염색질에 대한 단일-세포 검정(scATAC-seq), CyTOF/SCoP, E-MS/Abseq, miRNA-seq, CITE-seq 및 그 임의의 조합에 의해 결정됨.
1524
관심 생리학적 조건은 질환이고, 제1 복수의 세포는 복수의 주석화된 세포 상태에 의해 입증된 바와 같이 질환을 대표하는 세포 및 질환을 대표하지 않는 세포를 포함함.
1526
복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈 중 각각의 후보 세포 구성성분 모듈을 식별하며, 복수의 후보 구성성분 모듈 중 각각의 후보 세포 구성성분 모듈은 복수의 세포 구성성분의 서브세트를 포함하고, 복수의 세포 구성성분 모듈은 (i) 복수의 후보 세포 구성성분 모듈 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 잠재 표현으로 배열되고, 복수의 세포 구성성분 모듈은 10개를 초과하는 세포 구성성분 모듈을 포함함.
B

도면14c


B
1526
(계속)
1528
복수의 벡터를 사용하여 복수의 후보 세포 구성성분 모듈 중 각각의 후보 세포 구성성분 모듈을 식별하는 것은 복수의 벡터 중 각각의 벡터의 각각의 상응하는 복수의 요소를 사용하여 상관 모델을 복수의 벡터에 적용하는 것을 포함함. 일부 실시예들에서, 상관 모델은 그래프 클러스터링을 포함함(예를 들어, 그래프 클러스터링 방법은 피어슨-상관-기반 거리 메트릭에 대한 라이덴 클러스터링이고, 그래프 클러스터링 방법은 로뱅 클러스터링인 등)
1532
복수의 세포 구성성분 모듈은 10개 내지 2,000개의 세포 구성성분 모듈로 구성됨.
1534
복수의 세포 구성성분은 100개 내지 8,000개의 세포 구성성분으로 구성됨.
1536
복수의 구성성분 모듈 중 각각의 후보 세포 구성성분 모듈은 2개 내지 300개의 세포 구성성분으로 구성됨.
1538
하나 이상의 제2 데이터세트를 전자 형태로 획득. 하나 이상의 제2 데이터세트는 제2 복수의 세포 중 각각의 개별 세포에 대해(제2 복수의 세포는 20개 이상의 세포를 포함하고, 집합적으로 관심 생리학적 조건을 알리는 복수의 공변량을 나타냄): 복수의 세포 구성성분의 각각의 개별 세포 구성성분에 대해, 각각의 세포 내의 각각의 세포 구성성분의 상응하는 풍부도를 포함하거나 집합적으로 포함하여, (i) 제2 복수의 세포 및 (ii) 복수의 세포 구성성분 또는 그 표현에 의해 차원화된 세포 구성성분 카운트 데이터 구조를 획득함.
1540
복수의 공변량은 세포 배치, 세포 공여자, 세포 유형, 질환 상태 또는 화학적 화합물에 대한 노출을 포함함.
C

도면14d


C
1542
복수의 세포 구성성분 또는 그 표현을 공통 차원으로서 사용하여 세포 구성성분 카운트 데이터 구조 및 잠재 표현을 조합함으로써 활성화 데이터 구조를 형성하며, 활성화 데이터 구조는 복수의 세포 구성성분 모듈 중 각각의 세포 구성성분 모듈에 대해, 제2 복수의 세포 중 각각의 세포에 대해 각각의 활성화 가중치를 포함함.
1544
(i) 후보 모델로의 활성화 데이터 구조의 입력시 활성화 데이터 구조에 표현된 각각의 세포 구성성분 모듈의 복수의 공변량 중 각각의 공변량의 부재 또는 존재의 예측과 (ii) 각각의 세포 구성성분 모듈의 각각의 공변량의 실제 부재 또는 존재 사이의 차이를 사용하여 후보 세포 구성성분 모듈을 훈련시키며, 훈련은 차이에 응답하여 후보 세포 구성성분 모델과 연관된 복수의 공변량 가중치를 조정하고, 복수의 공변량 가중치는 복수의 세포 구성성분 모듈 중 각각의 개별 세포 구성성분 모듈에 대해, 각각의 개별 공변량에 대해, 각각의 공변량이 활성화 데이터 구조에 걸쳐 각각의 세포 구성성분 모듈과 상관되는지 여부를 나타내는 상응하는 가중치를 포함함.
1546
후보 세포 구성성분 모델을 훈련시키는 것은 멀티-태스크 공식화에서 범주형 교차-엔트로피 손실을 사용하여 수행되며, 복수의 공변량에서의 각각의 공변량은 복수의 비용함수에서의 비용 함수에 상응하고, 복수의 비용 함수의 각각의 개별 비용 함수는 공통 가중 인자를 가짐.
1548
후보 세포 구성성분 모델의 훈련시 복수의 공변량 가중치를 사용하여, 복수의 후보 세포 구성성분 모듈에서 제1 세포 구성성분 모듈을 식별.