(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2024-525155
(P2024-525155A)
(43)公表日　令和6年7月10日(2024.7.10)

(51)国際特許分類		FI			テーマコード（参考）
Ｃ１２Ｑ	*1/02 (2006.01)*	Ｃ１２Ｑ	1/02		４Ｂ０６３
Ｃ１２Ｑ	*1/68 (2018.01)*	Ｃ１２Ｑ	1/68	１００Ｚ	４Ｂ０６５
Ｃ１２Ｎ	*5/071(2010.01)*	Ｃ１２Ｎ	5/071		

審査請求　未請求　予備審査請求　未請求　（全139頁）

(21)出願番号　特願2023-577304(P2023-577304)
(86)(22)出願日　令和4年6月15日(2022.6.15)
(85)翻訳文提出日　令和6年2月6日(2024.2.6)
(86)国際出願番号　PCT/US2022/033685
(87)国際公開番号　WO2022/266259
(87)国際公開日　令和4年12月22日(2022.12.22)
(31)優先権主張番号　63/210,930
(32)優先日　令和3年6月15日(2021.6.15)
(33)優先権主張国・地域又は機関
米国(US)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,
最終頁に続く

(71)出願人　520445473
フラッグシップ パイオニアリング イノベーションズ シックス，エルエルシー
アメリカ合衆国マサチューセッツ州０２１４２，ケンブリッジ，ケンブリッジ・パークウェイ ５５，エイス・フロア
(74)代理人　100079108
弁理士　稲葉 良幸
(74)代理人　100109346
弁理士　大貫 敏史
(74)代理人　100117189
弁理士　江口 昭彦
(74)代理人　100134120
弁理士　内藤 和彦
(72)発明者　ウォルフ，ファビアン アレクサンダー
最終頁に続く

(54)【発明の名称】　フィンガープリント分析を使用して化合物を生理学的状態と関連付けるためのシステム及び方法

(57)【要約】
【課題】化合物を生理学的状態と関連付けるためのシステム及び方法が提供される。
【解決手段】化合物化学構造のフィンガープリントが得られ、１つ以上の計算された活性化スコアを出力するモデルに入力される。各活性化スコアは、モジュールのセットにおける細胞構成要素モジュールを表し、各モジュールは、細胞構成要素のサブセットを含み、モジュールのセットにおける第１のモジュールは、生理学的状態と関連付けられる。第１のモジュールについての活性化スコアが閾値基準を満たす場合、化合物は、生理学的状態と関連付けられると識別される。いくつかの態様において、各活性化スコアは、生理学的状態と関連付けられた摂動シグネチャを表し、化合物は、第１の摂動シグネチャについての活性化スコアが閾値基準を満たす場合に識別される。化合物を生理学的状態と関連付けるモデルを訓練するためのシステム及び方法も提供される。
【選択図】図１０Ａ

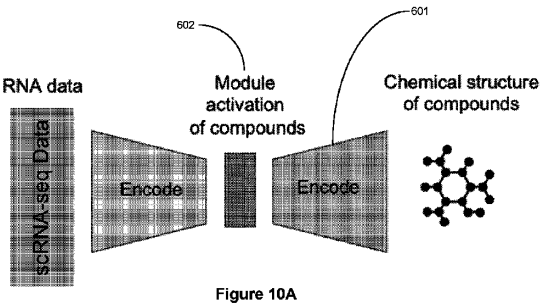


Figure 10A

【特許請求の範囲】
【請求項１】
試験化学化合物を目的の生理学的状態と関連付ける方法であって、前記方法が、
（Ａ）前記試験化学化合物の化学構造のフィンガープリントを得ることと、
（Ｂ）細胞構成要素モジュールのセットにアクセスすることであって、
前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素のそれぞれの独立したサブセットを含み、
前記複数の細胞構成要素のそれぞれの独立したサブセットの各々についての対応する複数の細胞ベースのアッセイ存在量値が、前記生理学的状態と関連付けられた複数の異なる状態にわたって別々に相関し、
前記細胞構成要素モジュールのセットにおける第１の細胞構成要素モジュールが、前記目的の生理学的状態と関連付けられる、アクセスすることと、
（Ｃ）前記化学構造の前記フィンガープリントを、１００以上のパラメータを含むモデルに入力することに応答して、前記モデルからの出力として、前記細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々についてのそれぞれの活性化スコアを取得することと、
（Ｄ）前記第１の細胞構成要素モジュールについての前記活性化スコアが、第１の閾値基準を満たす場合、前記試験化学化合物を前記目的の生理学的状態と関連付けることと、を含む、方法。
【請求項２】
前記細胞ベースのアッセイ存在量値が、臓器の細胞のものである、請求項１に記載の方法。
【請求項３】
前記臓器が、心臓、肝臓、肺、筋肉、脳、膵臓、脾臓、腎臓、小腸、子宮、又は膀胱である、請求項２に記載の方法。
【請求項４】
前記細胞ベースのアッセイ存在量値が、組織の細胞のものである、請求項１に記載の方法。
【請求項５】
前記組織が、骨、軟骨、関節、気管、脊髄、角膜、眼、皮膚、又は血管である、請求項４に記載の方法。
【請求項６】
前記細胞ベースのアッセイ存在量値が、複数の幹細胞の細胞のものである、請求項１に記載の方法。
【請求項７】
前記複数の幹細胞が、複数の胚性幹細胞、複数の成体幹細胞、又は複数の人工多能性幹細胞（ｉＰＳＣ）である、請求項６に記載の方法。
【請求項８】
前記細胞ベースのアッセイ存在量値が、複数の初代ヒト細胞の細胞のものである、請求項１に記載の方法。
【請求項９】
前記複数の初代ヒト細胞が、複数のＣＤ３４＋細胞、複数のＣＤ３４＋造血幹、複数の前駆細胞（ＨＳＰＣ）、複数のＴ細胞、複数の間葉系幹細胞（ＭＳＣ）、複数の気道基底幹細胞、又は複数の人工多能性幹細胞である、請求項８に記載の方法。
【請求項１０】
前記細胞ベースのアッセイ存在量値が、臍帯血中、末梢血中、又は骨髄中の細胞のものである、請求項１に記載の方法。
【請求項１１】
前記細胞ベースのアッセイ存在量値が、固体組織中の細胞のものである、請求項１に記載の方法。

【請求項１２】
前記固体組織が、胎盤、肝臓、心臓、脳、腎臓、又は胃腸管である、請求項１１に記載の方法。
【請求項１３】
前記細胞ベースのアッセイ存在量値が、複数の分化細胞のものである、請求項１に記載の方法。
【請求項１４】
前記複数の分化細胞が、複数の巨核球、複数の骨芽細胞、複数の軟骨細胞、複数の脂肪細胞、複数の肝細胞、複数の肝中皮細胞、複数の胆管上皮細胞、複数の肝星細胞、複数の肝類洞内皮細胞、複数のクッパー細胞、複数のピット細胞、複数の血管内皮細胞、複数の膵管上皮細胞、複数の膵管細胞、複数の腺房中心細胞、複数の腺房細胞、複数のランゲルハンス島、複数の心筋細胞、複数の線維芽細胞、複数のケラチノサイト、複数の平滑筋細胞、複数のＩ型肺胞上皮細胞、複数のＩＩ型肺胞上皮細胞、複数のクララ細胞、複数の線毛上皮細胞、複数の基底細胞、複数の杯細胞、複数の神経内分泌細胞、複数のクルチッキー（ｋｕｌｔｓｃｈｉｔｚｋｙ）細胞、複数の尿細管上皮細胞、複数の尿路上皮細胞、複数の円柱上皮細胞、複数の糸球体上皮細胞、複数の糸球体内皮細胞、複数の有足細胞、複数のメサンギウム細胞、複数の神経細胞、複数の星状膠細胞、複数の小膠細胞、又は複数の乏突起膠細胞である、請求項１３に記載の方法。
【請求項１５】
前記対応する複数の細胞ベースのアッセイ存在量値が、複数の細胞の単一細胞リボ核酸（ＲＮＡ）配列決定（ｓｃＲＮＡ－ｓｅｑ）データである、請求項１～１４のいずれか一項に記載の方法。
【請求項１６】
前記生理学的状態と関連付けられた前記複数の異なる状態が、細胞のアリコートが前記生理学的状態に影響を与えることが知られている化合物に曝露されている対照状態に加えて、前記生理学的状態に影響を与えることが知られている１つ以上の参照化合物に異なる細胞のアリコートを曝露することによって導出される、請求項１５に記載の方法。
【請求項１７】
前記対応する複数の細胞ベースのアッセイ存在量値が、バルクＲＮＡ配列に由来する、請求項１～１４のいずれか一項に記載の方法。
【請求項１８】
前記対応する複数の細胞ベースのアッセイ存在量値が、単一細胞ＲＮＡ配列決定に由来する、請求項１～１４のいずれか一項に記載の方法。
【請求項１９】
前記細胞構成要素モジュールのセットが、前記第１の細胞構成要素モジュールからなる、請求項１～１８のいずれか一項に記載の方法。
【請求項２０】
前記細胞構成要素モジュールのセットが、複数の細胞構成要素モジュールを含み、前記モデルが、複数のコンポーネントモデルを含むアンサンブルモデルであり、前記複数のコンポーネントモデルにおけるコンポーネントモデルの各々が、前記化学構造の前記フィンガープリントを前記複数のコンポーネントモデルにおけるコンポーネントモデルの各々に入力することに応答して、前記細胞構成要素モジュールのセットにおける異なる細胞構成要素モジュールについての活性化スコアを提供する、請求項１～１８のいずれか一項に記載の方法。
【請求項２１】
前記方法が、前記試験化学化合物の単純化された分子入力ラインエントリーシステム（ＳＭＩＬＥＳ）文字列表現から前記フィンガープリントを計算することを更に含む、請求項１～２０のいずれか一項に記載の方法。
【請求項２２】
前記複数のコンポーネントモデルにおけるコンポーネントモデルの各々が、対応するニ

ューラルネットワークである、請求項２０又は２１に記載の方法。
【請求項２３】
前記対応するニューラルネットワークが、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである、２２に記載の方法。
【請求項２４】
前記複数のコンポーネントモデルにおけるコンポーネントモデルが、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルである、請求項２０又は２１に記載の方法。
【請求項２５】
前記対応するニューラルネットワークが、対応する完全に接続されたニューラルネットワーク及び対応するメッセージパッシングニューラルネットワークの組み合わせであり、
前記対応する完全に接続されたニューラルネットワークの第１の出力及び前記対応するメッセージパッシングニューラルネットワークの第２の出力が、前記化学構造の前記フィンガープリントを前記対応する完全に接続されたニューラルネットワーク及び前記対応するメッセージパッシングニューラルネットワークに入力することに応答して、組み合わされ、前記細胞構成要素モジュールのセットにおける前記対応する細胞構成要素モジュールについての１つ以上の計算された活性化スコアにおける活性化スコアを決定する、請求項２２に記載の方法。
【請求項２６】
前記細胞構成要素モジュールのセットが、複数の細胞構成要素モジュールであり、
前記第１の細胞構成要素モジュールを含む前記複数の細胞構成要素モジュールの第１のサブセットが、前記目的の生理学的状態と関連付けられ、
前記複数の細胞構成要素モジュールの第２のサブセットが、前記目的の生理学的状態と関連付けられず、
前記第１の細胞構成要素モジュールについてのそれぞれの計算された活性化スコアが、前記第１の閾値基準を満たし、前記複数の細胞構成要素モジュールの前記第２のサブセットにおける細胞構成要素モジュールについてのそれぞれの計算された活性化スコアが、前記第１の閾値基準以外の第２の閾値基準を満たす場合、前記試験化学化合物が、前記目的の生理学的状態と識別される、請求項１に記載の方法。
【請求項２７】
前記方法が、
電子形式で１つ以上の第１のデータセットを得、前記１つ以上の第１のデータセットが、
第１の複数の細胞におけるそれぞれの細胞の各々について、前記第１の複数の細胞が、２０個以上の細胞を含み、複数の注釈付きの細胞状態を集合的に表し、
前記複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、前記複数の細胞構成要素が、１０個以上の細胞構成要素を含み、
前記それぞれの細胞における前記それぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含み、
それによって、複数のベクトルにアクセスするか、又はそれらを形成することであって、前記複数のベクトルにおけるそれぞれのベクトルの各々が、（ｉ）前記複数の構成要素におけるそれぞれの細胞構成要素に対応し、（ｉｉ）対応する複数のエレメントを含み、前記対応する複数のエレメントにおけるそれぞれのエレメントの各々が、前記第１の複数の細胞における前記それぞれの細胞における前記それぞれの細胞構成要素の前記対応する存在量を表す対応するカウントを有する、複数のベクトルにアクセスするか、又はそれらを形成することと、
前記複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細

胞構成要素モジュールの各々を識別することであって、前記複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々が、前記複数の細胞構成要素のサブセットを含み、前記複数の細胞構成要素モジュールが、（ｉ）前記複数の候補細胞構成要素モジュール及び（ｉｉ）前記複数の細胞構成要素又はその表現によって次元決定された潜在表現で配置され、前記複数の細胞構成要素モジュールが、１０を超える細胞構成要素モジュールを含む、識別することと、

電子形式で１つ以上の第２のデータセットを得、前記１つ以上の第２のデータセットが、

第２の複数の細胞におけるそれぞれの細胞の各々について、前記第２の複数の細胞が、２０個以上の細胞を含み、前記目的の生理学的状態を通知する複数の共変量を集合的に表し、

前記複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、

前記それぞれの細胞における前記それぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含み、

それによって、（ｉ）前記第２の複数の細胞及び（ｉｉ）前記複数の細胞構成要素又はその前記表現によって次元決定された細胞構成要素カウントデータ構造を得ることと、

前記複数の細胞構成要素又はその前記表現を共通次元として使用して前記細胞構成要素カウントデータ構造及び前記潜在表現を組み合わせることによって活性化データ構造を形成することであって、前記活性化データ構造が、前記複数の細胞構成要素モジュールにおける細胞構成要素モジュールの各々について、

前記第２の複数の細胞における細胞の各々について、それぞれの活性化重みを含む、形成することと、

前記複数の共変量におけるそれぞれの共変量の各々について、（ｉ）前記共変量のフィンガープリントの候補細胞構成要素モデルへの入力時に、候補細胞構成要素モデルによって表される細胞構成要素モジュールの各々に対する計算された活性化と、（ｉｉ）前記候補細胞構成要素モデルによって表される細胞構成要素モジュールの各々に対する実際の活性化との間の差を使用して、前記候補細胞構成要素モデルを訓練することであって、前記訓練することが、前記差に応答して、前記候補細胞構成要素モデルと関連付けられた複数の共変量パラメータを調整する、訓練することと、を含む、プロセスによって前記第１の細胞構成要素モジュールを識別することを更に含む、請求項１～２６のいずれか一項に記載の方法。

【請求項２８】

前記複数の共変量パラメータが、

前記複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、

それぞれの共変量の各々について、

前記それぞれの共変量が、前記第２の複数の細胞にわたって、前記それぞれの細胞構成要素モジュールと相関するかどうかを示す対応するパラメータを含み、前記方法が、

前記候補細胞構成要素モデルを訓練する際に前記複数の共変量パラメータを使用して、前記複数の候補細胞構成要素モジュールにおける前記第１の細胞構成要素モジュールを識別することを更に含む、請求項２７に記載の方法。

【請求項２９】

前記複数の注釈付きの細胞状態における注釈付きの細胞状態が、曝露条件下での化合物への前記第１の複数の細胞における細胞の曝露である、請求項２７又は２８に記載の方法。

【請求項３０】

前記曝露条件が、曝露期間、前記化合物の濃度、又は曝露期間及び前記化合物の濃度の組み合わせである、請求項２９に記載の方法。

【請求項３１】

前記複数の細胞構成要素における細胞構成要素の各々が、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせである、請求項１～３０のいずれか一項に記載の方法。

【請求項３２】

前記複数の細胞構成要素における細胞構成要素の各々が、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせであり、

前記第１又は第２の複数の細胞における前記それぞれの細胞における前記それぞれの細胞構成要素の前記対応する存在量が、比色測定、蛍光測定、発光測定、又は共鳴エネルギー移動（ＦＲＥＴ）測定によって決定される、請求項２７～３０のいずれか一項に記載の方法。

【請求項３３】

前記複数の細胞構成要素における細胞構成要素の各々が、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせであり、

前記第１又は第２の複数の細胞における前記それぞれの細胞における前記それぞれの細胞構成要素の前記対応する存在量が、単一細胞リボ核酸（ＲＮＡ）配列決定（ｓｃＲＮＡ－ｓｅｑ）、ｓｃＴａｇ－ｓｅｑ、配列決定を使用したトランスポザーゼ－アクセス可能なクロマチンのための単一細胞アッセイ（ｓｃＡＴＡＣ－ｓｅｑ）、ＣｙＴＯＦ／ＳＣｏＰ、Ｅ－ＭＳ／Ａｂｓｅｑ、ｍｉＲＮＡ－ｓｅｑ、ＣＩＴＥ－ｓｅｑ、又はそれらの任意の組み合わせによって決定される、請求項１１に記載の方法。

【請求項３４】

前記複数のベクトルを使用して、前記複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することが、前記複数のベクトルにおけるベクトルの各々の対応する複数のエレメントの各々を使用して、前記複数のベクトルに相関モデルを適用することを含む、請求項１～３０又は３２～３３のいずれか一項に記載の方法。

【請求項３５】

前記相関モデルが、グラフクラスタリングを含む、請求項３４に記載の方法。

【請求項３６】

前記グラフクラスタリングが、ピアソン相関ベースの距離メトリック上のライデン（Ｌｅｉｄｅｎ）クラスタリングである、請求項３４に記載の方法。

【請求項３７】

前記グラフクラスタリングが、ルーバン（Ｌｏｕｖａｉｎ）クラスタリングである、請求項３４に記載の方法。

【請求項３８】

前記複数の細胞構成要素モジュールが、１０～２０００個の細胞構成要素モジュールからなる、請求項２７～３７のいずれか一項に記載の方法。

【請求項３９】

前記複数の細胞構成要素が、１００～８，０００個の細胞構成要素からなる、請求項２７～３７のいずれか一項に記載の方法。

【請求項４０】

前記複数の構成要素モジュールにおける候補細胞構成要素モジュールの各々が、２００～３００個の細胞構成要素からなる、請求項２７～３７のいずれか一項に記載の方法。

【請求項４１】

前記目的の生理学的状態が、疾患である、請求項１～４０のいずれか一項に記載の方法。

【請求項４２】

前記目的の生理学的状態が、疾患であり、前記第１の複数の細胞が、前記複数の注釈付きの細胞状態によって示されるように、前記疾患を代表する細胞、及び前記疾患を代表しない細胞を含む、請求項２７に記載の方法。

【請求項４３】
前記複数の共変量が、細胞バッチ、細胞ドナー、細胞型、疾患状態、化学化合物への曝露、又はそれらの任意の組み合わせを含む、請求項２７に記載の方法。
【請求項４４】
前記候補細胞構成要素モデルを前記訓練することが、マルチタスク策定におけるカテゴリ交差エントロピー損失を使用して実施され、前記複数の共変量における共変量の各々が、複数のコスト関数におけるコスト関数に対応し、前記複数のコスト関数におけるそれぞれのコスト関数の各々が、共通の重み付け係数を有する、請求項２７に記載の方法。
【請求項４５】
前記試験化学化合物が、２０００ダルトン未満の分子量を有する有機化合物である、請求項１～４４のいずれか一項に記載の方法。
【請求項４６】
前記試験化学化合物が、５つの基準のリピンスキーの法則の各々を満たす有機化合物である、請求項４５に記載の方法。
【請求項４７】
前記試験化学化合物が、５つの基準の前記リピンスキーの法則のうちの少なくとも３つの基準を満たす有機化合物である、請求項４５に記載の方法。
【請求項４８】
前記モデルが、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む、請求項１～１９のいずれか一項に記載の方法。
【請求項４９】
前記方法が、Ｄａｙｌｉｇｈｔ、ＢＣＩ、ＥＣＦＰ４、ＥｃＦＣ、ＭＤＬ、ＡＰＦＰ、ＴＴＦＰ、ＵＮＩＴＹ　２Ｄフィンガープリント、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、前記試験化学化合物の化学構造から前記フィンガープリントを生成することを更に含む、請求項１～４８のいずれか一項に記載の方法。
【請求項５０】
前記細胞構成要素モジュールのセットが、５つ以上の細胞構成要素モジュールを含む、請求項１～１８又は２０～４９のいずれか一項に記載の方法。
【請求項５１】
前記細胞構成要素モジュールのセットが、１０個以上の細胞構成要素モジュールを含む、請求項１～１８又は２０～５０のいずれか一項に記載の方法。
【請求項５２】
前記細胞構成要素モジュールのセットが、１００個以上の細胞構成要素モジュールを含む、請求項１～１８又は２０～５０のいずれか一項に記載の方法。
【請求項５３】
前記それぞれの細胞構成要素モジュールにおける前記複数の細胞構成要素の前記独立したサブセットが、５つ以上の細胞構成要素を含む、請求項１～５２のいずれか一項に記載の方法。
【請求項５４】
前記それぞれの細胞構成要素モジュールにおける前記複数の細胞構成要素の前記独立したサブセットが、前記目的の生理学的状態と関連付けられた分子経路における２～２０個の細胞構成要素からなる、請求項１～５２のいずれか一項に記載の方法。
【請求項５５】
前記第１の閾値基準が、前記第１の細胞構成要素モジュールが閾値活性化スコアを有することが必要である、請求項１～５４のいずれか一項に記載の方法。
【請求項５６】
１つ以上のプロセッサ及びメモリを含むコンピュータシステムであって、前記メモリが、試験化学化合物を目的の生理学的状態と関連付けるための方法を実施するための命令を

格納し、前記方法が、
（Ａ）前記試験化学化合物の化学構造のフィンガープリントを得ることと、
（Ｂ）細胞構成要素モジュールのセットにアクセスすることであって、
前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素のそれぞれの独立したサブセットを含み、
前記複数の細胞構成要素のそれぞれの独立したサブセットの各々についての対応する複数の細胞ベースのアッセイ存在量値が、前記生理学的状態と関連付けられた複数の異なる状態にわたって別々に相関し、
前記細胞構成要素モジュールのセットにおける第１の細胞構成要素モジュールが、前記目的の生理学的状態と関連付けられる、アクセスすることと、
（Ｃ）前記化学構造の前記フィンガープリントを、１００以上のパラメータを含むモデルに入力することに応答して、前記モデルからの出力として、前記細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々についてのそれぞれの活性化スコアを取得することと、
（Ｄ）前記第１の細胞構成要素モジュールについての前記活性化スコアが、第１の閾値基準を満たす場合、前記試験化学化合物を前記目的の生理学的状態と関連付けることと、を含む、コンピュータシステム。
【請求項５７】
試験化学化合物を目的の生理学的状態と関連付けるための、コンピュータによって実行可能な１つ以上のコンピュータプログラムを格納する非一時的なコンピュータ可読媒体であって、前記コンピュータが、１つ以上のプロセッサ及びメモリを含み、前記１つ以上のコンピュータプログラムが、
（Ａ）前記試験化学化合物の化学構造のフィンガープリントを得ることと、
（Ｂ）細胞構成要素モジュールのセットにアクセスすることであって、
前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素のそれぞれの独立したサブセットを含み、
前記複数の細胞構成要素のそれぞれの独立したサブセットの各々についての対応する複数の細胞ベースのアッセイ存在量値が、前記生理学的状態と関連付けられた複数の異なる状態にわたって別々に相関し、
前記細胞構成要素モジュールのセットにおける第１の細胞構成要素モジュールが、前記目的の生理学的状態と関連付けられる、アクセスすることと、
（Ｃ）前記化学構造の前記フィンガープリントを、１００以上のパラメータを含むモデルに入力することに応答して、前記モデルからの出力として、前記細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々についてのそれぞれの活性化スコアを取得することと、
（Ｄ）前記第１の細胞構成要素モジュールについての前記活性化スコアが、第１の閾値基準を満たす場合、前記試験化学化合物を前記目的の生理学的状態と関連付けることと、を含む、方法を実行するコンピュータによって実行可能な命令を集合的に符号化する、非一時的なコンピュータ可読媒体。
【請求項５８】
試験化学化合物を目的の生理学的状態と関連付ける方法であって、前記方法が、
（Ａ）前記試験化学化合物の化学構造のフィンガープリントを得ることと、
（Ｂ）摂動シグネチャのセットにアクセスすることであって、
前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、複数の細胞構成要素のそれぞれの独立したサブセットを含み、
前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、前記それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、前記それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、前記それぞれの第１の細胞状態及び第２

の細胞状態のうちの一方が、非摂動細胞状態であり、前記それぞれの第１の細胞状態及び前記第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である、アクセスすることと、

（Ｃ）前記フィンガープリントをモデルに入力することであって、

前記モデルが、１００以上のパラメータを含み、

前記モデルが、前記フィンガープリントの前記モデルへの前記入力に応答して１つ以上の計算された活性化スコアを出力し、

前記１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々が、前記摂動シグネチャのセットにおける対応する摂動シグネチャを表す、入力することと、

（Ｄ）前記摂動シグネチャのセットにおける第１の摂動シグネチャについての前記それぞれの計算された活性化スコアが、第１の閾値基準を満たす場合、前記化学化合物を前記目的の生理学的状態と関連付けることと、を含む、方法。

【請求項５９】

前記方法が、前記試験化学化合物の単純化された分子入力ラインエントリーシステム（ＳＭＩＬＥＳ）文字列表現から前記フィンガープリントを計算することを更に含む、請求項５８に記載の方法。

【請求項６０】

前記モデルが、ニューラルネットワークを含む、請求項５８又は５９に記載の方法。

【請求項６１】

前記ニューラルネットワークが、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである、請求項６０に記載の方法。

【請求項６２】

前記モデルが、複数のコンポーネントモデルを含むアンサンブルモデルであり、前記複数のコンポーネントモデルにおけるコンポーネントモデルの各々が、前記化学構造の前記フィンガープリントを複数のコンポーネントモデルのセットにおけるコンポーネントモデルの各々に入力することに応答して、前記摂動シグネチャのセットにおける異なる摂動シグネチャについての活性化スコアを提供する、請求項５８～６１のいずれか一項に記載の方法。

【請求項６３】

前記複数のコンポーネントモデルが、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む、請求項６２に記載の方法。

【請求項６４】

前記複数のコンポーネントモデルにおけるコンポーネントモデルの各々が、対応するニューラルネットワークである、請求項６２又は６３に記載の方法。

【請求項６５】

前記対応するニューラルネットワークが、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである、請求項６４に記載の方法。

【請求項６６】

前記複数のコンポーネントモデルにおけるコンポーネントモデルが、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルである、請求項６３又は６４に記載の方法。

【請求項６７】

前記対応するニューラルネットワークが、完全に接続されたニューラルネットワーク及

びメッセージパッシングニューラルネットワークの組み合わせであり、
前記第１のニューラルネットワークの第１の出力及び前記第２のニューラルネットワークの第２の出力が、前記化学構造の前記フィンガープリントを前記完全に接続されたニューラルネットワーク及び前記メッセージパッシングニューラルネットワークに入力することに応答して、組み合わされて、前記摂動シグネチャのセットにおける第１の摂動シグネチャについての前記１つ以上の計算された活性化スコアにおける活性化スコアを決定する、請求項６５に記載の方法。
【請求項６８】
前記摂動シグネチャのセットが、複数の摂動シグネチャであり、
前記第１の摂動シグネチャを含む、前記複数の摂動シグネチャの第１のサブセットが、前記目的の生理学的状態と関連付けられ、
前記複数の摂動シグネチャの第２のサブセットが、前記目的の生理学的状態と関連付けられておらず、
前記第１の摂動シグネチャについての前記それぞれの計算された活性化スコアが、前記第１の閾値基準を満たし、前記複数の摂動シグネチャの前記第２のサブセットにおける摂動シグネチャについての前記それぞれの計算された活性化スコアが、前記第１の閾値基準以外の第２の閾値基準を満たす場合、前記試験化学化合物が、前記目的の生理学的状態と識別される、請求項５８に記載の方法。
【請求項６９】
前記目的の生理学的状態が、疾患である、請求項５８～６８のいずれか一項に記載の方法。
【請求項７０】
前記試験化学化合物が、２０００ダルトン未満の分子量を有する有機化合物である、請求項５８に記載の方法。
【請求項７１】
前記試験化学化合物が、５つの基準のリピンスキーの法則の各々を満たす有機化合物である、請求項７０に記載の方法。
【請求項７２】
前記試験化学化合物が、５つの基準の前記リピンスキーの法則のうちの少なくとも３つの基準を満たす有機化合物である、請求項７０に記載の方法。
【請求項７３】
前記モデルが、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む、請求項５８に記載の方法。
【請求項７４】
前記方法が、Ｄａｙｌｉｇｈｔ、ＢＣＩ、ＥＣＦＰ４、ＥｃＦＣ、ＭＤＬ、ＡＰＦＰ、ＴＴＦＰ、ＵＮＩＴＹ　２Ｄフィンガープリント、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、前記試験化学化合物の化学構造から前記フィンガープリントを生成することを更に含む、請求項５８～７３のいずれか一項に記載の方法。
【請求項７５】
前記摂動シグネチャのセットが、前記第１の摂動シグネチャからなる、請求項５８～７４のいずれか一項に記載の方法。
【請求項７６】
前記摂動シグネチャのセットが、５つ以上の摂動シグネチャを含む、請求項５８～７４のいずれか一項に記載の方法。
【請求項７７】
前記摂動シグネチャのセットが、１０個以上の摂動シグネチャを含む、請求項５８～７４のいずれか一項に記載の方法。
【請求項７８】

前記摂動シグネチャのセットが、１００個以上の摂動シグネチャを含む、請求項５８～７４のいずれか一項に記載の方法。
【請求項７９】
前記第１の閾値基準が、前記第１の摂動シグネチャが閾値活性化スコアを有することが必要である、請求項５８～７４のいずれか一項に記載の方法。
【請求項８０】
１つ以上のプロセッサ及びメモリを含むコンピュータシステムであって、前記メモリが、試験化学化合物を目的の生理学的状態と関連付けるための方法を実施するための命令を格納し、前記方法が、
（Ａ）前記試験化学化合物の化学構造のフィンガープリントを得ることと、
（Ｂ）摂動シグネチャのセットにアクセスすることであって、
前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、複数の細胞構成要素のそれぞれの独立したサブセットを含み、
前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、前記それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、前記それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、前記それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、前記それぞれの第１の細胞状態及び前記第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である、アクセスすることと、
（Ｃ）前記フィンガープリントをモデルに入力することであって、
前記モデルが、１００以上のパラメータを含み、
前記モデルが、前記フィンガープリントの前記モデルへの前記入力に応答して１つ以上の計算された活性化スコアを出力し、
前記１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々が、前記摂動シグネチャのセットにおける対応する摂動シグネチャを表す、入力することと、
（Ｄ）前記摂動シグネチャのセットにおける第１の摂動シグネチャについての前記それぞれの計算された活性化スコアが、第１の閾値基準を満たす場合、前記化学化合物を前記目的の生理学的状態と関連付けることと、を含む、コンピュータシステム。
【請求項８１】
試験化学化合物を目的の生理学的状態と関連付けるための、コンピュータによって実行可能な１つ以上のコンピュータプログラムを格納する非一時的なコンピュータ可読媒体であって、前記コンピュータが、１つ以上のプロセッサ及びメモリを含み、前記１つ以上のコンピュータプログラムが、
（Ａ）前記試験化学化合物の化学構造のフィンガープリントを得ることと、
（Ｂ）摂動シグネチャのセットにアクセスすることであって、
前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、複数の細胞構成要素のそれぞれの独立したサブセットを含み、
前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、前記それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、前記それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、前記それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、前記それぞれの第１の細胞状態及び前記第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である、アクセスすることと、
（Ｃ）前記フィンガープリントをモデルに入力することであって、
前記モデルが、１００以上のパラメータを含み、

前記モデルが、前記フィンガープリントの前記モデルへの前記入力に応答して１つ以上の計算された活性化スコアを出力し、

前記１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々が、前記摂動シグネチャのセットにおける対応する摂動シグネチャを表す、入力することと、

（Ｄ）前記摂動シグネチャのセットにおける第１の摂動シグネチャについての前記それぞれの計算された活性化スコアが、第１の閾値基準を満たす場合、前記化学化合物を前記目的の生理学的状態と関連付けることと、を含む、方法を実行するコンピュータによって実行可能な命令を集合的に符号化する、非一時的なコンピュータ可読媒体。

【請求項８２】

化学化合物を目的の生理学的状態と関連付ける方法であって、前記方法が、

メモリ及び１つ以上のプロセッサを含むコンピュータシステムにおいて、

（Ａ）複数の化合物におけるそれぞれの化合物の各々の対応する化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることと、

（Ｂ）前記複数の化合物における化合物の各々についての細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの数値的活性化スコアを電子形式で得ることであって、前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素の独立したサブセットを含む、得ることと、

（Ｃ）

前記複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、

前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々について、

（ｉ）前記それぞれの化合物の前記化学構造の前記フィンガープリントを訓練されていないモデルに入力したときの前記それぞれの細胞構成要素モジュールについてのそれぞれの計算された活性化スコアと、（ｉｉ）前記細胞構成要素モジュールのセットにおける前記それぞれの化合物についての前記それぞれの細胞構成要素モジュールの前記それぞれの数値的活性化スコアとの間のそれぞれの差を使用して前記訓練されていないモデルを訓練することであって、前記訓練すること（Ｃ）が、前記差に応答して前記訓練されていないモデルと関連付けられた複数のパラメータを調整し、前記複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を前記目的の生理学的状態と関連付ける訓練されたモデルを得る、訓練することと、を含む、方法。

【請求項８３】

前記細胞構成要素モジュールのセットが、単一の細胞構成要素モジュールからなる、請求項８２に記載の方法。

【請求項８４】

前記細胞構成要素モジュールのセットが、複数の細胞構成要素モジュールを含む、請求項８２に記載の方法。

【請求項８５】

前記細胞構成要素モジュールのセットが、２００～５００個の細胞構成要素モジュールからなる、請求項８２に記載の方法。

【請求項８６】

前記複数の化合物が、１０～1×10^{6}個の化合物からなる、請求項８２に記載の方法。

【請求項８７】

前記複数の化合物が、１００～１００，０００個の化合物からなる、請求項８２に記載の方法。

【請求項８８】

前記複数の化合物が、１０００～１００，０００個の化合物からなる、請求項８２に記載の方法。
【請求項８９】
前記訓練すること（Ｃ）が、回帰アルゴリズムに従って、前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々についてのそれぞれの化合物の各々と関連付けられた差の各々に応答して、前記訓練されていないモデルと関連付けられた前記複数のパラメータを調整する、請求項８２～８８のいずれか一項に記載の方法。
【請求項９０】
前記回帰アルゴリズムが、前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々についてのそれぞれの化合物の各々と関連付けられた差の各々の最小二乗誤差を最適化する、請求項８９に記載の方法。
【請求項９１】
前記訓練されたモデルが、ニューラルネットワークを含む、請求項８２～９０のいずれか一項に記載の方法。
【請求項９２】
前記ニューラルネットワークが、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである、請求項９１に記載の方法。
【請求項９３】
前記訓練されたモデルが、複数のコンポーネントモデルのアンサンブルモデルであり、前記複数のコンポーネントモデルにおけるそれぞれのコンポーネントモデルの各々が、前記複数の細胞構成要素モジュールにおける異なる細胞構成要素モジュールについて計算された活性化スコアを出力する、請求項８２～９０のいずれか一項に記載の方法。
【請求項９４】
前記複数のコンポーネントモデルが、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む、請求項９３に記載の方法。
【請求項９５】
前記複数のコンポーネントモデルにおけるコンポーネントモデルの各々が、対応するニューラルネットワークである、請求項９３に記載の方法。
【請求項９６】
前記対応するニューラルネットワークが、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである、請求項９５に記載の方法。
【請求項９７】
前記細胞構成要素モジュールのセットが、複数の細胞構成要素モジュールであり、
前記複数の細胞構成要素モジュールの第１のサブセットが、前記目的の生理学的状態と関連付けられ、
前記複数の細胞構成要素モジュールの第２のサブセットが、前記目的の生理学的状態と関連付けられていない、請求項８２～９６のいずれか一項に記載の方法。
【請求項９８】
前記方法が、
電子形式で１つ以上の第１のデータセットを得、前記１つ以上の第１のデータセットが、
第１の複数の細胞におけるそれぞれの細胞の各々について、前記第１の複数の細胞が、２０個以上の細胞を含み、複数の注釈付きの細胞状態を集合的に表し、
前記複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、前記複数の細胞構成要素が、１０個以上の細胞構成要素を含み、
前記それぞれの細胞における前記それぞれの細胞構成要素の対応する存在量を含

むか、又は集合的に含み、
それによって、複数のベクトルにアクセスするか、又はそれらを形成することであって、前記複数のベクトルにおけるそれぞれのベクトルの各々が、（ｉ）前記複数の構成要素におけるそれぞれの細胞構成要素に対応し、（ｉｉ）対応する複数のエレメントを含み、前記対応する複数のエレメントにおけるそれぞれのエレメントの各々が、前記第１の複数の細胞における前記それぞれの細胞における前記それぞれの細胞構成要素の前記対応する存在量を表す対応するカウントを有する、複数のベクトルにアクセスするか、又はそれらを形成することと、
前記複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することであって、前記複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々が、前記複数の細胞構成要素のサブセットを含み、前記複数の細胞構成要素モジュールが、（ｉ）前記複数の候補細胞構成要素モジュール及び（ｉｉ）前記複数の細胞構成要素又はその表現によって次元決定された潜在表現で配置され、前記複数の細胞構成要素モジュールが、１０を超える細胞構成要素モジュールを含む、識別することと、
電子形式で１つ以上の第２のデータセットを得、前記１つ以上の第２のデータセットが、
第２の複数の細胞におけるそれぞれの細胞の各々について、前記第２の複数の細胞が、２０個以上の細胞を含み、前記目的の生理学的状態を通知する複数の共変量を集合的に表し、
前記複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、
前記それぞれの細胞における前記それぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含み、
それによって、（ｉ）前記第２の複数の細胞及び（ｉｉ）前記複数の細胞構成要素又はその前記表現によって次元決定された細胞構成要素カウントデータ構造を得ることと、
前記複数の細胞構成要素又はその前記表現を共通次元として使用して前記細胞構成要素カウントデータ構造及び前記潜在表現を組み合わせることによって活性化データ構造を形成することであって、前記活性化データ構造が、前記複数の細胞構成要素モジュールにおける細胞構成要素モジュールの各々について、
前記第２の複数の細胞における細胞の各々について、それぞれの活性化重みを含む、形成することと、
（ｉ）前記活性化データ構造を候補モデルに入力したときに、前記活性化データ構造内に表される細胞構成要素モジュールの各々における前記複数の共変量における各共変量の不在又は存在の予測と、（ｉｉ）細胞構成要素モジュールの各々における各共変量の実際の不在又は存在との間の差を使用して、候補細胞構成要素モデルを訓練することであって、前記訓練することが、前記差に応答して、前記候補細胞構成要素モデルと関連付けられた複数の共変量パラメータを調整する、訓練することと、を含む、プロセスによって前記複数の細胞構成要素モジュールにおける細胞構成要素モジュールを識別することを更に含む、請求項８２～９７のいずれか一項に記載の方法。
【請求項９９】
前記複数の共変量パラメータが、
前記複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、
それぞれの共変量の各々について、
前記それぞれの共変量が、前記第２の複数の細胞にわたって、前記それぞれの細胞構成要素モジュールと相関するかどうかを示す対応するパラメータを含み、
前記候補細胞構成要素モデルを訓練する際に前記複数の共変量パラメータを使用して、前記複数の候補細胞構成要素モジュールにおける前記細胞構成要素モジュールを識別する、請求項９８に記載の方法。
【請求項１００】

前記複数の注釈付きの細胞状態における注釈付きの細胞状態が、曝露条件下での化合物への前記第１の複数の細胞における細胞の曝露である、請求項９９に記載の方法。
【請求項１０１】
前記曝露条件が、曝露期間、前記化合物の濃度、又は曝露期間及び前記化合物の濃度の組み合わせである、請求項９９に記載の方法。
【請求項１０２】
前記複数の細胞構成要素における細胞構成要素の各々が、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせである、請求項８２～１０１のいずれか一項に記載の方法。
【請求項１０３】
前記第１又は第２の複数の細胞における前記それぞれの細胞における前記それぞれの細胞構成要素の前記対応する存在量が、比色測定、蛍光測定、発光測定、又は共鳴エネルギー移動（ＦＲＥＴ）測定によって決定される、請求項９８に記載の方法。
【請求項１０４】
前記第１又は第２の複数の細胞における前記それぞれの細胞における前記それぞれの細胞構成要素の前記対応する存在量が、単一細胞リボ核酸（ＲＮＡ）配列決定（ｓｃＲＮＡ－ｓｅｑ）、ｓｃＴａｇ－ｓｅｑ、配列決定を使用したトランスポザーゼ－アクセス可能なクロマチンのための単一細胞アッセイ（ｓｃＡＴＡＣ－ｓｅｑ）、ＣｙＴＯＦ／ＳＣｏＰ、Ｅ－ＭＳ／Ａｂｓｅｑ、ｍｉＲＮＡ－ｓｅｑ、ＣＩＴＥ－ｓｅｑ、又はそれらの任意の組み合わせによって決定される、請求項９８に記載の方法。
【請求項１０５】
前記複数のベクトルを使用して、前記複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することが、前記複数のベクトルにおけるベクトルの各々の対応する複数のエレメントの各々を使用して、前記複数のベクトルに相関モデルを適用することを含む、請求項９８に記載の方法。
【請求項１０６】
前記相関モデルが、グラフクラスタリングを含む、請求項１０５に記載の方法。
【請求項１０７】
前記グラフクラスタリング方法が、ピアソン相関ベースの距離メトリック上のライデン（Ｌｅｉｄｅｎ）クラスタリングであるか、又はルーバン（Ｌｏｕｖａｉｎ）クラスタリングである、請求項１０６に記載の方法。
【請求項１０８】
前記複数の細胞構成要素が、１００～８，０００個の細胞構成要素からなる、請求項８２～１０７のいずれか一項に記載の方法。
【請求項１０９】
前記複数の構成要素モジュールにおける候補細胞構成要素モジュールの各々が、２００～３００個の細胞構成要素からなる、請求項９８に記載の方法。
【請求項１１０】
前記目的の生理学的状態が、疾患である、請求項８２～１０９のいずれか一項に記載の方法。
【請求項１１１】
前記生理学的状態が、疾患であり、前記第１の複数の細胞が、前記複数の注釈付きの細胞状態によって示されるように、前記疾患を代表する細胞、及び前記疾患を代表しない細胞を含む、請求項９８のいずれか一項に記載の方法。
【請求項１１２】
前記複数の共変量が、細胞バッチ、細胞ドナー、細胞型、疾患状態、又は化学化合物への曝露を含む、請求項９８に記載の方法。
【請求項１１３】
前記候補細胞構成要素モデルを前記訓練することが、マルチタスク策定におけるカテゴリ交差エントロピー損失を使用して実施され、前記複数の共変量における共変量の各々が

、複数のコスト関数におけるコスト関数に対応し、前記複数のコスト関数におけるそれぞれのコスト関数の各々が、共通の重み付け係数を有する、請求項９８に記載の方法。
【請求項１１４】
前記複数の化学化合物における化学化合物の各々が、２０００ダルトン未満の分子量を有する有機化合物である、請求項８２～１１３のいずれか一項に記載の方法。
【請求項１１５】
前記複数の化学化合物における化学化合物の各々が、５つの基準のリピンスキーの法則の各々を満たす、請求項８２～１１３のいずれか一項に記載の方法。
【請求項１１６】
前記複数の化学化合物における化学化合物の各々が、５つの基準の前記リピンスキーの法則のうちの少なくとも３つの基準を満たす、請求項８２～１１３のいずれか一項に記載の方法。
【請求項１１７】
前記訓練されたモデルが、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む、請求項８２～１１６のいずれか一項に記載の方法。
【請求項１１８】
前記方法が、Ｄａｙｌｉｇｈｔ、ＢＣＩ、ＥＣＦＰ４、ＥｃＦＣ、ＭＤＬ、ＡＰＦＰ、ＴＴＦＰ、ＵＮＩＴＹ　２Ｄフィンガープリント、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、前記対応する化学構造からそれぞれのフィンガープリントの各々を生成することを更に含む、請求項８２～１１７のいずれか一項に記載の方法。
【請求項１１９】
前記細胞構成要素モジュールのセットが、５つ以上の細胞構成要素モジュールを含む、請求項８２に記載の方法。
【請求項１２０】
前記細胞構成要素モジュールのセットが、１０個以上の細胞構成要素モジュールを含む、請求項８２に記載の方法。
【請求項１２１】
前記細胞構成要素モジュールのセットが、１００個以上の細胞構成要素モジュールを含む、請求項８２に記載の方法。
【請求項１２２】
１つ以上のプロセッサ及びメモリを含むコンピュータシステムであって、前記メモリが、化学化合物を目的の生理学的状態と関連付けるための方法を実施するための命令を格納し、前記方法が、
（Ａ）複数の化合物におけるそれぞれの化合物の各々の対応する化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることと、
（Ｂ）前記複数の化合物における化合物の各々についての細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの数値的活性化スコアを電子形式で得ることであって、前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素の独立したサブセットを含む、得ることと、
（Ｃ）
前記複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、
前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々について、
（ｉ）前記それぞれの化合物の前記化学構造の前記フィンガープリントを訓練

されていないモデルに入力したときの前記それぞれの細胞構成要素モジュールについてのそれぞれの計算された活性化スコアと、（ｉｉ）前記細胞構成要素モジュールのセットにおける前記それぞれの化合物についての前記それぞれの細胞構成要素モジュールの前記それぞれの数値的活性化スコアとの間のそれぞれの差を使用して前記訓練されていないモデルを訓練することであって、前記訓練すること（Ｃ）が、前記差に応答して前記訓練されていないモデルと関連付けられた複数のパラメータを調整し、前記複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を前記目的の生理学的状態と関連付ける訓練されたモデルを得ることと、を含む、コンピュータシステム。

【請求項１２３】

化学化合物を目的の生理学的状態と関連付けるための、コンピュータによって実行可能な１つ以上のコンピュータプログラムを格納する非一時的なコンピュータ可読媒体であって、前記コンピュータが、１つ以上のプロセッサ及びメモリを含み、前記１つ以上のコンピュータプログラムが、

（Ａ）複数の化合物におけるそれぞれの化合物の各々の対応する化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることと、

（Ｂ）前記複数の化合物における化合物の各々についての細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの数値的活性化スコアを電子形式で得ることであって、前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素の独立したサブセットを含む、得ることと、

（Ｃ）

前記複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、

前記細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々について、

（ｉ）前記それぞれの化合物の前記化学構造の前記フィンガープリントを訓練されていないモデルに入力したときの前記それぞれの細胞構成要素モジュールについてのそれぞれの計算された活性化スコアと、（ｉｉ）前記細胞構成要素モジュールのセットにおける前記それぞれの化合物についての前記それぞれの細胞構成要素モジュールの前記それぞれの数値的活性化スコアとの間のそれぞれの差を使用して前記訓練されていないモデルを訓練することであって、前記訓練すること（Ｃ）が、前記差に応答して前記訓練されていないモデルと関連付けられた複数のパラメータを調整し、前記複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を前記目的の生理学的状態と関連付ける訓練されたモデルを得ることと、を含む、方法を実行するコンピュータによって実行可能な命令を集合的に符号化する、非一時的なコンピュータ可読媒体。

【請求項１２４】

化学化合物を目的の生理学的状態と関連付ける方法であって、前記方法が、

メモリ及び１つ以上のプロセッサを含むコンピュータシステムにおいて、

（Ａ）複数の化合物におけるそれぞれの化合物の各々の対応する化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることと、

（Ｂ）前記複数の化合物における対応する化合物の各々についての摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々のそれぞれの数値的活性化スコアを電子形式で得ることであって、前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、前記それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、前記それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、前記それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、前記それぞれの

第１の細胞状態及び前記第２の細胞状態のうちの他方が、前記対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である、得ることと、

（Ｃ）

前記複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、

前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々について、

（ｉ）前記それぞれの化合物の前記化学構造の前記フィンガープリントを訓練されていないモデルに入力したときの前記それぞれの摂動シグネチャについてのそれぞれの計算された活性化スコアと、（ｉｉ）前記摂動シグネチャのセットにおける前記対応する化合物についての前記それぞれの摂動シグネチャの前記それぞれの数値的活性化スコアとの間のそれぞれの差を使用して前記訓練されていないモデルを訓練することであって、前記訓練すること（Ｃ）が、前記差に応答して、前記訓練されていないモデルと関連付けられた複数のパラメータを調整し、前記複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を前記目的の生理学的状態と関連付ける訓練されたモデルを得る、訓練することと、を含む、方法。

【請求項１２５】

前記摂動シグネチャのセットが、単一の摂動シグネチャからなる、請求項１２４に記載の方法。

【請求項１２６】

前記摂動シグネチャのセットが、２００～５００個の摂動シグネチャからなる、請求項１２４に記載の方法。

【請求項１２７】

前記複数の化合物が、１０～1×10^6個の化合物からなる、請求項１２４～１２６のいずれか一項に記載の方法。

【請求項１２８】

前記複数の化合物が、１００～１００，０００個の化合物からなる、請求項１２４～１２６のいずれか一項に記載の方法。

【請求項１２９】

前記複数の化合物が、１０００～１００，０００個の化合物からなる、請求項１２４～１２６のいずれか一項に記載の方法。

【請求項１３０】

前記訓練すること（Ｃ）が、回帰アルゴリズムに従って、前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々についての対応する化合物の各々と関連付けられた差の各々に応答して、前記訓練されていないモデルと関連付けられた前記複数のパラメータを調整する、請求項１２４～１２９のいずれか一項に記載の方法。

【請求項１３１】

前記回帰アルゴリズムが、前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々についての対応する化合物の各々と関連付けられた差の各々の最小二乗誤差を最適化する、請求項１３０に記載の方法。

【請求項１３２】

前記訓練されたモデルが、ニューラルネットワークを含む、請求項１２４～１３１のいずれか一項に記載の方法。

【請求項１３３】

前記ニューラルネットワークが、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである、請求項１３２に記載の方法。

【請求項１３４】

前記訓練されたモデルが、複数のコンポーネントモデルのアンサンブルモデルであり、前記複数のコンポーネントモデルにおけるそれぞれのコンポーネントモデルの各々が、そ

れぞれの化学構造のフィンガープリントを複数のコンポーネントモデルのセットにおけるコンポーネントモデルの各々に入力することに応答して、前記複数の摂動シグネチャのセットにおける異なる摂動シグネチャのセットについて計算された活性化スコアを出力する、請求項１２４に記載の方法。

【請求項１３５】

前記複数のコンポーネントモデルが、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む、請求項１３４に記載の方法。

【請求項１３６】

前記複数のコンポーネントモデルにおけるコンポーネントモデルの各々が、対応するニューラルネットワークである、請求項１３４に記載の方法。

【請求項１３７】

前記対応するニューラルネットワークが、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである、請求項１３６に記載の方法。

【請求項１３８】

前記摂動シグネチャのセットが、複数の摂動シグネチャを含み、

前記複数の摂動シグネチャの第１のサブセットが、前記目的の生理学的状態と関連付けられ、

前記複数の摂動シグネチャの第２のサブセットが、前記目的の生理学的状態と関連付けられていない、請求項１２４～１３７のいずれか一項に記載の方法。

【請求項１３９】

前記目的の生理学的状態が、疾患である、請求項１２４～１３８のいずれか一項に記載の方法。

【請求項１４０】

前記複数の化学化合物における化学化合物の各々が、２０００ダルトン未満の分子量を有する有機化合物である、請求項１２４～１３９のいずれか一項に記載の方法。

【請求項１４１】

前記複数の化学化合物における化学化合物の各々が、５つの基準のリピンスキーの法則の各々を満たす、請求項１２４～１４０のいずれか一項に記載の方法。

【請求項１４２】

前記複数の化学化合物における化学化合物の各々が、５つの基準の前記リピンスキーの法則のうちの少なくとも３つの基準を満たす、請求項１２４～１４０のいずれか一項に記載の方法。

【請求項１４３】

前記訓練されたモデルが、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシン、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む、請求項１２４に記載の方法。

【請求項１４４】

前記方法が、Ｄａｙｌｉｇｈｔ、ＢＣＩ、ＥＣＦＰ４、ＥｃＦＣ、ＭＤＬ、ＡＰＦＰ、ＴＴＦＰ、ＵＮＩＴＹ　２Ｄフィンガープリント、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、前記対応する化学構造からそれぞれのフィンガープリントの各々を生成することを更に含む、請求項１２４～１４３のいずれか一項に記載の方法。

【請求項１４５】

前記摂動シグネチャのセットが、５つ以上の摂動シグネチャを含む、請求項１２４に記載の方法。

【請求項１４６】

前記摂動シグネチャのセットが、１０個以上の摂動シグネチャを含む、請求項１２４に

記載の方法。
【請求項１４７】
前記摂動シグネチャのセットが、１００個以上の摂動シグネチャを含む、請求項１２４に記載の方法。
【請求項１４８】
前記方法が、
変化していない細胞状態と変化した細胞状態との間の差次的細胞構成要素存在量の尺度を表す単一細胞遷移シグネチャに電子形式でアクセスすることであって、
前記変化した細胞状態が、前記変化していない細胞状態から前記変化した細胞状態への前記細胞遷移を通して発生し、
（ｉ）前記変化していない細胞状態、（ｉｉ）前記変化した細胞状態、及び（ｉｉｉ）前記変化していない細胞状態から前記変化した細胞状態への前記遷移のうちの少なくとも１つが、前記目的の生理学的状態と関連付けられ、
前記単一細胞遷移シグネチャが、参照の複数の細胞構成要素の識別と、前記複数の参照細胞構成要素におけるそれぞれの細胞構成要素の各々について、前記それぞれの細胞構成要素の存在量の変化と、前記変化していない細胞状態と前記変化した細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する第１の有意性スコアとを含む、アクセスすることと、
前記単一細胞遷移シグネチャと前記それぞれの摂動シグネチャとを比較し、それによって前記それぞれの摂動シグネチャの前記それぞれの数値的活性化スコアを決定することと、を含む、手順によって前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアを得ることを更に含む、請求項１２４に記載の方法。
【請求項１４９】
前記単一細胞遷移シグネチャと前記摂動シグネチャとを前記比較して、前記それぞれの摂動シグネチャの前記それぞれの数値的活性化スコアを決定することが、前記単一細胞遷移シグネチャの前記参照の複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、
前記それぞれの摂動シグネチャにおける前記対応する細胞構成要素の前記対応する有意性スコアに対する前記それぞれの細胞構成要素の前記第１の有意性スコアを比較することを含む、請求項１４８に記載の方法。
【請求項１５０】
前記それぞれの摂動シグネチャの前記活性化スコアが、前記摂動シグネチャのセットにおける他の摂動シグネチャと比較して、前記単一細胞遷移シグネチャに対する前記それぞれの摂動シグネチャの関連性の相対的なランキングである、請求項１４８又は１４９に記載の方法。
【請求項１５１】
前記相対的なランキングが、ウィルコクソンの順位和検定、ｔ検定、ロジスティック回帰、又は一般化線形モデルによって決定される、請求項１５０に記載の方法。
【請求項１５２】
前記単一細胞遷移シグネチャの前記変化していない細胞状態が、前記それぞれの摂動シグネチャの前記第１の細胞状態又は前記第２の細胞状態と同じである、請求項１４８～１５１のいずれか一項に記載の方法。
【請求項１５３】
前記単一細胞遷移シグネチャの前記変化していない細胞状態が、前記それぞれの摂動シグネチャの前記第１の細胞状態及び前記第２の細胞状態の両方とは異なる、請求項１４８～１５１のいずれか一項に記載の方法。
【請求項１５４】
前記方法が、
前記単一細胞遷移シグネチャの前記参照の複数の細胞構成要素、及び前記それぞれの摂動シグネチャの前記それぞれの複数の細胞構成要素を剪定して、転写因子と比較するこ

とを制限することを更に含む、請求項１４８～１５３のいずれか一項に記載の方法。
【請求項１５５】
前記複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの前記摂動細胞状態が、前記複数の化合物における化合物に曝露されていない対照細胞によって表される、請求項１２４～１５４のいずれか一項に記載の方法。
【請求項１５６】
前記複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの前記摂動細胞状態が、前記それぞれの摂動シグネチャと関連付けられた前記化合物以外の前記複数の化学化合物における化学化合物に曝露されている無関係の摂動細胞にわたる平均によって表される、請求項１２４～１５４のいずれか一項に記載の方法。
【請求項１５７】
１つ以上のプロセッサ及びメモリを含むコンピュータシステムであって、前記メモリが、化学化合物を目的の生理学的状態と関連付けるための命令を格納し、前記方法が、
（Ａ）複数の化合物におけるそれぞれの化合物の各々の対応する化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることと、
（Ｂ）前記複数の化合物における対応する化合物の各々についての摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々のそれぞれの数値的活性化スコアを電子形式で得ることであって、前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、前記それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、前記それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、前記それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、前記それぞれの第１の細胞状態及び前記第２の細胞状態のうちの他方が、前記対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である、得ることと、
（Ｃ）
前記複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、
前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々について、
（ｉ）前記それぞれの化合物の前記化学構造の前記フィンガープリントを訓練されていないモデルに入力したときの前記それぞれの摂動シグネチャについてのそれぞれの計算された活性化スコアと、（ｉｉ）前記摂動シグネチャのセットにおける前記対応する化合物についての前記それぞれの摂動シグネチャの前記それぞれの数値的活性化スコアとの間のそれぞれの差を使用して前記訓練されていないモデルを訓練することであって、前記訓練すること（Ｃ）が、前記差に応答して、前記訓練されていないモデルと関連付けられた複数のパラメータを調整し、前記複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を前記目的の生理学的状態と関連付ける訓練されたモデルを得る、訓練することと、を含む、コンピュータシステム。
【請求項１５８】
化学化合物を目的の生理学的状態と関連付けるための、コンピュータによって実行可能な１つ以上のコンピュータプログラムを格納する非一時的なコンピュータ可読媒体であって、前記コンピュータが、１つ以上のプロセッサ及びメモリを含み、前記１つ以上のコンピュータプログラムが、
（Ａ）複数の化合物におけるそれぞれの化合物の各々の対応する化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることと、
（Ｂ）前記複数の化合物における対応する化合物の各々についての摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々のそれぞれの数値的活性化スコアを電子

形式で得ることであって、前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、前記それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、前記それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、前記それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、前記それぞれの第１の細胞状態及び前記第２の細胞状態のうちの他方が、前記対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である、得ることと、

（Ｃ）

前記複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、

前記摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々について、

（ｉ）前記それぞれの化合物の前記化学構造の前記フィンガープリントを訓練されていないモデルに入力したときの前記それぞれの摂動シグネチャについてのそれぞれの計算された活性化スコアと、（ｉｉ）前記摂動シグネチャのセットにおける前記対応する化合物についての前記それぞれの摂動シグネチャの前記それぞれの数値的活性化スコアとの間のそれぞれの差を使用して前記訓練されていないモデルを訓練することであって、前記訓練すること（Ｃ）が、前記差に応答して、前記訓練されていないモデルと関連付けられた複数のパラメータを調整し、前記複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を前記目的の生理学的状態と関連付ける訓練されたモデルを得る、訓練することと、を含む、方法を実行するコンピュータによって実行可能な命令を集合的に符号化する、非一時的なコンピュータ可読媒体。

【請求項１５９】

前記モデルが、リグレッサーである、先行請求項のいずれか一項に記載の方法。

【発明の詳細な説明】

【技術分野】

【０００１】

関連出願の相互参照

本出願は、２０２１年６月１５日に出願された「ＳＹＳＴＥＭＳ ＡＮＤ ＭＥＴＨＯＤＳ ＦＯＲ ＡＳＳＯＣＩＡＴＩＮＧ ＣＯＭＰＯＵＮＤＳ ＷＩＴＨ ＰＨＹＳＩＯＬＯＧＩＣＡＬ ＣＯＮＤＩＴＩＯＮＳ ＵＳＩＮＧ ＦＩＮＧＥＲＰＲＩＮＴ ＡＮＡＬＹＳＩＳ」と題される米国仮特許出願第６３／２１０，９３０号、及び２０２１年６月１５日に出願された「ＣＯＭＰＵＴＡＴＩＯＮＡＬ ＭＯＤＥＬＩＮＧ ＰＬＡＴＦＯＲＭ」と題される６３／２１０，６７９号の優先権を主張し、これらの各々は参照によりその全体が本明細書に組み込まれる。

【０００２】

本発明は、概して、化合物を生理学的状態と関連付けるためのシステム及び方法に関する。

【背景技術】

【０００３】

細胞メカニズムの研究は、疾患を理解するために重要である。

【０００４】

生物組織は、動的かつ高度にネットワーク化された多細胞系である。特定の細胞における細胞内ネットワークの機能障害は、細胞行動の全体像を変化させ、疾患状態につながる。現在の創薬の努力は、細胞を健康な状態から疾患の状態へと遷移させる分子メカニズムを特徴付けることを目指し、これらの遷移を逆転又は阻害する薬理学的アプローチを特定する。これまでの努力はまた、これらの遷移を特徴付ける分子的特徴を特定し、これらの特徴を逆転させる薬理学的アプローチを特定することを目指していた。

【０００５】

表面マーカーによって濃縮された組織又は細胞における細胞のバルク集合に関する分子データは、集団における個々の細胞の表現型及び分子多様性をマスクする。これらの細胞のバルク集合における細胞の不均一性は、疾患駆動メカニズムを解明することを目的とした現在の努力の結果を、誤解させるか、又は完全に不正確にさえさせる。単一細胞ＲＮＡ配列決定などの新しいアプローチは、分子レベルで個々の細胞を特徴付けることができる。これらのデータは、より高い解像度で様々な細胞状態を理解するための基質を提供し、細胞が有する豊富で顕著な状態の多様性を明らかにする。

【０００６】

単一細胞データ、すなわち、これらのデータのまばらさ、細胞内に存在する分子の存在の見落とし、及びノイズを解釈する際に、これらの分子測定の精度に不確実性を伴う重大な課題が存在する。したがって、個々の細胞状態を制御するための薬理学的アプローチへの洞察を導き出し、それに応じて疾患を解決するために、新しいアプローチが必要である。

【０００７】

更に、複雑な疾患は、多くの場合、単一又はいくつかの分子標的に分解することができない。インビトロ疾患モデルのためのハイスループットイメージング技術及びハイスループットスクリーニングの最近の進歩にもかかわらず、インビトロベースのスクリーニングアプローチから生成された候補標的を有効な薬物に変換することは、多くの場合、比較的遅く、非効率的な分子標的ベースの創薬アプローチへの回帰を伴うかなりのタスクである。

【０００８】

上記の背景を考慮すると、当該技術分野で必要とされるのは、創薬のための候補化合物を識別するためのシステム及び方法である。

【発明の概要】

【０００９】

本開示は、上記で特定された欠点に対処する。本開示は、少なくとも部分的に、目的の生理学的状態（例えば、表現型、疾患、細胞状態、及び／又は目的の細胞プロセス）に対応する細胞構成要素データ（例えば、遺伝子の存在量及び／又は摂動シグネチャ）、並びに潜在表現及び機械学習を使用して、細胞構成要素のモジュール（例えば、サブセット）と、目的の生理学的状態との間の関連性（例えば、重み及び／又は相関）を決定することによって、これらの欠点に対処する。特に、本開示は、疾患などの様々な生理学的状態の基礎となる分子メカニズムを解明するためのシステム及び方法を提供する。

【００１０】

本開示の一態様は、試験化学化合物を目的の生理学的状態と関連付ける方法を提供する。この方法は、（Ａ）試験化学化合物の化学構造のフィンガープリントを得ることを含む。

【００１１】

この方法は、（Ｂ）細胞構成要素モジュールのセットにアクセスすることを更に含む。細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々は、複数の細胞構成要素のそれぞれの独立したサブセットを含む。複数の細胞構成要素のそれぞれの独立したサブセットの各々についての対応する複数の細胞ベースのアッセイ存在量値は、生理学的状態と関連付けられた複数の異なる状態にわたって別々に相関する。細胞構成要素モジュールのセットにおける第１の細胞構成要素モジュールは、目的の生理学的状態と関連付けられる。

【００１２】

この方法は、（Ｃ）化学構造のフィンガープリントをモデルに入力することに応答して、モデルからの出力として、細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々についてのそれぞれの活性化スコアを取得することを更に含む。いくつかの実施形態において、モデルは、５０以上のパラメータ、１００以上のパラメータ、１０００以上のパラメータ、又は１０，０００以上のパラメータを含む。

【００１３】
この方法は、（Ｄ）第１の細胞構成要素モジュールについての活性化スコアが、第１の閾値基準を満たす場合、試験化学化合物を目的の生理学的状態と関連付けることを更に含む。
【００１４】
いくつかの実施形態において、細胞ベースのアッセイ存在量値は、臓器の細胞のものである。いくつかのそのような実施形態において、臓器は、心臓、肝臓、肺、筋肉、脳、膵臓、脾臓、腎臓、小腸、子宮、又は膀胱である。
【００１５】
いくつかの実施形態において、細胞ベースのアッセイ存在量値は、組織の細胞のものである。いくつかの実施形態において、組織は、骨、軟骨、関節、気管、脊髄、角膜、眼、皮膚、又は血管である。
【００１６】
いくつかの実施形態において、細胞ベースのアッセイ存在量値は、複数の幹細胞の細胞のものである。いくつかの実施形態において、複数の幹細胞は、複数の胚性幹細胞、複数の成体幹細胞、又は複数の人工多能性幹細胞（ｉＰＳＣ）である。
【００１７】
いくつかの実施形態において、細胞ベースのアッセイ存在量値は、複数の初代ヒト細胞の細胞のものである。いくつかのそのような実施形態において、複数の初代ヒト細胞は、複数のＣＤ３４＋細胞、複数のＣＤ３４＋造血幹、複数の前駆細胞（ＨＳＰＣ）、複数のＴ細胞、複数の間葉系幹細胞（ＭＳＣ）、複数の気道基底幹細胞、又は複数の人工多能性幹細胞である。
【００１８】
いくつかの実施形態において、細胞ベースのアッセイ存在量値は、臍帯血中、末梢血中、又は骨髄中の細胞のものである。
【００１９】
いくつかの実施形態において、細胞ベースのアッセイ存在量値は、固体組織中の細胞のものである。いくつかのそのような実施形態において、固体組織は、胎盤、肝臓、心臓、脳、腎臓、又は胃腸管である。
【００２０】
いくつかの実施形態において、細胞ベースのアッセイ存在量値は、複数の分化細胞のものである。いくつかのそのような実施形態において、複数の分化細胞は、複数の巨核球、複数の骨芽細胞、複数の軟骨細胞、複数の脂肪細胞、複数の肝細胞、複数の肝中皮細胞、複数の胆管上皮細胞、複数の肝星細胞、複数の肝類洞内皮細胞、複数のクッパー細胞、複数のピット細胞、複数の血管内皮細胞、複数の膵管上皮細胞、複数の膵管細胞、複数の腺房中心細胞、複数の腺房細胞、複数のランゲルハンス島、複数の心筋細胞、複数の線維芽細胞、複数のケラチノサイト、複数の平滑筋細胞、複数のＩ型肺胞上皮細胞、複数のＩＩ型肺胞上皮細胞、複数のクララ細胞、複数の線毛上皮細胞、複数の基底細胞、複数の杯細胞、複数の神経内分泌細胞、複数のクルチッキー（ｋｕｌｔｓｃｈｉｔｚｋｙ）細胞、複数の尿細管上皮細胞、複数の尿路上皮細胞、複数の円柱上皮細胞、複数の糸球体上皮細胞、複数の糸球体内皮細胞、複数の有足細胞、複数のメサンギウム細胞、複数の神経細胞、複数の星状膠細胞、複数の小膠細胞、又は複数の乏突起膠細胞である。
【００２１】
いくつかの実施形態において、対応する複数の細胞ベースのアッセイ存在量値は、複数の細胞の単一細胞リボ核酸（ＲＮＡ）配列決定（ｓｃＲＮＡ－ｓｅｑ）データである。いくつかのそのような実施形態において、生理学的状態に関連付けられた複数の異なる状態は、細胞のアリコートが生理学的状態に影響を与えることが知られている化合物に曝露されている対照状態に加えて、生理学的状態に影響を与えることが知られている１つ以上の参照化合物に異なる細胞のアリコートを曝露することによって導出される。
【００２２】

いくつかの実施形態において、対応する複数の細胞ベースのアッセイ存在量値は、バルクＲＮＡ配列に由来する。

【００２３】

いくつかの実施形態において、対応する複数の細胞ベースのアッセイ存在量値は、単一細胞ＲＮＡ配列決定に由来する。

【００２４】

いくつかの実施形態において、細胞構成要素モジュールのセットは、第１の細胞構成要素モジュールからなる。

【００２５】

いくつかの実施形態において、細胞構成要素モジュールのセットは、複数の細胞構成要素モジュールを含み、モデルは、複数のコンポーネントモデルを含むアンサンブルモデルである。複数のコンポーネントモデルにおけるコンポーネントモデルの各々は、化学構造のフィンガープリントを複数のコンポーネントモデルにおけるコンポーネントモデルの各々に入力することに応答して、細胞構成要素モジュールのセットにおける異なる細胞構成要素モジュールについての活性化スコアを提供する。

【００２６】

いくつかの実施形態において、この方法は、試験化学化合物の単純化された分子入力ラインエントリーシステム（ＳＭＩＬＥＳ）文字列表現からフィンガープリントを計算することを更に含む。

【００２７】

いくつかのそのような実施形態において、複数のコンポーネントモデルにおけるコンポーネントモデルの各々は、対応するニューラルネットワーク（例えば、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせ）である。いくつかの実施形態において、対応するニューラルネットワークは、対応する完全に接続されたニューラルネットワーク及び対応するメッセージパッシングニューラルネットワークの組み合わせであり、対応する完全に接続されたニューラルネットワークの第１の出力及び対応するメッセージパッシングニューラルネットワークの第２の出力は、化学構造のフィンガープリントを対応する完全に接続されたニューラルネットワーク及び対応するメッセージパッシングニューラルネットワークに入力することに応答して組み合わされ、細胞構成要素モジュールのセットにおける対応する細胞構成要素モジュールについての１つ以上の計算された活性化スコアにおける活性化スコアを決定する。

【００２８】

いくつかのそのような実施形態において、複数のコンポーネントモデルにおけるコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルである。

【００２９】

いくつかの実施形態において、細胞構成要素モジュールのセットは、複数の細胞構成要素モジュールであり、第１の細胞構成要素モジュールを含む複数の細胞構成要素モジュールの第１のサブセットは、目的の生理学的状態と関連付けられ、複数の細胞構成要素モジュールの第２のサブセットは、目的の生理学的状態と関連付けられず、第１の細胞構成要素モジュールについてのそれぞれの計算された活性化スコアが、第１の閾値基準を満たし、複数の細胞構成要素モジュールの第２のサブセットにおける細胞構成要素モジュールについてのそれぞれの計算された活性化スコアが、第１の閾値基準以外の第２の閾値基準を満たす場合、試験化学化合物は、目的の生理学的状態と識別される。

【００３０】

いくつかの実施形態において、この方法は、電子形式で１つ以上の第１のデータセットを得、１つ以上の第１のデータセットが、第１の複数の細胞におけるそれぞれの細胞の各々について、第１の複数の細胞が、２０個以上の細胞を含み、複数の注釈付きの細胞状態

を集合的に表し、複数の細胞構成要素（例えば、少なくとも１０、２０、３０、１００、若しくは１０００個以上の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含み、それによって複数のベクトルにアクセスするか、又はそれらを形成することであって、複数のベクトルにおけるそれぞれのベクトルの各々が、（ｉ）複数の構成要素におけるそれぞれの細胞構成要素に対応し、（ｉｉ）対応する複数のエレメントを含み、対応する複数のエレメントにおけるそれぞれのエレメントの各々が、第１の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を表す対応するカウントを有する、複数のベクトルにアクセスするか、又はそれらを形成すること、を含む、プロセスによって第１の細胞構成要素モジュールを識別することを更に含む。この方法は、複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することであって、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々が、複数の細胞構成要素のサブセットを含み、複数の細胞構成要素モジュールが、（ｉ）複数の候補細胞構成要素モジュール及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された潜在表現で配置され、複数の細胞構成要素モジュールが、１０を超える細胞構成要素モジュールを含む、識別すること、を更に含む。この方法は、電子形式で１つ以上の第２のデータセットを得、１つ以上の第２のデータセットが、第２の複数の細胞におけるそれぞれの細胞の各々について、第２の複数の細胞が、２０個以上の細胞を含み、目的の生理学的状態を通知する複数の共変量を集合的に表し、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含み、それによって、（ｉ）第２の複数の細胞、及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された細胞構成要素カウントデータ構造を得る。この方法は、複数の細胞構成要素又はその表現を共通次元として使用して細胞構成要素カウントデータ構造及び潜在表現を組み合わせることによって活性化データ構造を形成することであって、活性化データ構造が、複数の細胞構成要素モジュールにおける細胞構成要素モジュールの各々について、第２の複数の細胞における細胞の各々について、それぞれの活性化重みを含む、形成することと、複数の共変量におけるそれぞれの共変量の各々について、（ｉ）共変量のフィンガープリントの候補細胞構成要素モデルへの入力時に、候補細胞構成要素モデルによって表される細胞構成要素モジュールの各々に対する計算された活性化と、（ｉｉ）候補細胞構成要素モデルによって表される細胞構成要素モジュールの各々に対する実際の活性化との間の差を使用して、候補細胞構成要素モデルを訓練することであって、訓練することが、差に応答して、候補細胞構成要素モデルと関連付けられた複数の共変量パラメータを調整する、訓練することと、を更に含む。いくつかのそのような実施形態において、複数の共変量パラメータは、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの共変量の各々について、それぞれの共変量が、第２の複数の細胞にわたって、それぞれの細胞構成要素モジュールと相関するかどうかを示す対応するパラメータを含み、方法が、候補細胞構成要素モデルを訓練する際に複数の共変量パラメータを使用して、複数の候補細胞構成要素モジュールにおける第１の細胞構成要素モジュールを識別することを更に含む。いくつかのそのような実施形態において、この方法は、複数の注釈付きの細胞状態における注釈付きの細胞状態が、曝露条件下での化合物への第１の複数の細胞における細胞の曝露（例えば、曝露期間、化合物の濃度、又は曝露期間及び化合物の濃度の組み合わせ）である、注釈付きの細胞状態を更に含む。

【００３１】

いくつかの実施形態において、複数の細胞構成要素における細胞構成要素の各々は、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせである。

【００３２】

いくつかの実施形態において、複数の細胞構成要素における細胞構成要素の各々は、特

定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせであり、第１又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量が、比色測定、蛍光測定、発光測定、又は共鳴エネルギー移動（ＦＲＥＴ）測定によって決定される。

【００３３】

いくつかの実施形態において、複数の細胞構成要素における細胞構成要素の各々は、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせであり、第１又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量が、単一細胞リボ核酸（ＲＮＡ）配列決定（ｓｃＲＮＡ－ｓｅｑ）、ｓｃＴａｇ－ｓｅｑ、配列決定を使用したトランスポザーゼ－アクセス可能なクロマチンのための単一細胞アッセイ（ｓｃＡＴＡＣ－ｓｅｑ）、ＣｙＴＯＦ／ＳＣｏＰ、Ｅ－ＭＳ／Ａｂｓｅｑ、ｍｉＲＮＡ－ｓｅｑ、ＣＩＴＥ－ｓｅｑ、又はそれらの任意の組み合わせによって決定される。

【００３４】

いくつかの実施形態において、複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することは、複数のベクトルにおけるベクトルの各々の対応する複数のエレメントの各々を使用して、複数のベクトルに相関モデルを適用することを含む。いくつかのそのような実施形態において、相関モデルは、グラフクラスタリング（例えば、ピアソン相関ベースの距離メトリック上のライデン（Ｌｅｉｄｅｎ）クラスタリング、ルーバン（Ｌｏｕｖａｉｎ）クラスタリングなど）を含む。

【００３５】

いくつかの実施形態において、複数の細胞構成要素モジュールは、１０～２０００個の細胞構成要素モジュール、又は１００～８，０００個の細胞構成要素からなる。いくつかの実施形態において、複数の構成要素モジュールにおける候補細胞構成要素モジュールの各々は、２００～３００個の細胞構成要素からなる。

【００３６】

いくつかの実施形態において、目的の生理学的状態は、疾患である。

【００３７】

いくつかの実施形態において、目的の生理学的状態は、疾患であり、第１の複数の細胞が、複数の注釈付きの細胞状態によって示されるように、疾患を代表する細胞、及び疾患を代表しない細胞を含む。

【００３８】

いくつかの実施形態において、複数の共変量は、細胞バッチ、細胞ドナー、細胞型、疾患状態、化学化合物への曝露、又はそれらの任意の組み合わせを含む。

【００３９】

いくつかの実施形態において、候補細胞構成要素モデルを訓練することは、マルチタスク策定におけるカテゴリ交差エントロピー損失を使用して実施され、複数の共変量における共変量の各々が、複数のコスト関数におけるコスト関数に対応し、複数のコスト関数におけるそれぞれのコスト関数の各々が、共通の重み付け係数を有する。

【００４０】

いくつかの実施形態において、試験化学化合物は、２０００ダルトン未満の分子量を有する有機化合物である。いくつかのそのような実施形態において、試験化学化合物は、５つの基準のリピンスキーの法則の各々を満たす有機化合物である。いくつかの実施形態において、試験化学化合物は、５つの基準のリピンスキーの法則のうちの少なくとも３つの基準を満たす有機化合物である。いくつかの実施形態において、モデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。

【００４１】

いくつかの実施形態において、この方法は、Ｄａｙｌｉｇｈｔ、ＢＣＩ、ＥＣＦＰ４、ＥｃＦＣ、ＭＤＬ、ＡＰＦＰ、ＴＴＦＰ、ＵＮＩＴＹ　２Ｄフィンガープリント、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、試験化学化合物の化学構造からフィンガープリントを生成することを更に含む。

【００４２】

いくつかの実施形態において、細胞構成要素モジュールのセットは、５つ以上の細胞構成要素モジュール、１０個以上の細胞構成要素モジュール、又は１００個以上の細胞構成要素モジュールを含む。

【００４３】

いくつかの実施形態において、それぞれの細胞構成要素モジュールにおける複数の細胞構成要素の独立したサブセットは、５つ以上の細胞構成要素を含む。

【００４４】

いくつかの実施形態において、それぞれの細胞構成要素モジュールにおける複数の細胞構成要素の独立したサブセットは、目的の生理学的状態と関連付けられた分子経路における２～２０個の細胞構成要素からなる。

【００４５】

いくつかの実施形態において、第１の閾値基準は、第１の細胞構成要素モジュールが閾値活性化スコアを有することが必要である。

【００４６】

本開示の別の態様は、試験化学化合物を目的の生理学的状態と関連付ける方法を提供する。

【００４７】

この方法は、（Ａ）試験化学化合物の化学構造のフィンガープリントを得ることを含む。

【００４８】

この方法は、（Ｂ）摂動シグネチャのセットにアクセスすることであって、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、複数の細胞構成要素のそれぞれの独立したサブセットを含み、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、それぞれの第１の細胞状態及び第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である、アクセスすること、を更に含む。

【００４９】

この方法は、（Ｃ）フィンガープリントをモデルに入力することであって、モデルが、５０、１００、５００、１０００、又は１０，０００以上のパラメータを含み、モデルが、フィンガープリントのモデルへの入力に応答して１つ以上の計算された活性化スコアを出力し、１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々が、摂動シグネチャのセットにおける対応する摂動シグネチャを表す、入力することを更に含む。

【００５０】

この方法は、（Ｄ）摂動シグネチャのセットにおける第１の摂動シグネチャについてのそれぞれの計算された活性化スコアが、第１の閾値基準を満たす場合、化学化合物を目的の生理学的状態に関連付けることを更に含む。

【００５１】

いくつかの実施形態において、この方法は、試験化学化合物の単純化された分子入力ラインエントリーシステム（ＳＭＩＬＥＳ）文字列表現からフィンガープリントを計算することを更に含む。

【００５２】
いくつかの実施形態において、モデルは、ニューラルネットワークを含む。いくつかのそのような実施形態において、ニューラルネットワークは、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである。
【００５３】
いくつかの実施形態において、モデルは、複数のコンポーネントモデルを含むアンサンブルモデルであり、複数のコンポーネントモデルにおけるコンポーネントモデルの各々が、化学構造のフィンガープリントを複数のコンポーネントモデルのセットにおけるコンポーネントモデルの各々に入力することに応答して、摂動シグネチャのセットにおける異なる摂動シグネチャについての活性化スコアを提供する。
【００５４】
いくつかの実施形態において、複数のコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。
【００５５】
いくつかの実施形態において、複数のコンポーネントモデルにおけるコンポーネントモデルの各々は、対応するニューラルネットワークである（例えば、対応するニューラルネットワークは、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである）。
【００５６】
いくつかの実施形態において、複数のコンポーネントモデルにおけるコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルである。
【００５７】
いくつかの実施形態において、対応するニューラルネットワークは、完全に接続されたニューラルネットワーク及びメッセージパッシングニューラルネットワークの組み合わせであり、第１のニューラルネットワークの第１の出力及び第２のニューラルネットワークの第２の出力が、化学構造のフィンガープリントを完全に接続されたニューラルネットワーク及びメッセージパッシングニューラルネットワークに入力することに応答して、組み合わされて、摂動シグネチャのセットにおける第１の摂動シグネチャについての１つ以上の計算された活性化スコアにおける活性化スコアを決定する。
【００５８】
いくつかの実施形態において、摂動シグネチャのセットは、複数の摂動シグネチャであり、第１の摂動シグネチャを含む、複数の摂動シグネチャの第１のサブセットが、目的の生理学的状態と関連付けられ、複数の摂動シグネチャの第２のサブセットが、目的の生理学的状態と関連付けられておらず、第１の摂動シグネチャについてのそれぞれの計算された活性化スコアが、第１の閾値基準を満たし、複数の摂動シグネチャの第２のサブセットにおける摂動シグネチャについてのそれぞれの計算された活性化スコアが、第１の閾値基準以外の第２の閾値基準を満たす場合、試験化学化合物が、目的の生理学的状態と識別される。
【００５９】
いくつかの実施形態において、目的の生理学的状態は、疾患である。
【００６０】
いくつかの実施形態において、試験化学化合物は、２０００ダルトン未満の分子量を有する有機化合物である。
【００６１】

いくつかの実施形態において、試験化学化合物は、５つの基準のリピンスキーの法則の各々を満たす有機化合物である。いくつかのそのような実施形態において、試験化学化合物は、５つの基準のリピンスキーの法則のうちの少なくとも３つの基準を満たす有機化合物である。

【００６２】

いくつかの実施形態において、モデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。

【００６３】

いくつかの実施形態において、この方法は、Ｄａｙｌｉｇｈｔ、ＢＣＩ、ＥＣＦＰ４、ＥｃＦＣ、ＭＤＬ、ＡＰＦＰ、ＴＴＦＰ、ＵＮＩＴＹ　２Ｄフィンガープリント、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、使用して、試験化学化合物の化学構造からフィンガープリントを生成することを更に含む。

【００６４】

いくつかの実施形態において、摂動シグネチャのセットは、第１の摂動シグネチャからなる。

【００６５】

いくつかの実施形態において、摂動シグネチャのセットは、５つ以上の摂動シグネチャ、１０個以上の摂動シグネチャ、又は１００個以上の摂動シグネチャを含む。

【００６６】

いくつかの実施形態において、第１の閾値基準は、第１の摂動シグネチャが閾値活性化スコアを有することが必要である。

【００６７】

本開示の別の態様は、化学化合物を目的の生理学的状態と関連付ける方法を提供する。

【００６８】

この方法は、メモリ及び１つ以上のプロセッサを含むコンピュータシステムにおいて、（Ａ）複数の化合物におけるそれぞれの化合物の各々の対応する化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることを含む。

【００６９】

この方法は、（Ｂ）複数の化合物における化合物の各々についての細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの数値的活性化スコアを電子形式で得ることであって、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素の独立したサブセットを含む、得ること、を更に含む。

【００７０】

この方法は、（Ｃ）複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々について、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの細胞構成要素モジュールについてのそれぞれの計算された活性化スコアと、（ｉｉ）細胞構成要素モジュールのセットにおけるそれぞれの化合物についてのそれぞれの細胞構成要素モジュールのそれぞれの数値的活性化スコアとの間のそれぞれの差を使用して訓練されていないモデルを訓練することであって、訓練すること（Ｃ）が、差に応答して訓練されていないモデルと関連付けられた複数のパラメータを調整し、複数のパラメータが、５０、１００、２００、５００、１０００、又は１０，０００以上のパラメータを含み、それによって、化学化合物を目的の生理学的状態と関連付ける訓練されたモデルを得る、訓練すること、を更に含む。

【００７１】

いくつかの実施形態において、細胞構成要素モジュールのセットは、単一の細胞構成要

素モジュールからなる。
【００７２】
いくつかの実施形態において、細胞構成要素モジュールのセットは、複数の細胞構成要素モジュールを含む。
【００７３】
いくつかの実施形態において、細胞構成要素モジュールのセットは、２００～５００個の細胞構成要素モジュールからなる。
【００７４】
いくつかの実施形態において、複数の化合物は、１０～１×１０^{6}個の化合物からなる。
【００７５】
いくつかの実施形態において、複数の化合物は、１００～１００，０００個の化合物からなる。
【００７６】
いくつかの実施形態において、複数の化合物は、１０００～１００，０００個の化合物からなる。
【００７７】
いくつかの実施形態において、訓練すること（Ｃ）は、回帰アルゴリズムに従って、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々についてのそれぞれの化合物の各々と関連付けられた差の各々に応答して、訓練されていないモデルと関連付けられた複数のパラメータを調整する。いくつかのそのような実施形態において、回帰アルゴリズムは、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々についてのそれぞれの化合物の各々と関連付けられた差の各々の最小二乗誤差を最適化する。
【００７８】
いくつかの実施形態において、訓練されたモデルは、ニューラルネットワーク（例えば、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせ）を含む。
【００７９】
いくつかの実施形態において、訓練されたモデルは、複数のコンポーネントモデルのアンサンブルモデルであり、複数のコンポーネントモデルにおけるそれぞれのコンポーネントモデルの各々が、複数の細胞構成要素モジュールにおける異なる細胞構成要素モジュールについて計算された活性化スコアを出力する。いくつかのそのような実施形態において、複数のコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。
【００８０】
いくつかの実施形態において、複数のコンポーネントモデルにおけるコンポーネントモデルの各々は、対応するニューラルネットワークである。いくつかのそのような実施形態において、対応するニューラルネットワークは、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせである。
【００８１】
いくつかの実施形態において、細胞構成要素モジュールのセットは、複数の細胞構成要素モジュールであり、複数の細胞構成要素モジュールの第１のサブセットは、目的の生理学的状態と関連付けられ、複数の細胞構成要素モジュールの第２のサブセットは、目的の生理学的状態と関連付けられていない。
【００８２】
いくつかの実施形態において、この方法は、電子形式で１つ以上の第１のデータセットを得、１つ以上の第１のデータセットが、第１の複数の細胞におけるそれぞれの細胞の各

々について、第１の複数の細胞が、２０個以上の細胞を含み、複数の注釈付きの細胞状態を集合的に表し、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、複数の細胞構成要素が、５、１０、１５、２０、２５、５０、又は１００個以上の細胞構成要素を含み、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含み、それによって、複数のベクトルにアクセスするか、又はそれらを形成すること、を含む、プロセスによって複数の細胞構成要素モジュールにおける細胞構成要素モジュールを識別することを更に含む。複数のベクトルにおけるそれぞれのベクトルの各々は、（ｉ）複数の構成要素におけるそれぞれの細胞構成要素に対応し、（ｉｉ）対応する複数のエレメントを含む。対応する複数のエレメントにおけるそれぞれのエレメントの各々は、第１の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を表す対応するカウントを有する。複数のベクトルは、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別するために使用され、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々は、複数の細胞構成要素のサブセットを含む。複数の細胞構成要素モジュールは、（ｉ）複数の候補細胞構成要素モジュール、及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された潜在表現で配置され、複数の細胞構成要素モジュールは、３、５、１０、１５、２０、又は１００を超える細胞構成要素モジュールを含む。１つ以上の第２のデータセットは、電子形式で得られ、１つ以上の第２のデータセットは、第２の複数の細胞におけるそれぞれの細胞の各々について、第２の複数の細胞が、２０個以上の細胞を含み、目的の生理学的状態を通知する複数の共変量を集合的に表し、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含み、それによって、（ｉ）第２の複数の細胞、及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された細胞構成要素カウントデータ構造を得る。活性化データ構造は、複数の細胞構成要素又はその表現を共通次元として使用して、細胞構成要素カウントデータ構造及び潜在表現を組み合わせることによって形成され、活性化データ構造は、複数の細胞構成要素モジュールにおける細胞構成要素モジュールの各々について、第２の複数の細胞における細胞の各々について、それぞれの活性化重みを含む。候補細胞構成要素モデルは、（ｉ）活性化データ構造を候補モデルに入力したときに、活性化データ構造内に表される細胞構成要素モジュールの各々における複数の共変量における各共変量の不在又は存在の予測と、（ｉｉ）細胞構成要素モジュールの各々における各共変量の実際の不在又は存在との間の差を使用して訓練される。この訓練は、差に応答して、候補細胞構成要素モデルと関連付けられた複数の共変量パラメータを調整する。

【００８３】

いくつかの実施形態において、複数の共変量パラメータは、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの共変量の各々について、それぞれの共変量が、第２の複数の細胞にわたって、それぞれの細胞構成要素モジュールと相関するかどうかを示す対応するパラメータを含み、候補細胞構成要素モデルを訓練する際に複数の共変量パラメータを使用して、複数の候補細胞構成要素モジュールにおける細胞構成要素モジュールを識別する。

【００８４】

いくつかの実施形態において、複数の注釈付きの細胞状態における注釈付きの細胞状態は、曝露条件下での化合物への第１の複数の細胞における細胞の曝露である。

【００８５】

いくつかの実施形態において、曝露条件は、曝露期間、化合物の濃度、又は曝露期間及び化合物の濃度の組み合わせである。

【００８６】

いくつかの実施形態において、複数の細胞構成要素における細胞構成要素の各々は、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせである。

【００８７】
いくつかの実施形態において、第１又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、比色測定、蛍光測定、発光測定、又は共鳴エネルギー移動（ＦＲＥＴ）測定によって決定される。
【００８８】
いくつかの実施形態において、第１又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、単一細胞リボ核酸（ＲＮＡ）配列決定（ｓｃＲＮＡ－ｓｅｑ）、ｓｃＴａｇ－ｓｅｑ、配列決定を使用したトランスポザーゼ－アクセス可能なクロマチンのための単一細胞アッセイ（ｓｃＡＴＡＣ－ｓｅｑ）、ＣｙＴＯＦ／ＳＣｏＰ、Ｅ－ＭＳ／Ａｂｓｅｑ、ｍｉＲＮＡ－ｓｅｑ、ＣＩＴＥ－ｓｅｑ、又はそれらの任意の組み合わせによって決定される。
【００８９】
いくつかの実施形態において、複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することは、複数のベクトルにおけるベクトルの各々の対応する複数のエレメントの各々を使用して、複数のベクトルに相関モデルを適用することを含む。いくつかのそのような実施形態において、相関モデルは、グラフクラスタリング（例えば、ピアソン相関ベースの距離メトリック上のライデン（Ｌｅｉｄｅｎ）クラスタリング、又はルーバン（Ｌｏｕｖａｉｎ）クラスタリングである）を含む。
【００９０】
いくつかの実施形態において、複数の細胞構成要素は、１００～８，０００個の細胞構成要素からなる。
【００９１】
いくつかの実施形態において、複数の構成要素モジュールにおける候補細胞構成要素モジュールの各々は、２００～３００個の細胞構成要素からなる。
【００９２】
いくつかの実施形態において、目的の生理学的状態は、疾患である。
【００９３】
いくつかの実施形態において、生理学的状態は、疾患であり、第１の複数の細胞が、複数の注釈付きの細胞状態によって示されるように、疾患を代表する細胞、及び疾患を代表しない細胞を含む。
【００９４】
いくつかの実施形態において、複数の共変量は、細胞バッチ、細胞ドナー、細胞型、疾患状態、又は化学化合物への曝露を含む。
【００９５】
いくつかの実施形態において、候補細胞構成要素モデルを訓練することは、マルチタスク策定におけるカテゴリ交差エントロピー損失を使用して実施され、複数の共変量における共変量の各々が、複数のコスト関数におけるコスト関数に対応し、複数のコスト関数におけるそれぞれのコスト関数の各々が、共通の重み付け係数を有する。
【００９６】
いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、２０００ダルトン未満の分子量を有する有機化合物である。
【００９７】
いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、５つの基準のリピンスキーの法則の各々を満たす。いくつかのそのような実施形態において、複数の化学化合物における化学化合物の各々は、５つの基準のリピンスキーの法則のうちの少なくとも３つの基準を満たす。
【００９８】
いくつかの実施形態において、訓練されたモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最

近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。

【００９９】

いくつかの実施形態において、方法は、Ｄａｙｌｉｇｈｔ、ＢＣＩ、ＥＣＦＰ４、ＥＣＦＣ、ＭＤＬ、ＡＰＦＰ、ＴＴＦＰ、ＵＮＩＴＹ　２Ｄフィンガープリント、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、対応する化学構造からそれぞれのフィンガープリントの各々を生成することを更に含む。

【０１００】

いくつかの実施形態において、細胞構成要素モジュールのセットは、５つ以上の細胞構成要素モジュール、１０個以上の細胞構成要素モジュール、又は１００個以上の細胞構成要素モジュールを含む。

【０１０１】

本開示の別の態様は、化学化合物を目的の生理学的状態と関連付ける方法を提供する。この方法は、例えば、メモリ及び１つ以上のプロセッサを含むコンピュータシステムにおいて実施することができる。

【０１０２】

この方法は、（Ａ）複数の化合物におけるそれぞれの化合物の各々の対応する化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることを含む。

【０１０３】

この方法は、（Ｂ）複数の化合物における対応する化合物の各々についての摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々のそれぞれの数値的活性化スコアを電子形式で得ることであって、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含む、得ることを更に含む。それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、それぞれの第１の細胞状態及び第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である。

【０１０４】

この方法は、（Ｃ）複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々について、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの摂動シグネチャについてのそれぞれの計算された活性化スコアと、（ｉｉ）摂動シグネチャのセットにおける対応する化合物についてのそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアとの間のそれぞれの差を使用して訓練されていないモデルを訓練することを更に含む。訓練（Ｃ）は、差に応答して、訓練されていないモデルと関連付けられた複数のパラメータを調整し、それによって、化学化合物を目的の生理学的状態と関連付ける訓練されたモデルを得る。いくつかの実施形態において、複数のパラメータは、５０、１００、２００、５００、１０００、１０，０００、又は1×10^{6}以上のパラメータを含む。

【０１０５】

いくつかの実施形態において、摂動シグネチャのセットは、単一の摂動シグネチャからなる。

【０１０６】

いくつかの実施形態において、摂動シグネチャのセットは、２００～５００個の摂動シグネチャからなる。

【０１０７】

いくつかの実施形態において、複数の化合物は、１０～1×10^{6}個の化合物からなる

。いくつかの実施形態において、複数の化合物は、１００～１００，０００個の化合物からなる。いくつかの実施形態において、複数の化合物は、１０００～１００，０００個の化合物からなる。

【０１０８】

いくつかの実施形態において、訓練すること（Ｃ）は、回帰アルゴリズムに従って、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々についての対応する化合物の各々と関連付けられた差の各々に応答して、訓練されていないモデルと関連付けられた複数のパラメータを調整する。いくつかのそのような実施形態において、回帰アルゴリズムは、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々についての対応する化合物の各々と関連付けられた差の各々の最小二乗誤差を最適化する。

【０１０９】

いくつかの実施形態において、訓練されたモデルは、ニューラルネットワーク（例えば、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせ）を含む。

【０１１０】

いくつかの実施形態において、訓練されたモデルは、複数のコンポーネントモデルのアンサンブルモデルであり、複数のコンポーネントモデルにおけるそれぞれのコンポーネントモデルの各々が、それぞれの化学構造のフィンガープリントを複数のコンポーネントモデルのセットにおけるコンポーネントモデルの各々に入力することに応答して、複数の摂動シグネチャのセットにおける異なる摂動シグネチャのセットについて計算された活性化スコアを出力する。いくつかのそのような実施形態において、複数のコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。

【０１１１】

いくつかの実施形態において、複数のコンポーネントモデルにおけるコンポーネントモデルの各々は、対応するニューラルネットワーク（例えば、完全に接続されたニューラルネットワーク、メッセージパッシングニューラルネットワーク、又はそれらの組み合わせ）である。

【０１１２】

いくつかの実施形態において、摂動シグネチャのセットは、複数の摂動シグネチャを含み、複数の摂動シグネチャの第１のサブセットは、目的の生理学的状態と関連付けられ、複数の摂動シグネチャの第２のサブセットは、目的の生理学的状態と関連付けられていない。

【０１１３】

いくつかの実施形態において、目的の生理学的状態は、疾患である。

【０１１４】

いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、２０００ダルトン未満の分子量を有する有機化合物である。

【０１１５】

いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、５つの基準のリピンスキーの法則の各々を満たす。

【０１１６】

いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、５つの基準のリピンスキーの法則のうちの少なくとも３つの基準を満たす。

【０１１７】

いくつかの実施形態において、訓練されたモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシン、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジステ

ィック回帰モデル、線形モデル、又は線形回帰モデルを含む。

【０１１８】

いくつかの実施形態において、方法は、Ｄａｙｌｉｇｈｔ、ＢＣＩ、ＥＣＦＰ４、ＥｃＦＣ、ＭＤＬ、ＡＰＦＰ、ＴＴＦＰ、ＵＮＩＴＹ　２Ｄフィンガープリント、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、対応する化学構造からそれぞれのフィンガープリントの各々を生成することを更に含む。

【０１１９】

いくつかの実施形態において、摂動シグネチャのセットは、５つ以上の摂動シグネチャ、１０個以上の摂動シグネチャ、又は１００個以上の摂動シグネチャを含む。

【０１２０】

いくつかの実施形態において、方法は、変化していない細胞状態と変化した細胞状態との間の差次的細胞構成要素存在量の尺度を表す単一細胞遷移シグネチャに電子形式でアクセスすることであって、変化した細胞状態が、変化していない細胞状態から変化した細胞状態への細胞遷移を通して発生し、（ｉ）変化していない細胞状態、（ｉｉ）変化した細胞状態、及び（ｉｉｉ）変化していない細胞状態から変化した細胞状態への遷移のうちの少なくとも１つが、目的の生理学的状態と関連付けられ、単一細胞遷移シグネチャが、参照の複数の細胞構成要素の識別と、複数の参照細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、変化していない細胞状態と変化した細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する第１の有意性スコアとを含む、アクセスすること、を含む、手順によって摂動シグネチャのセットにおけるそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアを得ることを更に含む。更に、単一細胞遷移シグネチャ及びそれぞれの摂動シグネチャを比較し、それによってそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアを決定する。

【０１２１】

いくつかの実施形態において、単一細胞遷移シグネチャと摂動シグネチャとを比較して、それぞれの摂動シグネチャのそれぞれの数値的活性化スコアを決定することは、単一細胞遷移シグネチャの参照の複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの摂動シグネチャにおける対応する細胞構成要素の対応する有意性スコアに対するそれぞれの細胞構成要素の第１の有意性スコアを比較することを含む。

【０１２２】

いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、摂動シグネチャのセットにおける他の摂動シグネチャと比較して、単一細胞遷移シグネチャに対するそれぞれの摂動シグネチャの関連性の相対的なランキングである。

【０１２３】

いくつかの実施形態において、相対的なランキングは、ウィルコクソンの順位和検定、ｔ検定、ロジスティック回帰、又は一般化線形モデルによって決定される。

【０１２４】

いくつかの実施形態において、単一細胞遷移シグネチャの変化していない細胞状態が、それぞれの摂動シグネチャの第１の細胞状態又は第２の細胞状態と同じである。

【０１２５】

いくつかの実施形態において、単一細胞遷移シグネチャの変化していない細胞状態が、それぞれの摂動シグネチャの第１の細胞状態及び第２の細胞状態の両方とは異なる。

【０１２６】

いくつかの実施形態において、方法は、単一細胞遷移シグネチャの参照の複数の細胞構成要素、及びそれぞれの摂動シグネチャのそれぞれの複数の細胞構成要素を剪定して、転写因子と比較することを制限することを更に含む。

【０１２７】

いくつかの実施形態において、複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの摂動細胞状態は、複数の化合物における化合物に曝露されていない対照細胞によって表される。

【０１２８】
いくつかの実施形態において、複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの摂動細胞状態は、それぞれの摂動シグネチャと関連付けられた化合物以外の複数の化学化合物における化学化合物に曝露されている無関係の摂動細胞にわたる平均によって表される。
【０１２９】
開示された実施形態のいくつかにおいて、モデルは、リグレッサーである。
【０１３０】
本開示の別の態様は、１つ以上のプロセッサ、及び１つ以上のプロセッサによる実行のための１つ以上のプログラムを格納するメモリを有するコンピュータシステムを提供し、１つ以上のプログラムは、本明細書に開示される方法及び／又は実施形態のうちのいずれかを実施するための命令を含む。
【０１３１】
本開示の別の態様は、コンピュータによって実行するように構成された１つ以上のプログラムを格納する非一時的なコンピュータ可読記憶媒体を提供し、１つ以上のプログラムは、本明細書に開示される方法及び／又は実施形態のうちのいずれかを実行するための命令を含む。
【０１３２】
本開示の更なる態様及び利点は、以下の詳細な説明から当業者に容易に明らかになるであろう。それにおいては、本開示の例示的な実施形態のみが示され、説明される。理解されるように、本開示は、他の異なる実施形態が可能であり、そのいくつかの詳細は、全て本開示から逸脱することなく、様々な明白な点で修正が可能である。したがって、図面及び説明は、本質的に例示とみなされるべきであり、限定とみなされるべきではない。
【０１３３】
本明細書に開示される実施形態は、添付の図面の図において、限定としてではなく例として示される。同様の参照番号は、図面全体を通して対応する部分を指す。
【図面の簡単な説明】
【０１３４】
【図１】本開示の一実施形態による、例示的なシステム及びコンピューティングデバイスのブロック図を示す。
【図２Ａ】本開示の様々な実施形態による、複数の細胞構成要素を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートをまとめて提供する。
【図２Ｂ】本開示の様々な実施形態による、複数の細胞構成要素を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートをまとめて提供する。
【図３Ａ】本開示の様々な実施形態による、試験化合物を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートを提供し、破線のボックスは任意選択の要素を表す。
【図３Ｂ】本開示の様々な実施形態による、試験化合物を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートを提供し、破線のボックスは任意選択の要素を表す。
【図３Ｃ】本開示の様々な実施形態による、試験化合物を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートを提供し、破線のボックスは任意選択の要素を表す。
【図３Ｄ】本開示の様々な実施形態による、試験化合物を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートを提供し、破線のボックスは任意選択の要素を表す。
【図３Ｅ】本開示の様々な実施形態による、試験化合物を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートを提供し、破線のボックスは

任意選択の要素を表す。
【図４】本開示のいくつかの実施形態による、細胞構成要素の複数のベクトルの例及び細胞構成要素モジュールの潜在表現の例を示す。
【図５】本開示のいくつかの実施形態による、細胞構成要素カウントデータ構造及び例示的な活性化データ構造の例を示す。
【図６】本開示のいくつかの実施形態による、複数の化合物の重みを調整するためにモデルを訓練する方法の例を示す。
【図７】本開示のいくつかの実施形態による、試験化学化合物を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートを提供し、破線のボックスは任意選択の要素を表す。
【図８】本開示の一実施形態による、化学化合物を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートを提供し、破線のボックスは任意選択の要素を表す。
【図９】本開示の一実施形態による、化学化合物を目的の生理学的状態と関連付けるための例示的な方法のプロセス及び特徴のフローチャートを提供し、破線のボックスは任意選択の要素を表す。
【図１０Ａ】本開示の一実施形態による、脂肪酸関連細胞プログラムの活性化のための化学構造を予測するための例示的な方法の性能及び４倍の検証を示す。図１０Ａは、化学構造を予測するためのモデルアーキテクチャの概略図を示す。
【図１０Ｂ】本開示の一実施形態による、脂肪酸関連細胞プログラムの活性化のための化学構造を予測するための例示的な方法の性能及び４倍の検証を示す。図１０Ｂは、１，２００個のランダムに選択された化合物の試験セットにおける性能を示す。
【図１０Ｃ】本開示の一実施形態による、脂肪酸関連細胞プログラムの活性化のための化学構造を予測するための例示的な方法の性能及び４倍の検証を示す。図１０Ｃは、訓練セットとは異なる足場を有する１，２００個の化合物の試験セットにおける性能を示す。
【図１０Ｄ】本開示の一実施形態による、脂肪酸関連細胞プログラムの活性化のための化学構造を予測するための例示的な方法の性能及び４倍の検証を示す。図１０Ｄは、インビトロ前脂肪細胞アッセイにおける転写活性化に基づくベージング（ｂｅｉｇｅｉｎｇ）関連モジュールの検証を示す。
【図１０Ｅ】本開示の一実施形態による、脂肪酸関連細胞プログラムの活性化のための化学構造を予測するための例示的な方法の性能及び４倍の検証を示す。図１０Ｅは、標的モジュールに対する５００万個の化合物のデータベースから引き出された予測される化合物の最適化を示す。
【図１１】本開示の一実施形態による、胎児の赤血球生成及びＴ細胞枯渇に関連する細胞挙動の活性化のための化学構造を予測するための例示的な方法の検証を示す。
【図１２】本開示の一実施形態による、単一細胞ＲＮＡ配列決定（ｓｃＲＮＡ－ｓｅｑ）を使用したヒト前脂肪細胞遺伝子モジュール活性化に対する既知のピペリジン含有化合物（「ＫＰＣＣ」）及び６つの新たに合成されたヒット「合成ヒット」）の影響を評価するための例示的な方法の概略図を示す。
【図１３】本開示の一実施形態による、所望の転写変化の活性化に対するＫＰＣＣ及び６つの合成ヒットの効果を示す。
【図１４Ａ】任意選択の要素が破線のボックスによって示される細胞構成要素モジュールを識別するフローチャートを提供する。
【図１４Ｂ】任意選択の要素が破線のボックスによって示される細胞構成要素モジュールを識別するフローチャートを提供する。
【図１４Ｃ】任意選択の要素が破線のボックスによって示される細胞構成要素モジュールを識別するフローチャートを提供する。
【図１４Ｄ】任意選択の要素が破線のボックスによって示される細胞構成要素モジュールを識別するフローチャートを提供する。
【発明を実施するための形態】

【０１３５】
導入。
上記の背景を考慮すると、本開示は、疾患に重要な細胞プロセス及びプログラムを標的とする創薬へのアプローチを記載する。このアプローチは、いくつかの態様において、生理学的状態（例えば、細胞プログラム、細胞プロセス、及び／又は細胞状態）及び化合物の化学構造のコンピュータにより操作された表現を使用して、化学構造関連モダリティ及びそれらの特性を予測することによって実現される。次いで、符号化された化学構造を細胞プログラム及び／又は細胞状態の表現にマッピングし、それによって、化合物を生理学的状態に関連付けることができる。
【０１３６】
例えば、いくつかの態様において、本開示は、分子プロファイル（例えば、遺伝子モジュール）と、目的の生物学的プロセス（例えば、細胞プログラム及び／又は細胞状態）及び化合物の化学構造との間の関連性を得るためのシステム及び方法を提供する。これらの関連性を使用して、創薬のために、類似の機能的又は構造的特性を有するものなどの新しい化学構造を予測することができる。
【０１３７】
いくつかの実施形態において、予測能力を有する計算モデリングアーキテクチャは、１つ以上のドメイン及び／又はデータタイプにわたる生理学的に関連する化学構造の潜在表現の生成を通じて、これらの関連性を発見するために使用される。関連性は、例えば、細胞の１つ以上の化合物への曝露に応答して、差次的遺伝子発現又は細胞状態遷移などの細胞挙動のプロファイルを提供する摂動データに由来し得る。いくつかの実施態様において、方法は、潜在表現及び機械学習を使用して、様々なドメイン（例えば、分子、細胞、臨床、インビボ、インビトロ、知識ベースなど）及び／又は様々なデータタイプ（転写、遺伝的、エピジェネティック、共変量など）の間の相関を組み合わせて決定して、生理学的に関連する化学構造を予測する。
【０１３８】
例示的な実施形態において、本開示は、化合物についての潜在表現を使用するモデリングアプローチを提供する。複数の化合物におけるそれぞれの化合物の各々について、方法は、それぞれの化合物が、複数の生理学的状態における生理学的状態の各々を誘発する可能性を表すベクトルを格納する潜在表現を生成することを含む。生理学的状態は、特定の表現型、細胞プロセス、及び／又は疾患と関連付けられた細胞状態遷移及び／又は細胞構成要素モジュール（例えば、遺伝子モジュール）を含むことができる。したがって、方法は、例えば、ｎ＿化合物ｘ　ｎ＿細胞＿状態又はｎ＿化合物ｘ　ｎ＿遺伝子＿モジュールとして示される、化合物及び生理学的状態（例えば、細胞状態及び／又は遺伝子モジュール）によって次元決定されたモデルについてのマルチタスク訓練標識として機能するマトリックス表現を生成する。
【０１３９】
化合物を生理学的状態と関連付けるための機械学習モデルについての入力は、化合物の化学構造を符号化し、更にモデルを訓練するために使用される、各化合物の正準異性体ＳＭＩＬＥＳ表現及び／又はグラフベースの表現を含む。訓練標識は、各化合物を各生理学的状態と関連付ける数値的活性化スコアとして提供される。例えば、各化合物についてのベクトルは、複数の関連する重みを含むことができ、各重みは、化合物が、それぞれの細胞状態、細胞状態遷移、摂動シグネチャ、及び／又はそれぞれの遺伝子モジュールの活性化などのそれぞれの生理学的状態を誘導する可能性を示す。
【０１４０】
入力としてマトリックス表現を受信すると、モデルは、回帰問題を解決することによって化学構造から細胞状態（例えば、摂動シグネチャ）及び／又は遺伝子モジュール活性化を学習するように訓練される。２つの例示的なモデルアーキテクチャは、回帰問題を解決するために使用される。第１のモデルは、ＳＭＩＬＥＳ文字列の標準的なフィンガープリント上で完全に接続されたネットワークを利用し、ネットワークアーキテクチャは、Ｒｅ

ＬＵ活性化を伴う３層ネットワークである。第２のモデルは、ＤＧＬライブラリからのＭＰＮＮネットワークを含む。これらのモデルの各々は、回帰予測の最小二乗誤差を最適化することによって、互いに独立して訓練される。試験時間に、これらのモデルの予測は平均化され、したがって、第１及び第２のモデルを含むアンサンブルモデルを形成する。次いで、アンサンブルモデルを使用して、化合物と生理学的状態との間の関連性を決定することができ、これを更に適用して、化学構造から生理学的活性化の可能性の予測及び／又は特定の生理学的状態を誘発する可能性のある化学構造の予測を得ることができる。

【０１４１】

有利には、本明細書に開示されるシステム及び方法は、創薬のための体系的でスケーラブルなアプローチを提供することによって、上記の欠点に対処する。例えば、創薬に関連する従来の機械学習アプローチは、ディープラーニング方法及び高性能コンピューティングとペアリングされた３Ｄタンパク質及び化学構造表現を使用したインシリコ標的スクリーニング能力を利用して、標的のライブラリに対する候補化合物の作用方法を計算する。しかしながら、これらのアプローチは、生物学的プロセスの基礎となる動的かつ高度にネットワーク化された多細胞系の複雑さに適切に対処していない、標的に焦点を当てたスクリーニングパラダイムに該当する。創薬のための他の従来の方法は、トランスクリプトームデータ又はイメージングデータに基づいて、単一の細胞及び細胞株が摂動にどのように応答するかをモデル化するために機械学習アプローチを使用する。そのような方法において、ハイスループットデータセットは、疾患の表現型表現及びインビトロ細胞系の複合摂動を学習するために使用される。これらは、表現型疾患応答を誘発又は相殺する化合物を予測するために使用される。しかしながら、従来のハイスループットデータモデリングアプローチは、それにもかかわらず、キュレーションの欠如及び多数の候補標的の識別の可能性によって不利になっている。ハイスループットスクリーニングから得られる潜在的な候補の各々の検証は、多くの場合、分子標的ベースの最適化又はインビトロスクリーニングのための数百若しくは更に数千又は化合物の合成を必要とする、手間のかかるプロセスである。

【０１４２】

これらのアプローチとは対照的に、本開示は、次いで、生物学的プロセス（例えば、目的の生理学的状態に関与する遺伝子モジュール又は摂動シグネチャ）と関連付けられた細胞状態、摂動シグネチャ及び／又は細胞成分の表現にわたってマッピングされる、表現化学構造データ（例えば、化合物処理に対する細胞応答）を得るためのシステム及び方法を有利に提供する。それにもかかわらず、この標的に依存しないアプローチは、候補標的の体系的なキュレーション及び最適化を可能にし、したがって、標的発見とシステムにわたる予測翻訳との間のかなりのギャップを埋める。

【０１４３】

例えば、以下の実施例に例示されるように、脂肪酸代謝に関与する候補ファーマコフォアは、本明細書に開示されるシステム及び方法の実施形態を使用して特定された。実施例４に更に例示されるように、候補ファーマコフォアに基づく予測翻訳は、６つの新しい化学物質を生成し、それらの全ては、ヒト脂肪細胞で試験した場合、脂肪酸関連細胞プロセスに関与する遺伝子モジュールを活性化することが見出された。タンパク質標的に対するハイスループットスクリーニング、特定若しくは最適化、又は数百若しくは数千の新しい化合物の合成を必要とせずに、候補ファーマコフォアの特定及び６つの新しい化学物質の設計を行った。したがって、本明細書に提供されるシステム及び方法は、標的発見から予測翻訳及び検証まで、従来の分子標的ベース又は表現型ベースのアプローチよりも、創薬及び開発プロセスの容易さ及び効率を改善する。

【０１４４】

有利には、本開示は、化合物と生理学的状態との間の関連性（例えば、重み及び／又は相関）の標的化された決定のためのモデルの訓練及び使用を改善することによって、化合物と生理学的状態との関連性を改善する様々なシステム及び方法を更に提供する。機械学習モデルの複雑さは、時間の複雑性（所与の入力サイズｎに対する実行時間、又はアルゴ

リズムの速度の尺度）、空間の複雑性（空間要件、又は所与の入力サイズｎに対するアルゴリズムを実行するために必要なコンピューティングパワー若しくはメモリの量）、又は両方を含む。複雑性（及びその後の計算負担）は、所与のモデルの訓練及び所与のモデルによる予測の両方に適用される。

【０１４５】

いくつかの例では、計算の複雑性は、実装、追加のアルゴリズム若しくは交差検証方法の組み込み、及び／又は１つ以上のパラメータ（例えば、重み及び／又はハイパーパラメータ）によって影響を受ける。いくつかの例では、計算の複雑性は、入力サイズｎの関数として表され、入力データは、インスタンスの数（例えば、訓練試料の数）、次元ｐ（例えば、特徴の数）、ツリーn_{trees}の数（例えば、ツリーに基づく方法の場合）、サポートベクトルn_{sv}の数（例えば、サポートベクトルに基づく方法の場合）、隣接ｋの数（例えば、ｋ最近傍モデルの場合）、クラスｃの数、及び／又は層ｉにおけるニューロンn_iの数（例えば、ニューラルネットワークの場合）である。入力サイズｎに関して、次いで、（例えば、ビッグＯ表記での）計算の複雑性の近似は、入力サイズが増加するにつれて、実行時間及び／又は空間要件がどのように増加するかを示す。関数は、入力サイズの増加と比較して、より遅い速度又はより速い速度で複雑性を増加させることができる。計算の複雑性の様々な近似には、定数（例えば、Ｏ（１））、対数（例えば、Ｏ（ｌｏｇ　ｎ））、線形（例えば、Ｏ（ｎ））、対数線形（例えば、Ｏ（ｎ　ｌｏｇ　ｎ））、二次（例えば、Ｏ（n^2））、多項式（例えば、Ｏ（n^c））、指数（例えば、Ｏ（c^n））、及び／又は階乗（例えば、Ｏ（ｎ！））が含まれるが、これらに限定されない。いくつかの例では、定数関数の場合のように、入力サイズが増加するにつれて、より単純な関数はより低いレベルの計算の複雑性を伴うが、階乗関数などのより複雑な関数は、入力サイズのわずかな増加に応答して複雑性の大幅な増加を示すことができる。

【０１４６】

機械学習モデルの計算の複雑性は、同様に（例えば、ビッグＯ表記で）関数によって表すことができ、複雑性は、モデルのタイプ、１つ以上の入力若しくは次元のサイズ、使用法（例えば、訓練及び／若しくは予測）、並びに／又は時間若しくは空間の複雑性が評価されているかどうかに応じて変化し得る。例えば、決定木モデルにおける複雑性は、訓練のためのＯ（np^2）及び予測のためのＯ（ｐ）として近似され、一方、線形回帰モデルにおける複雑性は、訓練のためのＯ（p^2n+p^3）及び予測のためのＯ（ｐ）として近似される。ランダムフォレストモデルの場合、訓練の複雑性はＯ（np^2n_{trees}）として近似され、予測の複雑性はＯ（pn_{trees}）として近似される。勾配ブーストモデルの場合、複雑性は、訓練のためのＯ（npn_{trees}）及び予測のためのＯ（pn_{trees}）として近似される。カーネルサポートベクトルマシンの場合、複雑性は、訓練のためのＯ（n^2p+n^3）、及び予測のためのＯ（$n_{sv}p$）として近似される。ナイーブベイズモデルの場合、複雑性は、訓練のためのＯ（ｎｐ）、及び予測のためのＯ（ｐ）として表され、ニューラルネットワークの場合、複雑性は、予測のためのＯ（$pn_1+n_1n_2+...$）として近似される。Ｋ最近傍モデルの複雑性は、時間のためのＯ（ｋｎｐ）、及び空間のためのＯ（ｎｐ）として近似される。ロジスティック回帰モデルの場合、複雑性は、時間のためのＯ（ｎｐ）、及び空間のためＯ（ｐ）として近似される。ロジスティック回帰モデルの場合、複雑性は、時間のためのＯ（ｎｐ）、及び空間のためＯ（ｐ）として近似される。

【０１４７】

上述したように、機械学習モデルについて、計算の複雑性は、スケーラビリティを決定し、したがって、入力、特徴、及び／又はクラスサイズの増加、並びにモデルアーキテクチャのバリエーションのためのモデル（例えば、リグレッサー）の全体的な有効性及び有用性を決定する。大規模なデータセットの文脈において、少なくとも１０、少なくとも１００、少なくとも１０００、又はそれ以上の細胞に対して得られた少なくとも１０、少なくとも１００、少なくとも１０００、又はそれ以上の遺伝子の存在量を含む遺伝子発現データセットの場合と同様に、そのような大規模なデータセット上で実施される関数の計算

の複雑性は、多くの既存のシステムの能力に負担をかける可能性がある。更に、入力特徴の数（例えば、細胞構成要素（例えば、遺伝子）の数及び／又は化合物の数）及び／又はインスタンスの数（例えば、細胞の数、細胞状態注釈、摂動シグネチャ、モジュール、及び／又は共変量）が、技術的進歩とともに増加し、注釈の可用性を増加させ、下流の適用及び可能性を拡大するにつれて、任意の所与の分類モデルの計算の複雑性は、それぞれのシステムの仕様によって提供される時間及び空間容量を迅速に圧倒することができる。

【０１４８】

したがって、化合物を生理学的状態と関連付けるための、最小入力サイズ（例えば、少なくとも１０、少なくとも１００、少なくとも１０００、若しくはそれ以上の化合物；それぞれの細胞構成要素モジュールのための少なくとも１０、少なくとも５０、少なくとも１００、若しくはそれ以上の細胞構成要素；少なくとも５、少なくとも１０、少なくとも１００、若しくはそれ以上の摂動シグネチャ；及び／又は少なくとも５、少なくとも１０、少なくとも１００、若しくはそれ以上の細胞構成要素モジュール）及び／又は対応する最小数のパラメータ（例えば、少なくとも５０、少なくとも１００、若しくは少なくとも１０００のパラメータ及び／又は機械学習モデルに入力される特徴の全てのあらゆる可能なペアリングに対応するパラメータ）を有する機械学習モデルを使用することによって、計算の複雑性は、それが精神的に実施され得ないように比例して増加し、方法は、計算上の問題に対処する。例えば、本開示の一実施形態において、複数の少なくとも１０個の細胞構成要素モジュール及び複数の少なくとも５０個の化合物によって次元決定された活性化スコアマトリックスを得ることは、少なくとも５００のパラメータ（例えば、重み）を得ることを含む。本開示の別の実施形態において、複数の少なくとも１０個の摂動シグネチャにおける摂動シグネチャの各々について、複数の少なくとも５０個の化合物における各化合物についてのそれぞれの活性化重みを得ることは、少なくとも５００の活性化重みを得ることを含む。細胞状態遷移、細胞構成要素、細胞、化合物、共変量、試料、時点、複製、及び／又はバッチの数を含むがこれらに限定されない追加の入力特徴及び／又はインスタンスに同様の最小値を課すことは、同様に、方法の計算の複雑性に影響を与えるであろう。

【０１４９】

機械学習モデルにおける計算の複雑性に関する更なる詳細は、２０１８年４月１６日に公開され、ｔｈｅｋｅｒｎｅｌｔｒｉｐ．ｃｏｍ／ｍａｃｈｉｎｅ／ｌｅａｒｎｉｎｇ／ｃｏｍｐｕｔａｔｉｏｎａｌ－ｃｏｍｐｌｅｘｉｔｙ－ｌｅａｒｎｉｎｇ－ａｌｇｏｒｉｔｈｍｓで、オンラインで入手可能な“Ｃｏｍｐｕｔａｔｉｏｎａｌ ｃｏｍｐｌｅｘｉｔｙ ｏｆ ｍａｃｈｉｎｅ ｌｅａｒｎｉｎｇ ａｌｇｏｒｉｔｈｍｓ”、Ｈａｓｔｉｅ，２００１，Ｔｈｅ Ｅｌｅｍｅｎｔｓ ｏｆ Ｓｔａｔｉｓｔｉｃａｌ Ｌｅａｒｎｉｎｇ，Ｓｐｒｉｎｇｅｒ，Ｎｅｗ Ｙｏｒｋ、及びＡｒｏｒａ ａｎｄ Ｂａｒａｋ，２００９，Ｃｏｍｐｕｔａｔｉｏｎａｌ Ｃｏｍｐｌｅｘｉｔｙ：Ａ Ｍｏｄｅｒｎ Ａｐｐｒｏａｃｈ，Ｃａｍｂｒｉｄｇｅ Ｕｎｉｖｅｒｓｉｔｙ Ｐｒｅｓｓ，Ｎｅｗ Ｙｏｒｋに提供されており、それらの各々は、参照によりその全体が本明細書に組み込まれる。

【０１５０】

ここで、添付の図面にその例が示される実施形態を詳細に参照する。以下の詳細な説明には、本開示の完全な理解を提供するために、多くの特定の詳細が示されている。しかしながら、本開示は、これらの特定の詳細なしで実施され得ることが当業者に明らかであろう。他の場合では、実施形態の態様を不必要に曖昧にしないように、周知の方法、手順、構成要素、回路、及びネットワークは、詳細には説明されていない。

【０１５１】

単一のインスタンスとして本明細書に記載される構成要素、動作、又は構造には、複数のインスタンスが提供され得る。最後に、様々な構成要素、動作、及びデータストアの間の境界は、多少任意であり、特定の動作は、特定の例示的な構成の文脈で例示される。他の形態の機能性が想定され、実装の範囲内に含まれ得る。概して、例示的な構成において

別個の構成要素として提示される構造及び機能性は、組み合わされた構造又は構成要素として実装され得る。同様に、単一の構成要素として提示される構造及び機能性は、別個の構成要素として実装され得る。これら及び他の変形、修正、追加、及び改善は、実装の範囲内にある。

【０１５２】

また、「第１の」、「第２の」などの用語は、様々な要素を説明するために本明細書で使用され得るが、これらの要素はこれらの用語によって制限されるべきではないことも理解されたい。これらの用語は、ある要素を別の要素と区別するためにのみ使用される。例えば、第１のデータセットは、本発明の範囲から逸脱することなく、第２のデータセットと称され得、同様に、第２のデータセットは、第１のデータセットと称され得る。第１のデータセット及び第２のデータセットは両方ともデータセットであるが、同じデータセットではない。

【０１５３】

本明細書で使用される専門用語は、特定の実装を説明することのみを目的とし、特許請求の範囲を限定することを意図するものではない。実装及び添付の特許請求の範囲の説明で使用される場合、単数形の「ａ」、「ａｎ」及び「ｔｈｅ」は、文脈により明らかにそうではないと指示されない限り、複数の形態も同様に含むことが意図される。「及び／又は（ａｎｄ／ｏｒ）」という用語は、本明細書で使用するときに、関連する列挙した品目のうちの１つ以上の任意の及び全ての可能な組み合わせを指し、包含することも理解されるであろう。「含む（ｃｏｍｐｒｉｓｅｓ）」及び／又は「含む（ｃｏｍｐｒｉｓｉｎｇ）」という用語は、本明細書で使用される場合、記載された特徴、整数、ステップ、動作、要素、及び／又は成分の存在を指定するが、１つ以上の他の特徴、整数、ステップ、動作、要素、成分、及び／又はそれらの群の存在又は追加を排除しないことが更に理解されるであろう。

【０１５４】

本明細書で使用される場合、「～する場合（ｉｆ）」という用語は、文脈に応じて、「～するとき（ｗｈｅｎ）」又は「～した後（ｕｐｏｎ）」又は記載された先行する条件が真であることの「判定に応じて」又は「判定に従って」又は「検出に応じて」を意味すると解釈され得る。同様に、文脈に応じて、「（述べられた先行する条件が真であると）判定される場合」又は「（述べられた先行する条件が真である）場合」又は「（述べられた先行する条件が真である）とき」という語句は、述べられた先行する条件が真であることの「判定後」又は「判定に応じて」又は「判定に従って」又は「検出後」又は「検出に応じて」を意味すると解釈され得る。

【０１５５】

更に、参照番号が「ｉ番目」の表示を与えられるとき、参照番号は、一般的な成分、セット、又は実施形態を指す。例えば、「細胞成分ｉ」と称される細胞成分は、複数の細胞成分におけるｉ番目の細胞成分を指す。

【０１５６】

前述の説明は、例示的な実装を具現化する、例示的なシステム、方法、技術、命令シーケンス、及びコンピューティングマシンプログラム製品を含む。説明の目的において、本発明の主題の様々な実装の理解を提供するために、多くの特定の詳細が示されている。しかしながら、本発明の主題の実装は、これらの特定の詳細なしで実践され得ることは、当業者には明らかであろう。一般に、周知の命令インスタンス、プロトコル、構造、及び技術は、詳細に示されていない。

【０１５７】

説明の目的において、前述の説明は、特定の実装を参照して説明されている。しかしながら、以下の例示的な議論は、網羅的であることを意図するものではなく、又は、実装を開示される正確な形態に限定することを意図するものではない。上記の教示を考慮して、多くの修正及び変形が可能である。実装は、原理及びそれらの実際の用途を最もよく説明するために選択及び説明され、それによって、当業者が、企図される特定の使用に適した

実装及び様々な修正を伴う様々な実装を最もよく利用できるようにする。

【０１５８】

明確にするために、本明細書に記載される実装の慣例的特徴の全てが示され、説明されるわけではない。そのような任意の実際の実装の開発において、ユースケース及びビジネスに関連する制約への準拠など、設計者の特定の目標を達成するために多くの実装固有の決定が行われ、これらの特定の目標は、実装によって、及び設計者によって異なることが理解されるだろう。更に、そのような設計努力は複雑で時間がかかり得るが、それでも本開示の利益を得る当業者にとってはエンジニアリングの日常的な作業であることが理解されるであろう。

【０１５９】

本明細書のいくつかの部分は、情報に対する動作のアルゴリズム及び記号的表現の観点から、本発明の実施形態を説明する。これらのアルゴリズムの説明及び表現は、データ処理技術の当業者によって、それらの仕事の実質を当業者に効果的に伝達するために一般的に使用される。これらの動作は、機能的に、計算的に、又は論理的に説明されているが、コンピュータプログラム又は同等の電気回路、マイクロコードなどによって実装されることが理解される。

【０１６０】

本明細書で使用される言語は、可読性及び指示目的のために主に選択されており、本発明の主題を描写又は制限するために選択されていない場合がある。したがって、本発明の範囲は、この詳細な説明によって限定されるのではなく、それに基づく出願に関して生じる任意の特許請求の範囲によって限定されることが意図される。したがって、本発明の実施形態の開示は、本発明の範囲を例示することを意図するが、限定するものではない。

【０１６１】

一般に、特許請求の範囲及び本明細書で使用される用語は、当業者によって理解される平易な意味を有すると解釈されることが意図される。特定の用語は、追加の明確さを提供するために以下に定義される。明白な意味と提供される定義との間に矛盾がある場合、提供される定義が使用される。

【０１６２】

本明細書で直接定義されていない任意の用語は、本発明の技術分野内で理解されているように、それらに一般的に関連付けられた意味を有するものと理解されるべきである。ある特定の用語は、本発明の態様の組成物、デバイス、方法など、及びそれらを作製又は使用する方法を説明する際に、実践者に追加の指針を提供するために本明細書で議論される。同じことが複数の様式で言及され得ることが理解されるだろう。その結果、本明細書で議論される用語のうちの任意の１つ以上に対して、代替の言語及び同義語が使用され得る。用語が本明細書で詳述又は議論されるかどうかは重要ではない。いくつかの同義語又は置換可能な方法、材料などが提供される。１つ又はいくつかの同義語又は均等物の列挙は、それが明示的に述べられていない限り、他の同義語又は均等物の使用を排除しない。用語の例を含む例の使用は例示のみを目的とし、本明細書における本発明の態様の範囲及び意味を限定するものではない。

【０１６３】

定義。

本明細書で使用される場合、「約」又は「およそ」という用語は、当業者によって決定される特定の値に対する許容誤差範囲内であることを意味し、それは、部分的には、その値がどのように測定又は決定されるか、例えば、測定システムの限界に依存する。例えば、いくつかの実施形態において、「約」は、当該技術分野における慣例に従って、１以内又は１を超える標準偏差を意味する。いくつかの実施形態において、「約」は、所与の値の±２０％、±１０％、±５％、又は±１％の範囲を意味する。いくつかの実施形態において、「約」又は「およそ」という用語は、値の１桁以内、５倍以内、又は２倍以内であることを意味する。本出願及び特許請求の範囲において特定の値が記載される場合、別段の記載がない限り、特定の値について許容可能な誤差範囲内であることを意味する「約」

という用語が、想定され得る。本明細書の詳細な説明内の全ての数値は、「約」示される値によって修正され、当業者によって予想される実験誤差及び変動を考慮する。「約」という用語は、当業者によって一般的に理解される意味を有することができる。いくつかの実施形態において、「約」という用語は、±１０％を指す。いくつかの実施形態において、「約」という用語は、±５％を指す。

【０１６４】

本明細書で使用される場合、「存在量」、「存在量レベル」、又は「発現レベル」という用語は、１つ以上の細胞に存在する細胞構成要素（例えば、ＲＮＡ種、例えば、ｍＲＮＡ若しくはｍｉＲＮＡ、又はタンパク質分子などの遺伝子産物）の量、又は複数の細胞にわたって存在する細胞構成要素の平均量を指す。ｍＲＮＡ又はタンパク質発現を指す場合、この用語は、一般に、特定の遺伝子座、例えば、特定の遺伝子に対応する任意のＲＮＡ又はタンパク質種の量を指す。しかしながら、いくつかの実施形態において、存在量は、複数のｍＲＮＡ又はタンパク質アイソフォームを生じる特定の遺伝子に対応するｍＲＮＡ又はタンパク質の特定のアイソフォームの量を指すことができる。遺伝子座は、遺伝子名、染色体位置、又は任意の他の遺伝子マッピングメトリックを使用して識別することができる。

【０１６５】

本明細書で同義的に使用される場合、「細胞状態」又は「生物学的状態」は、細胞又は細胞集団の状態又は表現型を指す。例えば、細胞状態は、健康であってもよいか、又は疾患状態であってもよい。細胞状態は、複数の疾患のうちの１つであってもよい。細胞状態は、化合物治療及び／又は分化細胞系列に対する応答であってもよい。細胞状態は、１つ以上の遺伝子、１つ以上のタンパク質、及び／又は１つ以上の生物学的経路を含むが、これらに限定されない、１つ以上の細胞構成要素の尺度（例えば、活性化、発現、及び／又は存在量の尺度）によって特徴付けられ得る。

【０１６６】

本明細書で使用される場合、「細胞状態遷移」又は「細胞遷移」は、第１の細胞状態から第２の細胞状態への細胞の状態の遷移を指す。いくつかの実施形態において、第２の細胞状態は、変化した細胞状態（例えば、罹患した細胞状態への健康な細胞状態）である。いくつかの実施形態において、それぞれの第１の細胞状態及び第２の細胞状態のうちの一方は、非摂動状態であり、それぞれの第１の細胞状態及び第２の細胞状態のうちの他方は、状態への細胞の曝露によって引き起こされる摂動状態である。摂動状態は、化合物への細胞の曝露によって引き起こされ得る。細胞状態遷移は、細胞内の細胞構成要素存在量の変化によって、したがって細胞（例えば、摂動シグネチャ）によって産生される同一性及び量の細胞構成要素（例えば、ｍＲＮＡ、転写因子）によってマークされ得る。

【０１６７】

本明細書で使用される場合、細胞又は複数の細胞についての細胞構成要素存在量測定に関連する「データセット」という用語は、いくつかの文脈において、単一細胞（例えば、単一細胞構成要素存在量データセット）から収集された高次元のデータセットを指すことができる。他の文脈では、「データセット」という用語は、単一細胞から収集された複数の高次元のデータセット（例えば、複数の単一細胞構成要素存在量データセット）を指すことができ、複数の細胞のうちの１つの細胞から収集された複数のデータセットの各々を指すことができる。

【０１６８】

本明細書で使用される場合、「差次的存在量」又は「差次的発現」という用語は、第２の実体（例えば、第２の細胞、複数の細胞、及び／又は試料）と比較して、第１の実体（例えば、第１の細胞、複数の細胞、及び／又は試料）に存在する細胞構成要素の量及び／又は頻度の差を指す。いくつかの実施形態において、第１の実体は、第１の細胞状態（例えば、罹患した表現型）を特徴とする試料であり、第２の実体は、第２の細胞状態（例えば、正常又は健康な表現型）を特徴とする試料である。例えば、細胞構成要素は、第２の細胞状態を特徴とする実体と比較して、第１の細胞状態を特徴とする実体において高レベ

ル又は低レベルで存在するポリヌクレオチド（例えば、ｍＲＮＡ転写産物）であってもよい。いくつかの実施形態において、細胞構成要素は、第２の細胞状態を特徴とする実体と比較して、第１の細胞状態を特徴とする実体においてより高い頻度又はより低い頻度で検出されるポリヌクレオチドであってもよい。細胞構成要素は、量、頻度、又は両方の点で差次的に存在し得る。いくつかの場合において、一方の実体における細胞構成要素の量が、他方の実体における細胞構成要素の量と統計的に有意に異なる場合、細胞構成要素は、２つの実体の間で差次的に存在する。例えば、細胞構成要素は、他の実体に存在するものよりも、一方の実体において少なくとも約１２０％、少なくとも約１３０％、少なくとも約１５０％、少なくとも約１８０％、少なくとも約２００％、少なくとも約３００％、少なくとも約５００％、少なくとも約７００％、少なくとも約９００％、若しくは少なくとも約１０００％大きい場合、又は一方の実体において検出可能であり、他方の実体において検出不可能である場合、２つの実体において差次的に存在する。いくつかの場合において、実体の第１のサブセット（例えば、注釈付きの細胞状態の第１のサブセットを表す細胞）における細胞構成要素を検出する頻度が、実体の第２のサブセット（例えば、注釈付きの細胞状態の第２のサブセットを表す細胞）における頻度よりも統計的に有意に高い又は低い場合、細胞構成要素は、２つの実体のセットにおいて差次的に発現される。例えば、細胞構成要素は、一方の実体のセットにおいて、他の実体のセットよりも少なくとも約１２０％、少なくとも約１３０％、少なくとも約１５０％、少なくとも約１８０％、少なくとも約２００％、少なくとも約３００％、少なくとも約５００％、少なくとも約７００％、少なくとも約９００％、又は少なくとも約１０００％以上の頻度又は以下の頻度で観察される場合、２つの実体のセットにおいて差次的に発現される。

【０１６９】

本明細書で使用される場合、「健康な」という用語は、健康な状態（例えば、良好な健康を有する対象から得られた）を特徴とする試料を指す。健康な対象は、任意の悪性又は非悪性疾患の不在を示すことができる。「健康な」個体は、アッセイされる状態とは無関係であり、通常は「健康な」とみなすことができない他の疾患又は状態を有することができる。

【０１７０】

本明細書で使用される場合、細胞に関連する「摂動」という用語（例えば、細胞の摂動又は細胞摂動）は、１つ以上の化合物による治療などの１つ以上の状態への細胞の任意の曝露を指す。これらの化合物は、「ペルターバゲン（ｐｅｒｔｕｒｂａｇｅｎｓ）」と称され得る。いくつかの実施形態において、ペルターバゲンは、例えば、小分子、生物製剤、治療剤、タンパク質、小分子と組み合わされたタンパク質、ＡＤＣ、ｓｉＲＮＡ若しくは干渉ＲＮＡなどの核酸、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ｓｈＲＮＡ、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ガイドＲＮＡ（例えば、Ｃａｓ９系若しくは他の遺伝子編集系）、又は前述のいずれかの任意の組み合わせを含むことができる。摂動は、細胞の表現型の変化、及び／又は細胞内の１つ以上の細胞構成要素の発現若しくは存在量レベルの変化（例えば、摂動シグネチャ）を誘発し得るか、又はそれによって特徴付けることができる。例えば、摂動は、細胞の転写プロファイルの変化によって特徴付けることができる。

【０１７１】

本明細書で使用される場合、「試料」、「生体試料」、又は「患者試料」という用語は、対象に関連する生物学的状態を反映し得る、対象から採取された任意の試料を指す。試料の例としては、対象の血液、全血、血漿、血清、尿、脳脊髄液、糞便、唾液、汗、涙、胸膜液、心膜液、又は腹膜液が挙げられるが、これらに限定されない。試料は、生きている又は死んでいる対象に由来する任意の組織又は材料を含むことができる。試料は、無細胞試料であってもよい。試料は、１つ以上の細胞構成要素を含むことができる。例えば、試料は、核酸（例えば、ＤＮＡ若しくはＲＮＡ）若しくはその断片、又はタンパク質を含むことができる。「核酸」という用語は、デオキシリボ核酸（ＤＮＡ）、リボ核酸（ＲＮＡ）、又はそれらの任意のハイブリッド若しくは断片を指すことができる。試料中の核酸

は、無細胞核酸であってもよい。試料は、液体試料又は固体試料（例えば、細胞又は組織試料）であってもよい。試料は、体液であってもよい。試料は、糞便試料であってもよい。試料を処理して、組織又は細胞構造を物理的に破壊し（例えば、遠心分離及び／又は細胞溶解）、したがって、細胞内成分を、分析のために試料を調製するために使用され得る酵素、緩衝液、塩、洗剤などを更に含有し得る溶液中に放出することができる。

【０１７２】

本明細書で使用される場合、化合物のフィンガープリントのような「フィンガープリント」という用語は、化合物のデジタルダイジェストである。そのようなデジタルダイジェストの非限定的な例としては、Ｄａｙｌｉｇｈｔフィンガープリント、ＢＣＩフィンガープリント、ＥＣＦＣ４フィンガープリント、ＥＣＦＰ４フィンガープリント、ＥｃＦＣフィンガープリント、ＭＤＬフィンガープリント、原子対フィンガープリント（ＡＰＦＰフィンガープリント）、トポロジカル二面角フィンガープリント（ＴＴＦＰ）フィンガープリント、ＵＮＩＴＹ　２Ｄフィンガープリント、ＲＮＮＳ２Ｓフィンガープリント、又はＧｒａｐｈＣｏｎｖフィンガープリントが挙げられる。Ｆｒａｎｃｏ，２０１４，“Ｔｈｅ　Ｕｓｅ　ｏｆ　２Ｄ　ｆｉｎｇｅｒｐｒｉｎｔ　ｍｅｔｈｏｄｓ　ｔｏ　ｓｕｐｐｏｒｔ　ｔｈｅ　ａｓｓｅｓｓｍｅｎｔ　ｏｆ　ｓｔｒｕｃｔｕｒａｌ　ｓｉｍｉｌａｒｉｔｙ　ｉｎ　ｏｒｐｈａｎ　ｄｒｕｇ　ｌｅｇｉｓｌａｔｉｏｎ，”Ｊ．Ｃｈｅｍｉｎｆｏｒｍ　６，ｐ．５、及び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を参照されたく、それらの各々は参照により本明細書に組み込まれる。また、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、及びＦｒａｎｃｏ　ｅｔ　ａｌ．，２０１４，“Ｔｈｅ　ｕｓｅ　ｏｆ　２Ｄ　ｆｉｎｇｅｒｐｒｉｎｔ　ｍｅｔｈｏｄｓ　ｔｏ　ｓｕｐｐｏｒｔ　ｔｈｅ　ａｓｓｅｓｓｍｅｎｔ　ｏｆ　ｓｔｒｕｃｔｕｒａｌ　ｓｉｍｉｌａｒｉｔｙ　ｉｎ　ｏｒｐｈａｎ　ｄｒｕｇ　ｌｅｇｉｓｌａｔｉｏｎ”Ｊｏｕｒｎａｌ　ｏｆ　ｃｈｅｍｏｉｎｆｏｒｍａｔｉｃｓ　６（５）も参照されたく、それらの各々は参照により本明細書に組み込まれる。

【０１７３】

本明細書で使用される場合、「分類」という用語は、実体（例えば、細胞、試料、細胞構成要素、細胞構成要素モジュールなど）の特定の特性（例えば、細胞プロセス、共変量、細胞状態注釈など）に関連付けられた任意の数又は他の特徴を指すことができる。例えば、「＋」記号（又は単語「正」）は、実体が特定の特性に対して正として分類されていることを示すことができる（例えば、細胞構成要素モジュールは、目的の細胞プロセスに正に関連付けられている）。別の例では、「分類」という用語は、実体と特定の特性との間の相関（例えば、それぞれの共変量とそれぞれの細胞構成要素モジュールとの間の相関）の決定を指すことができる。いくつかの実施形態において、分類は相関係数及び／又は重みである。分類は、バイナリ（例えば、正又は負）であってもよいか、又はより多くのレベルの分類（例えば、１～１０又は０～１のスケール）を有してもよい。「カットオフ」及び「閾値」という用語は、動作で使用される所定の数を指すことができる。例えば、カットオフ値は、実体が除外される上記の値を参照することができる。閾値は、特定の分類が適用される値を上回るか、又は下回る値であってもよい。これらの用語のいずれかは、これらの文脈のいずれかで使用され得る。

【０１７４】

本明細書で互換的に使用される場合、「分類子」、「モデル」、アルゴリズム、「リグ

レッサー」、及び／」又は分類子」という用語は、機械学習モデル又はアルゴリズムを指す。いくつかの実施形態において、モデルは、教師なし学習アルゴリズムである。教師なし学習アルゴリズムの一例は、クラスタ分析である。

【０１７５】

いくつかの実施形態において、モデルは、教師あり機械学習である。教師あり学習アルゴリズムの非限定的な例としては、限定されないが、ロジスティック回帰、ニューラルネットワーク、サポートベクトルマシン、ナイーブベイズアルゴリズム、最近傍アルゴリズム、ランダムフォレストアルゴリズム、決定木アルゴリズム、ブーストツリーアルゴリズム、多項式ロジスティック回帰アルゴリズム、線形モデル、線形回帰、勾配ブースティング、混合モデル、隠れマルコフモデル、ガウシアンＮＢアルゴリズム、線形判別分析、又はそれらの任意の組み合わせが挙げられる。いくつかの実施形態において、モデルは、多項分類子アルゴリズムである。いくつかの実施形態において、モデルは、２段階の確率的勾配降下法（ＳＧＤ）モデルである。いくつかの実施形態において、モデルは、ディープニューラルネットワーク（例えば、ディープアンドワイド試料レベルモデル）である。いくつかの実施形態において、本開示の分類子又はモデルは、２５以上、１００以上、１０００以上、１０，０００以上、１００，０００以上、又は1×10^{6}以上のパラメータを有するため、モデルの計算を精神的に実施することができない。

【０１７６】

更に、本明細書で使用される場合、「パラメータ」という用語は、アルゴリズム、モデル、リグレッサー、及び／又は分類子における１つ以上の入力、出力、及び／又は機能に影響を与える（例えば、修正、適応、及び／又は調整する）ことができる、アルゴリズム、モデル、リグレッサー、及び／又は分類子における内部又は外部エレメント（例えば、重み及び／又はハイパーパラメータ）の任意の係数、又は同様に任意の値を指す。例えば、いくつかの実施形態において、パラメータは、アルゴリズム、モデル、リグレッサー、及び／又は分類子の挙動、学習、及び／又は性能を制御、修正、適応、及び／又は調整するために使用され得る任意の係数、重み、及び／又はハイパーパラメータを指す。いくつかの場合において、パラメータは、アルゴリズム、モデル、リグレッサー、及び／又は分類子への入力（例えば、特徴）の影響を増加又は減少させるために使用される。非限定的な例として、いくつかの実施形態において、パラメータは、ノード（例えば、ニューラルネットワーク）の影響を増加又は減少させるために使用され、ノードは、１つ以上の活性化関数を含む。特定の入力、出力、及び／又は関数へのパラメータの割り当ては、所与のアルゴリズム、モデル、リグレッサー、及び／又は分類子のための任意の１つのパラダイムに限定されるものではなく、所望の性能のための任意の好適なアルゴリズム、モデル、リグレッサー、及び／又は分類子アーキテクチャで使用することができる。いくつかの実施形態において、パラメータは、固定値を有する。いくつかの実施形態において、パラメータの値は、手動及び／又は自動的に調整可能である。いくつかの実施形態において、パラメータの値は、アルゴリズム、モデル、リグレッサー、及び／又は分類子のための検証及び／又は訓練プロセスによって（例えば、誤差最小化及び／又は逆伝搬方法によって）修正される。いくつかの実施形態において、本開示のアルゴリズム、モデル、リグレッサー、及び／又は分類子は、複数のパラメータを含む。いくつかの実施形態において、複数のパラメータはｎ個のパラメータであり、ここで、$n\geqq2$；$n\geqq5$；$n\geqq10$；$n\geqq25$；$n\geqq40$；$n\geqq50$；$n\geqq75$；$n\geqq100$；$n\geqq125$；$n\geqq150$；$n\geqq200$；$n\geqq225$；$n\geqq250$；$n\geqq350$；$n\geqq500$；$n\geqq600$；$n\geqq750$；$n\geqq1,000$；$n\geqq2,000$；$n\geqq4,000$；$n\geqq5,000$；$n\geqq7,500$；$n\geqq10,000$；$n\geqq20,000$；$n\geqq40,000$；$n\geqq75,000$；$n\geqq100,000$；$n\geqq200,000$；$n\geqq500,000$、$n\geqq1\times10^{6}$、$n\geqq5\times10^{6}$、又は$n\geqq1\times10^{7}$である。したがって、本開示のアルゴリズム、モデル、リグレッサー、及び／又は分類子は、精神的に実施することができない。いくつかの実施形態において、ｎは、１０，０００～1×10^{7}、１００，０００～5×10^{6}、又は５００，０００～1×10^{6}である。いくつかの実施形態において、本開示のアルゴリズム、モデル、リグレ

ッサー、及び／又は分類子は、ｋ次元空間で動作し、ここで、ｋは、５又はそれよりも大きい（例えば、５、６、７、８、９、１０など）正の整数である。したがって、本開示のアルゴリズム、モデル、リグレッサー、及び／又は分類子は、精神的に実施することができない。

【０１７７】

ニューラルネットワーク。いくつかの実施形態において、モデルはニューラルネットワーク（例えば、畳み込みニューラルネットワーク及び／又は残差ニューラルネットワーク）である。人工ニューラルネットワーク（ＡＮＮ）としても知られるニューラルネットワークモデルは、畳み込み及び／又は残差ニューラルネットワークモデル（ディープラーニングモデル）を含む。ニューラルネットワークは、入力データセットを出力データセットにマッピングするように訓練され得る機械学習モデルであり得、ニューラルネットワークは、ノードの複数の層に編成されたノードの相互接続されたグループを含む。例えば、ニューラルネットワークアーキテクチャは、少なくとも入力層、１つ以上の隠れ層、及び出力層を含み得る。ニューラルネットワークは、任意の総数の層、及び任意の数の隠れ層を含み得、隠れ層は、入力データのセットを出力値又は出力値のセットにマッピングすることを可能にする訓練可能な特徴抽出器として機能する。本明細書で使用される場合、ディープラーニングモデル（ＤＮＮ）は、複数の隠れ層、例えば、２つ以上の隠れ層を含むニューラルネットワークであり得る。ニューラルネットワークの各層は、いくつかのノード（又は「ニューロン」）を含むことができる。ノードは、入力データ又は前の層のノードの出力のいずれかから直接来る入力を受信し、特定の動作、例えば、合計動作を実施することができる。いくつかの実施形態において、入力からノードへの接続は、パラメータ（例えば、重み及び／又は重み係数）に関連付けられる。いくつかの実施形態において、ノードは、入力、x_i、及びそれらに関連付けられたパラメータの全ての対の積を合計してもよい。いくつかの実施形態において、重み付けされた合計は、バイアスｂでオフセットされる。いくつかの実施形態において、ノード又はニューロンの出力は、線形関数又は非線形関数であってもよい閾値関数又は活性化関数ｆを使用してゲートされてもよい。活性化関数は、例えば、整流化線形ユニット（ＲｅＬＵ）活性化関数、漏洩ＲｅＬＵ活性化関数、又は飽和双曲線正接、同一性、バイナリステップ、ロジスティック、ａｒｃＴａｎ、ソフトサイン、パラメトリック整流化線形ユニット、指数線形ユニット、ｓｏｆｔＰｌｕｓ、ベント同一性、ｓｏｆｔＥｘｐｏｎｅｎｔｉａｌ、正弦曲線、正弦、ガウシアン、若しくはシグモイド関数などの他の関数、又はそれらの任意の組み合わせであり得る。

【０１７８】

ニューラルネットワークの重み付け係数、バイアス値、及び閾値、又は他の計算パラメータは、訓練データの１つ以上のセットを使用して、訓練段階で「教示」又は「学習」され得る。例えば、パラメータは、ＡＮＮが計算する出力値が訓練データセットに含まれる例と一致するように、訓練データセットからの入力データ及び勾配降下又は後方伝搬法を使用して訓練され得る。パラメータは、逆伝搬ニューラルネットワーク訓練プロセスから取得され得る。

【０１７９】

様々なニューラルネットワークのいずれも、対象の画像を分析する際に使用するのに好適であり得る。例は、限定されないが、フィードフォワードニューラルネットワーク、放射基底関数ネットワーク、再帰ニューラルネットワーク、残差ニューラルネットワーク、畳み込みニューラルネットワーク、残差畳み込みニューラルネットワークなど、又はそれらの任意の組み合わせを含むことができる。いくつかの実施形態において、機械学習は、事前に訓練された及び／若しくは転移学習されたＡＮＮ又はディープラーニングアーキテクチャを利用する。畳み込み及び／又は残差ニューラルネットワークは、本開示に従って対象の画像を分析するために使用することができる。

【０１８０】

例えば、ディープニューラルネットワークモデルは、入力層、複数の個別にパラメータ化された（例えば、重み付けされた）畳み込み層、及び出力スコアラーを含む。畳み込み

層の各々のパラメータ（例えば、重み）並びに入力層は、ディープニューラルネットワークモデルと関連付けられた複数のパラメータ（例えば、重み）に寄与する。いくつかの実施形態において、少なくとも１００のパラメータ、少なくとも１０００のパラメータ、少なくとも２０００のパラメータ、又は少なくとも５０００のパラメータは、ディープニューラルネットワークモデルに関連付けられる。そのため、ディープニューラルネットワークモデルは、精神的に解決され得ないため、コンピュータを使用する必要がある。換言すれば、モデルへの入力が与えられた場合、そのような実施形態において、モデル出力は、精神的にではなく、コンピュータを使用して決定される必要がある。例えば、Ｋｒｉｚｈｅｖｓｋｙ ｅｔ ａｌ．，２０１２，“Ｉｍａｇｅｎｅｔ ｃｌａｓｓｉｆｉｃａｔｉｏｎ ｗｉｔｈ ｄｅｅｐ ｃｏｎｖｏｌｕｔｉｏｎａｌ ｎｅｕｒａｌ ｎｅｔｗｏｒｋｓ，”ｉｎ Ａｄｖａｎｃｅｓ ｉｎ Ｎｅｕｒａｌ Ｉｎｆｏｒｍａｔｉｏｎ Ｐｒｏｃｅｓｓｉｎｇ Ｓｙｓｔｅｍｓ ２，Ｐｅｒｅｉｒａ，Ｂｕｒｇｅｓ，Ｂｏｔｔｏｕ，Ｗｅｉｎｂｅｒｇｅｒ，ｅｄｓ．，ｐｐ．１０９７－１１０５，Ｃｕｒｒａｎ Ａｓｓｏｃｉａｔｅｓ，Ｉｎｃ．、Ｚｅｉｌｅｒ，２０１２“ＡＤＡＤＥＬＴＡ：ａｎ ａｄａｐｔｉｖｅ ｌｅａｒｎｉｎｇ ｒａｔｅ ｍｅｔｈｏｄ，”’ＣｏＲＲ，ｖｏｌ．ａｂｓ／１２１２．５７０１、及びＲｕｍｅｌｈａｒｔ ｅｔ ａｌ．，１９８８，“Ｎｅｕｒｏ ｃｏｍｐｕｔｉｎｇ：Ｆｏｕｎｄａｔｉｏｎｓ ｏｆ ｒｅｓｅａｒｃｈ，”ｃｈ．Ｌｅａｒｎｉｎｇ Ｒｅｐｒｅｓｅｎｔａｔｉｏｎｓ ｂｙ Ｂａｃｋ－ｐｒｏｐａｇａｔｉｎｇ Ｅｒｒｏｒｓ，ｐｐ．６９６－６９９，Ｃａｍｂｒｉｄｇｅ，ＭＡ，ＵＳＡ：ＭＩＴ Ｐｒｅｓｓを参照されたく、それらの各々は参照により本明細書に組み込まれる。

【０１８１】

モデルとしての使用に好適な畳み込みニューラルネットワークモデルを含む、ニューラルネットワークモデルは、例えば、Ｖｉｎｃｅｎｔ ｅｔ ａｌ．，２０１０，“Ｓｔａｃｋｅｄ ｄｅｎｏｉｓｉｎｇ ａｕｔｏｅｎｃｏｄｅｒｓ：Ｌｅａｒｎｉｎｇ ｕｓｅｆｕｌ ｒｅｐｒｅｓｅｎｔａｔｉｏｎｓ ｉｎ ａ ｄｅｅｐ ｎｅｔｗｏｒｋ ｗｉｔｈ ａ ｌｏｃａｌ ｄｅｎｏｉｓｉｎｇ ｃｒｉｔｅｒｉｏｎ，”Ｊ Ｍａｃｈ Ｌｅａｒｎ Ｒｅｓ １１，ｐｐ．３３７１－３４０８、Ｌａｒｏｃｈｅｌｌｅ ｅｔ ａｌ．，２００９，“Ｅｘｐｌｏｒｉｎｇ ｓｔｒａｔｅｇｉｅｓ ｆｏｒ ｔｒａｉｎｉｎｇ ｄｅｅｐ ｎｅｕｒａｌ ｎｅｔｗｏｒｋｓ，”Ｊ Ｍａｃｈ Ｌｅａｒｎ Ｒｅｓ １０，ｐｐ．１－４０、及びＨａｓｓｏｕｎ，１９９５，Ｆｕｎｄａｍｅｎｔａｌｓ ｏｆ Ａｒｔｉｆｉｃｉａｌ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋｓ，Ｍａｓｓａｃｈｕｓｅｔｔｓ Ｉｎｓｔｉｔｕｔｅ ｏｆ Ｔｅｃｈｎｏｌｏｇｙに開示されており、それらの各々は参照により本明細書に組み込まれる。モデルとしての使用に好適な更なる例示的なニューラルネットワークは、Ｄｕｄａ ｅｔ ａｌ．，２００１，Ｐａｔｔｅｒｎ Ｃｌａｓｓｉｆｉｃａｔｉｏｎ，Ｓｅｃｏｎｄ Ｅｄｉｔｉｏｎ，Ｊｏｈｎ Ｗｉｌｅｙ ＆ Ｓｏｎｓ，Ｉｎｃ．，Ｎｅｗ Ｙｏｒｋ、及びＨａｓｔｉｅ ｅｔ ａｌ．，２００１，Ｔｈｅ Ｅｌｅｍｅｎｔｓ ｏｆ Ｓｔａｔｉｓｔｉｃａｌ Ｌｅａｒｎｉｎｇ，Ｓｐｒｉｎｇｅｒ－Ｖｅｒｌａｇ，Ｎｅｗ Ｙｏｒｋに開示されており、それらの各々は参照によりその全体が本明細書に組み込まれる。モデルとしての使用に好適な更なる例示的なニューラルネットワークはまた、Ｄｒａｇｈｉｃｉ，２００３，Ｄａｔａ Ａｎａｌｙｓｉｓ Ｔｏｏｌｓ ｆｏｒ ＤＮＡ Ｍｉｃｒｏａｒｒａｙｓ，Ｃｈａｐｍａｎ ＆ Ｈａｌｌ／ＣＲＣ、及びＭｏｕｎｔ，２００１，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ：ｓｅｑｕｅｎｃｅ ａｎｄ ｇｅｎｏｍｅ ａｎａｌｙｓｉｓ，Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ Ｌａｂｏｒａｔｏｒｙ Ｐｒｅｓｓ，Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ，Ｎｅｗ Ｙｏｒｋにも記載されており、それらの各々は参照によりその全体が本明細書に組み込まれる。

【０１８２】

サポートベクトルマシン。いくつかの実施形態において、モデルはサポートベクトルマシン（ＳＶＭ）である。モデルとしての使用に好適なＳＶＭモデルは、例えば、Ｃｒｉｓｔｉａｎｉｎｉ ａｎｄ Ｓｈａｗｅ－Ｔａｙｌｏｒ，２０００，“Ａｎ Ｉｎｔｒｏｄ

uction to Support Vector Machines,”Cambridge University Press,Cambridge、Boser et al.,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、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.、Duda,Pattern Classification,Second Edition,2001,John Wiley & Sons,Inc.,pp.259,262－265、及びHastie,2001,The Elements of Statistical Learning,Springer,New York、及びFurey et al.,2000,Bioinformatics 16,906－914に記載されており、それらの各々は参照によりその全体が本明細書に組み込まれる。分類に使用される場合、ＳＶＭは、標識されたデータから、最大限に離れたハイパープレーンを使用して、バイナリ標識されたデータの所与のセットを分離する。線形分離が不可能な場合、ＳＶＭは、特徴空間への非線形マッピングを自動的に実現する「カーネル」の技術と組み合わせて機能することができる。特徴空間内のＳＶＭによって見出されるハイパープレーンは、入力空間内の非線形決定境界に対応し得る。いくつかの実施形態において、ＳＶＭに関連付けられた複数のパラメータ（例えば、重み）は、ハイパープレーンを定義する。いくつかの実施形態において、ハイパープレーンは、少なくとも１０、少なくとも２０、少なくとも５０、又は少なくとも１００のパラメータによって定義され、ＳＶＭモデルは、それが精神的に解決され得ないため、計算するのにコンピュータを必要とする。

【０１８３】

ナイーブベイズモデル。いくつかの実施形態において、モデルはナイーブベイズモデルである。モデルとしての使用に好適なナイーブベイズモデルは、例えば、Ｎｇ et al.,２００２,“On discriminative vs.generative classifiers：A comparison of logistic regression and naive Bayes,”Advances in Neural Information Processing Systems,１４に開示されており、これは参照により本明細書に組み込まれる。ナイーブベイズ分類子は、特徴間の強い（ナイーブな）独立性仮定を伴うベイズの定理を適用することに基づく、「確率的分類子」のファミリー内の任意の分類子である。いくつかの実施形態において、それらは、カーネル密度推定と結合される。例えば、Ｈａｓｔｉｅ et al.,２００１,The elements of statistical learning：data mining,inference,and prediction,eds.Tibshirani and Friedman,Springer,New Yorkを参照されたく、これは参照により本明細書に組み込まれる。

【０１８４】

最近傍モデル。いくつかの実施形態において、モデルは、最近傍モデルである。最近傍モデルは、メモリベースであり得、適合するモデルを含まない。最近傍に関して、クエリ点x_0（試験対象）が与えられると、ｋ個の訓練点$x_{(r)}$、ｒ、．．．、ｋ（ここでは訓練対象）が、x_0に最も近い距離で識別され、次に点x_0がｋ個の最近傍を使用して分類される。ここで、これらの隣接までの距離は、識別遺伝子セットの存在量値の関数である。いくつかの実施形態において、特徴空間内のユークリッド距離は、$d_{(i)}=||x_{(i)}-x_{(O)}||$として距離を決定するために使用される。典型的には、最近傍モデルが使用される場合、線形判別を計算するために使用される存在量データは、平均ゼロ及

び分散１を有するように標準化される。最近傍法則は、不等クラス事前判定、差異的誤分類コスト、及び特徴選択の問題に対処するために改良することができる。これらの改良の多くは、隣接のための何らかの形態の重み付け投票を伴う。最近傍分析に関する更なる情報については、Ｄｕｄａ，Ｐａｔｔｅｒｎ Ｃｌａｓｓｉｆｉｃａｔｉｏｎ，Ｓｅｃｏｎｄ Ｅｄｉｔｉｏｎ，２００１，Ｊｏｈｎ Ｗｉｌｅｙ ＆ Ｓｏｎｓ，Ｉｎｃ、及びＨａｓｔｉｅ，２００１，Ｔｈｅ Ｅｌｅｍｅｎｔｓ ｏｆ Ｓｔａｔｉｓｔｉｃａｌ Ｌｅａｒｎｉｎｇ，Ｓｐｒｉｎｇｅｒ，Ｎｅｗ Ｙｏｒｋを参照されたく、それらの各々は参照により本明細書に組み込まれる。

【０１８５】

ｋ最近傍モデルは、入力が特徴空間におけるｋ個の最も近い訓練例からなる非パラメトリック機械学習方法である。出力はクラスメンバーシップである。オブジェクトは、隣接するオブジェクトの複数の投票によって分類され、オブジェクトは、その最も近い隣接するｋ個の中で最も一般的なクラスに割り当てられる（ｋは、典型的には小さい正の整数である）。ｋ＝１の場合、オブジェクトは単にその単一の最も近い隣接のクラスに割り当てられる。参照により本明細書に組み込まれる、Ｄｕｄａ ｅｔ ａｌ．，２００１，Ｐａｔｔｅｒｎ Ｃｌａｓｓｉｆｉｃａｔｉｏｎ，Ｓｅｃｏｎｄ Ｅｄｉｔｉｏｎ，Ｊｏｈｎ Ｗｉｌｅｙ ＆ Ｓｏｎｓを参照されたい。いくつかの実施形態において、ｋ最近傍モデルを解くために必要な距離計算の数は、それが精神的に実施され得ないために、コンピュータが所与の入力についてのモデルを解くために使用されるようなものである。

【０１８６】

ランダムフォレスト、決定木、及びブーストツリーモデル。いくつかの実施形態において、モデルは、決定木である。モデルとしての使用に好適な決定木は、Ｄｕｄａ，２００１，Ｐａｔｔｅｒｎ Ｃｌａｓｓｉｆｉｃａｔｉｏｎ，Ｊｏｈｎ Ｗｉｌｅｙ ＆ Ｓｏｎｓ，Ｉｎｃ．，Ｎｅｗ Ｙｏｒｋ，ｐｐ．３９５－３９６に概説されており、これは参照により本明細書に組み込まれる。決定木に基づく方法は、特徴空間を長方形のセットに分割し、各々に（定数のような）モデルを適合させる。いくつかの実施形態において、決定木はランダムフォレスト回帰である。使用され得る１つの特定のモデルは、分類及び回帰木（ＣＡＲＴ）である。他の特定の決定木モデルには、ＩＤ３、Ｃ４．５、ＭＡＲＴ、及びランダムフォレストが含まれるが、これらに限定されない。ＣＡＲＴ、ＩＤ３、及びＣ４．５は、Ｄｕｄａ，２００１，Ｐａｔｔｅｒｎ Ｃｌａｓｓｉｆｉｃａｔｉｏｎ，Ｊｏｈｎ Ｗｉｌｅｙ ＆ Ｓｏｎｓ，Ｉｎｃ．，Ｎｅｗ Ｙｏｒｋ，ｐｐ．３９６－４０８ ａｎｄ ｐｐ．４１１－４１２に記載されており、これは参照により本明細書に組み込まれる。ＣＡＲＴ、ＭＡＲＴ、及びＣ４．５は、Ｈａｓｔｉｅ ｅｔ ａｌ．，２００１，Ｔｈｅ Ｅｌｅｍｅｎｔｓ ｏｆ Ｓｔａｔｉｓｔｉｃａｌ Ｌｅａｒｎｉｎｇ，Ｓｐｒｉｎｇｅｒ－Ｖｅｒｌａｇ，Ｎｅｗ Ｙｏｒｋ，Ｃｈａｐｔｅｒ ９に記載されており、これは参照によりその全体が本明細書に組み込まれる。ランダムフォレストは、Ｂｒｅｉｍａｎ，１９９９，“Ｒａｎｄｏｍ Ｆｏｒｅｓｔｓ－－Ｒａｎｄｏｍ Ｆｅａｔｕｒｅｓ，”Ｔｅｃｈｎｉｃａｌ Ｒｅｐｏｒｔ ５６７，Ｓｔａｔｉｓｔｉｃｓ Ｄｅｐａｒｔｍｅｎｔ，Ｕ．Ｃ．Ｂｅｒｋｅｌｅｙ，Ｓｅｐｔｅｍｂｅｒ １９９９に記載されており、これは参照によりその全体が本明細書に組み込まれる。いくつかの実施形態において、決定木モデルは、少なくとも１０、少なくとも２０、少なくとも５０、又は少なくとも１００のパラメータ（例えば、重み及び／又は決定）を含み、それが精神的に解決され得ないため、計算するのにコンピュータを必要とする。

【０１８７】

回帰。いくつかの実施形態において、モデルは回帰を使用する。回帰アルゴリズムは、任意のタイプの回帰であり得る。例えば、いくつかの実施形態において、回帰はロジスティック回帰である。いくつかの実施形態において、回帰は、ラッソ、Ｌ２又は弾性ネット正規化によるロジスティック回帰である。いくつかの実施形態において、閾値を満たすことに失敗する対応する回帰係数を有するこれらの抽出された特徴は、考慮から取り除かれる（削除される）。いくつかの実施形態において、マルチカテゴリ応答を扱うロジスティ

ック回帰モデルの一般化が、モデルとして使用される。ロジスティック回帰は、Ａｇｒｅｓｔｉ，Ａｎ　Ｉｎｔｒｏｄｕｃｔｉｏｎ　ｔｏ　Ｃａｔｅｇｏｒｉｃａｌ　Ｄａｔａ　Ａｎａｌｙｓｉｓ，１９９６，Ｃｈａｐｔｅｒ　５，ｐｐ．１０３－１４４，Ｊｏｈｎ　Ｗｉｌｅｙ　＆　Ｓｏｎ，Ｎｅｗ　Ｙｏｒｋに開示されており、これは参照により本明細書に組み込まれる。いくつかの実施形態において、モデルは、Ｈａｓｔｉｅ　ｅｔ　ａｌ．，２００１，Ｔｈｅ　Ｅｌｅｍｅｎｔｓ　ｏｆ　Ｓｔａｔｉｓｔｉｃａｌ　Ｌｅａｒｎｉｎｇ，Ｓｐｒｉｎｇｅｒ－Ｖｅｒｌａｇ，Ｎｅｗ　Ｙｏｒｋに開示されている回帰モデルを利用する。いくつかの実施形態において、ロジスティック回帰モデルは、少なくとも１０、少なくとも２０、少なくとも５０、少なくとも１００、又は少なくとも１０００のパラメータ（例えば、重み）を含み、それが精神的に解決され得ないため、計算するのにコンピュータを必要とする。

【０１８８】

線形判別分析。線形判別分析（ＬＤＡ）、正規判別分析（ＮＤＡ）、又は判別関数分析は、２つ以上のクラスの対象又はイベントを特徴付けるか、又は分離する特徴の線形組み合わせを見出すための統計学、パターン認識、及び機械学習で使用される方法であるフィッシャーの線形判別の一般化であり得る。得られる組み合わせは、本開示のいくつかの実施形態においてモデル（線形モデル）として使用され得る。

【０１８９】

混合モデル及び隠れマルコフモデル。いくつかの実施形態において、モデルは、ＭｃＬａｃｈｌａｎ　ｅｔ　ａｌ．，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ　１８（３）：４１３－４２２，２００２に記載されるような混合モデルである。いくつかの実施形態において、特に、時間コンポーネントを含むそれらの実施形態において、モデルは、Ｓｃｈｌｉｅｐ　ｅｔ　ａｌ．，２００３，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ　１９（１）：ｉ２５５－ｉ２６３に記載されるような隠れマルコフモデルである。

【０１９０】

クラスタリング。いくつかの実施形態において、モデルは教師なしクラスタリングモデルである。いくつかの実施形態において、モデルは教師ありクラスタリングモデルである。モデルとしての使用に好適なクラスタリングは、例えば、Ｄｕｄａ　ａｎｄ　Ｈａｒｔ，Ｐａｔｔｅｒｎ　Ｃｌａｓｓｉｆｉｃａｔｉｏｎ　ａｎｄ　Ｓｃｅｎｅ　Ａｎａｌｙｓｉｓ，１９７３，Ｊｏｈｎ　Ｗｉｌｅｙ　＆　Ｓｏｎｓ，Ｉｎｃ．，Ｎｅｗ　Ｙｏｒｋ（本明細書以下では、“Ｄｕｄａ　１９７３”）の２１１～２５６ページに記載されており、これは参照によりその全体が本明細書に組み込まれる。クラスタリング問題は、データセット内の自然なグルーピングを見出すことの１つとして記述することができる。自然なグルーピングを識別するために、２つの問題に対処することができる。第一に、２つの試料間の類似性（又は相違性）を測定する方法を決定することができる。このメトリック（例えば、類似性尺度）を使用して、１つのクラスタ内の試料が他のクラスタ内の試料よりも互いにより類似していることを確実にすることができる。第二に、類似性尺度を使用してデータをクラスタに分割するための機構を決定することができる。クラスタリング調査を開始する１つの方法は、距離関数を定義し、訓練セット内の試料の全てのペア間の距離のマトリックスを計算することであり得る。距離が類似性の良好な尺度である場合、同じクラスタ内の参照実体間の距離は、異なるクラスタ内の参照実体間の距離よりも有意に小さくてもよい。しかしながら、クラスタリングは、距離メトリックを使用しなくてもよい。例えば、ノンメトリック類似性関数ｓ（ｘ，ｘ’）を使用して、２つのベクトルｘ及びｘ’を比較することができる。ｓ（ｘ，ｘ’）は、ｘ及びｘ’が何らかの形で「類似している」ときに値が大きい対称関数であり得る。データセット内の点間の「類似性」又は「相違性」を測定するための方法が選択されると、クラスタリングは、データの任意のパーティションのクラスタリング品質を測定する基準関数を使用することができる。基準関数を極端化するデータセットのパーティションを使用して、データをクラスタリングすることができる。本開示で使用することができる特定の例示的なクラスタリング技術は、階層的クラスタリング（最近傍アルゴリズム、最遠傍アルゴリズム、平均リンケージアルゴリズム、

重心アルゴリズム、又は二乗和アルゴリズムを使用した凝集クラスタリング）、ｋ平均クラスタリング、ファジーｋ平均クラスタリング、及びジャービス・パトリック（Ｊａｒｖｉｓ－Ｐａｔｒｉｃｋ）クラスタリングを含むことができるが、これらに限定されない。いくつかの実施形態において、クラスタリングは、教師なしクラスタリング（例えば、事前に考えられた数のクラスタ及び／又はクラスタ割り当ての事前決定を伴わない）を含む。

【０１９１】

モデル及びブースティングのアンサンブル。いくつかの実施形態において、モデルのアンサンブル（２つ以上）が使用される。いくつかの実施形態において、ＡｄａＢｏｏｓｔなどのブースティング技術は、モデルの性能を改善するために、多くの他のタイプの学習アルゴリズムと併せて使用される。このアプローチでは、本明細書に開示されるモデルのいずれか、又はそれらの等価物の出力は、ブーストされたモデルの最終出力を表す加重合計に組み合わされる。いくつかの実施形態において、モデルからの複数の出力は、平均、中央値、モード、加重平均、加重中央値、加重モードなどを含むが、これらに限定されない、当該技術分野で既知の中心傾向の任意の尺度を使用して組み合わされる。いくつかの実施形態において、複数の出力は、投票方法を使用して組み合わされる。いくつかの実施形態において、モデルのアンサンブル内のそれぞれのモデルは、重み付けされるか、又は重み付けされない。

【０１９２】

本明細書で使用される場合、「訓練されていないモデル」（例えば、「訓練されていないリグレッサー」及び／又は「訓練されていない分類子」）という用語は、訓練データセットで訓練されていないリグレッサー又は分類子などの機械学習モデルを指す。本明細書で使用される場合、「モデルを訓練する」という用語は、訓練されていない、又は部分的に訓練されたモデルを訓練するプロセスを指す。例えば、いくつかの実施形態において、モデルを訓練することは、潜在表現で配置された複数の細胞構成要素モジュール及び以下で説明される細胞構成要素カウントデータ構造を得ることを含む。潜在表現及び細胞構成要素カウントデータ構造で配置された複数の細胞構成要素モジュールは、活性化データ構造（本明細書以下、「一次訓練データセット」）内の複数の細胞構成要素モジュールについての複数の共変量における各共変量の存在の実際の不在と併せて、訓練されていない又は部分的に訓練されたモデルに集合的な入力として適用される活性化データ構造を形成するために組み合わされて、共変量モジュール相関で訓練されていない又は部分的に訓練されたモデルを訓練し、それによって訓練されたモデルを得る。更に、「訓練されていないモデル」という用語は、転移学習技術が訓練されていないモデルのそのような訓練に使用される可能性を排除しないことを理解されたい。例えば、参照により本明細書に組み込まれる、Ｆｅｒｎａｎｄｅｓ ｅｔ ａｌ．，２０１７，“Ｔｒａｎｓｆｅｒ Ｌｅａｒｎｉｎｇ ｗｉｔｈ Ｐａｒｔｉａｌ Ｏｂｓｅｒｖａｂｉｌｉｔｙ Ａｐｐｌｉｅｄ ｔｏ Ｃｅｒｖｉｃａｌ Ｃａｎｃｅｒ Ｓｃｒｅｅｎｉｎｇ，”Ｐａｔｔｅｒｎ Ｒｅｃｏｇｎｉｔｉｏｎ ａｎｄ Ｉｍａｇｅ Ａｎａｌｙｓｉｓ：８ｔｈ Ｉｂｅｒｉａｎ Ｃｏｎｆｅｒｅｎｃｅ Ｐｒｏｃｅｅｄｉｎｇｓ，２４３－２５０は、そのような転移学習の非限定的な例を提供する。転移学習が使用される例では、上述の訓練されていないモデルは、一次訓練データセットのもの以上の追加データを提供される。すなわち、転移学習の実施形態の非限定的な例では、訓練されていないモデルは、（ｉ）一次訓練データセット及び（ｉｉ）追加のデータを受信する。典型的には、この追加データは、別の補助訓練データセットから学習された係数（例えば、回帰係数）の形態である。更に、単一の補助訓練データセットの説明が開示されているが、本開示における訓練されていないモデルを訓練する際に一次訓練データセットを補完するために使用され得る補助訓練データセットの数に制限がないことを理解されたい。例えば、いくつかの実施形態において、２つ以上の補助訓練データセット、３つ以上の補助訓練データセット、４つ以上の補助訓練データセット、又は５つ以上の補助訓練データセットは、転移学習を通じて一次訓練データセットを補完するために使用され、そのような各補助データセットは一次訓練データセットと

は異なる。そのような実施形態において、転移学習の任意の方法が使用され得る。例えば、一次訓練データセットに加えて、第１の補助訓練データセット及び第２の補助訓練データセットが存在する場合を考慮する。第１の補助訓練データセットから（第１の補助訓練データセットへの回帰などのモデルの適用によって）学習された係数は、転移学習技術（例えば、２次元マトリックス乗算）を使用して第２の補助訓練データセットに適用され得、これは、次いで、係数が一次訓練データセットに適用される訓練された中間モデルをもたらし得、これは、一次訓練データセット自体と併せて、訓練されていないモデルに適用される。あるいは、第１の補助訓練データセットから学習された係数の第１のセット（第１の補助訓練データセットへの回帰などのモデルの適用によって）及び第２の補助訓練データセットから学習された係数の第２のセット（第２の補助訓練データセットへの回帰などのモデルの適用によって）は、各々個別に一次訓練データセットの別個のインスタンスに適用され得（例えば、別個の独立マトリックス乗算によって）、一次訓練データセット自体（又は一次訓練データセットから学習された主要コンポーネント若しくは回帰係数などの一次訓練データセットのいくつかの縮小形式）と併せて一次訓練データセットのインスタンスを分離するための係数のそのような適用の両方が、次いで、訓練されていないモデルを訓練するために訓練されていないモデルに適用され得る。いずれかの例では、第１及び第２の補助訓練データセットから導出される共変量モジュール相関（例えば、追加の細胞状態注釈、追加の共変量、及び／又はその細胞構成要素存在量など）に関する知識を、共変量で標識された一次訓練データセットと併せて、訓練されていないモデルを訓練するために使用される。

【０１９３】

本明細書で互換的に使用される場合、「ニューロン」、「ノード」、「ユニット」、「隠れニューロン」、「隠れユニット」などの用語は、入力を受け入れ、活性化関数及び１つ以上のパラメータ（例えば、係数及び／又は重み）を介して出力を提供するニューラルネットワークのユニットを指す。例えば、隠れニューロンは、以前の層からの１つ以上の入力を受け入れ、後続の層についての入力として機能する出力を提供することができる。いくつかの実施形態において、ニューラルネットワークは、１つの出力ニューロンのみを含む。いくつかの実施形態において、ニューラルネットワークは、複数の出力ニューロンを含む。一般的に、出力は、共変量、細胞状態注釈、又は目的の細胞プロセスなどの目的の状態の確率若しくは尤度、バイナリ判定（例えば、存在又は不在、正又は負の結果）、及び／又は標識（例えば、分類及び／又は相関係数）などの予測値である。単一クラス分類モデルの場合、出力は、状態（例えば、共変量、細胞状態注釈、及び／又は目的の細胞プロセス）を有する入力特徴（例えば、１つ以上の細胞構成要素モジュール）の尤度（例えば、相関係数及び／又は重み）であり得る。マルチクラス分類モデルの場合、複数の予測値を生成することができ、各予測値は、目的の状態の各々についての入力特徴の尤度を示す。

【０１９４】

本明細書で使用される場合、「パラメータ」という用語は、モデル、分類子、又はアルゴリズムにおける１つ以上の入力、出力、及び／又は機能に影響を与える（例えば、修正、適応、及び／又は調整する）ことができる、モデル、分類子、又はアルゴリズムにおける内部又は外部エレメント（例えば、重み及び／又はハイパーパラメータ）の任意の係数、又は同様に任意の値を指す。いくつかの実施形態において、パラメータは、モデルにおける１つ以上の入力、出力、又は関数を調節する係数（例えば、重み）である。例えば、パラメータの値を使用して、モデルへの入力（例えば、特徴）の影響をアップウェイト又はダウンウェイトすることができる。特徴は、ロジスティック回帰、ＳＶＭ、又はナイーブベイズモデルなどのパラメータと関連付けることができる。パラメータの値は、代替的又は追加的に、ニューラルネットワークにおけるノード（例えば、ノードは、入力から出力への変換を定義する１つ以上の活性化関数を含む）、クラス、又はインスタンス（例えば、複数の細胞における細胞）の影響をアップウェイト又はダウンウェイトするために使用することができる。特定の入力、出力、機能、又は特徴へのパラメータの割り当ては、

所与のモデルのための任意の１つのパラダイムに限定されないが、最適な性能のための任意の適切なモデルアーキテクチャで使用され得る。いくつかの例では、モデルの入力、出力、機能、又は特徴と関連付けられたパラメータ（例えば、係数）への参照は、機械学習モデルの計算の複雑性のコンテキストなどにおいて、同じものの数、性能、又は最適化の指標として同様に使用され得る。いくつかの実施形態において、パラメータは、固定値を有する。いくつかの実施形態において、パラメータの値は、手動及び／又は自動的に（例えば、ハイパーパラメータ最適化方法を使用して）調整可能である。いくつかの実施形態において、パラメータの値は、モデル検証及び／又は訓練プロセスによって（例えば、本明細書の他の箇所に記載されるように、エラー最小化及び／又は逆伝搬方法によって）修正される。

【０１９５】

本明細書で使用される場合、「ベクトル」という用語は、エレメントの配列などのエレメントの列挙されたリストであり、各エレメントは割り当てられた意味を有する。したがって、本開示で使用される「ベクトル」という用語は、「テンソル」という用語と互換性がある。例として、ベクトルが存在量カウントを含む場合、複数の細胞において、それぞれの細胞構成要素について、複数の細胞の各々の１つについて、ベクトルに所定のエレメントが存在する。提示を容易にするために、いくつかの例では、ベクトルは、一次元であると説明され得る。しかしながら、本開示は、そのように限定されない。任意の次元のベクトルは、ベクトルにおける各エレメントが表すものの説明が定義されている（例えば、そのエレメント１は、複数の細胞の細胞１の存在量カウントなどを表す）ことを条件として、本開示で使用することができる。

【０１９６】

Ｉ．例示的なシステムの実施形態

本開示のいくつかの態様の概要及び本開示で使用されるいくつかの定義が提供されたので、例示的なシステムの詳細は、図１と併せて説明される。

【０１９７】

図１は、本開示のいくつかの実施形態によるシステム１００を示すブロック図を提供する。システム１００は、目的の細胞プロセスと関連付けられた複数の細胞構成要素モジュールにおける１つ以上の細胞構成要素モジュールの決定を提供する。図１では、システム１００はコンピューティングデバイスとして示されている。コンピュータシステム１００の他のトポロジが可能である。例えば、いくつかの実施形態において、システム１００は、実際には、ネットワーク内で一緒にリンクされるか、又はクラウドコンピューティング環境内で仮想マシン若しくはコンテナであるいくつかのコンピュータシステムを構成し得る。したがって、図１に示される例示的トポロジは、当業者に容易に理解されるような様式で、本開示の一実施形態の特徴を説明する役割を果たすだけである。

【０１９８】

図１を参照すると、いくつかの実施形態において、コンピュータシステム１００（例えば、コンピューティングデバイス）は、ネットワークインターフェース１０４を含む。いくつかの実施形態において、ネットワークインターフェース１０４は、１つ以上の通信ネットワークを通じて（例えば、ネットワーク通信モジュール１５８を通じて）、システム内のシステム１００コンピューティングデバイスを互いに、並びに任意選択の外部システム及びデバイスと相互接続する。いくつかの実施形態において、ネットワークインターフェース１０４は、インターネット、１つ以上のローカルエリアネットワーク（ＬＡＮ）、１つ以上のワイドエリアネットワーク（ＷＡＮ）、他のタイプのネットワーク、又はそのようなネットワークの組み合わせを介してネットワーク通信モジュール１５８を通じた通信を任意選択で提供する。

【０１９９】

ネットワークの例としては、ワールドワイドウェブ（ＷＷＷ）、イントラネット及び／又は無線ネットワーク、例えば携帯電話ネットワーク、無線ローカルエリアネットワーク（ＬＡＮ）及び／又は首都圏ネットワーク（ＭＡＮ）、並びに無線通信による他のデバイ

スが挙げられる。無線通信は、グローバルモバイルコミュニケーションシステム（ＧＳＭ）、エンハンスドデータＧＳＭ環境（ＥＤＧＥ）、高速ダウンリンクパケットアクセス（ＨＳＤＰＡ）、高速アップリンクパケットアクセス（ＨＳＵＰＡ）、エボリューション、データ専用（ＥＶ－ＤＯ）、ＨＳＰＡ、ＨＳＰＡ＋、デュアルセルＨＳＰＡ（ＤＣ－ＨＳＰＤＡ）、ロングタームエボリューション（ＬＴＥ）、近距離通信（ＮＦＣ）、広帯域コード分割多重アクセス（Ｗ－ＣＤＭＡ）、コード分割多重アクセス（ＣＤＭＡ）、時分割多重アクセス（ＴＤＭＡ）、Ｂｌｕｅｔｏｏｔｈ、ワイヤレスフィデリティ（Ｗｉ－Ｆｉ）（例えば、ＩＥＥＥ　８０２．１１ａ、ＩＥＥＥ　８０２．１１ａｃ、ＩＥＥＥ　８０２．１１ａｘ、ＩＥＥＥ　８０２．１１ｂ、ＩＥＥＥ　８０２．１１ｇ及び／若しくはＩＥＥＥ　８０２．１１ｎ）、ボイスオーバーインターネットプロトコル（ＶｏＩＰ）、Ｗｉ－ＭＡＸ、電子メール用プロトコル（例えば、インターネットメッセージアクセスプロトコル（ＩＭＡＰ）及び／若しくはポストオフィスプロトコル（ＰＯＰ））、インスタントメッセージング（例えば、エクステンシブルメッセージング及びプレゼンスプロトコル（ＸＭＰＰ）、インスタントメッセージング及びプレゼンスレバレッジ拡張機能のセッション開始プロトコル（ＳＩＭＰＬＥ）、インスタントメッセージング及びプレゼンスサービス（ＩＭＰＳ））、並びに／又はショートメッセージングサービス（ＳＭＳ）、あるいは本書の出願日の時点でまだ開発されていない通信プロトコルを含む任意の他の好適な通信プロトコルを含む、複数の通信規格、プロトコル及び技術のいずれかを任意選択で使用する。

【０２００】

いくつかの実施形態において、システム１００は、１つ以上の処理ユニット（ＣＰＵ）１０２（例えば、プロセッサ、処理コアなど）、１つ以上のネットワークインターフェース１０４、ユーザによって使用されるためのディスプレイ１０８及び入力システム１０５（例えば、入力／出力インターフェース、キーボード、マウスなど）を（任意選択で）含むユーザインターフェース１０６、メモリ（例えば、非永続的メモリ１０７、永続的メモリ１０９）、並びに前述のコンポーネントを相互接続するための１つ以上の通信バス１０３を含む。１つ以上の通信バス１０３は、システムコンポーネント間の通信を相互接続及び制御する回路（チップセットと呼ばれることもある）を任意選択で含む。非永続的メモリ１０７は、典型的には、ＤＲＡＭ、ＳＲＡＭ、ＤＤＲ　ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリなどの高速ランダムアクセスメモリを含み、一方、永続的メモリ１０９は、典型的には、ＣＤ－ＲＯＭ、デジタル汎用ディスク（ＤＶＤ）、又は他の光学ストレージ、磁気カセット、磁気テープ、磁気ディスクストレージ、又は他の磁気記憶デバイス、磁気ディスク記憶デバイス、光学ディスク記憶デバイス、フラッシュメモリデバイス、又は他の不揮発性固体記憶デバイスを含む。永続的メモリ１０９は、任意選択で、ＣＰＵ１０２から遠隔に位置する１つ以上の記憶デバイスを含む。永続的メモリ１０９、及び非永続的メモリ１０９内の不揮発性メモリデバイスは、非一時的コンピュータ可読記憶媒体を含む。いくつかの実施形態において、非永続的メモリ１０７又は代替的に、非一時的コンピュータ可読記憶媒体は、以下のプログラム、モジュール及びデータ構造、又はそれらのサブセットを、場合によっては永続的メモリ１０９と併せて格納する：

●任意選択のオペレーティングシステム１５６（例えば、ＡＮＤＲＯＩＤ、ｉＯＳ、ＤＡＲＷＩＮ、ＲＴＸＣ、ＬＩＮＵＸ、ＵＮＩＸ、ＯＳ　Ｘ、ＷＩＮＤＯＷＳ、又はＶｘＷｏｒｋｓなどの組み込みオペレーティングシステム）であって、様々な基本システムサービスを処理するための、及びハードウェア依存タスクを実施するための手順を含むオペレーティングシステム；

●システム１００を他のデバイス及び／又は通信ネットワーク１０４と接続するための任意選択のネットワーク通信モジュール（又は命令）１５８；

●複数の化合物における化合物の各々についてのそれぞれの化学構造１２２（例えば、１２２－１、．．．１２２－Ｒ）又はその表現（例えば、化学構造のフィンガープリント）を含む化合物構造データストア１２０；

●細胞構成要素モジュール１３２のセット（例えば、１３２－１、．．．１３２－Ｋ）

を含む細胞構成要素モジュールデータストア１３０であり、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素１３４のサブセット（例えば、１３４－１－１、．．．１３４－１－Ｚ）を含む；

●摂動シグネチャ１４２のセット（例えば、１４２－１、．．．１４２－Ｐ）を含む摂動データストア１４０であり、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別を含み、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応するシグネチャスコア１４４（例えば、１４４－１－１、．．．１４４－１－Ｑ）；

●複数の化合物におけるそれぞれの化合物の各々について、それぞれの化学構造１５２（例えば、１５２－１、．．．１５２－Ｒ）の各々について、

○任意選択で、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの数値的活性化スコア１５４（例えば、１５４－１－１、．．．１５４－１－Ｋ）、及び／又は

○任意選択で、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々について、それぞれの数値的活性化スコア１５６（例えば、１５６－１－１、．．．１５６－１－Ｐ）を含む、活性化データ構造１５０、並びに

●複数のパラメータ（例えば、１００以上のパラメータ）を含むモデルであって、複数のパラメータは、計算された活性化スコアと、それぞれの化学構造についての数値的活性化スコアとの間の差に応答して調整される、モデル。

【０２０１】

様々な実施形態において、上述の識別されたエレメントのうちの１つ以上は、前述のメモリデバイスのうちの１つ以上に格納され、上述の機能を実施するための命令のセットに対応する。上記の識別されたモジュール、データ、又はプログラム（例えば、命令のセット）は、別個のソフトウェアプログラム、手順、データセット、又はモジュールとして実装される必要はなく、したがって、これらのモジュール及びデータの様々なサブセットは、様々な実装で組み合わされてもよく、又は別様に再配置されてもよい。いくつかの実施態様において、非永続的メモリ１０７は、任意選択で、上記で識別されたモジュール及びデータ構造のサブセットを格納する。更に、いくつかの実施形態において、メモリは、上述されていない追加のモジュール及びデータ構造を格納する。いくつかの実施形態において、上記の識別されたエレメントのうちの１つ以上は、システム１００のもの以外のコンピュータシステムに格納され、システム１００によってアドレス指定可能であり、したがってシステム１００は、必要なときにそのようなデータの全部又は一部を取り出すことができる。

【０２０２】

図１は、「システム１００」を描写するが、図は、本明細書に記載される実装の構造的な概略図ではなく、コンピュータシステムに存在し得る様々な特徴の機能的な説明としてより意図される。実際には、当業者によって認識されるように、別個に示される項目は組み合わされてもよく、いくつかの項目は分離されてもよい。更に、図１は、非永続的メモリ１０７内のある特定のデータ及びモジュールを描写するが、これらのデータ及びモジュールの一部又は全ては、代わりに、永続的メモリ１０９内又は２つ以上のメモリ内に格納されてもよい。例えば、いくつかの実施形態において、少なくとも化合物構造データストア１２０及び活性化データ構造１５０は、クラウドベースのインフラストラクチャの一部であり得るリモート記憶デバイスに格納される。いくつかの実施形態において、少なくとも化合物構造データストア１２０及び活性化データ構造１５０は、クラウドベースのインフラストラクチャ上に格納される。いくつかの実施形態において、化合物構造データストア１２０及び活性化データ構造１５０もまた、リモート記憶デバイスに格納され得る。

【０２０３】

本開示によるシステムは、図１を参照して開示されているが、本開示による方法２００

、３００、７００、８００、９００、及び１５００は、図２、図３、図７、図８、図９、及び図１４を参照してここで詳細に説明される。

【０２０４】

ＩＩ．試験化学化合物を目的の生理学的状態と関連付ける方法

生理学的状態。

図３Ａ～図３Ｅを参照すると、本開示の一態様は、試験化学化合物を目的の生理学的状態と関連付ける方法３００を提供する。

【０２０５】

いくつかの実施形態において、目的の生理学的状態は、疾患である。

【０２０６】

いくつかの実施形態において、疾患は、感染性又は寄生虫性疾患、腫瘍、血液又は造血器官の疾患、免疫系の疾患、内分泌疾患、栄養疾患又は代謝疾患、精神障害、行動障害又は神経発達障害、睡眠覚醒障害、神経系の疾患、視覚系の疾患、耳又は乳様突起の疾患、循環器系の疾患、呼吸器系の疾患、消化器系の疾患、皮膚の疾患、筋骨格系又は結合組織の疾患、泌尿生殖器系の疾患、性的健康に関連する状態、妊娠、出産又は産褥期に関連する疾患、周産期に起因する特定の状態、及び発達異常からなる群から選択される。いくつかの実施形態において、疾患は、ＩＣＤ－１１　ＭＭＳ、又は国際疾病分類の１つ以上の項目である。ＩＣＤは、疾患、負傷、及び死因を分類する方法を提供する。世界保健機関（ＷＨＯ）は、診断された疾患の事例を記録及び追跡する方法を標準化するためにＩＣＤを発行している。

【０２０７】

いくつかの実施形態において、目的の生理学的状態は、疾患の前提条件又は併存疾患などの疾患刺激性である。

【０２０８】

いくつかの実施形態において、目的の生理学的状態は、細胞系で発生するか、又は細胞系の文脈で測定される。いくつかの実施形態において、目的の生理学的状態は、１つ以上の細胞において生じるか、又は１つ以上の細胞の文脈において測定され、１つ以上の細胞は、単一細胞、細胞株、生検試料細胞、及び／又は培養された初代細胞を含む。いくつかの実施形態において、目的の生理学的状態は、ヒト細胞において生じる生理学的状態である。いくつかの実施形態において、目的の生理学的状態は、本明細書に記載される試料（例えば、定義：試料を参照されたい）のいずれかなどの試料において生じる生理学的状態である。いくつかの実施形態において、目的の生理学的状態は、ヒト又は動物などの対象において生じる生理学的状態である。

【０２０９】

いくつかの実施形態において、目的の生理学的状態は、目的の細胞プロセスであるか、又はそれに関連する。

【０２１０】

いくつかの実施形態において、目的の細胞プロセスは、異常な細胞プロセスである。いくつかの実施形態において、目的の細胞プロセスは、疾患と関連付けられた細胞プロセスである。例えば、上記のように、いくつかの実施形態において、この方法は、疾患に重要な細胞プロセス及びプログラムの標的化及び解明を提供する。いくつかの実施形態において、目的の細胞プロセスは、疾患の発症、進行、症状、重症度、及び／又は解消を含むが、これらに限定されない疾患の特徴のうちのいずれかの基礎となる機構を示すか、又はそれに関連する。いくつかの実施形態において、目的の細胞プロセスは、機能的経路である。いくつかの実施形態において、目的の細胞プロセスは、シグナル伝達経路である。いくつかの実施形態において、目的の細胞プロセスは、（例えば、化合物、小分子、及び／又は治療剤の）作用機序である。いくつかの実施形態において、目的の細胞プロセスは、転写ネットワーク（例えば、遺伝子調節ネットワーク）によって特徴付けられ、かつ／又は調節される。いくつかの実施形態において、目的の細胞プロセスは、第１の細胞状態と第２の細胞状態との間の遷移の間に生じる細胞プロセスである。

【０２１１】
いくつかの実施形態において、目的の細胞プロセスは、遺伝子セット濃縮アッセイ（ＧＳＥＡ）注釈、遺伝子オントロジー注釈、機能的及び／若しくはシグナル伝達経路注釈、並びに／又は細胞シグネチャ注釈などの注釈である。注釈は、ＮＩＨ Ｇｅｎｅ Ｅｘｐｒｅｓｓｉｏｎ Ｏｍｎｉｂｕｓ（ＧＥＯ）、ＥＢＩ ＡｒｒａｙＥｘｐｒｅｓｓ、ＮＣＢＩ、ＢＬＡＳＴ、ＥＭＢＬ－ＥＢＩ、ＧｅｎＢａｎｋ、Ｅｎｓｅｍｂｌ、ＫＥＧＧ経路データベース、Ｌｉｂｒａｒｙ ｏｆ Ｉｎｔｅｇｒａｔｅｄ Ｎｅｔｗｏｒｋ－ｂａｓｅｄ Ｃｅｌｌｕｌａｒ Ｓｉｇｎａｔｕｒｅｓ（ＬＩＮＣＳ）Ｌ１０００データセット、Ｒｅａｃｔｏｍｅ経路データベース、Ｇｅｎｅ Ｏｎｔｏｌｏｇｙプロジェクト、及び／又は任意の疾患特異的データベースを含むが、これらに限定されない、任意の公知のデータベースから得ることができる。

【０２１２】
したがって、いくつかの実施形態において、目的の生理学的状態は、本明細書に記載される任意のそれぞれの疾患、機能的経路、シグナル伝達経路、作用機序、転写ネットワーク、不一致、及び／又は細胞若しくは生物学的プロセスである。

【０２１３】
いくつかの実施形態において、目的の生理学的状態は表現型である。例えば、いくつかの実施形態において、目的の生理学的状態は、化合物、小分子、及び／又は治療剤、例えば、疾患の毒性及び／又は解消などの生理学的兆候である。いくつかの実施形態において、生理学的状態は、フローサイトメトリーの読み出し、イメージング及び顕微鏡注釈（例えば、Ｈ＆Ｅスライド、ＩＨＣスライド、放射線画像、及び／又は他の医学的イメージング）、並びに／又は細胞構成要素データを含むが、これらに限定されない実験データを使用して測定される表現型である。

【０２１４】
いくつかの実施形態において、目的の生理学的状態は、毒性の尺度である。いくつかの実施形態において、生理学的状態は、核受容体の阻害若しくは活性化、及び／又は核受容体の阻害の量若しくは活性化の量である。いくつかの実施形態において、生理学的状態は、阻害若しくは活性化、並びに／又は生物学的経路（例えば、ストレス応答経路）の阻害の量若しくは活性化の量である。本開示で使用され得る例示的な核受容体及び例示的なストレス応答経路、並びにこれらの核受容体及び例示的なストレス応答経路の阻害又は活性化データは、参照により本明細書に組み込まれるＨｕａｎｇ ｅｔ ａｌ．，２０１６，“Ｍｏｄｅｌｌｉｎｇ ｔｈｅ Ｔｏｘ２１ １０Ｋ ｃｈｅｍｉｃａｌ ｐｒｏｆｉｌｅｓ ｆｏｒ ｉｎ ｖｉｖｏ ｔｏｘｉｃｉｔｙ ｐｒｅｄｉｃｔｉｏｎ ａｎｄ ｍｅｃｈａｎｉｓｍ ｃｈａｒａｃｔｅｒｉｚａｔｉｏｎ，”Ｎａｔ Ｃｏｍｍｕｎ．７，ｐ．１０４２５に記載のように、およそ１０，０００個の化合物に対して説明されている。

【０２１５】
いくつかの実施形態において、目的の生理学的状態は、細胞構成要素のセット（例えば、細胞構成要素モジュール）の活性化及び／又は摂動シグネチャ（例えば、摂動に応答する複数の分析物の差次的発現プロファイル）を特徴とする。

【０２１６】
例えば、いくつかの実施形態において、目的の生理学的状態は、細胞構成要素のセットを含む細胞構成要素モジュールである。任意の種類の分析物（例えば、遺伝子、転写物、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせ）は、それぞれの細胞構成要素モジュールにおける細胞構成要素のセットで使用することが企図される。いくつかの実施形態において、細胞構成要素モジュールは、当業者に明白であろうように、当業者に既知の任意の細胞又は生物学的プロセス、並びにその任意の異常と関連付けられる。本明細書に開示されるシステム及び方法とともに使用するのに好適な細胞構成要素モジュールは、以下の「細胞構成要素及び細胞構成要素モジュール」と題されるセクションに更に記載される。

【０２１７】

いくつかの実施形態において、目的の生理学的状態は、第１の細胞状態と第２の細胞状態との間の不一致を特徴とする摂動シグネチャ（例えば、細胞状態遷移シグネチャ）である。

【０２１８】

いくつかのそのような実施形態において、目的の生理学的状態は、疾患状態（例えば、疾患対象及び／又は疾患組織から得られた細胞）と健康な状態（例えば、健康又は対照の対象及び／又は組織から得られた細胞）との間の不一致によって識別される。例えば、いくつかの実施形態において、疾患状態は、細胞の機能の喪失、細胞の機能の獲得、細胞の進行（例えば、細胞の分化状態への遷移）、細胞の静止（例えば、細胞が分化状態に遷移することができない）、細胞の侵入（例えば、異常な位置における細胞の出現）、細胞の消失（例えば、細胞が通常存在する位置における細胞の不在）、細胞の障害（例えば、細胞内及び／又は周囲の構造、形態、及び／又は空間的変化）、細胞のネットワークの損失（例えば、子孫細胞又は細胞の下流の細胞における正常な効果を排除する細胞の変化）、細胞のネットワークの獲得（例えば、細胞の下流の細胞の子孫細胞における新しい下流の効果を引き起こす細胞の変化）、細胞の余剰（例えば、細胞の過剰）、細胞の不足（例えば、臨界閾値を下回る細胞の密度）、細胞内の細胞構成要素比及び／若しくは量の差、細胞における遷移速度の差、又はこれらの任意の組み合わせによって識別される。

【０２１９】

本明細書に開示されるシステム及び方法とともに使用するのに好適な摂動シグネチャは、以下の「摂動シグネチャ」と題されるセクションに更に記載される。

【０２２０】

いくつかの実施形態において、目的の生理学的状態は、複数の生理学的状態（例えば、細胞プロセス、細胞構成要素モジュール、及び／又は摂動シグネチャ）を含む。いくつかの実施形態において、目的の生理学的状態は、少なくとも３、少なくとも４、少なくとも５、少なくとも６、少なくとも７、少なくとも８、少なくとも９、少なくとも１０、少なくとも１５、少なくとも２０、少なくとも３０、少なくとも４０、少なくとも５０、少なくとも６０、少なくとも７０、少なくとも８０、少なくとも９０、又は少なくとも１００の生理学的状態を含む。いくつかの実施形態において、目的の生理学的状態は、２００以下、１００以下、９０以下、８０以下、７０以下、６０以下、５０以下、２０以下、又は１０以下の生理学的状態を含む。いくつかの実施形態において、目的の生理学的状態は、１～５、５～１０、２～２０、１０～５０、又は２０～１００の生理学的状態を含む。いくつかの実施形態において、目的の生理学的状態は、３以上の生理学的状態から始まり、２００以下の生理学的状態で終わる別の範囲内にある複数の生理学的状態を含む。

【０２２１】

いくつかの実施形態において、本開示の化合物は、５つの基準のリピンスキーの法則を満たす化学化合物である。いくつかの実施形態において、本開示の化合物は、５つのリピンスキーの法則のうちの２つ以上の法則、３つ以上の法則、又は４つ全ての法則を満たす有機化合物である。（ｉ）５つ以下の水素結合ドナー（例えば、ＯＨ及びＮＨ基）、（ｉｉ）１０個以下の水素結合アクセプター（例えば、Ｎ及びＯ）、（ｉｉｉ）５００ダルトン未満の分子量、及び（ｉｖ）５未満のＬｏｇＰ。４つの基準のうち３つが５という数字を含むため、「５つの法則」と呼ばれる。Ｌｉｐｉｎｓｋｉ，１９９７，Ａｄｖ．Ｄｒｕｇ　Ｄｅｌ．Ｒｅｖ．２３，３を参照されたく、これは参照によりその全体が本明細書に組み込まれる。いくつかの実施形態において、本開示の化合物は、５つのリピンスキーの法則に加えて、１つ以上の基準を満たす。例えば、いくつかの実施形態において、本開示の化合物は、５個以下の芳香族環、４個以下の芳香族環、３個以下の芳香族環、又は２個以下の芳香族環を有する。

【０２２２】

ブロック３０２を参照すると、方法３００は、試験化学化合物の化学構造のフィンガープリントを得ることを含む。

【０２２３】

例えば、いくつかの実施態様では、試験化学化合物を機械学習アプローチに適用することは、分子データ（例えば、化合物の化学構造）を機械学習モデルによって読み取り及び操作可能な形式に変換することを含む。

【０２２４】

図３Ａのブロック３０４を参照すると、化学構造を機械学習読み取り可能なフォーマットに変換するための１つのアプローチは、テキストの文字列として分子を表す単純化された分子入力ラインエントリーシステム（ＳＭＩＬＥＳ）を使用して化学構造の「フィンガープリント」を決定することを含む。したがって、いくつかの実施形態において、この方法は、試験化学化合物の単純化された分子入力ラインエントリーシステム（ＳＭＩＬＥＳ）文字列表現からフィンガープリントを計算することを更に含む。ＳＭＩＬＥＳ文字列を使用した分子フィンガープリンティングは、例えば、Ｈｏｎｄａ ｅｔ ａｌ．，２０１９，“ＳＭＩＬＥＳ Ｔｒａｎｓｆｏｒｍｅｒ：Ｐｒｅ－ｔｒａｉｎｅｄ Ｍｏｌｅｃｕｌａｒ Ｆｉｎｇｅｒｐｒｉｎｔ ｆｏｒ Ｌｏｗ Ｄａｔａ Ｄｒｕｇ Ｄｉｓｃｏｖｅｒｙ，”ａｒＸｉｖ：１９１１．０４７３８に更に記載されており、これは参照によりその全体が本明細書に組み込まれる。

【０２２５】

化学構造を機械学習読み取り可能フォーマットに変換するための別のアプローチは、グラフベースの分子フィンガープリントを決定することを含む。グラフベースの分子フィンガープリンティングでは、元の分子構造は、ノードが個々の原子を表し、エッジが原子間の結合を表すグラフによって表される。グラフベースのアプローチは、より低いサイズ要件を有する複数のサブ構造を効率的に符号化する能力、したがってより低い計算負荷、並びにフィンガープリント間の構造的類似性の表示を符号化する能力を含むいくつかの利点を提供する。グラフベースのフィンガープリンティングは、例えば、Ｄｕｖｅｎａｕｄ ｅｔ ａｌ．，２０１５，“Ｃｏｎｖｏｌｕｔｉｏｎａｌ ｎｅｔｗｏｒｋｓ ｏｎ ｇｒａｐｈｓ ｆｏｒ ｌｅａｒｎｉｎｇ ｍｏｌｅｃｕｌａｒ ｆｉｎｇｅｒｐｒｉｎｔｓ，”ＮｅｕｒＩＰＳ，２２２４－２２３２に更に記載され、これは参照によりその全体が本明細書に組み込まれる。いくつかの実施形態において、フィンガープリントはグラフ畳み込みネットワークから生成される。いくつかの実施形態において、フィンガープリントは、グラフアテンションネットワーク（ＧＡＴ）、グラフ同形ネットワーク（ＧＩＮ）、又はグラフ下部構造インデックスベース近似グラフ（ＳＡＧＡ）などの空間的グラフ畳み込みネットワークから生成される。いくつかの実施形態において、フィンガープリントは、チェビシェフ（Ｃｈｅｂｙｓｈｅｖ）多項式フィルタリングを使用するスペクトルグラフ畳み込みなどのスペクトルグラフ畳み込みネットワークから生成される。

【０２２６】

図３Ａのブロック３０６を参照すると、いくつかの実施形態において、フィンガープリントは、ＳＭＩＬＥＳ Ｔｒａｎｓｆｏｒｍｅｒ、ＥＣＦＰ４、ＲＮＮＳ２Ｓ、及び／又はＧｒａｐｈＣｏｎｖを使用して化学構造から生成される。

【０２２７】

モデルアーキテクチャ。

図３Ｂのブロック３０８を参照すると、方法は、フィンガープリントをモデルに入力することを含む。いくつかの実施形態において、モデルは、複数（例えば、１００、２００、３００、５００、１０００、１０，０００又はそれ以上）のパラメータを含む。

【０２２８】

いくつかの実施形態において、モデルは、複数のパラメータ（例えば、重み及び／又はハイパーパラメータ）を含む。いくつかの実施形態において、モデルについての複数のパラメータは、少なくとも１０、少なくとも５０、少なくとも１００、少なくとも５００、少なくとも１０００、少なくとも２０００、少なくとも５０００、少なくとも１０，０００、少なくとも２０，０００、少なくとも５０，０００、少なくとも１００，０００、少なくとも２００，０００、少なくとも５００，０００、少なくとも１００万、少なくとも２００万、少なくとも３００万、少なくとも４００万、又は少なくとも５００万のパラメ

ータを含む。いくつかの実施形態において、モデルについての複数のパラメータは、８００万以下、５００万以下、４００万以下、１００万以下、５００，０００以下、１００，０００以下、５０，０００以下、１０，０００以下、５０００以下、１０００以下、又は５００以下のパラメータを含む。いくつかの実施形態において、モデルについての複数のパラメータは、１０～５０００、５００～１０，０００、１０，０００～５００，０００、２０，０００～１００万、又は１００万～５００万のパラメータを含む。いくつかの実施形態において、モデルについての複数のパラメータは、１０以上のパラメータから始まり、８００万以下のパラメータで終わる別の範囲内にある。

【０２２９】

いくつかの実施形態において、モデルの訓練は、１つ以上のハイパーパラメータ（例えば、訓練中に合わせられ得る１つ以上の値）によって更に特徴付けられる。いくつかの実施形態において、ハイパーパラメータ値は、訓練中に合わせられる（例えば、調整される）。いくつかの実施形態において、ハイパーパラメータ値は、訓練データセット及び／又は１つ以上の入力（例えば、細胞、細胞構成要素モジュール、共変量など）の特定のエレメントに基づいて決定される。いくつかの実施形態において、ハイパーパラメータ値は、実験的最適化を使用して決定される。いくつかの実施形態において、ハイパーパラメータ値は、ハイパーパラメータスイープを使用して決定される。いくつかの実施形態において、ハイパーパラメータ値は、以前のテンプレート又はデフォルト値に基づいて割り当てられる。

【０２３０】

いくつかの実施形態において、１つ以上のハイパーパラメータのそれぞれのハイパーパラメータは、学習速度を含む。いくつかの実施形態において、学習速度は、少なくとも０．０００１、少なくとも０．０００５、少なくとも０．００１、少なくとも０．００５、少なくとも０．０１、少なくとも０．０５、少なくとも０．１、少なくとも０．２、少なくとも０．３、少なくとも０．４、少なくとも０．５、少なくとも０．６、少なくとも０．７、少なくとも０．８、少なくとも０．９、又は少なくとも１である。いくつかの実施形態において、学習速度は、１以下、０．９以下、０．８以下、０．７以下、０．６以下、０．５以下、０．４以下、０．３以下、０．２以下、０．１以下　０．０５以下、０．０１以下、又はそれ未満を含む。いくつかの実施形態において、学習速度は、０．０００１～０．０１、０．００１～０．５、０．００１～０．０１、０．００５～０．８、又は０．００５～１である。いくつかの実施形態において、学習速度は、０．０００１以上から始まり、１以下で終わる別の範囲内に入る。いくつかの実施形態において、１つ以上のハイパーパラメータは、正規化強度（例えば、Ｌ２重みペナルティ、中断率など）を更に含む。例えば、いくつかの実施形態において、モデル（例えば、ニューラルネットワーク）は、複数の隠れニューロンにおける隠れニューロンの各々の対応するパラメータ（例えば、重み）に対する正規化を使用して訓練される。いくつかの実施形態において、正規化は、Ｌ１又はＬ２ペナルティを含む。

【０２３１】

いくつかの実施形態において、１つ以上のハイパーパラメータのそれぞれのハイパーパラメータは、損失関数である。いくつかの実施形態において、損失関数は、平均平方誤差、平滑化平均平方誤差、二次損失、平均絶対誤差、平均バイアス誤差、ヒンジ、マルチクラスサポートベクトルマシン、及び／又は交差エントロピーである。いくつかの実施形態において、損失関数は、勾配降下アルゴリズム及び／又は最小化関数である。

【０２３２】

いくつかの実施形態において、モデルは、１つ以上の活性化関数と関連付けられる。いくつかの実施形態において、１つ以上の活性化関数における活性化関数は、ｔａｎｈ、シグモイド、ｓｏｆｔｍａｘ、ガウシアン、ボルツマン（Ｂｏｌｔｚｍａｎｎ）加重平均化、絶対値、線形、整流化線形ユニット（ＲｅＬＵ）、有界整流化線形、ソフト整流化線形、パラメータ化整流化線形、平均、最大、最初、サイン、平方、平方根、多重二乗、逆二乗、逆多重二乗、多調和スプライン、スウィッシュ（ｓｗｉｓｈ）、ミッシュ（ｍｉｓｈ

）、ガウシアン誤差線形ユニット（ＧｅＬＵ）、及び／又は薄板スプラインである。モデルは、フィンガープリントのモデルへの入力に応答して１つ以上の計算された活性化スコアを出力する。

【０２３３】

図３Ｂのブロック３１０を参照すると、いくつかの実施形態において、モデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。いくつかの実施形態において、モデルは、リグレッサーである。いくつかの実施形態において、モデルは、本明細書に開示されるモデルのいずれかである（例えば、定義：モデルを参照されたい）。

【０２３４】

図３Ｂのブロック３１２を参照すると、いくつかの実施形態において、モデルはニューラルネットワークを含む。

【０２３５】

いくつかの実施形態において、ニューラルネットワークは、ＲｅＬＵ活性化を有する完全に接続されたニューラルネットワークである。例えば、いくつかの実施形態において、モデルは、対応する１つ以上の入力を含むニューラルネットワークであり、対応する１つ以上の入力における各入力は、試験化学化合物についての化学構造のためのものであり、対応する複数の隠れニューロンを含む対応する第１の隠れ層であり、対応する複数の隠れニューロンにおける隠れニューロンの各々は、（ｉ）複数の入力における入力の各々に完全に接続され、（ｉｉ）第１の活性化関数タイプと関連付けられ、（ｉｉｉ）ニューラルネットワークについての複数のパラメータにおける対応するパラメータ（例えば、重み）、及び１つ以上の対応するニューラルネットワーク出力と関連付けられ、対応する１つ以上のニューラルネットワーク出力におけるそれぞれのニューラルネットワーク出力の各々は、（ｉ）入力として、対応する複数の隠れニューロンにおける隠れニューロンの各々の出力を直接的又は間接的に受信し、かつ（ｉｉ）第２の活性化関数タイプと関連付けられる。いくつかのそのような実施形態において、ニューラルネットワークは、完全に接続されたネットワークである。

【０２３６】

いくつかの実施形態において、ニューラルネットワークは、複数の隠れ層を含む。上述したように、隠れ層は、（例えば、追加の複雑性を捕捉するために）入力層と出力層との間に位置する。複数の隠れ層が存在するいくつかの実施形態において、隠れ層の各々は、同じ又は異なるそれぞれの数のニューロンを有し得る。

【０２３７】

いくつかの実施形態において、隠れニューロンの各々（例えば、ニューラルネットワークにおけるそれぞれの隠れ層における）は、入力データに対して関数（例えば、線形関数又は非線形関数）を実施する活性化関数と関連付けられる。一般に、活性化関数の目的は、ニューラルネットワークが元のデータの表現について訓練され、その後、新しい（例えば、以前には見えなかった）データの追加の表現を「適合」又は生成することができるように、データに非線形性を導入することである。特定の活性化関数は、データセットの極端な端部（例えば、ｔａｎｈ及び／又はシグモイド関数）で飽和をもたらす可能性があるため、活性化関数（例えば、第１及び／又は第２の活性化関数）の選択は、ニューラルネットワークの使用例に依存する。例えば、いくつかの実施形態において、活性化関数（例えば、第１及び／又は第２の活性化関数）は、本明細書に開示される任意の活性化関数を含むが、これらに限定されない、当該技術分野で既知の任意の好適な活性化関数から選択される。

【０２３８】

いくつかの実施形態において、隠れニューロンの各々は、活性化関数に基づいて決定されたニューラルネットワークの出力に寄与するパラメータ（例えば、重み及び／又はバイ

アス値）と更に関連付けられる。いくつかの実施形態において、隠れニューロンは、任意のパラメータ（例えば、ランダム化された重み）によって初期化される。いくつかの代替的な実施形態において、隠れニューロンは、所定のパラメータのセットによって初期化される。

【０２３９】

いくつかの実施形態において、ニューラルネットワークにおける（例えば、１つ以上の隠れ層にわたる）複数の隠れニューロンは、少なくとも２個、少なくとも３個、少なくとも４個、少なくとも５個、少なくとも６個、少なくとも７個、少なくとも８個、少なくとも９個、少なくとも１０個、少なくとも１１個、少なくとも１２個、少なくとも１３個、少なくとも１４個、少なくとも１５個、少なくとも１６個、少なくとも１７個、少なくとも１８個、少なくとも１９個、少なくとも２０個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、又は少なくとも５００個のニューロンである。いくつかの実施形態において、複数の隠れニューロンは、少なくとも１００個、少なくとも５００個、少なくとも８００個、少なくとも１０００個、少なくとも２０００個、少なくとも３０００個、少なくとも４０００個、少なくとも５０００個、少なくとも６０００個、少なくとも７０００個、少なくとも８０００個、少なくとも９０００個、少なくとも１０，０００個、少なくとも１５，０００個、少なくとも２０，０００個、又は少なくとも３０，０００個のニューロンである。いくつかの実施形態において、複数の隠れニューロンは、３０，０００個以下、２０，０００個以下、１５，０００個以下、１０，０００個以下、９０００個以下、８０００個以下、７０００個以下、６０００個以下、５０００個以下、４０００個以下、３０００個以下、２０００個以下、１０００個以下、９００個以下、８００個以下、７００個以下、６００個以下、５００個以下、４００個以下、３００個以下、２００個以下、１００個以下、又は５０個以下のニューロンである。いくつかの実施形態において、複数の隠れニューロンは、２～２０個、２～２００個、２～１０００個、１０～５０個、１０～２００個、２０～５００個、１００～８００個、５０～１０００個、５００～２０００個、１０００～５０００個、５０００～１０，０００個、１０，０００～１５，０００個、１５，０００～２０，０００個、又は２０，０００～３０，０００個のニューロンである。いくつかの実施形態において、複数の隠れニューロンは、２個以上のニューロンから始まり、３０，０００個以下のニューロンで終わる別の範囲内にある。

【０２４０】

いくつかの実施形態において、ニューラルネットワークは、１～５０個の隠れ層を含む。いくつかの実施形態において、ニューラルネットワークは、１～２０個の隠れ層を含む。いくつかの実施形態において、ニューラルネットワークは、少なくとも２個、少なくとも３個、少なくとも４個、少なくとも５個、少なくとも６個、少なくとも７個、少なくとも８個、少なくとも９個、少なくとも１０個、少なくとも１１個、少なくとも１２個、少なくとも１３個、少なくとも１４個、少なくとも１５個、少なくとも１６個、少なくとも１７個、少なくとも１８個、少なくとも１９個、少なくとも２０個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、又は少なくとも１００個の隠れ層を含む。いくつかの実施形態において、ニューラルネットワークは、１００個以下、９０個以下、８０個以下、７０個以下、６０個以下、５０個以下、４０個以下、３０個以下、２０個以下、１０個以下、９個以下、８個以下、７個以下、６個以下、又は５個以下の隠れ層を含む。いくつかの実施形態において、ニューラルネットワークは、１～５個、１～１０個、１～２０個、１０～５０個、２～８０個、５～１００個、１０～１００個、５０～１００個、又は３～３０個の隠れ層を含む。いくつかの実施形態において、ニューラルネットワークは、１個以上の層から始まり、１００個以下の層で終わる別の範囲内にある複数の隠れ層を含む。

【０２４１】

いくつかの実施形態において、ニューラルネットワークは、浅いニューラルネットワー

クを含む。浅いニューラルネットワークは、少数の隠れ層を有するニューラルネットワークを指す。いくつかの実施形態において、そのようなニューラルネットワークアーキテクチャは、ニューラルネットワーク訓練の効率を改善し、訓練に関与する層の数の減少に起因して計算能力を節約する。いくつかの実施形態において、ニューラルネットワークは、１つの隠れ層を含む。いくつかの実施形態において、ニューラルネットワークは、２つ、３つ、４つ、又は５つの隠れ層を含む。

【０２４２】

いくつかの実施形態において、ニューラルネットワークは、メッセージパッシングニューラルネットワークである。メッセージパッシングニューラルネットワークは、グラフ（例えば、化学構造のグラフベースの表現）上の教師あり学習のためのフレームワークを指し、ノードは原子を表し、エッジは原子間の結合を表す。一般に、メッセージパッシングニューラルネットワークは、フォワードパスにおける２つのフェーズ、メッセージパッシングフェーズ及び読み出しフェーズを含む。メッセージパッシングフェーズは、Ｔ間隔の期間にわたって実行され、メッセージ関数M_t及び頂点更新関数U_tに従って、グラフ内の各ノードで隠された状態を更新することを含む。読み出しフェーズは、読み出し関数Ｒを使用してグラフについての特徴ベクトルを計算する。いくつかの実施形態において、メッセージパッシングニューラルネットワークは、畳み込みネットワーク（例えば、空間的グラフ畳み込みネットワーク及び／又はスペクトルグラフ畳み込みネットワーク）、ゲート付きグラフニューラルネットワーク（ＧＧ－ＮＮ）、相互作用ネットワーク、分子グラフ畳み込み、ディープテンソルニューラルネットワーク、及び／又はラプラシアンベースの方法を含む。例えば、Ｇｉｌｍｅｒ ｅｔ ａｌ．，２０１７，“Ｎｅｕｒａｌ Ｍｅｓｓａｇｅ Ｐａｓｓｉｎｇ ｆｏｒ Ｑｕａｎｔｕｍ Ｃｈｅｍｉｓｔｒｙ，”ａｒＸｉｖ：１７０４．０１２１２ｖ２を参照されたく、これは参照によりその全体が本明細書に組み込まれる。

【０２４３】

図３Ｂのブロック３１４を参照すると、いくつかの実施形態において、モデルは、複数のコンポーネントモデルのアンサンブルモデルである。例えば、ブロック３１６を参照すると、いくつかの実施形態において、１つ以上の計算された活性化スコアにおける計算された活性化スコアの各々は、複数のコンポーネントモデルにおけるコンポーネントモデルの各々の出力の中心傾向の測定値である。

【０２４４】

図３Ｂのブロック３１８を参照すると、いくつかの実施形態において、複数のコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、線形回帰モデル、又は複数のニューラルネットワークを含む。

【０２４５】

いくつかの実施形態において、アンサンブルモデルは、少なくとも２個、少なくとも３個、少なくとも４個、少なくとも５個、少なくとも６個、少なくとも７個、少なくとも８個、少なくとも９個、少なくとも１０個、少なくとも２０個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、又は少なくとも５００個のコンポーネントモデルを含む。いくつかの実施形態において、アンサンブルモデルは、５００個以下、４００個以下、３００個以下、２００個以下、又は１００個以下のコンポーネントモデルを含む。いくつかの実施形態において、アンサンブルモデルは、１００個以下、５０個以下、４０個以下、３０個以下、又は２０個以下のコンポーネントモデルを含む。いくつかの実施形態において、アンサンブルモデルは、１～５０個、２～２０個、５～５０個、１０～８０個、５～１５個、３～３０個、１０～５００個、２～１００個、又は５０～１００個のコンポーネントモデルを含む。いくつかの実施形態において、アンサンブルモデルは、２個以上のコンポー

ネントモデルから始まり、５００個以下のコンポーネントモデルで終わるコンポーネントモデルの別の範囲を含む。

【０２４６】

いくつかの実施形態において、アンサンブルモデルは、複数のコンポーネントモデルから得られた複数の出力（例えば、活性化スコア）を組み合わせることによって形成される。いくつかの実施形態において、分類子からの複数の出力（例えば、活性化スコア）は、平均、中央値、モード、加重平均、加重中央値、加重モード、算術平均、ミッドレンジ、ミッドヒンジ、トリミアン、及び／又はウィンザライズド平均を含むが、これらに限定されない、当該技術分野で既知の中心傾向の任意の尺度を使用して組み合わされる。例えば、アンサンブルモデルからの最終決定は、アンサンブルモデル内の全てのコンポーネントモデルにわたる出力の平均に基づいて得ることができる。

【０２４７】

いくつかの実施形態において、複数の出力は、投票方法を使用して組み合わされる。例えば、いくつかの実施形態において、複数の出力は、それぞれの化学構造と、目的のそれぞれの生理学的状態との間の関連性を示す、アンサンブルモデル内のコンポーネントモデルの各々からの出力の数（例えば、活性化スコア）を集計することによって組み合わされる。いくつかの実施形態において、コンポーネントモデルからの複数の出力（例えば、活性化スコア）は、多数決を使用して組み合わされる。いくつかのそのような実施形態において、関連性を示す出力の集計（例えば、閾値基準を超える活性化スコアの集計）が投票閾値よりも大きい場合、それぞれの化学構造とそれぞれの目的の生理学的状態との間の関連性を決定することによって、コンポーネントモデルからの複数の出力が組み合わされる。いくつかの実施形態において、投票閾値は、アンサンブルモデル内の複数のコンポーネントモデルからの総投票の少なくとも５０％である。いくつかの実施形態において、投票閾値は、アンサンブルモデル内の複数のコンポーネントモデルからの総投票の少なくとも２０％、少なくとも３０％、少なくとも４０％、少なくとも５０％、少なくとも６０％、少なくとも７０％、少なくとも８０％、少なくとも９０％、又は少なくとも９５％である。

【０２４８】

いくつかの実施形態において、アンサンブルモデル内のコンポーネントモデルの各々は、重み付けされない（例えば、各コンポーネントモデルは、アンサンブルモデル内で１票を有する）。いくつかの実施形態において、アンサンブルモデル内の１つ以上のコンポーネントモデルは、更に重み付けされる（例えば、アンサンブルモデル内で１票を超える投票を有する）。

【０２４９】

いくつかの実施形態において、方法は、単一のアンサンブルモデル又は複数のアンサンブルモデルを得ることを含む。当該技術分野で既知の任意のアーキテクチャが、アンサンブルモデルについて企図される。例えば、いくつかの実施形態において、複数のコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、及び／又は線形回帰モデルを含む。いくつかの実施形態において、複数のコンポーネントモデルは、複数のニューラルネットワークを含む。

【０２５０】

図３Ｂのブロック３２０を参照すると、いくつかの実施形態において、モデルは、複数のニューラルネットワークのアンサンブルモデルである。図３Ｂのブロック３２２を参照すると、いくつかの実施形態において、モデルは、複数のニューラルネットワークを含むアンサンブルモデルであり、複数のニューラルネットワークにおける第１のニューラルネットワークは、ＲｅＬＵ活性化を伴う完全に接続されたニューラルネットワークであり、複数のニューラルネットワークにおける第２のニューラルネットワークは、メッセージパッシングニューラルネットワークである。いくつかのそのような実施形態において、第１

のニューラルネットワークは、入力として、化学構造についての分子フィンガープリントをＳＭＩＬＥＳ表現として受け入れる完全に接続された３層ニューラルネットワークである。いくつかの実施形態において、第２のニューラルネットワークは、入力として、化学構造についての分子フィンガープリントをグラフベースの表現として受け入れるメッセージパッシングニューラルネットワーク（ＭＰＮＮ）である。

【０２５１】

細胞構成要素及び細胞構成要素モジュール。

上述のように、再びブロック３０８を参照すると、モデルへの化学構造についてのフィンガープリントの入力に応答して、モデルは、細胞構成要素モジュールのセットについて１つ以上の計算された活性化スコアを出力する。図３Ｃのブロック３２６を参照すると、１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、細胞構成要素モジュールのセットにおける対応する細胞構成要素モジュールを表す。

【０２５２】

図３Ｃのブロック３２８を参照すると、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々は、複数の細胞構成要素の独立したサブセットを含む。

【０２５３】

いくつかの実施形態において、細胞構成要素は、遺伝子、遺伝子産物（例えば、ｍＲＮＡ及び／又はタンパク質）、炭水化物、脂質、エピジェネティック特徴、代謝産物、及び／又はそれらの組み合わせである。いくつかの実施形態において、複数の細胞構成要素における細胞構成要素の各々は、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせである。いくつかの実施形態において、複数の細胞構成要素は、ＤＮＡ、修飾（例えば、メチル化）ＤＮＡ、コード（例えば、ｍＲＮＡ）又は非コードＲＮＡ（例えば、ｓｎｃＲＮＡ）を含むＲＮＡを含む核酸、転写後修飾タンパク質（例えば、リン酸化、グリコシル化、ミリスチル化などのタンパク質）を含むタンパク質、脂質、炭水化物、環状アデノシン一リン酸（ｃＡＭＰ）及び環状グアノシン一リン酸（ｃＧＭＰ）などの環状ヌクレオチドを含む、ヌクレオチド（例えば、アデノシン三リン酸（ＡＴＰ）、アデノシン二リン酸（ＡＤＰ）及びアデノシン一リン酸（ＡＭＰ））、酸化及び還元形態のニコチンアミドアデニンジヌクレオチド（ＮＡＤＰ／ＮＡＤＰＨ）などの他の小分子細胞構成要素、並びにそれらの任意の組み合わせを含む。

【０２５４】

いくつかの実施形態において、複数の細胞構成要素は、少なくとも５個、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも２５個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも６００個、少なくとも７００個、少なくとも８００個、少なくとも９００個、少なくとも１０００個、少なくとも２０００個、少なくとも３０００個、少なくとも４０００個、少なくとも５０００個、少なくとも６０００個、少なくとも７０００個、少なくとも８０００個、少なくとも９０００個、少なくとも１０，０００個、少なくとも２０，０００個、少なくとも３０，０００個、少なくとも５０，０００個、又は５０，０００個を超える細胞構成要素を含む。いくつかの実施形態において、複数の細胞構成要素は、７０，０００個以下、５０，０００個以下、３０，０００個以下、１０，０００個以下、５０００個以下、１０００個以下、５００個以下、２００個以下、１００個以下、９０個以下、８０個以下、７０個以下、６０個以下、５０個以下、又は４０個以下の細胞構成要素を含む。いくつかの実施形態において、複数の細胞構成要素は、２０～１０，０００個の細胞構成要素からなる。いくつかの実施形態において、複数の細胞構成要素は、１００～８，０００個の細胞構成要素からなる。いくつかの実施形態において、複数の細胞構成要素は、５～２０個、２０～５０個、５０～１００個、１００～２００個、２００～５００個、５００～１０００個、１０００～

５０００個、５０００～１０，０００個、又は１０，０００～５０，０００個の細胞構成要素を含む。いくつかの実施形態において、複数の細胞構成要素は、５個以上の細胞構成要素から始まり、７０，０００個以下の細胞構成要素で終わる別の範囲内にある。

【０２５５】

一例として、いくつかの実施形態において、複数の細胞構成要素は、ＲＮＡレベルで、任意選択で測定された複数の遺伝子を含む。いくつかの実施形態において、複数の遺伝子は、少なくとも５個、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも２５個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも６００個、少なくとも７００個、少なくとも８００個、少なくとも９００個、又は少なくとも１０００個の遺伝子を含む。いくつかの実施形態において、複数の遺伝子は、少なくとも１０００個、少なくとも２０００個、少なくとも３０００個、少なくとも４０００個、少なくとも５０００個、少なくとも１０，０００個、少なくとも３０，０００個、少なくとも５０，０００個、又は５０，０００個を超える遺伝子を含む。いくつかの実施形態において、複数の遺伝子は、５～２０個、２０～５０個、５０～１００個、１００～２００個、２００～５００個、５００～１０００個、１０００～５０００個、５０００～１０，０００個、又は１０，０００～５０，０００個の遺伝子を含む。

【０２５６】

別の例として、いくつかの実施形態において、複数の細胞構成要素は、複数のタンパク質を含む。いくつかの実施形態において、複数のタンパク質は、少なくとも５個、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも２５個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも６００個、少なくとも７００個、少なくとも８００個、少なくとも９００個、又は少なくとも１０００個のタンパク質を含む。いくつかの実施形態において、複数のタンパク質は、少なくとも１０００個、少なくとも２０００個、少なくとも３０００個、少なくとも４０００個、少なくとも５０００個、少なくとも１０，０００個、少なくとも３０，０００個、少なくとも５０，０００個、又は５０，０００個を超えるタンパク質を含む。いくつかの実施形態において、複数のタンパク質は、５～２０個、２０～５０個、５０～１００個、１００～２００個、２００～５００個、５００～１０００個、１０００～５０００個、５０００～１０，０００個、又は１０，０００～５０，０００個のタンパク質を含む。

【０２５７】

細胞構成要素モジュールにおける細胞構成要素の各々が一意であるという要件はない。例えば、細胞構成要素モジュールＡが、細胞構成要素１、３及び１０を含有する場合を考慮する。細胞構成要素モジュールのセットにおける他の細胞構成要素モジュールは、これらの細胞構成要素も含有し得る。ここで、「独立した」という用語は、特定の細胞構成要素モジュールにおける複数の細胞構成要素のサブセットが全体として一意であることを意味する。したがって、上記の例示的な細胞構成要素モジュールＡを考慮すると、細胞構成要素モジュールのセットにおける別の細胞構成要素モジュールは、細胞構成要素モジュールＡが含有しない他の細胞構成要素を更に含有することを条件として、細胞構成要素１、３及び１０を含有し得る。上記の例示的な細胞構成要素モジュールＡを更に考慮すると、細胞構成要素モジュールのセットにおける別の細胞構成要素モジュールは、細胞構成要素モジュールＡが含有しない他の細胞構成要素を更に含有するという要件なしに提供される細胞構成要素１、３及び１０のサブセットに限定され得る（しかしながら、そのような追加の細胞構成要素も有し得る）。

【０２５８】

いくつかの実施形態において、細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々は、複数の細胞構成要素のそれぞれの独立したサブセットにおける同じ

又は異なる数の細胞構成要素を含む。いくつかの実施形態において、それぞれの細胞構成要素モジュールの各々に対応する細胞構成要素のそれぞれの独立したサブセットの各々は、細胞構成要素の固有のサブセットである（例えば、非重複であり、複数の細胞構成要素における細胞構成要素の各々は、１つ以下のモジュールにグループ化される）。いくつかの実施形態において、第１の細胞構成要素モジュールは、第２の細胞構成要素モジュールに対応する細胞構成要素の第２のサブセットと重複する細胞構成要素の第１のサブセットを有する（例えば、重複しており、複数の細胞構成要素における少なくとも１つの細胞構成要素は、２つ以上の異なるモジュールに共通している）。

【０２５９】

図３Ｃのブロック３３０を参照すると、いくつかの実施形態において、それぞれの細胞構成要素モジュールにおける複数の細胞構成要素の独立したサブセットは、５つ以上の細胞構成要素を含む。いくつかの実施形態において、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールにおける複数の細胞構成要素の独立したサブセットは、少なくとも２個、少なくとも５個、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも２５個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも６００個、少なくとも７００個、少なくとも８００個、少なくとも９００個、少なくとも１０００個、少なくとも２０００個、又は少なくとも３０００個の細胞構成要素を含む。いくつかの実施形態において、複数の細胞構成要素の独立したサブセットは、５０００個以下、３０００個以下、１０００個以下、５００個以下、２００個以下、１００個以下、９０個以下、８０個以下、７０個以下、６０個以下、又は５０個以下の細胞構成要素を含む。いくつかの実施形態において、複数の細胞構成要素の独立したサブセットは、５～１００個、２～３００個、２０～５００個、２００～１０００個、又は１０００～５０００個の細胞構成要素を含む。いくつかの実施形態において、複数の細胞構成要素の独立したサブセットは、２個以上の細胞構成要素から始まり、５０００個以下の細胞構成要素で終わる別の範囲内にある。

【０２６０】

いくつかの実施形態において、それぞれの細胞構成要素モジュールにおける複数の細胞構成要素の独立したサブセットは、目的の生理学的状態と関連付けられた細胞プロセス（例えば、分子経路）における細胞構成要素からなる。例えば、図３Ｃのブロック３３２を参照すると、いくつかの実施形態において、それぞれの細胞構成要素モジュールにおける複数の細胞構成要素の独立したサブセットは、目的の生理学的状態と関連付けられた分子経路における２～２０個の細胞構成要素からなる。

【０２６１】

図３Ｄのブロック３３４を参照すると、細胞構成要素モジュールのセットにおける少なくとも第１の細胞構成要素モジュールは、目的の生理学的状態と関連付けられる。実際に、多数の細胞構成要素モジュールは、目的の生理学的状態と関連付けられ得る。

【０２６２】

図３Ｄのブロック３３６を参照すると、１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、細胞構成要素モジュールのセットにおける対応する細胞構成要素モジュールを表す。

【０２６３】

図３Ｄのブロック３３８を参照すると、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々は、複数の細胞構成要素の独立したサブセットを含む。

【０２６４】

図３Ｄのブロック３４０を参照すると、いくつかの実施形態において、細胞構成要素モジュールのセットは、複数の細胞構成要素モジュールである。第１の細胞構成要素モジュールを含む複数の細胞構成要素モジュールの第１のサブセットは、目的の生理学的状態と

関連付けられる。すなわち、このような細胞構成要素モジュールは、目的の生理学的状態に関与する細胞構成要素を表す。例えば、このような細胞構成要素モジュールのこのような細胞構成要素は、いくつかのベースライン、野生型状態の細胞と比較して、目的の生理学的状態を表す細胞において下方制御又は上方制御され得る。更に、複数の細胞構成要素モジュールの第２のサブセットは、目的の生理学的状態と関連付けられていない。すなわち、このような細胞構成要素モジュールの細胞構成要素は、目的の生理学的状態に関与していない細胞構成要素を表す。例えば、このような細胞構成要素は、いくつかのベースライン、野生型状態の細胞と比較して、目的の生理学的状態を表す細胞において下方制御又は上方制御されない。このような実施形態において、（細胞構成要素モジュールの第１のサブセット内にある）第１の細胞構成要素モジュールについてのそれぞれの計算された活性化スコアが第１の閾値基準を満たし、複数の細胞構成要素モジュールの第２のサブセットにおける細胞構成要素モジュールについてのそれぞれの計算された活性化スコアが第２の閾値基準を満たす場合、化学化合物は、目的の生理学的状態と識別される。例示的な第１の閾値基準は、図３Ｅのブロック３４８に関して以下で説明される。一般に、求められるのは、（第１の閾値を満たす計算された活性化スコアを有することによって示されるように）細胞構成要素モジュールの第１のサブセットにおける細胞構成要素モジュールと識別するが、（第２の閾値を満たす計算された活性化スコアを有することによって示されるように）細胞構成要素モジュールの第２のサブセットにおける細胞構成要素モジュールと識別しない化学化合物である。例えば、いくつかの実施形態において、第１の閾値の達成は、第１の所定の数値を上回る活性化スコアを必要とするが、第２の閾値の達成は、第２の所定の数値を下回る活性化スコアを必要とし、正確な第１及び第２の所定の数値は、適用に依存する。

【０２６５】

上記に示されるように、いくつかの実施態様において、方法は、１つ以上のタイプの分子データ（例えば、細胞構成要素）を使用して、目的の生理学的状態（例えば、細胞プロセス）を特徴付けることを含む。そのような分子データは、オミクスプロファイリング（例えば、トランスクリプトミクス、プロテオミクス、メタボロミクスなど）などの測定可能な属性（例えば、存在量及び／又は発現レベル）を有する任意の分析物を含むことができる。

【０２６６】

一般に、細胞プロセスと関連付けられる場合、細胞構成要素（例えば、遺伝子）の細胞構成要素モジュールは、同様の時間にスイッチする細胞構成要素（例えば、遺伝子）が一緒にモジュールを形成する、一連のスイッチングイベントから生じると考えられ得る。したがって、例えば、いくつかの実施形態において、それぞれの細胞構成要素モジュールは、複数の細胞構成要素のそれぞれのサブセットを含み、細胞構成要素のサブセットは、目的のそれぞれの生理学的状態（例えば、目的の細胞プロセス）と関連付けられた挙動の類似性に基づいてグループ化される。一例では、目的のそれぞれの生理学的状態と関連付けられた細胞構成要素モジュールは、それぞれの生理学的状態を有する複数の細胞型にわたって同様に挙動する（例えば、同様の発現プロファイルを示す）遺伝子のサブセットを含むことができる。

【０２６７】

図３Ｄのブロック３４２を参照すると、いくつかの実施形態において、細胞構成要素モジュールのセットは、第１の細胞構成要素モジュールからなる。

【０２６８】

図３Ｄのブロック３４４を参照すると、いくつかの実施形態において、細胞構成要素モジュールのセットは、５つ以上の細胞構成要素モジュールを含む。いくつかの実施形態において、細胞構成要素モジュールのセットは、２０個以上、３０個以上、４０個以上、５０個以上、６０個以上、７０個以上、８０個以上　９０個以上、又は１００個以上の細胞構成要素モジュールを含む。

【０２６９】

いくつかの実施形態において、細胞構成要素モジュールのセットは、少なくとも５個、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも２５個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも６００個、少なくとも７００個、少なくとも８００個、少なくとも９００個、少なくとも１０００個、少なくとも２０００個、少なくとも３０００個、少なくとも４０００個、又は少なくとも５０００個の細胞構成要素モジュールを含む。いくつかの実施形態において、細胞構成要素モジュールのセットは、１０，０００個以下、５０００個以下、２０００個以下、１０００個以下、５００個以下、３００個以下、２００個以下、１００個以下、９０個以下、８０個以下、７０個以下、６０個以下、又は５０個以下の細胞構成要素モジュールを含む。いくつかの実施形態において、細胞構成要素モジュールのセットは、１０～２０００個の細胞構成要素モジュールからなる。いくつかの実施形態において、細胞構成要素モジュールのセットは、５０～５００個の細胞構成要素モジュールからなる。いくつかの実施形態において、細胞構成要素モジュールのセットは、５～２０個、２０～５０個、５０～１００個、１００～２００個、２００～５００個、５００～１０００個、１０００～５０００個、又は５０００～１０，０００個の細胞構成要素モジュールを含む。いくつかの実施形態において、細胞構成要素モジュールのセットは、５個以上の細胞構成要素モジュールから始まり、１０，０００個以下の細胞構成要素モジュールで終わる別の範囲内にある。

【０２７０】

いくつかの実施形態において、方法は、目的の生理学的状態と関連付けられた細胞構成要素モジュールを識別することを更に含む。そのような方法は、図１４Ａ～図１４Ｄと併せて、細胞構成要素モジュールの識別と題されたセクションで以下に説明される。

【０２７１】

活性化スコア。

図３Ｂのブロック３０８に記載されるように、モデルは、モデルへのフィンガープリントの入力に応答して、１つ以上の計算された活性化スコアを出力する。一般に、訓練されたモデル（ブロック３０８のモデル）の出力は、標識（例えば、数値的活性化スコア）を含む訓練データセット上で学習するプロセスを通じて定義され、訓練されたモデルの出力が検証ステップなどを介して性能の最小レベルを満たすまで、複数のパラメータを調整する。訓練モデルは、「モデル訓練」と題されたセクションで以下に更に開示される。

【０２７２】

いくつかの実施形態において、１つ以上の計算された活性化スコアにおける活性化スコアは、それぞれの化合物に対応するそれぞれの細胞構成要素モジュールについてのそれぞれの活性化重みである。例えば、いくつかの実施形態において、活性化スコアは、図２Ａ～図２Ｂ及び図１４Ａ～図１４Ｄを参照して、並びに図５の活性化データ構造に示される「細胞構成要素モジュールの識別」と題された以下のセクションに記載されるように得られる活性化重みであり、活性化スコアは、それぞれの（例えば、第１の）細胞構成要素モジュールの活性化（例えば、誘導及び／又は差次的発現）を示し、それぞれの化合物による治療に相関する及び／又は応答する。

【０２７３】

したがって、いくつかのそのような実施形態において、訓練されたモデルは、出力として、試験化学化合物と目的の生理学的状態（例えば、目的の生理学的状態と関連付けられた第１の細胞構成要素モジュール）との関連性を示す計算された活性化スコアを提供する。次いで、図３Ｅのブロック３４８を参照すると、方法は、第１の細胞構成要素モジュールについてのそれぞれの計算された活性化スコアが第１の閾値基準を満たす場合、化学化合物を目的の生理学的状態で識別する（例えば、関連性を決定する）ことを含む。

【０２７４】

図３Ｅのブロック３５０を参照すると、いくつかの実施形態において、第１の閾値基準は、第１の細胞構成要素モジュールが閾値活性化スコアを有することが必要である。一般

的に、求められるのは、（第１の閾値を満たす計算された活性化スコアを有することによって示されるように）目的の生理学的状態で識別する化学化合物である。例えば、いくつかの実施形態において、第１の閾値の達成は、第１の所定の数値を超える活性化スコアを必要とする。

【０２７５】

例えば、いくつかの実施形態において、活性化スコアは、「０」と「１」との間の正規化された連続値（又はＡ及びＢが２つの異なる数である場合、いくつかの他の範囲の「Ａ」から「Ｂ」）として表され、ここで、「１」に近い値（例えば、０．８９、０．９０、０．９１、０．９２など）は、細胞構成要素モジュール（及び細胞構成要素モジュールが表す化学化合物）と目的の生理学的状態との間の強い関連性を示す。「０」に近い値（例えば、０．０１、０．０２、０．０３、０．０４など）は、細胞構成要素モジュール（及び細胞構成要素が表す化学化合物）と目的の生理学的状態との間に関連性がないことを示す。そのような例では、第１の閾値は、「０」と「１」（又はＡ及びＢが２つの異なる数であるいくつかの他の範囲の「Ａ」から「Ｂ」）との間で選択され、細胞構成要素モジュール（及びそれが表す化学構造）は、活性化スコアが第１の閾値を上回る場合に目的の生理学的状態と関連付けられているとみなされ、一方、細胞構成要素モジュール（及びそれが表す化学構造）は、活性化スコアが第１の閾値を下回る場合に目的の生理学的状態と関連付けられていないとみなされる。いくつかのそのような実施形態において、活性化スコアは、「０」と「１」（又はＡ及びＢが２つの異なる数であるいくつかの他の範囲の「Ａ」から「Ｂ」）との間の連続的な尺度における正規化された値として表され、第１の閾値は、０と１との間、０．１０と０．９０との間、０．２０と０．８０との間、０．３０と０．７０との間、０．５０と０．９９との間、０．６０と０．９９との間、０．７０と０．９９との間、０．８０と０．９９との間、又は０．９０と０．９９との間の値である。

【０２７６】

別の例として、いくつかの実施形態において、活性化スコアは、「０」と「１」（又はＡ及びＢが２つの異なる数である場合、いくつかの他の範囲の「Ａ」から「Ｂ」）との間の連続的な尺度における正規化された値として表され、「１」に近い値（例えば、０．８９、０．９０、０．９１、０．９２など）は、細胞構成要素モジュール（及び細胞構成要素モジュールが表す化学化合物）と目的の生理学的状態との間に関連性がないことを示す。「０」に近い値（例えば、０．０１、０．０２、０．０３、０．０４など）は、細胞構成要素モジュール（及び細胞構成要素が表す化学化合物）と目的の生理学的状態との間の関連性を示す。そのような例では、第１の閾値は、「０」と「１」（又はＡ及びＢが２つの異なる数であるいくつかの他の範囲の「Ａ」から「Ｂ」）との間で選択され、細胞構成要素モジュール（及びそれが表す化学構造）は、活性化スコアが第１の閾値を下回る場合に目的の生理学的状態と関連付けられているとみなされ、一方、細胞構成要素モジュール（及びそれが表す化学構造）は、活性化スコアが第１の閾値を上回る場合に目的の生理学的状態と関連付けられていないとみなされる。いくつかのそのような実施形態において、活性化スコアは、「０」と「１」（又はＡ及びＢが２つの異なる数であるいくつかの他の範囲の「Ａ」から「Ｂ」）との間の連続的な尺度における正規化された値として表され、第１の閾値は、０と１との間、０．１０と０．９０との間、０．２０と０．８０との間、０．３０と０．７０との間、０．５０と０．９９との間、０．６０と０．９９との間、０．７０と０．９９との間、０．８０と０．９９との間、又は０．９０と０．９９との間の値である。

【０２７７】

図３Ｅのブロック３５２を参照すると、いくつかの実施形態において、細胞構成要素モジュールのセットは、複数の細胞構成要素モジュール（例えば、２～１０００個、１０～１００個、２～１００個、４～５０個の細胞構成要素モジュール）であり、ブロック３４８の識別は、細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの計算された活性化スコアが第１の閾値基準を満たすことを必要とする。例えば、細胞構成要素モジュールのセットが２つの細胞構成要素モジュール：Ａ及びＢからなる

場合を考慮する。図３Ｅのブロック３５２は、細胞構成要素モジュールＡ及びＢの活性化スコアの各々が、第１の閾値条件を満たすことを必要とする。例えば、細胞構成要素モジュールＡが０．２５の計算された活性化スコアを有し、細胞構成要素モジュールＢが０．７５の計算された活性化スコアを有し、第１の閾値条件の達成が、各活性化スコアが０．４よりも大きいことを必要とする場合を考慮する。この例では、各活性化スコアが０．４の閾値要件を超えないため、細胞構成要素モジュールのセットは、図３Ｅのブロック３５２の要件を満たさない。

【０２７８】

図３Ｅのブロック３５４を参照すると、いくつかの実施形態において、細胞構成要素モジュールのセットは、複数の細胞構成要素モジュール（例えば、２～１０００個、１０～１００個、２～１００個、４～５０個の細胞構成要素モジュール）であり、ブロック３４８の識別は、細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの計算された活性化スコアにわたる中心傾向の測定値が第１の閾値基準を満たすことを必要とする。例えば、細胞構成要素モジュールのセットが２つの細胞構成要素モジュール：Ａ及びＢからなる場合を考慮する。図３Ｅのブロック３５４は、細胞構成要素モジュールＡ及びＢの活性化スコアの中心傾向のいくつかの測定値が第１の閾値条件を満たすことを必要とする。例えば、中心傾向の測定値が平均化されており、細胞構成要素モジュールＡが０．２５の計算された活性化スコアを有し、細胞構成要素モジュールＢが０．７５の計算された活性化スコアを有し、第１の閾値条件の達成が、平均活性化スコアが０．４よりも大きいことを必要とする場合を考慮する。この例では、細胞構成要素モジュールのセットは、それらが０．４の閾値要件よりも大きい０．２５＋０．７５／２又は０．５の平均活性化スコアを有するため、図３Ｅのブロック３５４の要件を満たす。いくつかの実施形態において、中心傾向の測定値は、細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの計算された活性化スコアの各々の算術平均、加重平均、ミッドレンジ、ミッドヒンジ、トリミアン、ウィンザライズド平均、平均、又はモードである。

【０２７９】

化合物。

いくつかの実施形態において、試験化学化合物は、小分子、生物製剤、タンパク質、小分子と組み合わされたタンパク質、ＡＤＣ、ｓｉＲＮＡ若しくは干渉ＲＮＡなどの核酸、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ｓｈＲＮＡ、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ガイドＲＮＡ（例えば、Ｃａｓ９系若しくは他の細胞成分編集系）、並びに／又は前述のいずれかの任意の組み合わせである。

【０２８０】

いくつかの実施形態において、試験化学化合物は、無機又は有機である。

【０２８１】

例えば、図３Ｅのブロック３５６を参照すると、いくつかの実施形態において、試験化学化合物は、２０００ダルトン（Ｄａ）未満の分子量を有する有機化合物である。いくつかの実施形態において、試験化学化合物は、少なくとも１０Ｄａ、少なくとも２０Ｄａ、少なくとも５０Ｄａ、少なくとも１００Ｄａ、少なくとも２００Ｄａ、少なくとも５００Ｄａ、少なくとも１ｋＤａ、少なくとも２ｋＤａ、少なくとも３ｋＤａ、少なくとも５ｋＤａ、少なくとも１０ｋＤａ、少なくとも２０ｋＤａ、少なくとも３０ｋＤａ、少なくとも５０ｋＤａ、少なくとも１００ｋＤａ、又は少なくとも５００ｋＤａの分子量を有する。いくつかの実施形態において、試験化学化合物は、１０００ｋＤａ以下、５００ｋＤａ以下、１００ｋＤａ以下、５０ｋＤａ以下、１０ｋＤａ以下、５ｋＤａ以下、２ｋＤａ以下、１ｋＤａ以下、５００Ｄａ以下、３００Ｄａ以下、１００Ｄａ以下、又は５０Ｄａ以下の分子量を有する。いくつかの実施形態において、試験化学化合物は、１０Ｄａ～９００Ｄａ、５０Ｄａ～１０００Ｄａ、１００Ｄａ～２０００Ｄａ、１ｋＤａ～１０ｋＤａ、５ｋＤａ～５００ｋＤａ、又は１００ｋＤａ～１０００ｋＤａの分子量を有する。いくつかの実施形態において、試験化学化合物は、１０ダルトン以上から始まり、１０００ｋＤ

ａ以下で終わる別の範囲内にある分子量を有する。
【０２８２】
図３Ｅのブロック３５８を参照すると、いくつかの実施形態において、試験化学化合物は、５つの基準のリピンスキーの法則の各々を満たす有機化合物である。５つ（例えば、ＲＯ５）の基準のリピンスキーの法則は、それぞれの薬理学的又は生物学的活性を有するそれぞれの化合物が、ヒトへの投与に好適な対応する化学的又は物理的特性を有するかどうかを決定するなどの、ドラッグライクネスを評価するために使用されるガイドラインのセットである。５つのリピンスキーの法則は、化合物のドラッグライクネスを決定するための以下の基準を含む。（ｉ）５００Ｄａ未満の分子量、（ｉｉ）５個以下の水素結合ドナー、（ｉｉｉ）１０個以下の水素結合アクセプター、及び（ｉｖ）５個以下のオクタノール－水分配係数ｌｏｇ　Ｐ。
【０２８３】
図３Ｅのブロック３６０を参照すると、いくつかの実施形態において、試験化学化合物は、５つの基準のリピンスキーの法則の少なくとも２つ、３つ、又は４つの基準を満たす有機化合物である。いくつかの実施形態において、試験化学化合物は、５つの基準のリピンスキーの法則のゼロ、１つ、２つ、３つ、又は４つ全ての基準を満たす有機化合物である。
【０２８４】
いくつかの実施形態において、試験化学化合物は、データベースから選択される。薬物スクリーニング、注釈、及び／又は化合物標的及び化合物の化学特性などの一般的な情報からの結果を提供する好適な化合物データベースの例としては、限定されないが、Ｇｅｎｏｍｉｃｓ　ｏｆ　Ｄｒｕｇ　Ｓｅｎｓｉｔｉｖｉｔｙ　ｉｎ　Ｃａｎｃｅｒ、Ｃａｎｃｅｒ　Ｔｈｅｒａｐｅｕｔｉｃｓ　Ｒｅｓｐｏｎｓｅ　Ｐｏｒｔａｌ、Ｃｏｎｎｅｃｔｉｖｉｔｙ　Ｍａｐ、ＰｈａｒｍａｃｏＤＢ、Ｂａｓｅ　ｏｆ　Ｂｉｏｉｓｏｓｔｅｒｉｃａｌｌｙ　Ｅｘｃｈａｎｇｅａｂｌｅ　Ｒｅｐｌａｃｅｍｅｎｔｓ（ＢｏＢＥＲ）、及び／又はＤｒｕｇＢａｎｋが挙げられる。いくつかの実施形態において、試験化学化合物は、遺伝子及び遺伝子産物、摂動誘発細胞構成要素シグネチャ、及び／又は経路注釈に関する情報を提供するデータベースから選択される。好適なデータベースの例としては、限定されないが、ＮＩＨ　Ｇｅｎｅ　Ｅｘｐｒｅｓｓｉｏｎ　Ｏｍｎｉｂｕｓ（ＧＥＯ）、ＥＢＩ　ＡｒｒａｙＥｘｐｒｅｓｓ、ＮＣＢＩ、ＢＬＡＳＴ、ＥＭＢＬ－ＥＢＩ、ＧｅｎＢａｎｋ、Ｅｎｓｅｍｂｌ、ＫＥＧＧ経路データベース、Ｌｉｂｒａｒｙ　ｏｆ　Ｉｎｔｅｇｒａｔｅｄ　Ｎｅｔｗｏｒｋ－ｂａｓｅｄ　Ｃｅｌｌｕｌａｒ　Ｓｉｇｎａｔｕｒｅｓ（ＬＩＮＣＳ）Ｌ１０００データセット、Ｒｅａｃｔｏｍｅ経路データベース、及び／又はＧｅｎｅ　Ｏｎｔｏｌｏｇｙプロジェクトが挙げられる。
【０２８５】
方法３００の結果を実際の適用に使用する。
いくつかの実施形態において、図３と併せて上述した方法３００を使用して、目的の生理学的状態に対して複数の試験化合物を評価する。そのような実施形態において、複数の試験化合物における試験化合物の各々は、図３の方法３００により実行される。したがって、１００個の試験化合物及び１つの目的の生理学的状態が存在する場合、そのような実施形態において、方法３００を１００回実行し、１００回の各事例は、試験化合物の異なる１つに対してである。
【０２８６】
更に、いくつかの実施形態において、図３と併せて上述した方法３００を使用して、目的の複数の生理学的状態に対して複数の化合物を評価する。そのような実施形態において、目的の生理学的状態の各々について、複数の試験化合物におけるそれぞれの各試験化合物の各々は、図３の方法３００により実行される。したがって、１００個の試験化合物及び２つの目的の生理学的状態が存在する場合、そのような実施形態において、方法３００は２００回実行され、２００回の各事例は、目的の第１の生理学的状態又は第２の生理学的状態のいずれかに対する試験化合物の異なる１つに対してである。

【０２８７】
いくつかの実施形態において、複数の試験化合物は、少なくとも５個、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも８００個、少なくとも１０００個、少なくとも２０００個、少なくとも３０００個、少なくとも４０００個、少なくとも５０００個、少なくとも８０００個、少なくとも１０，０００個、少なくとも２０，０００個、少なくとも３０，０００個、少なくとも５０，０００個、少なくとも８０，０００個、少なくとも１００，０００個、少なくとも２００，０００個、少なくとも５００，０００個、少なくとも８００，０００個、少なくとも１００万個、又は少なくとも２００万個の試験化合物を含み、単一の目的の生理学的状態が存在する。いくつかのそのような実施形態において、方法３００は、少なくとも５回、少なくとも１０回、少なくとも１５回、少なくとも２０回、少なくとも３０回、少なくとも４０回、少なくとも５０回、少なくとも１００回、少なくとも２００回、少なくとも３００回、少なくとも４００回、少なくとも５００回、少なくとも８００回、少なくとも１０００回、少なくとも２０００回、少なくとも３０００回、少なくとも４０００回、少なくとも５０００回、少なくとも８０００回、少なくとも１０，０００回、少なくとも２０，０００回、少なくとも３０，０００回、少なくとも５０，０００回、少なくとも８０，０００回、少なくとも１００，０００回、少なくとも２００，０００回、少なくとも５００，０００回、少なくとも８００，０００回、少なくとも１００万回、又は少なくとも２００万回実行されて、少なくとも５、少なくとも１０、少なくとも１５、少なくとも２０、少なくとも３０、少なくとも４０、少なくとも５０、少なくとも１００、少なくとも２００、少なくとも３００、少なくとも４００、少なくとも５００、少なくとも８００、少なくとも１０００、少なくとも２０００、少なくとも３０００、少なくとも４０００、少なくとも５０００、少なくとも８０００、少なくとも１０，０００、少なくとも２０，０００、少なくとも３０，０００、少なくとも５０，０００、少なくとも８０，０００、少なくとも１００，０００、少なくとも２００，０００、少なくとも５００，０００、少なくとも８００，０００、少なくとも１００万、又は少なくとも２００万の活性化スコアを実現し、各試験化合物に対して１つである。

【０２８８】
いくつかの実施形態において、複数の化合物は、１０００万個以下、５００万個以下、１００万個以下、５００，０００個以下、１００，０００個以下、５０，０００個以下、１０，０００個以下、８０００個以下、５０００個以下、２０００個以下、１０００個以下、８００個以下、５００個以下、２００個以下、又は１００個以下の試験化合物を含む。いくつかの実施形態において、複数の化合物は、１０～５００個、１００～１０，０００個、５０００～２００，０００個、又は１０，０００～１００万個の試験化合物からなる。

【０２８９】
いくつかの実施形態において、複数の試験化合物は、$10 \sim 1 \times 10^{6}$個の試験化合物である。いくつかの実施形態において、複数の試験化合物は、１００～１００，０００個の試験化合物である。いくつかの実施形態において、複数の試験化合物は、１０００～１００，０００個の試験化合物である。

【０２９０】
したがって、方法３００を使用して、多数の試験化合物についての活性化スコアを得ることができる。これらの活性化スコアに対する第１の閾値の適用を使用して、目的の生理学的状態と関連付けられる試験された多くの試験化合物の中から、試験化合物を識別することができる。典型的な実施形態において、選択された数の試験化合物は、それらが目的の生理学的状態と関連付けられることを示す活性化スコアを有するが、他のものはそうではない。選択された数の試験化合物の分析を使用して、目的の生理学的状態との関連性をもたらす化合物を試験するための分子特性を決定することができる。例えば、目的の生理学的状態と関連付けられていることを示す活性化スコアを有する選択された数の試験化合

物の化学構造を、目的の生理学的状態と関連付けられていない試験化合物と区別する構造の類似性について視覚的に検査することができる。次いで、そのような分子特性は、モデル６０１によって評価された元の試験分子に含まれず、モデル６０１を訓練するために使用されなかった新しい試験分子に組み込むことができる。

【０２９１】

更に、より正式なアプローチを使用して、試験化合物を分析することができる（方法３００によって課された第１の閾値を満たすものと満たさないものとの両方）。例えば、部分構造マイニングを使用して、そのような化合物を目的の生理学的状態と関連付けるようにする試験化合物内の部分構造を特定することができる。部分構造マイニングの例としては、ＭＯＳＳ（参照により本明細書に組み込まれる、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、及びＭＯＦＡ（参照により本明細書に組み込まれる、Ｍｅｉｎｌ ａｎｄ Ｗｏｒｌｅｉｎ，２００６ “Ｍｉｎｉｎｇ Ｍｏｌｅｃｕｌａｒ Ｄａｔａｓｅｔｓ ｏｎ Ｓｙｍｍｅｔｒｉｃ Ｐｒｏｃｅｓｓｏｒ Ｓｙｓｔｅｍｓ，”Ｉｎｔｅｒｎａｔｉｏｎａｌ ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｓｙｓｔｅｍｓ，ｍａｎ ａｎｄ Ｃｙｂｅｒｎｅｔｉｃｓ ２，ｐｐ．１２６９－１２７４）が挙げられるが、これらに限定されない。

【０２９２】

また、最大共通部分構造（ＭＣＳ）分析を使用して、そのような化合物を目的の生理学的状態と関連付ける試験化合物内の部分構造を識別することができる。ＭＣＳ分析の例としては、ＬＩＢＭＣＳ（Ｃｈｅｍａｘｏｎ，Ｌｉｂｒａｒｙ ＭＣＳ，２００８）、ＭＣＳＳ（ＯＥＣｈｅｍ ＴＫ ｖｅｒｓｉｏｎ ２．０．０，ＯｐｅｎＥｙｅ Ｓｃｉｅｎｔｉｆｉｃ Ｓｏｆｔｗａｒｅ，Ｓａｎｔａ Ｆｅ，ＮＭ．ｈｔｔｐ：／／ｗｗｗ．ｅｙｅｓｏｐｅｎ．ｃｏｍ）、及びＣｎｃＭＣＳ（ｈｔｔｐ：／／ｗｗｗ．ｃｈｅｍｎａｖｉｇａｔｏｒ．ｃｏｍ／ｃｎｃ／ｐｒｏｄｕｃｔｓ／ｄｏｗｎｌｏａｄｓ．ａｓｐ）が挙げられるが、これらに限定されない。

【０２９３】

また、ＳＭＡＲＴＳを使用して、そのような化合物を目的の生理学的状態と関連付ける試験化合物内の部分構造を識別することができる。ＳＭＡＲＴ分析の例は、ＣＤＫ Ｄｅｓｃｒｉｐｔｏｒ ＧＵＩである。

【０２９４】

また、頻出部分グラフマイニングを使用して、そのような化合物を目的の生理学的状態と関連付けるようにする試験化合物内の部分構造を識別することができる。頻出部分グラフマイニングの例は、ＰａｒＭｏｌ（Ｕｎｉ Ｅｒｌａｎｇｅｎ）である。

【０２９５】

また、グラフ及び化学マイニングを使用して、そのような化合物を目的の生理学的状態と関連付けるようにする試験化合物内の部分構造を識別することができる。グラフ及び化学マイニングの例は、ＰＡＦＩ／ＡＦＧｅｎ（Ｋａｒｙｐｉｓ Ｌａｂ ＵＭＮ）である。

【０２９６】

摂動シグネチャ。

上記のように、いくつかの実施形態において、目的の生理学的状態は、摂動シグネチャ（例えば、摂動に応答して第１の細胞状態と第２の細胞状態との間の不一致を特徴とする）である。したがって、本開示の別の態様は、試験化学化合物を目的の生理学的状態と関連付ける方法７００を提供する。いくつかの実施形態において、目的の生理学的状態は、疾患である。

【０２９７】

ブロック７０２を参照すると、方法は、試験化学化合物の化学構造のフィンガープリントを得ることを含む。「生理学的状態」及び「化合物」と題する上記のセクションに開示

されるような、生理学的状態、化合物、フィンガープリント、及び／又はフィンガープリントを得る方法の任意の好適な実施形態は、当業者には明白であろうように、それらの任意の置換、修飾、追加、欠失、及び／又は組み合わせを含むことが企図される。

【０２９８】

例えば、いくつかの実施形態において、試験化学化合物は、２０００ダルトン未満の分子量を有する有機化合物である。いくつかの実施形態において、試験化学化合物は、５つの基準のリピンスキーの法則の各々を満たす有機化合物である。いくつかの実施形態において、試験化学化合物は、５つの基準のリピンスキーの法則のうちの少なくとも３つの基準を満たす有機化合物である。いくつかの実施形態において、方法は、試験化学化合物の単純化された分子入力ラインエントリーシステム（ＳＭＩＬＥＳ）文字列表現からフィンガープリントを計算することを更に含む。いくつかの実施形態において、フィンガープリントは、ＳＭＩＬＥＳ　Ｔｒａｎｓｆｏｒｍｅｒ、ＥＣＦＰ４、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、化学構造から生成される。

【０２９９】

ブロック７０４を参照すると、方法は、フィンガープリントをモデルに入力することを更に含み、モデルは１００以上のパラメータを含み、モデルは、フィンガープリントのモデルへの入力に応答して１つ以上の計算された活性化スコアを出力し、１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、摂動シグネチャのセットにおける対応する摂動シグネチャを表す。

【０３００】

「モデルアーキテクチャ」と題された上記のセクションに開示されるものなどのモデルの任意の好適な実施形態が企図され、当業者には明らかであろうように、それらの任意の置換、修飾、追加、削除、及び／又は組み合わせが企図される。例えば、いくつかの実施形態において、モデルは、ニューラルネットワークを含む。いくつかのそのような実施形態において、ニューラルネットワークは、ＲｅＬＵ活性化を有する完全に接続されたニューラルネットワークである。いくつかの実施形態において、ニューラルネットワークは、メッセージパッシングニューラルネットワークである。

【０３０１】

いくつかの実施形態において、モデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。

【０３０２】

いくつかの実施形態において、モデルは、複数のコンポーネントモデルのアンサンブルモデルであり、１つ以上の計算された活性化スコアにおける計算された活性化スコアの各々は、複数のコンポーネントモデルにおけるコンポーネントモデルの各々の出力の中心傾向の測定値である。

【０３０３】

いくつかの実施形態において、複数のコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。

【０３０４】

いくつかの実施形態において、複数のコンポーネントモデルは、複数のニューラルネットワークを含む。いくつかのそのような実施形態において、複数のニューラルネットワークにおける第１のニューラルネットワークは、ＲｅＬＵ活性化を伴う完全に接続されたニューラルネットワークであり、複数のニューラルネットワークにおける第２のニューラルネットワークは、メッセージパッシングニューラルネットワークである。

【０３０５】

上記で定義されるように、摂動は、１つ以上の化合物による治療などの１つ以上の状態

への細胞の任意の曝露を指す。いくつかの実施形態において、摂動シグネチャは、摂動によって誘発される細胞内の１つ以上の細胞構成要素の発現又は存在量レベルの変化である。

【０３０６】

例示的な摂動には、限定されないが、遺伝子ノックダウン、刺激に対する細胞応答、組織成長及び再生、並びに／又は化合物による治療若しくは化合物への曝露が含まれる。例示的なペルターバゲンには、小分子、生物製剤、治療剤、タンパク質、小分子と組み合わされたタンパク質、ＡＤＣ、ｓｉＲＮＡ若しくは干渉ＲＮＡなどの核酸、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ｓｈＲＮＡ、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ガイドＲＮＡ（例えば、Ｃａｓ９系若しくは他の遺伝子編集系）、又は前述のいずれかの任意の組み合わせが含まれるが、これらに限定されない。

【０３０７】

いくつかの実施形態において、摂動は、システムレベル（例えば、結合又はドッキング活性）、並びに／又は下流効果及び臓器レベルの表現型に関して特徴付けられる。いくつかの実施形態において、摂動は、分子、細胞、及び／又は組織レベルでのペルターバゲンに対する応答を駆動する又はその基礎となる機構の機能として特徴付けられる（例えば、摂動の前又は後にバイオマーカー、細胞生存率、及び／又は薬物タンパク質相互作用を識別又は測定することによって）。例えば、摂動の測定値は、表現型の測定値（例えば、ＩＣ５０値）及び／又は細胞構成要素シグネチャ（例えば、オミクスプロファイリング）を含むことができる。

【０３０８】

いくつかの実施形態において、それぞれの摂動及び／又は対応する摂動シグネチャは、Ｇｅｎｏｍｉｃｓ ｏｆ Ｄｒｕｇ Ｓｅｎｓｉｔｉｖｉｔｙ ｉｎ Ｃａｎｃｅｒ、Ｃａｎｃｅｒ Ｔｈｅｒａｐｅｕｔｉｃｓ Ｒｅｓｐｏｎｓｅ Ｐｏｒｔａｌ、Ｃｏｎｎｅｃｔｉｖｉｔｙ Ｍａｐ、ＰｈａｒｍａｃｏＤＢ、Ｂａｓｅ ｏｆ Ｂｉｏｉｓｏｓｔｅｒｉｃａｌｌｙ Ｅｘｃｈａｎｇｅａｂｌｅ Ｒｅｐｌａｃｅｍｅｎｔｓ（ＢｏＢＥＲ）、ＤｒｕｇＢａｎｋ、Ｈｕｍａｎ Ｃｅｌｌ Ａｔｌａｓ、Ｍｏｌｅｃｕｌａｒ Ｓｉｇｎａｔｕｒｅｓ Ｄａｔａｂａｓｅ（ＭＳｉｇＤＢ）、及び／又はＥｎｒｉｃｈｒなどの公的に利用可能なデータベースから得られる。摂動データを得ることができる他の好適なデータベースとしては、ＮＩＨ Ｇｅｎｅ Ｅｘｐｒｅｓｓｉｏｎ Ｏｍｎｉｂｕｓ（ＧＥＯ）、ＥＢＩ ＡｒｒａｙＥｘｐｒｅｓｓ、ＮＣＢＩ、ＢＬＡＳＴ、ＥＭＢＬ－ＥＢＩ、ＧｅｎＢａｎｋ、Ｅｎｓｅｍｂｌ、ＫＥＧＧ経路データベース、Ｌｉｂｒａｒｙ ｏｆ Ｉｎｔｅｇｒａｔｅｄ Ｎｅｔｗｏｒｋ－ｂａｓｅｄ Ｃｅｌｌｕｌａｒ Ｓｉｇｎａｔｕｒｅｓ（ＬＩＮＣＳ）Ｌ１０００データセット、Ｒｅａｃｔｏｍｅ経路データベース、及び／又はＧｅｎｅ Ｏｎｔｏｌｏｇｙプロジェクトが挙げられる。

【０３０９】

摂動データを得る方法には、例えば、ｐｅｒｔｕｒｂ－ｓｅｑ、ＣＲＩＳＰ－ｓｅｑ、ＣＲＯＰ－ｓｅｑ、ＣＲＩＳＰＲｉ、ＴＡＰ－ｓｅｑ、ＣＲＩＳＰＲａ、ｐｅｒｔｕｒｂ－ＣＩＴＥ－ｓｅｑ、ｓｃｉ－Ｐｌｅｘ、ｍｕｌｔｉｐｌｅｘｅｄ、ＭＩＸ－ｓｅｑ、ＣｙＴＯＦ、及び／又はｓｃＲＮＡ－ｓｅｑを使用した細胞構成要素データの測定が含まれる。摂動データを得る方法には、更に、質量分析（例えば、ＬＣＭＳ、ＧＣＭＳ）、フローサイトメトリー、定量的ポリメラーゼ連鎖反応（ｑＰＣＲ）、ゲル電気泳動、遺伝子チップ分析、マイクロアレイ、細胞蛍光分析、蛍光顕微鏡、共焦点レーザースキャニング顕微鏡、レーザースキャニングサイトメトリー、親和性クロマトグラフィー、手動バッチモード分離、電界懸濁、配列決定、及び／又はそれらの任意の組み合わせを含む、オミクスデータを得る任意の方法が含まれる。いくつかの実施形態において、本明細書に開示される細胞構成要素存在量値を得るための方法のうちのいずれかは、摂動データを得る際に（例えば、摂動シグネチャのために）使用するために企図される。

【０３１０】

いくつかの実施形態において、摂動シグネチャのセットは、第１の摂動シグネチャから

なる。いくつかの実施形態において、摂動シグネチャのセットは、５つ以上の摂動シグネチャを含む。いくつかの実施形態において、摂動シグネチャのセットは、１０個以上の摂動シグネチャを含む。いくつかの実施形態において、摂動シグネチャのセットは、１００個以上の摂動シグネチャを含む。

【０３１１】

いくつかの実施形態において、摂動シグネチャのセットは、少なくとも２つ、少なくとも３つ、少なくとも４つ、少なくとも５つ、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも８００個、少なくとも１０００個、少なくとも２０００個、又は少なくとも５０００個の摂動シグネチャを含む。いくつかの実施形態において、摂動シグネチャのセットは、１０，０００個以下、５０００個以下、１０００個以下、８００個以下、５００個以下、２００個以下、１００個以下、５０個以下、又は２０個以下の摂動シグネチャを含む。いくつかの実施形態において、摂動シグネチャのセットは、５～５０個、２～１００個、２０～５００個、１０～１０００個、８００～５０００個、又は５０～２０００個の摂動シグネチャを含む。いくつかの実施形態において、摂動シグネチャのセットは、２つ以上の摂動シグネチャから始まり、１０，０００個以下の摂動シグネチャで終わる別の範囲内にある。

【０３１２】

ブロック７０６を参照すると、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、それぞれの第１の細胞状態及び第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である。

【０３１３】

いくつかの実施形態において、複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの摂動状態は、複数の化合物における化合物に曝露されていない対照細胞によって表される。いくつかの実施形態において、複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの摂動状態は、それぞれの摂動シグネチャに関連付けられた化合物以外の複数の化学化合物における化学化合物に曝露されている無関係の摂動細胞にわたる平均によって表される。

【０３１４】

いくつかの実施形態において、細胞状態の変化は、変化していない細胞状態と変化した細胞状態との間の変化を指し、変化した細胞状態は、変化していない細胞状態から変化した細胞状態への細胞遷移を通じて生じる。更に、（ｉ）変化していない細胞状態、（ｉｉ）変化した細胞状態、及び（ｉｉｉ）変化していない細胞状態から変化した細胞状態への遷移のうちの少なくとも１つが、目的の生理学的状態と関連付けられる。

【０３１５】

いくつかの実施形態において、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャは、非限定的な例として、参照により本明細書に組み込まれる、２０１９年７月１５日に出願された「Ｍｅｔｈｏｄｓ　ｏｆ　Ａｎａｌｙｚｉｎｇ　Ｃｅｌｌｓ」と題された米国特許出願第１６／５１１，６９１号に開示された方法のいずれかを使用して決定され得る。

【０３１６】

ある特定の実施形態において、摂動（例えば、特定の化学組成物への細胞の曝露）の共変量が存在し得る。例えば、化学組成物の共変量は、化学組成物の特定の用量、化学組成

物に曝露された細胞が細胞構成要素を定量化するために測定される時間、及び／又は化学組成物に曝露された細胞の同一性（例えば、細胞株）を含み得る。いくつかの実施形態において、摂動（例えば、特定の化学組成物への細胞の曝露）は、その共変量の閾値量も特定の細胞遷移に影響すると予測される場合にのみ、特定の細胞遷移に影響すると予測される。言い換えれば、いくつかの実施形態において、特定の摂動シグネチャの計算された活性化スコアは、特定の摂動シグネチャの化学組成物の共変量が、目的の生理学的状態と関連付けられた特定の細胞遷移にも影響を与えると予測されるかどうかによって少なくとも部分的に決定される。

【０３１７】

一般に、上述のように、訓練されたモデルの出力は、標識（例えば、数値的活性化スコア）を含む訓練データセット上で学習するプロセスを通じて定義され、訓練されたモデルの出力が検証ステップなどを介して性能の最小レベルを満たすまで、複数のパラメータを調整する。訓練モデルは、「モデル訓練」と題されたセクションで以下に更に開示される。したがって、いくつかのそのような実施形態において、訓練されたモデルは、出力として、試験化学化合物と目的の生理学的状態との関連性を示す第１の摂動シグネチャについて計算された活性化スコアを提供する（例えば、第１の摂動シグネチャは、目的の生理学的状態と関連付けられた細胞状態遷移と関連付けられる）。

【０３１８】

次にブロック７０８を参照すると、方法は、摂動シグネチャのセットにおける第１の摂動シグネチャについてのそれぞれの計算された活性化スコアが第１の閾値基準を満たす場合、化学化合物を目的の生理学的状態と識別することを含む。

【０３１９】

「活性化スコア」と題する上記のセクションに開示されるような活性化スコアの任意の好適な実施形態は、１つ以上の計算された活性化スコアを得るために企図され、活性化スコアの各々は、当業者に明らかであろうように、それらの任意の置換、修飾、追加、欠失、及び／又は組み合わせを含む、摂動シグネチャのセットにおける対応する摂動シグネチャを表す。

【０３２０】

一般的に、求められるのは、（第１の閾値基準を満たす計算された活性化スコアを有することによって示されるように）目的の生理学的状態と識別する化学化合物である。例えば、いくつかの実施形態において、第１の閾値の達成は、第１の所定の数値を超える活性化スコアを必要とする。

【０３２１】

例えば、いくつかの実施形態において、活性化スコアは、「０」と「１」（又はＡ及びＢが２つの異なる数である場合、いくつかの他の範囲の「Ａ」から「Ｂ」）との間の連続的な尺度における正規化された値として表され、ここで、「１」に近い値（例えば、０．８９、０．９０、０．９１、０．９２など）は、摂動シグネチャ（及び摂動シグネチャが表す化学化合物）と目的の生理学的状態との間の強い関連性を示す。「０」に近い値（例えば、０．０１、０．０２、０．０３、０．０４など）は、摂動シグネチャ（及び摂動シグネチャが表す化学化合物）と目的の生理学的状態との間に関連性がないことを示す。そのような例では、第１の閾値は、「０」と「１」（又はＡ及びＢが２つの異なる数であるいくつかの他の範囲の「Ａ」から「Ｂ」）との間で選択され、摂動シグネチャ（及びそれが表す化学構造）は、活性化スコアが第１の閾値を上回る場合に目的の生理学的状態と関連付けられているとみなされ、一方、摂動シグネチャ（及びそれが表す化学構造）は、活性化スコアが第１の閾値を下回る場合に目的の生理学的状態と関連付けられていないとみなされる。いくつかのそのような実施形態において、活性化スコアは、「０」と「１」（又はＡ及びＢが２つの異なる数であるいくつかの他の範囲の「Ａ」から「Ｂ」）との間の連続的な尺度における正規化された値として表され、第１の閾値は、０と１との間、０．１０と０．９０との間、０．２０と０．８０との間、０．３０と０．７０との間、０．５０と０．９９との間、０．６０と０．９９との間、０．７０と０．９９との間、０．８０

と０．９９との間、又は０．９０と０．９９との間の値である。
【０３２２】
別の例として、いくつかの実施形態において、活性化スコアは、「０」と「１」（又はＡ及びＢが２つの異なる数である場合、いくつかの他の範囲の「Ａ」から「Ｂ」）との間の連続的な尺度における正規化された値として表され、ここで、「１」に近い値（例えば、０．８９、０．９０、０．９１、０．９２など）は、摂動シグネチャ（及び摂動シグネチャが表す化学化合物）と目的の生理学的状態との間に関連性がないことを示す。「０」に近い値（例えば、０．０１、０．０２、０．０３、０．０４など）は、摂動シグネチャ（及び摂動シグネチャが表す化学化合物）と目的の生理学的状態との間の関連性を示す。そのような例では、第１の閾値は、「０」と「１」（又はＡ及びＢが２つの異なる数であるいくつかの他の範囲の「Ａ」から「Ｂ」）との間で選択され、摂動シグネチャ（及びそれが表す化学構造）は、活性化スコアが第１の閾値を下回る場合に目的の生理学的状態と関連付けられているとみなされ、一方、摂動シグネチャ（及びそれが表す化学構造）は、活性化スコアが第１の閾値を上回る場合に目的の生理学的状態と関連付けられていないとみなされる。いくつかのそのような実施形態において、活性化スコアは、「０」と「１」（又はＡ及びＢが２つの異なる数であるいくつかの他の範囲の「Ａ」から「Ｂ」）との間の連続的な尺度における正規化された値として表され、第１の閾値は、０と１との間、０．１０と０．９０との間、０．２０と０．８０との間、０．３０と０．７０との間、０．５０と０．９９との間、０．６０と０．９９との間、０．７０と０．９９との間、０．８０と０．９９との間、又は０．９０と０．９９との間の値である。
【０３２３】
いくつかの実施形態において、第１の閾値基準は、第１の摂動シグネチャが閾値活性化スコアを有することが必要である。
【０３２４】
いくつかの実施形態において、第１の閾値基準は、第１の摂動シグネチャが、摂動シグネチャのセットにおける少なくとも閾値ランクを有することが必要であり、摂動シグネチャのセットは、摂動シグネチャのセットにおける摂動シグネチャの各々と参照シグネチャ（例えば、単一細胞遷移シグネチャ）との比較に基づいてランク付けされる。化学化合物を生理学的状態と関連付ける際の使用に好適な参照シグネチャ（例えば、単一細胞遷移シグネチャ）に対する摂動シグネチャの比較方法は、以下の「摂動シグネチャについての数値的活性化スコア」と題されたセクションに更に詳細に記載されている。
【０３２５】
いくつかの実施形態において、識別は、摂動シグネチャのセットにおける摂動シグネチャの各々のそれぞれの計算された活性化スコアが閾値基準を満たすことを必要とする。いくつかの実施形態において、識別は、摂動シグネチャのセットにおける摂動シグネチャの各々のそれぞれの計算された活性化スコアにわたる中心傾向の測定値が閾値基準を満たすことを必要とする。いくつかの実施形態において、中心傾向の測定値は、摂動シグネチャのセットにおける摂動シグネチャの各々のそれぞれの計算された活性化スコアの各々の算術平均、加重平均、ミッドレンジ、ミッドヒンジ、トリミアン、ウィンザライズド平均、平均、又はモードである。
【０３２６】
いくつかの実施形態において、摂動シグネチャのセットは、２～１００個の摂動シグネチャであり、識別は、摂動シグネチャのセットにおける摂動シグネチャの各々のそれぞれの計算された活性化スコアが閾値基準を満たすことを必要とする。いくつかの実施形態において、摂動シグネチャのセットは、２～１００個の摂動シグネチャであり、識別は、摂動シグネチャのセットにおける摂動シグネチャの各々のそれぞれの計算された活性化スコアにわたる中心傾向の測定値が閾値基準を満たすことを必要とする。いくつかの実施形態において、中心傾向の測定値は、摂動シグネチャのセットにおける摂動シグネチャの各々のそれぞれの計算された活性化スコアの各々の算術平均、加重平均、ミッドレンジ、ミッドヒンジ、トリミアン、ウィンザライズド平均、平均、又はモードである。

【０３２７】
いくつかの実施形態において、摂動シグネチャのセットは、複数の摂動シグネチャであり、第１の摂動シグネチャを含む、複数の摂動シグネチャの第１のサブセットが、目的の生理学的状態と関連付けられ、複数の摂動シグネチャの第２のサブセットが、目的の生理学的状態と関連付けられておらず、第１の摂動シグネチャについてのそれぞれの計算された活性化スコアが、第１の閾値基準を満たし、複数の摂動シグネチャの第２のサブセットにおける摂動シグネチャについてのそれぞれの計算された活性化スコアが、第２の閾値基準を満たす場合、試験化学化合物が、目的の生理学的状態と識別される。
【０３２８】
いくつかの実施形態において、第２の閾値基準は、複数の摂動シグネチャの第２のサブセットにおける摂動シグネチャについてのそれぞれの計算された活性化スコアが閾値活性化スコアを有することを必要とする。
【０３２９】
いくつかの実施形態において、第２の閾値基準は、複数の摂動シグネチャの第２のサブセットにおける摂動シグネチャについてのそれぞれの計算された活性化スコアが、摂動シグネチャのセットにおける少なくとも閾値ランクを有することを必要とし、摂動シグネチャのセットは、摂動シグネチャのセットにおける摂動シグネチャの各々と参照シグネチャ（例えば、単一細胞遷移シグネチャ）との比較に基づいてランク付けされる。
【０３３０】
いくつかの実施形態において、識別は、摂動シグネチャの第２のサブセットにおける摂動シグネチャの各々のそれぞれの計算された活性化スコアが第２の閾値基準を満たすことを必要とする。いくつかの実施形態において、識別は、摂動シグネチャの第２のサブセットにおける摂動シグネチャの各々のそれぞれの計算された活性化スコアにわたる中心傾向の測定値が第２の閾値基準を満たすことを必要とする。いくつかの実施形態において、中心傾向の測定値は、摂動シグネチャのセットにおける摂動シグネチャの各々のそれぞれの計算された活性化スコアの各々の算術平均、加重平均、ミッドレンジ、ミッドヒンジ、トリミアン、ウィンザライズド平均、平均、又はモードである。
【０３３１】
ＩＩＩ．化学化合物を目的の生理学的状態と関連付ける方法
モデル訓練。
本開示の別の態様は、化学化合物を目的の生理学的状態と関連付ける方法８００を提供する。いくつかの実施形態において、目的の生理学的状態は、疾患である。
【０３３２】
ブロック８０２を参照すると、方法は、複数の化合物における化合物の各々の化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得る。「生理学的状態」及び「化合物」と題する上記のセクションに開示されるような、生理学的状態、化合物、フィンガープリント、及び／又はフィンガープリントを得る方法の任意の好適な実施形態は、当業者には明白であろうように、それらの任意の置換、修飾、追加、欠失、及び／又は組み合わせを含むことが企図される。
【０３３３】
例えば、いくつかの実施形態において、複数の化合物は、１０～１×１０^{6}個の化合物である。いくつかの実施形態において、複数の化合物は、１００～１００，０００個の化合物である。いくつかの実施形態において、複数の化合物は、１０００～１００，０００個の化合物である。
【０３３４】
いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、２０００ダルトン未満の分子量を有する有機化合物である。いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、５つの基準のリピンスキーの法則の各々を満たす。いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、５つの基準のリピンスキーの法則のうちの少なくとも３つの基準を満たす。いくつかの実施

形態において、それぞれのフィンガープリントの各々は、ＳＭＩＬＥＳ　Ｔｒａｎｓｆｏｒｍｅｒ、ＥＣＦＰ４、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、化学構造から生成される。

【０３３５】

ブロック８０４を参照すると、方法は、複数の化合物における化合物の各々についての細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの数値的活性化スコアを電子形式で得ることを含み、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素の独立したサブセットを含む。上記「細胞構成要素及び細胞構成要素モジュール」、及び以下「細胞構成要素モジュールの識別」と題するセクションに開示されるように、細胞構成要素、細胞構成要素モジュール、及び／又は細胞構成要素モジュールを識別する方法の任意の好適な実施形態が企図され、当業者には明らかであろうように、それらの任意の置換、修飾、追加、欠失、及び／又は組み合わせが含まれる。

【０３３６】

例えば、いくつかの実施形態において、細胞構成要素モジュールのセットは、単一の細胞構成要素モジュールである。いくつかの実施形態において、細胞構成要素モジュールのセットは、複数の細胞構成要素モジュールである。いくつかの実施形態において、細胞構成要素モジュールのセットは、２００～５００個の細胞構成要素モジュールである。いくつかの実施形態において、細胞構成要素モジュールのセットは、単一の細胞構成要素モジュールからなる。いくつかの実施形態において、細胞構成要素モジュールのセットは、５つ以上の細胞構成要素モジュールを含む。いくつかの実施形態において、細胞構成要素モジュールのセットは、１０個以上の細胞構成要素モジュールを含む。いくつかの実施形態において、細胞構成要素モジュールのセットは、１００個以上の細胞構成要素モジュールを含む。いくつかの実施形態において、細胞構成要素モジュールのセットは、複数の細胞構成要素モジュールであり、複数の細胞構成要素モジュールの第１のサブセットは、目的の生理学的状態と関連付けられ、複数の細胞構成要素モジュールの第２のサブセットは、目的の生理学的状態と関連付けられていない。

【０３３７】

いくつかの実施形態において、図２Ａ～図２Ｂの例示的なワークフローによって示されるように、方法は、電子形式で１つ以上の第１のデータセットを得ることであって、１つ以上の第１のデータセットが、第１の複数の細胞におけるそれぞれの細胞の各々について、第１の複数の細胞が、２０個以上の細胞を含み、複数の注釈付きの細胞状態を集合的に表し、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、複数の細胞構成要素が、１０個以上の細胞構成要素を含み、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含む、得ることを含む、プロセスによって複数の細胞構成要素モジュールにおける細胞構成要素モジュールを識別することを更に含む。したがって、方法は、複数のベクトルにアクセスするか、又はそれらを形成し、複数のベクトルにおけるそれぞれのベクトルの各々が、（ｉ）複数の構成要素におけるそれぞれの細胞構成要素に対応し、（ｉｉ）対応する複数のエレメントを含み、対応する複数のエレメントにおけるそれぞれのエレメントの各々が、第１の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を表す対応するカウントを有する。複数のベクトルは、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別するために使用される。複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々は、複数の細胞構成要素のサブセットを含み、複数の細胞構成要素モジュールは、（ｉ）複数の候補細胞構成要素モジュール、及び（ｉｉ）複数の細胞構成要素、又はその表現によって次元決定された潜在表現で配置され、複数の細胞構成要素モジュールは、１０を超える細胞構成要素モジュールを含む。

【０３３８】

１つ以上の第２のデータセットは、電子形式で得られ、１つ以上の第２のデータセットは、第２の複数の細胞におけるそれぞれの細胞の各々について、第２の複数の細胞が、２

０個以上の細胞を含み、目的の生理学的状態を通知する複数の共変量を集合的に表し、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含む。したがって、（ｉ）第２の複数の細胞及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された細胞構成要素カウントデータ構造が得られる。活性化データ構造は、複数の細胞構成要素又はその表現を共通次元として使用して、細胞構成要素カウントデータ構造及び潜在表現を組み合わせることによって形成され、活性化データ構造は、複数の細胞構成要素モジュールにおける細胞構成要素モジュールの各々について、第２の複数の細胞における細胞の各々について、それぞれの活性化重みを含む。

【０３３９】

候補細胞構成要素モデルは、（ｉ）活性化データ構造を候補モデルに入力したときに、活性化データ構造内に表される細胞構成要素モジュールの各々における複数の共変量における各共変量の不在又は存在の予測と、（ｉｉ）細胞構成要素モジュールの各々における各共変量の実際の不在又は存在との間の差を使用して訓練され、訓練は、差に応答して、候補細胞構成要素モデルと関連付けられた複数の共変量重みを調整し、複数の共変量重みは、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの共変量の各々について、それぞれの共変量が、活性化データ構造にわたって、それぞれの細胞構成要素モジュールと相関するかどうかを示す対応する重みを含む。候補細胞構成要素モデルを訓練する際に、複数の共変量重みを使用して、複数の候補細胞構成要素モジュールにおける細胞構成要素モジュール（例えば、目的の生理学的状態と関連付けられるもの）を識別する。

【０３４０】

いくつかの実施形態において、目的の生理学的状態は、疾患であり、第１の複数の細胞が、複数の注釈付きの細胞状態によって立証されるように、疾患を代表する細胞、及び疾患を代表しない細胞を含む。いくつかの実施形態において、複数の注釈付きの細胞状態における注釈付きの細胞状態は、曝露条件下での化合物への第１の複数の細胞における細胞の曝露である。いくつかの実施形態において、曝露条件は、曝露期間、化合物の濃度、又は曝露期間及び化合物の濃度の組み合わせである。

【０３４１】

いくつかの実施形態において、複数の細胞構成要素における細胞構成要素の各々は、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせである。いくつかの実施形態において、第１又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、比色測定、蛍光測定、発光測定、又は共鳴エネルギー移動（ＦＲＥＴ）測定によって決定される。いくつかの実施形態において、第１又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、単一細胞リボ核酸（ＲＮＡ）配列決定（ｓｃＲＮＡ－ｓｅｑ）、ｓｃＴａｇ－ｓｅｑ、配列決定を使用したトランスポザーゼ－アクセス可能なクロマチンのための単一細胞アッセイ（ｓｃＡＴＡＣ－ｓｅｑ）、ＣｙＴＯＦ／ＳＣｏＰ、Ｅ－ＭＳ／Ａｂｓｅｑ、ｍｉＲＮＡ－ｓｅｑ、ＣＩＴＥ－ｓｅｑ、及びそれらの任意の組み合わせによって決定される。いくつかの実施形態において、複数の細胞構成要素は、１００～８，０００個の細胞構成要素からなる。

【０３４２】

いくつかの実施形態において、複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することは、複数のベクトルにおけるベクトルの各々の対応する複数のエレメントの各々を使用して、複数のベクトルに相関モデルを適用することを含む。いくつかの実施形態において、相関モデルは、グラフクラスタリングを含む。いくつかの実施形態において、グラフクラスタリング方法は、ピアソン相関ベースの距離メトリック上のライデン（Ｌｅｉｄｅｎ）クラスタリングであるか、又はルーバン（Ｌｏｕｖａｉｎ）クラスタリングである。

【０３４３】
いくつかの実施形態において、複数の細胞構成要素モジュールは、１０～２０００個の細胞構成要素モジュールからなる。いくつかの実施形態において、複数の構成要素モジュールにおける候補細胞構成要素モジュールの各々は、２００～３００個の細胞構成要素からなる。
【０３４４】
いくつかの実施形態において、複数の共変量は、細胞バッチ、細胞ドナー、細胞型、疾患状態、又は化学化合物への曝露を含む。
【０３４５】
いくつかの実施形態において、候補細胞構成要素モデルを訓練することは、マルチタスク策定におけるカテゴリ交差エントロピー損失を使用して実施され、複数の共変量における共変量の各々が、複数のコスト関数におけるコスト関数に対応し、複数のコスト関数におけるそれぞれのコスト関数の各々が、共通の重み付け係数を有する。
【０３４６】
ブロック８０６を参照すると、方法は、複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々について、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの細胞構成要素モジュールについてのそれぞれの計算された活性化スコアと、（ｉｉ）細胞構成要素モジュールのセットにおけるそれぞれの化合物についてのそれぞれの細胞構成要素モジュールのそれぞれの数値的活性化スコアとの間のそれぞれの差を使用して訓練されていないモデルを訓練することを更に含む。
【０３４７】
いくつかの実施形態において、１つ以上の計算された活性化スコアにおける活性化スコアは、それぞれの化合物に対応するそれぞれの細胞構成要素モジュールについてのそれぞれの活性化重みである。例えば、いくつかの実施形態において、活性化スコアは、図２Ａ～図２Ｂに記載され、図５の活性化データ構造に示されるように得られた活性化重みであり、活性化スコアは、それぞれの（例えば、第１の）細胞構成要素モジュールの活性化（例えば、誘導及び／又は差次的発現）を示し、それぞれの化合物による処置に相関及び／又は応答している。
【０３４８】
「モデルアーキテクチャ」と題された上記のセクションに開示されるものなどのモデルの任意の好適な実施形態が企図され、当業者には明らかであろうように、それらの任意の置換、修飾、追加、削除、及び／又は組み合わせが企図される。例えば、いくつかの実施形態において、訓練されたモデルは、ニューラルネットワークを含む。いくつかの実施形態において、ニューラルネットワークは、ＲｅＬＵ活性化を有する完全に接続されたニューラルネットワークである。いくつかの実施形態において、ニューラルネットワークは、メッセージパッシングニューラルネットワークである。いくつかの実施形態において、訓練されたモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。
【０３４９】
いくつかの実施形態において、訓練されたモデルは、複数のコンポーネントモデルのアンサンブルモデルであり、それぞれの計算された活性化スコアは、複数のコンポーネントモデルにおけるコンポーネントモデルの各々の出力の中心傾向の測定値である。いくつかの実施形態において、複数のコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。いくつかの実施形態

において、複数のコンポーネントモデルは、複数のニューラルネットワークを含む。いくつかの実施形態において、複数のニューラルネットワークにおける第１のニューラルネットワークは、ＲｅＬＵ活性化を伴う完全に接続されたニューラルネットワークであり、複数のニューラルネットワークにおける第２のニューラルネットワークは、メッセージパッシングニューラルネットワークである。

【０３５０】

ブロック８０８を参照すると、訓練は、差に応答して訓練されていないモデルと関連付けられた複数のパラメータを調整し、複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を目的の生理学的状態と関連付ける訓練されたモデルを得る。

【０３５１】

いくつかの実施形態において、モデルへの入力は、複数の活性化スコアを含み、それぞれの活性化スコアの各々は、複数の化合物における化合物の各々について、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールに対応する。それぞれの化合物の各々についてのそれぞれの細胞構成要素モジュールの各々に対応する活性化スコアは、モジュールと化合物との間の関連性（例えば、重み及び／又は相関）を識別するためにマルチタスクモデルを訓練するための標識（例えば、モジュールと化合物との間の関連性の実際の存在又は不在を示す数値的活性化スコア）として機能する。例えば、上述のように、いくつかの実施形態において、複数の細胞構成要素モジュールの第１のサブセットは、目的の生理学的状態に関連付けられ、複数の細胞構成要素モジュールの第２のサブセットは、目的の生理学的状態に関連付けられていない。したがって、いくつかのそのような実施形態において、関連性の実際の存在は、複数の細胞構成要素モジュールの第１のサブセットを標識として使用して訓練データセットに含めることができ、関連性の実際の不在は、複数の細胞構成要素モジュールの第２のサブセットを標識として使用して訓練データセットに含めることができる。

【０３５２】

いくつかの実施形態において、複数の化合物は、少なくとも５個、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも８００個、少なくとも１０００個、少なくとも２０００個、少なくとも３０００個、少なくとも４０００個、少なくとも５０００個、少なくとも８０００個、少なくとも１０，０００個、少なくとも２０，０００個、少なくとも３０，０００個、少なくとも５０，０００個、少なくとも８０，０００個、少なくとも１００，０００個、少なくとも２００，０００個、少なくとも５００，０００個、少なくとも８００，０００個、少なくとも１００万個、又は少なくとも２００万個の化合物を含み、複数の化合物における化合物の各々について、モデルへの入力は、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの活性化スコアを含む。

【０３５３】

いくつかの実施形態において、複数の化合物は、１０００万個以下、５００万個以下、１００万個以下、５００，０００個以下、１００，０００個以下、５０，０００個以下、１０，０００個以下、８０００個以下、５０００個以下、２０００個以下、１０００個以下、８００個以下、５００個以下、２００個以下、又は１００個以下の化合物を含み、複数の化合物における化合物の各々について、モデルへの入力は、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの活性化スコアを含む。いくつかの実施形態において、複数の化合物は、１０～５００個、１００～１０，０００個、５０００～２００，０００個、又は１０，０００～１００万個の化合物からなり、複数の化合物における化合物の各々について、モデルへの入力は、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの活性化スコアを含む。

【０３５４】
いくつかの実施形態において、上述したように、複数の数値的活性化スコアにおけるそれぞれの数値的活性化スコアは、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、複数の化合物における化合物の各々についての活性化重みである（例えば、図５の活性化データ構造に示される）。
【０３５５】
上記のように、いくつかの実施形態において、モデルの出力は、複数の化合物におけるそれぞれの化合物（例えば、試験化学化合物）が、複数の細胞構成要素モジュールにおけるそれぞれの１つ以上の細胞構成要素モジュールと相関するかどうかを示す１つ以上の計算された活性化スコアを含む。
【０３５６】
一般に、モデル（例えば、ニューラルネットワーク）を訓練することは、逆伝搬（例えば、勾配降下）を通してそれぞれのモデルについての複数のパラメータ（例えば、重み）を更新することを含む。第一に、入力データ（例えば、複数のモジュールにおけるそれぞれの細胞構成要素モジュールの各々について、複数の化合物におけるそれぞれの化合物の各々についての複数の活性化スコア）がニューラルネットワークに受け入れられ、選択された活性化関数及びパラメータの初期セット（例えば、重み及び／又はハイパーパラメータ）に基づいて出力が計算される、順方向伝搬が実施される。いくつかの実施形態において、パラメータ（例えば、重み及び／又はハイパーパラメータ）は、訓練されていないか、又は部分的に訓練されたモデルに対してランダムに割り当てられる（例えば、初期化される）。いくつかの実施形態において、パラメータは、以前に保存された複数のパラメータから、又は事前に訓練されたモデルから（例えば、転移学習によって）転送される。
【０３５７】
次いで、後方パスが、各層におけるそれぞれのユニットの各々に対応するそれぞれのパラメータの各々についての誤差勾配を計算することによって実施され、各パラメータについての誤差は、ネットワーク出力（例えば、計算された活性化スコアとしてのそれぞれの化合物とそれぞれの細胞構成要素モジュールとの間の関連性の予測された不在又は存在）及び入力データ（例えば、期待値又は真の標識、数値的活性化スコアとしてのそれぞれの化合物とそれぞれの細胞構成要素モジュールとの間の関連性の実際の不在又は存在）に基づいて損失（例えば、誤差）を計算することによって決定される。次いで、パラメータ（例えば、重み）は、計算された損失に基づいて値を調整することによって更新され、それによってモデルを訓練する。
【０３５８】
例えば、機械学習のいくつかの一般的な実施形態において、逆伝搬は、複数の重み（例えば、埋め込み）を含む隠れ層を有するネットワークを訓練する方法である。訓練されていないモデルの出力（例えば、計算された活性化スコアとしての関連性の予測された不在又は存在）は、最初に任意に選択された初期重みのセットを使用して生成される。次いで、（例えば、損失関数を使用して）誤差を計算するために誤差関数を評価することによって、出力を元の入力（例えば、数値的活性化スコアとしての関連性の実際の不在又は存在）と比較する。次いで、重みは、（例えば損失関数に従って）誤差が最小化されるように更新される。いくつかの実施形態において、当業者には明らかであろうように、様々な逆伝搬アルゴリズム及び／又は方法のいずれか１つが、複数の重みを更新するために使用される。
【０３５９】
いくつかの実施形態において、損失関数は、平均平方誤差、二次損失、平均絶対誤差、平均バイアス誤差、ヒンジ、マルチクラスサポートベクトルマシン、及び／又は交差エントロピーである。いくつかの実施形態において、訓練されていないか、又は部分的に訓練されたモデルを訓練することは、勾配降下アルゴリズム及び／又は最小化関数に従って誤差を計算することを含む。いくつかの実施形態において、訓練されていないか、又は部分的に訓練されたモデルを訓練することは、複数の損失関数を使用して複数の誤差を計算す

ることを含む。いくつかの実施形態において、複数の損失関数における損失関数の各々は、同じ又は異なる重み付け係数を受け取る。

【０３６０】

図６は、本開示のいくつかの実施形態による、モデルを訓練するための方法の例を示す。活性化データ構造（上部パネル）は、複数のＫ細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々と、複数のＧ細胞における細胞の各々との間の関連性を示す複数の活性化スコアを含むモデルへの入力を提供し、細胞の各々は、複数の化合物におけるそれぞれの化合物を表す。複数の細胞構成要素モジュール（中央パネル）におけるそれぞれの細胞構成要素モジュールの各々について、訓練の前に、複数の細胞（例えば、Ｗ化合物）によって集合的に表される複数の化合物におけるそれぞれの化合物の各々について、対応する重みを初期化する（例えば、ランダムな重みにする）。したがって、複数の化合物重みは、化合物重みマトリックス（中央パネル）を含む。複数の化合物重みの調整は、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの細胞構成要素モジュールについてのそれぞれの計算された活性化スコア（例えば、予測）と、（ｉｉ）細胞構成要素モジュールのセットにおけるそれぞれの化合物についてのそれぞれの細胞構成要素モジュールのそれぞれの数値的活性化スコア（例えば、実際）との間の差を使用して実施される（下部パネル）。いくつかの実施形態において、実際の活性化は、例えば、図２Ａ～図２Ｂ及び図１４Ａ～図１４Ｄを参照して、以下の「細胞構成要素モジュールの識別」と題されるセクションに記載の細胞構成要素モジュールを識別するための方法を使用して得られ、複数の共変量は複数の化合物を含む。次いで、訓練（例えば、化合物重みの調整）は、訓練されたモデルが形成されるまで（例えば、最小数の調整の完了及び／又は最小性能閾値の達成によって）実施することができる。

【０３６１】

いくつかの実施形態において、誤差関数は、計算された損失に比例する量によって１つ以上のパラメータの値を調整することによって、モデル（例えば、ニューラルネットワーク）における１つ以上のパラメータ（例えば、重み）を更新するために使用され、それによってモデルを訓練する。いくつかの実施形態において、パラメータが調整される量は、パラメータが更新される程度又は重大度（例えば、より小さい又はより大きい調整）を指示する学習率ハイパーパラメータによって計測される。したがって、いくつかの実施形態において、訓練は、学習率に基づいて、複数のパラメータの全て又はサブセットを更新する。いくつかの実施形態において、学習率は、差次的学習率である。

【０３６２】

いくつかの実施形態において、モデル（例えば、ニューラルネットワーク）を訓練することは、対応する複数の隠れニューロンにおける隠れニューロンの各々の対応するパラメータに対する正規化を更に使用する。例えば、いくつかの実施形態において、正規化は、損失関数にペナルティを追加することによって実施され、ペナルティは、ニューラルネットワークにおけるパラメータの値に比例する。一般に、正規化は、１つ以上のパラメータにペナルティを追加することによってモデルの複雑性を低減し、それらのパラメータと関連付けられたそれぞれの隠れニューロンの重要性を低下させる。そのような実践は、より一般化されたモデルをもたらし、データの過剰適合を低減することができる。いくつかの実施形態において、正規化は、Ｌ１又はＬ２ペナルティを含む。例えば、いくつかの好ましい実施形態において、正規化は、より低い及びより高いパラメータに対するＬ２ペナルティを含む。いくつかの実施形態において、正規化は、空間正規化（例えば、先験的及び／又は実験的知識に基づいて決定される）又はドロップアウト正規化を含む。いくつかの実施形態において、正規化は、独立して最適化されるペナルティを含む。

【０３６３】

いくつかの実施形態において、モデルに関連付けられた複数の化合物重みを調整すること（例えば、予測された標識と実際の標識との間の差に応答する）を含む訓練プロセスは、複数の訓練インスタンスにおける訓練インスタンスの各々に対して繰り返される。

【０３６４】
いくつかの実施形態において、複数の訓練インスタンスは、少なくとも３、少なくとも４、少なくとも５、少なくとも６、少なくとも７、少なくとも８、少なくとも９、少なくとも１０、少なくとも５０、少なくとも１００、少なくとも５００、少なくとも１０００、少なくとも２０００、少なくとも３０００、少なくとも４０００、少なくとも５０００、又は少なくとも７５００の訓練インスタンスを含む。いくつかの実施形態において、複数の訓練インスタンスは、１０，０００以下、５０００以下、１０００以下、５００以下、１００以下、又は５０以下の訓練インスタンスを含む。いくつかの実施形態において、複数の訓練インスタンスは、３～１０、５～１００、１００～５０００、又は１０００～１０，０００の訓練インスタンスを含む。いくつかの実施形態において、複数の訓練インスタンスは、３以上の訓練インスタンスから始まり、１０，０００以下の訓練インスタンスで終わる別の範囲内にある。

【０３６５】
いくつかのそのような実施形態において、訓練は、複数の訓練インスタンスにわたって（例えば、逆伝搬を介して）モデルのパラメータの調整を繰り返すことを含み、したがって、それぞれの化合物がそれぞれの細胞構成要素モジュールと相関するかどうかを示す際のモデルの精度を増加させる。

【０３６６】
いくつかの実施形態において、訓練は、転移学習を含む。転移学習は、例えば、定義のセクションに更に記載されている（上記の「訓練されていないモデル」を参照されたい）。

【０３６７】
いくつかの実施形態において、訓練されていないか、又は部分的に訓練されたモデルを訓練することは、誤差関数の第１の評価に続いて訓練されたモデルを形成する。いくつかのそのような実施形態において、訓練されたモデルは、誤差関数の第１の評価に基づいて、１つ以上のパラメータの第１の更新に続いて形成される。いくつかの代替の実施形態において、訓練されたモデルは、誤差関数の少なくとも１回、少なくとも２回、少なくとも３回、少なくとも４回、少なくとも５回、少なくとも６回、少なくとも７回、少なくとも８回、少なくとも９回、少なくとも１０回、少なくとも２０回、少なくとも３０回、少なくとも４０回、少なくとも５０回、少なくとも１００回、少なくとも５００回、少なくとも１０００回、少なくとも１０，０００回、少なくとも５０，０００回、少なくとも１００，０００回、少なくとも２００，０００回、少なくとも５００，０００回、又は少なくとも１００万回の評価に続いて形成される。いくつかのそのような実施形態において、訓練されたモデルは、誤差関数の少なくとも１回、少なくとも２回、少なくとも３回、少なくとも４回、少なくとも５回、少なくとも６回、少なくとも７回、少なくとも８回、少なくとも９回、少なくとも１０回、少なくとも２０回、少なくとも３０回、少なくとも４０回、少なくとも５０回、少なくとも１００回、少なくとも５００回、少なくとも１０００回、少なくとも１０，０００回、少なくとも５０，０００回、少なくとも１００，０００回、少なくとも２００，０００回、少なくとも５００，０００回、又は少なくとも１００万回の評価に基づいて、１つ以上のパラメータの少なくとも１回、少なくとも２回、少なくとも３回、少なくとも４回、少なくとも５回、少なくとも６回、少なくとも７回、少なくとも８回、少なくとも９回、少なくとも１０回、少なくとも２０回、少なくとも３０回、少なくとも４０回、少なくとも５０回、少なくとも１００回、少なくとも５００回、少なくとも１０００回、少なくとも１０，０００回、少なくとも５０，０００回、少なくとも１００，０００回、少なくとも２００，０００回、少なくとも５００，０００回、又は少なくとも１００万回の更新に続いて形成される。

【０３６８】
いくつかの実施形態において、訓練されたモデルは、モデルが最小性能要件を満たす場合に形成される。例えば、いくつかの実施形態において、訓練されたモデルについて計算された誤差が、誤差関数の評価（例えば、各化合物と各細胞構成要素モジュールとの間の

予測された関連性と実際の関連性との間の差）に続いて、誤差閾値を満たす場合に、訓練されたモデルが形成される。いくつかの実施形態において、誤差関数によって計算される誤差は、誤差が２０パーセント未満、１８パーセント未満、１５パーセント未満、１０パーセント未満、５パーセント未満、又は３パーセント未満である場合に誤差閾値を満たす。

【０３６９】

例示的な実施形態において、モデルを訓練することは、マルチタスク策定におけるカテゴリ交差エントロピー損失を使用して実施され、複数の共変量における共変量の各々が、複数のコスト関数におけるコスト関数に対応し、複数のコスト関数におけるそれぞれのコスト関数の各々が、共通の重み付け係数を有する。

【０３７０】

いくつかの実施形態において、訓練することは、回帰モデルに従って、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々についてのそれぞれの化合物の各々と関連付けられた差の各々に応答して、訓練されていないモデルと関連付けられた複数のパラメータを調整する。いくつかの実施形態において、回帰モデルは、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々についてのそれぞれの化合物の各々と関連付けられた差の各々の最小二乗誤差を最適化する。

【０３７１】

モデル訓練の前述の説明は、化合物と細胞構成要素モジュールとの間の関連性を示す活性化スコアを得て、使用することを記載しているが、実際には、化合物と目的の任意の他の生理学的状態、又はその任意の細胞プロセスとの間の関連性を示す活性化スコアは、化合物を生理学的状態と関連付けるための訓練及びモデルの使用において企図されている。例えば、以下のセクションで説明されるように、本開示の別の態様は、摂動シグネチャを使用してモデルを訓練することを含む。具体的には、いくつかの実施形態において、モデルは、訓練標識として摂動シグネチャのための数値的活性化スコアを使用して訓練される。次いで、訓練されたモデルは、方法７００において記載されるように、出力として、モデルへの化学構造フィンガープリントの入力に応答して、１つ以上の計算された活性化スコアを得るために使用され、１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、摂動シグネチャのセットにおける対応する摂動シグネチャを表す。

【０３７２】

摂動シグネチャについての数値的活性化スコアの取得。

したがって、本開示の別の態様は、化学化合物を目的の生理学的状態と関連付けるための方法９００を提供する。いくつかの実施形態において、目的の生理学的状態は、疾患である。

【０３７３】

ブロック９０２を参照すると、方法は、複数の化合物における化合物の各々の化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得る。「生理学的状態」及び「化合物」と題する上記のセクションに開示されるような、生理学的状態、化合物、フィンガープリント、及び／又はフィンガープリントを得る方法の任意の好適な実施形態は、当業者には明白であろうように、それらの任意の置換、修飾、追加、欠失、及び／又は組み合わせを含むことが企図される。

【０３７４】

例えば、いくつかの実施形態において、複数の化合物は、１０～1×10^{6}個の化合物である。いくつかの実施形態において、複数の化合物は、１００～１００，０００個の化合物である。いくつかの実施形態において、複数の化合物は、１０００～１００，０００個の化合物である。いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、２０００ダルトン未満の分子量を有する有機化合物である。いくつかの実施形態において、複数の化学化合物における化学化合物の各々は、５つの基準のリピンスキーの法則の各々を満たす。いくつかの実施形態において、複数の化学化合物における化学化

合物の各々は、５つの基準のリピンスキーの法則のうちの少なくとも３つの基準を満たす。いくつかの実施形態において、それぞれのフィンガープリントの各々は、ＳＭＩＬＥＳＴｒａｎｓｆｏｒｍｅｒ、ＥＣＦＰ４、ＲＮＮＳ２Ｓ、又はＧｒａｐｈＣｏｎｖを使用して、化学構造から生成される。

【０３７５】

ブロック９０４を参照すると、方法は、複数の化合物における対応する化合物の各々についての摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々のそれぞれの数値的活性化スコアを電子形式で得ることを含む。「摂動シグネチャ」と題された上記のセクションに開示される摂動シグネチャの任意の好適な実施形態は、当業者に明白であろうように、それらの任意の置換、修正、追加、欠失、及び／又は組み合わせを含むことが企図される。

【０３７６】

例えば、いくつかの実施形態において、摂動シグネチャのセットは、単一の摂動シグネチャである。いくつかの実施形態において、摂動シグネチャのセットは、複数の摂動シグネチャである。いくつかの実施形態において、摂動シグネチャのセットは、２００～５００個の摂動シグネチャである。いくつかの実施形態において、摂動シグネチャのセットは、５つ以上の摂動シグネチャを含む。いくつかの実施形態において、摂動シグネチャのセットは、１０個以上の摂動シグネチャを含む。いくつかの実施形態において、摂動シグネチャのセットは、１００個以上の摂動シグネチャを含む。いくつかの実施形態において、摂動シグネチャのセットは、複数の摂動シグネチャであり、複数の摂動シグネチャの第１のサブセットは、目的の生理学的状態と関連付けられ、複数の摂動シグネチャの第２のサブセットは、目的の生理学的状態と関連付けられていない。

【０３７７】

ブロック９０６を参照すると、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々が、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、それぞれの第１の細胞状態及び第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である。

【０３７８】

いくつかの実施形態において、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアは、変化していない細胞状態と、変化した細胞状態との間の差次的な細胞構成要素存在量の測定値を表す単一細胞遷移シグネチャに電子形式でアクセスすることを含む手順によって得られる。変化した細胞状態は、変化していない細胞状態から変化した細胞状態への細胞遷移を通じて生じ、（ｉ）変化していない細胞状態、（ｉｉ）変化した細胞状態、及び（ｉｉｉ）変化していない細胞状態から変化した細胞状態への遷移のうちの少なくとも１つは、目的の生理学的状態と関連付けられる。単一細胞遷移シグネチャは、参照の複数の細胞構成要素の識別と、複数の参照細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、変化していない細胞状態と変化した細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する第１の有意性スコアとを含む。それぞれの摂動シグネチャのそれぞれの数値的活性化スコアを決定するために、単一細胞遷移シグネチャ及びそれぞれの摂動シグネチャを比較する。

【０３７９】

いくつかの実施形態において、単一細胞遷移シグネチャと摂動シグネチャとを比較して、それぞれの摂動シグネチャのそれぞれの数値的活性化スコアを決定することは、単一細胞遷移シグネチャの参照の複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの摂動シグネチャにおける対応する細胞構成要素の対応する有意性スコ

アに対する単一細胞遷移シグネチャにおけるそれぞれの細胞構成要素の第１の有意性スコアを比較することを含む。

【０３８０】

いくつかの実施形態において、単一細胞遷移シグネチャと摂動シグネチャとを比較して、それぞれの摂動シグネチャのそれぞれの数値的活性化スコアを決定することは、単一細胞遷移シグネチャの参照の複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの摂動シグネチャにおける複数の細胞構成要素における対応する細胞構成要素の各々の対応する有意性スコアに対する単一細胞遷移シグネチャにおける複数の参照細胞構成要素におけるそれぞれの細胞構成要素の各々の有意性スコアを比較することを含む。

【０３８１】

いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、摂動シグネチャのセットにおける他の摂動シグネチャと比較して、単一細胞遷移シグネチャに対するそれぞれの摂動シグネチャの関連性の相対的なランキングである。いくつかの実施形態において、相対的なランキングは、ウィルコクソンの順位和検定、ｔ検定、ロジスティック回帰、又は一般化線形モデルによって決定される。いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、ランキングに基づいていない。

【０３８２】

いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、それぞれの摂動シグネチャについてのそれぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々についての、対応する有意性スコアの中心傾向の測定値である。いくつかの実施形態において、中心傾向の測定値は、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々についての対応する有意性スコアの算術平均、加重平均、ミッドレンジ、ミッドヒンジ、トリミアン、ウィンザライズド平均、平均、又はモードである。

【０３８３】

いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、（ｉ）それぞれの摂動シグネチャについての、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々についての、対応する有意性スコアの中心傾向の測定値と、（ｉｉ）単一細胞遷移シグネチャについての、複数の参照細胞構成要素におけるそれぞれの細胞構成要素の各々についての、対応する第１の有意性スコアの中心傾向の測定値との間の差である。

【０３８４】

いくつかの実施形態において、単一細胞遷移シグネチャの変化していない細胞状態が、それぞれの摂動シグネチャの第１の細胞状態又は第２の細胞状態と同じである。いくつかの実施形態において、単一細胞遷移シグネチャの変化していない細胞状態が、それぞれの摂動シグネチャの第１の細胞状態及び第２の細胞状態の両方とは異なる。

【０３８５】

いくつかの実施形態において、方法は、単一細胞遷移シグネチャの参照の複数の細胞構成要素、及びそれぞれの摂動シグネチャのそれぞれの複数の細胞構成要素を剪定して、転写因子と比較することを制限することを更に含む。いくつかの実施形態において、方法は、単一細胞遷移シグネチャの参照の複数の細胞構成要素、及びそれぞれの摂動シグネチャのそれぞれの複数の細胞構成要素を剪定して、別の細胞構成要素の種類（例えば、遺伝子、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、及び／又はそれらの組み合わせ）との比較を制限することを更に含む。いくつかの実施形態において、参照の複数の細胞構成要素及びそれぞれの複数の細胞構成要素は、剪定されない。

【０３８６】

いくつかの実施形態において、複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの摂動状態は、複数の化合物における化合物に曝露されていない対照細胞によって表される。いくつかの実施形態において、複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの摂動状態は、それぞれの摂動シグネチャに関連付けられた化合物以外の複数の化

学化合物における化学化合物に曝露されている無関係の摂動細胞にわたる平均によって表される。

【０３８７】

上述のように、いくつかの実施形態において、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャは、非限定的な例として、参照により本明細書に組み込まれる、２０１９年７月１５日に出願された「Ｍｅｔｈｏｄｓ　ｏｆ　Ａｎａｌｙｚｉｎｇ　Ｃｅｌｌｓ」と題された米国特許出願第１６／５１１，６９１号に開示された方法のいずれかを使用して決定され得る。

【０３８８】

それぞれの摂動シグネチャは、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含む。それぞれの第１の細胞状態及び第２の細胞状態のうちの一方は、非摂動細胞状態であり、他方は、それぞれの摂動シグネチャに対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である。更に、上述したように、それぞれの摂動シグネチャは、数値的活性化スコアを含む。いくつかの実施形態において、それぞれの摂動シグネチャについての数値的活性化スコアは、連続スケール上の絶対値である。いくつかの実施形態において、それぞれの摂動シグネチャについての数値的活性化スコアは、以下でより詳細に説明されるように、相対的なランキングである。

【０３８９】

いくつかの実施形態において、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアは、変化していない細胞状態と、変化した細胞状態との間の差次的な細胞構成要素存在量の測定値を表す単一細胞遷移シグネチャに電子形式でアクセスすることを含む手順によって得られる。ここで、変化した細胞状態は、変化していない細胞状態から変化した細胞状態への細胞遷移を通して発生する。更に、（ｉ）変化していない細胞状態、（ｉｉ）変化した細胞状態、及び（ｉｉｉ）変化していない細胞状態から変化した細胞状態への遷移のうちの少なくとも１つが、目的の生理学的状態と関連付けられる。

【０３９０】

単一細胞遷移シグネチャは、参照の複数の細胞構成要素の識別と、複数の参照細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、変化していない細胞状態と変化した細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する第１の有意性スコアとを含む。いくつかの実施形態において、単一細胞遷移シグネチャは、参照により本明細書に組み込まれる、２０１９年７月１５日に出願された「Ｍｅｔｈｏｄｓ　ｏｆ　Ａｎａｌｙｚｉｎｇ　Ｃｅｌｌｓ」と題される米国特許出願第１６／５１１，６９１号に開示される方法のいずれかを使用して決定される。

【０３９１】

一度得られると、単一細胞遷移シグネチャは、それぞれの摂動シグネチャと比較され、それによってそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアを決定する。いくつかの実施形態において、２０１９年７月１５日に出願された「Ｍｅｔｈｏｄｓ　ｏｆ　Ａｎａｌｙｚｉｎｇ　Ｃｅｌｌｓ」と題された米国特許出願第１６／５１１，６９１号に開示された、単一細胞遷移シグネチャをそれぞれの摂動シグネチャと比較して、複数の摂動シグネチャにおける他の摂動シグネチャに対して、それぞれの摂動シグネチャの相対的なランキングを決定するための方法のいずれかを使用することができ、例えば、そのような相対的なランキングは、次いで、それぞれの摂動シグネチャのそれぞれの数値的な活性化スコアとみなされるであろう。

【０３９２】

いくつかの実施形態において、単一細胞遷移シグネチャ及び摂動シグネチャを比較して

、それぞれの摂動シグネチャのそれぞれの数値的活性化スコアを決定することは、単一細胞遷移シグネチャの参照の複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の第１の有意性スコアを、それぞれの摂動シグネチャにおける対応する細胞構成要素の対応する有意性スコアと比較することを含む。いくつかのそのような実施形態において、それぞれの摂動シグネチャの活性化スコアは、摂動シグネチャのセットにおける他の摂動シグネチャと比較して、単一細胞遷移シグネチャに対するそれぞれの摂動シグネチャの関連性の相対的なランキングである。いくつかのそのような実施形態において、相対的なランキングは、ウィルコクソンの順位和検定、ｔ検定、ロジスティック回帰、又は一般化線形モデルによって決定される。いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、それぞれの摂動シグネチャの関連性の相対的なランキングではなく、むしろ、単一細胞遷移シグネチャに対する他の摂動シグネチャのランキングとは独立して決定される。

【０３９３】

いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、ランキングに基づいていない。例えば、いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、それぞれの摂動シグネチャについてのそれぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々についての対応する有意性スコアを含む複数の有意性スコアである。

【０３９４】

いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、それぞれの摂動シグネチャについてのそれぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々についての、対応する有意性スコアの中心傾向の測定値である。いくつかの実施形態において、中心傾向の測定値は、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々についての対応する有意性スコアの算術平均、加重平均、ミッドレンジ、ミッドヒンジ、トリミアン、ウィンザライズド平均、平均、又はモードである。

【０３９５】

いくつかの実施形態において、それぞれの摂動シグネチャの活性化スコアは、（ｉ）それぞれの摂動シグネチャについての、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々についての、対応する有意性スコアの中心傾向の測定値と、（ｉｉ）単一細胞遷移シグネチャについての、複数の参照細胞構成要素におけるそれぞれの細胞構成要素の各々についての、対応する第１の有意性スコアの中心傾向の測定値との間の差である。

【０３９６】

一実施形態において、単一細胞遷移シグネチャとそれぞれの摂動シグネチャとの間の比較を実施するために、摂動シグネチャの細胞構成要素は、マトリックスとして表される。マトリックスの各行は、単一の摂動（例えば、複数の化合物における単一の化合物）と関連付けられる。マトリックス上の各列は、それぞれの状態間の差次的な存在量を示す細胞構成要素のうちの１つと関連付けられる。マトリックスの各エントリーは、特定の摂動シグネチャについて識別された細胞構成要素についての有意性スコア（例えば、ｐ値、ｔスコア）を含む。このマトリックスは、単一細胞遷移シグネチャにある細胞構成のみを含むようにフィルタリングされる。このフィルタリングは、閾値ｐ値、細胞構成要素の閾値数の使用などを使用して達成され得る。

【０３９７】

マトリックス内の各有意性スコアは、個別のマッチングスコアと置き換えられる。各有意性スコアを個別のマッチングスコアと置き換えるために、細胞遷移についての有意に上方制御された細胞構成要素及び細胞遷移についての有意に下方制御された細胞構成要素を識別する。単一細胞遷移シグネチャによって識別される有意に上方制御された細胞構成要素の各々について、細胞構成要素がその摂動（例えば、化学組成物）についての摂動シグネチャについても有意に上方制御されている場合、その細胞構成要素／摂動の組み合わせについてのマトリックスにおける有意性スコアは、「１」の個別のマッチングスコアと置

き換えられる。細胞構成要素が単一細胞遷移シグネチャと比較して摂動シグネチャに対して有意に下方制御されている場合、その細胞構成要素／摂動の組み合わせについてのマトリックスにおける有意性スコアは、「－２」の個別のマッチングスコアと置き換えられる。細胞構成要素が摂動シグネチャに対して有意に上方制御又は下方制御されていない場合、細胞構成要素／摂動の組み合わせについてのマトリックスにおける有意性スコアは、「０」の個別のマッチングスコアと置き換えられる。

【０３９８】

逆に、単一細胞遷移シグネチャにおいて識別された有意に下方制御された細胞構成要素の各々について、細胞構成要素が摂動についても有意に下方制御されている場合、その細胞構成要素／摂動の組み合わせについてのマトリックスにおける有意性スコアは、「－１」の個別のマッチングスコアと置き換えられる。細胞構成要素が摂動に対して有意に上方制御されている場合、その細胞構成要素／摂動の組み合わせについてのマトリックスにおける有意性スコアは、「２」の個別のマッチングスコアと置き換えられる。細胞構成要素が摂動シグネチャに対して有意に上方制御又は下方制御されていない場合、その細胞構成要素／摂動の組み合わせについてのマトリックスにおける有意性スコアは、「０」の個別のマッチングスコアと置き換えられる。当業者は、いくつかの実施形態において、これらの特定のスコア置換が他の数値で置換され得ることを理解するであろう。更に、上方制御又は下方制御の代わりに、細胞構成要素の各々についての閾値存在量値の使用が使用され得、次いで、所与の細胞構成要素が閾値存在量値を上回るか、又は下回るかどうかの考慮が、前述のクラス標識（例えば、「－１」、「２」、「０」など）をマトリックスの各エレメントに割り当てる際に行われる。

【０３９９】

結果は、摂動の数（複数の化学組成物における化学組成物の数、したがって複数の摂動シグネチャにおける摂動シグネチャの数）によって与えられる行の数と、上記のマトリックスエレメントエントリーがマッチングスコアを表す単一細胞遷移からの差次的細胞構成要素によって与えられる列の数とのマトリックスである。

【０４００】

上記のように、マトリックス内の有意性スコアを個別のマッチングスコアに置き換えた後、マトリックスの各行における個別のマッチングスコアを合計して、各行についての合計されたマッチングスコアを生成する。次いで、各々が摂動シグネチャに対応するマトリックスの行は、合計したマッチングスコアを減少させる順序でランク付けされる。最上位の行は、単一細胞遷移シグネチャの識別された細胞遷移と関連付けられる可能性が最も高い摂動シグネチャと関連付けられる。更に、行の各々のランキングは、行の各々に対応する摂動シグネチャについての活性化スコアとして使用することができる。

【０４０１】

いくつかの実施形態において、マトリックスにおける各行の合計マッチングスコアについて、偽の細胞構成要素発見率の推定は、参照により本明細書に組み込まれる、２０１９年７月１５日に出願された「Ｍｅｔｈｏｄｓ　ｏｆ　Ａｎａｌｙｚｉｎｇ　Ｃｅｌｌｓ」と題された米国特許出願第１６／５１１，６９１号で説明されるように推定される。

【０４０２】

ある特定の実施形態において、摂動（例えば、特定の化学組成物への細胞の曝露）の共変量が存在し得る。例えば、化学組成物の共変量は、化学組成物の特定の用量、化学組成物に曝露された細胞が細胞構成要素を定量化するために測定される時間、及び／又は化学組成物に曝露された細胞の同一性（例えば、細胞株）を含み得る。いくつかの実施形態において、摂動（例えば、特定の化学組成物への細胞の曝露）は、その共変量の閾値量も特定の細胞遷移に影響すると予測される場合にのみ、特定の細胞遷移に影響すると予測される。言い換えれば、いくつかの実施形態において、特定の摂動シグネチャの数値的活性化スコアは、特定の摂動シグネチャの化学組成物の共変量が、単一細胞遷移スコアと関連付けられた特定の細胞遷移にも影響を与えると予測されるかどうかによって少なくとも部分的に決定される。

【０４０３】
それぞれの摂動シグネチャを単一細胞遷移シグネチャと比較する代替方法を使用して、それぞれの摂動シグネチャの数値的活性化スコアを決定し得る。例えば、細胞構成要素は、ウェブインターフェースを使用してデータベースに適合され得る（例えば、ａｍｐ．ｐｈａｒｍ．ｍｓｓｍ．ｅｄｕ／Ｌ１０００ＣＤＳ２／＃／ｉｎｄｅｘのワールドワイドウェブ上のＬ１０００ＣＤＳ２．Ａｎ　ｕｌｔｒａ－ｆａｓｔ　ＬＩＮＣＳ　Ｌ１０００　Ｃｈａｒａｃｔｅｒｉｓｔｉｃ　Ｄｉｒｅｃｔｉｏｎ　Ｓｉｇｎａｔｕｒｅ　Ｓｅａｒｃｈ　Ｅｎｇｉｎｅなど）。
【０４０４】
いくつかの実施形態において、単一細胞遷移シグネチャの変化していない細胞状態が、それぞれの摂動シグネチャの第１の細胞状態又は第２の細胞状態と同じである。いくつかの実施形態において、単一細胞遷移シグネチャの変化していない細胞状態が、それぞれの摂動シグネチャの第１の細胞状態及び第２の細胞状態の両方とは異なる。
【０４０５】
いくつかの実施形態において、方法は、単一細胞遷移シグネチャの参照の複数の細胞構成要素、及びそれぞれの摂動シグネチャのそれぞれの複数の細胞構成要素を剪定して、転写因子と比較することを制限することを更に含む。いくつかの実施形態において、複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの摂動状態は、複数の化合物における化合物に曝露されていない対照細胞によって表される。
【０４０６】
いくつかの実施形態において、複数の摂動シグネチャにおけるそれぞれの摂動シグネチャの摂動状態は、それぞれの摂動シグネチャに関連付けられた化合物以外の複数の化学化合物における化学化合物に曝露されている無関係の摂動細胞にわたる平均によって表される。
【０４０７】
ブロック９０８を参照すると、方法は、複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々について、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの摂動シグネチャについてのそれぞれの計算された活性化スコアと、（ｉｉ）摂動シグネチャのセットにおける対応する化合物についてのそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアとの間のそれぞれの差を使用して訓練されていないモデルを訓練することを更に含む。
【０４０８】
「モデルアーキテクチャ」と題された上記のセクションに開示されるものなどのモデルの任意の好適な実施形態が企図され、当業者には明らかであろうように、それらの任意の置換、修飾、追加、削除、及び／又は組み合わせが企図される。例えば、いくつかの実施形態において、訓練されたモデルは、ニューラルネットワークを含む。いくつかの実施形態において、ニューラルネットワークは、ＲｅＬＵ活性化を有する完全に接続されたニューラルネットワークである。いくつかの実施形態において、ニューラルネットワークは、メッセージパッシングニューラルネットワークである。いくつかの実施形態において、訓練されたモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。
【０４０９】
いくつかの実施形態において、訓練されたモデルは、複数のコンポーネントモデルのアンサンブルモデルであり、それぞれの計算された活性化スコアは、複数のコンポーネントモデルにおけるコンポーネントモデルの各々の出力の中心傾向の測定値である。いくつかの実施形態において、複数のコンポーネントモデルは、ロジスティック回帰モデル、ニューラルネットワークモデル、サポートベクトルマシンモデル、ナイーブベイズモデル、最

近傍モデル、ブーストツリーモデル、ランダムフォレストモデル、決定木モデル、多項ロジスティック回帰モデル、線形モデル、又は線形回帰モデルを含む。いくつかの実施形態において、複数のコンポーネントモデルは、複数のニューラルネットワークを含む。いくつかの実施形態において、複数のニューラルネットワークにおける第１のニューラルネットワークは、ＲｅＬＵ活性化を伴う完全に接続されたニューラルネットワークであり、複数のニューラルネットワークにおける第２のニューラルネットワークは、メッセージパッシングニューラルネットワークである。

【０４１０】

ブロック９１０を参照すると、訓練は、差に応答して訓練されていないモデルと関連付けられた複数のパラメータを調整し、複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を目的の生理学的状態と関連付ける訓練されたモデルを得る。

【０４１１】

「モデル訓練」と題された上記のセクションに開示されるものなど、訓練されていないか、又は部分的に訓練されたモデルを訓練するための任意の好適な方法及び実施形態は、当業者に明らかであろうように、それらの任意の置換、修飾、追加、欠失、及び／又は組み合わせを含むことが企図される。

【０４１２】

いくつかの実施形態に関して、モデルへの入力は、複数の活性化スコアを含み、それぞれの活性化スコアの各々は、複数の化合物における化合物の各々について、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャに対応する。それぞれの化合物の各々についてのそれぞれの摂動シグネチャの各々に対応する活性化スコアは、摂動シグネチャと化合物との間の関連性（例えば、重み及び／又は相関）を識別するためにマルチタスクモデルを訓練するための標識（例えば、摂動シグネチャと化合物との間の関連性の実際の存在又は不在を示す数値的活性化スコア）として機能する。例えば、上述のように、いくつかの実施形態において、複数の摂動シグネチャの第１のサブセットは、目的の生理学的状態と関連付けられ、複数の摂動シグネチャの第２のサブセットは、目的の生理学的状態と関連付けられていない。したがって、いくつかのそのような実施形態において、関連性の実際の存在は、複数の摂動シグネチャの第１のサブセットを標識として使用して訓練データセットに含めることができ、関連性の実際の不在は、複数の摂動シグネチャの第２のサブセットを標識として使用して訓練データセットに含めることができる。

【０４１３】

いくつかの実施形態において、訓練することは、回帰モデルに従って、摂動シグネチャｓのセットにおけるそれぞれの摂動シグネチャの各々についての対応する化合物の各々と関連付けられた差の各々に応答して、訓練されていないモデルと関連付けられた複数のパラメータを調整する。いくつかの実施形態において、回帰モデルは、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々についての対応する化合物の各々と関連付けられた差の各々の最小二乗誤差を最適化する。

【０４１４】

いくつかの実施形態において、モデルは、細胞構成要素モジュール、摂動シグネチャ、又は両方についての活性化スコアに基づいて、化合物を目的の生理学的状態と関連付けるために訓練及び／又は使用される。いくつかの実施形態において、モデルは、複数のドメイン（例えば、モジュール及び／又は摂動シグネチャなどの標識タイプ）及び／又はデータタイプ（例えば、遺伝子発現プロファイル、メタボロミクス、プロテオミクス、エピジェネティクスなどの分析物及び／又は細胞構成要素）についての活性化スコアに基づいて、化合物を目的の生理学的状態と関連付けるために訓練及び／又は使用される。いくつかの実施形態において、モデルは、任意の１つ以上の目的の生理学的状態（例えば、化合物の毒性、疾患状態の解消など）についての活性化スコアに基づいて、化合物を目的の生理学的状態と関連付けるために訓練及び／又は使用される。いくつかの実施形態において、モデルは、複数のシステムにわたって訓練され、システムは、本明細書に開示される任意

の１つ以上の生理学的状態、任意の１つ以上のドメイン、及び／若しくは任意の１つ以上のデータタイプ、又は当業者に明白であろう任意の置換、修飾、追加、欠失、及び／又は組み合わせを指す。例えば、いくつかの実施形態において、モデルは、試験化学化合物、毒性の遺伝子モジュール特性の活性化、及び疾患解消を示す摂動シグネチャの間の関連性を集合的に決定するように共同訓練される。

【０４１５】

追加の実施形態。

本開示の別の態様は、１つ以上のプロセッサ及びメモリを含むコンピュータシステムを提供し、メモリは、試験化学化合物を目的の生理学的状態と関連付けるための方法を実施するための命令を格納する。方法は、試験化学化合物の化学構造のフィンガープリントを得ることと、フィンガープリントをモデルに入力することと、を含み、モデルは、１００以上のパラメータを含み、モデルは、フィンガープリントのモデルへの入力に応答して、１つ以上の計算された活性化スコアを出力し、１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、細胞構成要素モジュールのセットにおける対応する細胞構成要素モジュールを表し、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々は、複数の細胞構成要素の独立したサブセットを含み、細胞構成要素モジュールのセットにおける第１の細胞構成要素モジュールは、目的の生理学的状態と関連付けられる。方法は、第１の細胞構成要素モジュールについてのそれぞれの計算された活性化スコアが、第１の閾値基準を満たす場合、化学化合物を目的の生理学的状態と識別することを更に含む。

【０４１６】

本開示の別の態様は、試験化学化合物を目的の生理学的状態と関連付けるための、コンピュータによって実行可能な１つ以上のコンピュータプログラムを格納する非一時的なコンピュータ可読媒体を提供し、コンピュータは、１つ以上のプロセッサ及びメモリを含み、１つ以上のコンピュータプログラムは、方法を実施するコンピュータによって実行可能な命令を集合的に符号化する。方法は、試験化学化合物の化学構造のフィンガープリントを得ることと、フィンガープリントをモデルに入力することと、を含み、モデルは、１００以上のパラメータを含み、モデルは、フィンガープリントのモデルへの入力に応答して、１つ以上の計算された活性化スコアを出力し、１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、細胞構成要素モジュールのセットにおける対応する細胞構成要素モジュールを表し、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々は、複数の細胞構成要素の独立したサブセットを含み、細胞構成要素モジュールのセットにおける第１の細胞構成要素モジュールは、目的の生理学的状態と関連付けられる。方法は、第１の細胞構成要素モジュールについてのそれぞれの計算された活性化スコアが、第１の閾値基準を満たす場合、化学化合物を目的の生理学的状態と識別することを更に含む。

【０４１７】

本開示の更に別の態様は、１つ以上のプロセッサ及びメモリを含むコンピュータシステムを提供し、メモリは、試験化学化合物を目的の生理学的状態と関連付けるための方法を実施するための命令を格納する。方法は、試験化学化合物の化学構造のフィンガープリントを得、フィンガープリントをモデルに入力することを含み、モデルは１００以上のパラメータを含む。モデルは、フィンガープリントのモデルへの入力に応答して１つ以上の計算された活性化スコアを出力する。１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、摂動シグネチャのセットにおける対応する摂動シグネチャを表す。摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々は、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、それぞれの第１の細胞状態及び第２の細胞状

態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である。方法は、摂動シグネチャのセットにおける第１の摂動シグネチャについてのそれぞれの計算された活性化スコアが、第１の閾値基準を満たす場合、化学化合物を目的の生理学的状態と識別することを更に含む。

【０４１８】

本開示の別の態様は、試験化学化合物を目的の生理学的状態と関連付けるための、コンピュータによって実行可能な１つ以上のコンピュータプログラムを格納する非一時的なコンピュータ可読媒体を提供し、コンピュータは、１つ以上のプロセッサ及びメモリを含み、１つ以上のコンピュータプログラムは、方法を実施するコンピュータによって実行可能な命令を集合的に符号化する。方法は、試験化学化合物の化学構造のフィンガープリントを得、フィンガープリントをモデルに入力することを含み、モデルは１００以上のパラメータを含む。モデルは、フィンガープリントのモデルへの入力に応答して１つ以上の計算された活性化スコアを出力する。１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、摂動シグネチャのセットにおける対応する摂動シグネチャを表す。摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々は、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、それぞれの第１の細胞状態及び第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である。方法は、摂動シグネチャのセットにおける第１の摂動シグネチャについてのそれぞれの計算された活性化スコアが、第１の閾値基準を満たす場合、化学化合物を目的の生理学的状態と識別することを更に含む。

【０４１９】

本開示の更に別の態様は、１つ以上のプロセッサ及びメモリを含むコンピュータシステムを提供し、メモリは化学化合物を目的の生理学的状態と関連付けるための方法を実施するための命令を格納する。方法は、複数の化合物における化合物の各々の化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることを含む。方法は、複数の化合物における化合物の各々についての細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの数値的活性化スコアを電子形式で得ることであって、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素の独立したサブセットを含む、得ること、を含む。この方法は、複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々について、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの細胞構成要素モジュールについてのそれぞれの計算された活性化スコアと、（ｉｉ）細胞構成要素モジュールのセットにおけるそれぞれの化合物についてのそれぞれの細胞構成要素モジュールのそれぞれの数値的活性化スコアとの間のそれぞれの差を使用して訓練されていないモデルを訓練することを更に含む。訓練は、差に応答して訓練されていないモデルと関連付けられた複数のパラメータを調整し、複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を目的の生理学的状態と関連付ける訓練されたモデルを得る。

【０４２０】

本開示の別の態様は化学化合物を目的の生理学的状態と関連付けるための、コンピュータによって実行可能な１つ以上のコンピュータプログラムを格納する非一時的なコンピュータ可読媒体を提供し、コンピュータは、１つ以上のプロセッサ及びメモリを含み、１つ以上のコンピュータプログラムは、方法を実施するコンピュータによって実行可能な命令を集合的に符号化する。方法は、複数の化合物における化合物の各々の化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得る

ことを含む。方法は、複数の化合物における化合物の各々についての細胞構成要素モジュールのセットにおける細胞構成要素モジュールの各々のそれぞれの数値的活性化スコアを電子形式で得ることであって、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々が、複数の細胞構成要素の独立したサブセットを含む、得ること、を更に含む。この方法は、複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々について、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの細胞構成要素モジュールについてのそれぞれの計算された活性化スコアと、（ｉｉ）細胞構成要素モジュールのセットにおけるそれぞれの化合物についてのそれぞれの細胞構成要素モジュールのそれぞれの数値的活性化スコアとの間のそれぞれの差を使用して訓練されていないモデルを訓練することを更に含む。訓練は、差に応答して訓練されていないモデルと関連付けられた複数のパラメータを調整し、複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を目的の生理学的状態と関連付ける訓練されたモデルを得る。

【０４２１】

本開示の更に別の態様は、１つ以上のプロセッサ及びメモリを含むコンピュータシステムを提供し、メモリは、化学化合物を目的の生理学的状態と関連付けるための命令を格納し、方法は、複数の化合物における化合物の各々の化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ることを含む。方法は、複数の化合物における対応する化合物の各々についての摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々のそれぞれの数値的活性化スコアを電子形式で得ることを更に含む。摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々は、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、それぞれの第１の細胞状態及び第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である。方法は、複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々について、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの摂動シグネチャについてのそれぞれの計算された活性化スコアと、（ｉｉ）摂動シグネチャのセットにおける対応する化合物についてのそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアとの間のそれぞれの差を使用して訓練されていないモデルを訓練することを含む。訓練は、差に応答して訓練されていないモデルと関連付けられた複数のパラメータを調整し、複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を目的の生理学的状態と関連付ける訓練されたモデルを得る。

【０４２２】

本開示の別の態様は、化学化合物を目的の生理学的状態と関連付けるための、コンピュータによって実行可能な１つ以上のコンピュータプログラムを格納する非一時的なコンピュータ可読媒体を提供し、コンピュータは、１つ以上のプロセッサ及びメモリを含み、１つ以上のコンピュータプログラムは、複数の化合物における化合物の各々の化学構造のそれぞれのフィンガープリントを電子形式で得、それによって複数のフィンガープリントを得ること、を含む、方法を実施するコンピュータによって実行可能な命令を集合的に符号化する。方法は、複数の化合物における対応する化合物の各々についての摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々のそれぞれの数値的活性化スコアを電子形式で得ることを更に含む。摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々は、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化

と、それぞれの第１の細胞状態とそれぞれの第２の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコアと、を含み、それぞれの第１の細胞状態及び第２の細胞状態のうちの一方が、非摂動細胞状態であり、それぞれの第１の細胞状態及び第２の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である。方法は、複数の化合物におけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々について、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの摂動シグネチャについてのそれぞれの計算された活性化スコアと、（ｉｉ）摂動シグネチャのセットにおける対応する化合物についてのそれぞれの摂動シグネチャのそれぞれの数値的活性化スコアとの間のそれぞれの差を使用して訓練されていないモデルを訓練することを更に含む。訓練は、差に応答して訓練されていないモデルと関連付けられた複数のパラメータを調整し、複数のパラメータが、１００以上のパラメータを含み、それによって、化学化合物を目的の生理学的状態と関連付ける訓練されたモデルを得る。

【０４２３】

本開示の更に別の態様は、１つ以上のプロセッサ、及び１つ以上のプロセッサによる実行のための１つ以上のプログラムを格納するメモリを有するコンピュータシステムを提供し、１つ以上のプログラムは、本明細書に開示される方法及び／又は実施形態のうちのいずれかを実施するための命令を含む。いくつかの実施形態において、本開示の方法及び／又は実施形態のいずれかは、１つ以上のプロセッサ、及び１つ以上のプロセッサによって実行するための１つ以上のプログラムを格納するメモリを有するコンピュータシステムにおいて実施される。

【０４２４】

本開示の別の態様は、コンピュータによって実行するように構成された１つ以上のプログラムを格納する非一時的なコンピュータ可読記憶媒体を提供し、１つ以上のプログラムは、本明細書に開示される方法のいずれかを実行するための命令を含む。

【０４２５】

ＩＶ．細胞構成要素モジュールの識別

いくつかの実施形態において、目的の生理学的状態と関連付けられた細胞構成要素モジュール１３２が識別される。このような方法は、図２及び図１４と併せて本明細書で説明される。特に、図１４Ａのブロック１５００を参照すると、いくつかの実施形態において、方法は、目的の生理学的状態と関連付けられた第１の細胞構成要素モジュール１３２を識別することを更に含む。

【０４２６】

本開示のいくつかの実施形態に従って、細胞構成要素を目的の生理学的状態と関連付けるための方法２００の例示的なワークフローは、図２Ａ～図２Ｂを参照して提供される。

【０４２７】

図２Ａのブロック２０２及び図１４Ａのブロック１５０２を参照すると、方法は、１つ以上の第１のデータセットを電子形式で得ることを含む。図１４Ｂのブロック１５０４を参照すると、１つ以上の第１のデータセットは、第１の複数の細胞におけるそれぞれの細胞の各々について、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含む。このようにして、複数のベクトルが得られる。

【０４２８】

いくつかの実施形態において、目的の生理学的状態は、疾患であり、第１の複数の細胞が、複数の注釈付きの細胞状態によって立証されるように、疾患を代表する細胞、及び疾患を代表しない細胞を含む。

【０４２９】

いくつかの実施形態において、図３Ａのブロック３００の目的の生理学的状態は、疾患と関連付けられた異常な細胞プロセスであり、第１の複数の細胞は、注釈付きの細胞状態

によって立証されるように、疾患を代表する細胞、及び疾患を代表しない細胞を含む。

【０４３０】

いくつかの実施形態において、図３Ａのブロック３００の目的の生理学的状態は、疾患と関連付けられた異常な細胞プロセスであり、第１の複数の細胞は、注釈付きの細胞状態によって立証されるように、疾患状態を代表する細胞、及び健康又は対照状態を代表する細胞を含む。

【０４３１】

いくつかの実施形態において、図３Ａのブロック３００の目的の生理学的状態は、複数の疾患と関連付けられた異常な細胞プロセスであり、第１の複数の細胞は、複数の注釈付きの細胞状態によって立証されるように、複数の細胞のサブセット、複数の疾患におけるそれぞれの疾患を代表する細胞のそれぞれのサブセットの各々を含む。

【０４３２】

図１４Ｂのブロック１５０６を参照すると、いくつかの実施形態において、第１の複数の細胞は、２個、３個、４個、５個、６個、７個、８個、９個、１０個、１１個、１２個、１３個、１４個、１５個、１６個、１７個、１８個、１９個、２０個、３０個、４０個、５０個、６０個、７０個、８０個、１００個、２００個、又は１０００個以上の細胞を含み、複数（例えば、２個、３個、４個、５個、６個、７個、８個、９個、１０個、１１個、１２個、１３個、１４個、１５個、１６個、１７個、１８個、１９個、２０個、３０個、４０個、５０個、６０個、７０個、８０個、１００個、２００個、又は１０００個）の注釈付きの細胞状態を集合的に表す。

【０４３３】

図１４Ｂのブロック１５０８を参照すると、いくつかの実施形態において、複数の細胞構成要素は、２個、３個、４個、５個、６個、７個、８個、９個、１０個、１５個、２０個、２５個、３０個、３５個、５０個、１００個、５００個、１０００個、５０００個、１０，０００個以上の細胞構成要素を含む。いくつかの実施形態において、複数の細胞構成要素は、２～１０，０００個又は細胞構成要素からなる。いくつかの実施形態において、複数の細胞構成要素は、１００～１０，０００個又は細胞構成要素からなる。

【０４３４】

図２Ａのブロック２０４を参照すると、方法は、複数のベクトルにアクセスすること、又はそれを形成することを含む。図１４Ａのブロック１５１０を参照すると、複数のベクトルにおけるそれぞれのベクトルの各々は、（ｉ）複数の構成要素におけるそれぞれの細胞構成要素に対応し、（ｉｉ）対応する複数のエレメントを含む。図１４Ａのブロック１５１２を参照すると、対応する複数のエレメントにおけるそれぞれのエレメントの各々は、第１の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を表す対応するカウントを有する。

【０４３５】

ブロック２０６を参照すると、複数のベクトルは、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別するために使用される。複数の細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々は、複数の細胞構成要素のサブセットを含む。複数の細胞構成要素モジュールは、（ｉ）複数の候補細胞構成要素モジュール、及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された潜在表現で配置され、複数の細胞構成要素モジュールは、１０を超える細胞構成要素モジュールを含む。

【０４３６】

図１４Ｂのブロック１５１４を参照すると、いくつかの実施形態において、複数の注釈付きの細胞状態における注釈付きの細胞状態は、曝露条件下（例えば、曝露期間、化合物の濃度、又は曝露期間及び化合物の濃度の組み合わせ）の化合物への第１の複数の細胞における細胞の曝露である。

【０４３７】

図１４Ｂのブロック１５１８を参照すると、いくつかの実施形態において、複数の細胞

構成要素における細胞構成要素の各々は、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせである。

【０４３８】

図１４Ｂのブロック１５２０を参照すると、いくつかの実施形態において、第１又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、比色測定、蛍光測定、発光測定、又は共鳴エネルギー移動（ＦＲＥＴ）測定によって決定される。

【０４３９】

図１４Ｂのブロック１５２２を参照すると、いくつかの実施形態において、第１又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、単一細胞リボ核酸（ＲＮＡ）配列決定（ｓｃＲＮＡ－ｓｅｑ）、ｓｃＴａｇ－ｓｅｑ、配列決定を使用したトランスポザーゼ－アクセス可能なクロマチンのための単一細胞アッセイ（ｓｃＡＴＡＣ－ｓｅｑ）、ＣｙＴＯＦ／ＳＣｏＰ、Ｅ－ＭＳ／Ａｂｓｅｑ、ｍｉＲＮＡ－ｓｅｑ、ＣＩＴＥ－ｓｅｑ、又はそれらの任意の組み合わせによって決定される。

【０４４０】

図１４Ｂのブロック１５２４を参照すると、いくつかの実施形態、目的の生理学的状態は、疾患であり、第１の複数の細胞が、複数の注釈付きの細胞状態によって立証されるように、疾患を代表する細胞、及び疾患を代表しない細胞を含む。

【０４４１】

図１４Ｂのブロック１５２６を参照すると、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別するために使用され、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々は、複数の細胞構成要素のサブセットを含む。複数の細胞構成要素モジュールは、（ｉ）複数の候補細胞構成要素モジュール、及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された潜在表現で配置され、複数の細胞構成要素モジュールは、１０を超える細胞構成要素モジュールを含む。

【０４４２】

図１４Ｃのブロック１５２８を参照すると、いくつかの実施形態において、複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することは、複数のベクトルにおけるベクトルの各々の対応する複数のエレメントの各々を使用して、複数のベクトルに相関モデルを適用することを含む。いくつかの実施形態において、相関モデルは、グラフクラスタリングアルゴリズム（例えば、グラフクラスタリング方法は、ピアソン相関ベースの距離メトリック上のライデン（Ｌｅｉｄｅｎ）クラスタリングであり、グラフクラスタリング方法は、ルーバン（Ｌｏｕｖａｉｎ）クラスタリングなどである）。

【０４４３】

図１４Ｃのブロック１５３２を参照すると、いくつかの実施形態において、複数の細胞構成要素モジュールは、１０～２０００個、１００～１００００個、２０～５０００個、２～１５，０００個、８０～５０００個、１００～５００個の細胞構成要素モジュールからなる。いくつかの実施形態において、複数の細胞構成要素モジュールは、２～５００個の細胞構成要素モジュールである。

【０４４４】

図１４Ｃのブロック１５３４を参照すると、いくつかの実施形態において、複数の細胞構成要素は、１０～２０００個、１００～１００００個、２０～５０００個、２～１５，０００個、８０～５０００個、１００～５００個の細胞構成要素からなる。いくつかの実施形態において、複数の細胞構成要素は、２～５００個の細胞構成要素である。

【０４４５】

図１４Ｃのブロック１５３６を参照すると、いくつかの実施形態において、複数の構成

要素モジュールにおける候補細胞構成要素モジュールの各々は、２００～３００個の細胞構成要素からなる。

【０４４６】

図２Ａのブロック２０８及び図１４Ｃのブロック１５３８を参照すると、方法は、１つ以上の第２のデータセットを電子形式で得ることを含む。１つ以上の第２のデータセットは、第２の複数の細胞におけるそれぞれの細胞の各々について、第２の複数の細胞が、２０個以上の細胞を含み、目的の生理学的状態を通知する複数の共変量を集合的に表し、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含む。したがって、細胞構成要素カウントデータ構造が得られ、細胞構成要素カウントデータ構造は、（ｉ）第２の複数の細胞、及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定される。

【０４４７】

図１４Ｃのブロック１５４０を参照すると、いくつかの実施形態において、複数の共変量は、細胞バッチ、細胞ドナー、細胞型、疾患状態、又は化学化合物への曝露を含む。

【０４４８】

図２Ｂのブロック２１０及びブロック１５４２　図１４Ｄを参照すると、活性化データ構造は、複数の細胞構成要素又はその表現を共通次元として使用して、細胞構成要素カウントデータ構造及び潜在表現を組み合わせることによって形成される。活性化データ構造は、複数の細胞構成要素モジュールにおける細胞構成要素モジュールの各々について、第２の複数の細胞における細胞の各々について、それぞれの活性化重みを含む。

【０４４９】

図２Ｂのブロック２１２及び図１４Ｄのブロック１５４４を参照すると、方法は、（ｉ）活性化データ構造を候補モデルに入力したときに、活性化データ構造内に表される細胞構成要素モジュールの各々における複数の共変量における各共変量の不在又は存在の予測と、（ｉｉ）細胞構成要素モジュールの各々における各共変量の実際の不在又は存在との間の差を使用して候補細胞構成要素モデルを訓練することを更に含む。訓練することは、差に応答して候補細胞構成要素モデルと関連付けられた複数の共変量重みを調整し、複数の共変量重みは、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの共変量の各々について、それぞれの共変量が、活性化データ構造にわたって、それぞれの細胞構成要素モジュールと相関するかどうかを示す対応する重みを含む。

【０４５０】

図１４Ｄのブロック１５４６を参照すると、候補細胞構成要素モデルを訓練することは、マルチタスク策定におけるカテゴリ交差エントロピー損失を使用して実施され、複数の共変量における共変量の各々が、複数のコスト関数におけるコスト関数に対応し、複数のコスト関数におけるそれぞれのコスト関数の各々が、共通の重み付け係数を有する。

【０４５１】

したがって、図２Ｃのブロック２１４及び図１４Ｄのブロック１５４８を参照すると、複数の共変量重みは、候補細胞構成要素モデルを訓練する際に、複数の候補細胞構成要素モジュールにおける第１の細胞構成要素モジュールを識別するために使用され、複数の候補細胞構成要素モジュールにおける第１の細胞構成要素モジュールは、目的の生理学的状態と関連付けられる。

【０４５２】

いくつかの実施形態において、第１及び／又は第２の複数の細胞は、少なくとも５個、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも３０個、少なくとも４０個、少なくとも５０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも１０００個、少なくとも少なくとも２０００個、少なくとも３０００個、少なくとも４０００個、少なくとも５０００個、少なくとも１０，０００個、少なくとも２０，０００個、少なくとも３０，００

０個、少なくとも５０，０００個、少なくとも８０，０００個、少なくとも１００，０００個、少なくとも５００，０００個、又は少なくとも１００万個の細胞を含む。いくつかの実施形態において、第１及び／又は第２の複数の細胞は、５００万個以下、１００万個以下、５００，０００個以下、１００，０００個以下、５０，０００個以下、１０，０００個以下、５０００個以下、１０００個以下、５００個以下、２００個以下、１００個以下、又は５０個以下の細胞を含む。いくつかの実施形態において、第１及び／又は第２の複数の細胞は、５～１００個、１０～５０個、２０～５００個、２００～１０，０００個、１０００～１００，０００個、５０，０００～５００，０００個、又は１０，０００～１００万個の細胞を含む。いくつかの実施形態において、第１及び／又は第２の複数の細胞は、５個以上の細胞から始まり、５００万個以下の細胞で終わる別の範囲内にある。

【０４５３】

いくつかの実施形態において、第２の複数の細胞は、第１の複数の細胞に含まれる細胞を含まない。いくつかの実施形態において、第２の複数の細胞は、第１の複数の細胞に含まれる細胞の一部又は全てを含む。

【０４５４】

いくつかの実施形態において、複数の注釈付きの細胞状態は、細胞表現型、細胞挙動、疾患状態、遺伝子変異、遺伝子若しくは遺伝子産物の摂動（例えば、ノックダウン、サイレンシング、過剰発現など）、及び／又は化合物への曝露のうちの１つ以上を含む。いくつかの実施形態において、複数の注釈付きの細胞状態における注釈付きの細胞状態は、曝露条件下での化合物への第１の複数の細胞における細胞の曝露である。例えば、細胞の曝露は、１つ以上の化合物での細胞の任意の処理を含む。いくつかの実施形態において、１つ以上の化合物は、例えば、小分子、生物製剤、治療剤、タンパク質、小分子と組み合わされたタンパク質、ＡＤＣ、核酸（例えば、ｓｉＲＮＡ、干渉ＲＮＡ、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ｓｈＲＮＡ、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ガイドＲＮＡ（例えば、Ｃａｓ９系若しくは他の細胞成分編集系）など）、並びに／又は前述のいずれかの任意の組み合わせを含む。いくつかの実施形態において、曝露条件は、曝露期間、化合物の濃度、又は曝露期間及び化合物の濃度の組み合わせである。いくつかの実施形態において、化合物は、上記の「化合物」と題されるセクションにおいてなど、本明細書に記載される実施形態のいずれでもある。

【０４５５】

いくつかの実施形態において、複数の注釈付きの細胞状態は、細胞バッチ、細胞ドナー、細胞型、細胞株、疾患状態、時点、複製、及び／又は関連するメタデータの１つ以上の兆候を含む。いくつかの実施形態において、複数の注釈付きの細胞状態は、実験データ（例えば、フローサイトメトリーの読み出し、イメージング及び顕微鏡注釈、細胞構成要素データなど）を含む。いくつかの実施形態において、複数の注釈付きの細胞状態は、１つ以上の遺伝子マーカー（例えば、コピー数バリエーション、単一ヌクレオチドバリアント、多ヌクレオチド多型、挿入、欠失、遺伝子融合、マイクロサテライト不安定性状態、増幅、及び／又はアイソフォーム）を含む。いくつかの実施形態において、複数の注釈付きの細胞状態は、本明細書に開示される共変量のうちのいずれか及び／又は本明細書に開示される目的の生理学的状態のうちのいずれか、例えば、上記の「生理学的状態」と題されるセクションなどを含む。

【０４５６】

本明細書に開示される任意の細胞構成要素及び／又は任意の細胞構成要素モジュール、並びにそれらの任意の実施形態、置換、修飾、追加、欠失、及び／又は組み合わせは、上記の「細胞構成要素及び細胞構成要素モジュール」と題されたセクションに記載されるように、細胞構成要素モジュールの識別のために企図される。例えば、いくつかの実施形態において、複数の細胞構成要素における細胞構成要素の各々は、特定の遺伝子、遺伝子に関連する特定のｍＲＮＡ、炭水化物、脂質、エピジェネティック特徴、代謝産物、タンパク質、又はそれらの組み合わせである。いくつかの実施形態において、複数の細胞構成要素は、１００～８，０００個の細胞構成要素からなる。いくつかの実施形態において、複

数の細胞構成要素モジュールは、１０～２０００個の細胞構成要素モジュールからなる。いくつかの実施形態において、複数の構成要素モジュールにおける候補細胞構成要素モジュールの各々は、２００～３００個の細胞構成要素からなる。

【０４５７】

いくつかの実施形態において、それぞれの細胞構成要素の対応する存在量は、上記に開示される細胞構成要素のいずれかの存在量を含む。

【０４５８】

いくつかの存在量カウント技術（例えば、細胞構成要素測定技術）のうちのいずれか１つを使用して、それぞれの細胞の各々におけるそれぞれの細胞構成要素の各々についての対応する存在量を得ることができる。例えば、表１は、本開示のいくつかの実施形態に従う、単一細胞の細胞構成要素測定のための非限定的な技術を列挙する。

【０４５９】

いくつかの実施形態において、それぞれの細胞構成要素の対応する存在量は、蛍光、化学発光、電気シグナル検出、ポリメラーゼ連鎖反応（ＰＣＲ）、逆転写酵素ポリメラーゼ連鎖反応（ＲＴ－ＰＣＲ）、デジタル液滴ＰＣＲ（ｄｄＰＣＲ）、固体状態ナノポア検出、ＲＮＡスイッチ活性化、ノーザンブロット、及び／又は遺伝子発現の連続分析（ＳＡＧＥ）を介したマイクロアレイ分析を含む１つ以上の方法を使用して決定される。いくつかの実施形態において、第１又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、比色測定、蛍光測定、発光測定、又は共鳴エネルギー移動（ＦＲＥＴ）測定によって決定される。

【０４６０】

いくつかの実施形態において、第１及び／又は第２の複数の細胞におけるそれぞれの細胞における遺伝子発現は、細胞を配列決定し、次いで配列決定中に識別された各遺伝子転写物の量をカウントすることによって測定することができる。いくつかの実施形態において、配列決定及び定量化された遺伝子転写産物は、ｍＲＮＡなどのＲＮＡを含む。いくつかの実施形態において、配列決定及び定量化された遺伝子転写産物は、タンパク質（例えば、転写因子）などのｍＲＮＡの下流産物を含む。一般に、本明細書で使用される場合、「遺伝子転写産物」という用語は、翻訳後修飾を含む、遺伝子転写又は翻訳の任意の下流産物を示すために使用されてもよく、「遺伝子発現」は、一般に、遺伝子転写産物の任意の尺度を指すために使用されてもよい。

【０４６１】

いくつかの実施形態において、それぞれの細胞構成要素の対応する存在量はＲＮＡ存在量（例えば、遺伝子発現）であり、それぞれの細胞構成要素の存在量は、それぞれの遺伝子に対応する１つ以上の核酸分子のポリヌクレオチドレベルを測定することによって決定される。それぞれの遺伝子の転写レベルは、第１及び／又は第２の複数の細胞におけるそれぞれの細胞中に存在するｍＲＮＡ又はそれに由来するポリヌクレオチドの量から決定することができる。ポリヌクレオチドは、マイクロアレイ分析、ポリメラーゼ連鎖反応（ＰＣＲ）、逆転写酵素ポリメラーゼ連鎖反応（ＲＴ－ＰＣＲ）、ノーザンブロット、遺伝子発現の連続分析（ＳＡＧＥ）、ＲＮＡスイッチ、ＲＮＡフィンガープリンティング、リガーゼ連鎖反応、Ｑベータレプリカーゼ、等温増幅法、鎖置換増幅、転写ベース増幅システム、ヌクレアーゼ保護アッセイ（Ｓｉヌクレアーゼ又はＲＮＡｓｅ保護アッセイ）、及び／又は固体状態ナノポア検出を含むが、これらに限定されない、様々な方法によって検出及び定量することができる。例えば、Ｄｒａｇｈｉｃｉ，Ｄａｔａ Ａｎａｌｙｓｉｓ Ｔｏｏｌｓ ｆｏｒ ＤＮＡ Ｍｉｃｒｏａｒｒａｙｓ，Ｃｈａｐｍａｎ ａｎｄ Ｈａｌｌ／ＣＲＣ，２００３、Ｓｉｍｏｎ ｅｔ ａｌ．，Ｄｅｓｉｇｎ ａｎｄ Ａｎａｌｙｓｉｓ ｏｆ ＤＮＡ Ｍｉｃｒｏａｒｒａｙ Ｉｎｖｅｓｔｉｇａｔｉｏｎｓ，Ｓｐｒｉｎｇｅｒ，２００４、Ｒｅａｌ－Ｔｉｍｅ ＰＣＲ：Ｃｕｒｒｅｎｔ Ｔｅｃｈｎｏｌｏｇｙ ａｎｄ Ａｐｐｌｉｃａｔｉｏｎｓ，Ｌｏｇａｎ，Ｅｄｗａｒｄｓ，ａｎｄ Ｓａｕｎｄｅｒｓ ｅｄｓ．，Ｃａｉｓｔｅｒ Ａｃａｄｅｍｉｃ Ｐｒｅｓｓ，２００９、Ｂｕｓｔｉｎ Ａ－Ｚ ｏｆ Ｑｕａｎｔｉｔａｔｉｖｅ ＰＣＲ（ＩＵＬ Ｂｉｏ

technology，No．５），International University Line，２００４、Velculescu et al．，（１９９５）Science ２７０：４８４－４８７、Matsumura et al．，（２００５）Cell．Microbiol．７：１１－１８、Serial Analysis of Gene Expression（ＳＡＧＥ）：Methods and Protocols（Methods in Molecular Biology），Humana Press，２００８を参照されたく、これらの各々は参照によりその全体が本明細書に組み込まれる。

【０４６２】

いくつかの実施形態において、それぞれの細胞構成要素の対応する存在量は、発現ＲＮＡ又はそれに由来する核酸（例えば、ＲＮＡポリメラーゼプロモーターを組み込んだｃＤＮＡに由来するｃＤＮＡ又は増幅ＲＮＡ）から、天然核酸分子、及び合成核酸分子を含む、第１及び／又は第２の複数の細胞におけるそれぞれの細胞から得られる。したがって、いくつかの実施形態において、それぞれの細胞構成要素の対応する存在量は、総細胞ＲＮＡ、ポリ（Ａ）＋メッセンジャーＲＮＡ（ｍＲＮＡ）若しくはその画分、細胞質ｍＲＮＡ、又はｃＤＮＡから転写されたＲＮＡ（例えば、ｃＲＮＡ）などの非限定的な供給源から得られる。総ＲＮＡ及びポリ（Ａ）＋ＲＮＡを調製するための方法は、当該技術分野で周知であり、一般に、例えば、Ｓａｍｂｒｏｏｋ，ｅｔ ａｌ．，Ｍｏｌｅｃｕｌａｒ Ｃｌｏｎｉｎｇ：Ａ Ｌａｂｏｒａｔｏｒｙ Ｍａｎｕａｌ（３ｒｄ Ｅｄｉｔｉｏｎ，２００１）に記載されている。ＲＮＡは、グアニジンチオシアナート溶解後のＣｓＣｌ遠心分離（例えば、Ｃｈｉｒｇｗｉｎ ｅｔ ａｌ．，１９７９，Ｂｉｏｃｈｅｍｉｓｔｒｙ １８：５２９４－５２９９を参照されたい）、シリカゲルベースのカラム（例えば、ＲＮｅａｓｙ（Ｑｉａｇｅｎ、Ｖａｌｅｎｃｉａ，Ｃａｌｉｆ．）若しくはＳｔｒａｔａＰｒｅｐ（Ｓｔｒａｔａｇｅｎｅ、Ｌａ Ｊｏｌｌａ，Ｃａｌｉｆ．））を使用して、又はＡｕｓｕｂｅｌ ｅｔ ａｌ．，ｅｄｓ．，１９８９，Ｃｕｒｒｅｎｔ Ｐｒｏｔｏｃｏｌｓ Ｉｎ Ｍｏｌｅｃｕｌａｒ Ｂｉｏｌｏｇｙ，Ｖｏｌ．ＩＩＩ，Ｇｒｅｅｎ Ｐｕｂｌｉｓｈｉｎｇ Ａｓｓｏｃｉａｔｅｓ，Ｉｎｃ．，Ｊｏｈｎ Ｗｉｌｅｙ ＆ Ｓｏｎｓ，Ｉｎｃ．，Ｎｅｗ Ｙｏｒｋ，ｐｐ．１３．１２．１－１３．１２．５）に記載されているフェノール及びクロロホルムを使用して目的の細胞から抽出することができる。ポリ（Ａ）＋ＲＮＡは、例えば、オリゴ－ｄＴセルロースを用いた選択によって、又は代替的に、全細胞ＲＮＡのオリゴ－ｄＴプライミング逆転写によって選択することができる。ＲＮＡは、当該技術分野で既知の方法によって、例えば、ＺｎＣｌ２とのインキュベーションによって断片化して、ＲＮＡの断片を生成することができる。

【０４６３】

いくつかの実施形態において、第１及び／又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、配列決定によって決定される。いくつかの実施形態において、第１及び／又は第２の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、単一細胞リボ核酸（ＲＮＡ）配列決定（ｓｃＲＮＡ－ｓｅｑ）、ｓｃＴａｇ－ｓｅｑ、配列決定を使用したトランスポザーゼ－アクセス可能なクロマチンのための単一細胞アッセイ（ｓｃＡＴＡＣ－ｓｅｑ）、ＣｙＴＯＦ／ＳＣｏＰ、Ｅ－ＭＳ／Ａｂｓｅｑ、ｍｉＲＮＡ－ｓｅｑ、ＣＩＴＥ－ｓｅｑ、及びそれらの任意の組み合わせによって決定される。

【０４６４】

細胞構成要素存在量測定技術は、測定される所望の細胞構成要素に基づいて選択することができる。例えば、ｓｃＲＮＡ－ｓｅｑ、ｓｃＴａｇ－ｓｅｑ、及びｍｉＲＮＡ－ｓｅｑを使用して、ＲＮＡ発現を測定することができる。具体的には、ｓｃＲＮＡ－ｓｅｑはＲＮＡ転写産物の発現を測定し、ｓｃＴａｇ－ｓｅｑは希少なｍＲＮＡ種の検出を可能にし、ｍｉＲＮＡ－ｓｅｑはマイクロＲＮＡの発現を測定する。ＣｙＴＯＦ／ＳＣｏＰ及びＥ－ＭＳ／Ａｂｓｅｑを使用して、細胞内のタンパク質発現を測定することができる。ＣＩＴＥ－ｓｅｑは、細胞における遺伝子発現及びタンパク質発現の両方を同時に測定し、

ｓｃＡＴＡＣ－ｓｅｑは、細胞におけるクロマチンコンフォメーションを測定する。以下の表１は、上記の細胞構成要素存在量測定技術の各々を実施するための例示的なプロトコルを提供する。

【０４６５】

【表１】

表１－測定プロトコルの例

技術	プロトコル
RNA-seq	Olsen *et al.*, (2018), "Introduction to Single-Cell RNA Sequencing," Current protocols in molecular biology 122(1), pg.57.
Tag-seq	Rozenberg *et al.*, (2016), "Digital gene expression analysis with sample multiplexing and PCR duplicate detection: A straightforward protocol," BioTechniques, 61(1), pg.26.
ATAC-seq	Buenrostro *et al.,* (2015), "ATAC-seq: a method for assaying chromatic accessibility genome-wide," Current protcols in molecular biology, 109(1), pg.21.
miRNA-seq	Faridani *et al.*, (2016), "Single-cell sequencing of the small-RNA transcriptome," Nature biotechnology, 34(12), pg.1264.
CyTOF/SCoPE-MS/Abseq	Bandura *et al.,* (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), pg.6813. Budnik *et al.*, (2018), "SCoPE-ME: mass scpectrometry of single mammalian cells quantifies proteome heterogenity during cell differentiation," Genome biology, 19(1), pg.161. Shahi *et al.*, (2017), "Abseq: Ultrahigh-throughoutput single cell protein profiling with droplep microfluidic barcoding," Scientific reports, 7, pg.44447.
CITE-seq	Stoeckius *et al.*, (2017), "Simultaneous epitope and transcritome measurement in single cells," Nature Methods, 14(9), pg.856.

【０４６６】

いくつかの実施形態において、複数の細胞構成要素は、単一の時点で測定される。いくつかの実施形態において、複数の細胞構成要素は、複数の時点で測定される。例えば、いくつかの実施形態において、複数の細胞構成要素は、細胞状態遷移（例えば、分化プロセス、化合物への曝露に対する応答、発生プロセスなど）全体にわたる複数の時点で測定される。

【０４６７】

本開示は、細胞（例えば、単一細胞）から得られる他の細胞構成要素の測定値を使用する類似の方法を包含するため、これは例示であり、限定ではないことを理解されたい。本開示は、本開示に記載される方法を実施する個人又は組織によって実施される実験作業から直接得られた測定値を使用する方法、並びに例えば、他者によって実施される実験作業の結果の報告から間接的に得られ、第三者の出版物、データベース、請負業者によって実施されるアッセイ、又は開示される方法を実施するのに有用な好適な入力データの他の供給源で報告されたデータを含む、任意の手段又は機構を通じて利用可能にされた測定値を使用する方法を包含することを更に理解されたい。

【０４６８】

いくつかの実施形態において、第１及び／又は第２の複数の細胞（例えば、１つ以上の複数の第１のデータセット及び／又は１つ以上の第２のデータセット）における複数の細

胞構成要素に対する対応する存在量は、前処理される。いくつかの実施形態において、前処理は、フィルタリング、正規化、マッピング（例えば、参照配列に対する）、定量化、スケーリング、デコンボリューション、クリーニング、次元縮小、変換、統計分析、及び／又は集約のうちの１つ以上を含む。

【０４６９】

例えば、いくつかの実施形態において、複数の細胞構成要素は、所望の品質、例えば、核酸配列のサイズ及び／若しくは品質、又はそれぞれの細胞構成要素についての最小及び／若しくは最大存在量値に基づいてフィルタリングされる。いくつかの実施形態において、フィルタリングは、Ｓｋｅｗｅｒなどの様々なソフトウェアツールによって一部又はその全体が実施される。Ｊｉａｎｇ，Ｈ．ｅｔ ａｌ．，ＢＭＣ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ １５（１８２）：１－１２（２０１４）を参照されたい。いくつかの実施形態において、複数の細胞構成要素は、例えば、ＡｆｔｅｒＱＣ、Ｋｒａｋｅｎ、ＲＮＡ－ＳｅＱＣ、ＦａｓｔＱＣ、又は別の同様のソフトウェアプログラムなどの配列決定データＱＣソフトウェアを使用して、品質管理のためにフィルタリングされる。いくつかの実施形態において、複数の細胞構成要素は、例えば、プルダウン、増幅、及び／又は配列決定バイアス（例えば、マッパビリティ、ＧＣバイアスなど）を考慮するために正規化される。例えば、Ｓｃｈｗａｒｔｚ ｅｔ ａｌ．，ＰＬｏＳ ＯＮＥ ６（１）：ｅ１６６８５（２０１１）ａｎｄ Ｂｅｎｊａｍｉｎｉ ａｎｄ Ｓｐｅｅｄ，Ｎｕｃｌｅｉｃ Ａｃｉｄｓ Ｒｅｓｅａｒｃｈ ４０（１０）：ｅ７２（２０１２）を参照されたく、その内容は全ての目的のために参照によりその全体が本明細書に組み込まれる。いくつかの実施形態において、前処理は、細胞構成要素のサブセットを複数の細胞構成要素から除去する。いくつかの実施形態において、複数の細胞構成要素について対応する存在量を前処理することは、高い信号対ノイズ比を改善する（例えば、低下させる）。

【０４７０】

いくつかの実施形態において、前処理は、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量と参照の存在量との比較を実施することを含む。いくつかの実施形態において、参照存在量は、例えば、正常試料、適合した試料、参照存在量値を含む参照データセット、ハウスキーピング遺伝子などの参照細胞構成要素、及び／又は参照標準から得られる。いくつかの実施形態において、細胞構成要素存在量のこの比較は、平均検定、ウィルコクソンランクサム検定（マンホイットニーＵ検定）、ｔ検定、ロジスティック回帰、及び一般化線形モデルの差異を含むが、これらに限定されない任意の差次的発現試験を使用して実施される。当業者は、細胞構成要素存在量の比較及び／又は正規化のために他のメトリックも可能であることを理解するであろう。

【０４７１】

したがって、いくつかの実施形態において、１つ以上の第１のデータセット及び／又は１つ以上の第２のデータセットにおけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量は、限定されないが、生の存在量値、絶対存在量値（例えば、転写物数）、相対的な存在量値（例えば、相対的な蛍光単位、トランスクリプトーム分析、及び／若しくは遺伝子セット発現分析（ＧＳＥＡ））、化合物若しくは集合的な存在量値、変換された存在量値（例えば、ｌｏｇ２及び／若しくはｌｏｇ１０変換）、参照（例えば、通常の試料、適合した試料、参照データセット、ハウスキーピング遺伝子、及び／若しくは参照標準）に対する変化（例えば、倍数若しくはｌｏｇ変化）、標準化された存在量値、中心傾向の尺度（例えば、平均、中央値、モード、加重平均、加重中央値、及び／若しくは加重モード）、分散の尺度（例えば、不一致、標準偏差、及び／若しくは標準誤差）、調整された存在量値（例えば、正規化された、スケーリングされた、及び／若しくは誤差訂正された）、次元低減された存在量値（例えば、主成分ベクトル及び／若しくは潜在成分）、並びに／又はそれらの組み合わせを含む、様々な形態のうちのいずれか１つを含む。次元縮小技術を使用して細胞構成要素の存在量を得るための方法は、主成分分析、因子分析、線形判別分析、多次元スケーリング、等角特徴マッピング、局所線形埋め込み、ヘシアン固有マッピング、スペクトル埋め込み、ｔ分布確率論的隣接埋め込み、並びに／又

はそれらの任意の置換、追加、欠失、修飾、及び／若しくは組み合わせを含むが、これらに限定されない、当該技術分野において既知であり、以下で更に詳細に説明されることは、当業者に明らかであろう。例えば、参照によりその全体が本明細書に組み込まれる、Ｓｕｍｉｔｈｒａ ｅｔ ａｌ．，２０１５，“Ａ Ｒｅｖｉｅｗ ｏｆ Ｖａｒｉｏｕｓ Ｌｉｎｅａｒ ａｎｄ Ｎｏｎ Ｌｉｎｅａｒ Ｄｉｍｅｎｓｉｏｎａｌｉｔｙ Ｒｅｄｕｃｔｉｏｎ Ｔｅｃｈｎｉｑｕｅｓ，”Ｉｎｔ Ｊ Ｃｏｍｐ Ｓｃｉ ａｎｄ Ｉｎｆ Ｔｅｃｈ，６（３），２３５４－２３６０を参照されたい。

【０４７２】

いくつかの実施形態において、複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することは、複数のベクトルにおけるベクトルの各々の対応する複数のエレメントの各々を使用して、複数のベクトルに相関モデルを適用することを含む。

【０４７３】

いくつかの実施形態において、相関モデルは、クラスタリング方法（例えば、クラスタリングモデル）を含む。いくつかの実施形態において、相関モデルは、グラフクラスタリング方法（例えば、モデル）及び／又は非グラフクラスタリング方法を含む。いくつかの実施形態において、グラフクラスタリング方法は、ピアソン相関ベースの距離メトリック上のライデン（Ｌｅｉｄｅｎ）クラスタリングである。いくつかの実施形態において、グラフクラスタリング方法は、ルーバン（Ｌｏｕｖａｉｎ）クラスタリングである。

【０４７４】

例えば、いくつかの実施態様では、方法は、相関ベースのコスト関数の適用を含む。相関ベースのコスト関数を最適化することは、細胞構成要素（例えば、遺伝子）間の最近傍関係を定義する最近傍グラフを計算することと、各細胞構成要素を、各細胞内の細胞構成要素についての存在量カウント（例えば、発現値）を格納することによって形成されるベクトルによって表すことと、細胞構成要素間の相関を計算することと、を含む。互いに高い相関を有する細胞構成要素は、最近傍であると判定され、グラフクラスタリング方法（例えば、ライデン（Ｌｅｉｄｅｎ）及び／又はルーバン（Ｌｏｕｖａｉｎ））を使用してグラフをクラスタリングすることによって、細胞構成要素モジュールを形成するために使用される。

【０４７５】

いくつかのクラスタリング技術のうちの任意の１つを使用することができ、その例としては、階層的クラスタリング、ｋ平均クラスタリング、及び密度ベースのクラスタリングが含まれるが、これらに限定されない。一実施形態において、階層的密度ベースのクラスタリングが使用される（ＨＤＢＳＣＡＮと称され、例えば、Ｃａｍｐｅｌｌｏ ｅｔ ａｌ．，（２０１５）．Ｈｉｅｒａｒｃｈｉｃａｌ ｄｅｎｓｉｔｙ ｅｓｔｉｍａｔｅｓ ｆｏｒ ｄａｔａ ｃｌｕｓｔｅｒｉｎｇ，ｖｉｓｕａｌｉｚａｔｉｏｎ，ａｎｄ ｏｕｔｌｉｅｒ ｄｅｔｅｃｔｉｏｎ．ＡＣＭ Ｔｒａｎｓ Ｋｎｏｗｌ Ｄｉｓｃ Ｄａｔａ，１０（１），５を参照されたい）。別の実施形態において、ルーバン（Ｌｏｕｖａｉｎ）クラスタリングなどのコミュニティ検出に基づくクラスタリングが使用される（例えば、Ｂｌｏｎｄｅｌ ｅｔ ａｌ．，（２００８）．Ｆａｓｔ ｕｎｆｏｌｄｉｎｇ ｏｆ ｃｏｍｍｕｎｉｔｉｅｓ ｉｎ ｌａｒｇｅ ｎｅｔｗｏｒｋｓ．Ｊ ｓｔａｔ ｍｅｃｈ：ｔｈｅｏｒ ｅｘｐ，２００８（１０），Ｐ１０００８を参照されたい）。なお別の実施形態において、ライデン（Ｌｅｉｄｅｎ）クラスタリングが使用される。ライデン（Ｌｅｉｄｅｎ）アルゴリズムは、個々のノードをコミュニティ間で移動してパーティションを決定し、パーティションを洗練し、洗練されたパーティションに基づいて集約ネットワークを作成することによって進行する。集約ネットワークは、プロセスの以前のステップで決定された未洗練のパーティションに基づいて更に分割され、新しいパーティションは、各集約ネットワーク内の個々のノードを移動することによって洗練される。例えば、Ｔｒａａｇ ｅｔ ａｌ．，（２０１９），“Ｆｒｏｍ Ｌｏｕｖａｉｎ ｔｏ Ｌｅｉｄｅｎ：ｇｕａｒａｎｔｅｅｉｎｇ ｗｅｌｌ－ｃｏｎｎｅｃｔｅｄ ｃｏｍｍｕｎｉ

ｔｉｅｓ，”Ｓｃｉ Ｒｅｐ ９：５２３３，ｄｏｉ：１０．１０３８／ｓ４１５９８－０１９－４１６９５－ｚを参照されたい。なお別の実施形態において、拡散経路アルゴリズムが使用される。

【０４７６】

一般に、ルーバン（Ｌｏｕｖａｉｎ）クラスタリング及び／又はライデン（Ｌｅｉｄｅｎ）クラスタリングなどのクラスタリングは、ハードパーティショニング技術を使用し、各エレメント（例えば、各細胞構成要素）は、重複することなく単一のクラスタに一意に割り当てられる。しかしながら、任意の１つの特定の理論に拘束されることなく、細胞プロセス（例えば、目的の生理学的状態と関連付けられる）は、細胞内の細胞構成要素のネットワーク間の複雑かつ動的な相互作用によって特徴付けられ得、例えば、単一の遺伝子は、細胞内の２つ、３つ、４つ、又はそれ以上の細胞プロセスにおいて、任意の数の同じ又は異なるプロセス及び経路において同様に機能する任意の数の他の遺伝子と組み合わせて役割を果たすことができる。したがって、細胞内活性の複雑さと並行して、第１のモジュールへの細胞構成要素のクラスタリングは、必ずしも他のモジュールを除外する必要はない。したがって、いくつかの実施形態において、細胞構成要素モジュールの識別は、細胞構成要素の重複するサブセットを有するモジュールを得ることを含む。

【０４７７】

相関ベースのモデルを使用してハードパーティショニング技術を利用することの代わりに、又はそれに加えて、いくつかの実施形態において、複数のベクトルを使用して複数の細胞構成要素モジュールにおける細胞構成要素モジュールの各々を識別することは、複数の細胞構成要素の表現を複数の次元縮小構成要素として生成する辞書学習モデルを含む。いくつかの実施形態において、辞書学習モデルは、Ｌ０正規化オートエンコーダである。これらのモデルの利点は、モジュールと細胞構成要素との間に１：１の対応を強制しないが、細胞構成要素が同時にいくつかのモジュールに現れることを可能にすることである。

【０４７８】

例えば、いくつかの実施態様において、方法は、スペアオートエンコーダコスト関数の適用を含む。いくつかのそのような例では、スパースオートエンコーダのコスト関数を最適化することは、ｐｙｔｏｒｃｈ又はｔｅｎｓｏｒｆｌｏｗに実装されているような標準訓練を使用して、その重みのＬ０正規化、及び再構築損失を伴う１層オートエンコーダを訓練することを含む。

【０４７９】

限定されないが、ファジーＫ平均、重複Ｋ平均（ＯＫＭ）、重み付けＯＫＭ（ＷＯＫＭ）、重複分割クラスタ（ＯＰＣ）、及びマルチクラスタ重複Ｋ平均拡張（ＭＣＯＫＥ）、並びに／又はそれらの任意の変形若しくは組み合わせを含む、重複分割アルゴリズムの他の方法が可能である。

【０４８０】

いくつかの実施形態において、統計技術は、１つ以上の第１のデータセットに符号化された潜在情報の形状を維持しながら、高次元データ（例えば、複数の注釈付きの細胞状態を集合的に表す第１の複数の細胞における細胞の各々について、複数の細胞構成要素モジュールにわたる複数の細胞構成要素の存在量）を低次元空間に圧縮するために使用され得る。例えば、図４の上部パネルに示されるように、カウントマトリックスは、第１の複数の細胞における細胞の各々について、複数の細胞構成要素における細胞構成要素の各々について、対応するカウント（例えば、存在量）を含む。カウントマトリックスは、異なる注釈付きの細胞状態（例えば、細胞型、曝露条件、疾患など）の条件下でのそれらの対応する存在量の類似性に基づいて、第１の複数の細胞にわたる細胞構成要素のクラスタリングを表す低次元空間にデータが縮小される、図４の下部パネルに示される潜在表現に変換することができる。したがって、クラスタ化された細胞構成要素は、細胞構成要素モジュールとして表され、潜在表現では、複数の細胞状態にわたる挙動の類似性を符号化する。

【０４８１】

図４に示される潜在表現を再び参照すると、各行－列グループ化におけるエントリー内

の値は、元の入力データセットに基づいて次元数の減少によって決定される。例えば、各エントリーは、それぞれの列によって表されるそれぞれの細胞構成要素の各々について、それぞれの行によって表されるそれぞれの細胞構成要素モジュールに含まれる複数の細胞構成要素のサブセット（例えば、重み１－１、重み１－２など）におけるメンバーシップの表示を含むことができる。特に、いくつかの実施形態において、各エントリーは、それぞれの細胞構成要素がそれぞれのモジュールに含まれるかどうかを示す重みである。いくつかの実施態様において、重みは、メンバーシップの２値表示である（例えば、それぞれのモジュールにおける存在又は不在は、それぞれ１又は０で示される）。いくつかの実施態様において、重みは、それぞれのモジュールに対する細胞構成要素の相対的な重要性（例えば、メンバーシップの確率及び／又は相関）を示すようにスケーリングされる。

【０４８２】

いくつかの実施形態において、潜在表現におけるそれぞれの次元は、それぞれの細胞構成要素の表現に対応する。細胞構成要素の表現は、例えば、潜在表現マトリックス内のそれぞれのエントリー（例えば、重み）が複数の細胞構成要素に対応する場合などの、細胞構成要素の非線形表現から生じ得る。細胞構成要素の表現を含む他の実施形態は、主成分分析を使用して得られた潜在表現を含み、各主成分は、複数の細胞構成要素に対応するデータの分散及び／又は他の変換を表す。

【０４８３】

いくつかの実施形態において、次元数削減技術は、データのいくつかの非可逆圧縮をもたらす。しかしながら、結果として生じる潜在表現（例えば、潜在表現１１８）は、計算記憶サイズにおいてより小さく、したがって、モデル訓練などの他の下流技術と併せて分析するためのより少ない計算処理能力を必要とする。したがって、潜在表現における複数の細胞構成要素モジュールの配置は、現代のコンピューティングデバイスを使用して、現在開示されている方法の計算実現可能性を増加させる。

【０４８４】

様々な次元数削減技術を使用することができる。いくつかの実施形態において、次元数削減は、主成分（ＰＣＡ）、ランダム投影、独立成分分析、特徴選択、因子分析、Ｓａｍｍｏｎマッピング、曲線成分分析、確率的隣接埋め込み（ＳＮＥ）、アイソマップ、最大分散展開、局所線形埋め込み、ｔ－ＳＮＥ、非負のマトリックス因子分解、カーネル主成分分析、グラフベースのカーネル主成分分析、線形判別分析（ＬＤＡ）、一般化判別分析、一様多様体近似及び投影（ＵＭＡＰ）、ＬａｒｇｅＶｉｓ、Ｌａｐｌａｃｉａｎ Ｅｉｇｅｎｍａｐ、拡散マップ、ネットワーク（例えば、ニューラルネットワーク）技術、及び／又はフィッシャーの線形判別分析である。例えば、Ｆｏｄｏｒ，２００２，“Ａ ｓｕｒｖｅｙ ｏｆ ｄｉｍｅｎｓｉｏｎ ｒｅｄｕｃｔｉｏｎ ｔｅｃｈｎｉｑｕｅｓ，”Ｃｅｎｔｅｒ ｆｏｒ Ａｐｐｌｉｅｄ Ｓｃｉｅｎｔｉｆｉｃ Ｃｏｍｐｕｔｉｎｇ，Ｌａｗｒｅｎｃｅ Ｌｉｖｅｒｍｏｒｅ Ｎａｔｉｏｎａｌ，Ｔｅｃｈｎｉｃａｌ Ｒｅｐｏｒｔ ＵＣＲＬ－ＩＤ－１４８４９４、Ｃｕｎｎｉｎｇｈａｍ，２００７，“Ｄｉｍｅｎｓｉｏｎ Ｒｅｄｕｃｔｉｏｎ，”Ｕｎｉｖｅｒｓｉｔｙ Ｃｏｌｌｅｇｅ Ｄｕｂｌｉｎ，Ｔｅｃｈｎｉｃａｌ Ｒｅｐｏｒｔ ＵＣＤ－ＣＳＩ－２００７－７，Ｚａｈｏｒｉａｎ ｅｔ ａｌ．，２０１１，“Ｎｏｎｌｉｎｅａｒ Ｄｉｍｅｎｓｉｏｎａｌｉｔｙ Ｒｅｄｕｃｔｉｏｎ Ｍｅｔｈｏｄｓ ｆｏｒ Ｕｓｅ ｗｉｔｈ Ａｕｔｏｍａｔｉｃ Ｓｐｅｅｃｈ Ｒｅｃｏｇｎｉｔｉｏｎ，”Ｓｐｅｅｃｈ Ｔｅｃｈｎｏｌｏｇｉｅｓ．ｄｏｉ：１０．５７７２／１６８６３．ＩＳＢＮ ９７８－９５３－３０７－９９６－７、及びＬａｋｓｈｍｉ ｅｔ ａｌ．，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を参照されたく、それらの各々は参照により本明細書に組み込まれる。したがって、いくつかの実施形態において、次元数削減は、主成分分析（ＰＣＡ）であり、それぞれの抽出された次元数削減成分の各々は、ＰＣＡによって導出されたそれぞれの主成分を含む。そのような実施形態にお

いて、複数の主成分における主成分の数は、ＰＣＡによって計算された主成分の閾値数に制限され得る。主成分の閾値数は、例えば、少なくとも５、少なくとも１０、少なくとも２０、少なくとも５０、少なくとも１００、少なくとも１０００、少なくとも１５００、又は任意の他の数であり得る。いくつかの実施形態において、ＰＣＡによって計算された各主成分は、ＰＣＡによって固有値を割り当てられ、第１の複数の抽出された特徴の対応するサブセットは、最も高い固有値を割り当てられた主成分の閾値数に限定される。複数の細胞構成要素ベクトルにおけるそれぞれの細胞構成要素ベクトルの各々について、複数の次元数削減コンポーネントをそれぞれの細胞構成要素ベクトルに適用して、複数の次元数削減コンポーネントにおけるそれぞれの次元数削減コンポーネントの各々についての次元数削減コンポーネント値を含む対応する次元数削減ベクトルを形成する。これは、複数の細胞構成要素ベクトルから、対応する複数の次元数削減ベクトルを形成し、それによって、潜在表現に配置された複数の細胞構成要素モジュールを形成する。

【０４８５】

いくつかの実施形態において、方法は、潜在表現で配置された複数の細胞構成要素モジュールを使用して多様体学習を実施することを更に含む。一般に、多様体学習は、データセットにおける最大変動を決定することによって、高次元データの低次元構造を説明するために使用される。例としては、限定されないが、力指向レイアウト（Ｆｒｕｃｈｔｅｒｍａｎ，Ｔ．Ｍ．，＆ Ｒｅｉｎｇｏｌｄ，Ｅ．Ｍ．（１９９１）．Ｇｒａｐｈ ｄｒａｗｉｎｇ ｂｙ ｆｏｒｃｅ－ｄｉｒｅｃｔｅｄ ｐｌａｃｅｍｅｎｔ．Ｓｏｆｔｗａｒｅ：Ｐｒａｃｔｉｃｅ ａｎｄ ｅｘｐｅｒｉｅｎｃｅ，２１（１１），１１２９－１１６４）（例えば、Ｆｏｒｃｅ Ａｔｌａｓ ２）、ｔ分布型確率的近傍埋め込み法（ｔ－ＳＮＥ）、局所線形埋め込み（Ｒｏｗｅｉｓ，Ｓ．Ｔ．，＆ Ｓａｕｌ，Ｌ．Ｋ．（２０００）．Ｎｏｎｌｉｎｅａｒ ｄｉｍｅｎｓｉｏｎａｌｉｔｙ ｒｅｄｕｃｔｉｏｎ ｂｙ ｌｏｃａｌｌｙ ｌｉｎｅａｒ ｅｍｂｅｄｄｉｎｇ．Ｓｃｉｅｎｃｅ，２９０（５５００），２３２３－２３２６）、局所線形アイソメトリックマッピング（ＩＳＯＭＡＰ，Ｔｅｎｅｎｂａｕｍ，Ｊ．Ｂ．，Ｄｅ Ｓｉｌｖａ，Ｖ．，＆ Ｌａｎｇｆｏｒｄ，Ｊ．Ｃ．（２０００）．Ａ ｇｌｏｂａｌ ｇｅｏｍｅｔｒｉｃ ｆｒａｍｅｗｏｒｋ ｆｏｒ ｎｏｎｌｉｎｅａｒ ｄｉｍｅｎｓｉｏｎａｌｉｔｙ ｒｅｄｕｃｔｉｏｎ．Ｓｃｉｅｎｃｅ，２９０（５５００），２３１９－２３２３）、カーネルＰＣＡ、グラフベースのカーネルＰＣＡ、親和性ベースの軌道埋め込みのための熱拡散の可能性（Ｐｏｔｅｎｔｉａｌ ｏｆ Ｈｅａｔ－Ｄｉｆｆｕｓｉｏｎ ｆｏｒ Ａｆｆｉｎｉｔｙ Ｂａｓｅｄ Ｔｒａｊｅｃｔｏｒｙ Ｅｍｂｅｄｄｉｎｇ）（ＰＨＡＴＥ）、一般化判別分析（ＧＤＡ）、一様多様体近似及び投影（ＵＭＡＰ）、又はカーネル判別分析が挙げられるが、これらに限定されない。判別分析は、特に、各細胞の特定の細胞型についていくつかの情報が事前に知られている場合に使用され得る。力指向レイアウトは、基礎となる細胞プロセスから生じる基礎となるデータの非線形の態様を符号化する新しい低次元を識別する能力のために、様々な特定の実施形態において有用である。力指向レイアウトは、データを最もよく表す縮小された次元を決定するための機構として、物理ベースのモデルを使用する。例として、力指向レイアウトは、本実施形態において、１つ以上の第１のデータセットにおける各細胞に「反発」力が割り当てられ、第１の複数の細胞にわたって計算されるときに、これらの競合する「力」の下で一緒に「拡散」するデータのセクタを識別するグローバルな「重力」が存在する物理シミュレーションの形態を使用する。力指向レイアウトは、データの構造についての仮定をほとんど行わず、ノイズ除去アプローチを課さない。

【０４８６】

多様体学習は、例えば、Ｗａｎｇ ｅｔ ａｌ．，２００４，“Ａｄａｐｔｉｖｅ Ｍａｎｉｆｏｌｄ Ｌｅａｒｎｉｎｇ，”Ａｄｖａｎｃｅｓ ｉｎ Ｎｅｕｒａｌ Ｉｎｆｏｒｍａｔｉｏｎ Ｐｒｏｃｅｓｓｉｎｇ Ｓｙｓｔｅｍｓ １７に更に記載され、これはその全体が参照により本明細書に組み込まれる。

【０４８７】

いくつかの実施形態において、複数の共変量は、細胞バッチ、細胞ドナー、細胞型、疾患状態、又は化学化合物への曝露を含む。いくつかの実施形態において、複数の共変量は、第２の複数の細胞における１つ以上の細胞に関連する時点、複製、及び／又は関連するメタデータの１つ以上の表示を含む。いくつかの実施形態において、複数の共変量は、実験データ（例えば、フローサイトメトリーの読み出し、イメージング及び顕微鏡注釈、細胞構成要素データなど）を含む。いくつかの実施形態において、複数の共変量は、第２の複数の細胞における１つ以上の細胞に特徴的な１つ以上の遺伝子マーカー（例えば、コピー数バリエーション、単一ヌクレオチドバリアント、多ヌクレオチド多型、挿入、欠失、遺伝子融合、マイクロサテライト不安定性状態、増幅、及び／又はアイソフォーム）を含む。いくつかの実施形態において、複数の共変量は、第２の複数の細胞における１つ以上の細胞についての細胞表現型、細胞挙動、疾患状態、遺伝子変異、遺伝子若しくは遺伝子産物の摂動（例えば、ノックダウン、サイレンシング、過剰発現など）、及び／又は曝露条件のうちの１つ以上を含む。

【０４８８】

例えば、いくつかの実施形態において、共変量は、曝露条件下での化合物への第２の複数の細胞における細胞の曝露又は曝露に対する応答である。いくつかの実施形態において、細胞の曝露は、１つ以上の化合物での細胞の任意の処理を含む。いくつかの実施形態において、１つ以上の化合物は、例えば、小分子、生物製剤、治療剤、タンパク質、小分子と組み合わされたタンパク質、ＡＤＣ、核酸（例えば、ｓｉＲＮＡ、干渉ＲＮＡ、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ｓｈＲＮＡ、ｃＤＮＡ過剰発現野生型及び／若しくは変異体ガイドＲＮＡ（例えば、Ｃａｓ９系若しくは他の細胞成分編集系）など）、並びに／又は前述のいずれかの任意の組み合わせを含む。いくつかの実施形態において、曝露条件は、曝露期間、化合物の濃度、又は曝露期間及び化合物の濃度の組み合わせである。

【０４８９】

いくつかの実施形態において、共変量は、１つ以上の細胞（例えば、ペルターバゲン）において細胞状態遷移及び／又は摂動シグネチャを誘導する１つ以上の細胞に適用される化合物である。

【０４９０】

いくつかの実施形態において、共変量は、複数の細胞構成要素における細胞構成要素、又は第２の複数の細胞における細胞と関連付けられた知識用語（例えば、注釈）である。例えば、いくつかの実施形態において、共変量は、ゲノムワイド関連研究（ＧＷＡＳ）注釈、遺伝子セット濃縮アッセイ（ＧＳＥＡ）注釈、遺伝子オントロジー注釈、機能的及び／若しくはシグナル伝達経路注釈、並びに／又は細胞シグネチャ注釈である。いくつかの実施形態において、共変量は、ＮＩＨ遺伝子発現オムニバス（ＧＥＯ）、ＥＢＩ　ＡｒｒａｙＥｘｐｒｅｓｓ、ＮＣＢＩ、ＢＬＡＳＴ、ＥＭＢＬ－ＥＢＩ、ＧｅｎＢａｎｋ、Ｅｎｓｅｍｂｌ、ＫＥＧＧ経路データベース、及び／又は任意の疾患特異的データベースを含むが、これらに限定されない、当該技術分野で既知の任意の公知の知識データベースから得られる。いくつかの実施形態において、共変量は、摂動（例えば、小分子）誘導遺伝子発現シグネチャを提供するデータベース、例えば、Ｌｉｂｒａｒｙ　ｏｆ　Ｉｎｔｅｇｒａｔｅｄ　Ｎｅｔｗｏｒｋ－ｂａｓｅｄ　Ｃｅｌｌｕｌａｒ　Ｓｉｇｎａｔｕｒｅｓ（ＬＩＮＣＳ）Ｌ１０００データセットから得られる。例えば、Ｄｕａｎ，２０１６，“Ｌ１０００ＣＤＳ^{2}：Ａｎ　ｕｌｔｒａ－ｆａｓｔ　ＬＩＮＣＳ　Ｌ１０００　Ｃｈａｒａｃｔｅｒｉｓｔｉｃ　Ｄｉｒｅｃｔｉｏｎ　Ｓｉｇｎａｔｕｒｅ　Ｓｅａｒｃｈ　Ｅｎｇｉｎｅ，”Ｓｙｓｔｅｍｓ　Ｂｉｏｌｏｇｙ　ａｎｄ　Ａｐｐｌｉｃａｔｉｏｎｓ　２，ａｒｔｉｃｌｅ　１６０１５を参照されたく、これは参照によりその全体が本明細書に組み込まれる。

【０４９１】

いくつかの実施形態において、複数の共変量は、少なくとも３個、少なくとも５個、少なくとも１０個、少なくとも１５個、少なくとも２０個、少なくとも３０個、少なくとも

４０個、少なくとも５０個、少なくとも６０個、少なくとも７０個、少なくとも８０個、少なくとも９０個、少なくとも１００個、少なくとも２００個、少なくとも３００個、少なくとも４００個、少なくとも５００個、少なくとも６００個、少なくとも７００個、少なくとも８００個、少なくとも９００個、少なくとも１０００個、少なくとも２０００個、又は少なくとも３０００個の共変量を含む。いくつかの実施形態において、複数の共変量は、５０００個以下、１０００個以下、５００個以下、２００個以下、１００個以下、５０個以下、又は２０個以下の共変量を含む。いくつかの実施形態において、複数の共変量は、３～１０個、１０～５０個、２０～５００個、２００～１０００個、又は１０００～５０００個の共変量を含む。いくつかの実施形態において、複数の共変量は、３個以上の共変量から始まり、５０００個以下の共変量で終わる別の範囲内にある。

【０４９２】

いくつかの実施形態において、複数の共変量における共変量の各々は、細胞状態遷移及び／又は摂動シグネチャを誘導する１つ以上の細胞に適用される化合物であり、複数の共変量は複数の化合物である。いくつかの実施形態において、複数の共変量は、上記「化合物」と題されるセクションに開示されるように、複数の化合物からなる。

【０４９３】

図５は、細胞構成要素カウントデータ構造（例えば、目的の生理学的状態を通知する複数の共変量を集合的に表す第２の複数の細胞を使用して得られる）と、複数の細胞構成要素又はその表現を共通次元として使用する潜在表現とを組み合わせることによって形成される例示的な活性化データ構造を示す。これを達成するために、いくつかの実施形態において、第２の複数の細胞についてのカウントマトリックス（例えば、図４に示される第１の複数の細胞についてのカウントマトリックスと構造が類似している）及び潜在表現が一緒に乗算され、潜在表現マトリックスの重みがカウントマトリックスの正規化されたカウントによって乗算されるようにする。一般に、２つのマトリックスは、共通次元（例えば、第１のマトリックスのｘ軸及び第２のマトリックスのｙ軸）によって一緒に乗算され得る。第１及び第２のマトリックスのそれらの共通次元によるマトリックス乗算は、第１のマトリックス及び／若しくは第２のマトリックスに代替的に、又はそれに加えて、訓練されていないか、又は部分的に訓練されたモデルに適用することができる補助データの第３のマトリックスをもたらす。

【０４９４】

したがって、いくつかのそのような実施形態において、カウントマトリックスは、次元ｎ＿細胞ｘ　ｎ＿遺伝子を有し、潜在表現は、次元ｎ＿遺伝子ｘ　ｎ＿モジュールを有し、ｎ＿細胞は、第２の複数の細胞における細胞の数であり、ｎ＿遺伝子は、複数の細胞構成要素における細胞構成要素（例えば、遺伝子）の数、又はその表現であり、ｎ＿モジュールは、複数の細胞構成要素モジュールにおけるモジュールの数である。これは、カウントマトリックスにおける細胞構成要素の存在量を、各細胞（例えば、目的の１つ以上の共変量に対応する）がそのモジュール活性化によって特徴付けられ、得られたマトリックス表現（例えば、活性化データ構造）が、（例えば、ｎ＿遺伝子の共通次元を乗算した後の）次元ｎ＿細胞ｘ　ｎ＿モジュールを有する空間にマッピングする。

【０４９５】

例えば、マトリックス乗算を使用する潜在表現及び細胞構成要素カウントデータ構造の組み合わせ、並びにマトリックス形態での結果として生じる活性化データ構造は、図５にまとめて示される。潜在表現（図５の上部パネルに示される）は、次元Ｚ×Ｋを有し、ここで、Ｚは、細胞構成要素の数又はその表現であり、Ｋは、細胞構成要素モジュールの数である。細胞構成要素カウントデータ構造（左下のパネルに示される）は、次元Ｇ×Ｚを有し、ここで、Ｇは、第２の複数の細胞における細胞の数であり、潜在表現に関して、Ｚは、細胞構成要素の数又はその表現である。Ｚ（細胞構成要素の数又はその表現）を共通次元として使用するマトリックス乗算による組み合わせは、次元Ｇ×Ｋを有する結果として生じる活性化データ構造を生成する。それぞれの行の各々におけるそれぞれの列の各々についての各エントリーは、それぞれの列に対応する第２の複数の細胞におけるそれぞれ

の細胞におけるそれぞれの細胞構成要素モジュールの各々の活性化を示す活性化重みである。したがって、図５に示されるように、モジュール１に対応するカウントは、細胞１に対応する活性化重み$_{1-1}$、細胞Ｇに対応する活性化重み$_{1-G}$などを含む。

【０４９６】

いくつかの実施形態において、活性化データ構造における複数の活性化重みは、差次的モジュール活性化を含む。いくつかの実施形態において、差次的モジュール活性化（例えば、活性化データ構造における第２の複数の細胞における細胞間のそれぞれのモジュールの差次的活性化重み）は、関数（ｍｕ＿１－ｍｕ＿２）／（ｖａｒ＿１＋ｖａｒ＿２）$^{-0.5}$を使用してｖ－スコアを計算することによって得られ、ｍｕ＿ｉは、それぞれの条件ｉ（例えば、共変量ｉ）を有する細胞にわたるモジュール活性化の手段を示し、ｖａｒ＿ｉは、条件ｉにおけるモジュール活性化の分散を示す。Ｖ－スコアは、分母内の細胞の数によって正規化されないｔ－スコアとして説明することができる。

【０４９７】

いくつかの実施形態において、活性化データ構造における第２の複数の細胞におけるそれぞれの細胞の各々は、それぞれの共変量を表す。いくつかの実施形態において、活性化データ構造における第２の複数の細胞におけるそれぞれの細胞の各々は、細胞状態遷移及び／又は摂動シグネチャを誘導する１つ以上の細胞に適用されるそれぞれの化合物を表す。

【０４９８】

したがって、いくつかの実施形態において、活性化データ構造は、第２の複数の細胞によって表される複数の化合物における各化合物への曝露に対応する（例えば、相関する、及び／又はそれに応答する）それぞれの細胞構成要素モジュールの活性化（例えば、活性化のレベル又は程度）を示す。例えば、第２の複数の細胞におけるそれぞれの細胞の各々がそれぞれのペルターバゲン（例えば、１つ以上の細胞が曝露される化合物並びに／又は細胞状態遷移及び／若しくは摂動シグネチャを誘導する化合物）を表すいくつかの実施形態において、活性化データ構造は、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの化合物による処置に相関する及び／又はそれに応答して、それぞれの細胞構成要素モジュールの活性化（例えば、誘導及び／又は差次的発現）を示すそれぞれの活性化重みを含む。

【０４９９】

いくつかの実施形態において、候補細胞構成要素モデルは、上記「モデルアーキテクチャ」と題されるセクションに記載されるように、本明細書に開示されるモデルアーキテクチャのいずれをも含む。

【０５００】

いくつかの実施形態において、候補細胞構成要素モデルは、オートエンコーダ、スパースオートエンコーダ、及び／又はスパースマルチ読み出し、知識結合オートエンコーダである。いくつかの実施形態において、候補細胞構成要素モデルは、半教師ありモデルである。いくつかの実施形態において、候補細胞構成要素モデルは、１層ニューラルネットワーク（例えば、ＳｏｆｔＭａｘ及び／又はロジスティック回帰モデル）である。いくつかの実施形態において、候補細胞構成要素モデルは、一次元Ｈｕｂｅｒ　Ｏｕｔｌｉｅｒ　Ｒｅｇｒｅｓｓｏｒモデルである。

【０５０１】

いくつかの実施形態において、候補細胞構成要素モデルは、複数の層を含むスパースマルチ読み出し、知識結合オートエンコーダであり、第１の層は潜在表現を得るために使用され、第２の層は細胞構成要素モジュール知識構築物（例えば、共変量重みマトリックス）を得るために使用される。

【０５０２】

いくつかの実施形態において、候補細胞構成要素モデルを訓練することは、マルチタスク策定におけるカテゴリ交差エントロピー損失を使用して実施され、複数の共変量における共変量の各々が、複数のコスト関数におけるコスト関数に対応し、複数のコスト関数に

おけるそれぞれのコスト関数の各々が、共通の重み付け係数を有する。

【０５０３】

いくつかの実施形態において、候補細胞構成要素モデルを訓練して、目的の生理学的状態と関連付けられた細胞構成要素モジュールのセットにおける第１の細胞構成要素モジュールを識別するようにモデルを訓練する。モデルを訓練するための方法については、本明細書に更に詳細に記載される。本明細書に開示される方法及び／又は実施形態のいずれかは、上記「モデル訓練」と題されるセクションに記載されるように、候補細胞構成要素モデルの訓練において使用することが企図される。

【実施例】

【０５０４】

Ｖ．実施例

本明細書に提供されるのは、化合物を生理学的状態と関連付けるためのモデルの例示的な性能尺度及び治療上の適用である。

【０５０５】

実施例１．脂肪酸関連細胞プロセスの活性化のための化学構造の予測。

この実施例では、細胞構成要素モジュールを最初に定義した。これは、細胞が目的の生理学的状態と関連付けられた異なる状態を表す細胞についての発現データを得ることによって行われた。これは、当初出願された請求項２７を追跡する。細胞構成要素存在量値は、細胞の各々から測定され、このデータは、細胞構成要素をクラスタ化するために使用される。細胞によって表される様々な状態にわたって発現値が互いに相関しているこれらの細胞構成要素を、細胞構成要素モジュールにグループ化する。これは、いくつかの細胞構成要素モジュールをもたらし、それらの各々は、細胞構成要素試料の異なるサブセットを含む。いくつかの実施形態において、細胞構成要素モジュールの各々は、細胞構成要素の異なるサブセットを有するが、ある細胞構成要素モジュールにおける細胞構成要素と別の細胞構成要素モジュールにおける細胞構成要素との間に重複がある可能性がある。

【０５０６】

更に、この例では、追加の訓練データが、第２の訓練セットの形態で得られる。この第２の訓練セットはまた、細胞構成要素についての単一細胞存在量データを含む。しかしながら、この第２の訓練セットにおいて、各細胞は、複数の訓練化学化合物において異なる化学化合物に曝露されている。この訓練セットにおいて、既知の量は、それぞれ異なる化学化合物のフィンガープリントであり、そのような化合物に曝露された細胞の得られる細胞構成要素存在量データである。第２のデータセットについてのデータは、細胞構成要素同一性のための第１の軸、及び細胞同一性のための第２の軸を有する、カウントマトリックス５０２（図５に示される）として配置することができる。したがって、カウントマトリックス５０２における各エレメントは、所与の細胞内の所与の細胞構成要素の存在量である。更に、（特定の細胞に対応する）カウントマトリックス５０２におけるそれぞれの列の各々は、特定の細胞が曝露された特定の化合物で標識される。したがって、カウントマトリックス５０２の各列は、特定の化合物（例えば、訓練化合物）で標識されるが、各エレメントは、対応する細胞（Ｘ軸）についての対応する細胞構成要素（Ｙ軸）のカウントである。

【０５０７】

図５に示されるように、第１のデータセット（潜在表現４０４）及び第２のデータセット（カウントマトリックス５０２）からのデータは、組み合わされて、活性化データ構造（例えば、図５に示されるような活性化データ構造５０４）を形成する。例えば、これを達成する１つの方法は、第１の軸が細胞構成要素モジュールを表し、第２の軸が細胞構成要素の各々を表すように、潜在表現４０４における行として細胞構成要素モジュールを配置することである。このようにして、活性化データ構造５０４を生成するために、潜在表現４０４及びカウントマトリックス５０２は、マトリックス乗算を介して、それらの共通軸、細胞構成要素の数によって乗算されて、活性化データ構造５０４に到達する。活性化データ構造５０４は、カウントマトリックス５０２からの細胞同一性軸及び潜在表現５０

４からの細胞構成要素モジュール軸を保持する。異なる細胞型に対して異なる活性化構造を形成することができる。すなわち、カウントマトリックス５０２を形成するために使用される細胞は、目的の特定の疾患状態を表すことができる。したがって、異なる疾患状態又は目的の他の表現型について、異なる活性化データ構造５０４を形成することができる。

【０５０８】

図６を参照すると、いくつかの例では、活性化データ構造５０４の各行（図５から、現在は図６の上部にある）は、異なるモデル６０１についての訓練データとして機能する。例えば、モデル６０１が行６０４－１の重み（重み$_{1-1}$から重み$_{1-W}$）を含み、化合物１からＷがそれぞれ細胞構成要素モジュール１を活性化する程度を表す場合を考慮する。このモデル６０１は、活性化データ構造５０４の行６４０のエレメントについて訓練され、これは、訓練化合物１、．．．、Ｇの各々が細胞構成要素モジュール１を活性化する程度を提供する。この訓練では、まず、細胞１が曝露された化合物のフィンガープリント表現がモデル６０１に入力される。この入力に応答して、細胞構成要素モジュール１についてのモデル６０１は、Ｐｒｅｄと呼ばれる活性化値を出力する。図６の命名法における値$_{1}$。この出力活性化値は、活性化データ構造５０４のＡｃｔ$_{1-1}$である、実際の活性化値と比較される。次に、細胞２が曝露された化合物のフィンガープリント表現がモデル６０１に入力される。この入力に応答して、モデルは活性化値（Ｐｒｅｄ．値$_{2}$）を出力する。この出力活性化値は、活性化データ構造５０４のＡｃｔ$_{1-2}$である、化合物２についての実際の活性化値と比較される。このプロセスは、細胞Ｇを通じて進行する。細胞Ｇが曝露された化合物のフィンガープリント表現は、モデル６０１に入力される。これに応答して、モデルは活性化値（Ｐｒｅｄ．値$_{G}$）を出力する。この出力活性化値は、活性化データ構造５０４のＡｃｔ$_{1-G}$である、細胞Ｇについての実際の活性化値と比較される。この例では、Ｗ及びＧは同じ値を有する。このようにして、細胞構成要素モジュール１について図５に概説されるように、活性化データ構造を導出するために使用される化合物の訓練セットにおける各化合物について結果として生じる予測（Ｐｒｅｄ．値）が存在する。（活性化値の）上述の計算された予測を、これらの化合物の各々について上述の実際の活性化値と比較し、予測された活性化値と実際の活性化値との間の差を使用して、逆伝搬及び関連するモデル改良技術を使用してモデル６０１を更に訓練する。

【０５０９】

したがって、結果は、細胞構成要素モジュールの各々のうちの１つである、一連の訓練されたモデル６０１である。試験化合物のフィンガープリントは、訓練されたモデルの各々に入力され得、それぞれの訓練されたモデル６０１の各々は、予測された活性化値を出力し、その大きさは、それぞれの訓練されたモデルに対応する細胞構成要素モジュールが試験化合物によって活性化されるかどうかを示す。ここで、プロセスの概要が説明されたので、ステップの各々は、この例で使用される実験データと併せて説明される。

【０５１０】

以下のプロセスによって、第１の細胞構成要素モジュール（図１、図４　１３２－１）を識別する。電子形式で１つ以上の第１のデータセットを得る。１つ以上の第１のデータセットは、複数の注釈付き（例えば、標識されたか、又は既知の）細胞状態を集合的に表す第１の複数の細胞（例えば、２０個以上の細胞）に関するデータを含む。第１のデータセットは、第１の複数の細胞におけるそれぞれの細胞の各々について、複数の細胞構成要素（例えば、１０個以上の細胞構成要素）におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含む。例えば、各細胞についての転写データ。このようにして、複数のベクトルがアクセスされるか、又は形成される。複数のベクトルにおけるそれぞれのベクトルの各々は、複数の構成要素におけるそれぞれの細胞構成要素に対応し、対応する複数のエレメントを含む。ベクトルの対応する複数のエレメントにおけるそれぞれのエレメントの各々は、第１の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を表す対応するカウントを有する。したがって、いくつかのそのような実施形態において、複数の細胞状

態における細胞状態の各々についての転写データが得られる。

【０５１１】

例示するために、図４に例示される形態のカウントマトリックス４０２が形成される。この例では、前脂肪細胞において代謝活性プロセスを誘導することが知られている小分子ペルターバゲンを使用した。前脂肪細胞株のアリコートをペルターバゲンに２４時間曝露し、摂動状態における細胞株の曝露アリコートについてｓｃＲＮＡ－ｓｅｑ読み出しを得た。ｓｃＲＮＡ－ｓｅｑ読み出しは、ペルターバゲンに曝露されていない細胞株のアリコートについても得られ、これらの読み出しは、対照条件を表している。このようにして、図１４Ａのブロック１５０４に従って、第１のデータセットを得、これは、第１の複数の細胞におけるそれぞれの細胞の各々について、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含み、それによって、複数のベクトルにアクセスしたか、又はそれを形成した。すなわち、ペルターバゲンに曝露された細胞及びペルターバゲンに曝露されなかった細胞（対照細胞）の両方で測定された各細胞構成要素（例えば、遺伝子）の発現値は、図４に例示されるカウントマトリックス４０２のエレメントを形成した。図４に例示され、図１４Ａのブロック１５１０に記載されるように、カウントマトリックス４０２は、各細胞構成要素についてのベクトルを含み、したがって、複数のベクトルが存在する。複数のベクトルにおけるそれぞれのベクトルの各々は、（ｉ）複数の構成要素におけるそれぞれの細胞構成要素に対応し、（ｉｉ）対応する複数のエレメントを含む。

【０５１２】

例えば、細胞構成要素１（例えば、遺伝子１）について、カウント１－１、．．．．、カウント１－Ｎは、細胞１からＮにおける遺伝子１の発現の測定値であり、Ｎ細胞のいくつかは、ペルターバゲンに曝露されており、いくつかは、ペルターバゲンに曝露されておらず、これらのカウントは、細胞構成要素１についてのベクトルのエレメントを形成する。すなわち、図１４Ａのブロック１５１２に従って、細胞構成要素１のベクトルの対応する複数のエレメントにおけるそれぞれのエレメントの各々は、第１の複数の細胞におけるそれぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を表す対応するカウントを有する。この例は、２つの状態（ペルターバゲンに曝露されたか、又はされていない）を含むが、原則として、異なる濃度のペルターバゲン、曝露時間などの、任意の数の状態を包含することができる。

【０５１３】

図１４Ａのブロック１５１４によれば、この実施例１には、対照（ペルターバゲンへの曝露なし）及びペルターバゲンの曝露という２つの注釈付きの状態がある。すなわち、複数の注釈付きの細胞状態における１つの注釈付きの細胞状態は、曝露条件（例えば、曝露期間、ここでは２４時間）下での化合物（ここでは、ペルターバゲン）への第１の複数の細胞における細胞の曝露である。この例は、２つの状態（ペルターバゲンに曝露されたか、又はされていない）からなるが、原則として、異なる濃度のペルターバゲン、曝露時間などの、任意の数の状態を包含することができる。

【０５１４】

カウントマトリックス４０２は、フィルタリング及び正規化ステップを介して前処理され、高いシグナル対ノイズ比を有するいくつかの遺伝子を含有する前処理されたカウントマトリックスをもたらした。

【０５１５】

複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別する。複数の細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々は、複数の細胞構成要素のサブセットを含む。複数の細胞構成要素モジュールは、（ｉ）複数の候補細胞構成要素モジュール、及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された潜在表現で配置され、複数の細胞構成要素モジュールは、１０を超える細胞構成要素モジュールを含む。

【０５１６】

いくつかの実施形態において、候補細胞構成要素モジュールの各々は、候補転写フィンガープリントである。

【０５１７】

この実施例では、カウントマトリックス４０２を使用して、細胞構成要素モジュール１３２を識別した。これは、図１４Ｂのブロック１５２６に従って行われた。複数のベクトル（図４のカウントマトリックス４０２の各行）を使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別し、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々は、複数の細胞構成要素のサブセットを含む。

【０５１８】

これにより、（ｉ）複数の候補細胞構成要素モジュール、及び（ｉｉ）複数の細胞構成要素、又はその表現によって次元決定された潜在表現が得られ、複数の候補細胞構成要素モジュールは、１０を超える細胞構成要素モジュールを含む。この潜在表現の例は、図４の潜在表現４０４であり、それぞれの候補細胞構成要素モジュール１３２の各々について、どの細胞構成要素がそれぞれの候補細胞構成要素モジュール内にあるかを示す。

【０５１９】

潜在表現４０４は、図１４Ｃのブロック１５２８に従って形成され、複数のベクトル（カウントマトリックス４０２の細胞構成要素ベクトル）を使用して、複数のベクトルにおけるベクトルの各々の対応する複数のエレメントの各々を使用して複数のベクトルに相関モデルを適用することによって、複数の候補細胞構成要素モジュール（潜在表現４０４の）における候補細胞構成要素モジュールの各々を識別した。特に、相関ベースのコスト関数が最適化され、これは、細胞構成要素ベクトル間の最近傍関係を定義する最近傍グラフを計算し、カウントマトリックス４０２の細胞構成要素ベクトル間の相関を計算することに相当した。複数の細胞にわたって互いに高い相関を有する細胞構成要素（ここでは遺伝子）は、最終的に最近傍となり、ライデン（Ｌｅｉｄｅｎ）又は任意の他のグラフクラスタリング方法を使用してグラフをクラスタリングすることによって、潜在表現４０２内に細胞構成要素モジュールを形成した。スパースオートエンコーダコスト関数を最適化することは、ｐｙｔｏｒｃｈ又はｔｅｎｓｏｒｆｌｏｗに実装されているような標準訓練を使用して、その重みのＬ０正規化、及び再構築損失を伴う１層オートエンコーダの訓練に相当した）。この実施例では、これは、訓練中に１０８個の細胞構成要素モジュールが学習されることをもたらした。すなわち、図４の潜在表現４０４は、１０８個の細胞構成要素モジュール１３２を有し、各々が、カウントマトリックス４０２において発現データが利用可能であった細胞構成要素の独立したサブセットを有する。

【０５２０】

１０８個の細胞モジュールのうち、「モジュール７８」と称される細胞構成要素モジュール１３２は、摂動試料及び対照試料にわたって計算された細胞構成要素の各々についてのｔスコアを平均化するときに、最も強い活性化を示した。言い換えれば、カウントマトリックスデータ内の発現データを使用して、潜在表現４０４内のそれぞれの細胞構成要素モジュールの各々について、ペルターバゲンに曝露された細胞とペルターバゲンに曝露されていない細胞との間のそれぞれの細胞構成要素モジュールにおける細胞構成要素の各々の差次的発現に関するｔスコアを実施することによって、細胞構成要素を検証した。更に、モジュール７８は、脂肪酸及び脂質に関連する生物学的プロセスに関与する細胞構成要素で濃縮されている。要約すると、モジュール７８は、代謝活性のマーカーである、ＦＡＢＰ３を含む、２８個の遺伝子からなる。

【０５２１】

細胞構成要素モジュールに加えて、細胞を訓練化合物に曝露したときの細胞ベースの細胞構成要素応答データが必要である。

【０５２２】

したがって、１つ以上の第２のデータセットを電子形式で得た。１つ以上の第２のデータセットは、第２の複数の細胞からのデータを含む。第２の複数の細胞は、２０個以上の

細胞を含む。第２の複数の細胞は、目的の生理学的状態を通知する複数の共変量を集合的に表した。例えば、複数の共変量は、いくつかの例では、訓練化合物である。次に、第２の複数の細胞における細胞の各々について、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量が取得され、それによって、（ｉ）第２の複数の細胞、及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された細胞構成要素カウントデータ構造が得られる。

【０５２３】

これは、図１４Ｃのブロック１５３８に従っており、第２のデータセットが電子形式で得られた状態は、第２の複数の細胞におけるそれぞれの細胞の各々について、第２の複数の細胞が、２０個以上の細胞を含み、目的の生理学的状態を通知する複数の共変量（ここでは複数の異なる化学化合物）を集合的に表し、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含み、それによって、（ｉ）第２の複数の細胞、及び（ｉｉ）複数の細胞構成要素又はその表現によって次元決定された細胞構成要素カウントデータ構造を得る。

【０５２４】

このカウントマトリックスの形態の例示は、図５のカウントマトリックス５０２である。図５のカウントマトリックス５０２に示されるように、それぞれの細胞構成要素（例えば、遺伝子）の各々について、第２の複数の細胞における細胞の各々についての発現データが存在する。例えば、複数の遺伝子の各々の転写活性を、第２の複数の細胞にわたって測定する。細胞の各々は、共変量、ここでは訓練化学化合物に曝露されている。

【０５２５】

複数の細胞構成要素又はその表現を共通次元として使用して、細胞構成要素カウントデータ構造及び潜在表現を組み合わせることによって活性化データ構造を形成し、活性化データ構造は、複数の細胞構成要素モジュールにおける細胞構成要素モジュールの各々について、第２の複数の細胞における細胞の各々について、それぞれの活性化重みを含む。

【０５２６】

カウントマトリックス５０２は、図５に示される活性化データ構造５０４を得るために、潜在表現４０４によって乗算されたマトリックスであった。活性化データ構造５０４は、それぞれの細胞構成要素モジュールの各々について、第２の複数の細胞における細胞の各々について、活性化値Act_{K-G}を有し、その値は、カウントマトリックス５０２による潜在表現４０４の対応するマトリックス乗算によって決定される。

【０５２７】

（ｉ）活性化データ構造を候補モデルに入力したときに、活性化データ構造内に表される細胞構成要素モジュールの各々における複数の共変量における各共変量の不在又は存在の予測と、（ｉｉ）細胞構成要素モジュールの各々における各共変量の実際の不在又は存在との間の差を使用して、候補細胞構成要素モデルを訓練し、訓練することは、差に応答して、候補細胞構成要素モデルと関連付けられた複数の共変量重みを調整する。

【０５２８】

活性化データ構造５０２は、それ自体が次元Ｎ化合物ｘＭ細胞構成要素モジュールの潜在表現６０２である、図６のモデル６０１についての訓練データ（標識データ）として機能した。この実施例では、８０００個の異なる化合物及び１０８個の細胞構成要素モジュールを考慮した。したがって、図５の命名法では、Ｚは１０８であり、Ｇは８０００であった。活性化データ構造は、２つの方法で訓練及び試験セットに分割された。まず、１２００個の化合物を試験セットにグループ化し、残りの６８００個の化合物を訓練セットにグループ化した「ランダム分割」を選択した。また、「クロス骨格分割」は、試験セットが訓練セットとは異なる骨格を有する化合物を含むことを保証するオープンソースソフトウェアパッケージＲＤＫｉｔの機能を使用して定義された。

【０５２９】

図６に示されるように、活性化データ構造５０４のそれぞれの行の各々は、それぞれの行によって表される対応する細胞構成要素モジュールの細胞構成要素を誘導する可能性が

高い化合物を表すベクトルである。モデル６０１の各インスタンスは、活性化データ構造５０４の行で訓練された。活性化データ構造５０４は、６８００個の訓練化合物を使用して形成された。所与のモデル６０１について、特定の化学化合物のフィンガープリントがモデル６０１に入力され、この入力に応答して、対応する細胞構成要素モジュールについての予測された活性化値が計算される。この予測された活性化値は、活性化データ構造５０４内の対応するエレメントにおける実際の活性化値と直接比較され得る。したがって、このようにして、（ｉ）モデル６０１への活性化データ構造５０４の入力時に活性化データ構造５０４に表される細胞構成要素モジュールの各々についての訓練化合物における各化合物の不在又は存在の予測と、（ｉｉ）細胞構成要素モジュールの各々についての各化合物の実際の不在又は存在との間の差を計算し、差に応答して候補細胞構成要素モデルと関連付けられた複数の共変量重み６０４を調整することによってモデル６０１を訓練するために使用することができる。図６に示されるように、複数の共変量重みは、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの共変量の各々について、それぞれの共変量が、活性化データ構造にわたって、それぞれの細胞構成要素モジュールと相関するかどうかを示す対応する重みを含む。いくつかの実施形態において、細胞構成要素モジュールの各々について異なるモデル６０１が存在した。言い換えれば、図６を参照すると、いくつかの実施形態において、各行６０４は異なるモデル６０１内にある。したがって、そのような実施形態において、そのようなモデル６０１の各々は、活性化データ構造における対応する行（例えば、それぞれのモデル６０１と同じ細胞構成要素モジュールに対応する行）を使用して訓練される。

【０５３０】

図６に示すように、訓練されたモデル６０１（又は複数のモデル）は、各共変量（ここでは、訓練化学組成物）についての重みを提供する。すなわち、モデル６０１の潜在表現６０２は、各共変量（化学組成物）が細胞構成要素モジュールの活性化にどの程度関連するかを説明する重み（例えば、図６の重み$_{1-1}$又は行６０４－１）を提供する。そのような重みは、細胞構成要素モジュールのセットにおけるそれぞれの化合物についてのそれぞれの細胞構成要素モジュールのそれぞれの数値的活性化スコアとみなされる。細胞構成要素モジュールの各々についての異なるモデル６０１が形成される実施形態において、潜在表現６０２は、各モデル６０１の集合潜在表現である。いくつかの実施形態において、表現の各重みは分類的である（例えば、化合物は細胞構成要素モジュール「０」に影響を及ぼすか、又は化合物は細胞構成要素モジュール「１」に影響を及ぼさない。他の実施形態において、各重みは、連続スケール上にあり、スケールの一端は、訓練化合物が細胞構成要素モジュールに大きく影響することを示し、スケールの他端は、訓練化合物が細胞構成要素モジュールに影響しないことを示す。本明細書で使用される場合、「影響する」という用語は、用途依存性であるが、概して、化合物の不在又は存在が、細胞構成要素モジュールにおける細胞構成要素の存在量を変化させることを意味する。

【０５３１】

モデル６０１の訓練のために、この例では、図６の活性化データ構造５０４に表される化合物のＳＭＩＬＥＳ表現は、ＥＣＦＰ４フィンガープリント表現、更にはグラフ表現に変換される。その後、２つのモデルが訓練される。すなわち、モデル６０１は、この例では、２つの異なるモデルのアンサンブルである。Ａ）完全に接続されたニューラルネットワークアーキテクチャを使用してＥＣＦＰ４表現を訓練し、Ｂ）メッセージパッシングニューラルネットワーク（ＭＰＮＮ）を使用してグラフ表現を訓練する。この訓練を実施するために、オープンソースソフトウェアパッケージｐｙｔｏｒｃｈ及びＤＧＬを使用した。訓練されていないモデル６０１は、訓練セットにおけるそれぞれの化合物の各々のそれぞれの化学構造の各々について、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々について、（ｉ）それぞれの化合物の化学構造のフィンガープリントを訓練されていないモデルに入力したときのそれぞれの細胞構成要素モジュールについてのそれぞれの計算された活性化スコアと、（ｉｉ）細胞構成要素モジュールのセットにおけるそれぞれの化合物についてのそれぞれの細胞構成要素モジュールのそれぞれ

の数値的活性化スコア（活性化データ構造５０４から得た）との間のそれぞれの差を使用して訓練され、訓練することが、差に応答して訓練されていないモデル６０１と関連付けられた複数のパラメータを調整し、複数のパラメータが、１００以上のパラメータを含み、それによって、訓練されたモデルを得る。

【０５３２】

上述のように、この例では、モデル６０１は、（ｉ）ＳＭＩＬＥＳ文字列の標準フィンガープリント上の完全に接続されたネットワークであって、ネットワークアーキテクチャは、ＲｅＬＵ活性化を有する３層ネットワークである、完全に接続されたネットワーク、及び（ｉｉ）ＤＧＬライブラリからのＭＰＮＮネットワークのアンサンブルである。化学構造情報を入力すると、モデル６０１は、それが訓練された細胞構成要素モジュール１３２の各々の活性化スコアを提供する。

【０５３３】

実際に、いくつかの実施形態において、この例では、細胞構成要素モジュールの各々のための別個のアンサンブルモデル６０１が存在する。言い換えれば、モデル６０１は、化学構造の入力時に複数の細胞構成要素モジュールの各々に対して別個の活性化スコアを提供するマルチタスクエンコーダであった。なお更に、いくつかの実施形態において、上で説明したように、それぞれの細胞構成要素モジュールの各々について別個のモデル６０１が存在する。そのような実施形態において、そのようなそれぞれのモデル６０１の各々は、対応する細胞構成要素モデルに対する化合物の各々についての活性化重みを含む。

【０５３４】

現在訓練されているそれぞれのモデル６０１の各々は、訓練セットの一部であるか否かにかかわらず、任意の化合物について、その対応する細胞構成要素モジュールについての活性化スコアを提供する。すなわち、各モデル６０１は、その対応する細胞構成要素モジュールが試験化合物と関連付けられているかどうかを報告することができる。それがある場合、モデルは、その対応する細胞構成要素モジュールが試験化合物と関連付けられていることを示すスコアを出力する。いくつかの実施形態において、このスコアは分類的である（例えば、対応する細胞構成要素モジュールが試験化合物と関連付けられている場合は「１」であり、関連付けられていない場合は「０」である）。いくつかの実施形態において、このスコアは、例えば、１に近い数（例えば、０．８５）が、対応する細胞構成要素モジュールが試験化合物と関連付けられている可能性を示す、０～１のスケールでの確率又は尤度である。いくつかの実施形態において、このスコアは、「Ａ」から「Ｂ」の連続スケール上にあり、Ａ及びＢは２つの異なる数である。各々が異なる細胞構成要素モジュールに対応するいくつかのモデル６０１が存在するので、試験化合物をいくつかの異なるモデル６０１に対して実行して、どの細胞構成要素モジュールが化合物によって活性化されるか（それと関連付けられるか）を決定する。各例において、化学構造は、上で説明したようにフィンガープリントに変換され、各モデルに適用されるのはこのフィンガープリントである。生物学的観点から、所与の試験化合物は、任意の数の異なる細胞構成要素モジュール（例えば、１、２、３、４、５、又はそれよりも多い）を活性化し得ることが予想され得ることに留意されたい。更に、本開示に記載のアプローチは、モデル６０１が訓練されていないが、どの細胞構成要素モジュールが試験化合物によって活性化されるべきかが知られている化合物を試験することによって検証することができる。これは、以下に示すようにこの例で行われた。特に、化合物を生理学的状態と関連付けるための訓練されたモデル６０１は、この例では４倍に検証された。この試験は、当初出願された請求項１を追跡する。

【０５３５】

まず、モデル６０１からのモデル予測は、ハイスループットスクリーニングから上記の１２００個のランダムに選択された目に見えない化合物によって、また６８００個の化合物訓練セットに対して上記の１２００個の重複しない骨格を有する化合物によって誘導された脂肪酸生成関連細胞構成要素モジュールの活性化について得られた。ランダムに選択された化合物について得られたそれぞれのモデル６０１の予測（予測された細胞構成要素

活性化スコア）を図１０Ｂに示す。すなわち、図１０Ｂは、２つの異なるモデル６０１、すなわち、１つは細胞構成要素モジュール７８「モジュール７８」、１つは細胞構成要素モジュール「９０」からの結果を示す。モジュール７８は、細胞代謝にとって重要な脂肪酸関連細胞プロセスを表し、その対応する訓練されたモデル６０１は、高い決定係数を示した（R^2＝０．２８）。対照的に、同じｓｃＲＮＡ－ｓｅｑデータセットから学習した細胞構成要素「モジュール９０」のための訓練されたモデル６０１は、細胞代謝とは無関係であり（モジュール９０における細胞構成要素は脂肪酸関連プロセスとは関連しない）、低い決定係数を有した（R^2＝０．０８）。全てのベンチマークは、非常に有意な相関をもたらした（それぞれ、ピアソン相関係数p_s＝約０．５及び約０．２）。

【０５３６】

当初出願された請求項１に記載の言語において、この第１の検証アプローチは、試験化学化合物（ハイスループットスクリーニングから、また６８００個の化合物訓練セットに対する上記の１２００個の重複していない骨格を有する化合物によって、記載された１２００個のランダムに選択された目に見えない化合物のうちの１つ）を目的の生理学的状態（ここでは、この例では、細胞代謝にとって重要な脂肪酸関連細胞プロセス）と関連付ける方法を提供する。この方法は、メモリ及び１つ以上のプロセッサを含むコンピュータシステムにおいて、試験化学化合物の化学構造のフィンガープリントを得ることを含む。したがって、試験化学化合物の化学構造のフィンガープリントが得られ、それはこの例で図１の各モデル６０１に入力されるものである。当初出願された請求項１の文脈において、モデルは、モデルと称される。このモデルは、アンサンブルモデルを包含し、アンサンブルモデルにおける各コンポーネントモデルは、図６のモデル６０１について列挙されたパラメータの単一の行を含み、行は、コンポーネントモデルと関連付けられた所与の細胞構成要素モジュールについての重みに関するパラメータである。図６において、そのような重みは単一の行として表されるが、それらがアンサンブルモデルのコンポーネントモデルにおいて行の形式であるという要件はなく、その任意の等価物は本開示の範囲内であることを理解されたい。更に、図６のモデル６０１は、それが訓練された各化合物についての単一の重みを含み、これは回帰に基づいてモデル６０１に好適であるが、いくつかの実施形態において、モデル６０１の重みの数と、モデルが訓練された化合物の数との間に明確な関係は存在しない。いくつかの実施形態において、モデル６０１は、１００以上、１０００以上、１０，０００以上、又は１００，０００以上のパラメータを含む。

【０５３７】

当初出願された請求項１によれば、試験化合物のフィンガープリントは、モデルに入力される。当初出願された請求項１に記載されているように、モデルは、１００以上のパラメータを含む。言い換えれば、モデル出力の計算は、試験化合物のフィンガープリントを入力すると、精神的に実施することができない。モデルは、フィンガープリントのモデルへの入力に応答して１つ以上の計算された活性化スコアを出力する。１つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、細胞構成要素モジュールのセットにおける対応する細胞構成要素モジュールを表す。この例では、モデルはモデル６０１のアンサンブルであり、各々が異なる細胞構成要素モジュールを表し、したがって、アンサンブル内の各モデル６０１は、細胞構成要素モジュールのセットにおける単一の対応する細胞構成要素モジュールを表す１つ以上の計算された活性化スコアにおける計算された活性化スコアを出力する。この点で、及び上述したように、細胞構成要素モジュールのセットにおけるそれぞれの細胞構成要素モジュールの各々は、複数の細胞構成要素の独立したサブセットを含む。更に、細胞構成要素モジュールのセットにおける少なくとも第１の細胞構成要素モジュールは、目的の生理学的状態と関連付けられる。この実施例では、モジュール７８は、目的の生理学的状態と関連付けられる。図１０Ｂに示されるように、モジュール７８を正しく活性化し、したがって、モジュール７８の目的の生理学的状態（細胞代謝に重要な脂肪酸関連細胞プロセス）と関連する化合物は、（例えば、第１の閾値基準を満たす第１の細胞構成要素モジュールについてのそれぞれの計算された活性化スコアによって）識別される。

【０５３８】
特許請求されるアプローチの第２の検証として、次に、モジュール７８及び９０についてのそれぞれの訓練されたモデルを、訓練中に図６のモデル６０１に導入されていない別の試験セットである前脂肪細胞に曝露された特定の小分子である「合成ヒット」のｓｃＲＮＡ－ｓｅｑ特性評価に適用した。図１０Ｄは、合成ヒットによるモジュール７８についての訓練されたモデル６０１によって示される活性化の高い相関及び忠実な予測を、合成ヒットによるモジュール９０についての訓練されたモデル６０１によって示されるほとんどない活性化、又は全くない活性化と比較して示す。
【０５３９】
第三に、モジュール７８のための訓練されたモデル６０１を使用して、公開データベース内の５００万個の化合物からサンプリングされた２００，０００個の化合物のランダムなサブセットについて、細胞構成要素モジュール７８（モジュール７８）についての細胞構成要素活性化スコアを予測した。このことから、細胞構成要素モジュール７８を高度に活性化すると予測される上位５０個の化合物を選択し、ＬＩＮＣＳ　Ｌ１０００データセットからの化合物及び本明細書において既知のピペリジン含有化合物（「ＫＰＣＣ」）と称される既知の化合物の化学構造に由来する合成ヒット類似体を含むデータベース内の化合物のセットと比較した。この比較の分布を図１０Ｅに示す。分布の末端で、細胞構成要素モジュール７８についての訓練されたモデル６０１について得られた予測は、ＬＩＮＣＳ及び合成ヒットにおける全ての化合物を有意に上回る化合物を識別した。このアプローチは、特定の所望の細胞プロセスに対して化学構造を最適化するための方法を強調する。
【０５４０】
第４に、上位５０の予測で識別された化学構造を視覚的に検査し、既知の脂肪組織標的化ファーマコフォア）を表す明白な化学構造を含み、したがって、モジュール７８と関連付けられた細胞構成要素モジュールを正当に活性化することが見出された。
【０５４１】
この第１の例はまた、当初出願された請求項５８を追跡する。請求項１と請求項５８との間の違いは、細胞構成要素モジュールに対する摂動シグネチャの１つである。摂動シグネチャは、摂動にさらされた細胞とそうではない細胞の発現を比較することによって得られる。したがって、前脂肪細胞において代謝活性プロセスを誘導することが知られている低分子ペルターバゲンを使用することができる。前脂肪細胞株をペルターバゲンに２４時間曝露し、ｓｃＲＮＡ－ｓｅｑ読み出しを摂動状態及び対照状態について得ることができる。このことから、摂動シグネチャを得ることができる。あるいは、別個の摂動シグネチャは、第２のデータセットに使用される化学共変量のうちのいずれか１つに曝露された細胞の細胞発現を比較することによって得ることができる。実際に、第２のデータセットに使用される化学共変量の各々について、このようにして別個の摂動シグネチャを得ることができる。そのような摂動シグネチャの各々は、そのような重みの各々が現在２値スケールではなく連続スケール上にあることを除いて、潜在表現４０４内の行の形態を有する。例えば、いくつかの実施形態において、各重みは、０から１（又はいくつかの他の範囲の「Ａ」から「Ｂ」であり、Ａ及びＢは、－１００及び１００などの２つの異なる数字である）の間の連続スケール上の値である。そこから、訓練のプロセスは、潜在表現４０４、カウントマトリックス５０２、活性化データ構造、及びコンポーネントモデル６０１の訓練の使用に関して上述したものと同一であり、そのようなモデルの各々は、現在、摂動シグネチャのセットにおける異なる摂動シグネチャを表す。
【０５４２】
実施例２．胎児赤血球生成プログラムを活性化し、Ｔ細胞枯渇をブロックするための化学構造の予測。
２つの追加の例では、胎児赤血球生成及びＴ細胞枯渇に関連する２つのｓｃＲＮＡ－ｓｅｑデータセット上の２つのモデルを訓練した。
【０５４３】
胎児赤血球生成のために、ＣＤ３４造血幹細胞をツール化合物ＣＬＴ－ＡＡＡ－１２で

処理し、これに関して、胎児赤血球生成のエンドポイントマーカー、特にフローサイトメトリーによる読み出しとしてのアッセイにおけるＦ細胞の数が誘導されることが以前に確立されている。

【０５４４】

Ｔ細胞枯渇のために、ナイーブＴ細胞を枯渇誘導培地で処置した。

【０５４５】

両方の細胞系は、ｓｃＲＮＡ－ｓｅｑで特徴付けられる。その後、薬物リフレクターモデル（参照により本明細書に組み込まれる、２０１９年７月１５日に出願された「Ｍｅｔｈｏｄｓ　ｏｆ　Ａｎａｌｙｚｉｎｇ　Ｃｅｌｌｓ」と題された米国特許出願第１６／５１１，６９１号を参照されたい）を、摂動細胞対対照細胞によって定義される細胞状態遷移をそれらのそれぞれの試料に入力することによって、ｓｃＲＮＡ－ｓｅｑデータセットに適用した。薬物リフレクターは、薬物リフレクター潜在表現における８０００個の化合物の各々について細胞状態活性化スコアを割り当てる。これにより、両方の遷移（胎児ヘモグロビン及びＴ細胞枯渇）について細胞状態活性化スコアを有する２つのベクトルが生じる。これらの２つのベクトルは、モデル６０１についての訓練データとして機能する。

【０５４６】

このモデルを使用して、造血幹細胞における胎児赤血球生成を活性化する化合物及びＴ細胞枯渇を予測した。造血幹細胞における胎児赤血球生成は、近年、鎌状赤血球疾患に対する画期的なＣＲＩＳＰＲ療法につながった細胞プロセスであり、一方、Ｔ細胞枯渇は、がんに対するチェックポイント阻害剤療法のより広範な成功を妨げる重要な機構である。

【０５４７】

予測は、公開データベース内の５００万個の化合物からサンプリングされた２，０００個の化合物のサブセットを使用して実施され、サブセットは、ランダムに又は骨格上で分割された。図１１の上部パネルは、ランダムに分割された、及び骨格上にある２，０００個の化合物の試験セット上でのこの実施例のモデルの性能を示し、造血幹細胞における胎児赤血球生成に関連するヒット化合物ＣＬＴ－ＡＡＡ－１２の摂動シグネチャとともに、サンプリングした化合物の有意なR^2及び相関係数p_sを示している。図１１の下部パネルは、ランダムに分割された、及び骨格上にある２，０００個の化合物の試験セットの性能を示し、Ｔ細胞枯渇に関連する細胞遷移シグネチャとともに、サンプリングした化合物の有意なR^2及び相関係数p_sを示している。したがって、図１１は、モデル６０１が、目的の摂動シグネチャ及び／又は細胞遷移シグネチャと同じ細胞挙動効果を誘導する新しい骨格を予測することができることを実証する。

【０５４８】

実施例３．疾患クリティカル細胞挙動に基づく特徴属性：新しい分子の設計のためのファーマコフォアの予測。

実施例１に記載されるように、本明細書に開示されるシステム及び方法に従って予測される化学構造を使用して、目的の生理学的状態（例えば、脂肪組織標的化）に潜在的に関連している、ファーマコフォアなどの分子的特徴を識別することができる。実施例１と同様に、これらのファーマコフォアは、既知の化学構造によって検証することができるか、又は更なる検証のために新規の構造を提示することができる。例えば、ファーマコフォアに基づくアルゴリズムの例示的な使用事例は、Ｂａｓｅ　ｏｆ　Ｂｉｏｉｓｏｓｔｅｒｉｃａｌｌｙ　Ｅｘｃｈａｎｇｅａｂｌｅ　Ｒｅｐｌａｃｅｍｅｎｔｓ（ＢｏＢＥＲ）データベースを含む、以前に文献に記載された機能的意味を有するファーマコフォアのデータベースを活用することを含む。使用事例の別の例は、摂動に対するシステムの複雑な応答における識別されたファーマコフォアの役割に関する直感を得るために、薬学者などによる専門知識を適用することを含む。

【０５４９】

新しい分子の設計のためのファーマコフォアを予測するためのモデルを実施し、モデルは、Ｔｅｖｅｒｓｋｙ類似性を使用してスコアに基づいて選択された介入ライブラリからの低分子の特徴化を含み、ファーマコフォアが化学構造に含まれているかどうかを示す表

現を達成した。この表現（化学フィンガープリント）を、実施例１のモジュール７８のモデル６０１に入力した。実施例１で識別された脂肪標的化ファーマコフォアを使用して、実施例１のモジュール７８についてのモデルを使用して、既知のピペリジン含有化合物（「ＫＰＣＣ」）の脂肪標的化ファーマコフォアの関連性を決定し、分離して、０．０４０６４～０．０４６３３の範囲の活性化スコアを有する脂肪酸モジュールの転写活性化を観察した。

【０５５０】

実施例４．潜在的な細胞挙動に基づく合成ヒット化合物の生成。

試験事例として、本明細書において「６つの合成ヒット」と称される新たに合成された低分子ヒットのうちの６つは、インビトロ及びインビボで検証された脂肪細胞ベージング化合物及びその潜在空間表現に基づいて設計された。６つの合成ヒットの各々は、ヒト前脂肪細胞上の所望の細胞挙動変化を誘発した。まず、ＫＰＣＣクラスタのファーマコフォアを識別した。次いで、分子を、新規の生物学的等価体の組み込みとともに、このクラスタにおけるファーマコフォアの濃縮によって設計し、これが６つの合成ヒットの最終的な設計につながった。これらの６つの構造的に多様な合成ヒットの目標は、ＫＰＣＣを含む既知の化学物質（ＫＣＥ）と同じ細胞挙動効果を誘導することであった。図１３の概略図に示されるように、細胞挙動効果は、ヒト前脂肪細胞を１μＭのＫＰＣＣ及び６つの合成ヒットで２４時間処理し、ｓｃＲＮＡ－ｓｅｑを使用して遺伝子発現を測定し、上記実施例１に記載される脂肪代謝遺伝子モジュールの変化によって発現される細胞応答を評価することによって決定した（モジュール７８）。例えば、脂肪代謝モジュールにおける遺伝子は、とりわけ、ＦＡＢＰ３、ＦＤＰＳ、及びＬＰＩＮ１を含む。

【０５５１】

前脂肪細胞に対するこれらの化合物の影響の評価は、各々の合成ヒットが、ＫＰＣＣと同じ脂肪代謝遺伝子モジュールを活性化したことを明らかにした（図１３；モジュール７８は、ボックス１３０２で強調されている）。すなわち、強調表示されたボックス１３０２は、図１３のグラフのＹ軸に列挙されたモデルに化合物のフィンガープリントを入力すると、実施例１のモジュール７８についてのモデルによって出力される活性化スコアを示す。これらの結果は、所望の細胞挙動を予測可能に標的とするモデルプラットフォームに基づいて合成ヒットを生成する能力に高い信頼性を提供する。特に、本開示のモデル６０１（例えば、実施例１のモジュール７８についてのモデル６０１）を使用して、ハイスループットスクリーニング、分子標的に基づく識別若しくは最適化、又は検証のための数百又は数千の新規化合物の合成を必要とせずに、生理学的状態と関連する遺伝子モジュールを標的とする合成ヒットを予測することができる。

【０５５２】

引用文献及び代替の実施形態

本明細書で引用される全ての参考文献は、個々の刊行物又は特許又は特許出願の各々が、全ての目的において参照によりその全体が組み込まれることが具体的かつ個々に示されたのと同等に、全ての目的において参照によりその全体が本明細書に組み込まれる。

【０５５３】

本発明は、非一時的なコンピュータ可読記憶媒体に埋め込まれたコンピュータプログラム機構を含むコンピュータプログラム製品として実装され得る。例えば、コンピュータプログラム製品は、図１～図３及び図７～図９の任意の組み合わせで示されるプログラムモジュールを含み得る。これらのプログラムモジュールは、ＣＤ－ＲＯＭ、ＤＶＤ、磁気ディスクストレージ製品、又は任意の他の非一時的コンピュータ可読データ若しくはプログラムストレージ製品に格納され得る。

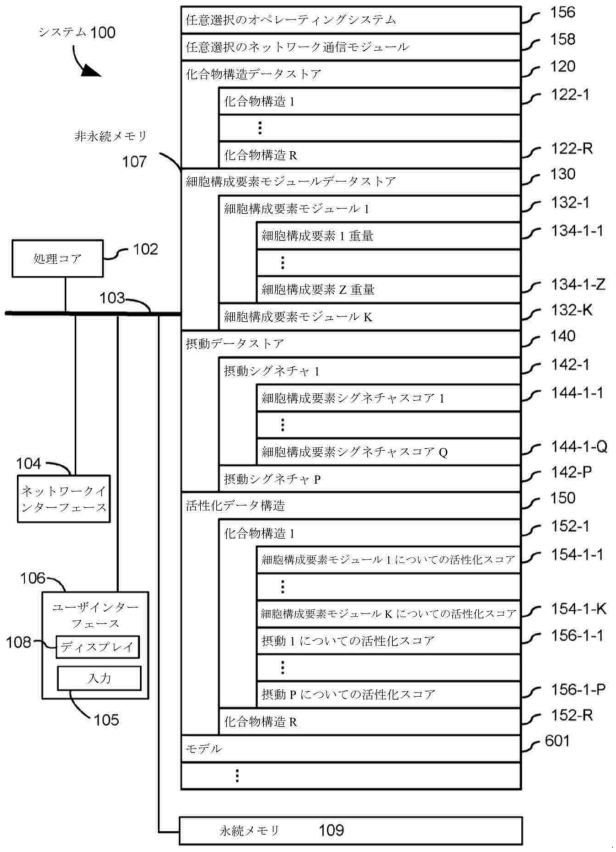
【０５５４】

当業者には明らかなように、本発明の多くの修正及び変形を、その趣旨及び範囲から逸脱することなく行うことができる。本明細書に記載される特定の実施形態は、例としてのみ提供される。実施形態は、本発明の原理及びその実際の用途を最もよく説明するために選択及び説明され、それによって、当業者が、本発明及び企図される特定の使用に適した

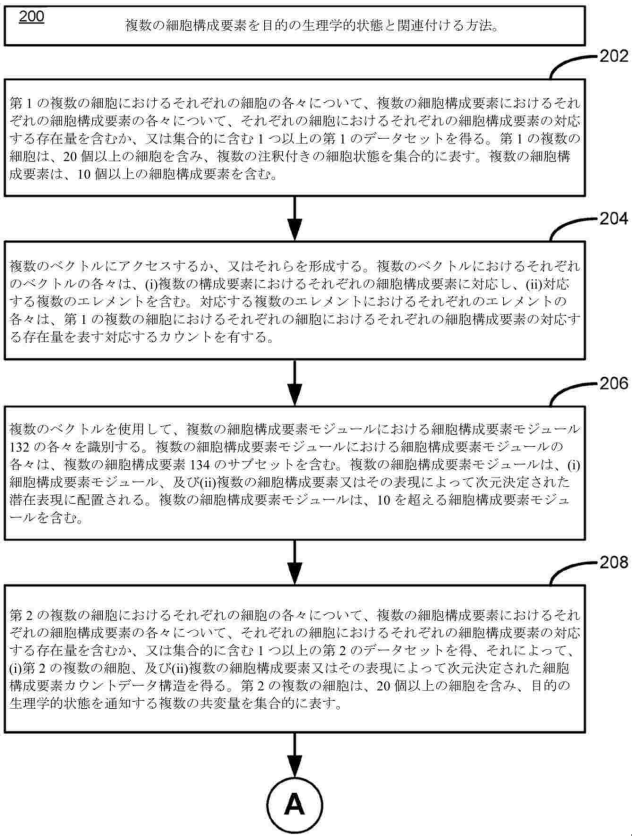
様々な修正を伴う様々な実施形態を最もよく利用できるようにする。本発明は、添付の特許請求の範囲の用語、及びそのような特許請求の範囲が権利を有する等価物の全範囲によってのみ限定される。

【図面】

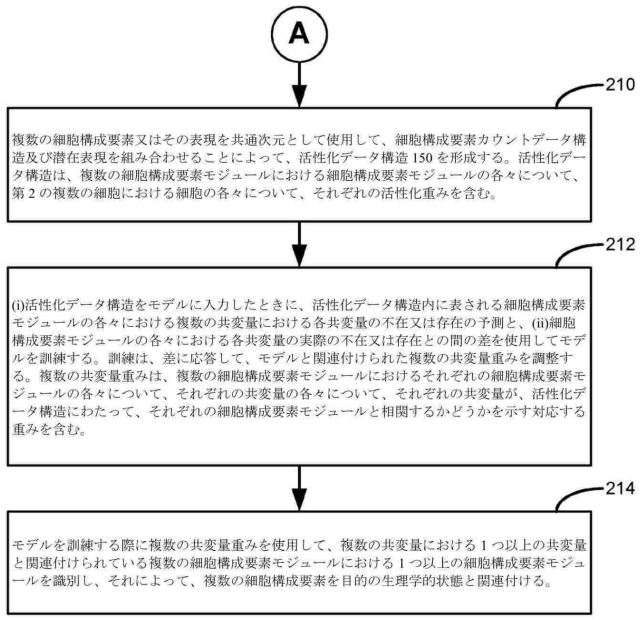
【図１】



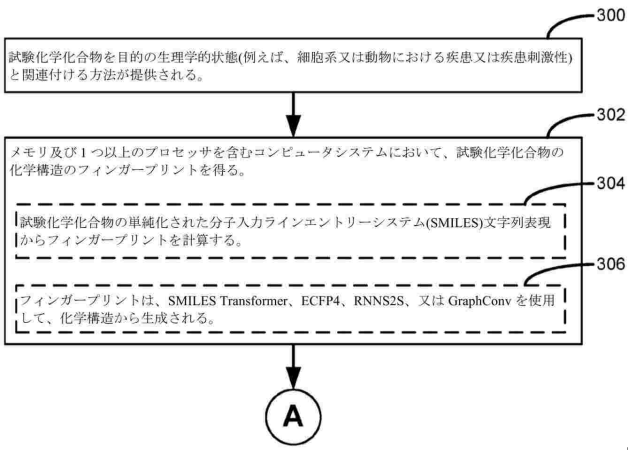
【図２Ａ】



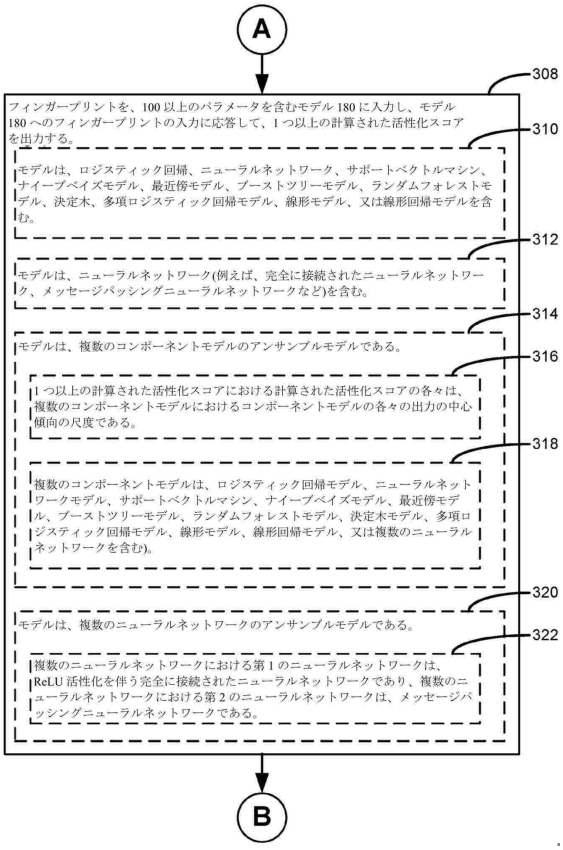
【図２Ｂ】



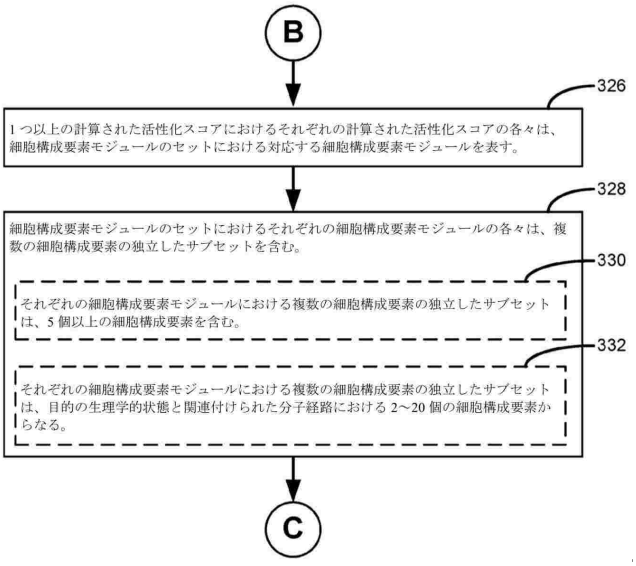
【図３Ａ】



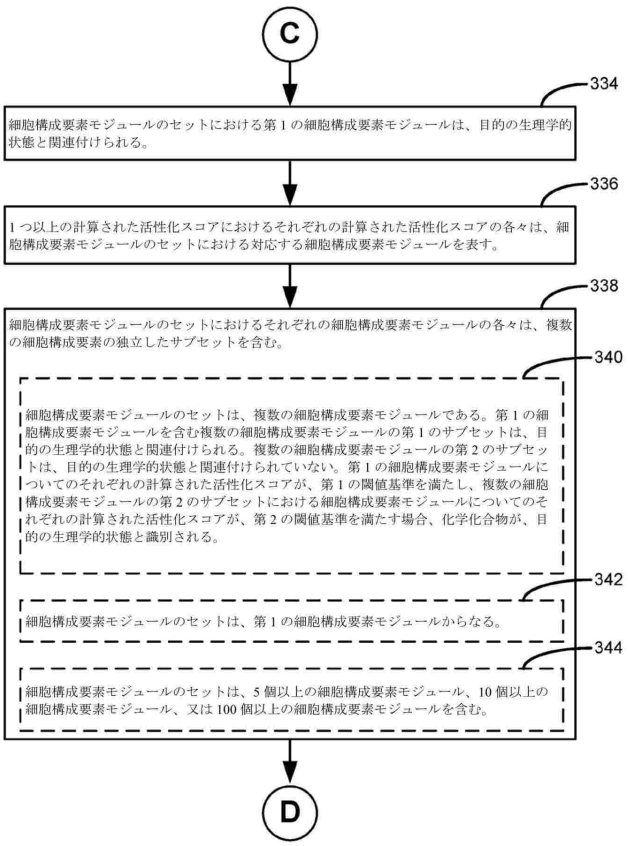
【図３Ｂ】



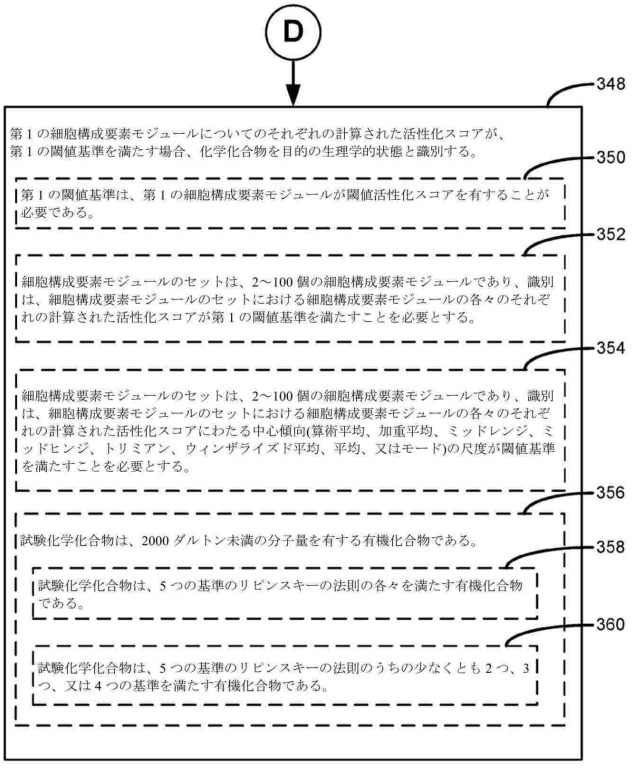
【図３Ｃ】



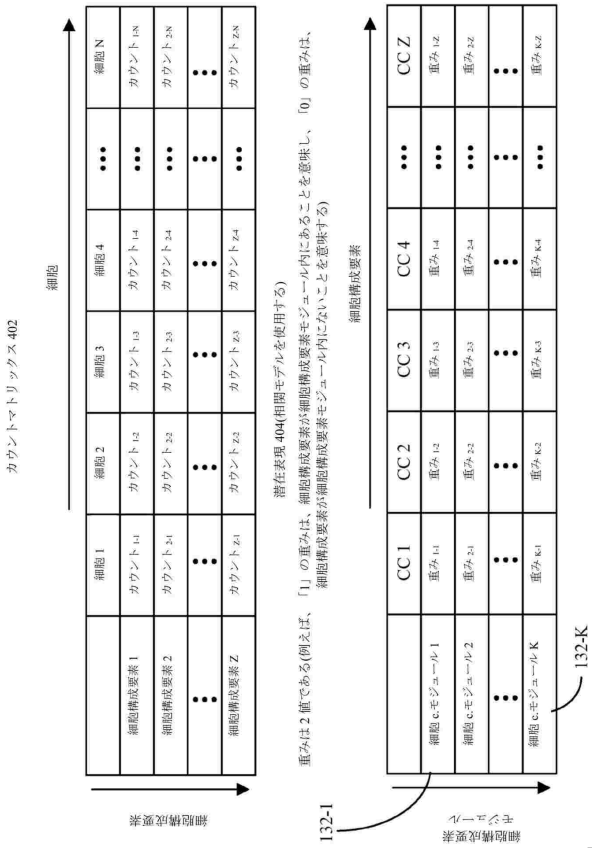
【図３Ｄ】



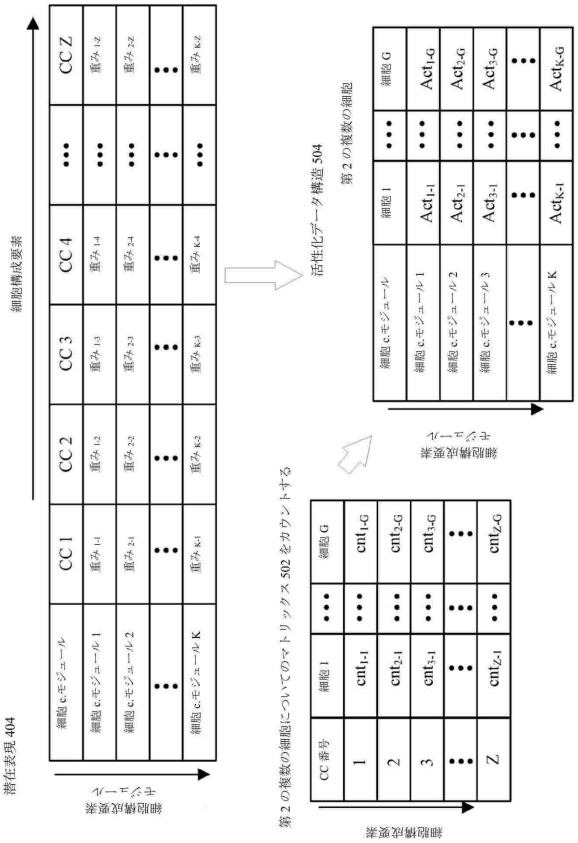
【図３Ｅ】



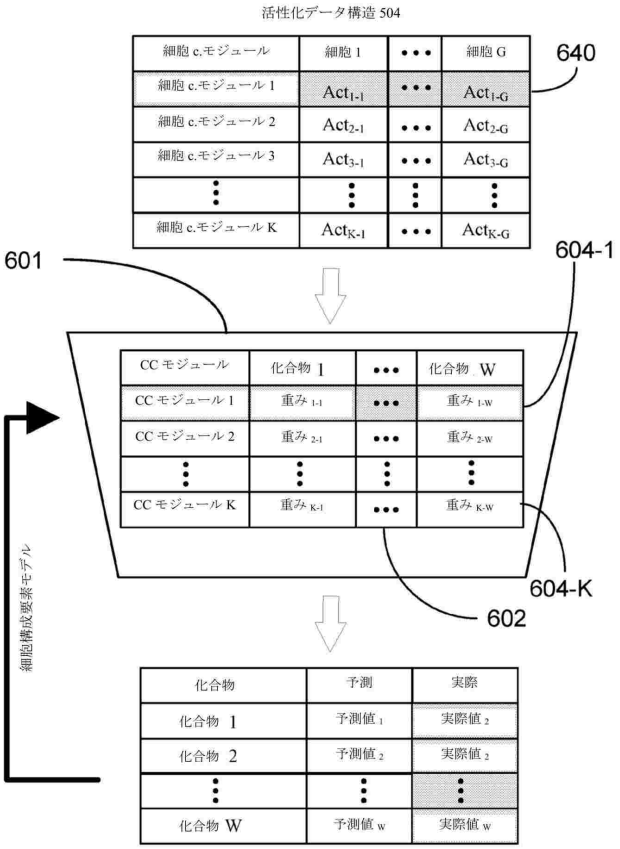
【図４】


カウントマトリックス 402
細胞
細胞構成要素
細胞構成要素 1
細胞構成要素 2
細胞構成要素 Z
細胞 1
細胞 2
細胞 3
細胞 4
細胞 N
カウント 1-1
カウント 1-2
カウント 1-3
カウント 1-4
カウント 1-N
カウント 2-1
カウント 2-2
カウント 2-3
カウント 2-4
カウント 2-N
カウント Z-1
カウント Z-2
カウント Z-3
カウント Z-4
カウント Z-N
潜在表現 404(相関モデルを使用する)
重みは 2 値である(例えば、「1」の重みは、細胞構成要素が細胞構成要素モジュール内にあることを意味し、「0」の重みは、細胞構成要素が細胞構成要素モジュール内にないことを意味する)
細胞構成要素
CC 1
CC 2
CC 3
CC 4
CC Z
細胞 c.モジュール 1
細胞 c.モジュール 2
細胞 c.モジュール K
重み 1-1
重み 1-2
重み 1-3
重み 1-4
重み 1-Z
重み 2-1
重み 2-2
重み 2-3
重み 2-4
重み 2-Z
重み K-1
重み K-2
重み K-3
重み K-4
重み K-Z
132-K
132-1
モジュール
細胞構成要素

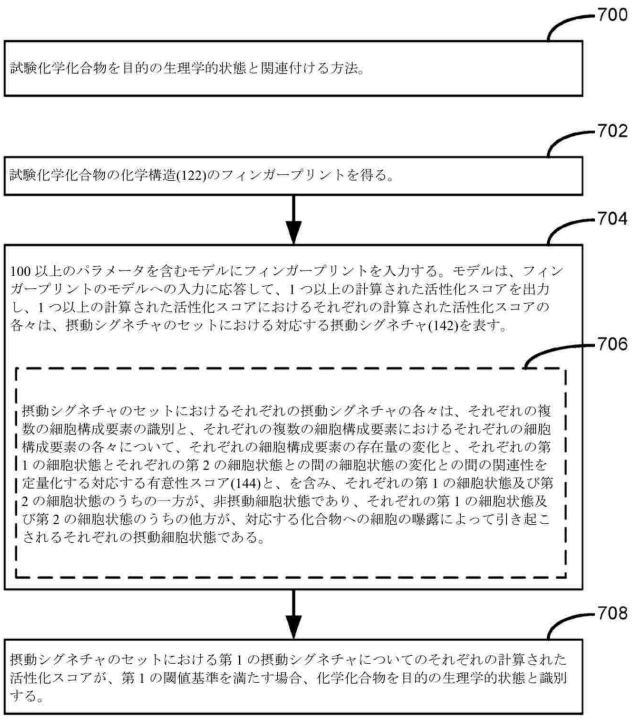
【図５】


細胞構成要素
CC 1
CC 2
CC 3
CC 4
CC Z
細胞 c.モジュール
細胞 c.モジュール 1
細胞 c.モジュール 2
細胞 c.モジュール K
重み 1-1
重み 1-2
重み 1-3
重み 1-4
重み 1-Z
重み 2-1
重み 2-2
重み 2-3
重み 2-4
重み 2-Z
重み K-1
重み K-2
重み K-3
重み K-4
重み K-Z
潜在表現 404
モジュール
細胞構成要素
活性化データ構造 504
第 2 の複数の細胞
細胞
細胞 1
細胞 G
Act1-1
Act1-G
Act2-1
Act2-G
Act3-1
Act3-G
ActK-1
ActK-G
第 2 の複数の細胞についてのマトリックス 502 をカウントする
CC 番号
cnt1-1
cnt1-G
cnt2-1
cnt2-G
cnt3-1
cnt3-G
cntZ-1
cntZ-G

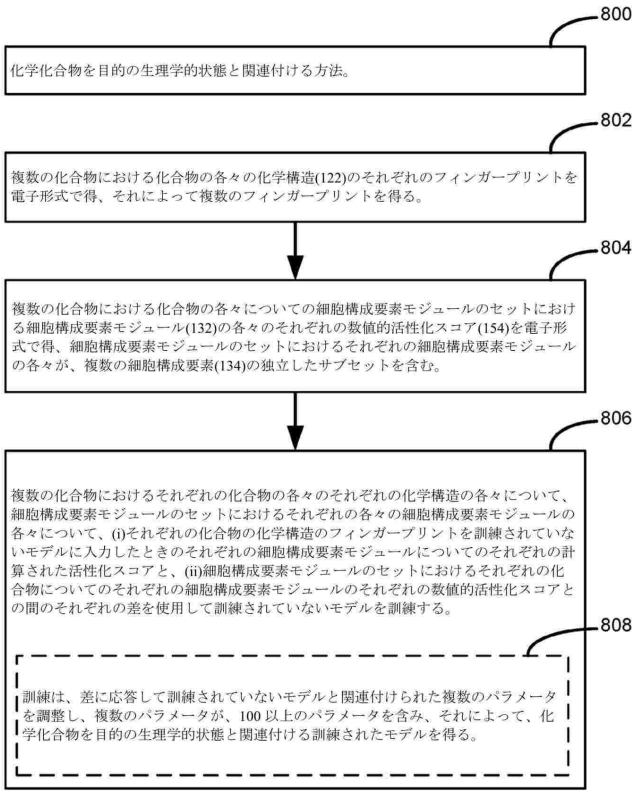
【図６】


活性化データ構造 504
細胞 c.モジュール
細胞 1
細胞 G
640
Act1-1
Act1-G
Act2-1
Act2-G
Act3-1
Act3-G
ActK-1
ActK-G
601
604-1
CC モジュール
化合物 1
化合物 W
重み 1-1
重み 1-W
重み 2-1
重み 2-W
重み K-1
重み K-W
604-K
602
細胞構成要素モデル
化合物
予測
実際
予測値 1
実際値 1
予測値 2
実際値 2
予測値 W
実際値 W

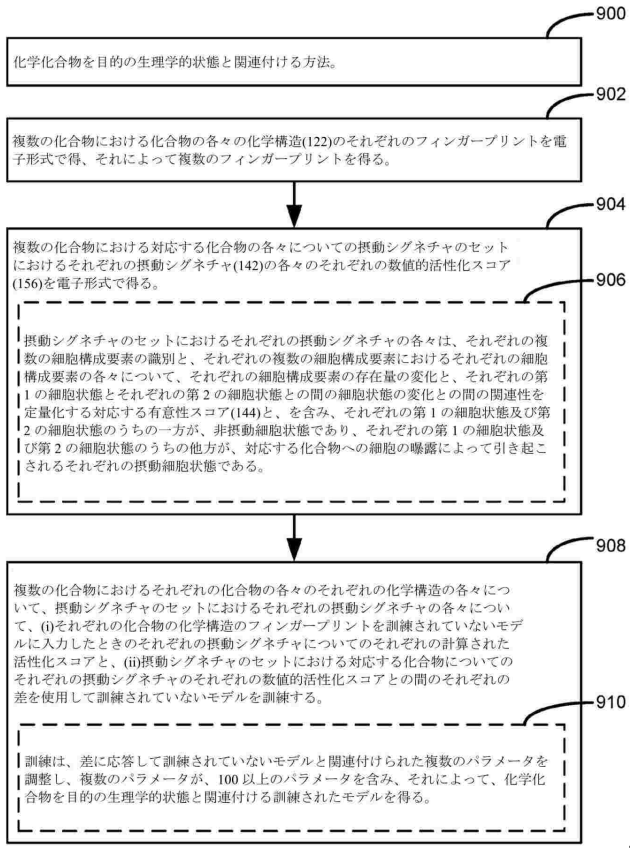
【図７】


700
試験化学化合物を目的の生理学的状態と関連付ける方法。
702
試験化学化合物の化学構造(122)のフィンガープリントを得る。
704
100 以上のパラメータを含むモデルにフィンガープリントを入力する。モデルは、フィンガープリントのモデルへの入力に応答して、1 つ以上の計算された活性化スコアを出力し、1 つ以上の計算された活性化スコアにおけるそれぞれの計算された活性化スコアの各々は、摂動シグネチャのセットにおける対応する摂動シグネチャ(142)を表す。
706
摂動シグネチャのセットにおけるそれぞれの摂動シグネチャの各々は、それぞれの複数の細胞構成要素の識別と、それぞれの複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞構成要素の存在量の変化と、それぞれの第 1 の細胞状態とそれぞれの第 2 の細胞状態との間の細胞状態の変化との間の関連性を定量化する対応する有意性スコア(144)と、を含み、それぞれの第 1 の細胞状態及び第 2 の細胞状態のうちの一方が、非摂動細胞状態であり、それぞれの第 1 の細胞状態及び第 2 の細胞状態のうちの他方が、対応する化合物への細胞の曝露によって引き起こされるそれぞれの摂動細胞状態である。
708
摂動シグネチャのセットにおける第 1 の摂動シグネチャについてのそれぞれの計算された活性化スコアが、第 1 の閾値基準を満たす場合、化学化合物を目的の生理学的状態と識別する。

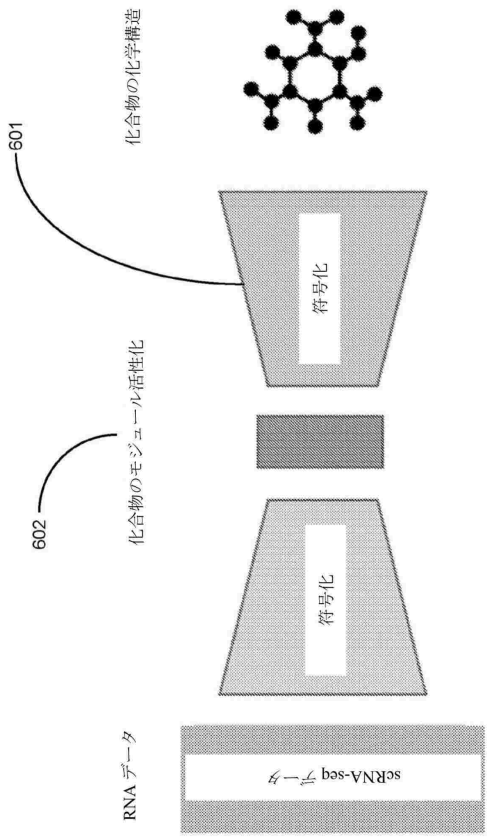
【図８】



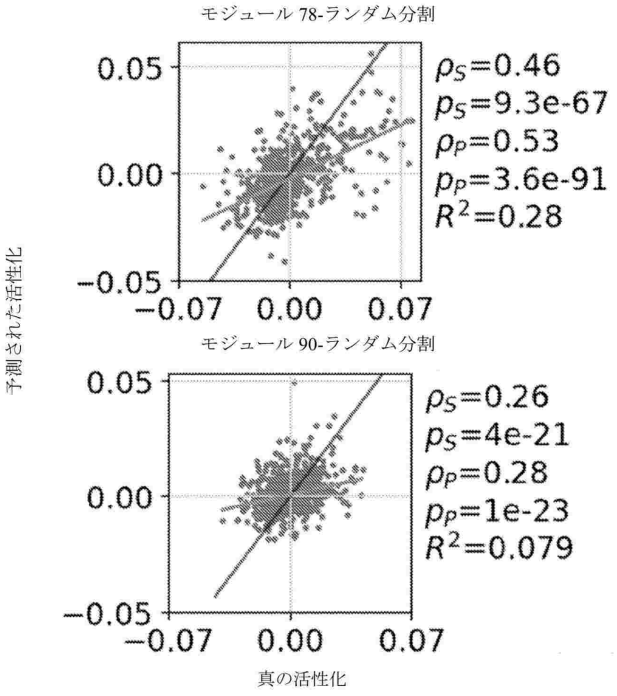
【図９】



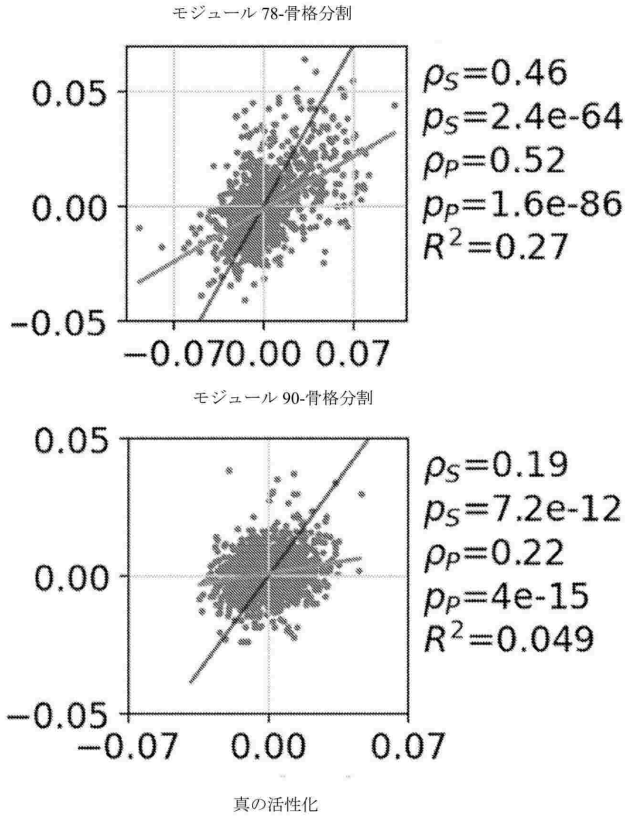
【図１０Ａ】



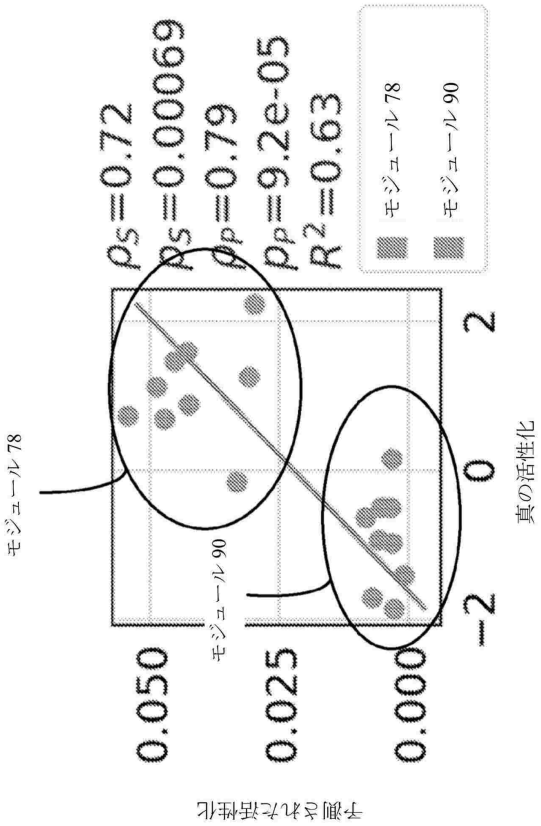
【図１０Ｂ】



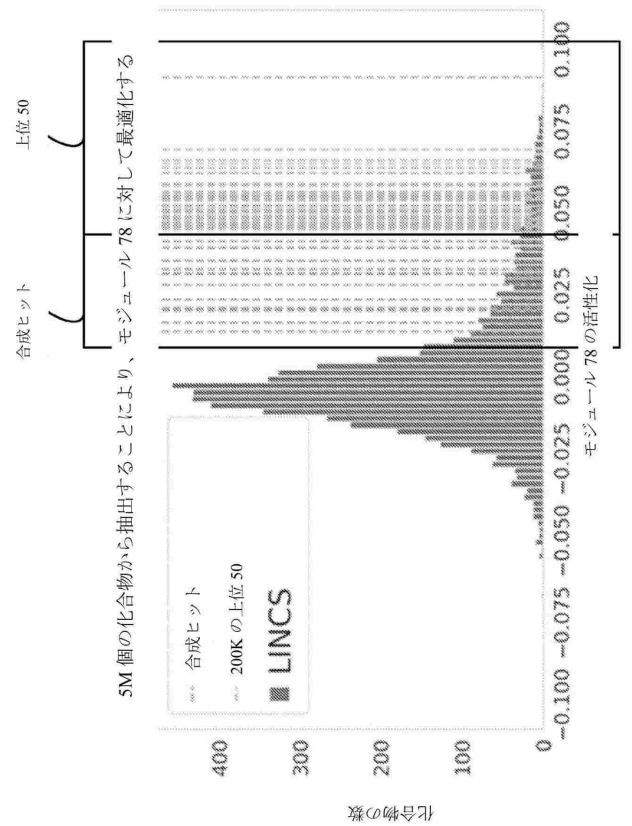
【図１０Ｃ】



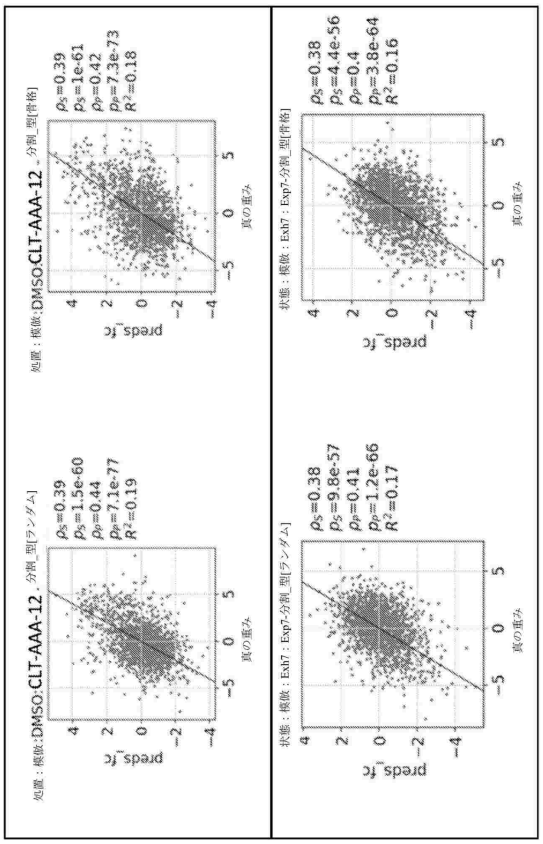
【図１０Ｄ】



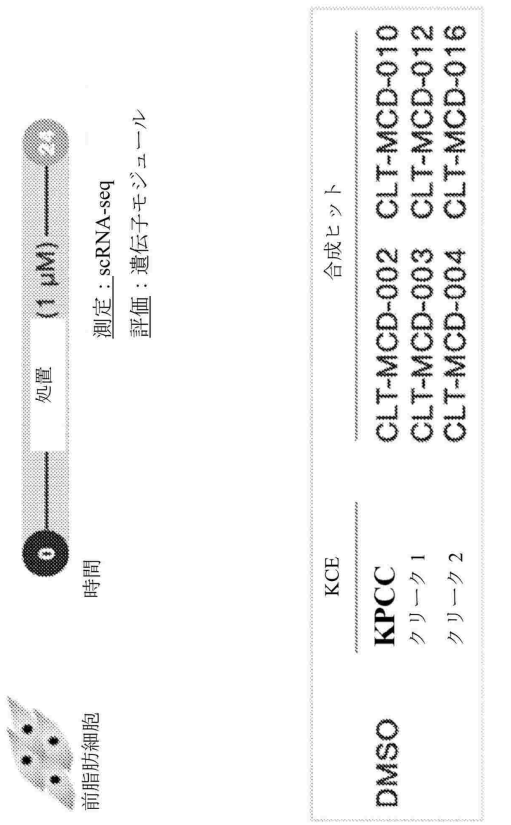
【図１０Ｅ】



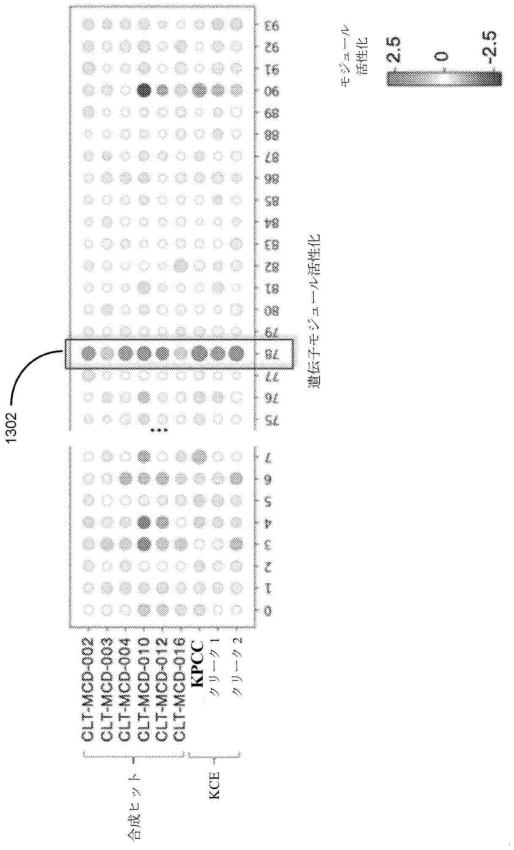
【図１１】



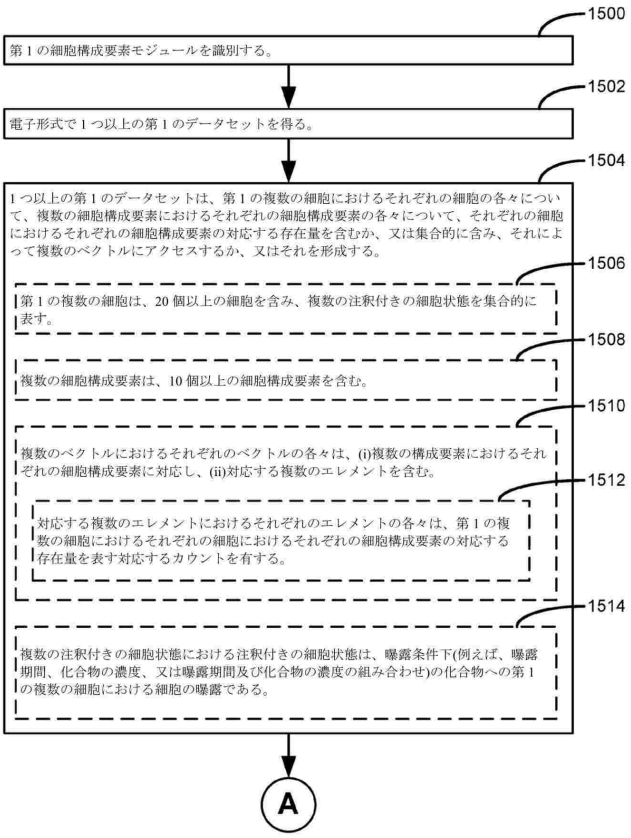
【図１２】



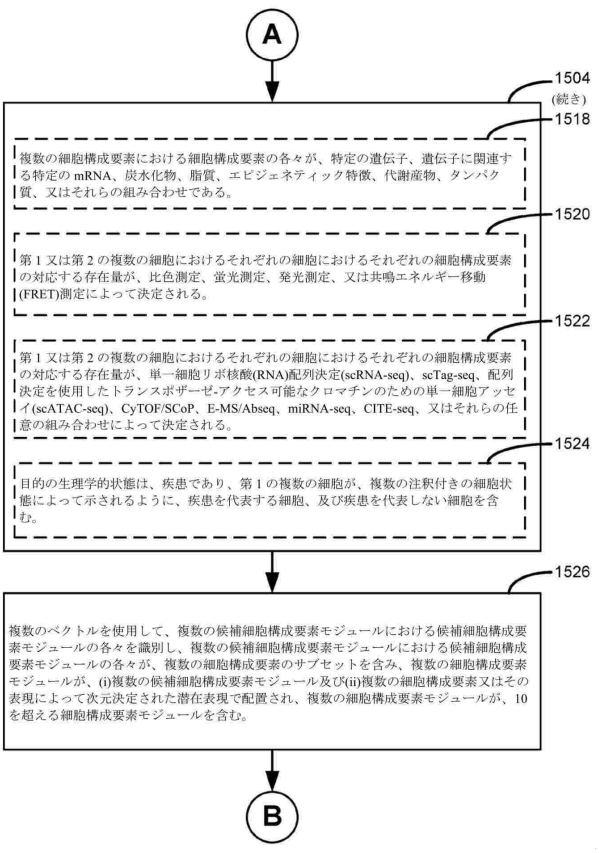
【図１３】



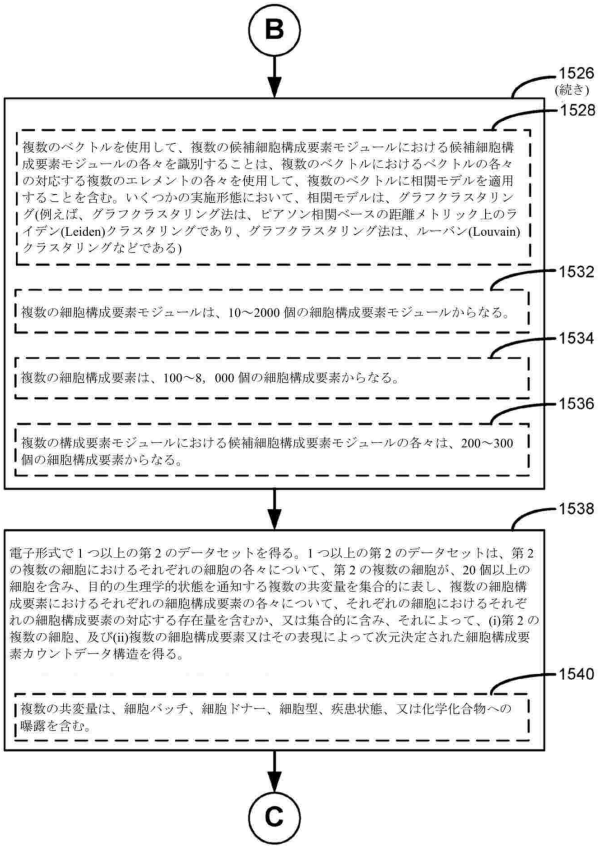
【図１４Ａ】



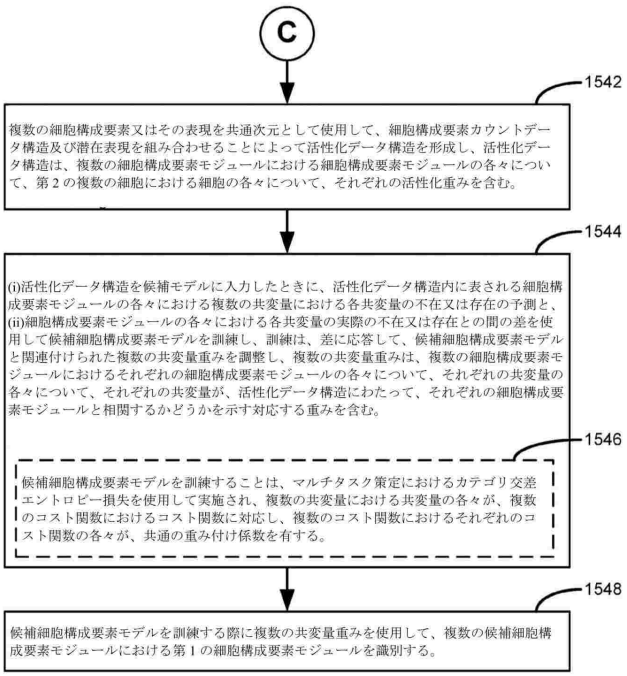
【図１４Ｂ】



【図１４Ｃ】


B
1526
(続き)
1528
複数のベクトルを使用して、複数の候補細胞構成要素モジュールにおける候補細胞構成要素モジュールの各々を識別することは、複数のベクトルにおけるベクトルの各々の対応する複数のエレメントの各々を使用して、複数のベクトルに相関モデルを適用することを含む。いくつかの実施形態において、相関モデルは、グラフクラスタリング(例えば、グラフクラスタリング法は、ピアソン相関ベースの距離メトリック上のライデン(Leiden)クラスタリングであり、グラフクラスタリング法は、ルーバン(Louvain)クラスタリングなどである)
1532
複数の細胞構成要素モジュールは、10～2000 個の細胞構成要素モジュールからなる。
1534
複数の細胞構成要素は、100～8，000 個の細胞構成要素からなる。
1536
複数の構成要素モジュールにおける候補細胞構成要素モジュールの各々は、200～300 個の細胞構成要素からなる。
1538
電子形式で 1 つ以上の第 2 のデータセットを得る。1 つ以上の第 2 のデータセットは、第 2 の複数の細胞におけるそれぞれの細胞の各々について、第 2 の複数の細胞が、20 個以上の細胞を含み、目的の生理学的状態を通知する複数の共変量を集合的に表し、複数の細胞構成要素におけるそれぞれの細胞構成要素の各々について、それぞれの細胞におけるそれぞれの細胞構成要素の対応する存在量を含むか、又は集合的に含み、それによって、(i)第 2 の複数の細胞、及び(ii)複数の細胞構成要素又はその表現によって次元決定された細胞構成要素カウントデータ構造を得る。
1540
複数の共変量は、細胞バッチ、細胞ドナー、細胞型、疾患状態、又は化学化合物への曝露を含む。
C

【図１４Ｄ】


C
1542
複数の細胞構成要素又はその表現を共通次元として使用して、細胞構成要素カウントデータ構造及び潜在表現を組み合わせることによって活性化データ構造を形成し、活性化データ構造は、複数の細胞構成要素モジュールにおける細胞構成要素モジュールの各々について、第 2 の複数の細胞における細胞の各々について、それぞれの活性化重みを含む。
1544
(i)活性化データ構造を候補モデルに入力したときに、活性化データ構造内に表される細胞構成要素モジュールの各々における複数の共変量における各共変量の不在又は存在の予測と、(ii)細胞構成要素モジュールの各々における各共変量の実際の不在又は存在との間の差を使用して候補細胞構成要素モデルを訓練し、訓練は、差に応答して、候補細胞構成要素モデルと関連付けられた複数の共変量重みを調整し、複数の共変量重みは、複数の細胞構成要素モジュールにおけるそれぞれの細胞構成要素モジュールの各々について、それぞれの共変量の各々について、それぞれの共変量が、活性化データ構造にわたって、それぞれの細胞構成要素モジュールと相関するかどうかを示す対応する重みを含む。
1546
候補細胞構成要素モデルを訓練することは、マルチタスク策定におけるカテゴリ交差エントロピー損失を使用して実施され、複数の共変量における共変量の各々が、複数のコスト関数におけるコスト関数に対応し、複数のコスト関数におけるそれぞれのコスト関数の各々が、共通の重み付け係数を有する。
1548
候補細胞構成要素モデルを訓練する際に複数の共変量重みを使用して、複数の候補細胞構成要素モジュールにおける第 1 の細胞構成要素モジュールを識別する。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/033685

A. CLASSIFICATION OF SUBJECT MATTER
INV. G16B15/30 G16B5/00 G16B25/10 G16B40/20
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G16B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Lotfollahi Mohammad ET AL: "Learning interpretable cellular responses to complex perturbations in high-throughput screens", bioRxiv, 15 April 2021 (2021-04-15), XP055962485, DOI: 10.1101/2021.04.14.439903 Retrieved from the Internet: URL:https://www.biorxiv.org/content/10.1101/2021.04.14.439903v1.full.pdf [retrieved on 2022-09-19] the whole document in particular "results" and "methods" sections; figures 1-3 ----- -/--	1-159

[x] Further documents are listed in the continuation of Box C. [x] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 September 2022	28/09/2022

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Rákossy, Z

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/033685

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	DURAN-FRIGOLA MIQUEL ET AL: "Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker", NATURE BIOTECHNOLOGY, NATURE PUBLISHING GROUP US, NEW YORK, vol. 38, no. 9, 18 May 2020 (2020-05-18), pages 1087-1096, XP037524011, ISSN: 1087-0156, DOI: 10.1038/S41587-020-0502-7 [retrieved on 2020-05-18] the whole document in particular "methods" section; page 1093 - page 1095; figures 1, 2, 5; table 1	1-159
X	JO JEONGHEE ET AL: "The message passing neural networks for chemical property prediction on SMILES", METHODS, ACADEMIC PRESS, NL, vol. 179, 21 May 2020 (2020-05-21), pages 65-72, XP086236801, ISSN: 1046-2023, DOI: 10.1016/J.YMETH.2020.05.009 [retrieved on 2020-05-21] the whole document in particular sections 2 and 3; figure 1	1-159
X	US 2020/020419 A1 (KAHVEJIAN AVAK [US] ET AL) 16 January 2020 (2020-01-16) the whole document in particular paragraph [0115] - paragraph [0226]; figures 2, 3	1-159

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2022/033685

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2020020419 A1	16-01-2020	CA 3103677 A1	23-01-2020
		CN 112424866 A	26-02-2021
		EP 3824080 A1	26-05-2021
		JP 2022501011 A	06-01-2022
		KR 20210031708 A	22-03-2021
		US 2020020419 A1	16-01-2020
		WO 2020018519 A1	23-01-2020

Form PCT/ISA/210 (patent family annex) (April 2005)

フロントページの続き

MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IQ,IR,IS,IT,JM,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

（特許庁注：以下のものは登録商標）

１．ＢＬＵＥＴＯＯＴＨ
２．ＡＮＤＲＯＩＤ
３．Ｌｉｎｕｘ
４．ＵＮＩＸ
５．ＯＳ Ｘ
６．ＷＩＮＤＯＷＳ
７．ＶＸＷＯＲＫＳ
８．ＴＥＮＳＯＲＦＬＯＷ
９．ＷＣＤＭＡ

(72)発明者 アメリカ合衆国，マサチューセッツ州 ０２１４２，ケンブリッジ，ケンブリッジ パークウェイ ５５ エイス フロア フラッグシップ パイオニアリング イノベーションズ シックス，エルエルシー
ハダド，レイジー
(72)発明者 アメリカ合衆国，マサチューセッツ州 ０２１４２，ケンブリッジ，ケンブリッジ パークウェイ ５５ エイス フロア フラッグシップ パイオニアリング イノベーションズ シックス，エルエルシー
プラギス，ニコラス マッカートニー
アメリカ合衆国，マサチューセッツ州 ０２１４２，ケンブリッジ，ケンブリッジ パークウェイ ５５ エイス フロア フラッグシップ パイオニアリング イノベーションズ シックス，エルエルシー

Ｆターム（参考） 4B063 QA01 QA13 QQ08 QQ53 QR08 QR42 QR62 QS34 QX02
4B065 AA93X CA46