(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2023-547571
(P2023-547571A)
(43)公表日　令和5年11月10日(2023.11.10)

(51)国際特許分類　FI
G16C 20/50 (2019.01)　G16C 20/50

審査請求　未請求　予備審査請求　未請求　（全53頁）

(21)出願番号　特願2023-549140(P2023-549140)
(86)(22)出願日　令和3年10月22日(2021.10.22)
(85)翻訳文提出日　令和5年6月21日(2023.6.21)
(86)国際出願番号　PCT/GB2021/052753
(87)国際公開番号　WO2022/084696
(87)国際公開日　令和4年4月28日(2022.4.28)
(31)優先権主張番号　2016884.5
(32)優先日　令和2年10月23日(2020.10.23)
(33)優先権主張国・地域又は機関
英国(GB)
(31)優先権主張番号　2109633.4
(32)優先日　令和3年7月2日(2021.7.2)
(33)優先権主張国・地域又は機関
英国(GB)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA
最終頁に続く

(71)出願人　523152031
エクセンシア・エイアイ・リミテッド
ＥＸＳＣＩＥＮＴＩＡ ＡＩ ＬＩＭＩＴＥＤ
イギリス ディディ１ ３ジェイティ ダンディー ウェスト・ヴィクトリア・ドック・ロード ５ ダンディー・ワン・リヴァー・コート レベル ３
ＬＥＶＥＬ ３，ＤＵＮＤＥＥ ＯＮＥ ＲＩＶＥＲ ＣＯＵＲＴ，５ ＷＥＳＴ ＶＩＣＴＯＲＩＡ ＤＯＣＫ ＲＯＡＤ，ＤＵＮＤＥＥ ＤＤ１ ３ＪＴ，ＵＮＩＴＥＤ ＫＩＮＧＤＯＭ
(74)代理人　110001818
弁理士法人Ｒ＆Ｃ
最終頁に続く

(54)【発明の名称】　アクティブラーニングによる薬剤の最適化

(57)【要約】
本発明は、アクティブラーニングによる薬剤最適化の方法を提供する。この方法は、各化合物が一つ以上の分子特徴を有する複数の化合物の集団を定義することと、一つ以上の生物学的特性が既知である集団から化合物のトレーニングセットを定義することとを含む。この方法は、トレーニングセットに含まれない集団からの化合物のサブセットを選択すること、選択されたサブセットの化合物に存在する分子特徴に基づいて選択されたサブセットのスコアを決定すること、および決定されたサブセットスコアに基づいて選択されたサブセットを評価することを含む。サブセットスコアは、集団内における、およびトレーニングセットおよび選択されたサブセットを含むサンプリングされたセット内における分子特徴の頻度に基づいて決定される。

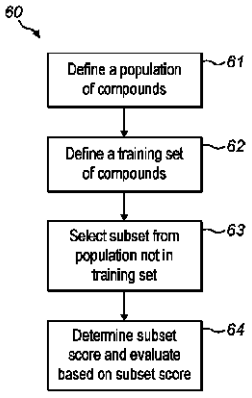


FIG. 6

【特許請求の範囲】
【請求項１】
各化合物が一つ以上の分子特性を有する複数の化合物の集団を定義する工程、
一つ以上の生物学的特性が既知である前記集団からの化合物のトレーニングセットを定義する工程、
前記トレーニングセットに含まれない前記集団からの一つ以上の化合物のサブセットを選択する工程、および
前記選択されたサブセット内の前記一つ以上の化合物の分子特性に応じて前記選択されたサブセットのサブセットスコアを決定し、当該決定されたサブセットスコアに基づいて前記選択されたサブセットを評価する工程、を含み、
前記サブセットスコアが、前記集団における前記分子特性の頻度と、前記トレーニングセットおよび前記選択されたサブセットを含むサンプリングされたセットにおける前記分子特性の頻度とに応じて決定されるものである、コンピュータによる薬剤設計のための方法。
【請求項２】
決定工程が、前記化合物の１つ以上の分子特性に応じて前記選択されたサブセットの前記１つ以上の化合物のそれぞれについて化合物スコアを決定することを含み、前記サブセットスコアが、前記選択されたサブセット内の各化合物の前記決定された化合物スコアに基づいて決定される、請求項１に記載の方法。
【請求項３】
前記サブセットスコアが、前記選択されたサブセット内の前記化合物の前記それぞれの化合物スコアの合計として決定される、請求項２に記載の方法。
【請求項４】
前記選択されたサブセット内の前記化合物の一つの前記化合物スコアを決定することが、前記集団における前記それぞれの分子特性の頻度、および前記サンプリングされたセットにおける前記それぞれの分子特性の頻度に応じて前記化合物の前記一つ以上の分子特性のそれぞれの分子特性スコアを決定することを含み、前記化合物の前記化合物スコアが、前記化合物の前記一つ以上の分子特性の前記決定されたスコアに基づくものである、請求項２または３に記載の方法。
【請求項５】
前記化合物の前記化合物スコアが、前記化合物の前記一つ以上の分子特性の前記決定された分子特性スコアの合計として決定される、請求項４に記載の方法。
【請求項６】
前記一つ以上の分子特性のそれぞれの前記分子特性スコアが、前記サンプリングされたセット内にある前記分子特性の正規化された確率に応じて決定され、前記正規化された確率が、前記集団および前記サンプリングされたセットにおける前記分子特性の頻度に応じて決定される、請求項４または５に記載の方法。
【請求項７】
前記正規化された確率が、前記集団内の化合物の数に対する、前記サンプリングされたセット内の化合物の数に応じて決定される、請求項６に記載の方法。
【請求項８】
前記正規化された確率が、ラプラシアン補正された正規化確率である、請求項７に記載の方法。
【請求項９】
前記ラプラシアン補正された正規化確率Ｐｃｏｒｒが、次式で与えられる、請求項８に記載の方法：
【数１】

$$P_{\mathrm{corr}} = \frac{F_{\mathrm{sampled}} + 1}{F_{\mathrm{set}} + 1/P_{\mathrm{base}}}$$

式中、$F_{sampled}$は、前記サンプリングされたセットにおける前記分子特性の頻度であり、F_{set}は、前記集団における前記分子特性の頻度であり、P_{base}は、前記サンプリングされたセットにおける化合物の数を前記集団における化合物の数で割ったものである。

【請求項１０】

前記一つ以上の分子特性のそれぞれの前記分子特性スコアが、前記サンプリングされたセット内の化合物の数に対する、前記分子特性が存在する前記サンプリングされたセット内の化合物の数に応じて決定される、請求項４から９のいずれかに記載の方法。

【請求項１１】

前記分子特性スコアが、前記サンプリングされたセットにおける前記分子特性の正規化されたシャノンエントロピー値に応じて決定される、請求項１０に記載の方法。

【請求項１２】

前記正規化されたシャノンエントロピー値が次の式で与えられる、請求項１１に記載の方法：

【数２】

$$SC = \frac{-f\ln(f) - (1-f)\ln(1-f)}{\ln(2)}$$

式中、ｆは、前記分子特性が存在する前記サンプリングされたセット内の化合物の数を前記サンプリングされたセット内の化合物の数で割ったものである。

【請求項１３】

前記分子特性スコアCov_{final}が次式で与えられる、請求項１２に記載の方法：

【数３】

$$Cov_{\text{final}} = \begin{cases} Cov * SC, & Cov \geq 0 \\ Cov * (2 - SC), & Cov < 0 \text{ and } f > 0.5 \end{cases}$$

式中、Ｃｏｖは以下の通りである。

【数４】

$$Cov = -\ln(P_{\text{corr}}/P_{\text{base}})$$

【請求項１４】

前記サブセットが所定数の化合物を含む、請求項１～１３のいずれか一項に記載の方法。

【請求項１５】

前記サブセット内で選択される化合物の数を定義することを含む、請求項１４に記載の方法。

【請求項１６】

前記評価する工程は、前記サブセットスコアが所定の条件を満たすかどうかを決定することを含む、請求項１～１５のいずれか一項に記載の方法。

【請求項１７】

前記所定の条件は、前記サブセットスコアが所定の最小閾値スコアより大きいことである、請求項１６に記載の方法。

【請求項１８】

前記所定の条件が満たされる場合、前記選択されたサブセット内の前記化合物の少なくとも一部を合成して、前記化合物の一つ以上の生物学的特性を決定することを含む、請求項１６または１７に記載の方法。

【請求項１９】

前記合成された化合物を前記トレーニングセットに加えることを含む、請求項１８に記載の方法。
【請求項２０】
前記選択されたサブセットが初期の選択されたサブセットであり；
前記トレーニングセットに含まれない前記集団からの一つ以上の化合物を含む、前記初期の選択されたサブセットとは異なる、第２サブセットを選択する工程、および
前記選択された第２サブセットの前記サブセットスコアを決定し、当該決定されたスコアに基づいて前記選択された第２サブセットを評価する工程、を含む、請求項１～１９のいずれか一項に記載の方法。
【請求項２１】
前記第２サブセットを選択しそのスコアを決定する工程が、前記所定の条件が満たされない場合に実行される、請求項１６に従属する場合の請求項２０に記載の方法。
【請求項２２】
前記第２サブセットを選択する工程が、前記初期の選択されたサブセット内の一つ以上の化合物を、前記トレーニングセットに含まれない前記集団からの一つ以上の新しい化合物で置換することを含む、請求項２０または２１に記載の方法。
【請求項２３】
置換される前記初期の選択されたサブセットから前記一つ以上の化合物を、前記初期の選択されたサブセット内の前記一つ以上の化合物の前記それぞれの決定された化合物スコアに基づいて、特定することを含む、請求項２に従属する場合の請求項２２に記載の方法。
【請求項２４】
最も低い決定された化合物スコアを有する前記初期の選択されたサブセット内の前記一つ以上の化合物が、置換のために特定される、請求項２３に記載の方法。
【請求項２５】
前記トレーニングセットに含まれない前記集団からの一つ以上の化合物を含む、前の反復で選択されたサブセットとは異なる、新しいサブセットを選択する工程と、
前記選択された新しいサブセットの前記サブセットスコアを決定し、当該決定されたスコアに基づいて前記選択された新しいサブセットを評価する工程と、を停止条件が満たされるまで、反復的に実行することを含む、請求項２０から２４のいずれかに記載の方法。
【請求項２６】
前記停止条件が、最大回数の反復が実行されたこと；
前記反復の一つにおいて選択された前記サブセットの前記サブセットスコアが所定の条件を満たすこと；および、
連続する反復における前記選択されたサブセットの前記それぞれのサブセットスコアの間の差が、所定の差の閾値未満であること；のうちの少なくとも一つを含む、請求項２５に記載の方法。
【請求項２７】
前記停止条件が満たされる前記反復で前記選択されたサブセットの前記化合物を合成して、前記化合物の一つ以上の生物学的特性を決定することを含む、請求項２５または２６に記載の方法。
【請求項２８】
各反復において複数の新しいサブセットを選択すること、
停止条件が満たされる前記反復において前記複数の選択されたサブセットのうちの一つを、前記それぞれの複数の選択されたサブセットの前記決定されたサブセットスコアに基づいて、特定すること、および
前記一つの特定されたサブセットの前記化合物を合成して、前記化合物の一つ以上の生物学的特性を決定すること、を含む、請求項２４から２７のいずれかに記載の方法。
【請求項２９】
前記特定されたサブセットが、前記停止条件が満たされる前記反復において、前記複数

のサブセットの中で最も高いサブセットスコアを有するサブセットである、請求項２８に記載の方法。
【請求項３０】
前記選択されたサブセットが第１のサブセットであり：
それぞれが前記トレーニングセットに含まれない前記集団からの複数の化合物を含む複数のサブセットを選択すること；
前記サブセットのそれぞれの前記サブセットスコアを決定すること；および
前記それぞれのサブセットの前記決定されたサブセットスコアに基づいて、前記複数のサブセットから前記第１のサブセットを選択すること；を含む、請求項１～２９のいずれかに記載の方法。
【請求項３１】
前記第１のサブセットが、前記複数のサブセットの中で最も高いサブセットスコアを有するサブセットとなるように選択される、請求項３０に記載の方法。
【請求項３２】
前記複数のサブセットがそれぞれ、同じ数の化合物を有する、請求項３０または３１に記載の方法。
【請求項３３】
前記評価する工程が、前記集団における前記化合物の活性レベルを予測するために、活性モデルから得られる前記選択されたサブセットの活性スコアに基づいて前記選択されたサブセットを評価することを含む、請求項１～３２のいずれかに記載の方法。
【請求項３４】
前記評価する工程が、前記決定されたサブセットスコアおよび前記活性スコアに基づいて、それらのスコアの所望のバランスに対して、前記選択されたサブセットを評価することを含む、請求項３３に記載の方法。
【請求項３５】
前記複数の新しいサブセットがそれぞれ、前記決定されたスコアと前記活性スコアとの間の異なるバランスを含む、請求項２８に従属する場合の請求項３３または３４に記載の方法。
【請求項３６】
前記複数の新しいサブセットが、停止条件が満たされる反復において、決定されたサブセットおよび活性スコアのパレートフロントを形成する、請求項３５に記載の方法。
【請求項３７】
前記トレーニングセットが、最初は空である、請求項１～３６のいずれかに記載の方法。
【請求項３８】
前記集団中の前記複数の化合物のそれぞれの前記分子特性が、前記化合物の構造的特徴を含む、請求項１～３７のいずれかに記載の方法。
【請求項３９】
前記集団中の前記複数の化合物のそれぞれの前記構造的特徴が、前記化合物中に存在するフラグメントに対応する、請求項３８に記載の方法。
【請求項４０】
前記複数の化合物のそれぞれに存在する前記フラグメントが、分子フィンガープリントとして表される、請求項３９に記載の方法。
【請求項４１】
前記分子フィンガープリントが、拡張接続フィンガープリント（ＥＣＦＰ）であり、任意にＥＣＦＰ０、ＥＣＦＰ２、ＥＣＦＰ４、ＥＣＦＰ６、ＥＣＦＰ８、ＥＣＦＰ１０またはＥＣＦＰ１２である、請求項４０に記載の方法。
【請求項４２】
前記集団中の前記複数の化合物のそれぞれの前記分子特性が、前記化合物の化学的特性を含む、請求項１～４１のいずれかに記載の方法。

【請求項４３】
前記集団中の前記複数の化合物のそれぞれの前記分子特性が、前記化合物の構造的特徴および化学的特性を含む、請求項１～４２のいずれかに記載の方法。
【請求項４４】
前記化学的特性が、前記それぞれの化合物が所定の標的分子に結合するときに示される相互作用のタイプに対応する、請求項４２または４３に記載の方法。
【請求項４５】
前記集団中の前記化合物の少なくとも一部のものの化学的性質が、前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプの予測に対応する、請求項４４に記載の方法。
【請求項４６】
前記予測が、前記それぞれの化合物が前記所定の標的分子に結合するときに、一つ以上の所定のタイプの相互作用のうちのどれが示されるかについての予測を含む、請求項４５に記載の方法。
【請求項４７】
前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプの予測を取得することを含む、請求項４５または４６に記載の方法。
【請求項４８】
各化合物についての予測を取得する工程が、
前記化合物の三次元表現を生成すること、および
前記生成された三次元表現を使用してドッキングプロセスを実行して、前記化合物が前記所定の標的分子に結合するときの好ましいドッキングポーズを予測すること、を含んでおり、
相互作用の前記示されるタイプが、前記ドッキングプロセスの結果に基づいて予測される、請求項４７に記載の方法。
【請求項４９】
前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプが、相互作用フィンガープリントとして、任意に、タンパク質－リガンド相互作用フィンガープリント（ＰＬＩＦ）として表される、請求項４４から４８のいずれかに記載の方法。
【請求項５０】
前記相互作用のタイプとして、水素結合相互作用、弱い水素結合相互作用、イオン相互作用、疎水性相互作用、面と面との芳香族相互作用、端と面との芳香族相互作用、π－カチオン相互作用、および金属錯体形成相互作用のうちの一つ以上が含まれる、請求項４４から４９のいずれかに記載の方法。
【請求項５１】
前記集団中の前記化合物のそれぞれがリガンドであり、前記所定の標的分子がタンパク質である、請求項４４から５０のいずれかに記載の方法。
【請求項５２】
前記一つ以上の生物学的特性が、活性、選択性、毒性、吸収、分布、代謝および排出のうちの一つ以上を含む、請求項１～５１のいずれかに記載の方法。
【請求項５３】
前記生物学的特性の一つ以上が、それぞれの所望の生物学的特性に対して定義される、請求項１～５２のいずれかに記載の方法。
【請求項５４】
前記集団内の化合物の一つ以上の生物学的特性を、前記化合物の前記一つ以上の分子特性の関数として近似するためのマシンラーニングモデルを定義すること、および
化合物の前記トレーニングセットを使用して前記マシンラーニングモデルをトレーニングすること、を含んでなる、請求項１～５３のいずれかに記載の方法。
【請求項５５】

一つ以上の化合物が前記トレーニングセットに加えられるたびにトレーニング工程を実行することを含む、請求項５４に記載の方法。
【請求項５６】
前記マシンラーニングモデルが、ベイズ最適化モデル、回帰モデル、クラスタリングモデル、デシジョンツリーモデル、ランダムフォレストモデルおよびニューラルネットワークモデルのうちの少なくとも一つである、請求項５４または５５に記載の方法。
【請求項５７】
トレーニング工程の後に、前記マシンラーニングモデルを実行して、一つ以上の所望の生物学的特性を有する前記集団中の一つ以上の化合物を予測することを含む、請求項５４から５６のいずれかに記載の方法。
【請求項５８】
前記一つ以上の予測化合物の少なくとも一つを合成することをさらに含む、請求項５７に記載の方法。
【請求項５９】
前記一つ以上の予測化合物が、所定の標的分子に対して所望の生物学的、生化学的、生理学的および／または薬理学的活性を有する候補薬剤または治療分子である、請求項５７または５８に記載の方法。
【請求項６０】
前記所定の標的分子が、インビトロおよび／またはインビボの治療、診断または実験アッセイ標的である、請求項５９に記載の方法。
【請求項６１】
前記候補薬剤または治療分子が、医学において、例えば、ヒトまたはヒト以外の動物などの動物の治療方法において、使用されるためのものである、請求項５９または６０に記載の方法。
【請求項６２】
請求項１～６１のいずれかの方法によって特定された化合物。
【請求項６３】
コンピュータプロセッサによって実行されるときに、当該コンピュータプロセッサに請求項１～６１のいずれかに記載の方法を実行させる命令を記憶する非一時的なコンピュータ可読記憶媒体。
【請求項６４】
各化合物が一つ以上の分子特性を有する複数の化合物の集団を示すデータを受け取り、一つ以上の生物学的特性が既知である前記集団からの化合物のトレーニングセットを示すデータを受け取るように構成されたインプット部；
前記トレーニングセットに含まれない前記集団からの一つ以上の化合物のサブセットを選択し、前記選択されたサブセット内の前記一つ以上の化合物の分子特性に応じて前記選択されたサブセットのサブセットスコアを決定し、および、前記決定されたサブセットスコアに基づいて前記選択されたサブセットを評価するように構成されたプロセッサ；および
前記評価の結果を出力するように構成されたアウトプット部；を含んでなり、
前記サブセットスコアが、前記集団における前記分子特性の頻度と、前記トレーニングセットおよび前記選択されたサブセットを含むサンプリングされたセットにおける前記分子特性の頻度と、に応じて決定されるものである、コンピュータによる薬剤設計のためのコンピューティングデバイス。
【請求項６５】
前記プロセッサは、請求項１～６１のいずれかに記載の方法を実行するように構成されている、請求項６４に記載のコンピューティングデバイス。
【発明の詳細な説明】
【技術分野】
【０００１】

本発明は、薬剤などの化合物のコンピュータによる設計のための方法およびシステムに関する。特に、本発明は、選択された標的分子と相互作用する薬剤の設計に使用されるアクティブラーニングによるコンピュータによるモデルの最適化方法、およびこれらのシステムおよび方法を使用して設計された前記薬剤に関する。

【背景技術】

【０００２】

創薬は、前臨床試験などの医薬品開発の次の段階に進むための候補化合物を特定するプロセスである。このような候補化合物は、さらなる開発のために特定の基準を満たす必要がある。現代の創薬には、初期のスクリーニングの「ヒット」化合物の特定と最適化が含まれる。特に、このような化合物は、多数の異なる特性の最適化を含む、必要な基準に対して最適化する必要がある。最適化される特性には、例えば：所望の標的に対する有効性／効力；望ましくない標的に対する選択性；毒性の可能性が低いこと；および、良好な薬剤代謝と薬剤動態特性（ＡＤＭＥ）、が含まれる。前記指定された要件を満たす化合物のみが、創薬プロセスに進むことができる候補化合物となる。

【０００３】

前記創薬プロセスでは、最初のスクリーニングヒットから候補化合物までの最適化中に、かなりの数の化合物の製造／合成が含まれる場合がある。特に、合成された化合物は、生物学的活性などの特性を決定するために測定される。しかし、特定の創薬プロジェクトの一環として作るべき化合物の数は、合成して試験できる化合物の数をおそらく桁違いに上回る。したがって、合成された化合物の測定の結果は、分析され、候補化合物に必要なさまざまな基準に対してさらに改善された特性を備えた化合物が得られる可能性を最大化するために、次にどの化合物を合成するかの決定を与えるために使用される。

【０００４】

特定の段階での一つ以上の化合物の合成とそれに続く生物学的活性の測定は、前記創薬プロセスの設計サイクル（または反復）と呼ばれる。通常、該プロセスの設計サイクルごとに一連の化合物が合成され試験されるが、これは、一回に一つの化合物を合成して試験するよりも効率的だからである。ただし、利用可能なリソースのレベルは、通常、既定の設計サイクルで合成することができるセット内の化合物の数には上限があることを意味している。

【０００５】

創薬プロジェクトでは、候補化合物が見つかるまでに、通常、数百、さらには数千の化合物が数回の設計サイクルにわたって合成される。これは、時間と費用がかかり、非効率的なプロセスであり：単一の化合物の合成には数千ポンドの費用がかかり、単一の候補化合物を得るには平均して３から５年かかることがある。

【０００６】

コンピュータによる手法を使用すると、医薬品化学者が単独で実行できる分析と比較して、すでに合成された化合物に対して実行できる分析のレベルが大幅に向上する。特に、マシンラーニング（ＭＬ）、人工知能（Ａｌ）、またはその他の数学的手法を使用して、人間の能力を超えたレベルで多数の設計パラメータを並行して評価し、パラメータと生物活性レベルのような望まれる特性との間の関係を特定できる。数学的手法は、これらの特定された関係を使用して、候補化合物の必要な基準と比較して、どの化合物がより多くの／レベルの望ましい特性を示す可能性が高いかについて、より適切な予測を行うことができる。これは、このような数学的手法を使用して設計サイクルの数を減らし、候補化合物に必要な特性の望ましい組み合わせを達成する化合物を得るために合成する必要がある化合物の数を減らすことができることを意味し、それにより創薬プロジェクトにかかるコストと時間の削減を実現する。

【０００７】

既に合成および試験された化合物のみを使用して、どの化合物が、望ましい特性、例えば、最高の生物活性を示す可能性が最も高いか予測するように設計されたＭＬモデルをトレーニングすることができる。したがって、望ましい特性を最適化するために次の設計サ

イクルでどの化合物を合成するのが最適であるかの予測は、前記ＭＬモデルをトレーニングするために利用できるデータ、すなわち以前に合成された化合物と同程度の精度しかない。特に、ＭＬモデルは、当該ＭＬモデルのトレーニングに使用されるセットに十分な数の化合物がある場合にのみ正確な予測を行い（可能性が高く）；このトレーニングセット内の前記化合物は、合成される化合物がそこから選択される化合物のプールを十分に代表している。

【発明の概要】

【発明が解決しようとする課題】

【０００８】

前述したように、規定の創薬プロジェクトで製造できる化合物の化学空間は広大であり得る。したがって、リソースを無駄にしないためには、ＭＬモデルの改善に最も効果的な化合物を合成用に選択し、それらが次の設計サイクルのトレーニングセットの一部となるようにする、すなわち、より良いＭＬモデルが以降の反復で利用できることが重要である。繰り返すと、コンピュータによる手法を使用して、前記ＭＬモデルの予測能力を最大限に向上させるために、トレーニングセットにどの化合物を加えるかを提案することができる。このようなコンピュータによる手法は、単独で使用する場合でも、医薬品化学者の専門知識と組み合わせて使用する場合でも、医薬品化学者の専門知識のみに依存する場合と比較して、ＭＬモデルの改善レベルを高めることができる。しかしながら、トレーニングセットに追加される最良のデータポイントを選択するための従来技術の方法は、創薬プロジェクトには最適ではない可能性がある。この理由の一つは、化学空間が、他の物理空間や理論空間とは異なり、等間隔ではないため、そのような仮定に基づく測定基準があまり効果的でない可能性があるからである。

【０００９】

本発明はこのような背景に基づいて設定される。

【００１０】

発明の概要

本発明は、概して、ＭＬモデルのトレーニングを最適化して、最終的には、該トレーニングされたモデルを使用して、必要な基準に対して最適化された化合物をより高い精度で設計および自動的に選択することができる、化学空間における化合物の設計および自動選択のためのコンピュータによる方法およびシステムに関する。

【課題を解決するための手段】

【００１１】

本発明の一態様によれば、コンピュータによる薬剤設計の方法が提供される。この方法は、各化合物が一つ以上の分子特性を有する複数の化合物の集団を定義することを含む。この方法は、一つ以上の生物学的特性が既知である前記集団からの化合物のトレーニングセットを定義することを含む。この方法は、前記トレーニングセットに含まれない前記集団からの一つ以上の化合物のサブセットを選択することを含む。この方法は、前記選択されたサブセット中の前記一つ以上の化合物の分子特性に応じて、前記選択されたサブセットのサブセットスコアを決定することと、当該決定されたサブセットスコアに基づいて前記選択されたサブセットを評価することと、を含む。前記サブセットスコアは、前記集団における前記分子特性の頻度、および前記トレーニングセットおよび前記選択されたサブセットを含むサンプリングされたセットにおける前記分子特性の頻度に応じて決定される。

【００１２】

前記決定工程は、前記選択されたサブセットの前記一つ以上の化合物のそれぞれについて、前記化合物の一つ以上の分子特性に応じて、化合物スコアを決定することを含むことができ、前記サブセットスコアは、前記選択されたサブセットにおける各化合物の前記決定された化合物スコアに基づいて決定される。

【００１３】

前記サブセットスコアは、前記選択されたサブセット内の前記化合物の前記それぞれの

化合物スコアの合計として決定され得る。

【００１４】

前記選択されたサブセット内の前記化合物の一つの前記化合物スコアを決定することは、前記化合物の前記一つ以上の分子特性のそれぞれの分子特性スコアを、前記集団における前記それぞれの分子特性の頻度、および前記サンプリングされたセットにおける前記それぞれの分子特性の頻度に応じて、決定することを含んでよく、前記化合物の前記化合物スコアは、前記化合物の前記一つ以上の分子特性の前記決定されたスコアに基づくものである。

【００１５】

前記化合物の前記化合物スコアは、前記化合物の前記一つ以上の分子特性の前記決定された分子特性スコアの合計として決定してよい。

【００１６】

前記一つ以上の分子特性のそれぞれの前記分子特性スコアは、前記サンプリングされたセット内に存在する前記分子特性の正規化された確率に応じて決定してよく、前記正規化された確率は、前記集団および前記サンプリングされたセット内における前記分子特性の頻度に応じて決定される。

【００１７】

前記正規化された確率は、前記集団内の化合物の数に対する、前記サンプリングされたセット内の化合物の数に応じて決定され得る。

【００１８】

前記正規化確率は、ラプラシアン補正された正規化確率であってよい。

【００１９】

前記ラプラシアン補正された正規化確率Ｐｃｏｒｒは、次の式で与えられる。

【００２０】

【数１】

$$P_{\mathrm{corr}} = \frac{F_{\mathrm{sampled}} + 1}{F_{\mathrm{set}} + 1/P_{\mathrm{base}}}$$

【００２１】

式中、Ｆｓａｍｐｌｅｄは、前記サンプリングされたセットにおける前記分子特性の頻度であり、Ｆｓｅｔは、前記集団における前記分子特性の頻度であり、Ｐｂａｓｅは、前記サンプリングされたセットにおける化合物の数を前記集団における化合物の数で割ったものである。

【００２２】

前記一つ以上の分子特性のそれぞれの前記分子特性スコアは、前記サンプリングされたセット内の化合物の数に対する、前記分子特性が存在する前記サンプリングされたセット内の化合物の数に応じて決定され得る。

【００２３】

前記分子特性スコアは、前記サンプリングされたセットにおける前記分子特性の正規化されたシャノンエントロピー値に応じて決定され得る。

【００２４】

前記正規化されたシャノンエントロピー値は、次式により与えられる：

【００２５】

【数２】

$$SC = \frac{-f\ln(f) - (1-f)\ln(1-f)}{\ln(2)}$$

【００２６】

式中、ｆは、前記分子特性が存在する前記サンプリングされたセット内の化合物の数を、前記サンプリングされたセット内の化合物の数で割ったものである。

【００２７】

前記分子特性スコアＣｏｖ$_{final}$は、次式で与えられる：

【００２８】

【数３】

$$Cov_{\text{final}} = \begin{cases} Cov * SC, & Cov \geq 0 \\ Cov * (2 - SC), & Cov < 0 \text{ and } f > 0.5 \end{cases}$$

【００２９】

式中、Ｃｏｖは以下の通りである。

【００３０】

【数４】

$$Cov = -\ln(P_{\text{corr}}/P_{\text{base}})$$

【００３１】

前記サブセットには、所定数の化合物が含まれていてよい。

【００３２】

前記方法は、前記サブセット内で選択される化合物の数を定義することを含んでよい。

【００３３】

前記評価工程は、前記サブセットスコアが所定の条件を満たすかどうかを判定することを含んでよい。

【００３４】

前記所定の条件は、前記サブセットスコアが所定の最小閾値スコアより大きいことであってよい。

【００３５】

前記所定の条件が満たされる場合、前記方法は、前記選択されたサブセット内の前記化合物の少なくとも一部を合成して、前記化合物の一つ以上の生物学的特性を決定することを含み得る。

【００３６】

前記方法は、前記合成された化合物を前記トレーニングセットに加えることを含んでよい。

【００３７】

前記選択されたサブセットは初期の選択されたサブセットであってよく、前記方法は：前記トレーニングセットに含まれない前記集団からの一つ以上の化合物を含む、前記初期の選択されたサブセットとは異なる、第２サブセットを選択すること；および、前記選択された第２サブセットの前記サブセットスコアを決定し、当該決定されたスコアに基づいて前記選択された第２サブセットを評価することを含んでよい。

【００３８】

前記第２サブセットを選択し、そのスコアを決定する前記工程は、前記所定の条件が満たされない場合に、実行されてよい。

【００３９】

前記第２サブセットを選択することは、前記初期の選択されたサブセット内の一つ以上の化合物を、前記トレーニングセットに含まれない前記集団からの一つ以上の新しい化合物で置き換えることを含み得る。

【００４０】

前記方法は、前記初期の選択されたサブセット内の前記一つ以上の化合物の前記それぞれ決定された化合物スコアに基づいて、置換される前記初期の選択されたサブセットから

の前記一つ以上の化合物を特定することを含み得る。

【００４１】

最低の決定された化合物スコアを有する前記初期の選択されたサブセット内の前記一つ以上の化合物が、置換のために特定され得る。

【００４２】

前記方法は、次の工程を反復的に実行することを含んでよい：前記トレーニングセットに含まれない前記集団からの一つ以上の化合物を含む、前の反復で選択されたサブセットとは異なる新しいサブセットを選択する工程；および、前記選択された新しいサブセットの前記サブセットスコアを決定し、停止条件が満たされるまで、前記決定されたスコアに基づいて前記選択された新しいサブセットを評価する工程。

【００４３】

前記停止条件には、最大回数の反復が実行されたこと；前記反復の一つで選択された前記サブセットの前記サブセットスコアが前記所定の条件を満たすこと；および、連続する反復における前記選択されたサブセットの前記それぞれのサブセットスコアの間の差が、所定の差の閾値未満である；のうちの少なくとも一つを含んでよい。

【００４４】

前記方法は、前記停止条件が満たされる前記反復で前記選択されたサブセットの前記化合物を合成して、前記化合物の一つ以上の生物学的特性を決定することを含んでよい。

【００４５】

前記方法は、反復ごとに複数の新しいサブセットを選択すること；前記それぞれの複数の選択されたサブセットの前記決定されたサブセットスコアに基づいて、前記停止条件が満たされる前記反復で前記複数の選択されたサブセットのうちの一つを特定すること；および前記一つの特定されたサブセットの前記化合物を合成して、前記化合物の一つ以上の生物学的特性を決定すること；を含んでよい。

【００４６】

前記特定されたサブセットは、前記停止条件が満たされる前記反復において前記複数のサブセットの中で最も高いサブセットスコアを有するサブセットであってよい。

【００４７】

前記選択されたサブセットは第１のサブセットであってよく、前記方法は以下の工程を含み得る：それぞれが前記トレーニングセットに含まれない前記集団からの複数の化合物を含む複数のサブセットを選択する工程；前記サブセットのそれぞれの前記サブセットスコアを決定する工程；および、前記それぞれのサブセットの前記決定されたサブセットスコアに基づいて、前記複数のサブセットから前記第１のサブセットを選択する工程。

【００４８】

前記第１のサブセットは、前記複数のサブセットの中で最も高いサブセットスコアを有するサブセットとなるように選択され得る。

【００４９】

前記複数のサブセットはそれぞれ、同じ数の化合物を有し得る。

【００５０】

前記評価工程は、前記集団中の前記化合物の活性レベルを予測するために、活性モデルから得られた前記選択されたサブセットの活性スコアに基づいて、前記選択されたサブセットを評価することを含み得る。

【００５１】

前記評価工程は、前記決定されたサブセットスコアおよび前記活性スコアに基づいて、それらのスコアの所望のバランスに対して、前記選択されたサブセットを評価することを含んでよい。

【００５２】

前記複数の新しいサブセットはそれぞれ、前記決定されたスコアと前記活性スコアとの間の異なるバランスを含み得る。

【００５３】

前記複数の新しいサブセットは、前記停止条件が満たされる前記反復において、決定されたサブセットと活性スコアとのパレートフロントを形成することができる。

【００５４】

前記トレーニングセットは、最初は空であってよい。

【００５５】

前記集団内の前記複数の化合物のそれぞれの前記分子特性には、前記化合物の構造的特徴が含まれ得る。

【００５６】

前記集団中の前記複数の化合物のそれぞれの前記構造的特徴は、前記化合物中に存在するフラグメントに対応し得る。

【００５７】

前記複数の化合物のそれぞれに存在する前記フラグメントは、分子フィンガープリントとして表すことができる。

【００５８】

前記分子フィンガープリントは、拡張接続フィンガープリント（ＥＣＦＰ）であってもよく、任意にＥＣＦＰ０，ＥＣＦＰ２，ＥＣＦＰ４，ＥＣＦＰ６，ＥＣＦＰ８，ＥＣＦＰ１０またはＥＣＦＰ１２であってよい。

【００５９】

前記集団内の前記複数の化合物のそれぞれの前記分子特性には、前記化合物の化学的特性が含まれ得る。

【００６０】

前記集団内の前記複数の化合物のそれぞれの前記分子特性には、前記化合物の構造的特徴および化学的特性が含まれ得る。

【００６１】

前記化学的特性は、前記それぞれの化合物が所定の標的分子に結合するときに示される相互作用のタイプに対応し得る。

【００６２】

前記集団内の前記化合物の少なくともいくつかの前記化学的特性は、前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプの予測に対応することがある。

【００６３】

前記予測は、前記それぞれの化合物が前記所定の標的分子に結合するときに、一つ以上の所定のタイプの相互作用のうちのどれが示されるかについての予測を含んでよい。

【００６４】

前記方法は、前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプの予測を取得することを含んでよい。

【００６５】

各化合物について前記予測を取得することは：前記化合物の三次元表現を生成すること；および、前記化合物が前記所定の標的分子に結合するときの好ましいドッキングポーズを予測するために、前記生成された三次元表現を使用してドッキングプロセスを実行すること；を含んでよく、前記示される相互作用のタイプは、前記ドッキングプロセスの結果に基づいて予測される。

【００６６】

前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプは、相互作用フィンガープリントとして表すことができる。任意に、前記相互作用フィンガープリントは、タンパク質－リガンド相互作用フィンガープリント（ＰＬＩＦ）である。

【００６７】

前記タイプの相互作用には、水素結合相互作用、弱い水素結合相互作用、イオン相互作用、疎水性相互作用、面と面との芳香族相互作用、端と面との芳香族相互作用、π－カチ

オン相互作用および金属錯体形成相互作用のうちの一つ以上を含み得る。
【００６８】
前記集団内の前記化合物のそれぞれはリガンドであってよく、前記所定の標的分子はタンパク質である。
【００６９】
前記集団内の前記複数の化合物のそれぞれの前記分子特性には、前記化合物の物理的特性が含まれ得る。
【００７０】
前記一つ以上の生物学的特性には、活性、選択性、毒性、吸収、分布、代謝および排出のうちの一つ以上が含まれ得る。
【００７１】
前記生物学的特性の一つ以上が、それぞれの所望の生物学的特性に対して定義され得る。
【００７２】
前記方法は、前記集団内の化合物の一つ以上の生物学的特性を、前記化合物の前記一つ以上の分子特性の関数として近似するためのマシンラーニングモデルを定義すること；および、化合物の前記トレーニングセットを使用して前記マシンラーニングモデルをトレーニングすることを含み得る。
【００７３】
前記方法は、一つ以上の化合物が前記トレーニングセットに加えられるたびに前記トレーニング工程を実行することを含んでよい。
【００７４】
前記マシンラーニングモデルは、ベイズ最適化モデル、回帰モデル、クラスタリングモデル、デシジョンツリーモデル、ランダムフォレストモデル、およびニューラルネットワークモデルのうちの少なくとも一つであってよい。
【００７５】
前記方法は、前記トレーニング工程後に、前記マシンラーニングモデルを実行して、一つ以上の所望の生物学的特性を有する前記集団中の一つ以上の化合物を予測することを含んでよい。
【００７６】
前記方法は、前記一つ以上の予測化合物のうちの少なくとも一つを合成することをさらに含んでよい。
【００７７】
前記一つ以上の予測化合物は、所定の標的分子に対して所望の生物学的、生化学的、生理学的および／または薬理学的活性を有する候補薬剤または治療分子であり得る。
【００７８】
前記所定の標的分子は、インビトロおよび／またはインビボの治療、診断、または実験アッセイの標的であり得る。
【００７９】
前記候補薬剤または治療分子は医学で使用してよく、例えば、ヒトまたはヒト以外の動物などの動物の治療方法において使用してよい。
【００８０】
本発明の別の態様によれば、上記の方法によって特定された化合物が提供される。
【００８１】
本発明の別の態様によれば、コンピュータプロセッサによって実行されると、当該コンピュータプロセッサに上述の方法を実行させる命令を記憶する非一時的なコンピュータ可読記憶媒体が提供される。
【００８２】
本発明の別の態様によれば、コンピュータによる薬剤設計のためのコンピューティングデバイスが提供される。当該コンピューティングデバイスは、各化合物が一つ以上の分子

特性を有する複数の化合物の集団を示すデータを受け取り、一つ以上の生物学的特性が知られている前記集団からの化合物のトレーニングセットを示すデータを受け取るように構成されたインプット部を含む。前記コンピューティングデバイスは、前記トレーニングセットに含まれない前記集団からの一つ以上の化合物のサブセットを選択し、前記選択されたサブセットにおける前記一つ以上の化合物の分子特性に応じて前記選択されたサブセットのサブセットスコアを決定し、前記決定されたサブセットスコアに基づいて前記選択されたサブセットを評価するように構成されたプロセッサを含む。前記コンピューティングデバイスは、前記評価の結果を出力するように構成されたアウトプット部を含む。前記サブセットスコアは、前記集団における前記分子特性の頻度と、前記トレーニングセットおよび前記選択されたサブセットを含むサンプリングされたセットにおける前記分子特性の頻度とに応じて決定される。

【００８３】

前記プロセッサは、上述の方法を実行するように構成され得る。

【００８４】

次に、以下の図面を参照して本発明の実施例を説明する。

【図面の簡単な説明】

【００８５】

【図１】アスピリン分子のＥＣＦＰ２フィンガープリントを概略的に示した図である。

【図２】化合物の例のセットにおける少なくともいくつかの化合物に存在する構造的特徴のシャノンエントロピースコアを示す表である。

【図３】図２の化合物の例のセットに異なる構造的特徴が存在する頻度に対するシャノンエントロピースコアのプロットを示した図である。

【図４】化合物の例のセットにおける化合物の構造的特徴をシャノンエントロピースコアの順にリストした図である。

【図５】化合物の一例のセットまたは集団における化合物の事前セットと化合物の選択セットとの間の関係を模式的に示した図である。

【図６】本発明による方法の工程を要約した図である。

【図７（ａ）】化学相互作用空間（「相互作用空間」）における化合物の集団からの化合物のサブセットの選択を示した図であり、図６の方法の例に従って選択が行われている。

【図７（ｂ）】化学構造空間（「化学空間」）における化合物の集団からの化合物のサブセットの選択を示した図であり、図６の方法の例に従って選択が行われている。

【図８（ａ）】化学相互作用空間（「相互作用空間」）において、従来の方法によるサブセット選択に対して、図７（ａ）および図７（ｂ）からのサブセット選択の一つを示している。

【図８（ｂ）】化学構造空間（「化学空間」）において、従来の方法によるサブセット選択に対して、図７（ａ）および図７（ｂ）からのサブセット選択の一つを示している。

【発明を実施するための形態】

【００８６】

詳細な説明

分子または薬剤の設計は、知識を進歩させるために仮説の生成と実験のサイクルを使用する多次元の最適化課題と考えることができる。それぞれの化合物の設計は、実験によって反証された仮説であると考えることができる。実験結果は構造活性関係として表され、どの化学構造が望ましい特性を含む可能性が高いかに関する仮説のランドスケープを構築する。各プロジェクトは、望ましい指定された属性の製品プロファイル、すなわちターゲット機能から始まるため、薬剤設計のプロセスも最適化の問題である。しかし、たとえ目的が正確に記述できたとしても、最適な解決を見つけることはこれまで費用がかかり、困難な課題であった。このタイプの問題の特に難しい点の一つは、実験結果の比較的限られた知識ベースから、実行可能な解決策の広大な空間にわたる仮説のランドスケープを効果的に構築することである。

【００８７】

前記創薬プロセスは通常、設計サイクルとして知られる反復で実行される。各反復で一連の分子または化合物が合成され、それらの生物学的特性が測定される。活性が分析され、以前の反復から学んだことに基づいて新しい化合物のセットが提案される。このプロセスは、臨床候補が見つかるまで繰り返される。活性だけでなく、測定される生物学的特性には、選択性，毒性，親和性，吸収，分布，代謝および排出のうちの一つ以上が含まれる場合がある。

【００８８】

プロセスの特定の段階では、一連の化合物が合成または製造されており、その生物学的活性は既知である。このプロセスの目的は、合成可能であるものの、その集団からの化合物のサブセットを合成するためのリソースおよび／または時間しかない大規模な集団または化合物のプールから一つ以上の最適な化合物を見つけることである。

【００８９】

自動化またはコンピュータによる薬剤設計プロセスでは、数学的モデル、例えば、マシンラーニング（ＭＬ）モデルを使用して、製造される可能性のある化合物の集団のうちどの化合物が最適な化合物、例えば、生物学的活性を最大化する化合物であるかを予測または仮説立てる。前記ＭＬモデルは、実験結果、すなわちすでに合成および試験された集団内の化合物から得られる利用可能な構造活性関係を使用してトレーニングされる。ＭＬモデルを使用して、考えられる化合物の集団から予測活性が最も高い化合物を合成用に選択する戦略またはアプローチは、「活用：エクスプロイテーション（exploitation）」と呼ばれる。活用戦略は、プロセスの使用フェーズとみなすことができる。

【００９０】

このアプローチは、前記ＭＬモデルの予測能力が十分に正確である場合、すなわち前記ＭＬモデルが十分にトレーニングされている場合にのみ成功する。合成および試験された集団からの各化合物は、前記ＭＬモデルのトレーニングに使用される化合物のトレーニングセットに追加される。特定の反復でトレーニングセットに追加される分子または化合物の数は、通常、リソースによって制限される。すなわち、各反復で合成される化合物のサブセット内の化合物の数は、所定の最大数で定義される。

【００９１】

前記ＭＬモデルの予測能力は、トレーニングセットに十分な数の化合物がある場合にのみ十分に正確になる。そのため、前記ＭＬモデルが十分にトレーニングされる前に、特定の回数の反復または設計サイクル（例えば、各反復で規定の最大数の化合物がトレーニングセットに追加される）を実行する必要がある場合がある。

【００９２】

また、前記ＭＬモデルの予測能力は、トレーニングセット内の化合物が、合成用に選択できる化合物の集団全体を十分に代表している場合にのみ、十分に正確になる。したがって、前記ＭＬモデルが十分にトレーニングされる前に、前記ＭＬモデルの改善に最も役立つ化合物、すなわち最も代表的な化合物が、特定の反復で合成されるサブセットに含まれていることが重要である。これに基づいて合成する化合物を選択することを「探索（exploration）」と呼ばれる。探索戦略は、プロセスのラーニングフェーズまたはトレーニングフェーズとみなされる場合がある。

【００９３】

したがって、創薬プロセスの特定の繰り返しで合成する化合物のサブセットを選択する場合、活用戦略と探索戦略には競合するニーズがある。実際、どの戦略が適切であるかという選択は、創薬プロセスの特定の段階に応じて変わる可能性がある。例えば、創薬プロジェクトの初期の段階では、十分にトレーニングされたモデルがまだ構築されている可能性は低くなる。したがって、探索の報酬は最終的にはよりよくトレーニングされ、したがってより正確なモデルとなるため、この段階での探索戦略は最も適切な戦略である可能性がある。活用戦略は、トレーニングセットの代表性を高めるための特に優れた戦略ではないため、この段階では限られたリソースを最大限に活用することはできない。一方、ＭＬモデルがすでに十分にトレーニングされている場合（例えば、創薬プロジェクトの後期段

階）、合成用モデルによって選択された化合物のサブセットが望ましい特性、例えば高い生物学的活性レベルと比較して、最適な化合物である可能性が高くなるので、その場合、活用が適切な戦略であろう。現段階では、探索戦略は、望ましい特性を持つ可能性が高い化合物を選択するための最適な戦略ではないため、限られたリソースを最大限に活用することはできない。

【００９４】

前述したように、ＭＬモデルは次の場合にのみ正確な予測を行う（可能性が高い）：ＭＬモデルのトレーニングに使用されるセット内に十分な数の化合物がある；および、このトレーニングセット内の化合物は、合成する化合物が選択される化合物のプールを十分に代表している。これらの一つ目は、十分な数の合成化合物を取得するために、一定数の設計サイクルを実行する必要がある可能性があることを意味する（十分な数の以前に合成された化合物に関するデータがすでに利用可能な場合を除く）。二番目は、創薬プロジェクトの初期の段階の初期の設計サイクルでは、設計されたＭＬモデルを（単独で）使用して合成するセットにどの化合物を含めるかを決定するのが望ましくない可能性があることを意味する。これは、ＭＬモデルは、まだ十分なレベルにトレーニングされていないモデルに従ってどの化合物が高活性であるかを予測するため、予測が正確である可能性が低くなる。さらに、ＭＬモデルの予測は化合物のトレーニングセットからすでに特定されている関係／情報にさらに焦点を当てているため、このような予測に従って化合物を合成しても、その後の設計サイクルでＭＬモデルを改善するのには役に立たない。特に、ＭＬモデルの予測は、次の設計サイクルに向けてＭＬモデルを改善する目的でどの化合物を合成するかを提案するのには役に立たない。

【００９５】

創薬プロジェクトに関連する時間とコストを削減するには、望ましい特性を持つ候補または最適な化合物を発見するために必要な反復または設計サイクルの数を最小限に抑える必要がある。したがって、望ましい特性を持つ化合物を予測するための十分にトレーニングされたモデルをできるだけ早く構築できること、すなわちトレーニングセットに必要な化合物をできるだけ少なくすることが重要である。そのため、候補化合物がこのような戦略を採用している反復から出現する可能性は低いため、（少なくともある程度の）探索が必要な反復回数を最小限に抑えるために、プロジェクトの初期の段階で最も代表的な化合物を合成用に選択することが重要である。

【００９６】

さらに、創薬プロセスの各反復では、探索と活用の組み合わせが実行される場合がある。すなわち、所与の反復での合成のために選択される化合物のサブセットにおいて、一部の化合物は探索戦略に従って選択され、一部の化合物は活用戦略に従って選択され得る。例えば、ＭＬモデルの精度は連続する反復ごとに向上する可能性が高いため、実行される反復回数が増加するにつれて、探索戦略に従ってサブセットに対して選択される化合物の数が減少する可能性がある。対照的に、活用戦略に従ってサブセットに対して選択される化合物の数は、実行される反復数が増加するにつれて増加する可能性がある。

【００９７】

本発明は、ＭＬモデルを十分なレベルまでトレーニングするための時間と費用が削減されるように、探索戦略の一部として、場合によっては活用戦略と組み合わせて、合成する化合物を選択するための改良されたコンピュータによる薬剤設計方法を提供するという点で有利である。

【００９８】

本発明によれば、コンピュータによる薬剤設計方法の第１の工程は、複数の化合物または分子の集団を定義することである。特に、この集団は、特定の創薬プロジェクト中に合成のために選択できる化合物のセットである。前記集団は、例えば既知の計算方法および／または人間の入力を介して、任意の適切な方法で定義または取得することができる。例えば、集団は、生成または進化設計アルゴリズムから取得された化合物のセットである場合がある。特に、進化的設計アルゴリズムは、本方法が使用されることになる特定のプロ

ジェクトに最適な化合物の望ましい特性の少なくとも一部を有する、一つ以上の既知の化合物の初期のセット（既存の薬剤など）に基づいて、多数の新規化合物を生成する可能性がある。あるいは、多数の新規化合物を任意の適切な方法で生成することもできる。少なくともいくつかの所望の特徴を有する生成された新規化合物は、さらなる分析のために保持され得る。一例では、既知の方法を追加して、特定のプロジェクトに少なくともいくつかの所望の特徴を備えた特定の化合物を手元に置いておくことによって、出発化合物群（例えば、数百万の化合物を含む）の数を減らすことができる。一つ以上のフィルターを保持された化合物に適用して、望ましくない化合物を除去することができる。これらのフィルターは、望ましくない化合物から望ましい化合物を選択（またはフィルタリング）するための任意の適切な基準に従って定義できる。例えば、一つの有用なフィルターを適用して、重複した化合物を除去することができる。別のフィルターを適用して、特定のレベルの毒性を持つ化合物を除去することもできる。次いで、フィルタリングされた化合物のセットが、合成のための選択が行われる集団を形成することができる。

【００９９】

前記集団には、任意の適切な数の化合物が含まれ得る。一般に、前記集団には、例えば、利用可能なリソースの理由から、特定の創薬プロジェクトの一部として合成できる化合物の数よりも多くの、そしておそらくかなり多くの化合物が含まれる。しかしながら、この集団には、一般に、本発明による集団のコンピュータ分析が実行不可能であるほど多くの化合物が含まれることもない。例えば、前記集団内の化合物の数は通常、数百または数千の化合物のオーダーである可能性があるが、任意のプロジェクトでは集団がこれよりも大きい場合も小さい場合もあることは理解される。

【０１００】

コンピュータによる薬剤設計方法の次の工程は、一つ以上の生物学的特性がわかっている前記集団からの化合物のトレーニングセットを定義することである。トレーニングセットは、特定の創薬プロジェクトの構造活性相関を評価するためのＭＬモデルをトレーニングするために使用される。トレーニングセットには、一つ以上の生物学的特性を決定するために実験的に合成および試験された集団からの化合物が含まれる。そのため、創薬プロジェクトが進行するにつれて、すなわち反復または設計サイクルが実行されるにつれて、トレーニングセット内の化合物の数が増加する。薬剤設計方法の開始時、すなわち集団内の化合物が試験される前、トレーニングセットは（最初は）空である可能性がある、すなわちトレーニングセットは化合物を含まない可能性がある。あるいは、トレーニングセットには、生物学的特性が事前に知られている化合物、例えば、別のプロジェクトの一部として以前に試験されており、検討中の特定のプロジェクトに従って最適な化合物の望ましい特性の少なくとも一部を有する化合物が含まれてもよい。

【０１０１】

計算論的設計法の次の工程は、集団から少なくとも一つの化合物のサブセットを選択することを含み、サブセット内の化合物はトレーニングセットには含まれない。選択される化合物の数は、利用可能なリソースを考慮して、薬剤設計プロジェクトの任意の反復または設計サイクルで試験できる化合物の数に基づく。したがって、サブセット内で選択される化合物の数、または少なくともその数の上限は、事前に決定することができる。一般に、この方法には、サブセット内で選択される化合物の数を指定することが含まれる。一つ以上の化合物のサブセットが選択される方法については、以下でより詳細に説明する。選択されたサブセットのサイズ、すなわち選択されたサブセット内の化合物の数は、集団のサイズよりも大幅に小さい可能性がある。例えば、選択されたサブセット内の化合物の数は、集団内の化合物の数より少なくとも一桁低くてもよく、場合により少なくとも一桁以上低くてもよい。

【０１０２】

最も広い意味では、本発明の計算による設計方法は、以下に記載するように、薬剤設計プロジェクトの所与の反復または設計サイクルにおける改良された探索戦略に従って化合物（のサブセット）を選択するための方法を提供するものとみなすことができる。しかし

ながら、任意の所与の設計サイクルにおいて、これを異なる戦略（例えば、異なる探索戦略または活用戦略）に従った化合物の選択と組み合わせることができることが理解されよう。

【０１０３】

本方法の探索戦略は情報理論に基づいている。特に、化合物の選択は、試験時にどの化合物が最大量の情報（例えば、集団における構造活性関係について）を提供するかに基づいて行われるべきである。化合物または化合物のサブセットが提供する情報の量またはタイプは、化合物の特徴に基づいて決定される。

【０１０４】

集団内の各化合物には、結合して化学構造を形成する多数の構造的特徴が含まれている。このような構造的特徴は、任意の適切な方法で表現することができる。例えば、化合物または分子の構造を記述する一つの方法は、フィンガープリンティングによるものである。特に、特定の化合物のフィンガープリントは、その化合物にどの特定の構造的特徴または部分構造が存在するか存在しないかを反映する数学的オブジェクト（例えば、一連のビットまたは整数のリスト）として表すことができる。

【０１０５】

フィンガープリントには、トポロジカルフィンガープリント、構造的フィンガープリント、円形フィンガープリントなど、いくつかの異なるクラスがある。一般的な循環フィンガープリンティング方法は、拡張接続フィンガープリンティング（ＥＣＦＰ）である。ＥＣＦＰ０，ＥＣＦＰ２，ＥＣＦＰ４，ＥＣＦＰ６など、数多くのＥＣＦＰメソッドが知られている。当技術分野で知られているように、化合物のフィンガープリントを決定することは、一般に、化合物内の各原子に識別子を割り当てること、隣接する原子に基づいてこれらの識別子を更新すること、重複を除去すること、次いで識別子のリストからベクトルを形成することを含む。

【０１０６】

純粋に説明を目的として、分子とそのＥＣＦＰ２フィンガープリント特徴の例を図１に示す。特に、説明として選択した分子はアスピリンである。この化合物には１７個のフィンガープリント特徴が含まれており、それぞれが該化合物の一個または一部を表しており、各特徴が（正または負の整数の）数値として保管されていることがわかる。

【０１０７】

創薬プロジェクトの一部として構築されるＭＬモデルは、化合物に存在する可能性のある構造的特徴を望ましい特性（生物学的活性）と関連付けることがある。例えば、標的に対して高い活性が望まれ、特定の化合物が高いレベルの活性を示すプロジェクトでは、その化合物のどの特徴が高い活性レベルに寄与しているのかを判断することが課題となる。この方法の目的は、構造活性関係に関して得られる情報量を最大化するような試験用の化合物を選択することである可能性がある。

【０１０８】

情報理論におけるシャノンエントロピー（または「情報エントロピー」、または単に「エントロピー」）は、情報内容の尺度である。一般に、データセットについて質問する場合、一部の質問は他の質問よりも有益である。シャノンエントロピーを使用すると、抽出された情報を最大化するためにどのような質問をするのが最適かを判断できる。尋ねるべき最適な（バイナリ）質問は、データセットを二つに均等に分割する質問である。本方法の文脈では、集団内の化合物の個々の特徴（例えば、フィンガープリント特徴）のエントロピーを決定することができる。特に、特定の特徴のエントロピーは、その特定の特徴が集団内にいくつの化合物に存在するかによって決まる。集団内の化合物の半分に存在し、残りの半分には存在しない特徴は、最も高いエントロピー値を持つ。一方、集団内の各化合物に存在する特徴、または集団内の各化合物に存在しない特徴は、最も低いエントロピー値（実際にはゼロ値）を持つ。比較的高いエントロピー値の特徴を有する化合物を試験することにより、例えば、どの特徴が高い活性レベルに寄与しているかをより容易に推定することができる。

【０１０９】
化合物の集団における特徴ｘのシャノンエントロピーＨは、次のように表すことができる。
【０１１０】
【数５】

$$H(x) = -\sum_i p(x_i) \ln p(x_i)$$

【０１１１】
式中、ｐ（ｘ$_i$）は、集団の化合物における特徴のさまざまな状態の確率、すなわち、化合物に存在するか、化合物に存在しないかの確率である。
【０１１２】
説明的な例として、２４００個の化合物を含む集団を考えてみる。第１の特徴（フィンガープリント特徴など）は、２４００個の化合物のうち１２００個に存在する。したがって、この第１の特徴のシャノンエントロピーは次のように計算できる。
【０１１３】
【数６】

$$H = -\left(\frac{1}{2}\right)\ln\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\ln\left(\frac{1}{2}\right) \cong 0.69.$$

【０１１４】
第２の特徴は、２４００個の化合物のうち５００個に存在する。したがって、この第２の特徴のシャノンエントロピーは次のようになる。
【０１１５】
【数７】

$$H = -\left(\frac{5}{24}\right)\ln\left(\frac{5}{24}\right) - \left(\frac{19}{24}\right)\ln\left(\frac{19}{24}\right) \cong 0.51.$$

【０１１６】
第１の特徴のシャノンエントロピーが第２の特徴のシャノンエントロピーよりも大きいことがわかる。実際、第１の特徴は集団内の化合物のちょうど半分に存在するため、第１の特徴は特徴レベルで最大の情報内容を提供する。
【０１１７】
したがって、集団内の特徴のシャノンエントロピーは、集団の化合物における特定の特徴の頻度、すなわち、特定の特徴が存在する集団内の化合物の数に依存する。例えば、Pickettら（２０１１）「遺伝的アルゴリズムを使用したＭＭＰ－１２阻害剤の自動リード最適化」、ACS Medicinal Chemistry Letters、２（１）、２８－３３に示されている２５００個の化合物のデータセットを使用すると、集団におけるシャノンエントロピースコアまたはさまざまな特徴の値（ＥＣＦＰフィンガープリントとして表される）を図２に示す。図２は、芳香環に結合したアミドの構造的特徴が、セットまたは集団、特にセット内の２５００個の化合物のうち５０個で比較的まれに発生することを示している。そのため、この特徴のシャノンスコアは０．０９８と比較的低くなる。スケールのもう一方の端では、ヒドロキシ（カルボン酸の部分構造）がセット内で非常に頻繁に発生し、実際にセット内の２５００個の化合物のそれぞれに存在しており、これは、そのシャノンスコアが０であることを意味する。図２に表す特徴では、エーテル酸素は、特徴が存在するかどうかという点で化合物のセットを二つに均等に分割するのに最も近いため、シャノンスコアが最も高くなる。すなわち、頻度０．５の特徴が最適な特徴となり、シャノンスコ

アが最大になる。

【０１１８】

Pickettらの文献で提示された上記の一連の化合物におけるＥＣＦＰ６特徴のシャノンエントロピースコアの広がりは、図３に示されているとおりである。特に、セット内の化合物の約半数に存在するフィンガープリントの特徴（ドットで表されている）が最も高いシャノンエントロピースコアを持っているのに対し、これらの特徴はごく少数の化合物に存在するか、またはほぼすべての化合物でシャノンエントロピースコアが最も低くなることがわかる。また、フィンガープリントの特徴の大部分が、セット内の化合物の半分未満に存在していることもわかる。図３では、単一のドットが複数の特徴を表す場合があることに注意されたい。

【０１１９】

情報理論を利用した探索戦略では、特徴のシャノンエントロピー値を使用して試験する化合物を選択することが一つのオプションである。特に、集団内の各化合物のシャノンエントロピーは、化合物の特徴のシャノンエントロピー値に基づいて計算できる。例えば、化合物のシャノンエントロピーは、前記化合物中に存在する特徴のシャノンエントロピー値の合計であり得る。しかしながら、化合物における特徴のシャノンエントロピー値は、化合物のエントロピー値を得るために任意の適切な方法で組み合わせることができる。したがって、探索戦略に従って最初の反復または設計サイクルの一部として選択される最初の化合物は、抽出された情報内容を最大化するために最も高いシャノンスコアを有する集団内の化合物となるように選択され得る。ただし、同じアプローチを使用して試験用に後続の化合物を選択した場合（同じ反復で試験される追加化合物として、すなわち選択されたサブセットの一部として、または後続の反復で試験される化合物として）、同じ有益な結果は得られない。特に、シャノンスコアを最大化することによる第２の化合物の選択は、一番番目と第２の化合物が同じ要因に基づいて選択されることを意味する可能性がある、すなわち、最初の化合物が高いスコアを持った要因は、第２の化合物に高スコアをもたらした要因と同じである可能性もある。これは、同様の質問を複数回行うこととみなされるため、抽出される情報量が減少する。したがって、必要なのは、一方ではシャノンエントロピースコアを最大化し、他方では選択された化合物に存在する特徴の重複を最小化することのバランスをとる化合物選択戦略である。

【０１２０】

選択プロセスにおいてさまざまな特徴がどの程度「アンダーサンプリング」されているかを決定するためのメトリクスまたは尺度を提供するアプローチを以下に説明する。アンダーサンプリングされた特徴は、高いシャノンエントロピースコアと、すでに選択されている特徴との低レベルの重複のバランスがとれた特徴である。図４は、化合物のセットまたは集団から５個の化合物のサブセットが選択された例を示している。特に、図４は、フィンガープリントの特徴（この場合はＥＣＦＰ４の特徴）をシャノンスコアの順に、すなわちセット内の化合物の総数に対して特徴を持つセット内の化合物の数に基づいてリストしている。図４には、選択した５個の化合物のうちのいくつに特徴が存在するかも示されている。シャノンスコアとサブセット内で特徴がサンプリングされる回数のバランスをとるスコアが定義される。ただし、二つの比率の単純な考慮に基づくスコアでは、入手可能な情報の一部が無視される。例えば、図４を参照すると、サブセット内で特徴がサンプリングされた回数と、特徴が存在する化合物の総数との比率の考慮は、２／２００か３／３００かは無視され、例えば、現在の状況ではより重要である。

【０１２１】

したがって、比率と有意性の両方をスコアまたは尺度に統合するメトリクスを定義できる。特に、定義されたメトリクスは、集団の化学空間全体にわたる特徴について抽出された情報の範囲を示すものであるため、集団内の（フィンガープリント）特徴の「カバレッジスコア」と呼ばれる。プロジェクトの探索フェーズでは、集団全体を広範囲にカバーする情報を抽出することは、例えば、構造活性相関を記述または予測するＭＬモデルを十分にトレーニングするための時間またはデータポイントの数を削減するのに有益である。

【０１２２】
特徴のカバレッジスコアが計算され（以下に概要を説明する）、これを使用して特定の化合物のカバレッジスコア、および実際に化合物の集団から選択（試験用など）された化合物のサブセットのカバレッジスコアを計算できる。図５を参照すると、化合物の所与の集団またはセット５０には、例えば、事前またはトレーニングセット５１が、現在の反復の一部として選択されたサブセット５２とは別のものである、すなわち、選択されたサブセット５２内の化合物が、事前またはトレーニングセット５１内の化合物とは異なっている、関連する創薬プロジェクトの以前の反復で、すでに（以前に）選択され試験された化合物の事前セットまたはトレーニングセット５１が存在する。サンプリングされた構造５１，５２または化合物Ｎｓａｍｐｌｅｄの総数は、化合物の事前セットと選択されたサブセットとの合計、すなわち、以下の如くである。

【０１２３】
【数８】

$$N_{\text{sampled}} = N_{\text{prior}} + N_{\text{subset}}$$

【０１２４】
式中、Ｎｐｒｉｏｒは、サブセットを選択する前に、例えば、前の設計サイクルですでに選択されている化合物の数（≧０）であり、Ｎｓｕｂｓｅｔは、所望のサブセットのサイズである。セットまたは集団からランダムな化合物をサンプリングする可能性または確率Ｐｂａｓｅは、次の式で与えられる。

【０１２５】
【数９】

$$P_{\text{base}} = \frac{N_{\text{sampled}}}{N_{\text{total}}}$$

【０１２６】
式中、Ｎｔｏｔａｌはセットまたは集団内の化合物の数である。

【０１２７】
セットまたは集団内の各化合物の特徴（フィンガープリント）が決定され、各グループ内の各特徴ｉの出現頻度が次のように求められる：Ｆｓｅｔ，ｉは、（全）セットまたは集団内の特徴ｉの頻度である、すなわち、特徴ｉが存在するセット内の化合物の数である；Ｆｐｉｏｒ，ｉは、事前またはトレーニング（サブ）セット内の特徴ｉの頻度であり；Ｆｓｕｂｓｅｔ，ｉは、（現在）選択されているサブセット内の特徴ｉの頻度であり、Ｆｓａｍｐｌｅｄ，ｉ＝Ｆｓｕｂｓｅｔ，ｉ＋Ｆｐｒｉｏｒ，ｉは、事前セットと選択されたサブセットとの組み合わせである所謂「サンプリングされたセット」における特徴ｉの頻度である。

【０１２８】
サンプリングされたセット内の特徴ｉの（ラプラシアン補正された）正規化確率Ｐｃｏｒｒ，ｉは、次のように計算される。

【０１２９】
【数１０】

$$P_{\text{corr},i} = \frac{F_{\text{sampled},i} + 1}{F_{\text{set},i} + 1/P_{\text{base}}}$$

【０１３０】
特徴ｉの「未補正」カバレッジスコアＣｏｖ｡は次のように定義できる。

【０１３１】

【数１１】

$$Cov_i = -\ln\left(P_{\mathrm{corr},i}/P_{\mathrm{base}}\right)$$

【０１３２】

このようにして、特徴がセットまたは集団内に存在する回数と比較して、特徴がサンプリングされた回数の尺度が提供される。ただし、さまざまな特徴によって提供される可能性のある情報内容も考慮して、未補正のカバレッジスコアに補正を適用する必要がある。このようにして、（サンプリングされた化合物のグループ内で）特徴が「オーバーサンプリング」されないことが保証される。特に、この補正は、サンプリングされたグループまたは化合物のセット内の特徴ｉのシャノンエントロピースコアに基づいている。サンプリングされた化合物グループ内の特徴ｉの頻度または割合は次のようになる。

【０１３３】

【数１２】

$$f_i = \frac{F_{\mathrm{sampled},i}}{N_{\mathrm{sampled}}}$$

【０１３４】

特徴ｉの（正規化された確率に対する）‘シャノン補正’S_iは、次のように与えられる。

【０１３５】

【数１３】

$$SC_i = \frac{-f_i\ln(f_i) - (1-f_i)\ln(1-f_i)}{\ln(2)}$$

【０１３６】

式中、分母は$0 < SC_i \leqq 1$となるようにシャノン補正を正規化する。シャノン補正は特徴がサンプリングされた頻度に依存するため、シャノン補正は事前セットおよび選択されたサブセット内の特定の化合物に応じて変化することに注意されたい（例えば、創薬プロジェクトの反復ごとに異なる）。これは、集団内の特徴について上述したシャノンスコアとは異なり、集団内の特定の化合物が創薬プロジェクトの異なる反復を通じて同じままであるため、これは一定である。

【０１３７】

いくつかの例では、未補正のカバレッジスコアがゼロより大きいか小さいかに応じて、シャノン補正がわずかに異なる方法で適用されてもよい。したがって、いくつかの例では、特徴ｉの最終（修正）カバレッジスコア$Cov_{final,i}$は次のように定義できる。

【０１３８】

【数１４】

$$Cov_{\mathrm{final},i} = \begin{cases} Cov_i * SC_i, & Cov_i \geq 0 \\ Cov_i * (2 - SC_i), & Cov_i < 0 \text{ and } f_i > 0.5 \end{cases}$$

【０１３９】

いくつかの例では、化合物のカバレッジスコアは、その特徴のカバレッジスコアの合計として計算され得、選択されたサブセットのカバレッジスコアは、選択されたサブセット内の化合物のカバレッジスコアの合計として計算され得る。特徴の（特徴）カバレッジスコアは、サンプリングされたセット（選択されたサブセットと事前セットを含む）内の特徴の頻度に依存するため、選択されたサブセット内の化合物の（化合物）カバレッジスコ

アは、選択されたサブセット（および以前のセット）内に他のどの化合物（および特にその特徴）が含まれるかによって決まる。すなわち、選択されたサブセットに複数の化合物が含まれており、それらの化合物の一つが事前セットに含まれない集団の別の化合物と置換される場合、（更新された）選択されたサブセットの（サブセット）カバレッジスコアを決定するには、（更新された）選択されたサブセット内の各化合物の（化合物）カバレッジスコアを再計算する必要がある。

【０１４０】

サンプリングされたセット内で特徴が「オーバーサンプリング」されると、そのカバレッジスコアが低下し、その特徴が存在する化合物がサブセット内に選択される可能性が低くなる。一つの化合物が選択されると（すなわち、選択されたサブセットに一つの化合物が含まれる）、その化合物に存在する特徴のカバレッジスコアが変化し、マイナスになる可能性がある。この意味で、選択した化合物に存在する特徴は、集団の化合物に存在する他の特徴と比較して「オーバーサンプリング」されていると見なすことができるため、これらの特徴を再度(他の特徴に優先して)選択すると、構造活性関係のコンテキストでの情報内容抽出には最適ではない可能性があるため、カバレッジスコアが低下する。第１の化合物の選択は、任意の適切な方法で、例えば最高のカバレッジスコアによって実行されてもよく、これは、第１の化合物の選択に関して、集団全体の最高のシャノンスコアと同等であり得、化合物のカバレッジスコアがその特徴のカバレッジスコアの合計として決定され得る。

【０１４１】

集団レベルでの化合物のシャノンスコアは反復間で静的または一定であるが、化合物のカバレッジスコアは各特徴がサンプリングされた回数に応じて反復間で動的または変動する。具体的には、他の特徴と比較して「過剰にサンプリングされた」特徴を持つ化合物はカバレッジスコアが低いため、反復ごとに、以前にサンプリングされた化合物を考慮して情報ゲインを最大化する化合物を選択できる。

【０１４２】

多数の化合物がサンプリングされた後、例えば創薬プロジェクトを何度も繰り返した後、または複数の化合物を含む選択されたサブセットの試験後、集団レベルでより高いシャノンスコアを持つ特徴の多くが何度かサンプリングされている可能性がある。そのため、これらの特徴は、集団レベルでシャノンスコアが低いが、この時点までそれほど頻繁にサンプリングされていない可能性がある特徴と比較して、この段階または反復までにカバレッジスコアが低くなる傾向がある可能性がある。特に、セットまたは集団内のレアな特徴、すなわち、集団内の比較的少数の化合物に存在する特徴はより魅力的であり、これは、この段階では比較的高いカバレッジスコアに反映されている可能性があり、それは、それらのレアな特徴を含む化合物が選択される可能性が高くなることを意味している。

【０１４３】

したがって、最も広い意味では、本発明の工程は：集団内にどの化合物があるか；および、トレーニングセットと選択されたサブセットとを含むサンプリングされたセット内にどの化合物があるか：の両方に基づいて、前もってサンプリングされていない、すなわち生物学的活性が知られている化合物のトレーニングセットに含まれていない、集団からの一つ以上の化合物のサブセットを選択することを含む。より具体的には、一つ以上の化合物のサブセットの選択（例えば、特定の反復における）は、サブセットの化合物に存在する構造的特徴が集団の化合物に現れる頻度に、および、それらの構造的特徴がサンプリングされたセットに現れる頻度に基づく。別の言い方をすると、一つ以上の（選択された）化合物のサブセットの選択は、一つ以上の選択された化合物の構造的特徴のそれぞれについて、それぞれの構造的特徴を含む集団内の化合物の数と、それぞれの構造的特徴を含むサンプリングされたセット内の化合物の数の考慮に依存する。一般に、サンプリングされたセット内の比較的多数の化合物にその特徴が存在する化合物が選択される可能性が低くなる場合がある。

【０１４４】

上述したように、化合物のサブセットを上記の考慮事項に応じて選択する一つの方法は、これらの考慮事項を定量化するためにサブセットにスコアを割り当てることである。特に、スコアは、選択された化合物の特徴が集団内で見つかる頻度と、選択された化合物の特徴がサンプリングされたセット内で見つかる頻度とのバランスをとることができる。化合物のサブセットが、サブセットの（サブセット）カバレッジスコアを最大化する目的で選択される場合、サンプリングされたセット内で集団内の有病率と比較して「アンダーサンプリング」されている構造的特徴を持つ化合物であるが、（上で定義したシャノン補正に従って）比較的高レベルの情報内容を提供し、より高い（化合物）カバレッジスコアを持つため、そのような化合物が選択されたサブセットに保持される可能性が高くなる。特定の構造的特徴が現れるサンプリングされたセット内の化合物の数が多ければ多いほど、それらの構造的特徴を含む化合物のスコアは減少する傾向にある。スコアは、正規化された確率を使用して上記の方程式に従って計算されるものとして上で説明されているが、これは、現在説明されている要素と考慮事項、すなわち、抽出された情報内容を最大化しながら集団内の特徴を比例的にサンプリングすることを考慮してスコアがどのように決定されるかの一例にすぎず、他の方程式またはアプローチも使用できることを示していると理解されるであろう。

【０１４５】

本発明の工程によれば、選択されたサブセット内の化合物のカバレッジスコアは選択されたサブセット内の他の化合物に依存するため、選択されたサブセットの（サブセット）カバレッジスコアは、サブセット内の選択された化合物の（化合物）カバレッジスコアに基づいて決定され得る。選択されたサブセットは、決定されたサブセットカバレッジスコアに応じて評価される。サブセットスコアは、選択されたサブセット、集団、およびサンプリングされたセットにおける化合物の構造的特徴の頻度に応じて決定される。本発明の方法は、集団およびサンプリングされたセットにおけるそれぞれの構造的特徴の頻度に応じて、サブセット内の選択された化合物の例えばフィンガープリントである一つ以上の構造的特徴のそれぞれの（特徴）カバレッジスコアを決定することを含むことができ、選択された各化合物の（化合物）カバレッジスコアは、前記選択された化合物の一つ以上の構造的特徴の決定されたスコアに基づくことができる。例えば、選択された各化合物のスコアは、前記選択された化合物の一つ以上の構造的特徴の決定されたスコアの合計として決定され得る。任意に、合計は、例えば、重みが特定の特徴および／または化合物のサイズ（例えば、特徴の数）に基づく場合、特徴スコアの加重合計であってよい。

【０１４６】

評価工程は、一つ以上の化合物の選択されたサブセットが特定の目的に適しているかどうか、例えばサブセット内の化合物の生物学的特性を決定するために合成に供されるかどうか、または化合物の異なるサブセットが選択されるかどうかを評価することを含み得る。例えば、評価工程は、選択されたサブセットの決定されたスコア（例えば、上述のカバレッジスコア）が所定の条件を満たすかどうか、例えばスコアが所定の最小閾値スコアより大きいかどうかを判定することを含み得る。規定の条件が満たされる場合、または評価工程により異なる／更新された化合物のサブセットを選択しないと決定される場合、方法は、サブセット内の選択された化合物を合成して、前記選択された化合物の一つ以上の生物学的特性を決定することを含み得る。次いで、一つ以上の合成された化合物をトレーニングセットに加えることができる。

【０１４７】

選択されたサブセットが初期の選択されたサブセットである場合、方法は、初期の選択されたサブセットとは異なる第２サブセットを集団から選択することを含み得、第２サブセット内の化合物もトレーニングセットには含まれない。選択された第２サブセットのスコアは、最初の第１のサブセットのスコアに対応する方法で決定され（第１と第２サブセットの両方に共通する化合物のスコアは再計算する必要があることに注意されたい）、次に、選択された第２サブセットが評価され得る。例えば、評価は、所定の条件が満たされるかどうかを判定することであってよい。上で示したように、第２サブセットを選択し、

そのスコアを決定する工程は、第１のサブセットに関して所定の条件が満たされない場合にのみ実行され得る。初期のサブセットは集団からランダムに選択することも、任意の適切な代替方法を使用して選択することもできる。

【０１４８】

第１（初期）、第２、およびその後のサブセットの選択は、所望の条件を満たす化合物のサブセットを取得するための反復プロセスの一部であり、したがって、創薬プロジェクトの特定の反復または設計サイクルでの合成に適している可能性がある。このような方法またはプロセスには、停止条件が満たされるまで、一つ以上の化合物の新しいサブセットを（決定されたスコアに基づいて）繰り返し選択および評価することが含まれ得る。選択された新しいサブセットはそれぞれ、前の反復で選択されたサブセットとは異なり、選択された新しいサブセット内の化合物は集団からのものであり、トレーニングセットには含まれていない。停止条件は、新しいサブセットのさらなる選択が実行されないような任意の適切な条件であり得る。例えば、停止条件は、反復プロセスによって最大数の新しいサブセットが選択されたこと、すなわち、最大数の反復が実行されたこととすることができる。あるいは、停止条件は、反復の一つで選択されたサブセットのスコアが所定の条件を満たすこととすることもできる。停止条件は、連続する反復における選択されたサブセットのそれぞれのスコア間の差が所定の差閾値未満であることとすることもできる。停止条件には、これらの条件例の任意の組み合わせが含まれ、および／または他の適切な条件が含まれる可能性がある。次いで、この方法は、停止条件が満たされる反復で選択されたサブセット内の化合物の一部またはすべてを合成して、前記選択された化合物の一つ以上の生物学的特性を決定することを含み得る。

【０１４９】

一般に、創薬プロセスの反復または設計サイクルで試験できる化合物の数には規定の上限があり、これが選択されたサブセットに含まれる化合物の数を知らせる可能性がある。サブセットの選択は、任意の適切な方法で実行することができる。探索戦略の一部として、高いカバレッジスコアを持つサブセットを選択することが望ましい場合がある。ただし、集団全体のカバレッジスコアを最適化するサブセットを決定するのは難しい場合がある。これは、サブセット内の個々の化合物のカバレッジスコアがサブセット内の他の化合物に依存するためであり、また、集団からサブセットを形成することが一般的に可能である化合物の異なる組み合わせが膨大にあるためでもある。純粋に説明のための例として、サブセットには約１０、２０または３０個の化合物が含まれる場合があるが、サブセットには、集団からの任意の適切な数の化合物を含めることができることが理解されるであろう。独自のサブセットの数は、サブセットサイズや集団サイズの増加に伴って指数関数的に増加するため、考えられるすべてのサブセットを列挙し、最良の（最高スコアの）サブセットを選択することが常に可能であるとは限らない。

【０１５０】

一つのオプションは、化合物の一つ以上の初期のサブセットを生成または選択し、カバレッジスコアを向上させることを目的として、それぞれのサブセットを一つ以上の化合物で置き換えることによってこれらを変更することである。例えば、一つ以上の初期のサブセットが選択される場合、この方法の評価工程は、選択されたサブセットのいずれかのスコアが所定の条件を満たすかどうかを判定することを含み得る。所定の条件は、スコアが所定の最小閾値スコアより大きいこととすることができる。所定の条件が満たされる場合、方法は、所定の条件を満たす選択されたサブセット内の化合物を合成して、前記化合物の一つ以上の生物学的特性を決定することを含み得る。次に、合成された化合物をトレーニングセットに加えることができる。このプロセスは、遺伝的アルゴリズムを使用して実行される場合があるが、これは、すべてのオプションの完全なスキャンが不可能または実行可能な場合に、最適に近い解決を見つける良い方法である。

【０１５１】

所定の条件が最初に選択された一つ以上のサブセットによって満たされない場合、一つ以上の第２サブセットが選択され、それらが所定の条件を満たすかどうかを確認すること

ができる。実際、所望のサブセットが得られるまで、一つ以上のサブセットを反復的に生成することができる。特に、複数のサブセットは、例えばランダムに、または進化的もしくは遺伝的アルゴリズムを使用して、最初に（並行して）生成され得る。任意の適切な数のサブセット、例えば１００未満、５０未満、または１０未満が生成され得ることが理解されるであろう。次いで、これらの生成されたサブセットのそれぞれのカバレッジスコアが決定され、最も高い決定されたスコアを有する一つ以上のサブセットを反復して、そのスコアをさらに増加させることができる。この段階では、特定の化合物が複数のサブセットのうちの一つ以上に含まれる可能性があることに注意されたい。また、サブセット内の化合物のカバレッジスコアはサブセット内の他の化合物に依存するため、サブセットのカバレッジスコアを最大化するために反復プロセス中にサブセット内の一つ以上の化合物が置換される場合、カバレッジスコアはサブセット内の残りの化合物の割合は反復ごとに変化するため、サブセットのスコアを決定するには反復ごとに再計算する必要があることにも注意されたい。すなわち、サブセット内で高スコアの化合物が選択されている場合、類似の化合物がトレーニングセットまたはサブセット(すなわち、サンプリングされたセット)に追加されると、そのスコアは低下するが、これは、類似した化合物には共通の特徴があり、そのためそれらの特徴がより多くサンプリングされるため、値が減少するからである。そのため、サブセット内の一つの化合物をより高いカバレッジスコアを持つ別の化合物に置き換えるだけで、サブセットの全体的なカバレッジスコアが必ずしも増加するとは限らない。したがって、最適化する必要があるのは、個々の化合物のスコアではなく、サブセットのスコアである。例えば、遺伝的アルゴリズムを使用して、このようにサブセットを最適化できる。

【０１５２】

サブセットのそのような反復は、任意の適切な方法で実行することができる。例えば、サブセット内の一つ以上の化合物を、トレーニングセットに含まれない集団からの一つ以上の新しい化合物と置き換えるか、置換することができる。一例では、方法は、初期の選択されたサブセット内の複数の化合物のそれぞれの決定されたスコアに基づいて、初期の選択されたサブセットから置換される一つ以上の化合物を特定することを含み得る。任意に、初期の選択されたサブセット内の最も低い決定スコアを有する一つ以上の化合物が、置換のために特定される。

【０１５３】

創薬プロジェクトの特定の設計サイクルでは、停止条件が満たされるまでサブセットが反復される場合があり、その停止条件は、上記の条件の一つであり得る。カバレッジスコアを最大化するときに反復ごとに複数のサブセットが生成される場合、方法は、複数の選択されたサブセットのそれぞれの決定されたスコアに基づいて、停止条件が満たされる反復で複数の選択されたサブセットのうちの一つを特定することを含むことができる。次いで、特定された化合物のサブセットを合成に供することができる。特定されたサブセットは、停止条件が満たされる反復において複数のサブセットの中で最も高いスコアを有するサブセットとなるように選択され得る。

【０１５４】

カバレッジスコアが最適化された化合物のサブセットの取得、すなわち純粋に探索的な戦略について説明した。ただし、サブセットの選択にある程度の活用（エクスプロイテーション）を組み込むことは可能である。これを行うために、化合物の活性が予測される活性モデルが定義される。例えば、ベイズジアンモデルまたは回帰モデルをこの目的に使用できる。化合物の活性は、最大阻害濃度の半分（ＩＣ５０）を参照して定義できる。例えば、化合物は、そのＩＣ５０値が閾値活性レベルを上回るか下回るかに応じて、単純に活性または不活性として分類される場合がある。あるいは、最も高いＩＣ５０値を有するセットからの所定数の化合物を活性として分類し、残りを不活性として分類することもできる。次いで、活性モデルからの各サブセット内の各化合物の活性スコア、例えばベイジアンモデルスコアに基づいて選択されたサブセットの活性スコアが、選択されたサブセットの（サブセット）カバレッジスコアに対してバランスを取って、探索と活用とが所望の組

み合わせでバランスのとれたサブセットを取得する。ここでも、進化的アルゴリズムまたは遺伝的アルゴリズムを使用して、探索と活用の望ましい組み合わせに従ってサブセットを最適化することができる。特に、複数のサブセットが並行して生成される場合、探索と活用の異なるバランスに従って、特定の設計サイクルで個々のサブセットを最適化できる。所定の設計サイクルで進化的アルゴリズムまたは遺伝的アルゴリズムを十分に反復すると、探索的重み付けが最も高いサブセットから活用的重み付けが最も高いサブセットまで、最適化されたサブセットのパレートフロントが出現する。次いで、活用（カバレッジスコアを犠牲にしてより高いモデルスコア）と探索（モデルスコアを犠牲にしてより高いカバレッジスコア）の所望のバランスを有する特定のサブセットを、例えば合成のために必要に応じて選択することができる。

【０１５５】

化合物のトレーニングセットは、ターゲットと比較して望ましい特性を示す可能性が高い集団内の化合物を予測または決定するために使用されるマシンラーニング（ＭＬ）モデルをトレーニングするために使用される。特に、本発明は、集団内の化合物の一つ以上の生物学的特性をそれらの化合物の一つ以上の構造的特徴の関数として近似するためのマシンラーニングモデルを定義することを含み得る。ＭＬモデルは、ベイズ最適化モデル、回帰モデル、クラスタリングモデル、デシジョンツリーモデル、ランダムフォレストモデル、ニューラルネットワークモデル、またはその他の適切なタイプのＭＬモデルであってよい。次に、ＭＬモデルは、化合物のトレーニングセットを使用してトレーニングできる。創薬プロジェクトの反復または設計サイクルごとに、トレーニングセット内の化合物の数が増加する。ＭＬモデルのトレーニング段階は、一つ以上の化合物がトレーニングセットに追加されるたびに実行できる。トレーニングセット内の化合物の数が増加するにつれて、より適切にトレーニングされたモデル、すなわち、どの化合物が特定のプロジェクトに必要な所望の特性、例えば、高い活性レベルを有するかをより正確に予測できるモデルが得られる可能性がある。具体的には、ＭＬモデルのトレーニングに使用されるトレーニングセットに追加される化合物の少なくとも一部が上記の探索方法を使用して選択されている場合、より短い時間でトレーニングされる、および／またはより高い精度の予測を提供するＭＬモデルが取得される。ＭＬモデルは、一つ以上の所望の生物学的特性を有する集団中の一つ以上の化合物を予測するために実行され得る。ＭＬモデルは、各設計サイクルまたは反復後に実行することも、モデルが特定のレベルまでトレーニングされた後にのみ実行することもできる。プロジェクトの特定の設計サイクルでは、特定の反復で合成および試験される一つ以上の化合物が、活用戦略の一部としてＭＬモデルによって選択される場合があることに注意されたい。例えば、プロジェクトの初期の段階、すなわち反復が比較的少ない段階では、この時点ではモデルが特に十分にトレーニングされていない可能性があるため、合成用に選択された化合物のサブセットの一部のみがＭＬモデルを使用して選択されることがあり、サブセットの残りの部分は、ＭＬモデルを改善するために、上記の探索戦略によって選択される。ただし、プロジェクトの後の段階でＭＬモデルがより良いレベルにトレーニングされると、合成用のサブセット内の化合物の大部分またはすべてがＭＬモデルによって選択される場合がある。次いで、これらの化合物を合成して、所望の生物学的活性、生理学的活性、または薬理学的活性を有する候補薬剤化合物を提供することができる。

【０１５６】

図６は、本発明によるコンピュータによる薬剤設計方法６０の工程を要約したものである。工程６１では、複数の化合物の集団５０が定義され、各化合物は、例えばフィンガープリント特徴として記述される一つ以上の構造特徴を有する。工程６２では、集団５０からの化合物のトレーニングセット５１が定義され、トレーニングセット５１内の化合物の一つ以上の生物学的特性が既知である。工程６３では、一つ以上の化合物のサブセット５２が集団５０から選択されるが、サブセット５２内の一つ以上の化合物はまだトレーニングセットに含まれない。工程６４では、選択されたサブセット５２のカバレッジスコアが、選択されたサブセット５２内の一つ以上の化合物の構造的特徴に応じて決定され、選択

されたサブセット５２は、決定されたサブセットスコアに基づいて評価または分析される。サブセットスコアは、集団５０内の各構造的特徴の頻度と、トレーニングセット５１および選択されたサブセット５２を含むサンプリングされたセット５１、５２内の各構造的特徴の頻度とに応じて決定される。化合物のサブセットの選択および評価は、例えば所定の条件が満たされるまで、例えば選択されたサブセットが十分に高いスコアを有するまで、反復プロセスの一部であってよい。

【０１５７】

本発明の方法は、例えば一つ以上のコンピュータプロセッサ上に実装された一つ以上の機能ユニットまたはモジュールによって、任意の適切なコンピューティングデバイス上で実行することができる。このような機能ユニットは、従来のまたは顧客のプロセッサおよびメモリを使用する任意の適切なコンピューティング基板上で実行される適切なソフトウェアによって提供され得る。一つ以上の機能ユニットは、共通のコンピューティング基板（例えば、同じサーバ上で実行することができる）または別個の基板を使用することができ、または一方または両方自体が複数のコンピューティングデバイス間で分散されることもできる。コンピュータメモリは、この方法を実行するための命令を記憶することができ、プロセッサは、記憶された命令を実行して、この方法を実行することができる。

【０１５８】

添付の条項および特許請求の範囲を特に参照して本明細書で定義される本発明の精神および範囲から逸脱することなく、上述の例に対して多くの修正を加えることができる。

【０１５９】

本発明の例は、創薬プロジェクトの一部として標的に対して最適化された化合物または分子を特定するためのより効率的な方法を提供するという点で有利である。特に、本発明は、集団またはセット内の最も代表的な分子を特定するための改良された技術を提供し、したがって、この技術は、特定のプロジェクトの特定の望ましい特性を示す集団内の一つ以上の分子を予測するために使用されるマシンラーニングモデルをトレーニングするのに最適である。本発明は、情報理論を有利に使用して、分子の集団について最大量の情報を提供する構造的特徴を有する分子を選択する。集団内での蔓延と比較して「過剰サンプリング」されていないが、比較的高レベルの情報内容を提供する特徴を持つ分子に焦点を当てることで、例えば、特定の特徴が寄与しているか、または特定の分子が示す一つ以上の望ましい特性を伴うかを判断することが容易になる。したがって、本発明の例は、有利なことに、分子のサブセットの情報またはシャノンエントロピーを最大化することと、それらの分子の特徴がすでに選択／試験されている頻度、すなわち同じ質問を繰り返し質問しないこと、および各分子がどれだけの特徴を持っているか、すなわち、同時に多くの質問をしないこととの間のトレードオフとみなすことができる。すなわち、本発明は、分子のどの特徴が重要であるかを特定するアプローチを提供するが、これは同じ（良い）質問を複数回するのと同じであるため、それらの特徴をあまりサンプリングしない。したがって、本発明によって説明されるアプローチの使用は、臨床候補分子を取得するために必要な反復数または設計サイクルの数を有利に減少させることができ、それによって時間および／またはコストを節約することができる。本発明の方法はまた、トレーニングセットを生成し、一つ以上の適切な臨床候補に到達するために選択、合成、および試験しなければならない化合物の数を減らすこともできる。このように、本発明の方法は、薬剤を最適化するためにアクティブラーニングまたはマシンラーニングを使用する。

【０１６０】

探索戦略を実施するための他のいくつかのアプローチとは異なり、本発明は、何らかの距離測定基準に基づいて多様な範囲の分子を選択しようとするために、（不均等な）化学空間内で「類似した」分子をクラスタリングすることに依存しない。その代わりに、本発明は、分子の選択されたサブセットによって提供される情報の範囲を最適化するための基準を有利に提供する、すなわち、尋ねるべき最良の質問を特定するためのメカニズムを提供する。いくつかのクラスタリングアプローチは集団から外れ値を選択するが、本発明の例は化学系列内の差異を試験することに焦点を当てる。本発明はまた、記載されたアプロ

ーチが分子の任意の集団またはグループに適用可能であり、可変のＲグループを可能にし、例えば、他の既知のアプローチの場合に当てはまり得る、静的なコアを保持しながら分子を修飾することに限定されないという点でも有利である。

【０１６１】

上記の説明では、化合物の集団の化学空間にわたる構造的特徴（フラグメント）について抽出された情報の範囲を示す指標（すなわち、「範囲スコア」）が定義されている。プロジェクトの探索フェーズでは、集団全体を広範囲にカバーする情報を抽出することは、例えば、構造活性相関を記述または予測するＭＬモデルを十分にトレーニングするための時間またはデータポイントの数を削減するのに有益である。本発明の別の例では、化合物の構造的特徴以外の化合物の集団の特徴またはパラメータに関して抽出された情報の範囲の指標を取得することが望まれる場合がある。上記の例は、集団に存在する構造的特徴（フラグメント）へのカバレッジスコアメトリクスの適用に焦点を当てているが、カバレッジスコアメトリクスは、集団内の化合物の化学的または物理的特性などの他の分子特性に代替または追加で適用することもできる。特に、カバレッジスコアメトリクスは、集団内の化合物のさまざまな異なる分子特性と関連付けて使用されて、上記の分子特性を備えている集団内の化合物のそのような分子特性と活性との間に存在し得る関係を説明するためのより優れたＭＬモデルを構築することができる。

【０１６２】

一例では、化合物が標的分子と結合するか、そうでなければ標的分子と相互作用するときに、集団内の化合物が示す、または示すことが予想される／予測される相互作用のタイプに関する適用範囲情報を決定することが望ましい場合がある。プロジェクトの探索フェーズにおいて、同じ構造的特徴を持つ集団から多すぎる化合物をサンプリングすることは望ましくないという上記の例に対応して、この例で、集団から多すぎる化合物をサンプリングして標的分子と同じ相互作用が起こすことは、プロジェクトの探索フェーズにおいて望ましくない。広範囲の結合相互作用を示す化合物をサンプリングすることは、例えば相互作用と活性の関係を記述または予測するＭＬモデルを十分にトレーニングするための時間またはデータポイントの数を削減するのに有益である可能性がある。

【０１６３】

上述のカバレッジスコアアプローチを集団の特定の分子特性に適用するためには、分子特性を分析のために適切な形式で表す必要がある場合がある。上述の例では、集団内の化合物の構造的特徴は、それぞれのフィンガープリント、すなわち数値のリストまたは１次元ベクトルとして表される。特に、上記の例では、各化合物は２進数のリストとして表され、リストの各エントリの１または０は、それぞれの化合物における特定の構造的特徴（フラグメント）の有無を示す。さまざまなタイプの相互作用が検討対象の集団の分子特性である例では、以下で説明するように、この情報も同様に、個々の化合物ベースでフィンガープリント形式で表すことができ、これにより、上述の化合物の構造的特徴に関する例に対応する方法でカバレッジスコアメトリクスを適用できるようになる。

【０１６４】

本開示のカバレッジスコアアプローチを使用して集団内の異なるタイプの相互作用を分析するために、集団内の異なる化合物によって示される異なるタイプの相互作用を示す相互作用データを最初に取得する必要がある。このようなデータを取得するための一つのアプローチには、分子ドッキングプロセスの適用が含まれる。ドッキングは、結合ポケット内のリガンド配置の正確なモデリングを提供することを目的として、標的の結合部位内のリガンドの立体構造を予測する方法である。別の言い方をすれば、ドッキングは、化合物または分子が互いに結合して安定な複合体を形成するときの、別の（標的）分子に対する化合物または分子の好ましい配向および立体構造の予測を提供する。したがって、ドッキングは、リガンドとタンパク質の両方が柔軟な、特定の標的タンパク質に結合するリガンドの「最適な状態」を記述する最適化問題と見なすことができる。場合によっては、相互作用データの一部またはすべてが異なる方法で取得される場合がある。例えば、特定の化合物の相互作用データは実験結果や他のソースから入手できる場合がある。

【０１６５】
ドッキングを実行するために、集団内の異なる化合物がどのように標的の結合ポケットに収容されるかをシミュレートするために、標的の三次元表現または記述が生成され得る。集団内の化合物ごとに、リガンドとタンパク質のペアの配向と立体構造のスナップショットにそれぞれ対応する多数のドッキングポーズが生成される。特定のポーズが好ましい結合相互作用を表すかどうかの可能性を決定するために、ポーズをスコアリングすることができる。ドッキングプロセスの一部としてドッキングポーズを生成しスコアリングするためのさまざまな方法が当技術分野で知られている。ドッキングされた化合物は、受容体の基準系内に３次元座標を持っている可能性がある。
【０１６６】
したがって、異なるリガンド－タンパク質複合体の三次元結合相互作用情報は、ドッキングプロセスから取得できる可能性がある。次に、この３次元情報を１次元のバイナリ文字列、すなわちフィンガープリントに変換して、カバレッジスコア法を適用できるようにする。これらのフィンガープリントは、相互作用フィンガープリント、またはタンパク質－リガンド相互作用フィンガープリント（ＰＬＩＦ）と呼ばれる場合がある。上述の分子フィンガープリントに対応する方法で、相互作用フィンガープリントの各ビットは、実施されている特定の創薬プロジェクトにおいて関連する化合物が対象となる所定の標的分子と結合するときの、特異的結合相互作用の有無を表すことができる。標的分子、例えばタンパク質は、対象となる特定の疾患に本質的に関連していると特定され、治療効果を生み出すために薬剤、例えば集団からの化合物の標的となり得る体内の分子であり得る。したがって、相互作用フィンガープリントは、特定の化合物が受容体とどのような相互作用を行い、どのような残基と相互作用したかを記述する方法である。
【０１６７】
相互作用フィンガープリントは、集団内の化合物が所定の標的分子に結合するときに示される可能性のある特定の相互作用の所望の数および組み合わせを含むように定義することができる。フィンガープリントに含まれる特定の相互作用には、水素結合相互作用、弱い水素結合相互作用、イオン相互作用、疎水性相互作用、面と面との芳香族相互作用、端と面との芳香族相互作用、π－カチオン相互作用、および金属錯体形成相互作用のうちの一つ以上が含まれ得る。
【０１６８】
相互作用フィンガープリントが生成されると、構造的特徴について上述したアプローチに対応する方法でカバレッジスコア選択を使用して、相互作用の多様なセット、すなわちフィンガープリントの多様なセットを有する集団から多数の化合物を選択することができる。上述の例における「特徴スコア」の計算は、本例では「相互作用」スコアと呼ばれることがある（また、一般に、分子特性スコアと呼ばれることもある）ことに留意されたい。
【０１６９】
図７は、カバレッジスコアが集団内の構造的特徴と相互作用のタイプに近づく場合に選択される集団内のさまざまな化合物の例を示している。特に、図７は、２２５８個の化合物の集団からカバレッジスコアを使用して２０個の化合物のサブセットが選択される例を示している。図７（ａ）は、相互作用空間における化合物の集団と、選択された二つのサブセット（一つはＰＬＩＦに基づくもの、もう一つはＥＣＦＰ４フィンガープリントに基づくもの）を示している。図７（ｂ）は、図７（ａ）と同じ集団と選択されたサブセットを示しているが、化学構造空間内にある。サブセットは、停止条件が満たされるまでカバレッジスコアの反復アプローチに従って選択された。
【０１７０】
図８（ａ）および８（ｂ）は、図７（ａ）および７（ｂ）と同様にＰＬＩＦに適用された場合のカバレッジスコア選択サブセットの同じ例を示し、それぞれ相互作用空間および化学構造空間にプロットされている。図７とは異なり、図８では、ＰＬＩＦカバレッジスコアの選択が、集団からランダムに選択されたサブセット、およびＰＬＩＦに適用される

代替選択方法、すなわち多様性選択方法と比較される。

【０１７１】

上述したもの以外の化合物の三次元記述を代替的または追加的に使用して、カバレッジスコア選択を適用できるフィンガープリントを生成することができる。例えば、化合物は、フィンガープリントに変換された三次元ファーマコフォアまたは三次元形状として記述される場合がある。したがって、本発明によれば、図６に示すコンピュータ実行方法の工程は、定義された化合物の集団に存在する異なる分子特性の分析に適用できるように一般化することができる。集団内の各化合物は、それに関連する一つ以上の分子特性を持っている。上述のように、これらには、構造的特徴、それぞれの化合物が所定の標的分子に結合するときに示される相互作用のタイプ、または他の適切な分子特性が含まれ得る。相互作用のタイプを考慮する場合、この情報は、それぞれの化合物が所定の標的分子と相互作用するときの予測結合相互作用を取得するために、上記の分子ドッキングプロセスを実行することによって取得する必要がある場合がある。考慮中の特定の分子特性に関係なく、生物学的特性が既知である化合物を含む化合物のトレーニングセットが定義され、トレーニングセットに含まれない化合物のサブセットが集団から選択される。次いで、選択されたサブセットのカバレッジスコアが、選択されたサブセット内の化合物を考慮して特定の分子特性に応じて決定され、決定されたサブセットスコアに基づいて選択されたサブセットが評価される。サブセットスコアは、集団における考慮中の各分子特性の頻度と、トレーニングセットおよび選択されたサブセットを含むサンプリングされたセットにおける考慮中の各分子特性の頻度に応じて決定される。

【０１７２】

特定の化合物のフィンガープリントは、その化合物の複数の分子特性に関する情報を含むように定義できることに注意されたい。例えば、フィンガープリントの最初のビットのセットは、化合物に存在する構造的特徴に関連することがあり、最初のセットの後のビットのセットは、化合物が所定の標的分子に結合するときに化合物によって示される相互作用のタイプに関連することができる。カバレッジスコアの選択は、集団内の化合物のフィンガープリント表現に含まれる情報の一部またはすべてに基づいて行うことができる。

【０１７３】

本開示のさらなる態様および実施形態を、以下の条項に記載する。

【０１７４】

［条項］

（条項１）

各化合物が一つ以上の分子特性を有する複数の化合物の集団を定義する工程、

一つ以上の生物学的特性が既知である前記集団からの化合物のトレーニングセットを定義する工程、

前記トレーニングセットに含まれない前記集団からの一つ以上の化合物のサブセットを選択する工程、および

前記選択されたサブセット内の前記一つ以上の化合物の分子特性に応じて前記選択されたサブセットのサブセットスコアを決定し、当該決定されたサブセットスコアに基づいて前記選択されたサブセットを評価する工程、を含み、

前記サブセットスコアが、前記集団における前記分子特性の頻度と、前記トレーニングセットおよび前記選択されたサブセットを含むサンプリングされたセットにおける前記分子特性の頻度とに応じて決定されるものである、

コンピュータによる薬剤設計のための方法。

【０１７５】

（条項２）

決定工程が、前記化合物の１つ以上の分子特性に応じて前記選択されたサブセットの前記１つ以上の化合物のそれぞれについて化合物スコアを決定することを含み、前記サブセットスコアが、前記選択されたサブセット内の各化合物の前記決定された化合物スコアに基づいて決定される、条項１に記載の方法。

【０１７６】
（条項３）
前記サブセットスコアが、前記選択されたサブセット内の前記化合物の前記それぞれの化合物スコアの合計として決定される、条項２に記載の方法。
【０１７７】
（条項４）
前記選択されたサブセット内の前記化合物の一つの前記化合物スコアを決定することが、前記集団における前記それぞれの分子特性の頻度、および前記サンプリングされたセットにおける前記それぞれの分子特性の頻度に応じて前記化合物の前記一つ以上の分子特性のそれぞれの分子特性スコアを決定することを含み、前記化合物の前記化合物スコアが、前記化合物の前記一つ以上の分子特性の前記決定されたスコアに基づくものである、条項２または条項３に記載の方法。
【０１７８】
（条項５）
前記化合物の前記化合物スコアが、前記化合物の前記一つ以上の分子特性の前記決定された分子特性スコアの合計として決定される、条項４に記載の方法。
【０１７９】
（条項６）
前記一つ以上の分子特性のそれぞれの前記分子特性スコアが、前記サンプリングされたセット内にある前記分子特性の正規化された確率に応じて決定され、前記正規化された確率が、前記集団および前記サンプリングされたセットにおける前記分子特性の頻度に応じて決定される、条項４または条項５に記載の方法。
【０１８０】
（条項７）
前記正規化された確率が、前記集団内の化合物の数に対する、前記サンプリングされたセット内の化合物の数に応じて決定される、条項６に記載の方法。
【０１８１】
（条項８）
前記正規化された確率が、ラプラシアン補正された正規化確率である、条項７に記載の方法。
【０１８２】
（条項９）
前記ラプラシアン補正された正規化確率Ｐｃｏｒｒが、次式で与えられる、条項８に記載の方法：
【０１８３】
【数１５】

$$P_{\mathrm{corr}} = \frac{F_{\mathrm{sampled}} + 1}{F_{\mathrm{set}} + 1/P_{\mathrm{base}}}$$

【０１８４】
式中、Ｆｓａｍｐｌｅｄは、前記サンプリングされたセットにおける前記分子特性の頻度であり、Ｆｓｅｔは、前記集団における前記分子特性の頻度であり、Ｐｂａｓｅは、前記サンプリングされたセットにおける化合物の数を前記集団における化合物の数で割ったものである。
【０１８５】
（条項１０）
前記一つ以上の分子特性のそれぞれの前記分子特性スコアが、前記サンプリングされたセット内の化合物の数に対する、前記分子特性が存在する前記サンプリングされたセット内の化合物の数に応じて決定される、条項４から９のいずれかに記載の方法。

【０１８６】
（条項１１）
前記分子特性スコアが、前記サンプリングされたセットにおける前記分子特性の正規化されたシャノンエントロピー値に応じて決定される、条項１０に記載の方法。
【０１８７】
（条項１２）
前記正規化されたシャノンエントロピー値が次の式で与えられる、条項１１に記載の方法：
【０１８８】
【数１６】

$$SC = \frac{-f\ln(f) - (1-f)\ln(1-f)}{\ln(2)}$$

【０１８９】
式中、ｆは、前記分子特性が存在する前記サンプリングされたセット内の化合物の数を前記サンプリングされたセット内の化合物の数で割ったものである。
【０１９０】
（条項１３）
前記分子特性スコアＣｏｖ$_{final}$が次式で与えられる、条項１２に記載の方法：
【０１９１】
【数１７】

$$Cov_{\text{final}} = \begin{cases} Cov * SC, & Cov \geq 0 \\ Cov * (2 - SC), & Cov < 0 \text{ and } f > 0.5 \end{cases}$$

【０１９２】
式中、Ｃｏｖは以下の通りである。
【０１９３】
【数１８】

$$Cov = -\ln(P_{\text{corr}}/P_{\text{base}})$$

【０１９４】
（条項１４）
前記サブセットが所定数の化合物を含む、前述の条項のいずれかに記載の方法。
【０１９５】
（条項１５）
前記サブセット内で選択される化合物の数を定義することを含む、条項１４に記載の方法。
【０１９６】
（条項１６）
前記評価する工程は、前記サブセットスコアが所定の条件を満たすかどうかを決定することを含む、前述の条項のいずれかに記載の方法。
【０１９７】
（条項１７）
前記所定の条件は、前記サブセットスコアが所定の最小閾値スコアより大きいことである、条項１６に記載の方法。
【０１９８】
（条項１８）

前記所定の条件が満たされる場合、前記選択されたサブセット内の前記化合物の少なくとも一部を合成して、前記化合物の一つ以上の生物学的特性を決定することを含む、条項１６または条項１７に記載の方法。

【０１９９】

（条項１９）

前記合成された化合物を前記トレーニングセットに加えることを含む、条項１８に記載の方法。

【０２００】

（条項２０）

前記選択されたサブセットが初期の選択されたサブセットであり；

前記トレーニングセットに含まれない前記集団からの一つ以上の化合物を含む、前記初期の選択されたサブセットとは異なる、第２サブセットを選択する工程、および

前記選択された第２サブセットの前記サブセットスコアを決定し、当該決定されたスコアに基づいて前記選択された第２サブセットを評価する工程、を含む、前述の条項のいずれかに記載の方法。

【０２０１】

（条項２１）

前記第２サブセットを選択しそのスコアを決定する工程が、前記所定の条件が満たされない場合に実行される、条項１６に従属する場合の条項２０に記載の方法。

【０２０２】

（条項２２）

前記第２サブセットを選択する工程が、前記初期の選択されたサブセット内の一つ以上の化合物を、前記トレーニングセットに含まれない前記集団からの一つ以上の新しい化合物で置換することを含む、条項２０または条項２１に記載の方法。

【０２０３】

（条項２３）

置換される前記初期の選択されたサブセットから前記一つ以上の化合物を、前記初期の選択されたサブセット内の前記一つ以上の化合物の前記それぞれの決定された化合物スコアに基づいて、特定することを含む、条項２に従属する場合の条項２２に記載の方法。

【０２０４】

（条項２４）

最も低い決定された化合物スコアを有する前記初期の選択されたサブセット内の前記一つ以上の化合物が、置換のために特定される、条項２３に記載の方法。

【０２０５】

（条項２５）

前記トレーニングセットに含まれない前記集団からの一つ以上の化合物を含む、前の反復で選択されたサブセットとは異なる、新しいサブセットを選択する工程と、

前記選択された新しいサブセットの前記サブセットスコアを決定し、当該決定されたスコアに基づいて前記選択された新しいサブセットを評価する工程と、を、停止条件が満たされるまで、反復的に実行することを含む、条項２０から２４のいずれかに記載の方法。

【０２０６】

（条項２６）

前記停止条件が、

最大回数の反復が実行されたこと；

前記反復の一つにおいて選択された前記サブセットの前記サブセットスコアが前記所定の条件を満たすこと；および、

連続する反復における前記選択されたサブセットの前記それぞれのサブセットスコアの間の差が、所定の差の閾値未満であること；のうちの少なくとも一つを含む、条項２５に記載の方法。

【０２０７】

（条項２７）
前記停止条件が満たされる前記反復で前記選択されたサブセットの前記化合物を合成して、前記化合物の一つ以上の生物学的特性を決定することを含む、条項２５または条項２６に記載の方法。
【０２０８】
（条項２８）
各反復において複数の新しいサブセットを選択すること、
前記停止条件が満たされる前記反復において前記複数の選択されたサブセットのうちの一つを、前記それぞれの複数の選択されたサブセットの前記決定されたサブセットスコアに基づいて、特定すること、および
前記一つの特定されたサブセットの前記化合物を合成して、前記化合物の一つ以上の生物学的特性を決定すること、を含む、条項２４から２７のいずれかに記載の方法。
【０２０９】
（条項２９）
前記特定されたサブセットが、前記停止条件が満たされる前記反復において、前記複数のサブセットの中で最も高いサブセットスコアを有するサブセットである、条項２８に記載の方法。
【０２１０】
（条項３０）
前記選択されたサブセットが第１のサブセットであり：
それぞれが前記トレーニングセットに含まれない前記集団からの複数の化合物を含む複数のサブセットを選択すること；
前記サブセットのそれぞれの前記サブセットスコアを決定すること；および
前記それぞれのサブセットの前記決定されたサブセットスコアに基づいて、前記複数のサブセットから前記第１のサブセットを選択すること；を含む、前述の条項のいずれかに記載の方法。
【０２１１】
（条項３１）
前記第１のサブセットが、前記複数のサブセットの中で最も高いサブセットスコアを有するサブセットとなるように選択される、条項３０に記載の方法。
【０２１２】
（条項３２）
前記複数のサブセットがそれぞれ、同じ数の化合物を有する、条項３０または条項３１に記載の方法。
【０２１３】
（条項３３）
前記評価する工程が、前記集団における前記化合物の活性レベルを予測するために、活性モデルから得られる前記選択されたサブセットの活性スコアに基づいて前記選択されたサブセットを評価することを含む、前述の条項のいずれかに記載の方法。
【０２１４】
（条項３４）
前記評価する工程が、前記決定されたサブセットスコアおよび前記活性スコアに基づいて、それらのスコアの所望のバランスに対して、前記選択されたサブセットを評価することを含む、条項３３に記載の方法。
【０２１５】
（条項３５）
前記複数の新しいサブセットがそれぞれ、前記決定されたスコアと前記活性スコアとの間の異なるバランスを含む、条項２８に従属する場合の条項３３または条項３４に記載の方法。
【０２１６】

（条項３６）
前記複数の新しいサブセットが、前記停止条件が満たされる前記反復において、決定されたサブセットおよび活性スコアのパレートフロントを形成する、条項３５に記載の方法。
【０２１７】
（条項３７）
前記トレーニングセットが、最初は空である、前述の条項のいずれかに記載の方法。
【０２１８】
（条項３８）
前記集団中の前記複数の化合物のそれぞれの前記分子特性が、前記化合物の構造的特徴を含む、前述の条項のいずれかに記載の方法。
【０２１９】
（条項３９）
前記集団中の前記複数の化合物のそれぞれの前記構造的特徴が、前記化合物中に存在するフラグメントに対応する、前述の条項のいずれかに記載の方法。
【０２２０】
（条項４０）
前記複数の化合物のそれぞれに存在する前記フラグメントが、分子フィンガープリントとして表される、条項３９に記載の方法。
【０２２１】
（条項４１）
前記分子フィンガープリントが、拡張接続フィンガープリント（ＥＣＦＰ）であり、任意にＥＣＦＰ０、ＥＣＦＰ２、ＥＣＦＰ４、ＥＣＦＰ６、ＥＣＦＰ８、ＥＣＦＰ１０またはＥＣＦＰ１２である、条項４０に記載の方法。
【０２２２】
（条項４２ａ）
前記集団中の前記複数の化合物のそれぞれの前記分子特性が、前記化合物の化学的特性を含む、前述の条項のいずれかに記載の方法。
【０２２３】
（条項４２ｂ）
前記集団中の前記複数の化合物のそれぞれの前記分子特性が、前記化合物の構造的特徴および化学的特性を含む、前述の条項のいずれかに記載の方法。
【０２２４】
（条項４３）
前記化学的特性が、前記それぞれの化合物が所定の標的分子に結合するときに示される相互作用のタイプに対応する、条項４２ａまたは条項４２ｂに記載の方法。
【０２２５】
（条項４４）
前記集団中の前記化合物の少なくとも一部のものの化学的性質が、前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプの予測に対応する、条項４３に記載の方法。
【０２２６】
（条項４５）
前記予測が、前記それぞれの化合物が前記所定の標的分子に結合するときに、一つ以上の所定のタイプの相互作用のうちのどれが示されるかについての予測を含む、条項４４に記載の方法。
【０２２７】
（条項４６）
前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプの予測を取得することを含む、条項４４または条項４５に記載の方法。

【０２２８】
（条項４７）
各化合物についての予測を取得する工程が、
前記化合物の三次元表現を生成すること、および
前記生成された三次元表現を使用してドッキングプロセスを実行して、前記化合物が前記所定の標的分子に結合するときの好ましいドッキングポーズを予測すること、を含んでおり、
相互作用の前記示されるタイプが、前記ドッキングプロセスの結果に基づいて予測される、条項４６に記載の方法。
【０２２９】
（条項４８）
前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプが、相互作用フィンガープリントとして、任意に、タンパク質－リガンド相互作用フィンガープリント（ＰＬＩＦ）として表される、条項４３から４７のいずれかに記載の方法。
【０２３０】
（条項４９）
前記相互作用のタイプとして、水素結合相互作用、弱い水素結合相互作用、イオン相互作用、疎水性相互作用、面と面との芳香族相互作用、端と面との芳香族相互作用、π－カチオン相互作用、および金属錯体形成相互作用のうちの一つ以上が含まれる、条項４３から４８のいずれかに記載の方法。
【０２３１】
（条項５０）
前記集団中の前記化合物のそれぞれがリガンドであり、前記所定の標的分子がタンパク質である、条項４３から４９のいずれかに記載の方法。
【０２３２】
（条項５１）
前記一つ以上の生物学的特性が、活性、選択性、毒性、吸収、分布、代謝、および排出のうちの一つ以上を含む、前述の条項のいずれかに記載の方法。
【０２３３】
（条項５２）
前記生物学的特性の一つ以上が、それぞれの所望の生物学的特性に対して定義される、前述の条項のいずれかに記載の方法。
【０２３４】
（条項５３）
前記集団内の化合物の一つ以上の生物学的特性を、前記化合物の前記一つ以上の分子特性の関数として近似するためのマシンラーニングモデルを定義すること、および
化合物の前記トレーニングセットを使用して前記マシンラーニングモデルをトレーニングすること、を含む、前述の条項のいずれかに記載の方法。
【０２３５】
（条項５４）
一つ以上の化合物が前記トレーニングセットに加えられるたびにトレーニング工程を実行することを含む、条項５３に記載の方法。
【０２３６】
（条項５５）
前記マシンラーニングモデルが、ベイズ最適化モデル、回帰モデル、クラスタリングモデル、デシジョンツリーモデル、ランダムフォレストモデル、およびニューラルネットワークモデルのうちの少なくとも一つである、条項５３または条項５４に記載の方法。
【０２３７】
（条項５６）

トレーニング工程の後に、前記マシンラーニングモデルを実行して、一つ以上の所望の生物学的特性を有する前記集団中の一つ以上の化合物を予測することを含む、条項５３から５５のいずれかに記載の方法。

【０２３８】

（条項５７）

前記一つ以上の予測化合物の少なくとも一つを合成することをさらに含む、条項５６に記載の方法。

【０２３９】

（条項５８）

前記一つ以上の予測化合物が、所定の標的分子に対して所望の生物学的、生化学的、生理学的および／または薬理学的活性を有する候補薬剤または治療分子である、条項５６または条項５７に記載の方法。

【０２４０】

（条項５９）

前記所定の標的分子が、インビトロおよび／またはインビボの治療、診断、または実験アッセイ標的である、条項５８に記載の方法。

【０２４１】

（条項６０）

前記候補薬剤または治療分子が、医学において、例えばヒトまたはヒト以外の動物などの動物の治療方法において、使用されるためのものである、条項５８または条項５９に記載の方法。

【０２４２】

（条項６１）

前項のいずれかの方法によって特定された化合物。

【０２４３】

（条項６２）

コンピュータプロセッサによって実行されるときに、当該コンピュータプロセッサに条項１から６０のいずれかに記載の方法を実行させる命令を記憶する非一時的なコンピュータ可読記憶媒体。

【０２４４】

（条項６３）

各化合物が一つ以上の分子特性を有する複数の化合物の集団を示すデータを受け取り、一つ以上の生物学的特性が既知である前記集団からの化合物のトレーニングセットを示すデータを受け取るように構成されたインプット部；

前記トレーニングセットに含まれない前記集団からの一つ以上の化合物のサブセットを選択し、前記選択されたサブセット内の前記一つ以上の化合物の分子特性に応じて前記選択されたサブセットのサブセットスコアを決定し、および、前記決定されたサブセットスコアに基づいて前記選択されたサブセットを評価するように構成されたプロセッサ；および

前記評価の結果を出力するように構成されたアウトプット部；を含み、

前記サブセットスコアが、前記集団における前記分子特性の頻度と、前記トレーニングセットおよび前記選択されたサブセットを含むサンプリングされたセットにおける前記分子特性の頻度とに応じて決定されるものである、コンピュータによる薬剤設計のためのコンピューティングデバイス。

【０２４５】

（条項６４）

前記プロセッサは、条項１から条項６０のいずれかに記載の方法を実行するように構成されている、条項６３に記載のコンピューティングデバイス。

【０２４６】

（条項６５）

複数の化合物の集団を定義すること、
前記複数の化合物のそれぞれについて、前記それぞれの化合物が所定の標的分子に結合するときに示される相互作用のタイプを示す相互作用データを取得すること、
一つ以上の生物学的特性が既知である前記集団からの化合物のトレーニングセットを定義すること、
前記トレーニングセットに含まれない前記集団からの一つ以上の化合物のサブセットを選択すること、および
前記選択されたサブセット内の前記一つ以上の化合物の前記得られた相互作用データに応じて前記選択されたサブセットのサブセットスコアを決定し、当該決定されたサブセットスコアに基づいて前記選択されたサブセットを評価すること、を含み、
前記サブセットスコアが、前記集団における前記タイプの相互作用の頻度と、前記トレーニングセットおよび前記選択されたサブセットを含むサンプリングされたセットにおける前記タイプの相互作用の頻度とに応じて決定されるものである、薬剤設計のためのコンピュータ実行方法。

【０２４７】

（条項６６）

前記集団内の前記複数の化合物の少なくともいくつかについての前記得られた相互作用データが、前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプの予測を示すデータである、条項６５に記載の方法。

【０２４８】

（条項６７）

前記予測が、前記それぞれの化合物が前記所定の標的分子に結合するときに、一つ以上の所定のタイプの相互作用のうちのどれが示されるかについての予測を含む、条項６６に記載の方法。

【０２４９】

（条項６８）

前記それぞれの化合物が前記所定の標的分子に結合するときに示される相互作用のタイプの予測を取得することを含む、条項６６または条項６７に記載の方法。

【０２５０】

（条項６９）

各化合物についての予測を取得する工程が、
前記化合物の三次元表現を生成すること、および
前記生成された三次元表現を使用してドッキングプロセスを実行して、前記化合物が前記所定の標的分子に結合するときの好ましいドッキングポーズを予測すること、を含み、
前記示されたタイプの相互作用が、前記ドッキングプロセスの結果に基づいて予測されるものである、条項６８に記載の方法。

【０２５１】

（条項７０）

前記それぞれの化合物が前記所定の標的分子に結合するときに示される前記タイプの相互作用が相互作用フィンガープリントとして、任意に、タンパク質－リガンド相互作用フィンガープリント（ＰＬＩＦ）として表される、条項６５から６９のいずれかに記載の方法。

【０２５２】

（条項７１）

前記タイプの相互作用として、水素結合相互作用、弱い水素結合相互作用、イオン相互作用、疎水性相互作用、面と面との芳香族相互作用、端と面との芳香族相互作用、π－カチオン相互作用、および金属錯体形成相互作用のうちの一つ以上が含まれる、条項６５から７０のいずれかに記載の方法。

【０２５３】

（条項７２）
前記集団中の前記化合物のそれぞれがリガンドであり、前記所定の標的分子がタンパク質である、条項６５から７１のいずれかに記載の方法。
【０２５４】
（条項７３）
前記決定する工程が、前記化合物の前記相互作用データにおける一つ以上のタイプの相互作用に応じて、前記選択されたサブセットの前記一つ以上の化合物のそれぞれについて化合物スコアを決定することを含み、前記サブセットスコアが、前記選択されたサブセット内の各化合物の前記決定された化合物スコアに基づいて決定される、条項６５から７２のいずれかに記載の方法。
【０２５５】
（条項７４）
前記サブセットスコアが、前記選択されたサブセット内の前記化合物の前記それぞれの化合物スコアの合計として決定される、条項７３に記載の方法。
【０２５６】
（条項７５）
前記選択されたサブセット内の前記化合物の一つの前記化合物スコアを決定することが、前記化合物の前記相互作用データにおける前記一つ以上のタイプの相互作用のそれぞれの相互作用スコアを、前記集団における前記それぞれのタイプの相互作用の頻度と、前記サンプリングされたセットにおける前記それぞれのタイプの相互作用の頻度とに応じて決定することを含み、前記化合物の前記化合物スコアが、前記化合物の前記相互作用データにおける前記一つ以上のタイプの相互作用の前記決定されたスコアに基づくものである、条項７３または条項７４に記載の方法。
【０２５７】
（条項７６）
前記化合物の前記化合物スコアが、前記化合物の前記相互作用データにおける前記一つ以上のタイプの相互作用の前記決定された相互作用スコアの合計として決定される、条項７５に記載の方法。
【０２５８】
（条項７７）
前記相互作用データにおける前記一つ以上のタイプの相互作用のそれぞれの相互作用スコアが、前記サンプリングされたセット内に存在する前記タイプの相互作用の正規化された確率に応じて決定され、前記正規化された確率が、前記集団および前記サンプリングされたセットにおける前記タイプの相互作用の頻度に応じて決定される、条項７５または条項７６に記載の方法。
【０２５９】
（条項７８）
前記正規化された確率が、前記集団内の化合物の数に対する、前記サンプリングされたセット内の化合物の数に応じて決定される、条項７７に記載の方法。
【０２６０】
（条項７９）
前記正規化された確率が、ラプラシアン補正された正規化された確率である、条項７８に記載の方法。
【０２６１】
（条項８０）
前記正規化確率P_{corr}が、次式で与えられる、条項７９に記載の方法：
【０２６２】
【数１９】

$$P_{\mathrm{corr}} = \frac{F_{\mathrm{sampled}} + 1}{F_{\mathrm{set}} + 1/P_{\mathrm{base}}}$$

【０２６３】
式中、Ｆｓａｍｐｌｅｄは前記サンプリングされたセット内の前記タイプの相互作用の頻度、Ｆｓｅｔは前記集団内の前記タイプの相互作用の頻度、Ｐｂａｓｅは前記サンプリングされたセット内の化合物の数を前記集団内の化合物の数で割ったものである。
【０２６４】
（条項８１）
前記相互作用データにおける前記一つ以上のタイプの相互作用のそれぞれの前記相互作用スコアが、前記サンプリングされたセット内の化合物の数に対する、前記タイプの相互作用が存在する前記サンプリングされたセット内の化合物の数に応じて決定される、条項７５から８０のいずれかに記載の方法。
【０２６５】
（条項８２）
前記相互作用スコアが、前記サンプリングされたセットにおける前記タイプの相互作用の正規化されたシャノンエントロピー値に応じて決定される、条項８１に記載の方法。
【０２６６】
（条項８３）
前記正規化されたシャノンエントロピー値が、次の式で与えられる、条項８２に記載の方法：
【０２６７】
【数２０】

$$SC = \frac{-f\ln(f) - (1-f)\ln(1-f)}{\ln(2)}$$

【０２６８】
式中、ｆは、前記分子特性が存在する前記サンプリングされたセット内の化合物の数を前記サンプリングされたセット内の化合物の数で割ったものである。
【０２６９】
（条項８４）
前記相互作用スコアＣｏｖｆｉｎａｌが次の式で与えられる、条項８３に記載の方法：
【０２７０】
【数２１】

$$Cov_{\mathrm{final}} = \begin{cases} Cov * SC, & Cov \geq 0 \\ Cov * (2 - SC), & Cov < 0 \text{ and } f > 0.5 \end{cases}$$

【０２７１】
式中、Ｃｏｖは以下の通りである。
【０２７２】
【数２２】

$$Cov = -\ln(P_{\mathrm{corr}}/P_{\mathrm{base}})$$

【０２７３】
（条項８５）
前記サブセットが所定数の化合物を含む、条項６５から８４のいずれかに記載の方法。

【０２７４】
（条項８６）
前記サブセット内で選択される化合物の数を定義することを含む、条項８５に記載の方法。
【０２７５】
（条項８７）
前記評価する工程が、前記サブセットスコアが所定の条件を満たすかどうかを判定することを含む、条項６５から８６のいずれかに記載の方法。
【０２７６】
（条項８８）
前記所定の条件は、前記サブセットスコアが所定の最小閾値スコアよりも大きいことである、条項８７に記載の方法。
【０２７７】
（条項８９）
前記所定の条件が満たされる場合、前記選択されたサブセット内の前記化合物の少なくとも一部を合成して、前記化合物の一つ以上の生物学的特性を決定することを含む、条項８７または条項８８に記載の方法。
【０２７８】
（条項９０）
前記合成された化合物を前記トレーニングセットに加えることを含む、条項８９に記載の方法。
【０２７９】
（条項９１）
前記選択されたサブセットが初期の選択されたサブセットであり：
前記トレーニングセットに含まれない前記集団からの一つ以上の化合物を含む、前記初期の選択されたサブセットとは異なる、第２サブセットを選択すること；および
前記選択された第２サブセットの前記サブセットスコアを決定し、前記決定されたスコアに基づいて前記選択された第２サブセットを評価すること；を含む、条項６５から９０のいずれかに記載の方法。
【０２８０】
（条項９２）
前記所定の条件が満たされない場合、前記第２サブセットを選択し、そのスコアを決定する工程が実行される、条項８７に従属する場合の条項９１に記載の方法。
【０２８１】
（条項９３）
前記第２サブセットを選択することが、前記初期の選択されたサブセット内の一つ以上の化合物を、前記トレーニングセットに含まれない前記集団からの一つ以上の新しい化合物で置換することを含む、条項９１または条項９２に記載の方法。
【０２８２】
（条項９４）
前記初期の選択されたサブセット内の前記一つ以上の化合物の前記それぞれ決定された化合物スコアに基づいて、置換される前記初期の選択されたサブセットから前記一つ以上の化合物を特定することを含む、条項７３に従属する場合の条項９３に記載の方法。
【０２８３】
（条項９５）
最も低い決定された化合物スコアを有する前記初期の選択されたサブセット内の前記一つ以上の化合物が置換のために特定される、条項９４に記載の方法。
【０２８４】
（条項９６）
前記トレーニングセットに含まれない前記集団からの一つ以上の化合物を含む、前の反

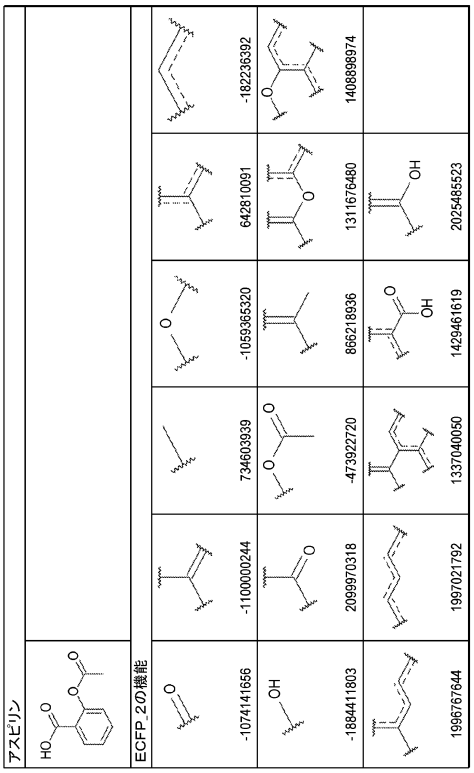
復で選択されたサブセットとは異なる新しいサブセットを選択する工程と、
前記選択された新しいサブセットの前記サブセットスコアを決定し、当該決定されたスコアに基づいて前記選択された新しいサブセットを評価する工程と、を停止条件が満たされるまで、反復的に実行することを含む、条項９１から９５のいずれかに記載の方法。
【０２８５】
（条項９７）
前記停止条件が：
最大回数の反復が実行されたこと；反復の一つで選択された前記サブセットの前記サブセットスコアが前記所定の条件を満たすこと；および、
連続する反復における前記選択されたサブセットの前記それぞれのサブセットスコアの間の差が所定の差の閾値未満であること；のうちの少なくとも一つを含む、条項９６に記載の方法。
【０２８６】
（条項９８）
前記停止条件が満たされる反復において前記選択されたサブセットの前記化合物を合成して、前記化合物の一つ以上の生物学的特性を決定することを含む、条項９６または条項９７に記載の方法。
【０２８７】
（条項９９）
各反復において複数の新しいサブセットを選択すること；前記それぞれの複数の選択されたサブセットの前記決定されたサブセットスコアに基づいて、前記停止条件が満たされる反復において前記複数の選択されたサブセットのうちの一つを特定すること；および、前記一つの特定されたサブセットの前記化合物を合成して、前記化合物の一つ以上の生物学的特性を決定すること；を含んでなる、条項９５から９８のいずれかに記載の方法。
【０２８８】
（条項１００）
前記特定されたサブセットが、前記停止条件が満たされる反復において、前記複数のサブセットの中で最も高いサブセットスコアを有する前記サブセットである、条項９９に記載の方法。
【０２８９】
（条項１０１）
前記選択されたサブセットが第１のサブセットであり：
それぞれが前記トレーニングセットに含まれない前記集団から複数の化合物を含む複数のサブセットを選択すること；
前記サブセットのそれぞれの前記サブセットスコアを決定すること；および、
前記それぞれのサブセットの前記決定されたサブセットスコアに基づいて、前記複数のサブセットから前記第１のサブセットを選択すること；を含む、条項６５から１００のいずれかに記載の方法。
【０２９０】
（条項１０２）
前記第１のサブセットが、前記複数のサブセットの中で最も高いサブセットスコアを有するサブセットとなるように選択される、条項１０１に記載の方法。
【０２９１】
（条項１０３）
前記複数のサブセットが、それぞれ同じ数の化合物を有する、条項１０２または条項１０３に記載の方法。
【０２９２】
（条項１０４）
前記評価する工程が、前記集団における前記化合物の活性レベルを予測するために、活性モデルから得られた前記選択されたサブセットの活性スコアに基づいて、前記選択され

たサブセットを評価することを含む、条項６５から１０３のいずれかに記載の方法。
【０２９３】
（条項１０５）
前記評価する工程が、前記決定されたサブセットスコアおよび前記活性スコアに基づいて、それらのスコアの所望のバランスに対して、前記選択されたサブセットを評価することを含む、条項１０４に記載の方法。
【０２９４】
（条項１０６）
前記複数の新しいサブセットが、それぞれ、前記決定されたスコアと前記活動スコアとの間の異なるバランスを含む、条項１０１に従属する場合の条項１０４または条項１０５に記載の方法。
【０２９５】
（条項１０７）
前記複数の新しいサブセットが、前記停止条件が満たされる反復において、前記決定されたサブセットおよび前記活性スコアのパレートフロントを形成する、条項１０６に記載の方法。
【０２９６】
（条項１０８）
前記トレーニングセットが最初は空である、条項６５から１０７のいずれかに記載の方法。
【０２９７】
（条項１０９）
前記一つ以上の生物学的特性が、活性、選択性、毒性、吸収、分布、代謝および排出のうちの一つ以上を含む、条項６５から１０８のいずれかに記載の方法。
【０２９８】
（条項１１０）
前記生物学的特性の一つ以上が、それぞれの所望の生物学的特性に対して定義される、条項６５から１０９のいずれかに記載の方法。
【０２９９】
（条項１１１）
集団内の化合物の一つ以上の生物学的特性を、前記化合物の得られた相互作用データにおける一つ以上のタイプの相互作用の関数として近似するためのマシンラーニングモデルを定義する工程、および
化合物のトレーニングセットを使用してマシンラーニングモデルをトレーニングする工程、を含む、条項６５から条項１１０のいずれかに記載の方法。
【０３００】
（条項１１２）
前記一つ以上の化合物がトレーニングセットに加えられるたびにトレーニング工程を実行することを含む、条項１１１に記載の方法。
【０３０１】
（条項１１３）
前記マシンラーニングモデルが、ベイズ最適化モデル、回帰モデル、クラスタリングモデル、デシジョンツリーモデル、ランダムフォレストモデルおよびニューラルネットワークモデルのうちの少なくとも一つである、条項１１１または条項１１２に記載の方法。
【０３０２】
（条項１１４）
前記トレーニング工程の後に、マシンラーニングモデルを実行して、一つ以上の所望の生物学的特性を有する集団中の一つ以上の化合物を予測することを含む、条項１１１から１１３のいずれかに記載の方法。
【０３０３】

（条項１１５）
一つ以上の予測化合物の少なくとも一つを合成することをさらに含む、条項１１４に記載の方法。
【０３０４】
（条項１１６）
前記一つ以上の予測化合物が、所定の標的分子に対して所望の生物学的、生化学的、生理学的および／または薬理学的活性を有する候補薬剤または治療分子である、条項１１４または条項１１５に記載の方法。
【０３０５】
（条項１１７）
前記所定の標的分子が、インビトロおよび／またはインビボの治療、診断、または実験アッセイ標的である、条項１１６に記載の方法。
【０３０６】
（条項１１８）
候補薬剤または治療分子が、医学において、例えば、ヒトまたはヒト以外の動物などの動物の治療方法において、使用されるためのものである、条項１１７または条項１１８に記載の方法。
【０３０７】
（条項１１９）
条項６５から条項１１８のいずれかに記載の方法により特定された化合物。
【０３０８】
（条項１２０）
コンピュータプロセッサによって実行されるときに当該コンピュータプロセッサに条項６５から条項１１８のいずれかに記載の方法を実施させる命令を記憶する非一時的なコンピュータ可読記憶媒体。
【０３０９】
（条項１２１）
複数の化合物の集団を示す集団データと、
それぞれの化合物が所定の標的分子に結合するときに示される相互作用のタイプを示す、前記複数の化合物のそれぞれについての、相互作用データと、
一つ以上の生物学的特性が知られている前記集団からの化合物のトレーニングセットを示す、トレーニングセットデータと、を受け取るように構成されたインプット部；
前記トレーニングセットに含まれない前記集団からの一つ以上の化合物のサブセットを選択し、前記選択されたサブセットにおける前記一つ以上の化合物の前記相互作用データの相互作用のタイプに応じて前記選択されたサブセットのサブセットスコアを決定し、前記決定されたサブセットスコアに基づいて前記選択されたサブセットを評価するように構成されたプロセッサ；および
前記評価の結果を出力するように構成されたアウトプット部；を含み、
前記サブセットスコアが、前記集団における前記相互作用データにおける前記タイプの相互作用の頻度と、前記トレーニングセットおよび前記選択されたサブセットを含むサンプリングされたセットにおける前記相互作用データにおける前記タイプの相互作用の頻度とに応じて決定されるものである、コンピュータによる薬剤設計のためのコンピューティングデバイス。
【０３１０】
（条項１２２）
前記プロセッサが、条項６５から条項１１８のいずれかに記載の方法を実行するように構成されている、条項１２１に記載のコンピューティングデバイス。

【図面】

【図１】



【図２】

	芳香環に結合したアミド	セット内の頻度	シャノンH
	フェニル／ピリジル	338/2500	0.396
	エーテル酸素	960/2500	0.666
	ヒドロキシ（カルボン酸の一部）	2500/2500	0.000

【図３】

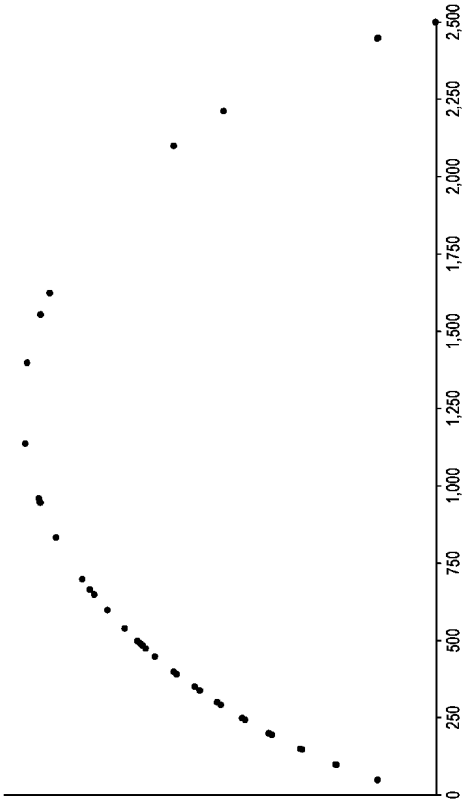
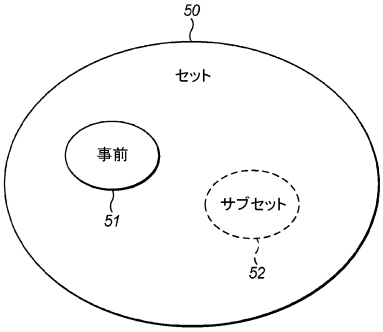


FIG. 3

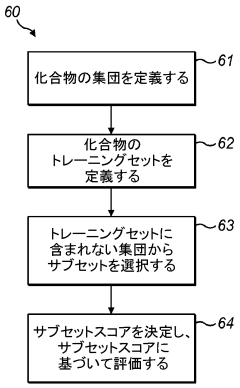
【図４】

ECFP_4	シャノンスコア	セット内の数	選択数
-1939757055	0.69315	300	4
2029508750	0.69315	300	4
2026249478	0.69315	300	3
-196050067	0.69315	300	3
657586427	0.69315	300	3
655739385	0.69092	320	4
200746375	0.69092	320	3
-1331920947	0.68593	264	4
1215378785	0.68593	264	2
-154530762	0.68423	340	4
-1672647522	0.67301	360	4
-1660340418	0.67301	240	4
734603939	0.66241	226	4
1559650422	0.64745	210	2
203547503	0.63651	200	3
-786013480	0.63651	200	2
-167460056	0.63179	404	5

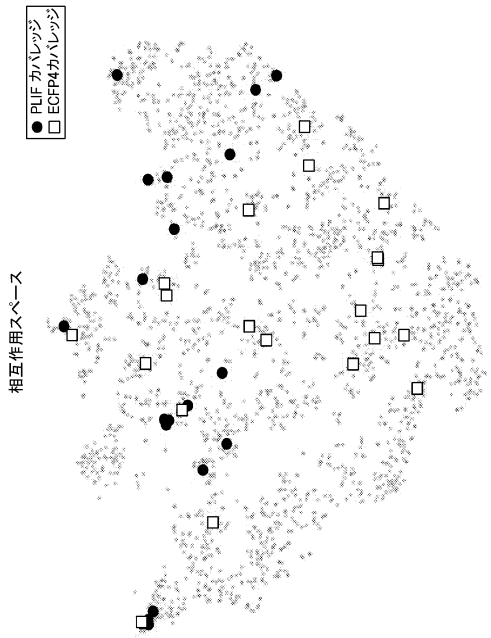
【図５】


50
セット
事前
51
サブセット
52

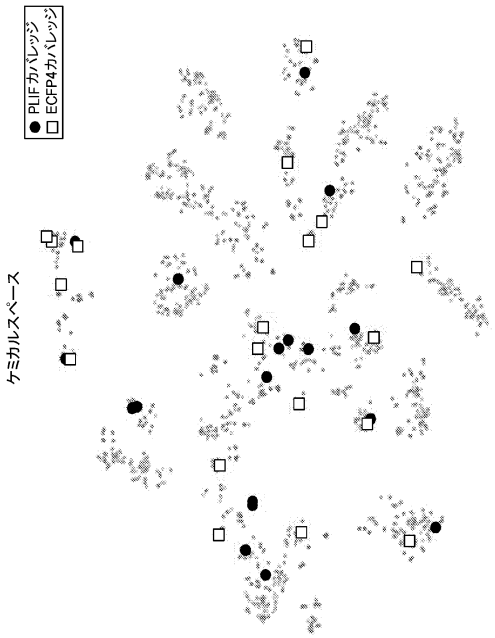
【図６】


60
化合物の集団を定義する
61
化合物の
トレーニングセットを
定義する
62
トレーニングセットに
含まれない集団から
サブセットを選択する
63
サブセットスコアを決定し、
サブセットスコアに
基づいて評価する
64

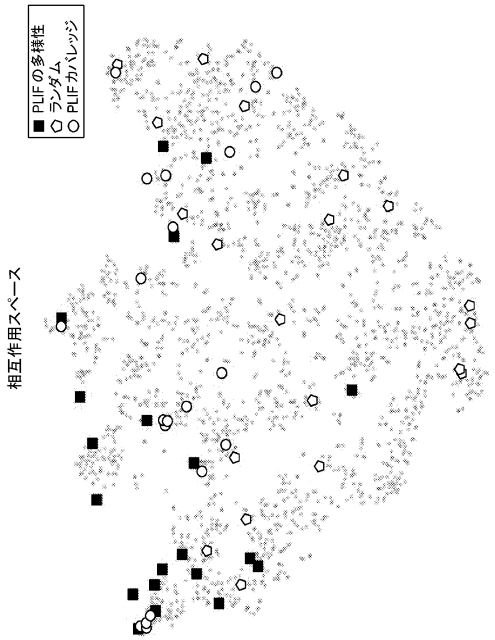
【図７ａ】


PLIF カバレッジ
ECFP4カバレッジ
相互作用スペース

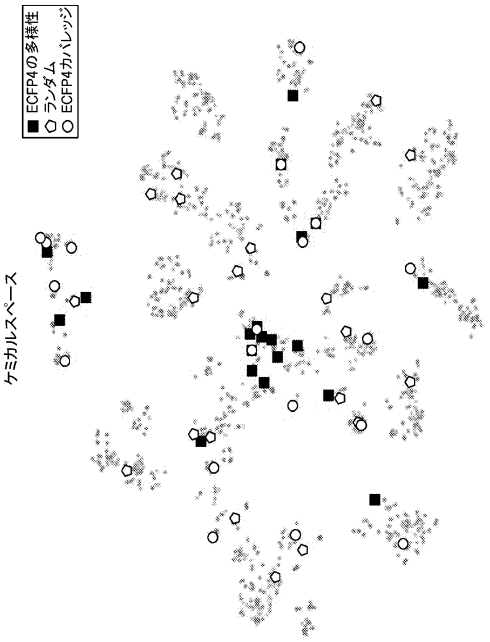
【図７ｂ】


PLIF カバレッジ
ECFP4カバレッジ
ケミカルスペース

【図８ａ】


PLIFの多様性
ランダム
PLIFカバレッジ
相互作用スペース

【図８ｂ】


ECFP4の多様性
ランダム
ECFP4カバレッジ
ケミカルスペース

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2021/052753

A. CLASSIFICATION OF SUBJECT MATTER
INV. G16C20/62
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)
G16C

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	LEONARD J. THOMAS ET AL: "On Selection of Training and Test Sets for the Development of Predictive QSAR models", QSAR & COMBINATORIAL SCIENCE, vol. 25, no. 3, 1 March 2006 (2006-03-01), pages 235-251, XP055892504, ISSN: 1611-020X, DOI: 10.1002/qsar.200510161 abstract ----- WO 2019/186193 A2 (BENEVOLENTAI TECH LIMITED [GB]) 3 October 2019 (2019-10-03) paragraphs [0003], [0004], [0062], [0064], [0065], [0092] ----- -/--	1-65 1-61, 63-65

[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
16 February 2022	28/02/2022

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	Nurmi, Jussi

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

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2021/052753

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WEBER LUTZ: "Current Status of Virtual Combinatorial Library Design", QSAR & COMBINATORIAL SCIENCE, vol. 24, no. 7, 19 September 2005 (2005-09-19), pages 809-823, XP055778966, ISSN: 1611-020X, DOI: 10.1002/qsar.200510120 page 814, column 2, paragraph 4 - page 815, column 1, paragraph 3 page 817, column 1, paragraph 2 - column 2, paragraph 2 -----	1

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

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2021/052753

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2019186193 A2	03-10-2019	CN 112136180 A	25-12-2020
		EP 3776562 A2	17-02-2021
		US 2021027864 A1	28-01-2021
		WO 2019186193 A2	03-10-2019

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

フロントページの続き

,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,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,IR,IS,IT,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

(72)発明者　ファン・ホールン，ウィレム・パウル
イギリス オーエックス４　４ジーイー オックスフォード オックスフォード・サイエンス・パーク（無番地）　ザ・シュレディンガー・ビルディング　エクセンシア・リミテッド内