(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2019-10095
(P2019-10095A)
(43)公開日　平成31年1月24日(2019.1.24)

(51)Int.Cl.		F I			テーマコード（参考）
C12M 1/34	(2006.01)	C12M	1/34		4B029
G06N 20/00	(2019.01)	G06N	99/00	153	

審査請求 未請求　請求項の数 8　OL　（全 16 頁）

(21)出願番号　特願2018-122565 (P2018-122565)
(22)出願日　平成30年6月28日 (2018.6.28)
(31)優先権主張番号　特願2017-129823 (P2017-129823)
(32)優先日　平成29年6月30日 (2017.6.30)
(33)優先権主張国　日本国(JP)

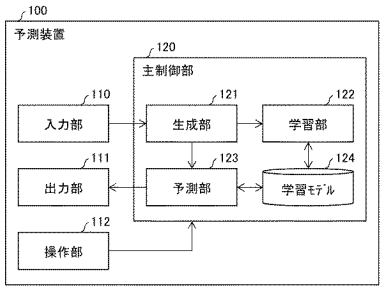
(71)出願人　505082350
学校法人　明治薬科大学
東京都清瀬市野塩２丁目５２２番１
(74)代理人　110000338
特許業務法人HARAKENZO WORLD PATENT & TRADEMARK
(72)発明者　植沢　芳広
東京都清瀬市野塩二丁目５２２番１　学校法人明治薬科大学内

Fターム(参考)　4B029 AA07　BB15　BB20　FA15

(54)【発明の名称】予測装置、予測方法、予測プログラム、学習モデル入力データ生成装置および学習モデル入力データ生成プログラム

(57)【要約】
【課題】対象化合物の構造に基づいて、対象化合物の活性を好適に予測する。
【解決手段】予測装置（１００）は、仮想カメラによって対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成部（１２１）と、学習モデル（１２４）を用いて前記生成部が生成した前記複数の撮像画像から前記対象化合物の活性を予測する予測部（１２３）と、を備えている。
【選択図】図１

図１



【特許請求の範囲】
【請求項１】
対象化合物の構造に基づいて、前記対象化合物の活性を予測する予測装置であって、
仮想カメラによって前記対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成部と、
学習モデルを用いて前記生成部が生成した前記複数の撮像画像から前記対象化合物の活性を予測する予測部と、を備えていることを特徴とする予測装置。
【請求項２】
前記予測部は、少なくとも、機械学習を行う学習モデルであって、前記複数の撮像画像を入力とする学習モデルを用いることを特徴とする請求項１に記載の予測装置。
【請求項３】
前記生成部は、前記仮想カメラを、前記構造モデルに対して少なくとも１つの軸を中心に相対的に回転させながら前記構造モデルを撮像することを特徴とする請求項１または２に記載の予測装置。
【請求項４】
前記構造モデルでは、前記対象化合物の原子の色は、当該原子の種類に応じて異なることを特徴とする請求項１～３の何れか一項に記載の予測装置。
【請求項５】
対象化合物の構造に基づいて、前記対象化合物の活性を予測する予測方法であって、
コンピュータが、仮想カメラによって前記対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成ステップと、
コンピュータが、学習モデルを用いて前記生成ステップにおいて生成された前記複数の撮像画像から前記対象化合物の活性を予測する予測ステップと、を包含することを特徴とする予測方法。
【請求項６】
請求項１～４の何れか一項に記載の予測装置としてコンピュータを機能させるための予測プログラムであって、上記生成部および上記予測部としてコンピュータを機能させるための予測プログラム。
【請求項７】
学習モデルの入力データを生成する学習モデル入力データ生成装置であって、
前記学習モデルは、仮想カメラによって対象化合物の構造モデルが相対的に複数の方向から撮像された複数の撮像画像を入力とし、当該対象化合物の活性の予測情報を出力とする学習モデルであり、
仮想カメラによって前記対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成部を備えていることを特徴とする学習モデル入力データ生成装置。
【請求項８】
請求項７に記載の学習モデル入力データ生成装置としてコンピュータを機能させるための学習モデル入力データ生成プログラムであって、上記生成部としてコンピュータを機能させるための学習モデル入力データ生成プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、学習モデルを利用する予測装置、予測方法および予測プログラム、ならびに、学習モデル入力データ生成装置および学習モデル入力データ生成プログラムに関する。
【背景技術】
【０００２】
化学物質毎の生理活性の相違は、化学構造に由来すると考えることができる。定量的構造活性相関（ＱＳＡＲ：Quantitative Structure-Activity Relationship）予測モデルは、化学構造と生理活性の間に成立するルールを数学的モデルとして表現したものであり、

定量的構造活性相関予測モデルを構築することによって、生理活性が未知の化学物質であっても実験をせずにその活性を予測することができる（特許文献１～４参照）。

【０００３】

従来の定量的構造活性相関モデルの構築法においては、まず、下記表１に示すように、化学構造を化学構造記述子と呼ばれる多様な数値群に変換する。その後、化学構造記述子から統計解析または機械学習によって数学的モデルを構築する。化学構造記述子は、通常、専用のソフトウェアを用いて数百から数千種類を計算する。化学構造記述子の組合せは、定量的構造活性相関予測モデルの汎化性能の高さに直結し、例えば、人の手によって選択される。

【０００４】

【表１】

化合物	MW	ALOGP	TPSA(Tot)	nH	nC	nN	nO	G(N..N)	G(O..O)	…
化合物１	130.09	-1.102	65.72	3	4	2	2	0	4.511	…
化合物２	151.18	0.683	49.33	9	8	1	2	0	6.605	…
…	…	…	…	…	…	…	…	…	…	…

また、より優れた定量的構造活性相関予測モデルの構築を競う国際的な活性予測コンペティション（Tox21DataChallenge2014）が知られている。

【先行技術文献】

【特許文献】

【０００５】

【特許文献１】米国特許第７７０２４６７号明細書

【特許文献２】米国特許第７７５１９８８号明細書

【特許文献３】米国特許出願公開第２００４／０００９５３６号明細書

【特許文献４】米国特許出願公開第２００４／０１９９３３４号明細書

【発明の概要】

【発明が解決しようとする課題】

【０００６】

従来技術では、上述したように、予測の精度を向上させるために、化学構造記述子の組合せを注意深く選定する必要がある。化学構造記述子の組合せを選定することなく、予測の精度を向上させることができれば、非常に有用である。

【０００７】

本発明の一態様は、上記課題に鑑みてなされたものであり、対象化合物の構造に基づいて、対象化合物の活性を好適に予測するための新規な技術を提供することを目的とする。

【課題を解決するための手段】

【０００８】

上記の課題を解決するために、本発明の一態様に係る予測装置は、対象化合物の構造に基づいて、前記対象化合物の活性を予測する予測装置であって、仮想カメラによって前記対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成部と、学習モデルを用いて前記生成部が生成した前記複数の撮像画像から前記対象化合物の活性を予測する予測部と、を備えている。

【０００９】

また、本発明の一態様に係る予測方法は、対象化合物の構造に基づいて、前記対象化合物の活性を予測する予測方法であって、コンピュータが、仮想カメラによって前記対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成ステップと、コンピュータが、学習モデルを用いて前記生成ステップにおいて生成された前記複数の撮像画像から前記対象化合物の活性を予測する予測ステップと、を包含する。

【００１０】

また、本発明の一態様に係る学習モデル入力データ生成装置は、学習モデルの入力デー

タを生成する学習モデル入力データ生成装置であって、前記学習モデルは、仮想カメラによって対象化合物の構造モデルが相対的に複数の方向から撮像された複数の撮像画像を入力とし、当該対象化合物の活性の予測情報を出力とする学習モデルであり、仮想カメラによって前記対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成部を備えている。

【発明の効果】

【００１１】

本発明の一態様によれば、対象化合物の構造に基づいて、対象化合物の活性を好適に予測することができる。

【図面の簡単な説明】

【００１２】

【図１】本発明の一実施形態に係る予測装置の概略構成の一例を示す機能ブロック図である。

【図２】本発明の一実施形態における画像生成の一例を概略的に説明する模式図である。

【図３】本発明の一実施形態における画像生成の一例を詳細に説明する模式図である。

【図４】本発明の一実施形態における学習処理の流れの一例を説明するフローチャートである。

【図５】本発明の一実施形態における予測処理の流れの一例を説明するフローチャートである。

【図６】本発明の一実施形態における予測結果の一例を示すグラフである。

【図７】本発明の一実施形態における予測結果の一例を示すグラフである。

【図８】本発明の一実施形態における予測結果の一例を示すグラフである。

【発明を実施するための形態】

【００１３】

〔実施形態１〕

以下、本発明の一実施形態について、詳細に説明する。図１は、本発明の一実施形態に係る予測装置１００の概略構成の一例を示す機能ブロック図である。予測装置１００は、入力部１１０、出力部１１１、操作部１１２および主制御部１２０を備えている。主制御部１２０は、生成部１２１、学習部１２２、予測部１２３および学習モデル１２４を備えている。

【００１４】

予測装置１００は、対象化合物の構造に基づいて、対象化合物の活性を予測する予測装置である。一態様において、予測装置１００は、入力部１１０から入力された対象化合物の構造を示すデータに基づいて、学習モデル１２４を用いて対象化合物の活性を予測し、その結果を出力部１１１が出力する。また、一態様において、予測装置１００は、入力部１１０から入力された参照化合物の構造を示すデータおよび参照化合物の活性を示すデータに基づいて、学習モデル１２４の学習を行う。なお、本明細書において、学習モデル１２４に学習させる情報の源となる化合物を参照化合物とし、学習モデル１２４によって活性を予測する化合物を対象化合物とする。

【００１５】

また、一態様において、予測装置１００は、学習モデル１２４に入力する入力データを生成する学習モデル入力データ生成装置としても機能する。さらに、一変形例において入力部１１０および生成部１２１を備えた学習モデル入力データ生成装置と、学習部１２２、予測部１２３および学習モデル１２４を備えた学習モデル装置とによって、予測装置を構成するようにしてもよい。

【００１６】

（入力部）

入力部１１０は、予測装置１００に対する、対象化合物の構造を示すデータ、または、参照化合物の構造を示すデータおよび参照化合物の活性を示すデータの入力を受け付けるものである。入力部１１０は、記憶媒体に記憶されたデータファイルを読み込むこと、ま

たは、有線または無線のネットワークを介して他の装置からデータを受信することによって、上述したデータの入力を受け付ける。

【００１７】

（化合物の構造を示すデータ）

対象化合物および参照化合物として用いる化合物の構造、由来、物性等は特に限定されず、例えば、天然化合物、合成化合物、高分子化合物、低分子化合物等であり得る。化合物の構造を示すデータは、ＰｕｂＣｈｅｍ（http://pubchem.ncbi.nlm.nih.gov）のような公開データベースから取得してもよいし、新たに作成したものであってもよい。化合物の構造を示すデータの形式は特に限定されず、例えば、ＳＤＦ形式等の公知のデータ形式であり得る。

【００１８】

化合物の構造を示すデータを作成する場合、例えば、二次元化学構造から三次元構造を生成する公知のソフトウェア（例えば、Corina（http://www.mn-am.com/products/corina）等）を用いることができる。三次元構造を生成する際の種々の条件（例えば、真空中であるか水溶液中であるか、温度条件、ｐＨ等）は特に限定されず、例えば、特定の条件（例えば、真空中で最も安定）を満たす三次元構造を示すデータを作成してもよい。また、公知のドッキングアルゴリズム（例えば、ＤＯＣＫ等）により、所望のタンパク質と結合状態となる三次元構造を推定し、当該三次元構造を示すデータを作成してもよい。これにより、より高度な予測を行うことができる。

【００１９】

また、一態様において、１つの化合物に対し、三次元構造を示すデータを複数生成してもよい。例えば、水溶液中などにおける原子間の結合の自由度を考慮し、一分子毎に分子内の回転可能な官能基を回転させることによって多様な三次元構造を生成してもよい。また、分子動力学（ＭＤ）シミュレーションによって熱エネルギーによる分子振動を考慮して多様な三次元構造を生成してもよい。これにより、後述する生成部１２１によってより多くの画像を生成することができ、より精度の高い予測を行うことができる。

【００２０】

（化合物の活性を示すデータ）

参照化合物の活性を示すデータは、例えば、ＰｕｂＣｈｅｍ（http://pubchem.ncbi.nlm.nih.gov）のような公開データベースから取得してもよいし、実験的に求めたものであってもよい。参照化合物の活性を示すデータの形式は、特に限定されないが、所望の活性を有するか否かの二値を示すデータであってもよいし、複数のカテゴリー値から選択される値を示すデータであってもよいし、連続変数を示すデータであってもよい。

【００２１】

所望の活性は、特に限定されず、薬学的な活性、生理学的な活性、生化学的な活性、毒性等、様々な活性であり得る。

【００２２】

（出力部）

出力部１１１は、予測部１２３による対象化合物の活性の予測結果を出力するものである。例えば、一態様において、出力部１１１は、予測結果を画像データまたは文字データとして表示装置に出力するものであってもよいし、予測結果を示す画像データ、文字データまたはバイナリデータを含むデータファイルを出力するものであってもよいし、予測結果を示す画像データ、文字データまたはバイナリデータを、有線または無線のネットワークを介して他の装置に送信するものであってもよい。

【００２３】

（操作部）

操作部１１２は、予測装置１００に対するユーザの操作を受け付ける。操作部１１２は、例えば、キーボード、マウス、トラックボール、タッチパッド（タッチパネルを含む）、光学センサ、音声入力のためのマイク等であり得る。

【００２４】

（主制御部）
主制御部１２０は、一つ以上のコンピュータによって構成されている。主制御部１２０が複数のコンピュータによって構成されている場合、複数のコンピュータは互いに有線または無線接続されており、主制御部１２０の機能を分担するものであってもよい。
【００２５】
（学習モデル）
学習モデル１２４としては、機械学習を行う学習モデルであって、仮想カメラによって対象化合物の構造モデルが複数の方向から撮像された複数の撮像画像を入力とし、当該対象化合物の活性の予測情報を出力とする学習モデルであることが好ましく、深層学習（Deep Learning）を行う学習モデルを用いることがより好ましく、例えば、AlexNet、CaffeNet、GoogLeNet、VGG net等の畳み込みニューラルネットワークを用いることができる。
【００２６】
対象化合物の活性の予測情報としては、特に限定されないが、対象化合物が所望の活性を有している確率を示す情報、対象化合物が所望の活性を有しているか否かの予測結果を示す情報、対象化合物が所望の活性を有している可能性に対応するスコア等であり得る。
【００２７】
また、一態様において、学習モデル１２４は、複数の学習モデルの組み合わせであってもよい。すなわち、学習モデル１２４は、仮想カメラによって対象化合物の構造モデルが複数の方向から撮像された複数の撮像画像を入力とし、特徴ベクトルを出力する第１の学習モデルと、特徴ベクトルを入力とし、当該対象化合物の活性の予測情報を出力とする第２の学習モデルとを組み合わせたものであってもよい。この場合、第１の学習モデルとしては、対象化合物の構造モデルが複数の方向から撮像された複数の撮像画像を入力とする学習モデルであればよいが、深層学習を行う学習モデルを用いることが好ましい。また、第２の学習モデルとしては、深層学習を行う学習モデルを用いてもよいし、深層学習を行わない学習モデル等を用いてもよい。
【００２８】
（生成部）
生成部１２１は、仮想カメラによって対象化合物または参照化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像（スナップショット）を生成する。図２は、本実施形態における画像生成の一例を概略的に説明する模式図である。図２に示すように、生成部１２１は、仮想空間に配置した対象化合物の構造モデル１０を回転させ、仮想カメラによって相対的に複数の方向から撮像して撮像画像を生成する（図２の（ａ）～（ｃ）に示す画像）。なお、生成部１２１は、構造モデル１０を回転させる代わりに、仮想カメラを移動させることによって、構造モデル１０に対して相対的に複数の方向から撮像してもよい。なお、本明細書において「撮像画像」とは、スナップショットとも称され、仮想空間に配置した構造モデルを仮想カメラによって撮像して得られる画像を意味し、当該画像と同一の内容の画像であれば、構造モデルを構築せずに座標データから直接算出した画像も含まれる。
【００２９】
構造モデルの生成および仮想カメラによる撮像は、分子構造の三次元的な表示および仮想カメラによる撮像が可能な公知のソフトウェア（例えば、Ｊｍｏｌ（http://jmol.sourceforge.net/）、ＶＭＤ（http://www.ks.uiuc.edu/Research/vmd/）、ＵＣＳＦ　Ｃｈｉｍｅｒａ（http://www.cgl.ucsf.edu/chimera/）、Ｒａｓｍｏｌ（http://www.umass.edu/microbio/rasmol/）、ＰｙＭＯＬ（http://www.pymol.org/）等）を用いることができる。
【００３０】
一態様において、生成する撮像画像の画像ファイルは、例えば、ＲＧＢ三色のドットの集合として入力され、二次元平面の位置情報と３種の色情報が数値化されているものであり得る。生成部１２１が生成する撮像画像のサイズは特に限定されず、対象化合物および参照化合物の大きさ等に応じて適宜調整すればよいが、例えば、１２８画素×１２８画素

、２５６画素×２５６画素、５１２画素×５１２画素、１０２４画素×１０２４画素といったサイズとすることができる。また、色深度は、特に限定されず、例えば、１～６４ｂｐｐの範囲とすることができるが、好ましくは、８～３２ｂｐｐの範囲であり得る。

【００３１】

図３は、本実施形態における画像生成の一例を詳細に説明する模式図である。図３では、構造モデル２０を、Ball and Stick表示している。なお、Ball and Stick表示とは、原子を球で、結合を棒で示す表示である。ただし、本実施形態はこれに限定されず、構造モデルを、結合のみによって示すWireframe表示、原子によって空間を充填するSpacefill表示、水溶液に接する分子の表面を表示するSurface表示、タンパク質の構造を模式的に示すRibbons表示等によって表示してもよい。

【００３２】

図３の（ａ）に示すように、構造モデル２０には、原子２１、結合２２および水素原子２３が含まれている。なお、原子２１は、水素原子以外の原子を示す。一態様において、水素原子２３は、構造モデル２０に含めなくともよい。構造モデル２０では、原子２１の色は、当該原子の種類に応じて異なっているが、これに限定されず、原子２１の色は同一であってもよいし、原子の種類を適宜グループ分けし、原子２１の色は、当該原子が属するグループに応じて異なっているようにしてもよい。

【００３３】

また、原子２１の半径は特に限定されず、例えば、半径の上限を、Van der Waals半径の５０％以下、４０％以下、３０％以下、２０％以下、１０％以下、５％以下、３％以下、１％以下とすることができ、半径の下限を、Van der Waals半径の０．１％以上、０．３％以上、０．７％以上、１％以上とすることができるが、０．１％以上３０％以下とすることが好ましく、０．１％以上１０％以下とすることがより好ましく、０．１％以上３％以下とすることが特に好ましい。

【００３４】

また、結合２２の太さは特に限定されず、例えば、太さの上限を、３００ミリオングストローム以下、２００ミリオングストローム以下、１００ミリオングストローム以下、５０ミリオングストローム以下、３０ミリオングストローム以下、２０ミリオングストローム以下とすることができ、太さの下限を、１ミリオングストローム以上、２ミリオングストローム以上、５ミリオングストローム以上、１０ミリオングストローム以上とすることができるが、１ミリオングストローム以上、２００ミリオングストローム以下とすることが好ましく、２ミリオングストローム以上、１００ミリオングストローム以下とすることがより好ましく、２ミリオングストローム以上、３０ミリオングストローム以下とすることが特に好ましい。

【００３５】

そして、一態様において、生成部１２１は、仮想カメラを、構造モデル２０に対して少なくとも１つの軸を中心に相対的に回転させながら構造モデル２０を撮像する。軸としては、特に限定されないが、例えば、構造モデル２０が配置された仮想空間のＸ軸、Ｙ軸およびＺ軸から選択される１つ以上の軸とすることができる。例えば、図３の（ｂ）は、構造モデル２０を、図３の（ａ）に示すＸ軸を中心に４５度回転させて撮像した撮像画像を示し、図３の（ｃ）は、構造モデル２０を、図３の（ａ）に示すＹ軸を中心に４５度回転させて撮像した撮像画像を示し、図３の（ｄ）は、構造モデル２０を、図３の（ａ）に示すＸ軸およびＹ軸に直交するＺ軸を中心に４５度回転させて撮像した撮像画像を示す。

【００３６】

なお、回転角度は、特に限定されず、１度～１８０度の範囲の任意の角度、好ましくは、１度～９０度の範囲の任意の角度、より好ましくは、１度～４５度の任意の角度毎に撮像すればよく、撮像毎に回転角度を変更してもよいか、例えば、３０度毎、４５度毎、９０度毎に撮像することができる。複数の軸を中心に回転させる場合には、各軸について取り得る角度を網羅するように撮像する。すなわち、Ｘ軸およびＹ軸を中心に９０度毎に撮像する場合には、１化合物あたりの撮像画像数は４×４＝１６枚となる。また、Ｘ軸、Ｙ

軸およびＺ軸を中心に４５度毎に撮像する場合には、１化合物あたりの撮像画像数は８×８×８＝５１２枚となる。このように網羅的に撮像することにより、あらゆる方向から視認した構造モデル２０のスナップショットを撮影することができる。

【００３７】

（学習部）

学習部１２２は、公知の方法により、生成部１２１が生成した参照化合物の各撮像画像と当該参照化合物の活性との対応を学習モデル１２４に学習させる。一態様において、学習部１２２は、公知の深層学習アルゴリズムを用いて、学習モデル１２４に、生成部１２１が生成した参照化合物の各撮像画像と当該参照化合物の活性との対応を学習させる。学習部１２２は、例えば、Ｄｉｇｉｔｓ（ＮＶＩＤＩＡ社）等の公知の深層学習フレームワークを利用してもよい。

【００３８】

（予測部）

予測部１２３は、生成部１２１が生成した対象化合物の各撮像画像と当該対象化合物の活性との対応を学習した学習モデル１２４を用いて、生成部１２１が生成した対象化合物の各撮像画像から対象化合物の活性を予測する。予測部１２３は、例えば、Ｄｉｇｉｔｓ（ＮＶＩＤＩＡ社）等の公知の深層学習フレームワークを利用してもよい。

【００３９】

一態様において、対象化合物の各撮像画像を入力したときの学習モデル１２４の出力が、対象化合物が所望の活性を有する確率を示す値である場合には、予測部１２３は、対象化合物の各撮像画像を入力したときの学習モデル１２４の各出力値の代表値（例えば、中央値、平均値、合計）を取得し、当該代表値を閾値と比較することにより、対象化合物が所望の活性を有しているか否かを予測することができる。

【００４０】

閾値としては、任意の値を用いることができるが、学習済みの学習モデル１２４に対し、参照化合物の各撮像画像を入力したときの出力値をＲＯＣ解析することによって算出した閾値を用いることが好ましい。

【００４１】

（学習処理）

図４は、本発明の一実施形態における学習処理の流れの一例を説明するフローチャートである。まず、操作部１１２による操作等により学習処理が開始されると、生成部１２１は、入力部１１０を介して、参照化合物の構造を示すデータおよび参照化合物の活性を示すデータを取得する（ステップＳ１）。続いて、生成部１２１は、ステップＳ１において入力されたデータのうち、未処理の参照化合物の構造を示すデータに基づいて、未処理の参照化合物の構造モデルを生成する（ステップＳ２）。続いて、生成部１２１は、仮想カメラによって、ステップＳ２において生成した参照化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する（ステップＳ３）。一態様において、生成部１２１は、ステップＳ３において、仮想カメラを、ステップＳ２において生成した構造モデルに対して少なくとも１つの軸を中心に相対的に回転させながら構造モデルを撮像することにより、複数の撮像画像を生成する。以上の処理が、ステップＳ１において入力されたデータに含まれる全ての参照化合物について完了した場合（ステップＳ４のｙｅｓ）には、ステップＳ５に進み、完了していない場合（ステップＳ４のｎｏ）には、ステップＳ２に戻る。

【００４２】

次に、学習部１２２が、公知の機械学習アルゴリズム（特に、深層学習アルゴリズム）によって、ステップＳ３において生成した参照化合物の各撮像画像と、ステップＳ１において入力された当該参照化合物の活性との対応を、学習モデル１２４に学習させる（ステップＳ５）。なお、学習部１２２が、Ｄｉｇｉｔｓを利用している場合、予め参照化合物に割り振った教師データ（例えば、所望の活性有り＝１、無し＝０）毎に異なるフォルダに撮像画像を格納することにより、ステップＳ５を好適に実行することができる。また、

各撮像画像に対応する参照化合物の教師データを紐付けてもよい。ステップＳ５が、ステップＳ１において入力されたデータに含まれる全ての参照化合物について完了した場合（ステップＳ６のｙｅｓ）には、学習処理を終了し、完了していない場合（ステップＳ６のｎｏ）には、ステップＳ５に戻る。

【００４３】

以上により、予測装置１００は、学習モデル１２４を、仮想カメラによって化合物の構造モデルが複数の方向から撮像された複数の撮像画像を入力とし、当該化合物の活性の予測情報を出力とする学習済みモデルとすることができる。

【００４４】

（予測処理）

図５は、本発明の一実施形態における予測処理の流れの一例を説明するフローチャートである。まず、操作部１１２による操作等により予測処理が開始されると、生成部１２１は、入力部１１０を介して、対象化合物の構造を示すデータを取得する（ステップＳ１１）。続いて、生成部１２１は、ステップＳ１１において入力されたデータのうち、未処理の対象化合物の構造を示すデータに基づいて、未処理の対象化合物の構造モデルを生成する（ステップＳ１２）。続いて、生成部１２１は、仮想カメラによって、ステップＳ１２において生成した対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する（ステップＳ１３）。一態様において、生成部１２１は、ステップＳ３において、仮想カメラを、ステップＳ１２において生成した構造モデルに対して少なくとも１つの軸を中心に相対的に回転させながら構造モデルを撮像することにより、複数の撮像画像を生成する。以上の処理が、ステップＳ１１において入力されたデータに含まれる全ての対象化合物について完了した場合（ステップＳ１４のｙｅｓ）には、ステップＳ１５に進み、完了していない場合（ステップＳ１４のｎｏ）には、ステップＳ１２に戻る。

【００４５】

次に、予測部１２３が、学習モデル１２４に対して、ステップＳ１３において生成した対象化合物の各撮像画像を入力し、学習モデル１２４からの出力を取得する。一実施形態において、学習モデル１２４からの出力が、対象化合物が所望の活性を有する確率を示す値である場合、予測部１２３は、１つの対象化合物の各撮像画像を入力したときの学習モデル１２４からの出力値の中央値を取得する（ステップＳ１５）。そして、予測部１２３は、ステップＳ１５において取得した中央値と、閾値とを比較することにより、対象化合物が所望の活性を有しているか否かを予測する（ステップＳ１６）。ステップＳ１５～Ｓ１６が、ステップＳ１１において入力されたデータに含まれる全ての対象化合物について完了した場合（ステップＳ１７のｙｅｓ）には、予測処理を終了し、完了していない場合（ステップＳ１７のｎｏ）には、ステップＳ１５に戻る。

【００４６】

以上により、予測装置１００は、対象化合物が所望の活性を有しているか否かを予測することができる。

【００４７】

（本実施形態の効果）

本実施形態によれば、多数の化合物について、実験することなく、薬効、毒性、酵素阻害活性等の活性を予測することができる。

【００４８】

特に、本実施形態によれば、学習モデル１２４に対する入力が画像であることによって、鏡像異性体を識別可能となる。記述子を使用する従来法では、記述子では鏡像異性体間で同じ値を取るため、多様な化合物を使用する場合に鏡像異性体間の活性差を表現することが困難である。これに対し、本実施形態によれば、撮像画像には、鏡像異性体を識別するための情報が含まれているために、当該情報も学習モデル１２４によるパターン認識に使用され、鏡像異性体を識別可能となる。鏡像異性体間で異なる生理活性を有する事例は普遍的であるので、本実施形態は非常に有用である。

【００４９】
　また、学習モデル１２４において、深層学習を行う学習モデルを用いることにより、偏ったデータに対応可能である。すなわち、入力する参照化合物の所望の活性の有無の比率が、例えば、１対１０のような極端な比率であっても良好な精度を得ることができる。一方、従来法では、データにおける活性の有無の比率が１：１程度で最も良好な精度のモデルが構築できるが、偏りのあるデータの取扱いは困難である。毒性等は、一部の化合物のみが活性を示すため、本実施形態は非常に有用である。
【００５０】
　また、本実施形態によれば、学習モデル１２４に対する入力が、構造モデルを複数の方向から撮像した撮像画像であることによって、対象化合物の構造を網羅的に示す情報を含むデータを学習モデルに入力することができ、対象化合物の活性を好適に予測することができる。特に、学習モデル１２４に対する入力が、構造モデルを、一つ以上の軸を中心に仮想カメラを相対的に回転させながら撮像した撮像画像とすることによって、対象化合物の構造をより網羅的に示す情報を含むデータを学習モデルに入力することができ、対象化合物の活性をより好適に予測することができる。
【００５１】
　また、構造モデルにおける原子および結合のサイズを上述したように規定することにより、外側の原子または結合によって内側の原子または結合が隠されることを抑制し、撮像画像に内側の原子または結合に関する情報を含ませることができる。これにより、活性を好適に予測することができる。
【００５２】
　また、構造モデルにおける原子の色を、原子の種類によって異ならせることにより、撮像画像に原子の種類に関する情報を含ませることができる。これにより、活性を好適に予測することができる。
【００５３】
　〔変形例〕
　上述した実施形態では、予測部１２３は、学習モデル１２４を用いて、各撮像画像の夫々について、当該撮像画像の対象化合物が所望の活性を有するか否かを予測し、その結果を統合して、対象化合物の活性を予測しているが、本発明はこれに限定されない。例えば、学習部１２２は、学習モデル１２４に、参照化合物の各撮像画像を一体化したデータと、当該参照化合物の活性との対応を学習させ、予測部１２３は、学習モデル１２４に、対象化合物の各撮像画像を一体化したデータを入力し、当該対象化合物の活性を予測するようにしてもよい。
【００５４】
　また、上述した実施形態では、予測部１２３が、学習モデル１２４の各出力値の代表値を閾値と比較することにより、対象化合物の活性を予測しているが、本発明はこれに限定されない。例えば、学習部１２２は、別の学習モデルに、参照化合物の各撮像画像を入力したときの学習モデル１２４の出力値と、当該参照化合物の活性との対応を学習させ、予測部１２３は、学習モデル１２４の各出力値を当該別の学習モデルに入力することにより、当該対象化合物の活性を予測するようにしてもよい。
【００５５】
　以上のように、本発明は、一態様において、仮想カメラによって対象化合物の構造モデルを相対的に複数の方向から撮像した複数の撮像画像を学習モデルに入力し、その出力に基づいて対象化合物の活性を予測することをポイントとするものであり、その他の構成については様々な態様を取り得る。
【００５６】
　〔ソフトウェアによる実現例〕
　予測装置１００の制御ブロック（主制御部１２０、特に生成部１２１、学習部１２２および予測部１２３）は、集積回路（ＩＣチップ）等に形成された論理回路（ハードウェア）によって実現してもよいし、ソフトウェアによって実現してもよい。

【００５７】
後者の場合、予測装置１００は、各機能を実現するソフトウェアであるプログラムの命令を実行するコンピュータを備えている。このコンピュータは、例えば少なくとも１つのプロセッサ（制御装置）を備えていると共に、上記プログラムを記憶したコンピュータ読み取り可能な少なくとも１つの記録媒体を備えている。そして、上記コンピュータにおいて、上記プロセッサが上記プログラムを上記記録媒体から読み取って実行することにより、本発明の目的が達成される。上記プロセッサとしては、例えばＣＰＵ（Central Processing Unit）を用いることができる。上記記録媒体としては、「一時的でない有形の媒体」、例えば、ＲＯＭ（Read Only Memory）等の他、テープ、ディスク、カード、半導体メモリ、プログラマブルな論理回路などを用いることができる。また、上記プログラムを展開するＲＡＭ（Random Access Memory）などをさらに備えていてもよい。また、上記プログラムは、該プログラムを伝送可能な任意の伝送媒体（通信ネットワークや放送波等）を介して上記コンピュータに供給されてもよい。なお、本発明の一態様は、上記プログラムが電子的な伝送によって具現化された、搬送波に埋め込まれたデータ信号の形態でも実現され得る。
【００５８】
〔まとめ〕
本発明の態様１に係る予測装置（１００）は、対象化合物の構造に基づいて、前記対象化合物の活性を予測する予測装置であって、仮想カメラによって前記対象化合物の構造モデル（１０、２０）に対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成部（１２１）と、学習モデル（１２４）を用いて前記生成部が生成した前記複数の撮像画像から前記対象化合物の活性を予測する予測部（１２３）と、を備えている。上記の構成によれば、対象化合物の構造に基づいて、記述子の組み合わせを選択することなく、対象化合物の活性を好適に予測することができる。また、学習モデルに対する入力が画像であることによって、鏡像異性体を識別可能となる。
【００５９】
本発明の態様２に係る予測装置は、上記態様１において、前記予測部は、少なくとも、機械学習を行う学習モデルであって、前記複数の撮像画像を入力とする学習モデルを用いてもよい。上記の構成によれば、対象化合物の活性を好適に予測することができる。
【００６０】
本発明の態様３に係る予測装置は、上記態様１または２において、前記生成部は、前記仮想カメラを、前記構造モデルに対して少なくとも１つの軸を中心に相対的に回転させながら前記構造モデルを撮像してもよい。上記の構成によれば、対象化合物の構造を網羅的に示す撮像画像を生成することができるため、活性を好適に予測することができる。
【００６１】
本発明の態様４に係る予測装置は、上記態様１～３において、前記構造モデルでは、前記対象化合物の原子（２１）の色は、当該原子の種類に応じて異なってもよい。上記の構成によれば、対象化合物の原子の種類を示す情報を含む撮像画像を生成することができるため、活性を好適に予測することができる。
【００６２】
本発明の態様５に係る予測方法は、対象化合物の構造に基づいて、前記対象化合物の活性を予測する予測方法であって、コンピュータが、仮想カメラによって前記対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成ステップと、コンピュータが、学習モデルを用いて前記生成ステップにおいて生成された前記複数の撮像画像から前記対象化合物の活性を予測する予測ステップと、を包含する。上記の構成によれば、上記態様１と同等の効果を奏する。
【００６３】
本発明の各態様に係る予測装置は、コンピュータによって実現してもよく、この場合には、コンピュータを上記予測装置が備える各部（ソフトウェア要素）として動作させることにより上記予測装置をコンピュータにて実現させる予測装置の予測プログラム、および

それを記録したコンピュータ読み取り可能な記録媒体も、本発明の範疇に入る。

【００６４】

本発明の態様７に係る学習モデル入力データ生成装置（１００）は、学習モデルの入力データを生成する学習モデル入力データ生成装置であって、前記学習モデルは、仮想カメラによって対象化合物の構造モデルが相対的に複数の方向から撮像された複数の撮像画像を入力とし、当該対象化合物の活性の予測情報を出力とする学習モデル（１２４）であり、仮想カメラによって前記対象化合物の構造モデルに対して相対的に複数の方向から撮像して複数の撮像画像を生成する生成部（１２１）を備えている。上記の構成によれば、上記態様１と同等の効果を奏する。

【００６５】

本発明の各態様に係る学習モデル入力データ生成装置は、コンピュータによって実現してもよく、この場合には、コンピュータを上記学習モデル入力データ生成装置が備える各部（ソフトウェア要素）として動作させることにより上記学習モデル入力データ生成装置をコンピュータにて実現させる学習モデル入力データ生成装置の学習モデル入力データ生成プログラム、およびそれを記録したコンピュータ読み取り可能な記録媒体も、本発明の範疇に入る。

【００６６】

本発明は上述した各実施形態に限定されるものではなく、請求項に示した範囲で種々の変更が可能であり、異なる実施形態にそれぞれ開示された技術的手段を適宜組み合わせて得られる実施形態についても本発明の技術的範囲に含まれる。さらに、各実施形態にそれぞれ開示された技術的手段を組み合わせることにより、新しい技術的特徴を形成することができる。

【００６７】

〔実施例１〕

Tox21DataChallenge2014のサイト（https://tripod.nih.gov/tox21/challenge/data.jsp）において公開された７３２０種類の化合物に基づく学習用データ、および、学習用データの化合物とは重複しない５４３種類の化合物に基づくテスト用データを用いて、本発明の一態様を実施した。予測対象の所望の活性は、ミトコンドリア膜電位攪乱活性とした。

【００６８】

まず、Ｊｍｏｌ（http://jmol.sourceforge.net/）を利用し、ＳＤＦファイルに基づいて化合物の構造モデルを生成し、各構造モデルに対し、Ｘ軸、Ｙ軸、Ｚ軸それぞれを中心に４５度刻みで回転させて撮像した５１２枚の撮像画像（スナップショット、サイズ：５１２×５１２、２４ｂｐｐ）を生成するプログラム（学習モデル入力データ生成プログラム）を作成した。当該プログラムを実行し、学習用データのＳＤＦファイルを入力し、各化合物についての撮像画像を生成した。各化合物の撮像画像は、当該化合物がミトコンドリア膜電位攪乱活性を有するか否かに応じた所定のフォルダに格納し、Ｄｉｇｉｔｓ（ＮＶＩＤＩＡ社）を用いて未改変のAlexNet（トロント大学）を学習させた。学習では、Ｄｉｇｉｔｓの設定を、学習率＝０．００１、ｅｐｏｃｈ＝１とした。ｅｐｏｃｈは、１つの学習用データを繰り返して学習させる回数を示す。

【００６９】

さらに、テスト用データを用いて、外部検証法によって予測性能を確認した。具体的には、前記プログラムを実行し、テスト用データのＳＤＦファイルを入力し、各化合物についての撮像画像を生成した。各化合物の撮像画像を、学習済みのAlexNetに入力し、出力値の中央値を取得し、ＲＯＣ解析を行った。その結果を図６に示す。図６に示すように、ＲＯＣ曲線下面積（ＡＵＣ）は、０．９０９であり、０．９以上の高値となった。なお、ここで用いたデータセットは２０１４年にＮＩＨによって開催された「Tox21 data challenge 2014」に使用されたものと同一であり、AlexNetを調整していないにもかかわらず、上記のＡＵＣ値はコンペティションの上位１０位と同等の成績となった。

【００７０】

〔実施例２〕
Ｄｉｇｉｔｓの設定を、学習率＝０．０００１、ｅｐｏｃｈ＝８に変更した以外は実施例１と同様に、本発明の一態様を実施した。その結果、図７に示すように、ＲＯＣ＿ＡＵＣ値は、実施例１の０．９０９から0.９２１２２に向上した。AlexNetを調整していないにもかかわらず、上記のＡＵＣ値は「Tox21 data challenge 2014」の上位１０位以内の成績となった。

【００７１】
〔実施例３〕
文献（Derivation and Validation of Toxicophores for Mutagenicity Prediction. J. Med. Chem. 2005, 48, 312-320.）の付録資料から取得した、総計４３３７化合物の立体構造（ＳＤＦファイル形式）と、各化合物に対するＡＭＥＳ試験結果（陽性又は陰性）とを用いて、本発明の一態様を実施した。予測対象の所望の活性は、変異原性（ＡＭＥＳ試験結果）とした。詳細には、以下の手順で試験を行った。

【００７２】
まず、総計４３３７化合物を、予測モデルの学習用の化合物群（４１３７化合物）と、予測結果の外部検証用の化合物群（２００化合物）とに分割した。そして、Ｊｍｏｌ（http://jmol.sourceforge.net/）を利用し、学習用の化合物群のＳＤＦファイルに基づいて化合物の構造モデルを生成し、各構造モデルに対し、Ｘ軸、Ｙ軸、Ｚ軸それぞれを中心に４５度刻みで回転させて撮像した５１２枚の撮像画像（スナップショット、サイズ：５１２×５１２、２４ｂｐｐ）を生成するプログラム（学習モデル入力データ生成プログラム）を実行し、各化合物についての撮像画像を生成した。各化合物の撮像画像は、当該化合物のＡＭＥＳ試験の結果が陽性であったか陰性であったかに応じた所定のフォルダに格納し、Ｄｉｇｉｔｓ（ＮＶＩＤＩＡ社）を用いて未改変のAlexNet（トロント大学）を学習させた。学習では、Ｄｉｇｉｔｓの設定を、学習率＝０．００１、ｅｐｏｃｈ＝１０とした。

【００７３】
続いて、外部検証法によって予測性能を確認した。具体的には、前記プログラムを実行し、外部検証用の化合物群のＳＤＦファイルを入力し、各化合物についての撮像画像を生成した。各化合物の撮像画像を、学習済みのAlexNetに入力し、１分子当たり５１２画像の陽性確率予測結果の平均値を算出した。すなわち、２００分子に対して化合物毎の陽性確率平均値を算出した。そして、上記文献から取得したＡＭＥＳ試験の実験結果（陽性または陰性）と、算出した化合物毎の陽性確率平均値を用いて、ＲＯＣ解析を行った。その結果を図８に示す。図８に示すように、ＲＯＣ曲線下面積（ＡＵＣ）は、０．８５７であった。

【００７４】
本実施例によって得られたＲＯＣ－ＡＵＣ値（0.８５７）は、現在使用されている記述子を用いた、一般的な機械学習によるＱＳＡＲ識別モデルと比較しても、本方法が良好な汎化性能を有していることを示している。例えば、ＡＭＥＳ試験のＱＳＡＲ解析による予測結果を、ＲＯＣ－ＡＵＣ値によって評価している近年の論文（Benchmark Data Set for in Silico Prediction of Ames Mutagenicity, J. Chem. Inf. Model., 2009, 49 (9), pp 2077-2081、In silico Prediction of Chemical Ames Mutagenicity, J. Chem. Inf. Model., 2012, 52 (11), pp 2840-2847）では、最良値として０．８６が報告されている。当該論文では、検証は５分割交差検証で行われており、５分割交差検証は外部検証と比較して過学習を引き起こす可能性が高く、一般に外部検証よりも良好な結果を与えることを考慮すれば、実施例３で得られたＡＵＣ値は、上記論文の最良値に匹敵している。

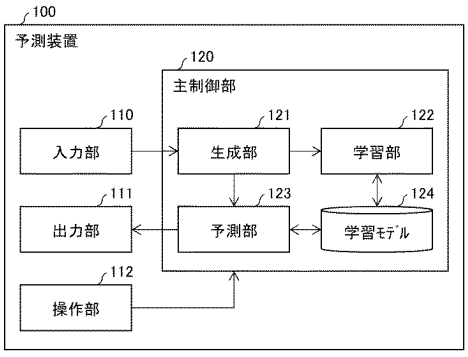
【産業上の利用可能性】
【００７５】
本発明は、化合物の毒性や活性等を予測するために利用することができる。

【符号の説明】
【００７６】

１０、２０：構造モデル　　２１：原子　　２２：結合　　２３：水素原子
１００：予測装置（学習モデル入力データ生成装置）　　１２１：生成部
１２２：学習部　　１２３：予測部　　１２４：学習モデル

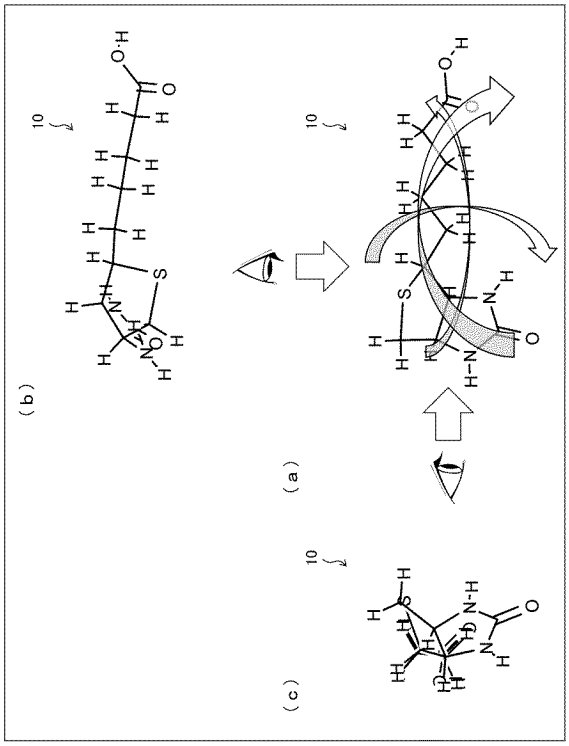
【図１】

図１



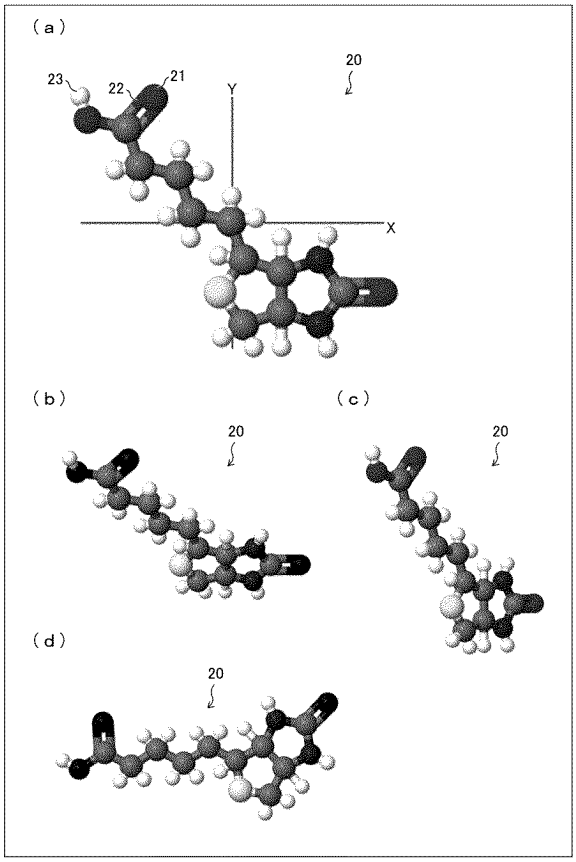
【図２】

図２



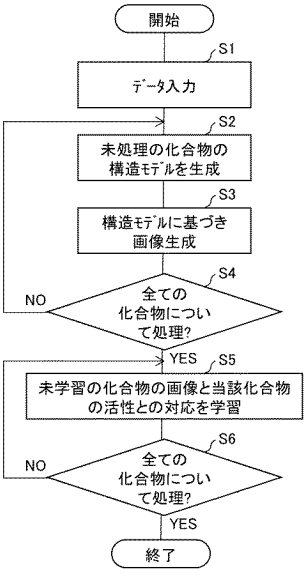
【図３】

図３



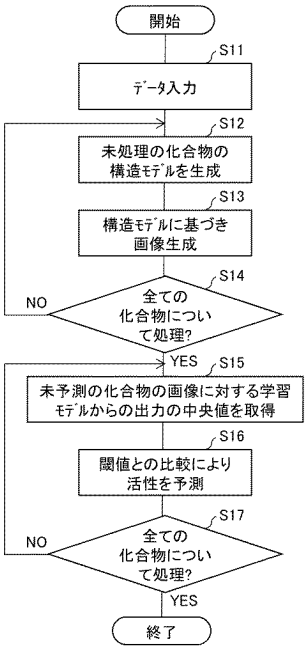
【図４】

図４



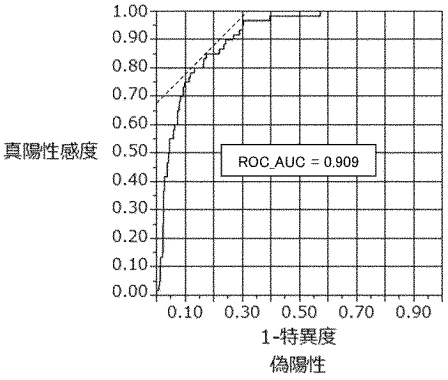
【図５】

図５



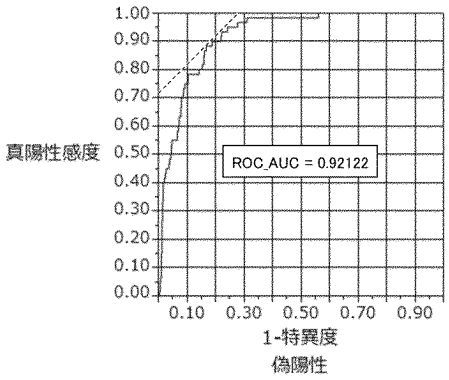
【図６】

図６



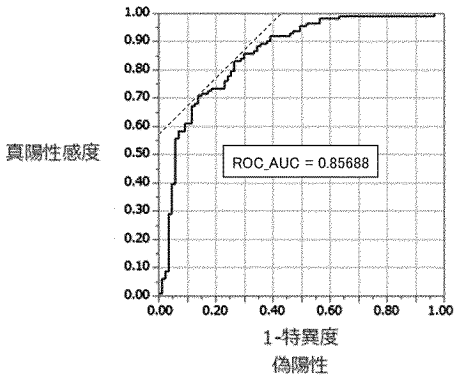
【図７】

図 7


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
真陽性感度
ROC_AUC = 0.92122
0.10
0.30
0.50
0.70
0.90
1-特異度
偽陽性

【図８】

図 8


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
真陽性感度
ROC_AUC = 0.85688
0.00
0.20
0.40
0.60
0.80
1.00
1-特異度
偽陽性