

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7201981号

(P7201981)

(45)発行日 令和5年1月11日(2023.1.11)

(24)登録日 令和4年12月27日(2022.12.27)

(51)国際特許分類

F I

G 1 6 C 20/30 (2019.01)

G 1 6 C 20/30

G 1 6 C 20/70 (2019.01)

G 1 6 C 20/70

C 1 2 M 1/34 (2006.01)

C 1 2 M 1/34

請求項の数 7 (全16頁)

(21)出願番号	特願2018-122565(P2018-122565)	(73)特許権者	505082350
(22)出願日	平成30年6月28日(2018.6.28)		学校法人 明治薬科大学
(65)公開番号	特開2019-10095(P2019-10095A)		東京都清瀬市野塩2丁目522番1
(43)公開日	平成31年1月24日(2019.1.24)	(74)代理人	100083116
審査請求日	令和3年6月17日(2021.6.17)		弁理士 松浦 憲三
(31)優先権主張番号	特願2017-129823(P2017-129823)	(74)代理人	100170069
(32)優先日	平成29年6月30日(2017.6.30)		弁理士 大原 一樹
(33)優先権主張国・地域又は機関	日本国(JP)	(74)代理人	100128635
			弁理士 松村 潔
		(74)代理人	100140992
			弁理士 松浦 憲政
		(74)代理人	110000338
			特許業務法人HARAKENZO WO
			RLD PATENT & TRADEM
			ARK

最終頁に続く

(54)【発明の名称】 予測装置、予測方法および予測プログラム

(57)【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、
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、

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、前記複数の画像を入力とする学習モデルを用いる予測装置。

【請求項2】

前記学習モデルは、深層学習を行う学習モデルを用いる請求項1に記載の予測装置。

10

【請求項3】

前記生成部は、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、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画像を生成する生成ステップと、

20

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、を包含し、

前記予測ステップは、少なくとも、機械学習を行う学習モデルであって、前記複数の画像を入力とする学習モデルを用いるステップを包含する予測方法。

【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】

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、
前記対象化合物の構造モデルに対して複数の方向からの画像を入力し、深層学習を行って特徴ベクトルを出力する第 1 の学習モデルと、
前記特徴ベクトルを入力し、前記対象化合物の活性の予測情報を出力する第 2 の学習モデルとを、組み合わせること、
を備えている、予測装置。

【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 (Q S A R : 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】

【文献】米国特許第 7 7 0 2 4 6 7 号明細書

米国特許第 7 7 5 1 9 8 8 号明細書

米国特許出願公開第 2 0 0 4 / 0 0 0 9 5 3 6 号明細書

米国特許出願公開第 2 0 0 4 / 0 1 9 9 3 3 4 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

10

20

30

40

50

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。

10

【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。

20

【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。

30

【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。

40

【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。

50

【 0 0 2 0 】

(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、P u b C h e m (<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。

【 0 0 2 1 】

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。

10

【 0 0 2 2 】

(出力部)

出力部 1 1 1 は、予測部 1 2 3 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 1 1 1 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。

【 0 0 2 3 】

(操作部)

操作部 1 1 2 は、予測装置 1 0 0 に対するユーザの操作を受け付ける。操作部 1 1 2 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。

20

【 0 0 2 4 】

(主制御部)

主制御部 1 2 0 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 1 2 0 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 1 2 0 の機能を分担するものであってもよい。

【 0 0 2 5 】

(学習モデル)

学習モデル 1 2 4 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。

30

【 0 0 2 6 】

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。

40

【 0 0 2 7 】

また、一態様において、学習モデル 1 2 4 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 1 2 4 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。

50

【 0 0 2 8 】

(生成部)

生成部 1 2 1 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 1 2 1 は、仮想空間に配置した対象化合物の構造モデル 1 0 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 1 2 1 は、構造モデル 1 0 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 1 0 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。

10

【 0 0 2 9 】

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.mass.edu/microbio/rasmol/>)、PyMOL (<http://www.pymol.org/>) 等) を用いることができる。

20

【 0 0 3 0 】

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 1 2 1 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、1 2 8 画素 × 1 2 8 画素、2 5 6 画素 × 2 5 6 画素、5 1 2 画素 × 5 1 2 画素、1 0 2 4 画素 × 1 0 2 4 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 ~ 6 4 b p p の範囲とすることができるが、好ましくは、8 ~ 3 2 b p p の範囲であり得る。

【 0 0 3 1 】

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 2 0 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。

30

【 0 0 3 2 】

図 3 の (a) に示すように、構造モデル 2 0 には、原子 2 1、結合 2 2 および水素原子 2 3 が含まれている。なお、原子 2 1 は、水素原子以外の原子を示す。一態様において、水素原子 2 3 は、構造モデル 2 0 に含めなくともよい。構造モデル 2 0 では、原子 2 1 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 2 1 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 2 1 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。

40

【 0 0 3 3 】

また、原子 2 1 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 5 0 % 以下、4 0 % 以下、3 0 % 以下、2 0 % 以下、1 0 % 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 3 0 % 以下とすることが好ましく、0 . 1 % 以上 1 0 % 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。

【 0 0 3 4 】

50

また、結合 2 2 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 1 2 1 は、仮想カメラを、構造モデル 2 0 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 2 0 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 2 0 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 2 0 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 2 0 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 2 0 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が、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 $4 \times 4 = 16$ 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 $8 \times 8 = 64$ 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 2 0 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 1 2 2 は、公知の方法により、生成部 1 2 1 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 1 2 4 に学習させる。一態様において、学習部 1 2 2 は、公知の深層学習アルゴリズムを用いて、学習モデル 1 2 4 に、生成部 1 2 1 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 1 2 2 は、例えば、D i g i t s (N V I D I A 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 1 2 3 は、生成部 1 2 1 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 1 2 4 を用いて、生成部 1 2 1 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 1 2 3 は、例えば、D i g i t s (N V I D I A 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 1 2 4 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 1 2 3 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 1 2 4 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 1 2 4 に対し

10

20

30

40

50

、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。

10

【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。

20

【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。

30

【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。

40

【0045】

50

次に、予測部 1 2 3 が、学習モデル 1 2 4 に対して、ステップ S 1 3 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 1 2 4 からの出力を取得する。一実施形態において、学習モデル 1 2 4 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 1 2 3 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 1 2 4 からの出力値の中央値を取得する（ステップ S 1 5）。そして、予測部 1 2 3 は、ステップ S 1 5 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 S 1 6）。ステップ S 1 5 ~ S 1 6 が、ステップ S 1 1 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 S 1 7 の y e s）には、予測処理を終了し、完了していない場合（ステップ S 1 7 の n o）には、ステップ S 1 5 に戻る。

10

【 0 0 4 6 】

以上により、予測装置 1 0 0 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。

【 0 0 4 7 】

（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。

【 0 0 4 8 】

特に、本実施形態によれば、学習モデル 1 2 4 に対する入力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 1 2 4 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。

20

【 0 0 4 9 】

また、学習モデル 1 2 4 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 1 0 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が困難である。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。

30

【 0 0 5 0 】

また、本実施形態によれば、学習モデル 1 2 4 に対する入力画像が、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 1 2 4 に対する入力画像が、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。

40

【 0 0 5 1 】

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。

【 0 0 5 2 】

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。

【 0 0 5 3 】

50

〔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。

【 0 0 5 9 】

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。

【 0 0 6 0 】

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。

【 0 0 6 1 】

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 (2 1) 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。

【 0 0 6 2 】

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。

【 0 0 6 3 】

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。

【 0 0 6 4 】

本発明の態様 7 に係る学習モデル入力データ生成装置 (1 0 0) 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 (1 2 4) 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 (1 2 1) 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。

【 0 0 6 5 】

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。

【 0 0 6 6 】

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。

【 0 0 6 7 】

〔 実施例 1 〕

Tox21DataChallenge2014 のサイト (<https://tripod.nih.gov/tox21/challenge/>)

10

20

30

40

50

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、24 b p p) 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。

10

【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。

20

【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。

30

【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、24 b p p) 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。

40

50

【 0 0 7 3 】

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 S D F 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 5 1 2 画像の陽性確率予測結果の平均値を算出した。すなわち、2 0 0 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 A M E S 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、R O C 解析を行った。その結果を図 8 に示す。図 8 に示すように、R O C 曲線下面積（A U C）は、0 . 8 5 7 であった。

【 0 0 7 4 】

本実施例によって得られた R O C - A U C 値（0 . 8 5 7）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 Q S A R 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、A M E S 試験の Q S A R 解析による予測結果を、R O C - A U C 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 . 8 6 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 A U C 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】

【 0 0 7 5 】

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】

【 0 0 7 6 】

1 0、2 0：構造モデル 2 1：原子 2 2：結合 2 3：水素原子
1 0 0：予測装置（学習モデル入力データ生成装置） 1 2 1：生成部
1 2 2：学習部 1 2 3：予測部 1 2 4：学習モデル

10

20

30

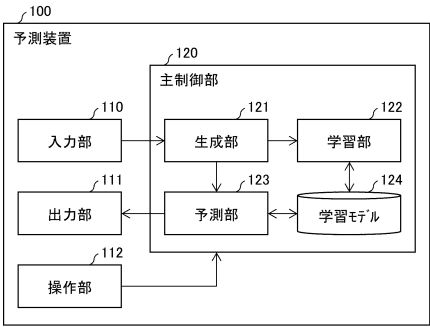
40

50

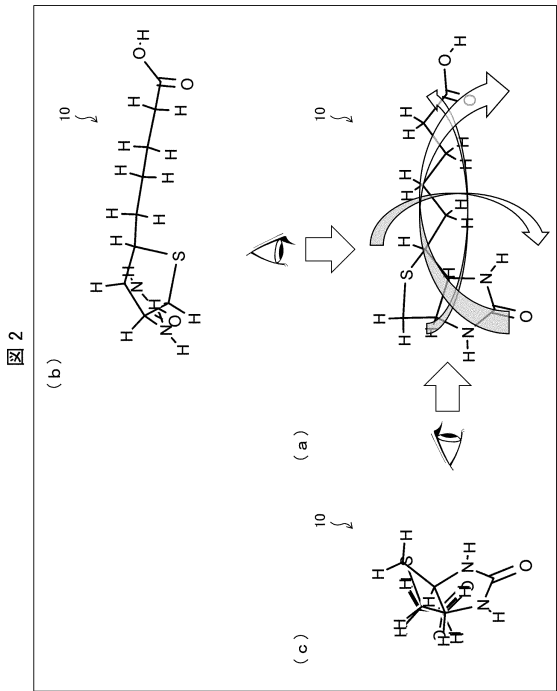
【図面】

【図 1】

図 1

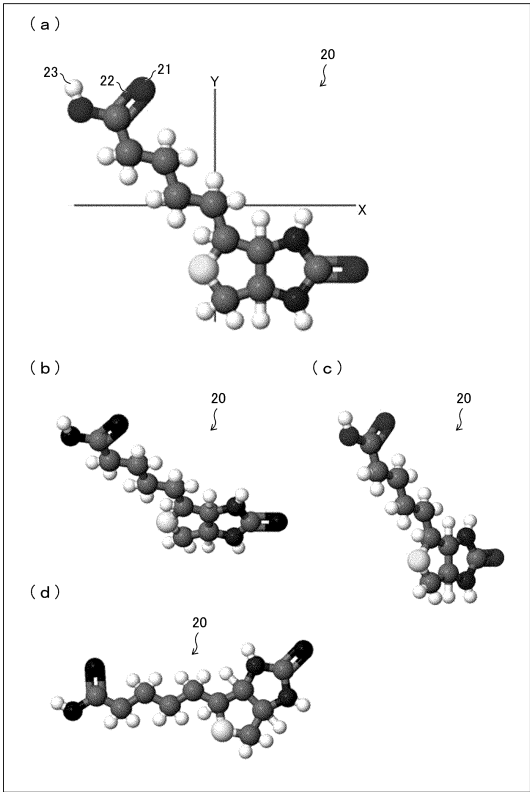


【図 2】



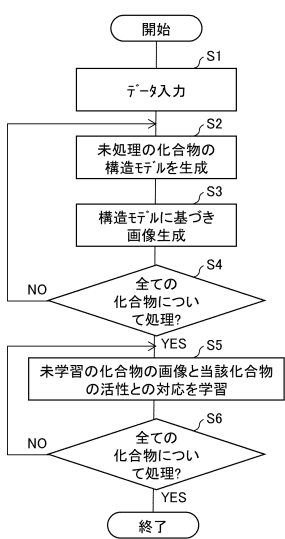
【図 3】

図 3



【図 4】

図 4



10

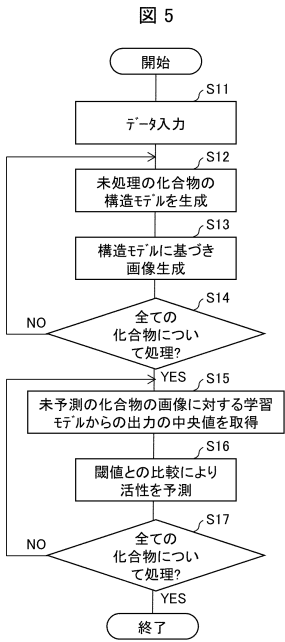
20

30

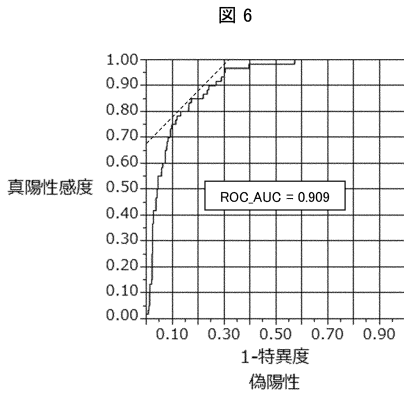
40

50

【 図 5 】

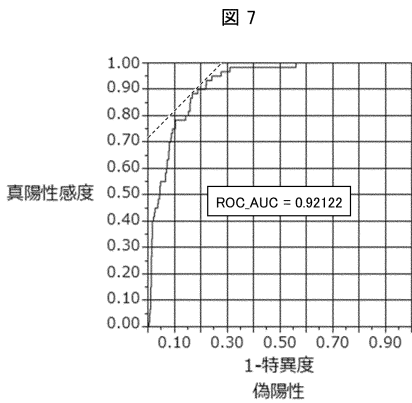


【 図 6 】

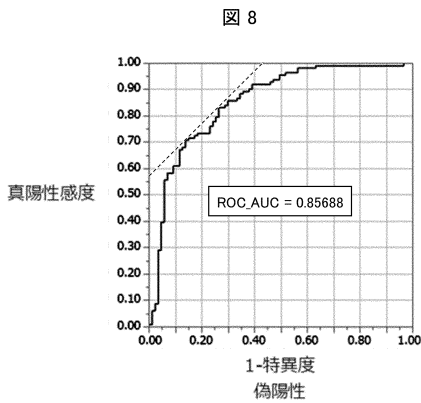


10

【 図 7 】



【 図 8 】



20

30

40

50

フロントページの続き

(72)発明者 植沢 芳広

東京都清瀬市野塩二丁目5-2番1 学校法人明治薬科大学内

審査官 関 博文

(56)参考文献 特開2005-250721(JP, A)

米国特許第05526281(US, A)

MA, Junshui, Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships, JOURNAL OF CHEMICAL INFORMATION AND MODELING, 米国, American Chemical Society, 2015年01月30日, Volume 55, p.263-274

TONG, Weida, Evaluation of Quantitative Structure-Activity Relationship Methods for Large-Scale Prediction of Chemicals, JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES, 米国, American Chemical Society, 1998年05月20日, Volume 38, p.669-677

NETZEVA, Tatiana I, Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity, The Report and Recommendations of ECVAM Workshop 52, イタリア, 2005年, p.1-19

(58)調査した分野 (Int.Cl., DB名)

G16C 10/00 - 99/00

C12M 1/34