

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7390250号
(P7390250)

(45)発行日 令和5年12月1日(2023.12.1)

(24)登録日 令和5年11月22日(2023.11.22)

(51)国際特許分類

F I

G 1 6 C 20/70 (2019.01)

G 1 6 C 20/70

G 0 6 N 20/00 (2019.01)

G 0 6 N 20/00

請求項の数 10 (全32頁)

(21)出願番号	特願2020-79790(P2020-79790)	(73)特許権者	000005108
(22)出願日	令和2年4月28日(2020.4.28)		株式会社日立製作所
(65)公開番号	特開2021-174401(P2021-174401 A)	(74)代理人	東京都千代田区丸の内一丁目6番6号 110001678
(43)公開日	令和3年11月1日(2021.11.1)		藤央弁理士法人
審査請求日	令和4年11月7日(2022.11.7)	(72)発明者	刑部 好弘
			東京都千代田区丸の内一丁目6番6号
		(72)発明者	株式会社日立製作所内
			浅原 彰規
			東京都千代田区丸の内一丁目6番6号
			株式会社日立製作所内
		審査官	塩田 徳彦

最終頁に続く

(54)【発明の名称】 化合物構造表現を生成するシステム

(57)【特許請求の範囲】

【請求項1】

化合物構造表現を生成するシステムであって、
1以上のプロセッサと、
1以上の記憶装置と、を含み、
前記1以上の記憶装置は、構造モデルと、構造物性関係モデルと、1以上の既知物質の化合物構造表現と、1種以上の物性値それぞれの1以上の目標値と、を格納し、
前記構造モデルは、
化合物構造表現を実数ベクトルに変換する、第1エンコーダと、
前記第1エンコーダにより変換された前記実数ベクトルから前記化合物構造表現を推定する、第1デコーダと、を含み、
前記構造物性関係モデルは、
前記第1エンコーダにより生成された前記実数ベクトルと前記1種以上の物性値の目標値を含む目標値ベクトルとを構成要素として含む入力ベクトルを、実数ベクトルに変換する、第2エンコーダと、
前記第2エンコーダにより生成された前記実数ベクトルから前記入力ベクトルを推定する第2デコーダと、を含み、
前記1以上のプロセッサは、前記1以上の既知物質の化合物構造表現及び前記1種以上の物性値それぞれの1以上の目標値に基づき、前記構造モデルの前記第1エンコーダを使用して、1以上の構造発生ベクトルを生成し、

10

20

前記 1 以上の構造発生ベクトルのそれぞれは、一つの既知物質の化合物構造表現の前記第 1 エンコーダにより生成された実数ベクトルと、前記 1 種以上の物性値それぞれの目標値を含む目標値ベクトルとを、構成要素として含み、

前記 1 以上のプロセッサは、

前記 1 以上の構造発生ベクトルの各構造発生ベクトルを、前記構造物性関係モデルに入力し、

前記構造物性関係モデルの前記第 2 デコーダの出力から、化合物構造表現に対応する実数ベクトルを抽出し、

前記抽出した実数ベクトルを前記構造モデルの前記第 1 デコーダに入力して、新たな化合物構造表現を生成する、システム。

10

【請求項 2】

請求項 1 に記載のシステムであって、

前記 1 以上の記憶装置は、

前記 1 種以上の物性値の測定値と関連づけられた第 1 化合物構造表現群を格納し、

前記 1 以上のプロセッサは、前記第 1 化合物構造表現群によって、前記構造モデルの学習を行い、

前記 1 種以上の物性値の測定値及び前記構造モデルによる前記第 1 化合物構造表現群の変換結果を使用して、前記構造物性関係モデルの学習を行う、システム。

【請求項 3】

請求項 1 に記載のシステムであって、

前記 1 以上の記憶装置は、

前記 1 種以上の物性値の測定値と関連づけられた第 1 化合物構造表現群と、

前記 1 種以上の物性値の測定値と関連付けのない第 2 化合物構造表現群と、を格納し、

前記 1 以上のプロセッサは、前記第 1 化合物構造表現群及び前記第 2 化合物構造表現群を使用して、前記構造モデルの学習を行い、

前記 1 種以上の物性値の測定値及び前記第 1 エンコーダによる前記第 1 化合物構造表現群の変換結果を使用して、前記構造物性関係モデルの学習を行う、システム。

20

【請求項 4】

請求項 1 に記載のシステムであって、

前記 1 以上の記憶装置は、前記 1 種以上の物性値それぞれの 1 以上の目標値に加えて、他の 1 種以上の物性値それぞれの 1 以上の目標値を含み、

前記構造物性関係モデルは、複数のオートエンコーダを含み、

前記複数のオートエンコーダの第 1 オートエンコーダは、前記第 2 エンコーダ及び前記第 2 デコーダで構成され、

前記複数のオートエンコーダにおいて、前記第 1 オートエンコーダ以外の各エンコーダは、他の一つのオートエンコーダのエンコーダとデコーダとの間に挟まれており、

前記第 1 オートエンコーダ以外の各オートエンコーダの入力は、前記他の一つのオートエンコーダのデコーダからの実数ベクトルと、前記他の 1 種以上の物性値から選択された 1 種以上の物性値それぞれの目標値を含む目標値ベクトルと、を構成要素として含む、システム。

30

40

【請求項 5】

請求項 4 に記載のシステムであって、

前記複数のオートエンコーダの間において、入力される目標値の物性値の種類は異なり、

前記複数のオートエンコーダそれぞれに入力される目標値の数は 1 である、システム。

【請求項 6】

請求項 4 に記載のシステムであって、

前記複数のオートエンコーダの間において、入力される目標値の物性値の種類は異なり、

前記複数のオートエンコーダは、入力される目標値の数が異なるオートエンコーダを含む、システム。

【請求項 7】

50

請求項 4 に記載のシステムであって、
前記 1 以上の記憶装置は、前記構造物性関係モデルの学習データを格納し、
前記 1 以上のプロセッサは、前記学習データを使用して、前記構造物性関係モデルの学習を行い、
前記学習データは、前記複数のオートエンコーダのそれぞれの学習に使用される複数のグループで構成され、
前記複数のグループの各グループは、複数の化合物構造表現のそれぞれを、1 以上の所定の物性値種類の測定値と関連付け、
前記複数のグループの二つのグループの間において、物性値種類数が多いグループは、物性値種類数が少ないグループの全ての物性値種類及び全ての化合物構造表現を含み、
前記物性値種類数が多いグループは、前記構造物性関係モデルにおけるより内側のオートエンコーダの学習に使用される、システム。

10

【請求項 8】

請求項 4 に記載のシステムであって、
前記複数のオートエンコーダにおいて、より外側のオートエンコーダの学習データ量が、より内側のオートエンコーダの学習データ量より多い、システム。

【請求項 9】

請求項 1 に記載のシステムであって、
前記 1 以上のプロセッサは、
前記構造モデル及び前記構造物性関係モデルを含むネットワーク構造を表示し、
前記ネットワーク構造に対するユーザ入力に応じて前記ネットワーク構造を修正する、システム。

20

【請求項 10】

システムが化合物構造表現を生成する方法であって、
前記システムは、
1 以上のプロセッサと、
1 以上の記憶装置と、を含み、
前記 1 以上の記憶装置は、構造モデルと、構造物性関係モデルと、1 以上の既知物質の化合物構造表現と、1 種以上の物性値それぞれの 1 以上の目標値と、を格納し、
前記構造モデルは、
化合物構造表現を実数ベクトルに変換する、第 1 エンコーダと、
前記第 1 エンコーダにより変換された前記実数ベクトルから前記化合物構造表現を推定する、第 1 デコーダと、を含み、
前記構造物性関係モデルは、
前記第 1 エンコーダにより生成された前記実数ベクトルと前記 1 種以上の物性値の目標値を含む目標値ベクトルとを構成要素として含む入力ベクトルを、実数ベクトルに変換する、第 2 エンコーダと、
前記第 2 エンコーダにより生成された前記実数ベクトルから前記入力ベクトルを推定する第 2 デコーダと、を含み、
前記方法は、前記 1 以上のプロセッサが、前記 1 以上の既知物質の化合物構造表現及び前記 1 種以上の物性値それぞれの 1 以上の目標値に基づき、前記構造モデルの前記第 1 エンコーダを使用して、1 以上の構造発生ベクトルを生成する、ことを含み、
前記 1 以上の構造発生ベクトルのそれぞれは、一つの既知物質の化合物構造表現の前記第 1 エンコーダにより生成された実数ベクトルと、前記 1 種以上の物性値それぞれの目標値を含む目標値ベクトルとを、構成要素として含み、
前記方法は、前記 1 以上のプロセッサが、
前記 1 以上の構造発生ベクトルの各構造発生ベクトルを、前記構造物性関係モデルに入力し、
前記構造物性関係モデルの前記第 2 デコーダの出力から、化合物構造表現に対応する実数ベクトルを抽出し、

30

40

50

前記抽出した実数ベクトルを前記構造モデルの前記第 1 デコーダに入力して、新たな化合物構造表現を生成する、ことを含む、方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、所望の物性値を持つことが期待される化合物構造表現の候補を生成するシステムに関する。

【背景技術】

【0002】

新材料探索タスクに対してバーチャルスクリーニングの手法が利用されている。バーチャルスクリーニングの手法の例は、例えば、非特許文献 1 に開示されている。バーチャルスクリーニングは、既知の化合物のデータに機械学習モデルを適用し、所定の表現形式で表された化学構造式を入力とした物性値推定モデルを構成する。次に、ランダムに生成した化学構造式に対して、上記物性値推定モデルを適用する。こうして計算された予測値をもとにスクリーニングを行い、閾値を超える物性値を持つと期待される化学構造式を候補として提示する。

10

【0003】

他の先行技術文献である非特許文献 2 は、画像分類タスクを実行する、スタックされた半教師あり学習モデルを開示する。非特許文献 2 は、スタックされたモデルにおける外側のモデルの学習をラベルなし学習データで行い、内側のモデルの学習をラベル付き学習データで行うことを開示する。

20

【先行技術文献】

【非特許文献】

【0004】

【文献】R. Gomez-Bombarelli et al., "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules," ACS Cent. Sci., vol. 4, no. 2, pp. 268-276, Feb. 2018.

【文献】D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised Learning with Deep Generative Models," NIPS 2014.

【発明の概要】

30

【発明が解決しようとする課題】

【0005】

従来の物性値推定モデルを使用した探索手法は、学習データの範囲内でしか推定することができない内挿的探索法であるため、既知材料の性能を超える物性値を持つ新材料を発見することを目的とする外挿的探索には不適切である。

【0006】

また、従来のバーチャルスクリーニング手法は、SMILES (Simplified Molecular Input Line Entry System) のような化学構造式の表現形式と物性値との関係を、ニューラルネットワークのようなモデルによって獲得する。これにより、所望の物性値を持つ化学構造式を生成することを目的とする。しかし、化学構造式の表現形式の文法規則と、化学構造式と物性値との間の関係との双方を学習により獲得するためには、化学構造式と物性値とがセットになった大量のデータが必要となる。しかし、実験データやシミュレーション結果のような、化学構造式と物性値がセットになったデータを大量に用意するのは困難である。

40

【0007】

したがって、学習データに含まれる物性値よりも良い物性値を持つ化合物の候補を提示することができる学習モデルを、少ない化学構造式と物性値がセットになったデータを使用して生成することができる技術が望まれる。

【課題を解決するための手段】

【0008】

50

本発明の一態様は、化合物構造表現を生成するシステムであって、1以上のプロセッサと、前記1以上のプロセッサが実行するプログラムを格納する1以上の記憶装置と、を含む。前記記憶装置は、構造モデルと、構造物性関係モデルと、1以上の既知物質の化合物構造表現と、1種以上の物性値それぞれの1以上の目標値と、を格納する。前記構造モデルは、化合物構造表現を実数ベクトルに変換する、第1エンコーダと、前記第1エンコーダにより変換された前記実数ベクトルから前記化合物構造表現を推定する、第1デコーダと、を含む。前記構造物性関係モデルは、前記第1エンコーダにより生成された前記実数ベクトルと前記1種以上の物性値の目標値を含む目標値ベクトルとを構成要素として含む入力ベクトルを、実数ベクトルに変換する、第2エンコーダと、前記第2エンコーダにより生成された前記実数ベクトルから前記入力ベクトルを推定する第2デコーダと、を含む。前記1以上のプロセッサは、前記1以上の既知物質の化合物構造表現及び前記1種以上の物性値それぞれの1以上の目標値に基づき、前記構造モデルの前記第1エンコーダを使用して、1以上の構造発生ベクトルを生成する。前記1以上の構造発生ベクトルのそれぞれは、一つの既知物質の化合物構造表現の前記第1エンコーダにより生成された実数ベクトルと、前記1種以上の物性値それぞれの目標値を含む目標値ベクトルとを、構成要素として含む。前記1以上のプロセッサは、前記1以上の構造発生ベクトルの各構造発生ベクトルを、前記構造物性関係モデルに入力する。前記1以上のプロセッサは、前記構造物性関係モデルの前記第2デコーダの出力から、化合物構造表現に対応する実数ベクトルを抽出する。前記1以上のプロセッサは、前記抽出した実数ベクトルを前記構造モデルの前記第1デコーダに入力して、新たな化合物構造表現を生成する。

10

20

【発明の効果】

【0009】

本発明の一態様によれば、学習データに含まれる物性値よりも良い物性値を持つ化合物の候補を提示することができる学習モデルを、少ない化学構造式と物性値がセットになったデータを使用して生成できる。

【図面の簡単な説明】

【0010】

【図1】本明細書の実施例に係る化学構造式生成モデルの構成例を模式的に示す。

【図2】実施例1に係る化学構造式生成システムの構成の一例を示す。

【図3】構造式生成装置のハードウェア構成例を示す。

30

【図4】カタログデータの構成例を示す。

【図5】実験データの構成例を示す。

【図6】構造式行列データベースに含まれるデータの例を示す。

【図7】学習用データベースに含まれるデータの例を示す。

【図8】初期パラメータに含まれる情報の例を示す。

【図9】表示装置において、表示部がユーザのために表示するネットワーク構造確認画面の例を示す。

【図10】モデルデータに含まれるモデルテーブルの構成例を示す。

【図11】構造式変換部の処理例のフローチャートを示す。

【図12】学習データ生成部の処理例のフローチャートを示す。

40

【図13】構造発生ベクトル群生成部の処理例のフローチャートを示す。

【図14】ネットワーク構造決定部の処理例のフローチャートを示す。

【図15】構造モデル学習部の処理例のフローチャートを示す。

【図16】構造モデル追加学習部の処理例のフローチャートを示す。

【図17】構造物性関係モデル学習部の処理例のフローチャートを示す。

【図18】新構造式生成部の処理例のフローチャートを示す。

【図19】構造式逆変換部の処理例のフローチャートを示す。

【図20】構造式整形部の処理例のフローチャートを示す。

【図21】実施例2に係る、表示部が表示装置においてユーザのために表示する、ネットワーク構造確認画面の例を示す。

50

【図 2 2】実施例 2 に係る、学習用データベースに含まれるデータの例を示す。

【図 2 3】実施例 2 に係る、モデルデータに含まれるモデルテーブルの構成例を示す。

【図 2 4】実施例 3 に係る、表示部が表示装置においてユーザのために表示する、ネットワーク構造確認画面の例を示す。

【図 2 5】実施例 3 に係る、学習用データベース DB に含まれるデータの例を示す。

【図 2 6】実施例 3 に係る、モデルデータ DB に含まれるモデルテーブルの構成例を示す。

【図 2 7】実験データにおけるレコードが含む物性値（の種類）が満たすべき条件を模式的に示す。

【図 2 8】2 種類の実験データで構成される構造物性関係モデル用学習データの例を模式的に示す。

【図 2 9】実施例 4 に係る、表示部が表示装置においてユーザのために表示する、ネットワーク構造確認画面の例を示す。

【図 3 0】実施例 3 に係る、モデルデータ DB に含まれるモデルテーブルの構成例を示す。

【図 3 1】学習データ生成部が、実験データから構造物性関係モデル用学習データを生成する例を示している。

【発明を実施するための形態】

【0011】

以下においては、便宜上その必要があるときは、複数のセクションまたは実施例に分割して説明するが、特に明示した場合を除き、それらは互いに無関係なものではなく、一方は他方の一部または全部の変形例、詳細、補足説明等の関係にある。また、以下において、要素の数等（個数、数値、量、範囲等を含む）に言及する場合、特に明示した場合及び原理的に明らかに特定の数に限定される場合等を除き、その特定の数に限定されるものではなく、特定の数以上でも以下でもよい。

【0012】

本システムは、物理的な計算機システム（一つ以上の物理的な計算機）でもよいし、クラウド基盤のような計算リソース群（複数の計算リソース）上に構築されたシステムでもよい。計算機システムあるいは計算リソース群は、1 以上のインタフェース装置（例えば通信装置及び入出力装置を含む）、1 以上の記憶装置（例えば、メモリ（主記憶）及び補助記憶装置を含む）、及び、1 以上のプロセッサを含む。

【0013】

プログラムがプロセッサによって実行されることで機能が実現される場合、定められた処理が、適宜に記憶装置及び/またはインタフェース装置等を用いながら行われるため、機能はプロセッサの少なくとも一部とされてもよい。機能を主語として説明された処理は、プロセッサあるいはそのプロセッサを有するシステムが行う処理としてもよい。プログラムは、プログラムソースからインストールされてもよい。プログラムソースは、例えば、プログラム配布計算機または計算機が読み取り可能な記憶媒体（例えば計算機読み取り可能な非一過性記憶媒体）であってもよい。各機能の説明は一例であり、複数の機能が一つの機能にまとめられたり、一つの機能が複数の機能に分割されたりしてもよい。

【0014】

〔概略〕

以下において、所望の物性値を有することが期待される化学構造式を推定する技術が開示される。化学構造式は、様々は表現形式で表され得る。化学構造式の表現形式は、例えば、一定の文法規則に従って記述された文字列や、行列であり得る。文法規則の例は、SMILES (Simplified Molecular Input Line Entry System) である。以下に説明する実施例は、化学構造式を記述する文法規則の一例として、SMILESを使用する。

【0015】

図 1 は、本明細書の実施例に係る化学構造式生成モデルの構成例を模式的に示す。化学構造式生成モデル 10 は、既知の化学構造式を表す表現と目標物性値を入力として受け付け、目標物性値に近い物性値を有することが期待される新たな化学構造式の表現を出力す

10

20

30

40

50

る。

【0016】

化学構造式生成モデル10は、化学構造式を学習対象とする構造モデル100と、化学構造式の特徴量と物性値との間の関係を学習する構造物性関係モデル104の、2種類のモデルを組み合わせる。構造モデル100は、一つの変分オートエンコーダ(VAE)で構成され、構造物性関係モデル104は1以上のVAEで構成される。図1の構成例において、構造物性関係モデル104は、単一のVAEで構成されている。

【0017】

VAEは、オートエンコーダの一種であり、エンコーダとデコーダの二つのニューラルネットワークから構成される深層生成モデルである。エンコーダは、入力(ベクトル)を実数ベクトルに変換する。実数ベクトルが属する空間は潜在空間と呼ばれ、所定の分布、例えば正規分布に従うことが仮定される。デコーダは、その実数ベクトルを逆変換し、入力と等しい次元のベクトルを出力する。

10

【0018】

エンコーダ及びデコーダは、入力と出力とが等しくなるように訓練される(学習する)。中間出力の実数ベクトルから入力を再構成できることは、入力の十分な特徴が実数ベクトルに反映されていることを意味する。潜在空間の次元は、入力の次元よりも小さくなるように設定される。そのため、エンコーダは、入力の特徴量を抽出すると共に、入力の次元を圧縮することができる。

【0019】

中間出力のベクトルは、潜在変数又は潜在表現と呼ばれ、化学構造式を表す構造式行列から抽出した特徴を表す抽象表現である。構造式行列は、例えば、材料の化学構造式を表す文字列から変換できる。潜在変数は、所定の分布、例えばガウス分布に従うことが仮定される。したがって、デコーダは、ノイズが加えられたベクトルを受け取った場合に、入力の構造式行列を高い精度で復元することができる。このように、VAEは、生成モデルとして高いロバスト性を有する。

20

【0020】

図1に示すように、化学構造式生成モデル10は入れ子構造を有する。具体的には、構造モデル(外側VAE)100のエンコーダ101とデコーダ102との間に構造物性関係モデル104が配置される。図1の例において、構造物性関係モデル104は、エンコーダ105及びデコーダ106で構成される。後述するように、構造物性関係モデル104は複数のVAE(内側VAE)を含むことができ、各内側VAEは、他のVAEのエンコーダとデコーダとの間に配置される。

30

【0021】

構造モデル100のエンコーダ101は、例えば、複数の1次元畳み込み層と、複数の全結合層で構成できる。エンコーダ101は、 $M \times N$ 次元の構造行列式(構造表現)を入力として受け取り、 L 次元ベクトルに変換する。デコーダ102は、例えば、複数の全結合層とRNN(Recurrent Neural Network)で構成できる。デコーダ102は、 L 次元ベクトルを入力として受け取り、 $M \times N$ 次元の構造行列式に逆変換する。

40

【0022】

構造物性関係モデル104の内側VAEのエンコーダ105及びデコーダ106は、例えば、複数の全結合層で構成することができる。エンコーダ105は、構造モデル100(外側VAE)の潜在変数(変換結果)である L 次元ベクトルと、 P 個の物性値の配列からなる P 次元ベクトル(目標値ベクトル)を構成要素として含む、 $L + P$ 次元ベクトル107(拡張ベクトル又は構造発生ベクトルとも呼ぶ)を入力として受け取る。エンコーダ105は、ベクトル107より小さい次元の中間ベクトル(潜在表現)108を出力する。構造物性関係モデル104の潜在空間は、構造的特徴と物性値的特徴の組み合わせが抽象化する潜在表現を与える。

【0023】

50

デコーダ 106 は、中間ベクトル 108 を入力として受け取り、 $L + P$ 次元ベクトル 109 を出力する。 $L + P$ 次元ベクトル 109 から抽出される P 個の要素は物性値の配列である。デコーダ 106 の出力から抽出される L 次元ベクトルは、構造モデル 100 のデコーダ 102 に入力され、 $M \times N$ 次元の構造行列式が出力される。

【0024】

例えば、システムは、既知化合物の化学構造式（の構造表現）と目標物性値を学習済みの化学構造式生成モデル 10 に順次入力する。これにより、目標物性値に近い値を示すと期待される新しい化学構造式を生成することができる。また、システムは、既知化合物のなかの性能上位の化合物の化学構造式それぞれと、所定の物性値及びその近傍の値（これらを目標値とも呼ぶ）それぞれとの組み合わせを、学習済みの化学構造式生成モデル 10 に順次入力してもよい。これにより、目標物性値に近い値を示すと新しい化学構造式が生成される蓋然性を高めることができる。

10

【0025】

本明細書の実施例が実行するタスクにおいて、化学構造式生成モデルが獲得すべき学習対象は、主に二つである。一つの学習対象は、化学構造式を表す表現形式の文法規則であり、もう一つの学習対象は、化学構造（化学構造式）と物性値との間の関係である。VAE の学習は、エンコーダの入力とデコーダの出力が等しくなるように損失関数を与え、エンコーダ及びデコーダのパラメータを更新（最適化）する。

【0026】

上記二つの学習対象において、物性値は、特徴と物性値との間の関係の学習においてのみ必要となる。物性値の種類は、物理的性質を表す種類と化学的性質を表す種類とを含む。いずれの種類の物性値も、主鎖構造や、末端構造、部分構造などの局所的な構造的特徴に強く影響を受ける。このため、化学構造と物性値との間の関係の学習は、化学構造式そのものに代えて、化学構造式から抽出した特徴量と物性値との間の関係を学習データとして利用することができる。

20

【0027】

本明細書の実施例は、主に以下のステップを実行する。まず、システムは、ユーザ設定とデータとを受け取って学習モデル（ネットワーク構造）を決定する。次に、システムは、化学構造式生成モデルの学習（訓練）を実行する。化学構造式生成モデルの学習は、構造モデル（外側 VAE）100 のカタログデータでの学習、構造モデル 100 の実験データの化学構造式での学習、そして構造物性関係モデル 104（1 以上の内側 VAE）の実験データでの学習を含む。システムは、学習済みの VAE から化学構造式生成モデルを構築し、新規構造式を生成する。

30

【実施例 1】

【0028】

図 2 に実施例 1 に係る化学構造式生成システムの構成の一例を示す。本システムは、ネットワークを介して互いに通信可能な、パラメータ設定装置 M01、データ保存装置 M02、モデル学習装置 M03、構造式生成装置 M04、及び表示装置 M05 を含む。

【0029】

パラメータ設定装置 M01 は、化学構造式生成システムの生成（学習を含む）のためのパラメータを含む、種々のデータを設定又は生成する。本例において、パラメータ設定装置 M01 は、構造式変換部 P01、学習データ生成部 P02、構造発生ベクトル群生成部 P03、及びネットワーク構造決定部 P04 を含む。これらはプログラムである。パラメータ設定装置 M01 は、さらに、カタログデータ DB10、実験データ DB11、構造式語彙データ DB12、及び初期パラメータ DB13 を格納する。

40

【0030】

データ保存装置 M02 は、他の装置が生成したデータ（情報）を含む、様々な種類のデータを格納することができる。本例において、データ保存装置 M02 は、構造式行列データベース DB14、学習用データベース DB15、構造発生ベクトルデータベース DB16、モデルデータ DB17、及び候補構造式データベース DB18 を格納している。

50

【 0 0 3 1 】

モデル学習装置 M 0 3 は、化学構造式生成システムに含まれる学習モデルの学習を行う。本例において、モデル学習装置 M 0 3 は、構造モデル学習部 P 0 5、構造モデル追加学習部 P 0 6、及び構造物性関係モデル学習部 P 0 7 を含む。これらはプログラムである。

【 0 0 3 2 】

構造式生成装置 M 0 4 は、学習済みの化学構造式生成モデルを使用して、所望の物性値を有することが期待される新たな物質の化学構造式を生成（推定）する。本例において、構造式生成装置 M 0 4 は、新構造式生成部 P 0 8、構造式逆変換部 P 0 9、および構造式整形部 P 1 0 を含む。これらはプログラムである。

【 0 0 3 3 】

表示装置 M 0 5 は、他の装置から取得した情報をユーザに対して提示すると共に、ユーザから入力データを受け取り、他の装置に送信することができる。表示装置 M 0 5 は、プログラムである、表示部 P 1 1 を含む。

【 0 0 3 4 】

図 3 は、構造式生成装置 M 0 4 のハードウェア構成例を示す。構造式生成装置 M 0 4 は、演算性能を有するプロセッサ U 1 1 1 と、プロセッサ U 1 1 1 が実行するプログラム及びデータを格納する揮発性一時記憶領域を与える D R A M U 1 1 2 と、を含む。

【 0 0 3 5 】

構造式生成装置 M 0 4 は、さらに、本システムにおける他の装置を含む他の装置とデータ通信をおこなう通信装置 U 1 1 3 と、H D D (H a r d D i s k D r i v e) やフラッシュメモリなどを利用した永続的な情報記憶領域を与える補助記憶装置 U 1 1 4 と、を含む。また、構造式生成装置 M 0 4 は、ユーザからの操作を受け付ける入力装置 U 1 1 5 と、各プロセスでの出力結果をユーザに提示するモニタ U 1 1 6 (出力装置の例) と、を含む。

【 0 0 3 6 】

例えば、補助記憶装置 U 1 1 4 は、新構造式生成部 P 0 8、構造式逆変換部 P 0 9、および構造式整形部 P 1 0 等のプログラムを格納する。プロセッサ U 1 1 1 が実行するプログラム及び処理対象のデータは、補助記憶装置 U 1 1 4 から D R A M U 1 1 2 にロードされる。

【 0 0 3 7 】

化学構造式生成システムに含まれる他の装置、具体的には、パラメータ設定装置 M 0 1、データ保存装置 M 0 2、モデル学習装置 M 0 3、及び表示装置 M 0 5 それぞれを構成するハードウェア要素は、構造式生成装置 M 0 4 と同様でよい。また、複数の装置に分かれている機能を一つの装置に統合してもよく、上記複数の装置機能をさらに多くの装置に分散してもよい。このように、化学構造式生成システムは、1 以上の記憶装置及び 1 以上のプロセッサを含む。

【 0 0 3 8 】

図 4 は、カタログデータ D B 1 0 の構成例を示す。カタログデータ D B 1 0 は、化合物構造式のデータベースであり、大量のレコードを含む。各レコードは、一つの化学構造式の情報を格納している。カタログデータ D B 1 0 のデータは、例えば、自由な二次利用が可能な状態で公開された、入手容易なオープンデータを含むことができる。本明細書の実施例において、化学構造式は、S M I L E S 記法に従った文字列（表現）で表される。

【 0 0 3 9 】

図 4 に示す、カタログデータ D B 1 0 の例において、T a b l e I D カラム T 0 C 1 は、そのテーブル（図 4 に示すテーブル）の識別子を示す。T a b l e T y p e カラム T 0 C 2 は、そのテーブルが格納しているデータのタイプを示す。T a b l e T y p e カラム T 0 C 2 は、本テーブルが、カタログデータのテーブルであることを示している。I D カラム T 0 C 3 は、化学構造式の識別子を示す。S M I L E S カラム T 0 C 4 は、化学構造式の S M I L E S 表現を示す。

【 0 0 4 0 】

10

20

30

40

50

図5は、実験データDB11の構成例を示す。実験データDB11は、化学構造式の注目している1以上の物性値を示す実験データを格納している。各レコードは、注目している1以上の物性値の実験結果と、一つの化学構造式の情報との組を含む。実験データDB11のレコード数は、カタログデータDB10のレコード数より少ないことが想定される。

【0041】

図5に示す実験データDB11の例において、TableIDカラムT1C1は、そのテーブル(図5に示すテーブル)の識別子を示す。TableTypeカラムT1C2は、そのテーブルが格納しているデータのタイプを示す。TableTypeカラムT1C2は、本テーブルが、実験データのテーブルであることを示している。

【0042】

IDカラムT1C3は、化学構造式の識別子を示す。SMILESカラムT1C4は、化学構造式のSMILES表現を示す。MWtカラムT1C5は、化学構造式で表される化合物の分子量を示す。logPカラムT1C6は、化学構造式で表される化合物の分配係数を示す。分子量及び分配係数は、化学構造式の物性値の例であり、実験データは、任意の物性値を含むことができる。

【0043】

図6は、構造式行列データベースDB14に含まれるデータの例を示す。構造式行列データベースDB14は複数のテーブルを含む。構造式行列データベースDB14は、カタログデータDB10や実験データDB11のテーブルそれぞれに対し、付与した元データIDと構造式(SMILES)を変換した構造式行列のカラムを追加したテーブルをまとめたものである。そのため、実施例1において、構造式行列データベースDB14は、二つのテーブルを含む。

【0044】

構造式行列データベースDB14は、構造式変換部P01によりSMILES表現から変換された化学構造式の行列を格納している。このように、本明細書の実施例において、化学構造式を表す文字列は、行列に変換される。行列の縦軸は元素記号のような記号種を示し、横軸は出現位置を示す。

【0045】

本明細書において、この行列を構造式行列と呼ぶ。記号種数をM、化学構造を表す文字列の長さをNとすると、構造式行列は、 $M \times N$ 次元となる。文字列の長さは、構造式に応じて変化し得る。そのため、負数やゼロ値によりパディングを行って、固定長行列を生成する。構造式行列は、どの位置にどの記号が出現するかについての情報を持つため、構造式が一意に決定され、構造式行列の逆変換により構造式を生成することができる。

【0046】

図6に例示するように、構造式行列データベースDB14は、カタログデータ構造式行列テーブル141及び実験データ構造式行列テーブル142を含む。カタログデータ構造式行列テーブル141は、カタログデータDB10から生成され、さらに構造式行列が追加されている。実験データ構造式行列テーブル142は、実験データDB11から生成され、さらに構造式行列が追加されている。

【0047】

カタログデータ構造式行列テーブル141において、TableIDカラムT3C1は、そのテーブル(図6に示すテーブル)の識別子を示す。TableTypeカラムT3C2は、そのテーブルが格納しているデータのタイプを示す。TableTypeカラムT3C2は、本テーブルが、カタログデータから生成されたテーブルであることを示している。IDカラムT3C3は、化学構造式の識別子を示す。SMILESカラムT3C4は、化学構造式のSMILES表現を示す。構造式行列カラムT3C5は、構造式変換部P01によりSMILES表現から変換された化学構造式の構造式行列を示す。

【0048】

実験データ構造式行列テーブル142において、TableIDカラムT4C1は、そのテーブル(図6に示すテーブル)の識別子を示す。TableTypeカラムT4C

10

20

30

40

50

2 は、そのテーブルが格納しているデータのタイプを示す。Table Type カラム T 4 C 2 は、本テーブルが、実験データから生成されたテーブルであることを示している。ID カラム T 4 C 3 は、化学構造式の識別子を示す。SMILES カラム T 4 C 4 は、化学構造式の SMILES 表現を示す。Mwt カラム T 4 C 5 は、化学構造式で表される化合物の分子量を示す。log P カラム T 4 C 6 は、化学構造式で表される化合物の分配係数を示す。構造式行列カラム T 4 C 7 は、構造式変換部 P 0 1 により SMILES 表現から変換された化学構造式の構造式行列を示す。

【0049】

図7は、学習用データベース DB 15 に含まれるデータの例を示す。学習用データベース DB 15 は、構造式行列データベース DB 14 から学習データ生成部 P 0 2 により生成された、化学構造式生成モデルの学習に使用するデータを格納する。本明細書の実施例における化学構造式生成モデルは、構造モデル及び構造物性関係モデルを含む。図6に示すように、学習用データベース DB 15 は、構造モデル用テーブル 15 1 及び構造物性関係モデル用テーブル 15 2 を含む。構造モデル用テーブル 15 1 は、物性値の測定値と関連付けのない化合物の構造式行列群（化合物構造表現群）を格納する。構造物性関係モデル用テーブル 15 2 は、物性値の測定値と関連づけられた化合物の構造式行列群（化合物構造表現群）を格納する。

【0050】

図7に示す構造モデル用テーブル 15 1 の例は、Table ID カラム T 5 C 1 を除き、カタログデータ構造式行列テーブル 14 1 と同一の情報を格納している。Table ID カラム T 5 C 1 は、図7に示す構造モデル用テーブル 15 1 の識別子を示す。カラム T 5 C 2 ~ T 5 C 5 は、それぞれ、カタログデータ構造式行列テーブル 14 1 における同名のカラム T 3 C 2 ~ T 3 C 5 と同様である。

【0051】

図7に示す構造物性関係モデル用テーブル 15 2 の例は、Table ID カラム T 6 C 1 を除き、実験データ構造式行列テーブル 14 2 と同一の情報を格納している。Table ID カラム T 6 C 1 は、図6に示す構造物性関係モデル用テーブル 15 2 の識別子を示す。カラム T 6 C 2 ~ T 6 C 7 は、それぞれ、実験データ構造式行列テーブル 14 2 における同名のカラム T 4 C 2 ~ T 4 C 7 と同様である。

【0052】

図8は、初期パラメータ DB 13 に含まれる情報の例を示す。初期パラメータ DB 13 は、ネットワーク構造の定義に必要なパラメータにおける初期値の全てを格納している。例えば、ニューラルネットワークの構造パラメータの初期値、学習用パラメータの初期値、そのほかのユーザ設定データの初期値が格納されている。更新されたパラメータは、モデルデータ DB 17 に格納される。

【0053】

なお、デフォルト値が事前に与えられている場合は、一部の初期パラメータは省略されていてもよい。例えば、ニューラルネットワークを構成する層の種類、数、順序、次元数、ニューロンの重み、重みの更新率等、ニューラルネットワークのネットワーク定義に一般的に必要とされるパラメータは省略されてもよい。

【0054】

ユーザは、いずれかの装置の入力装置を介して、初期パラメータ DB 13 を設定することができる。初期パラメータ DB 13 は、化学構造式生成モデルの構成のために必要な情報を含む。図8に示す例において、「Catalog Data Tables」は、使用するカタログデータを示す。「Experiment Data Tables」は、使用する実験データを示す。「Target Properties」は、注目する物性値の種類を示す。「Target Property Values」は、注目する物性値の種類の目標値を示す。

【0055】

「Number_of_vae_relation」は、構造物性関係モデルの VAE

10

20

30

40

50

(内側VAE)の段数を示す。「VAE__Initial__Params」は、化学構造式生成モデルのVAEそれぞれのパラメータの初期値を示す。より具体的には、「grammar__layer」は、構造モデルのVAE(外側VAE)層の数や次元数等の構成を示す。「vae__relation__layers」は、構造物性値関係モデルのVAEそれぞれの層の数や次元数等の構成を示す。「middle__dims」は、エンコーダ又はデコーダからの中間出力における次元数のリストを示す。

【0056】

図9は、表示装置M05において、表示部P11がユーザのために表示するネットワーク構造確認画面201の例を示す。表示部P11は、ネットワーク構造決定部P04から受信した化学構造式生成モデルの構成情報から、化学構造式生成モデルの構成図を生成し、モニタにおいて表示する。

10

【0057】

図9に示す構成例において、外側VAEである構造モデルは、外側エンコーダ#enc__01と外側デコーダ#dec__01で構成されている。内側VAEである構造物性関係モデルは、内側エンコーダ#enc__02と内側デコーダ#dec__02で構成されている。内側VAEは、外側エンコーダ#enc__01と外側デコーダ#dec__01とに挟まれている。

【0058】

外側エンコーダ#enc__01は、SMILES表現から生成された構造式行列を入力として受け付け、9次元の中間ベクトル(潜在表現)を出力する。内側エンコーダ#enc__02は、外側エンコーダ#enc__01の出力に二つの物性値(MWt及びlogP)を結合した11次元のベクトルを入力として受け付け、7次元の中間ベクトル(潜在表現)を出力する。

20

【0059】

内側デコーダ#dec__02は、内側エンコーダ#enc__02の出力を入力として受け付け、11次元のベクトルを出力する。このベクトルは、化学構造式に対応する9次元のベクトルと、二つの物性値を示す2次元のベクトルの結合である。外側デコーダ#dec__01は、内側デコーダ#dec__02の出力から抽出された9次元ベクトルを入力として受け付け、化学構造行列を示すベクトルを出力する。化学構造行列を逆変換することで、化学構造式のSMILES表現が得られる。

30

【0060】

ユーザは、ネットワーク構造確認画面201を参照することで、構成される化学構造式生成モデルが、所望の構成を有するか確認することができる。化学構造式生成モデルの構成の変更を望む場合、ユーザは、表示装置M05の入力装置から、修正のためのデータを入力することができる。

【0061】

表示部P11は、ネットワーク構造確認画面201の他に、化学構造式生成モデルが新たに生成した化学構造式及びそれに関連する情報を表示する。ユーザは、表示された化学構造式から、実際に実験を行う化学構造式を選択することができる。

【0062】

40

図10は、モデルデータDB17に含まれるモデルテーブル171の構成例を示す。モデルデータDB17は、化学構造式生成モデルのネットワーク構造の定義に必要なパラメータを格納する。初期パラメータには含まれないデフォルト値を含め、全てのパラメータが含まれる。例えば、ニューラルネットワークの構造パラメータ、学習用パラメータ、構造式行列の逆変換に必要な語彙データ、その他ユーザ設定データが含まれる。パラメータは、学習に従い逐次更新される。モデルデータDB17は、例えば、学習の開始時、学習中、学習終了時などのタイミングで読み書きされる。

【0063】

モデルテーブル171は、ネットワーク構造決定部P04に生成され、モデルデータDB17に含められる。図10のモデルテーブル171は、図9に示す化学構造式生成モデ

50

ルの構成図と対応している。図10の例において、Network IDカラムT7C1は、化学構造式生成モデルのエンコーダ又はデコーダの識別子を示す。Network OrderカラムT7C2は、エンコーダ又はデコーダの、入力からの順番を示す。Nested OrderカラムT7C3は、入れ子構造を有するVAEの入力からの順番を示す。TargetカラムT7C4は、VAEの学習に使用するデータの識別子を示す。

【0064】

図11は、構造式変換部P01の処理例のフローチャートを示す。構造式変換部P01は、カタログデータ及び実験データに含まれる化学構造式の文字列を、構造式行列に変換する。まず、構造式変換部P01は、初期パラメータDB13から、必要な初期パラメータを読み込む(S101)。構造式変換部P01は、さらに、構造式語彙データDB12を読み込む(S102)。構造式語彙データDB12は、構造式行列の縦に並ぶ元素の種類とSMILES表現における記号とを対応付ける。構造式行列の縦の次元数と語彙数とは一致する。

【0065】

次に、構造式変換部P01は、初期パラメータが示すカタログデータDB10及び実験データDB11から元データを読み込む(S103)。構造式変換部P01は、読み込んだデータにおける全ての構造式の末尾に、終了トークンを付加する(S104)。構造式変換部P01は、構造式語彙データDB12を参照して、全ての構造式それぞれを、構造式行列に変換する(S105)

【0066】

構造式変換部P01は、元データのテーブルそれぞれにカラムを追加し、変換した構造式行列を格納する(S106)。これにより、図6に示す、カタログデータ構造式行列テーブル141及び実験データ構造式行列テーブル142が生成される。構造式変換部P01は、生成したテーブル141、142を構造式行列データベースDB14に書出して追加する(S107)。さらに、構造式変換部P01は、変換に用いた構造式語彙データを、モデルデータDB17の一部(構造式語彙辞書)として書き出す(S108)。構造式語彙データは、構造式行列を逆変換して化学構造式のSMILES表現を得るために参照される。

【0067】

図12は、学習データ生成部P02の処理例のフローチャートを示す。学習データ生成部P02は、化学構造式生成モデルのVAEそれぞれの学習データを生成する。まず、学習データ生成部P02は、初期パラメータDB13から、必要な初期パラメータを読み込む(S151)。次に、学習データ生成部P02は、構造式行列データベースDB14を読み込む(S152)。学習データ生成部P02は、読み込んだデータのレコードそれぞれのTable Typeを判定する(S153)。

【0068】

学習データ生成部P02は、Table Typeに応じて異なる処理を実行する。Table Typeが「Catalog」であるレコード(S153:Catalog)に対する処理を説明する。学習データ生成部P02は、該当するレコードを抽出し、一つのテーブルに集約する(S154)。

【0069】

次に、Table Typeが「Experiment」であるレコード(S153:Experiment)に対する処理を説明する。学習データ生成部P02は、該当するレコードを抽出し(S155)、一つのテーブルに集約する(S156)。レコードが、欠損した物性値を含む場合、学習データ生成部P02は、その物性値のフィールドをNullで補完する。欠損した物性値は、他のいずれかのレコードが含み当該レコードに含まれない物性値である。

【0070】

次に、学習データ生成部P02は、初期パラメータが示す構造物性関係モデルの段数に従って、テーブルを生成する(S157)。各テーブルは、一つの内側VAEの学習デー

10

20

30

40

50

タを格納する。学習データ生成部 P 0 2 は、生成したテーブルの N u l l を含むカラムを削除する (S 1 5 8)。生成されるテーブルの物性値集合は、後述する包含関係を満たす (例えば実施例 4 を参照)。学習データ生成部 P 0 2 は、生成したテーブルに新しい T a b l e I D を付与し、T a b l e I D カラムを上書き更新する (S 1 5 9)。学習データ生成部 P 0 2 は、生成したテーブルを学習用データベースに書き出す (S 1 6 0)。

【 0 0 7 1 】

図 1 3 は、構造発生ベクトル群生成部 P 0 3 の処理例のフローチャートを示す。構造発生ベクトル群生成部 P 0 3 は、化学構造式生成モデルの学習後に、目標の物性値を有すると期待される新たな化学構造式を生成 (推定) するための入力データを生成する。

【 0 0 7 2 】

まず、構造発生ベクトル群生成部 P 0 3 は、初期パラメータ D B 1 3 から、必要な初期パラメータを読み込む (S 2 0 1)。次に、構造発生ベクトル群生成部 P 0 3 は、学習用データベース D B 1 5 を読み込む (S 2 0 2)。構造発生ベクトル群生成部 P 0 3 は、学習用データベースから、T a b l e T y p e が「E x p e r i m e n t」のテーブルを抽出する (S 2 0 3)。各テーブルは、対応する一つの内側 V A E の学習データを示す。

【 0 0 7 3 】

構造発生ベクトル群生成部 P 0 3 は、抽出した各テーブルにおいて、レコードを各物性値でソートして、各テーブルの各物性値の上位 S 件を抽出する。S は、初期パラメータが示す自然数である。テーブルが複数種類の物性値を含む場合、各種類の上位 S 件が抽出される。構造発生ベクトル群生成部 P 0 3 は、I D カラムと構造式行列カラムのみの上位化合物テーブルとして集約する (S 2 0 4)。なお、複数の同一 I D のレコードが抽出されている場合、それらの一つレコードのみが上位化合物テーブルに格納される。

【 0 0 7 4 】

上位化合物テーブルの生成は、上記方法に限定されない。例えば、一部のテーブル、例えば、最も多くの種類の物性値を含むテーブルのみからレコードを抽出してもよく、指定された種類の物性値のみの上位レコードを抽出してもよい。抽出する上位レコードの数は、物性値種類間で異なってもよい。

【 0 0 7 5 】

次に、構造発生ベクトル群生成部 P 0 3 は、初期パラメータに従って、各物性値目標値リストを生成する (S 2 0 5)。各目標値リストは、対応する物性値種類の複数の目標値を示す。初期パラメータは、複数の目標値を生成するための情報を示し、例えば、上記複数の目標値を示してもよく、又は、基準目標値、生成目標値数及び基準目標値から他の目標値を生成する式を示してもよい。

【 0 0 7 6 】

構造発生ベクトル群生成部 P 0 3 は、物性値種類それぞれの目標値リストの直積により目標値行列を生成する (S 2 0 6)。さらに、構造発生ベクトル群生成部 P 0 3 は、上位化合物テーブルと目標値行列の直積によって、構造発生ベクトル群を生成する (S 2 0 7)。構造発生ベクトル群生成部 P 0 3 は、生成した構造発生ベクトル群を、構造発生ベクトルデータベース D B 1 6 に書き出す (S 2 0 8)。

【 0 0 7 7 】

図 1 4 は、ネットワーク構造決定部 P 0 4 の処理例のフローチャートを示す。ネットワーク構造決定部 P 0 4 は、初期パラメータ及び実験データに含まれる考慮対象物性値等 x から、化学構造式生成モデルの構造を決定する (モデルデータ生成)。

【 0 0 7 8 】

まず、ネットワーク構造決定部 P 0 4 は、初期パラメータ D B 1 3 から、必要な初期パラメータを読み込む (S 2 5 1)。読み込まれる初期パラメータは、カタログデータ識別子、実験データ識別子、対象物性のカラム名、次元数リスト等を含む。

【 0 0 7 9 】

次に、ネットワーク構造決定部 P 0 4 は、構造モデルを構築して初期パラメータで初期化する (S 2 5 2)。ネットワーク構造決定部 P 0 4 は、学習用データベース D B 1 5 の

10

20

30

40

50

構造物性関係モデル用テーブルを読み込む (S 2 5 3)。ネットワーク構造決定部 P 0 4 は、構造物性関係モデルとして、構造物性関係モデル用テーブルの数だけ、エンコーダ・デコーダペア (内側 V A E) を構築し、初期パラメータで初期化する (S 2 5 4)。

【 0 0 8 0 】

ネットワーク構造決定部 P 0 4 は、構造モデルのエンコーダ、構造物性関係モデルのエンコーダ群、構造物性関係モデルのデコーダ群、構造モデルのデコーダを順にならべ、各ネットワークに対して入力側から順に連番 (N e t w o r k O r d e r) を付与する (S 2 5 5)。

【 0 0 8 1 】

ネットワーク構造決定部 P 0 4 は、構造物性関係モデル用テーブルそれぞれの物性値カラム名 (物性値種類) を集合化し、包含判定を行う (S 2 5 6)。任意の二つの構造物性関係モデル用テーブルの間において、物性値カラム名の包含関係が成立する。具体的には、物性値カラム数が多いテーブルは、物性値カラム数が少ないテーブルの全ての物性値カラム名を含む。このような包含関係が成立するように、構造物性関係モデル用テーブルは学習データ生成部 P 0 2 によって用意される。

【 0 0 8 2 】

ネットワーク構造決定部 P 0 4 は、含まれる物性値カラム数が少ないテーブル順に、T a b l e I D を降順にソートする (S 2 5 8)。ネットワーク構造決定部 P 0 4 は、より上位の T a b l e I D が、構造物性値関係モデルにおけるより外側のエンコーダ・デコーダペアと対応するように、各エンコーダ・デコーダペアと学習用テーブルを対応付ける。ネットワーク構造決定部 P 0 4 は、初期パラメータに応じて、各エンコーダ・デコーダペアの入出力の次元数を決定する (S 2 6 0)。

【 0 0 8 3 】

次に、ネットワーク構造決定部 P 0 4 は、モデル構造を表示する (S 2 6 1)。具体的には、ネットワーク構造決定部 P 0 4 は、モデル構造の情報を表示部 P 1 1 に送信する。表示部 P 1 1 は、受信した情報に従って、化学構造式生成モデルの構造画像を生成、表示する。

【 0 0 8 4 】

ネットワーク構造決定部 P 0 4 は、表示部 P 1 1 を介して、化学構造式生成モデルの構造についてのユーザ入力を受け付け、ネットワーク構造の修正の有無を判定する (S 2 6 2)。

【 0 0 8 5 】

ネットワーク構造の修正のユーザ指示を受信した場合 (S 2 6 2 : 修正あり)、ネットワーク構造決定部 P 0 4 は、ユーザ入力に従いネットワーク構造修正し (S 2 6 3)、修正したネットワーク構造を、表示部 P 1 1 を利用して表示する。

【 0 0 8 6 】

ネットワーク構造の修正が不要である場合 (S 2 6 2 : 修正なし)、ネットワーク構造決定部 P 0 4 は、エンコーダの入力とデコーダの出力が一致するエンコーダとデコーダをペアにし、各ペアに対して外側から順に連番 (N e s t O r d e r) を付与する (S 2 6 4)。各ペアが V A E を構成する。

【 0 0 8 7 】

ネットワーク構造決定部 P 0 4 は、全エンコーダ及びデコーダの全パラメータを、D B 1 7 の一部としてデータ保存装置 M 0 2 に出力する (S 2 6 5)。さらに、ネットワーク構造決定部 P 0 4 は、モデルテーブル 1 7 1 をモデルデータ D B 1 7 の一部としてデータ保存装置 M 0 2 に出力する (S 2 6 6)。

【 0 0 8 8 】

図 1 5 は、構造モデル学習部 P 0 5 の処理例のフローチャートを示す。構造モデル学習部 P 0 5 は、カタログデータから生成した構造モデル用テーブルを使用して、構造モデル (外側 V A E) の学習 (訓練とも呼ぶ) を実行する。カタログデータ D B 1 0 は、実験データ D B 1 1 より多くのレコード (データ) を格納している。化学構造式のみであれば学

10

20

30

40

50

習データとしてより多くのデータを用意することができるので、化学構造式生成モデル全体として効果的な学習が可能となる。

【0089】

まず、構造モデル学習部 P05 は、モデルデータ DB17 を読み込む (S301)。次に、構造モデル学習部 P05 は、モデルテーブルを参照し、Nest Order が 1 のモデルを特定する (S302)。Nest Order が 1 のモデルは、最も外側の構造モデルである。さらに、構造モデル学習部 P05 は、特定したモデルを構築する (S303)。

【0090】

構造モデル学習部 P05 は、学習用データベース DB15 を参照し、構造モデル用テーブルを読み込む (S304)。構造モデル学習部 P05 は、構造式行列を構造モデルに順次入力して、ニューラルネットワークの学習を行う。構造モデル学習部 P05 は、ネットワークのパラメータを更新、最適化する (S305)。構造モデル学習部 P05 は、学習後のパラメータを書き出し、モデルデータ DB17 を更新する (S306)。

【0091】

図 16 は、構造モデル追加学習部 P06 の処理例のフローチャートを示す。構造モデル追加学習により、構造モデルの特徴抽出精度を高めることができる。構造モデル追加学習部 P06 は、構造物性関係モデル用テーブルの構造式行列を使用して、構造モデルの追加学習を行う。まず、構造モデル追加学習部 P06 は、モデルデータ DB17 を読み込む (S351)。

【0092】

次に、構造モデル追加学習部 P06 は、モデルテーブルを参照し、Nest Order が 1 の学習済み構造モデルを再構築する (S352)。構造モデル追加学習部 P06 は、学習用データベース DB15 を参照し、全ての構造物性関係モデル用テーブルを読み込み (S353)。

【0093】

構造モデル追加学習部 P06 は、構造物性関係モデル用テーブルの構造式行列を構造モデルに順次入力して、学習済み構造モデルの追加学習を行う。構造モデル追加学習部 P06 は、ネットワークのパラメータを更新、最適化する (S354)。構造モデル追加学習部 P06 は、追加学習後の構造モデルパラメータを書き出し、モデルデータ DB17 を更新する (S355)。

【0094】

図 17 は、構造物性関係モデル学習部 P07 の処理例のフローチャートを示す。構造物性関係モデル学習部 P07 は、構造物性関係モデル内の VAE (以下モデルとも呼ぶ) それぞれの学習を実行する。内側 VAE の学習のため、その VAE より外側の全てのエンコーダが再構築及び接続される。

【0095】

まず、構造物性関係モデル学習部 P07 は、モデルデータ DB17 を読み込む (S401)。構造物性関係モデル学習部 P07 は、N を初期化して値を 2 に設定する (S402)。構造物性関係モデル学習部 P07 は、モデルテーブルの Nest Order の値が N に等しい行の Network ID 列を参照し、学習対象モデルの VAE を構築する (S403)。

【0096】

構造物性関係モデル学習部 P07 は、モデルテーブルの Nest Order の値が N に等しい行の Target 列を参照し、学習用データベース DB15 から、対応する学習用のテーブル (構造物性関係モデル用テーブル) を読み込む (S404)。

【0097】

構造物性関係モデル学習部 P07 は、追加学習済み構造モデルの (デコーダを構築することなく) エンコーダのみを再構築する (S405)。さらに、構造物性関係モデル学習部 P07 は、モデルテーブルの Nest Order の値が N より小さい Network

10

20

30

40

50

I D 列を参照し、(デコーダを構築することなく)学習済みエンコーダのみを再構築する(S 4 0 6)。構造物性関係モデル学習部 P 0 7 は、モデルテーブルの N e t w o r k O r d e r 列を参照し、構築した学習済みエンコーダを順番に接続する(S 4 0 7)。

【0 0 9 8】

構造物性関係モデル学習部 P 0 7 は、接続したエンコーダそれぞれに、構造式行列と対応する物性値を逐次入力し、学習対象ベクトルに変換する(S 4 0 8)。構造モデルには、構造式行列のみが入力される。構造物性関係モデル学習部 P 0 7 は、学習対象ベクトルを学習対象モデルである V A E に入力して、当該モデルの学習を行い、ネットワークのパラメータを最適化する(S 4 0 9)。N = 2 の場合、学習対象ベクトルは、構造モデルの構造行列の変換結果と物性値ベクトルを結合したベクトルである。構造物性関係モデル学習部 P 0 7 は、学習後の当該モデルのパラメータを書き出し、モデルデータ D B 1 7 を更新する(S 4 1 0)。

10

【0 0 9 9】

構造物性関係モデル学習部 P 0 7 は、構造物性関係モデルの全てのモデル(V A E)の学習が終了したか判定する(S 4 1 1)。未学習のモデルが残っている場合(S 4 1 1 : N O)、構造物性関係モデル学習部 P 0 7 は、N e s t O r d e r の値 N をインクリメントし(S 4 1 2)、ステップ S 4 0 3 に戻る。構造物性関係モデルの全てのモデルの学習が終了している場合(S 4 1 1 : Y E S)、本フローは終了する。

【0 1 0 0】

図 1 8 は、新構造式生成部 P 0 8 の処理例のフローチャートを示す。新構造式生成部 P 0 8 は、学習済みの化学構造式生成モデルを使用して、所望の物性値を有することが期待される新たな化学構造式の候補を生成(推定)する。

20

【0 1 0 1】

まず、新構造式生成部 P 0 8 は、モデルデータ D B 1 7 を読み込む(S 4 5 1)。新構造式生成部 P 0 8 は、学習済みの構造モデル及び構造物性値関係モデルを再構築し、化学構造式生成モデル(生成器)を構成する(S 4 5 2)。新構造式生成部 P 0 8 は、構造発生ベクトルデータベース D B 1 6 から、構造発生ベクトル群を読み込む(S 4 5 3)。

【0 1 0 2】

新構造式生成部 P 0 8 は、構造発生ベクトル群を化学構造式生成モデルに入力し、構造式行列を生成する(S 4 5 4)。新構造式生成部 P 0 8 は、構造式行列をまとめて候補構造式として、候補構造式データベース D B 1 8 に書き出す(S 4 5 5)。

30

【0 1 0 3】

図 1 9 は、構造式逆変換部 P 0 9 の処理例のフローチャートを示す。構造式逆変換部 P 0 9 は、化学構造式生成モデルが出力した構造式行列を、構造式の S M I L E S 表現(文字列)に変換する。

【0 1 0 4】

まず、構造式逆変換部 P 0 9 は、モデルデータ D B 1 7 から構造式語彙辞書を読み込む(S 5 0 1)。構造式逆変換部 P 0 9 は、候補構造式データベース D B 1 8 を読み込む(S 5 0 2)。構造式逆変換部 P 0 9 は、構造式行列を構造式(S M I L E S 表現)に変換する。(S 5 0 3)。構造式逆変換部 P 0 9 は、末尾の終了トークンを削除する。(S 5 0 4)。構造式逆変換部 P 0 9 は、構造式を候補構造式データベース 1 8 D B に書き出す。(S 5 0 5)。

40

【0 1 0 5】

図 2 0 は、構造式整形部 P 1 0 の処理例のフローチャートを示す。V A E の性質から、化学構造式生成モデルが生成した化学構造式は、S M I L E S の文法に適合しない化学構造式を含み得る。構造式整形部 P 1 0 は、S M I L E S の文法に適合しない化学構造式を補正し、さらに、補正できない化学構造式を除去する。

【0 1 0 6】

まず、構造式整形部 P 1 0 は、候補構造式データベース 1 8 D B を読み込む(S 5 5 1)。構造式整形部 P 1 0 は、各化学構造式について文法整合性判定する。(S 5 5 2)。

50

構造式整形部 P 1 0 は、化学構造式が文法整合性を満たさない場合、その化学構造式を補正する。(S 5 5 3)。構造式整形部 P 1 0 は、補正した化学構造式の文法整合性再判定する。(S 5 5 4)。構造式整形部 P 1 0 は、候補構造式を棄却する。(S 5 5 5)。構造式整形部 P 1 0 は、補正した化学構造式を、候補構造式データベース 1 8 D B に書きする(S 5 5 6)。

【 0 1 0 7 】

以上のように、入れ子構造を有する化学構造生成モデルは、学習データに含まれる物性値よりも良い物性値を持つ化合物の候補を提示することができ、少ない実験データを使用して生成することができる。また、実験データを使用した構造モデルの追加学習により、構造モデルの特徴抽出精度をさらに高めることができる。

10

【実施例 2】

【 0 1 0 8 】

以下において、実施例 2 を説明する。主に、実施例 1 との相違点を説明する。実施例 2 に係る構造物性値関係モデルは、複数段の V A E からなる入れ子構造を有している。また、構造物性値関係モデルの各エンコーダ(各 V A E)への入力は、前段エンコーダからの中間ベクトルと単一物性値とを結合したベクトルである。構造物性値関係モデルが複数段の V A E で構成されることで、例えば、分離することが好ましい物性値を異なる V A E の入力に含めることができる。

【 0 1 0 9 】

実施例 2 において、学習のために使用される実験データの全ての化学構造式(レコード)が、共通の物性値種類(物性値名)の実験データを有しているとする(実験データの欠損なし)。以下において、各化学構造式に対して、M W t と l o g P の二つの種類の物性値の実験データが関連付けられている例を説明する。

20

【 0 1 1 0 】

図 2 1 は、実施例 2 に係る、表示部 P 1 1 が表示装置 M 0 5 においてユーザのために表示する、ネットワーク構造確認画面 2 0 2 の例を示す。表示部 P 1 1 は、ネットワーク構造決定部 P 0 4 から受信した化学構造式生成モデルの構成情報から、化学構造式生成モデルの構成図を生成し、モニタにおいて表示する。

【 0 1 1 1 】

図 2 1 に示す構成例において、外側 V A E である構造モデルは、外側エンコーダ # e n c _ 0 1 と外側デコーダ # d e c _ 0 1 で構成されている。構造物性関係モデルは、二つの内側 V A E で構成されている。一つの内側 V A E は、内側エンコーダ # e n c _ 0 2 と内側デコーダ # d e c _ 0 2 で構成されている。もう一つの内側 V A E は、内側エンコーダ # e n c _ 0 3 と内側デコーダ # d e c _ 0 3 で構成されている。

30

【 0 1 1 2 】

内側エンコーダ # e n c _ 0 3 と内側デコーダ # d e c _ 0 3 は、内側エンコーダ # e n c _ 0 2 と内側デコーダ # d e c _ 0 2 の間に挟まれている。内側エンコーダ # e n c _ 0 3 及び内側デコーダ # d e c _ 0 3 並びに内側エンコーダ # e n c _ 0 2 及び内側デコーダ # d e c _ 0 2 は、外側エンコーダ # e n c _ 0 1 と外側デコーダ # d e c _ 0 1 とに挟まれている。

40

【 0 1 1 3 】

外側エンコーダ # e n c _ 0 1 は、S M I L E S 表現から生成された構造式行列を入力として受け付け、中間ベクトル(潜在表現)を出力する。内側エンコーダ # e n c _ 0 2 は、外側エンコーダ # e n c _ 0 1 の出力に一つの物性値(M W t)を示す 1 次元ベクトルを結合したベクトルを入力として受け付け、中間ベクトル(潜在表現)を出力する。内側エンコーダ # e n c _ 0 3 は、内側エンコーダ # e n c _ 0 2 の出力に一つの物性値(l o g P)を示す 1 次元ベクトルを結合したベクトルを入力として受け付け、中間ベクトル(潜在表現)を出力する。

【 0 1 1 4 】

内側デコーダ # d e c _ 0 3 は、内側エンコーダ # e n c _ 0 3 の出力を入力として受

50

付け、ベクトルを出力する。ベクトルの一部は、内側エンコーダ# e n c _ 0 3 への入力に対応し、他の一部は物性値 (l o g P) に対応する。内側デコーダ# d e c _ 0 3 の出力ベクトルから物性値ベクトルを除いたベクトルが、内側デコーダ# d e c _ 0 2 に入力される。内側デコーダ# d e c _ 0 2 からの出力されたベクトルの一部は、内側エンコーダ# e n c _ 0 2 への入力、つまり、化学構造式 (構造式行列) の特徴ベクトルであり、他の一部は物性値 (M W t) の物性値ベクトルである。

【 0 1 1 5 】

内側デコーダ# d e c _ 0 2 の出力ベクトルから物性値ベクトルを除いたベクトルが、外側デコーダ# d e c _ 0 1 に入力される。外側デコーダ# d e c _ 0 1 は、化学構造行列を示すベクトルを出力する。化学構造行列を逆変換することで、化学構造式の S M I L E S 表現が得られる。

10

【 0 1 1 6 】

図 2 2 は、実施例 2 に係る、学習用データベース D B 1 5 に含まれるデータの例を示す。図 2 2 は、構造物性関係モデルの学習用データを示す。構造物性関係モデルの学習用データは、構造物性関係モデル内の外側の第 1 V A E (# e n c _ 0 2 及び# d e c _ 0 2) 用学習テーブル 1 5 3 と、構造物性関係モデル内の内側の第 2 V A E (# e n c _ 0 3 及び# d e c _ 0 3) 用学習テーブル 1 5 4 と、を格納している。

【 0 1 1 7 】

第 1 V A E 用学習テーブル 1 5 3 は、図 7 に示す構造物性関係モデル用テーブル 1 5 2 から l o g P カラム T 6 C 6 を除いた構造を有している。T a b l e I D カラム T 8 C 1 は、第 1 V A E 用学習テーブル 1 5 3 の識別子を示す。第 1 V A E 用学習テーブル 1 5 3 のカラム T 8 C 2 ~ T 8 C 5、T 8 C 7 は、構造物性関係モデル用テーブル 1 5 2 のカラム T 6 C 2 ~ T 6 C 5、T 6 C 7 と、同じ情報を格納している。第 1 V A E (# e n c _ 0 2 及び# d e c _ 0 2) の学習が必要とする物性値は、第 1 V A E の入出力に含まれる M W t のみである。

20

【 0 1 1 8 】

第 2 V A E (# e n c _ 0 3 及び# d e c _ 0 3) 用学習テーブル 1 5 4 は、図 7 に示す構造物性関係モデル用テーブル 1 5 2 と同様の構造を有している。T a b l e I D カラム T 9 C 1 は、第 2 V A E 用学習テーブル 1 5 4 の識別子を示す。第 2 V A E 用学習テーブル 1 5 4 のカラム T 9 C 2 ~ T 9 C 7 は、構造物性関係モデル用テーブル 1 5 2 のカラム T 6 C 2 ~ T 6 C 7 と、同じ情報を格納している。

30

【 0 1 1 9 】

第 2 V A E の学習のためには、より外側の二つのエンコーダ# e n c _ 0 1、# e n c _ 0 2 を再構築して接続することが必要である。そのため学習データは、第 2 V A E の入出力物性値である l o g P に加え、エンコーダ# e n c _ 0 2 の入力物性値である M W t を含む。

【 0 1 2 0 】

一例において、第 1 V A E のための学習データは、第 2 V A E の学習データより多い。第 2 V A E より外側に配置された第 1 V A E の次元数がより多いため、構造物性関係モデルの学習を効果的に行うことができる。この点は、以下の実施例 3 及び 4 において同様である。

40

【 0 1 2 1 】

図 2 3 は、実施例 2 に係る、モデルデータ D B 1 7 に含まれるモデルテーブル 1 7 2 の構成例を示す。図 2 3 のモデルテーブル 1 7 2 は、図 2 1 に示す化学構造式生成モデルの構成図と対応している。図 2 3 の例において、N e t w o r k I D カラム T 1 0 C 1 は、化学構造式生成モデルの三つのエンコーダ及び三つデコーダそれぞれの識別子を示す。N e t w o r k O r d e r カラム T 1 0 C 2 は、三つエンコーダ及び三つのデコーダそれぞれの、入力からの順番を示す。N e s t O r d e r カラム T 1 0 C 3 は、エンコーダ及びデコーダが含まれる三つの V A E それぞれの入力から (外側から) の順番を示す。T a r g e t カラム T 1 0 C 4 は、V A E の学習に使用するデータの識別子及び対象物性

50

値を示す。

【実施例 3】

【0122】

以下において、実施例 3 を説明する。主に、実施例 1 及び 2 との相違点を説明する。実施例 3 に係る構造物性値関係モデルは、複数段の V A E からなる入れ子構造を有している。また、構造物性値関係モデルの各エンコーダ（各 V A E）への入力は、前段エンコーダからの中間ベクトルと単一又は複数物性値とを結合したベクトルである。構造物性値関係モデルが複数段の V A E で構成されることで、例えば、結合することが好ましい物性値を同一の V A E の入力に含め、分離することが好ましい物性値を異なる V A E の入力に含めることができる。

10

【0123】

実施例 3 において、学習のために使用される実験データの全ての化学構造式（レコード）が、共通の物性値種類（物性値名）の実験データを有しているとする（実験データの欠損なし）。以下において、各化学構造式に対して、P r o p 1、P r o p 2、P r o p 3 の、三つの種類の物性値の実験データが関連付けられている例を説明する。

【0124】

図 2 4 は、実施例 3 に係る、表示部 P 1 1 が表示装置 M 0 5 においてユーザのために表示する、ネットワーク構造確認画面 2 0 3 の例を示す。表示部 P 1 1 は、ネットワーク構造決定部 P 0 4 から受信した化学構造式生成モデルの構成情報から、化学構造式生成モデルの構成図を生成し、モニタにおいて表示する。

20

【0125】

図 2 4 に示すネットワーク構造は、実施例 2 の図 2 1 に示すネットワーク構造と比較して、最も内側の V A E（# e n c _ 0 3 及び # d e c _ 0 3）の物性値数が 2 である点が異なる。構造物性関係モデルにおけるより外側の V A E（# e n c _ 0 2 及び # d e c _ 0 2）の入出力物性値は、P r o p 1 であり、最も内側の V A E（# e n c _ 0 3 及び # d e c _ 0 3）の入出力物性値は、P r o p 1 及び P r o p 2 である。

【0126】

図 2 5 は、実施例 3 に係る、学習用データベース D B 1 5 に含まれるデータの例を示す。図 2 5 は、構造物性関係モデルの学習用データを示す。構造物性関係モデルの学習用データは、構造物性関係モデル内の第 1 V A E（# e n c _ 0 2 及び # d e c _ 0 2）用学習テーブル 1 5 5 と、構造物性関係モデル内の第 2 V A E（# e n c _ 0 3 及び # d e c _ 0 3）用学習テーブル 1 5 6 と、を格納している。

30

【0127】

第 1 V A E 用学習テーブル 1 5 5 は、図 2 2 に示す第 1 V A E 用学習テーブル 1 5 3 と同様の構造を有している。カラム T 1 1 C 1 ~ T 1 1 C 4、T 1 1 C 7 のカラム名は、第 1 V A E 用学習テーブル 1 5 3 と同様である。P r o p 1 カラム T 1 1 C 5 は、各化学構造式の P r o p 1 の測定値を示す。第 1 V A E（# e n c _ 0 2 及び # d e c _ 0 2）の学習が必要とする物性値は、第 1 V A E の入出力に含まれる P r o p 1 のみである。

【0128】

第 2 V A E（# e n c _ 0 3 及び # d e c _ 0 3）用学習テーブル 1 5 6 は、図 2 5 に示す第 1 V A E 用学習テーブル 1 5 5 に対して二つの物性値のカラムを追加した構造を有している。カラム T 1 2 C 1 ~ T 1 2 C 5、T 1 2 C 8 の情報は、第 1 V A E 用学習テーブル 1 5 5 のカラム T 1 1 C 1 ~ T 1 1 C 5、T 1 1 C 7 と同様である。追加されている P r o p 2 カラム T 1 2 C 6、P r o p 3 カラム T 1 2 C 7 は、それぞれ、各化学構造式の P r o p 2、P r o p 3 の実験測定値を示す。

40

【0129】

第 2 V A E の学習のためには、より外側の二つのエンコーダ # e n c _ 0 1、# e n c _ 0 2 を再構築して接続することが必要である。そのため学習データは、第 2 V A E の入出力物性値である P r o p 2、P r o p 3 に加え、エンコーダ # e n c _ 0 2 の入力物性値である P r o p 1 を含む。

50

【 0 1 3 0 】

図 2 6 は、実施例 3 に係る、モデルデータ D B 1 7 に含まれるモデルテーブル 1 7 3 の構成例を示す。図 2 6 のモデルテーブル 1 7 3 は、図 2 4 に示す化学構造式生成モデルの構成図と対応している。カラム T 1 3 C 1 ~ T 1 3 C 3 の情報は、実施例 2 に係るモデルテーブル 1 7 2 のカラム T 1 0 C 1 ~ T 1 0 C 3 の情報と同様である。T a r g e t カラム T 1 3 C 5 は、本例において V A E の学習に使用するデータの識別子及び対象物性値を示す。

【 実施例 4 】

【 0 1 3 1 】

以下において、実施例 4 を説明する。主に、上記他の実施例との相違点を説明する。本実施例の実験データは、実験により得られる物性値の欠損を有する化学構造式を含む。関連付けられている物性値の組み合わせに応じて、入れ子構造内の V A E それぞれに適用する学習データを実験データから構成することで、より適切な学習が可能となる。

10

【 0 1 3 2 】

図 2 7 は、実験データにおけるレコードが含む物性値（の種類）が満たすべき条件を模式的に示す。図 2 7 に示すように、レコードの物性値種類の組み合わせは、包含関係を満たすことが要求される。具体的には、より多くの物性値（の種類）を含むレコードは、より少ない物性値（の種類）を含むレコードの全ての物性値（の種類）を含む。例えば、3 種類の実験データが存在するとする。

【 0 1 3 3 】

20

第 1 の実験データは 1 種類の物性値の実験結果を含み、第 2 の実験データは 2 種類の物性値の実験結果を含み、第 3 の実験データは 3 種類の物性値の実験結果を含むとする。第 3 の実験データの 3 種類の物性値種類は、第 1 の実験データの物性値種類と、第 2 の実験データの二つの物性値種類で構成される。第 2 の実験データの物性値種類は、第 1 の実験データの物性値種類と、他の物性値種類で構成される。

【 0 1 3 4 】

第 1 の実験データの物性値集合（物性値カラム集合又は物性値種類集合）は、第 2 及び第 3 の実験データの物性値集合に包含され、第 2 の実験データの物性値集合は、第 3 の実験データの物性値集合に包含されている。第 1 の実験データにおいて 2 種類の物性値のデータが欠けており、第 2 の実験データにおいて 1 種類の物性値のデータが欠けている。構造物性関係モデル用学習データは、上述のような包含関係を満たすように実験データから前処理される。

30

【 0 1 3 5 】

図 2 8 は、2 種類の実験データで構成される構造物性関係モデル用学習データの例を模式的に示す。第 1 の実験データ 3 1 1 は、物性値 1（P r o p 1）の測定値のみ有している。第 2 の実験データ 3 1 2 は、物性値 1（P r o p 1）及び物性値 2（P r o p 2）の測定値を有している。つまり、第 1 の実験データの物性値集合は、第 2 の実験データの物性値集合に包含されている。

【 0 1 3 6 】

図 2 9 は、実施例 4 に係る、表示部 P 1 1 が表示装置 M 0 5 においてユーザのために表示する、ネットワーク構造確認画面 2 0 4 の例を示す。図 3 0 は、実施例 3 に係る、モデルデータ D B 1 7 に含まれるモデルテーブル 1 7 4 の構成例を示す。図 3 0 のモデルテーブル 1 7 4 は、図 2 9 に示す化学構造式生成モデルの構成図と対応している。

40

【 0 1 3 7 】

図 2 9 に示すネットワーク構造は、実施例 2 の図 2 1 に示すネットワーク構造と同様である。物性値 M W t が物性値 P r o p 1 に置き換えられ、物性値 l o g P が物性値 P r o p 2 に置き換えられている。図 3 0 のモデルテーブル 1 7 4 のカラム T 1 4 C 1 ~ T 1 4 C 3 の情報は、実施例 2 の図 2 3 に示すモデルテーブル 1 7 2 のカラム T 1 0 C 1 ~ T 1 0 C 3 の情報と同様である。T a r g e t カラム T 1 4 C 4 は、本実施例のテーブル名及び物性値名（P r o p 1、P r o p 2）を示す。

50

【 0 1 3 8 】

図 3 1 は、学習データ生成部 P 0 2 が、実験データから構造物性関係モデル用学習データを生成する例を示している。初期テーブル 1 5 0 は、実験データから生成され、物性値集合が包含関係を満たすように前処理された構造物性関係モデル用学習データを格納している。学習データ生成部 P 0 2 は、初期テーブル 1 5 0 から、V A E それぞれの学習データを生成する。初期テーブル 1 5 0 のカラム T 1 5 C 1 ~ T 1 5 C 4、T 1 5 C 7 は、実施例 2 の図 2 2 に示す学習テーブル 1 5 4 の同名のカラムと同様の種類の情報を示す。カラム T 1 5 C 5 及び T 1 5 C 6 は、それぞれ、P r o p 1 及び P r o p 2 の測定値を示す。

【 0 1 3 9 】

初期テーブル 1 5 0 は、物性値集合が異なるレコードを含む。T a b l e I D が「T b l _ E x p _ 0 1 1」の各レコードは、P r o p 1 及び P r o p 2 の測定値を含む。T a b l e I D が「T b l _ E x p _ 0 1 2」の各レコードは、P r o p 1 の測定値にのみを含む。T a b l e I D が「T b l _ E x p _ 0 1 3」のレコードは、P r o p 1 及び P r o p 1 の測定値を含むレコードと、P r o p 1 の測定値にのみを含むレコードを含む。

10

【 0 1 4 0 】

学習データ生成部 P 0 2 は、実験データから、P r o p 1 及び P r o p 2 を含むレコードと、P r o p 1 のみを含むレコードを抽出し、P r o p 1 のみを含むレコードに対して、P r o p 2 のフィールドに N u l l を格納する。学習データ生成部 P 0 2 は、これらレコードを初期テーブル 1 5 0 に格納して、N u l l の数に従って（例えば昇順で）レコードをソートする。

20

【 0 1 4 1 】

次に、学習データ生成部 P 0 2 は、初期テーブル 1 5 0 から、構造物性関係モデル内第 1 V A E 用学習テーブル 1 5 7 及び構造物性関係モデル内第 2 V A E 用学習テーブル 1 5 8 を生成する。

【 0 1 4 2 】

構造物性関係モデル内第 1 V A E 用学習テーブル 1 5 7 は、構造物性関係モデルの外側の V A E (# e n c _ 0 2 及び # d e c _ 0 2) の学習データである。第 1 V A E 用学習テーブル 1 5 7 は、初期テーブル 1 5 0 において P r o p 1 の測定値を含むレコード、つまり全てのレコードを含む。カラム T 1 6 C 1 ~ T 1 6 C 5、T 1 6 C 7 は、初期テーブルの同名のカラムと同種の情報を示す。第 1 V A E 用学習テーブル 1 5 7 において、初期

30

テーブル 1 5 0 の P r o p 2 カラムが削除されている。第 1 V A E (# e n c _ 0 2 及び # d e c _ 0 2) の学習が必要とする物性値は、第 1 V A E の入出力に含まれる P r o p 1 のみである。

【 0 1 4 3 】

構造物性関係モデル内第 2 V A E 用学習テーブル 1 5 8 は、構造物性関係モデルの内側側の V A E (# e n c _ 0 3 及び # d e c _ 0 3) の学習データである。第 2 V A E 用学習テーブル 1 5 8 は、初期テーブル 1 5 0 において P r o p 1 及び P r o p 2 の測定値を含むレコードで構成されている。カラム T 1 7 C 1 ~ T 1 7 C 7 は、初期テーブルの同名のカラムと同種の情報を示す。

【 0 1 4 4 】

第 2 V A E の学習のためには、より外側の二つのエンコーダ # e n c _ 0 1、# e n c _ 0 2 を再構築して接続することが必要である。そのため学習データは、第 2 V A E の入出力物性値である P r o p 2 に加え、エンコーダ # e n c _ 0 2 の入力物性値である P r o p 1 を含む。

40

【 0 1 4 5 】

実施例 2 - 4 で説明したように、構造物性関係モデルの学習データは、複数の V A E それぞれの学習に使用される複数の学習テーブル（グループ）で構成される。各学習テーブルは、化合物構造表現のそれぞれを 1 以上の所定の物性値種類の測定値と関連付ける。任意の二つ学習テーブルの間において、物性値種類数が多い学習テーブルは、物性値種類数が少ない学習テーブルの全ての物性値種類及び全ての化合物構造表現を含む。より物性値

50

種類数が多い学習テーブルは、構造物性関係モデルにおけるより内側の V A E の学習に使用される。

【 0 1 4 6 】

本実施例 4 で説明したように、物性値の欠損を有するレコードを含む実験データから、上記物性値集合の包含関係を有するレコードで、学習データを構成する。これにより、構造物性関係モデル内の V A E のための学習データを、当該 V A E の入出力物性値に加え、より外側のエンコーダに入力される全ての物性値を含むレコードで構成できる。この結果、各 V A E の適切な学習が可能となる。

【 0 1 4 7 】

なお、本発明は上記した実施例に限定されるものではなく、様々な変形例が含まれる。例えば、上記した実施例は本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明したすべての構成を備えるものに限定されるものではない。また、ある実施例の構成の一部を他の実施例の構成に置き換えることが可能であり、また、ある実施例の構成に他の実施例の構成を加えることも可能である。また、各実施例の構成の一部について、他の構成の追加・削除・置換をすることが可能である。

【 0 1 4 8 】

また、上記の各構成・機能・処理部等は、それらの一部又は全部を、例えば集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリや、ハードディスク、SSD (Solid State Drive) 等の記録装置、または、ICカード、SDカード等の記録媒体に置くことができる。

【 0 1 4 9 】

また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしもすべての制御線や情報線を示しているとは限らない。実際には殆どすべての構成が相互に接続されていると考えてもよい。

【符号の説明】

【 0 1 5 0 】

1 0 化学構造式生成モデル、1 0 0 構造モデル、1 0 1 エンコーダ、1 0 2 デコーダ、1 0 4 構造物性関係モデル、1 0 5 エンコーダ、1 0 6 デコーダ、M 0 1 パラメータ設定装置、M 0 2 データ保存装置、M 0 3 モデル学習装置、M 0 4 構造式生成装置、M 0 5 表示装置、P 0 1 構造式変換部、P 0 2 学習データ生成部、P 0 3 構造発生ベクトル群生成部、P 0 4 ネットワーク構造決定部、P 0 5 構造モデル学習部、P 0 6 構造モデル追加学習部、P 0 7 構造物性関係モデル学習部、P 0 8 新構造式生成部、P 0 9 構造式逆変換部、P 1 0 構造式整形部、DB 1 0 カタログデータ、DB 1 1 実験データ、DB 1 2 構造式語彙データ、DB 1 3 初期パラメータ、DB 1 4 構造式行列データベース、DB 1 5 学習用データベース、DB 1 6 構造発生ベクトルデータベース、DB 1 7 モデルデータ、DB 1 8 候補構造式データベース、U 1 1 1 プロセッサ、U 1 1 2 DRAM、U 1 1 3 通信装置、U 1 1 4 補助記憶装置、U 1 1 5 入力装置

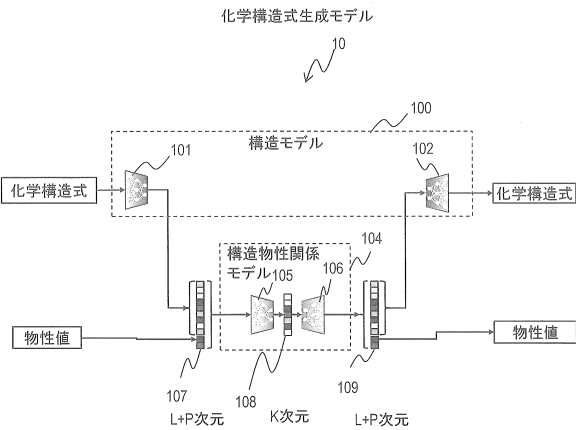
10

20

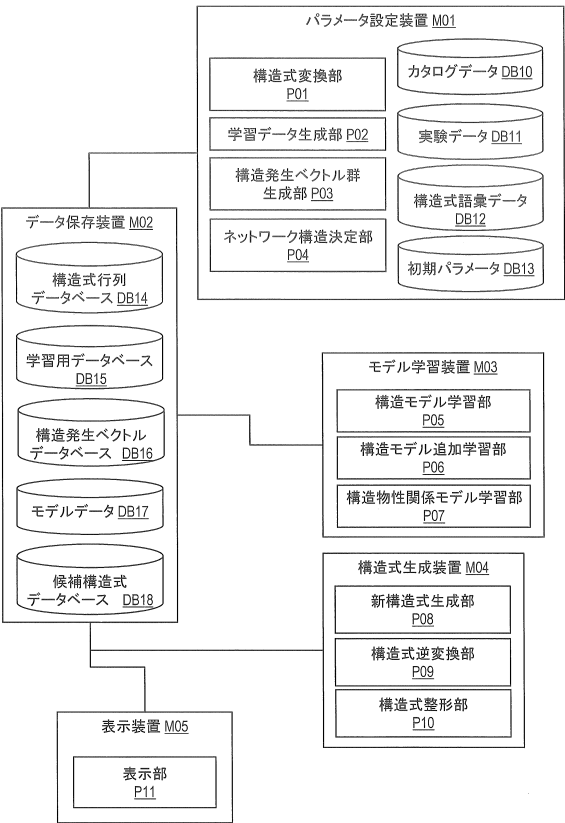
30

40

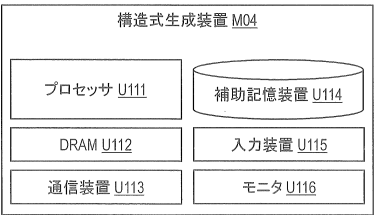
【図面】
【図 1】



【図 2】



【図 3】



【図 4】

カタログデータ DB10

TableID T0C1	TableType T0C2	ID T0C3	SMILES T0C4
Tbl_Cat_001	Catalog	ZINC000000001084	Cn1cnc2c1c(=O)n(C)c(=O)n2C
Tbl_Cat_001	Catalog	ZINC000000001447	Cn1c2nncnc2c(=O)n(C)c1=O
Tbl_Cat_001	Catalog	ZINC000000002151	Cn1cnc2c1c(=O)n(H)c(=O)n2C
Tbl_Cat_001	Catalog	ZINC000000008657	NC(=O)c1ncc(nH)c1=O
Tbl_Cat_001	Catalog	ZINC000000049143	Cc1[nH]c(=O)n(H)c(=O)c1CO
:	:	:	:

10

20

30

40

50

【図 5】

DB11

実験データ

TableID T1C1	TableType T1C2	ID T1C3	SMILES T1C4	MW T1C5	logP T1C6
Tbl_Exp_001	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	-1.368
Tbl_Exp_001	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	-1.275
Tbl_Exp_001	Experiment	ZINC000001534133	O=C(O)CNC(=O)C(=O)O	147.086	-1.728
Tbl_Exp_001	Experiment	ZINC000001555345	CNCC(N)=O	88.110	-1.309
Tbl_Exp_001	Experiment	ZINC000001556606	NC(=O)N1CCNCC1	129.163	-1.030
:	:	:	:	:	:

【図 6】

構造式行列データベース DB14

カタログデータ構造式行列テーブル

TableID T3C1	TableType T3C2	ID T3C3	SMILES T3C4	構造式行列 T3C5
Tbl_Cat_001	Catalog	ZINC000000001084	Cn1cnc2c1c(=O)n(C)c(=O)n2C	[[1,0,0,...], [0,1,0,...], [0,0,0,...], ...]
Tbl_Cat_001	Catalog	ZINC000000001447	Cn1c2nncnc2c(=O)n(C)c1=O	[[1,0,0,...], [0,1,0,...], [0,0,0,...], ...]
Tbl_Cat_001	Catalog	ZINC000000002151	Cn1cnc2c1c(=O)[nH]c(=O)n2C	[[1,0,0,...], [0,1,0,...], [0,0,0,...], ...]
Tbl_Cat_001	Catalog	ZINC000000008657	NC(=O)c1ncc[nH]c1=O	[[0,0,1,...], [0,0,0,...], [0,0,0,...], ...]
:	:	:	:	:

実験データ構造式行列テーブル

TableID T4C1	TableType T4C2	ID T4C3	SMILES T4C4	MW T4C5	logP T4C6	構造式行列 T4C7
Tbl_Exp_001	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	-1.368	[[0,0,0,...], [0,0,0,...], [0,0,0,...], ...]
Tbl_Exp_001	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	-1.275	[[0,0,0,...], [0,0,0,...], [0,0,0,...], ...]
Tbl_Exp_001	Experiment	ZINC000001534133	O=C(O)CNC(=O)C(=O)O	147.086	-1.728	[[0,0,0,...], [0,0,0,...], [0,0,0,...], ...]
Tbl_Exp_001	Experiment	ZINC000001555345	CNCC(N)=O	88.110	-1.309	[[0,0,0,...], [0,0,0,...], [0,0,0,...], ...]
:	:	:	:	:	:	:

10

20

【図 7】

学習用データベース DB15

構造モデル用テーブル

TableID T5C1	TableType T5C2	ID T5C3	SMILES T5C4	構造式行列 T5C5
Tbl_G_100	Catalog	ZINC000000001084	Cn1cnc2c1c(=O)n(C)c(=O)n2C	[[1,0,0,...], [0,1,0,...], [0,0,0,...], ...]
Tbl_G_100	Catalog	ZINC000000001447	Cn1c2nncnc2c(=O)n(C)c1=O	[[1,0,0,...], [0,1,0,...], [0,0,0,...], ...]
Tbl_G_100	Catalog	ZINC000000002151	Cn1cnc2c1c(=O)[nH]c(=O)n2C	[[1,0,0,...], [0,1,0,...], [0,0,0,...], ...]
Tbl_G_100	Catalog	ZINC000000008657	NC(=O)c1ncc[nH]c1=O	[[0,0,1,...], [0,0,0,...], [0,0,0,...], ...]
:	:	:	:	:

構造物性関係モデル用テーブル

TableID T6C1	TableType T6C2	ID T6C3	SMILES T6C4	MW T6C5	logP T6C6	構造式行列 T6C7
Tbl_R_100	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	-1.368	[[0,0,0,...], [0,0,0,...], [0,0,0,...], ...]
Tbl_R_100	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	-1.275	[[0,0,0,...], [0,0,0,...], [0,0,0,...], ...]
Tbl_R_100	Experiment	ZINC000001534133	O=C(O)CNC(=O)C(=O)O	147.086	-1.728	[[0,0,0,...], [0,0,0,...], [0,0,0,...], ...]
Tbl_R_100	Experiment	ZINC000001555345	CNCC(N)=O	88.110	-1.309	[[0,0,0,...], [0,0,0,...], [0,0,0,...], ...]
:	:	:	:	:	:	:

【図 8】

初期パラメータ DB13

```

{
  "CatalogData Tables": ["Tbl_Cat_001"],
  "ExperimentData Tables": ["Tbl_Exp_001"],
  "Target Properties": ["MW", "logP"],
  "Target Property Values": {"MW": 1.249, "logP": -1.121},
  "Number_of_vae_relation": 1,
  "VAE_Initial_Params": {
    "vae_grammar_layers": {
      "enc": {...},
      "dec": {...}
    },
    "vae_relation_layers": {
      "enc": {...},
      "dec": {...}
    },
    "middle_dims": [9,7],
    .....
  },
  .....
}

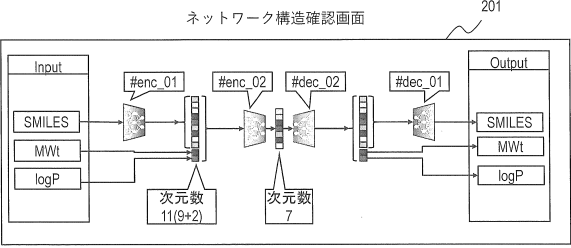
```

30

40

50

【図 9】



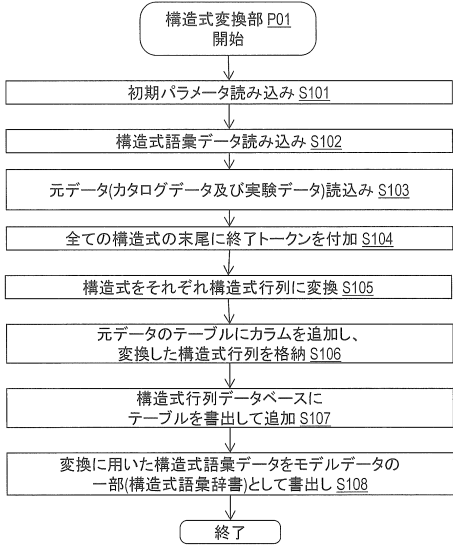
【図 10】

モデルテーブル

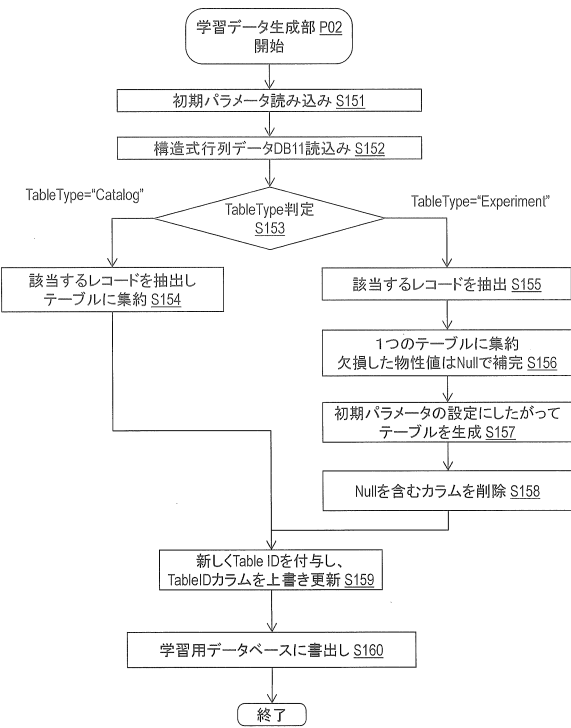
171

Network ID T7C1	Network Order T7C2	Nest Order T7C3	Target T7C4
#enc_01	1	1	{"Tbl_G_100":["構造式行列"]}
#enc_02	2	2	{"Tbl_R_100":["MWt","logP"]}
#dec_02	3	2	{"Tbl_R_100":["MWt","logP"]}
#dec_01	4	1	{"Tbl_G_100":["構造式行列"]}

【図 11】



【図 12】



10

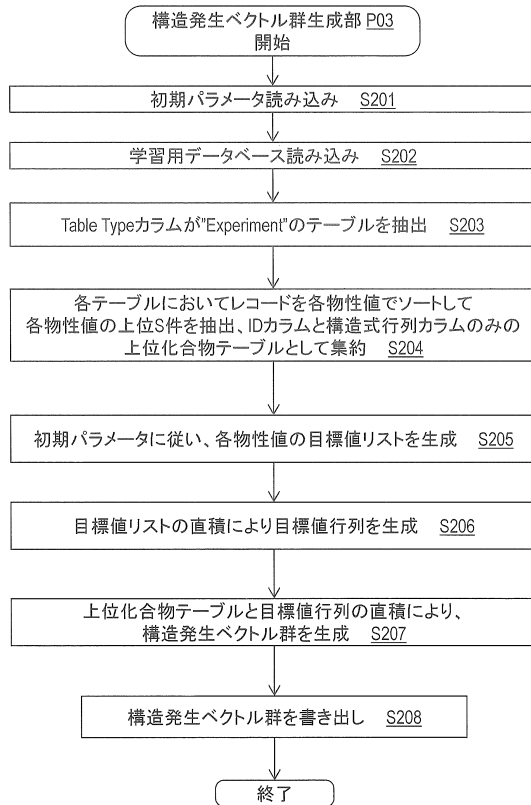
20

30

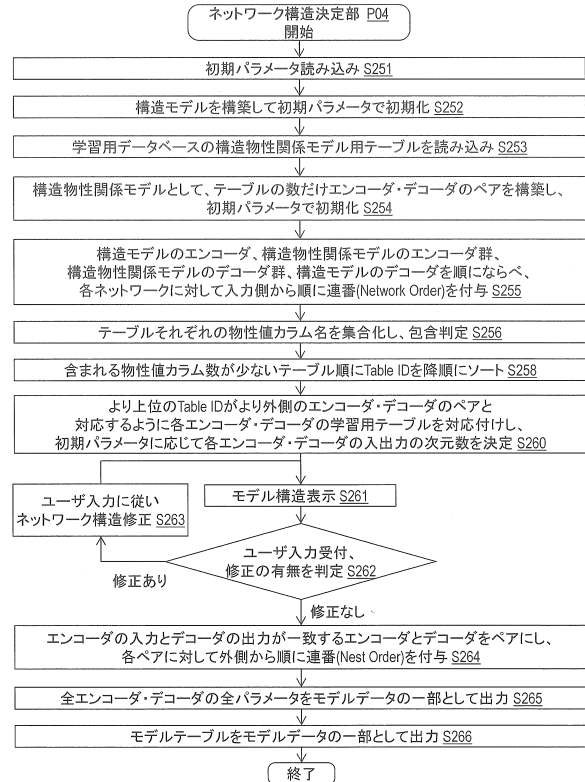
40

50

【図 13】



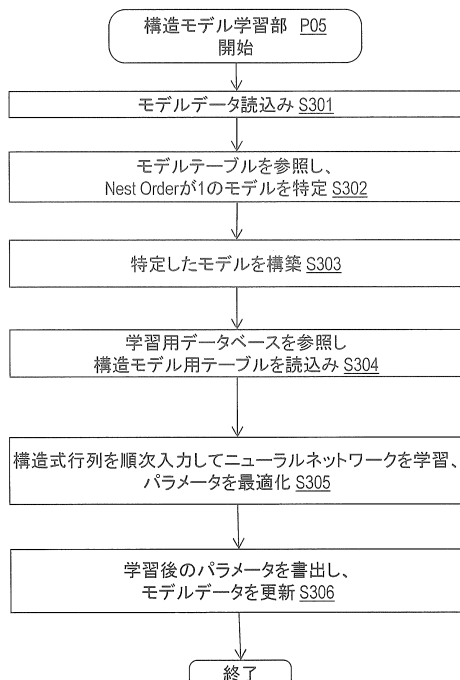
【図 14】



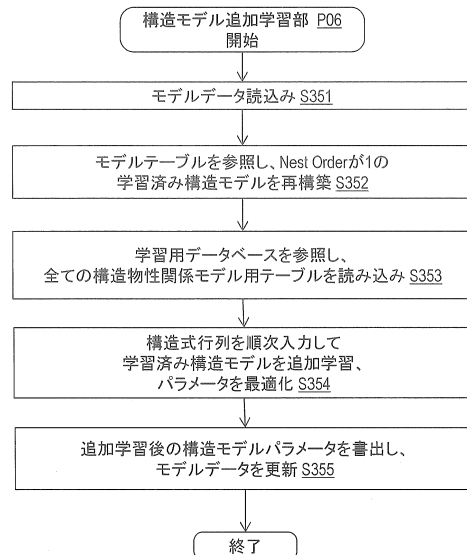
10

20

【図 15】



【図 16】

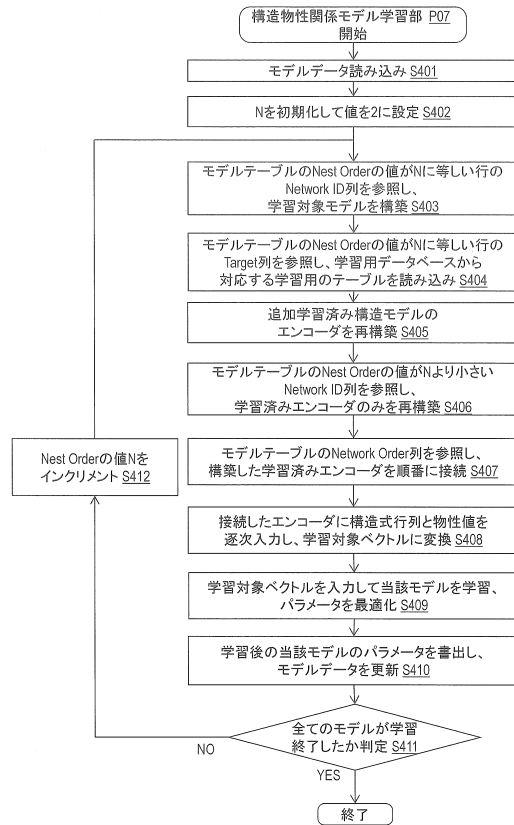


30

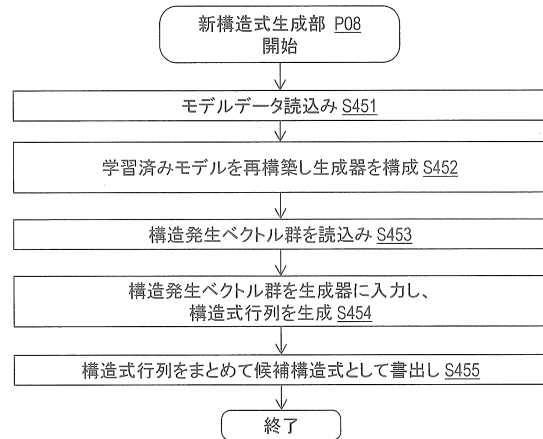
40

50

【図 17】



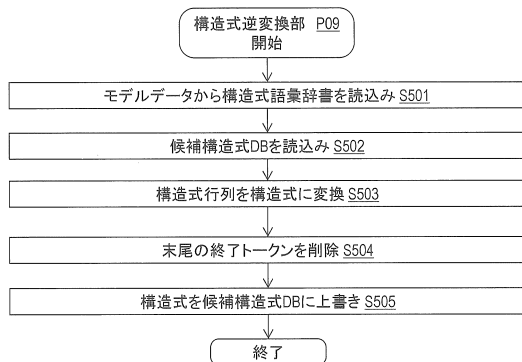
【図 18】



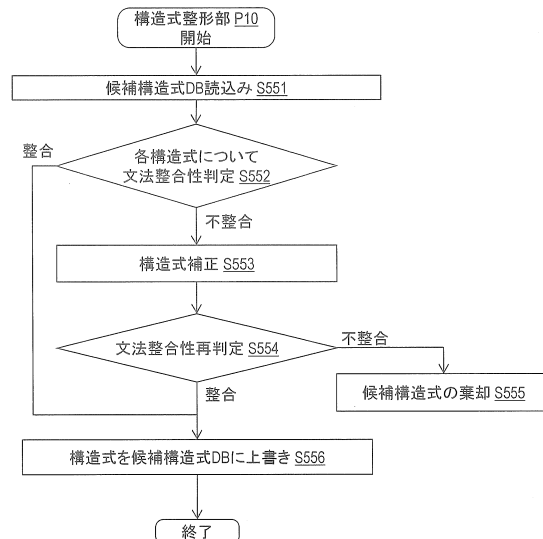
10

20

【図 19】



【図 20】

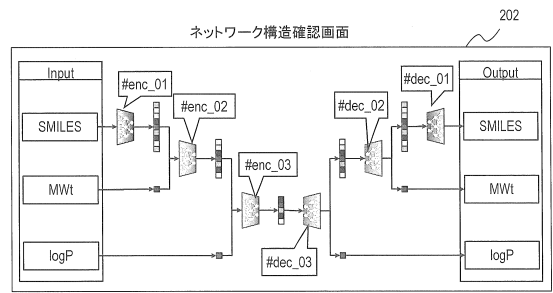


30

40

50

【図 2 1】



【図 2 2】

学習用データベース DB15

構造物性関係モデル内第1VAE用学習テーブル

TableID T8C1	TableType T8C2	ID T8C3	SMILES T8C4	MWt T8C5	構造式行列 T8C7
Tbl_R_201	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	[[0,0,...], [0,0,...], [0,0,...], ...]
Tbl_R_201	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	[[0,0,...], [0,0,...], [0,0,...], ...]
Tbl_R_201	Experiment	ZINC000001534133	O=C(O)CNC(=O)C(=O)O	147.086	[[0,0,...], [0,0,...], [0,0,...], ...]
Tbl_R_201	Experiment	ZINC000001555345	CNCC(N)=O	88.110	[[0,0,...], [0,0,...], [0,0,...], ...]
:	:	:	:	:	:

構造物性関係モデル内第2VAE用学習テーブル

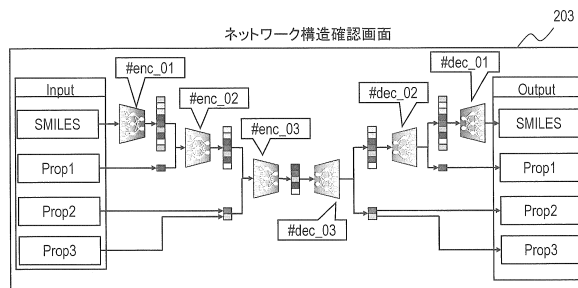
TableID T9C1	TableType T9C2	ID T9C3	SMILES T9C4	MWt T9C5	logP T9C6	構造式行列 T9C7
Tbl_R_202	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	-1.368	[[0,0,...], [0,0,...], [0,0,...], ...]
Tbl_R_202	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	-1.275	[[0,0,...], [0,0,...], [0,0,...], ...]
Tbl_R_202	Experiment	ZINC000001534133	O=C(O)CNC(=O)C(=O)O	147.086	-1.728	[[0,0,...], [0,0,...], [0,0,...], ...]
Tbl_R_202	Experiment	ZINC000001555345	CNCC(N)=O	88.110	-1.309	[[0,0,...], [0,0,...], [0,0,...], ...]
:	:	:	:	:	:	:

【図 2 3】

モデルテーブル

Network ID T10C1	Network Order T10C2	Nest Order T10C3	Target T10C4
#enc_01	1	1	{ "Tbl_G_100": ["構造式行列"] }
#enc_02	2	2	{ "Tbl_R_201": ["MWt"] }
#enc_03	3	3	{ "Tbl_R_202": ["logP"] }
#dec_03	4	3	{ "Tbl_R_202": ["logP"] }
#dec_02	5	2	{ "Tbl_R_201": ["MWt"] }
#dec_01	6	1	{ "Tbl_G_100": ["構造式行列"] }

【図 2 4】



10

20

30

40

50

【図 25】

学習用データベース DB15

構造物性関係モデル内第1VAE用学習テーブル

TableID T11C1	TableType T11C2	ID T11C3	SMILES T11C4	Prop1 T11C5	構造式行列 T11C7
Tbl_R_301	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	[...]
Tbl_R_301	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	[...]
:	:	:	:	:	:

155

構造物性関係モデル内第2VAE用学習テーブル

TableID T12C1	TableType T12C2	ID T12C3	SMILES T12C4	Prop1 T12C5	Prop2 T12C6	Prop3 T12C7	構造式 行列 T12C8
Tbl_R_302	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	-1.368	5.124	[...]
Tbl_R_302	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	-1.275	5.388	[...]
:	:	:	:	:	:	:	:

156

【図 26】

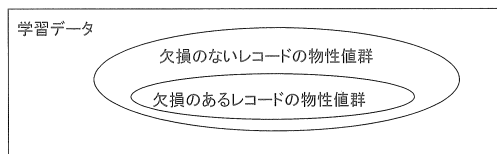
モデルテーブル

Network ID T13C1	Network Order T13C2	Nest Order T13C3	Target T13C4
#enc_01	1	1	{'Tbl_G_100': ['構造式行列']}
#enc_02	2	2	{'Tbl_R_301': ['Prop1']}
#enc_03	3	3	{'Tbl_R_302': ['Prop2', 'Prop3']}
#dec_03	4	3	{'Tbl_R_302': ['Prop2', 'Prop3']}
#dec_02	5	2	{'Tbl_R_301': ['Prop1']}
#dec_01	6	1	{'Tbl_G_100': ['構造式行列']}

173

10

【図 27】



【図 28】

構造物性関係モデル用学習データ

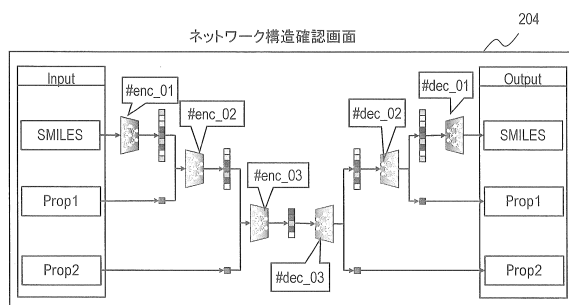
化学構造式	物性値 1 (Prop1)	物性値 2 (Prop2)
...
...
...
...
...
...

312

311

20

【図 29】



【図 30】

モデルテーブル

Network ID T14C1	Network Order T14C2	Nest Order T14C3	Target T14C4
#enc_01	1	1	{'Tbl_G_100': ['構造式行列']}
#enc_02	2	2	{'Tbl_R_401': ['Prop1']}
#enc_03	3	3	{'Tbl_R_402': ['Prop2']}
#dec_03	4	3	{'Tbl_R_402': ['Prop2']}
#dec_02	5	2	{'Tbl_R_401': ['Prop1']}
#dec_01	6	1	{'Tbl_G_100': ['構造式行列']}

174

30

40

50

【図 31】

150

TableID T15C1	TableType T15C2	ID T15C3	SMILES T15C4	Prop1 T15C5	Prop2 T15C6	構造式行列 T15C7
Tbl_Exp_011	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	-1.368	[...]
Tbl_Exp_011	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	-1.275	[...]
Tbl_Exp_013	Experiment	ZINC000001532449	O=C(O)CCCC(=O)O	123.044	-2.133	[...]
:	:	:	:	:	:	:
Tbl_Exp_012	Experiment	ZINC000001534133	O=C(O)CNC(=O)C(=O)O	147.086	Null	[...]
Tbl_Exp_012	Experiment	ZINC000001555345	CNCC(N)=O	88.110	Null	[...]
Tbl_Exp_012	Experiment	ZINC000001556606	NC(=O)N1CCNCCC1	129.163	Null	[...]
Tbl_Exp_013	Experiment	ZINC000001532451	O=C(OOC)CC(=OC)O	149.044	Null	[...]
:	:	:	:	:	:	:

157

構造物性関係モデル内第1VAE用学習テーブル

TableID T16C1	TableType T16C2	ID T16C3	SMILES T16C4	Prop1 T16C5	構造式行列 T16C7
Tbl_R_401	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	[...]
Tbl_R_401	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	[...]
Tbl_R_401	Experiment	ZINC000001532449	O=C(O)CCCC(=O)O	123.044	[...]
:	:	:	:	:	:
Tbl_R_401	Experiment	ZINC000001534133	O=C(O)CNC(=O)C(=O)O	147.086	[...]
Tbl_R_401	Experiment	ZINC000001555345	CNCC(N)=O	88.110	[...]
Tbl_R_401	Experiment	ZINC000001556606	NC(=O)N1CCNCCC1	129.163	[...]
Tbl_R_401	Experiment	ZINC000001532451	O=C(OOC)CC(=OC)O	149.044	[...]
:	:	:	:	:	:

158

構造物性関係モデル内第2VAE用学習テーブル

TableID T17C1	TableType T17C2	ID T17C3	SMILES T17C4	Prop1 T17C5	Prop2 T17C6	構造式行列 T17C7
Tbl_R_402	Experiment	ZINC000001532558	O=C(O)C(=O)CO	104.061	-1.368	[...]
Tbl_R_402	Experiment	ZINC000001532759	O=C(O)C(=O)C(=O)O	118.044	-1.275	[...]
Tbl_R_402	Experiment	ZINC000001532449	O=C(O)CCCC(=O)O	123.044	-2.133	[...]
:	:	:	:	:	:	:

10

20

30

40

50

フロントページの続き

- (56)参考文献 特開 2 0 2 0 - 9 2 0 3 (J P , A)
米国特許出願公開第 2 0 2 0 / 0 0 5 0 7 3 7 (U S , A 1)
特開 2 0 0 3 - 1 4 7 2 8 (J P , A)
- (58)調査した分野 (Int.Cl. , D B 名)
- G 1 6 C 1 0 / 0 0 - 9 9 / 0 0
G 0 6 N 2 0 / 0 0