(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116648753 A
(43)申请公布日 2023.08.25

(21)申请号 202080107270.6
(22)申请日 2020.12.18
(85)PCT国际申请进入国家阶段日
2023.05.17
(86)PCT国际申请的申请数据
PCT/JP2020/047562 2020.12.18
(87)PCT国际申请的公布数据
WO2022/130648 JA 2022.06.23
(71)申请人 富士通株式会社
地址 日本神奈川县川崎市
(72)发明人 片冈正弘 萩原稔 和田光人 松村量
(74)专利代理机构 北京三友知识产权代理有限公司 11127
专利代理师 朱丽娟
(51)Int.Cl.
G16C 20/10(2006.01)

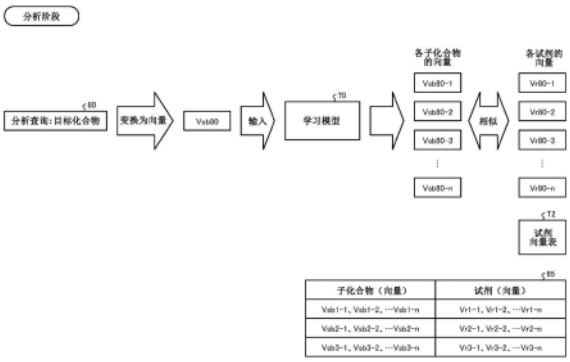
权利要求书3页 说明书15页 附图18页

(54)发明名称

信息处理程序、信息处理方法和信息处理装置

(57)摘要

信息处理装置基于学习数据，执行学习模型的学习，所述学习数据定义了与目标化合物对应的向量和与用于制造目标化合物的合成路径中包含的多个子化合物分别对应的向量之间的关系。信息处理装置在接受了分析对象的目标化合物的情况下，通过将分析对象的目标化合物的向量输入到学习模型，来计算与分析对象的目标化合物对应的多个子化合物的向量。



CN 116648753 A

1.一种信息处理程序,其特征在于,所述信息处理程序使计算机执行如下处理:

基于学习数据,执行学习模型的学习,所述学习数据定义了与目标化合物对应的向量和与用于制造所述目标化合物的合成路径中包含的多个子化合物分别对应的向量之间的关系;以及

在接受了分析对象的目标化合物的情况下,通过将所述分析对象的目标化合物的向量输入到所述学习模型,来计算与所述分析对象的目标化合物对应的多个子化合物的向量。

2.根据权利要求1所述的信息处理程序,其特征在于,所述信息处理程序还使计算机执行如下处理:根据通过进行所述计算的处理而计算出的多个子化合物的向量与成为代替候选的多个试剂的向量之间的相似程度,分析可代替所述分析对象的目标化合物的子化合物的试剂。

3.根据权利要求2所述的信息处理程序,其特征在于,在进行所述分析的处理中,检索试剂的示性式的信息作为所述可代替的试剂的信息,并输出检索结果。

4.根据权利要求1所述的信息处理程序,其特征在于,所述分析对象的目标化合物由组合了多个基团的信息来表示,所述信息处理方法还使计算机执行如下处理:通过对所述多个基团的向量进行累计,计算所述分析对象的目标化合物的向量。

5.一种信息处理程序,其特征在于,执行如下处理:

基于学习数据,执行学习模型的学习,所述学习数据定义了用于制造目标化合物的合成路径中包含的多个子化合物的载体与共同结构的载体之间的关系,所述共同结构表示子化合物的结构和试剂的结构中共同的结构;以及

在接受了分析对象的子化合物的情况下,通过将所述分析对象的子化合物的载体输入到所述学习模型,来计算与所述分析对象的子化合物对应的共同结构的载体。

6.根据权利要求5所述的信息处理程序,其特征在于,所述信息处理程序还执行如下处理:根据所述子化合物的载体与成为代替候选的多个试剂的载体之间的相似,检索与所述子化合物的载体相似的试剂的载体,根据检索出的试剂的载体与计算出的所述共同结构的载体,计算变换结构的载体,所述变换结构表示在所述子化合物的结构与检索出的试剂的结构中不同的部分的结构。

7.一种信息处理方法,所述信息处理方法由计算机执行,其特征在于,所述信息处理方法执行如下处理:

基于学习数据,执行学习模型的学习,所述学习数据定义了与目标化合物对应的载体和与用于制造所述目标化合物的合成路径中包含的多个子化合物分别对应的载体之间的关系;以及

在接受了分析对象的目标化合物的情况下,通过将所述分析对象的目标化合物的载体输入到所述学习模型,来计算与所述分析对象的目标化合物对应的多个子化合物的载体。

8.根据权利要求7所述的信息处理方法,其特征在于,所述信息处理方法还执行如下处理:根据通过进行所述计算的处理而计算出的多个子化合物的载体与成为代替候选的多个试剂的载体之间的相似程度,分析可代替所述分析对象的目标化合物的子化合物的试剂。

9.根据权利要求8所述的信息处理方法,其特征在于,在进行所述分析的处理中,检索试剂的示性式的信息作为所述可代替的试剂的信息,并输出检索结果。

10.根据权利要求9所述的信息处理方法,其特征在于,所述分析对象的目标化合物由

组合了多个基团的信息表示，所述信息处理方法还使计算机执行如下处理：通过对所述多个基团的载体进行累计来计算所述分析对象的目标化合物的载体。

11.一种信息处理方法，所述信息处理方法由计算机执行，其特征在于，所述信息处理方法执行如下处理：

基于学习数据，执行学习模型的学习，所述学习数据定义了用于制造目标化合物的合成路径中包含的多个子化合物的向量与共同结构的向量之间的关系，所述共同结构表示子化合物的结构和试剂的结构中共同的结构；以及

在接受了分析对象的子化合物的情况下，通过将所述分析对象的子化合物的向量输入到所述学习模型，来计算与所述分析对象的子化合物对应的共同结构的向量。

12.根据权利要求11所述的信息处理方法，其特征在于，所述信息处理方法还执行如下处理：

根据所述子化合物的向量与成为代替候选的多个试剂的向量之间的相似，检索与所述子化合物的向量相似的试剂的向量，根据检索出的试剂的向量与计算出的所述共同结构的向量，计算变换结构的向量，所述变换结构表示在所述子化合物的结构与检索出的试剂的结构中不同的部分的结构。

13.一种信息处理装置，其特征在于，所述信息处理装置具有：

学习部，其基于学习数据，执行学习模型的学习，所述学习数据定义了与目标化合物对应的向量和与用于制造所述目标化合物的合成路径中包含的多个子化合物分别对应的向量之间的关系；以及

计算部，其在接受了分析对象的目标化合物的情况下，通过将所述分析对象的目标化合物的向量输入到所述学习模型，来计算与所述分析对象的目标化合物对应的多个子化合物的向量。

14.根据权利要求13所述的信息处理装置，其特征在于，所述信息处理装置还具有分析部，该分析部根据由所述计算部计算出的多个子化合物的载体与成为代替候选的多个试剂的载体之间的相似程度，检索可代替所述分析对象的目标化合物的子化合物的试剂。

15.根据权利要求14所述的信息处理装置，其特征在于，分析部检索试剂的示性式的信息作为所述可代替的试剂的信息，并输出检索结果。

16.根据权利要求13所述的信息处理装置，其特征在于，检索对象的目标化合物由组合了多个基团的信息来表示，所述计算部还执行如下处理：通过对所述多个基团的载体进行累计来计算所述检索对象的目标化合物的载体。

17.一种信息处理装置，其特征在于，所述信息处理装置具有：

学习部，其基于学习数据，执行学习模型的学习，所述学习数据定义了用于制造目标化合物的合成路径中包含的多个子化合物的载体与共同结构的载体之间的关系，所述共同结构表示子化合物的结构和试剂的结构中共同的结构；以及

计算部，其在接受了分析对象的子化合物的情况下，通过将所述分析对象的子化合物的载体输入到所述学习模型，来计算与所述分析对象的子化合物对应的共同结构的载体。

18.根据权利要求17所述的信息处理装置，其特征在于，所述信息处理装置还具有分析部，该分析部根据所述子化合物的载体与成为代替候选的多个试剂的载体之间的相似，检索与所述子化合物的载体相似的试剂的载体，根据检索出的试剂的载体与计算出的所述共

同结构的载体,计算变换结构的载体,所述变换结构表示在所述子化合物的结构与检索出的试剂的结构中不同的部分的结构。

信息处理程序、信息处理方法和信息处理装置

技术领域

[0001] 本发明涉及信息处理程序等。

背景技术

[0002] 将自然界中存在的天然有机化合物作为药物开发的候选是非常有前景的，但是天然有机化合物是稀有的，难以直接使用该天然有机化合物来制造各种产品。因此，以廉价且容易获得的原料、试剂为基础，使用通用性高的变换反应，制造相当于稀有的天然有机化合物的有机化合物。在以下的说明中，将相当于天然有机化合物的有机化合物记为“目标化合物”。

[0003] 例如，存在如下现有技术：通过对天然有机化合物执行逆合成分析，设计表示为了制造目标化合物而进行变换反应的多个试剂(或原料)的组合、合成的顺序等的合成路径。基于通过该现有技术而设计的合成路径，使试剂依次反应，从而合成、制造目标化合物。

[0004] 图22是用于说明逆合成以及合成路径的一个例子的图。例如，对作为阿司匹林(镇痛剂)被公知的乙酰水杨酸1-1的逆合成进行说明。乙酰水杨酸1-1的官能团为酯和羧基。酯由羧酸和醇得到，因此乙酰水杨酸1-1之前的前体为水杨酸1-2，使用的试剂为乙酸酐。水杨酸1-2是由使二氧化碳在高压化的情况下与廉价的苯酚的钠盐反应的科尔贝-施密特(Kolbe-Schmitt)反应得到的，因此水杨酸的前体为苯酚1-3。基于该逆合成的结果，设计合成路径1-4，由苯酚1-3合成乙酰水杨酸1-1。

[0005] 现有技术文献

[0006] 专利文献

[0007] 专利文献1：日本特开2020-154442号公报

[0008] 专利文献2：日本特表2001-507675号公报

发明内容

[0009] 发明所要解决的问题

[0010] 为了制造目标化合物而通过逆合成分析得到的多个试剂在能够由具有相似特性的其他试剂代替的情况下，切换为容易获得且能够更廉价地进行变换反应的其他试剂来合成、制造目标化合物是有效的。但是，在现有技术中，难以从存在无数种的试剂的候选中筛选可替代的试剂来确定变换反应。

[0011] 在一个方面，本发明的目的在于提供一种信息处理程序、信息处理方法以及信息处理装置，能够检测与通过目标化合物的逆合成分析得到的试剂相似的试剂，并确定其变换反应。

[0012] 用于解决问题的手段

[0013] 在第一方案中，使计算机执行以下的处理。计算机基于学习数据，执行学习模型的学习，所述学习数据定义了与目标化合物对应的向量和与用于制造目标化合物的合成路径中包含的多个子化合物分别对应的向量之间的关系。计算机在接受了分析对象的目标化合

物的情况下,通过将分析对象的目标化合物的向量输入到学习模型,来计算与分析对象的目标化合物对应的多个子化合物的向量。

[0014] 发明效果

[0015] 可以检测与目标化合物的试剂相似的试剂。

附图说明

[0016] 图1是用于说明本实施例1的信息处理装置的学习阶段的处理的一例的图。

[0017] 图2是用于说明本实施例1的信息处理装置的分析阶段的处理的一例的图。

[0018] 图3是示出本实施例1的信息处理装置的结构的功能框图。

[0019] 图4是表示化学结构式文件的数据结构的一例的图。

[0020] 图5是示出基团辞典的一例的图。

[0021] 图6是示出试剂辞典的一例的图。

[0022] 图7A是示出子化合物辞典的一例的图。

[0023] 图7B是示出目标化合物辞典的一例的图。

[0024] 图7C是示出共同结构辞典的一例的图。

[0025] 图8是示出基团向量表的数据结构的一例的图。

[0026] 图9是示出试剂向量表的数据结构的一例的图。

[0027] 图10A是示出子化合物向量表的数据结构的一例的图。

[0028] 图10B是示出目标化合物向量表的数据结构的一例的图。

[0029] 图10C是示出共同结构向量表的数据结构的一例的图。

[0030] 图11是表示基团转置索引的数据结构的一例的图。

[0031] 图12是示出试剂转置索引的数据结构的一例的图。

[0032] 图13A是示出子化合物转置索引的数据结构的一例的图。

[0033] 图13B是示出目标化合物转置索引的数据结构的一例的图。

[0034] 图13C是示出共同结构转置索引的数据结构的一例的图。

[0035] 图14是示出逆合成分析表的数据结构的一例的图。

[0036] 图15是示出本实施例1的信息处理装置的处理过程的流程图(1)。

[0037] 图16是示出本实施例1的信息处理装置的处理过程的流程图(2)。

[0038] 图17是用于说明本实施例2的信息处理装置的学习阶段的处理的一例的图。

[0039] 图18是用于说明本实施例2的信息处理装置的处理的图。

[0040] 图19是示出本实施例2的信息处理装置的结构的功能框图。

[0041] 图20是示出本实施例2的信息处理装置的处理过程的流程图。

[0042] 图21是示出实现与实施例的信息处理装置同样的功能的计算机的硬件结构的一例的图。

[0043] 图22是用于说明逆合成和合成路径的一例的图。

具体实施方式

[0044] 在下文中,将参照附图详细说明本申请中公开的信息处理程序、信息处理方法和信息处理装置的实施例。然而,本发明不限于该实施例。

[0045] 实施例1

[0046] 将说明本实施例1的信息处理装置的处理的一例。本实施例1的信息处理装置通过预处理，执行计算目标化合物的向量(vector)的处理、分别计算与目标化合物对应的各子化合物(试剂)的向量的处理。需要说明的是，通过对目标化合物执行逆合成分析，设计用于制造目标化合物的合成路径，确定目标化合物与用于合成、制造该目标化合物的各试剂与变换反应的关系。

[0047] 图1是用于说明本实施例1的信息处理装置的学习阶段的处理的一例的图。如图1所示，信息处理装置使用学习数据65来执行学习模型70的学习。学习模型70对应于CNN(Convolutional Neural Network:卷积神经网络)、RNN(Recurrent Neural Network:循环神经网络)等。

[0048] 学习数据65定义目标化合物的向量与多个子化合物的向量之间的关系，该目标化合物具有逆合成分析和合成的实绩，该多个子化合物用于对目标化合物进行逆合成分析、合成。例如，目标化合物的向量对应于输入数据，多个子化合物的向量成为其输出数据的正解值。

[0049] 信息处理装置执行基于误差反向传播的学习，使得将目标化合物的向量输入到学习模型70时的输出接近各子化合物的向量。信息处理装置基于学习数据65所包含的目标化合物的向量与多个子化合物的向量的关系，反复执行上述处理，由此调整学习模型70的参数(执行机器学习)。

[0050] 图2是用于说明本实施例1的信息处理装置的分析阶段的处理的一例的图。信息处理装置在分析阶段中，使用在学习阶段中学习过的学习模型70来执行接下来的处理。

[0051] 信息处理装置当接受了指定了目标化合物的分析查询80时，将分析查询80的目标化合物变换为向量Vob80。信息处理装置通过将向量Vob80输入到学习模型70，来计算与各子化合物对应的多个向量(Vsb80-1、Vsb80-2、Vsb80-3、…Vsb80-n)。

[0052] 信息处理装置比较在试剂向量表T2中存储的与各试剂对应的多个向量(Vr80-1、Vr80-2、Vr80-3、・・・Vr80-n)和与各子化合物对应的多个向量(Vsb80-1、Vsb80-2、Vsb80-3、・・・Vsb80-n)的相似度，分析相似的子化合物和试剂。信息处理装置将相似的子化合物的向量与试剂的向量对应起来登记在子化合物・试剂表85中。

[0053] 如上所述，本实施例1的信息处理装置基于定义了目标化合物的向量与基于逆合成分析的各子化合物的向量的关系的学习数据65，执行学习模型70的学习。信息处理装置通过向学习完毕的学习模型70输入分析查询的向量，来计算与分析查询的目标化合物对应的各子化合物的向量。通过使用从学习模型70输出的各子化合物的向量，能够容易地执行检测与在目标化合物的合成路径中定义的子化合物相似的各试剂的处理。

[0054] 接下来，将说明本实施例1的信息处理装置的结构的一例。图3是示出本实施例1的信息处理装置的结构的功能框图。如图3所示，信息处理装置100包括通信部110、输入部120、显示部130、存储部140和控制部150。

[0055] 通信部110通过有线或无线与外部装置等连接，在与外部装置等之间进行信息的收发。例如，通信部110通过NIC(Network Interface Card:网络接口卡)等实现。通信部110也可以与未图示的网络连接。

[0056] 输入部120是将各种信息输入到信息处理装置100的输入装置。输入部120对应于

键盘、鼠标、触摸面板等。

[0057] 显示部130是显示从控制部150输出的信息的显示装置。显示部130对应于液晶显示器、有机EL(Electro Luminescence:电致发光)显示器、触摸面板等。

[0058] 存储部140具有化学结构式文件50、基团编码文件51、试剂编码文件52、子化合物编码文件53、目标化合物编码文件54、共同结构编码文件55。存储部140具有基团辞典D1、试剂辞典D2、子化合物辞典D3、目标化合物辞典D4、共同结构辞典D5。存储部140具有基团向量表T1、试剂向量表T2、子化合物表T3、目标化合物向量表T4、共同结构向量表T5。存储部140具有基团转置索引In1、试剂转置索引In2、子化合物转置索引In3、目标化合物索引In4、共同结构索引In5。存储部140具有逆合成分析结果表60、学习数据65、学习模型70、分析查询80、子化合物·试剂表85。

[0059] 存储部140例如通过RAM(Random Access Memory:随机存取存储器)、闪存(Flash Memory:闪存)等半导体存储器元件、或者硬盘、光盘等存储装置来实现。

[0060] 化学结构式文件50是包含多个官能团的示性式的信息,通过组合最小单位的官能团的示性式,成为一次结构或二次结构的示性式。在本实施例1中,作为一例,以一次结构的示性式对应于“子化合物”或“试剂”、二次结构(或高次结构)的示性式对应于“目标化合物(或天然有机化合物)”的情况进行说明。

[0061] 例如,化学结构式文件50分为记述了与各子化合物(或试剂)对应的示性式的子化合物(试剂)记述区域和记述了与各目标化合物对应的示性式的目标化合物记述区域。另外,化学结构式文件50也可以包含后述的逆合成分析结果表60的信息。

[0062] 图4是表示化学结构式文件的数据结构的一例的图。示性式(化学结构式)是表示构成化合物的元素的排列的式,可以用SMILES法等记述。

[0063] 官能团的基团编码文件51是将化学结构式文件50以基团为单位进行压缩而得到的文件。如后所述,基于化学结构式文件50和基团辞典D1生成基团编码文件51。

[0064] 试剂编码文件52是基于基团编码文件51的试剂压缩区域而生成的文件,是以试剂为单位压缩后的文件。1个试剂的压缩码与多个基团的压缩码的组合对应。如后所述,试剂编码文件52基于试剂压缩区域的压缩码和试剂辞典D2生成。

[0065] 子化合物编码文件53是基于基团编码文件51生成的文件,是以子化合物为单位压缩后的文件。1个子化合物的压缩码对应于多个基团的压缩码的组合。如后所述,子化合物编码文件53基于子化合物压缩区域的压缩码和子化合物辞典D3而生成。

[0066] 目标化合物编码文件54是基于基团编码文件51的目标化合物压缩区域而生成的文件,并且是以目标化合物为单位而压缩的文件。1个目标化合物的压缩码对应于多个基团的压缩码的组合。如后所述,目标化合物编码文件54是基于目标化合物压缩区域的压缩码和目标化合物辞典D4而生成的。

[0067] 共同结构编码文件55是基于基团编码文件51而生成的文件,并且是以共同结构为单位而压缩的文件。1个共同结构的压缩码对应于多个基团的压缩码的组合。如后所述,共同结构编码文件55是基于共同结构区域的压缩编码和共同结构辞典D5而生成的。

[0068] 基团辞典D1用示性式定义了基团的压缩码和构成基团的元素的排列。图5是表示基团辞典的一例的图。如图5所示,基团辞典D1将压缩码、名称和示性式对应起来。压缩码是分配给基团的压缩码。名称是相应的基团的名称的一例。示性式表示作为相应的基团的示

性式的排列。

[0069] 例如对“甲基”分配压缩码“D0008000h”。与压缩码“D0008000h”对应的示性式为“CH3”。“h”是表示压缩码为16进制数的记号。

[0070] 试剂辞典D2定义试剂的压缩码与构成该试剂的多个基团的压缩码的组合的关系。图6是表示试剂辞典的一例的图。如图6所示，试剂辞典D2将压缩码、名称和基团码序列对应起来。压缩码是分配给试剂的压缩码。名称是相应的试剂的名称的一例。基团码序列是组合了多个基团的压缩码的码序列。

[0071] 子化合物辞典D3定义目标化合物的压缩码与构成该目标化合物的多个基团的压缩码的组合的关系。图7A是表示子化合物辞典的一例的图。如图7A所示，子化合物辞典D3将压缩码、名称和基团码序列对应起来。压缩码是分配给子化合物的压缩码。名称是相应的子化合物的名称的一例。基团码序列是组合了多个基团的压缩码的码序列。

[0072] 目标化合物辞典D4定义目标化合物的压缩码与构成该目标化合物的多个基团的压缩码的组合的关系。图7B是表示目标化合物辞典的一例的图。如图7B所示，目标化合物辞典D4将压缩码、名称和基团码序列对应起来。压缩码是分配给目标化合物的压缩码。名称是相应的目标化合物的名称的一例。基团码序列是组合了多个基团的压缩码的码序列。

[0073] 共同结构辞典D5是多个试剂所包含的结构中共同的结构。共同结构辞典D5定义共同结构的压缩码与构成该共同结构的多个基团的压缩码的组合之间的关系。图7C是表示共同结构辞典的一例的图。如图7C所示，共同结构辞典D5将压缩码、名称和基团码序列对应起来。压缩码是分配给共同结构的压缩码。名称是相应的共同结构的名称的一例。基团码序列是组合了多个基团的压缩码的码序列。

[0074] 基团向量表T1是定义基团的向量的表。图8是表示基团向量表的数据结构的一例的图。如图8所示，在基团向量表T1中，将基团压缩码与分配给基团压缩码的向量对应起来。基团的向量是通过庞加莱嵌入(Poincaréembeddings)而算出的。

[0075] 试剂向量表T2是定义试剂的向量的表。图9是表示试剂向量表的数据结构的一例的图。如图9所示，该试剂向量表T2将试剂的压缩码与分配给该试剂的压缩码的向量对应起来。试剂的向量是对构成试剂的基团的压缩码的向量进行累计而得到的。试剂向量表T2也可以进一步对应地保持试剂的名称、试剂的示性式等特征。

[0076] 子化合物向量表T3是定义子化合物的向量的表。图10A是表示子化合物向量表的数据结构的一例的图。如图10A所示，该子化合物向量表T3将子化合物的压缩码与分配给该子化合物的压缩码的向量对应起来。子化合物的向量是对构成子化合物的基团的压缩码的向量进行累计而得到的。子化合物向量表T3也可以进一步将子化合物的名称、子化合物的示性式等特征对应起来保持。

[0077] 目标化合物向量表T4是定义目标化合物的向量的表。图10B是表示目标化合物向量表的数据结构的一例的图。如图10B所示，该目标化合物向量表T3将目标化合物的压缩码与分配给目标化合物的压缩码的向量对应起来。目标化合物的向量是对构成目标化合物的基团的压缩码的向量进行累计而得到的。

[0078] 共同结构向量表T5是定义共同结构的向量的表。图10C是示出共同结构向量表的数据结构的一个例子的图。如图10C所示，在共同结构向量表T5中，将共同结构的压缩码与分配给共同结构的压缩码的向量对应起来。共同结构的向量是对构成共同结构的基团的压

缩码的向量进行累计而得到的。

[0079] 基团转置索引In1针对基团的压缩码，表示基团编码文件51的出现位置(偏移)。图11是表示基团转置索引的数据结构的一例的图。如图11所示，基团转置索引In1的横轴是与偏移对应的轴。基团转置索引In1的纵轴是与基团的压缩码对应的轴。基团转置索引In1用“0”或“1”的位图来表示，在初始状态下全部的位图被设定为“0”。

[0080] 例如，将基团编码文件51的开头的基团的压缩码的偏移设为“0”。在从基团编码文件51的开头起第二个的位置中包含基团码“D008000h(甲基)”的情况下，基团转置索引In1的偏移为“1”的列与基团压缩码“D008000h(甲基)”的行交叉的位置的比特成为“1”。

[0081] 试剂转置索引In2表示针对试剂的压缩码，试剂编码文件52的出现位置(偏移)。图12是表示试剂转置索引的数据结构的一例的图。如图12所示，试剂转置索引In2的横轴是与偏移对应的轴。试验药转置索引In2的纵轴是与试剂的压缩码对应的轴。试剂转置索引In2用“0”或“1”的位图表示，在初始状态下全部的位图被设定为“0”。

[0082] 例如，将试剂编码文件52的开头的试剂的压缩码的偏移设为“0”。在试剂编码文件52的从开头起第九个的位置包含试剂的码“D0008000h”的情况下，试剂转置索引In2的偏移为“8”的列与试剂的压缩码“D0008000h”的行交叉的位置的比特成为“1”。

[0083] 子化合物转置索引In3表示针对子化合物的压缩码，子化合物编码文件53的出现位置(偏移)。图13A是表示子化合物转置索引的数据结构的一例的图。如图13A所示，子化合物转置索引In3的横轴是与偏移对应的轴。子化合物转置索引In3的纵轴是与子化合物的压缩码对应的轴。子化合物转置索引In3用“0”或“1”的位图表示，在初始状态下全部的位图被设定为“0”。

[0084] 例如，设子化合物编码文件53的开头的子化合物的压缩码的偏移为“0”。在子化合物编码文件53的从开头起的第九个的位置包含子化合物的码“D0008000h”的情况下，子化合物转置索引In3的偏移为“8”的列与子化合物的压缩码“D0008000h”的行交叉的位置的比特成为“1”。

[0085] 目标化合物转置索引In4表示针对目标化合物的压缩码，目标化合物编码文件54的出现位置(偏移)。图13B是表示目标化合物转置索引的数据结构的一例的图。如图13B所示，目标化合物转置索引In4的横轴是与偏移对应的轴。目标化合物药物转置索引In4的纵轴是与目标化合物的压缩码对应的轴。目标化合物转置索引In4由“0”或“1”的位图表示，在初始状态下全部的位图被设定为“0”。

[0086] 例如，将目标化合物编码文件54的开头的目标化合物的压缩码的偏移设为“0”。当在从目标化合物编码文件54的开头起的第九个的位置处包含目标化合物的码“D0008000h”时，目标化合物转置索引In4的偏移为“8”的列与目标化合物的压缩码“D0008000h”的行相交的位置处的比特是“1”。

[0087] 共同结构转置索引In5表示针对共同结构的压缩码，共同结构编码文件55的出现位置(偏移)。图13C是表示共同结构转置索引的数据结构的一例的图。如图13C所示，共同结构转置索引In5的横轴是与偏移对应的轴。共同结构转置索引In5的纵轴是与共同结构的压缩码对应的轴。共同结构转置索引In5由“0”或“1”的位图表示，在初始状态下全部的位图被设定为“0”。

[0088] 例如，将共同结构编码文件55的开头的共同结构的压缩码的偏移设为“0”。在从共

同结构编码文件55的开头起的第九个的位置处包含共同结构的码“D0008000h”的情况下，共同结构转置索引In4的偏移为“8”的列与子化合物的压缩码“D0008000h”的行交叉的位置的比特成为“1”。

[0089] 逆合成分析结果表60保存通过对目标化合物(相当于目标化合物的天然有机化合物)执行逆合成分析而得到的信息(合成路径)。图14是表示逆合成分析结果表的数据结构的一例的图。如图14所示，该逆合成分析结果表60将目标化合物的名称与通过对该目标化合物进行逆合成分析而得到的合成路径对应起来。在合计路径中包含在合成路径的中途进行反应的各试剂的名称。

[0090] 另外，在图14中，对将目标化合物的名称与各子化合物(试剂)的名称对应起来的情况进行了说明，但并不限定于此，也可以通过示性式将目标化合物与各子化合物(试剂)的名称对应起来。另外，逆合成分析结果表60的信息也可以是化学结构式文件50的一部分。

[0091] 学习数据65定义目标化合物的向量与用于制造目标化合物的多个子化合物(试剂)的向量的关系。学习数据65的数据结构对应于图1中说明的学习数据的数据结构。

[0092] 学习模型70是与CNN、RNN等对应的模型，参数被设定。

[0093] 分析查询80中包含成为试剂的分析对象的目标化合物的示性式的信息。

[0094] 子化合物·试剂表85是将相似的子化合物的向量与试剂的向量对应起来进行保存的表。子化合物·试剂表85的数据结构对应于图2中说明的子化合物·试剂表的数据结构。

[0095] 回到图3的说明。控制部150具有预处理部151、学习部152、计算部153、分析部154。控制部150例如由CPU(Central Processing Unit:中央处理单元)、MPU(Micro Processing Unit:微处理单元)实现。另外，控制部150例如也可以通过ASIC(Application Specific Integrated Circuit:专用集成电路)、FPGA(Field Programmable Gate Array:现场可编程门阵列)等集成电路来执行。

[0096] 预处理部151通过执行下述的各种处理，计算目标化合物的向量及子化合物(试剂)的向量等。

[0097] 例如，预处理部151执行生成基团编码文件51的处理、生成基团向量表T1、基团转置索引In1的处理、生成试剂编码文件52、试剂向量表T2、试剂转置索引In2的处理。预处理部151执行生成子化合物编码文件53、子化合物向量表T3、子化合物转置索引In3的处理。预处理部151执行生成目标化合物编码文件54、目标化合物向量表T4和目标化合物转置索引In4的处理。预处理部151执行生成学习数据65的处理。

[0098] 对预处理部151生成基团编码文件51的处理的一例进行说明。预处理部151基于化学结构式文件50和基团辞典D1，反复执行确定化学结构式文件50中包含的基团的示性式，将所确定的基团的示性式置换为压缩码的处理，由此生成基团编码文件51。例如，基团编码文件51中包含试剂压缩区域、子化合物压缩区域和目标化合物压缩区域。

[0099] 预处理部151通过对基团编码文件51的试剂记述区域中包含的各示性式执行上述处理，生成试剂压缩区域的基团编码序列。预处理部151通过对基团编码文件51的子化合物记述区域中包含的各示性式执行上述处理，生成子化合物压缩区域的基团编码序列。预处理部151通过对基团编码文件51的目标化合物记述区域中包含的各个示性式执行上述处理来生成目标化合物压缩区域的基团编码序列。

[0100] 对预处理部151生成基团向量表T1、基团转置索引In1的处理的一例进行说明。预处理部151在生成基团向量表T1时，执行庞加莱嵌入。

[0101] 预处理部151通过将基团的压缩码嵌入到庞加莱空间，来计算基团(基团的压缩码)的向量。嵌入到庞加莱空间中来计算向量的处理是被称为庞加莱嵌入(Poincare Embeddings)的技术。庞加莱嵌入例如使用非专利文献“Valentin Khrulkov1 et al. “Hyperbolic Image Embeddings”Cornell Universality，2019April 3”等中记载的技术即可。

[0102] 在庞加莱嵌入中，根据嵌入到庞加莱空间中的位置来分配向量，另外，具有越是相似的信息，则嵌入到越近的位置的特征。因此，具有相似的特征的各基团在庞加莱空间中被嵌入到各自接近的位置，因此被分配相似的向量。虽然省略了图示，但是预处理部151参照定义了相似的基团彼此的基团相似表，将各基团的压缩码嵌入到庞加莱空间，计算各基团的压缩码的向量。预处理部151也可以预先对在基团辞典D1中定义的各基团的压缩码执行庞加莱嵌入。

[0103] 预处理部151通过将基团(基团的压缩码)与基团的向量对应起来来生成基团向量表T1。预处理部151基于基团的向量与基团编码文件51中的基团(基团的压缩码)的位置之间的关系，生成基团转置索引In1。

[0104] 对预处理部151生成试剂编码文件52、试剂向量表T2、试剂转置索引In2的处理的一例进行说明。预处理部151基于基团编码文件51中包含的试剂压缩区域的基团编码序列和试剂辞典D2，反复执行将与试剂对应的基团编码序列置换为试剂的压缩编码的处理，由此生成试剂编码文件52。

[0105] 预处理部151通过比较与试剂对应的基团编码序列和基团向量表T1，确定基团编码序列中包含的各基团的压缩码，通过对所确定的各基团的压缩码的向量进行累计，计算与试剂对应的向量。

[0106] 预处理部151通过将试剂(试剂的压缩码)与试剂的向量对应起来，生成试剂向量表T2。预处理部151基于试剂的向量与试剂编码文件52中的试剂(试剂的压缩码)的位置之间的关系，生成试剂转置索引In2。

[0107] 对预处理部151生成子化合物编码文件53、子化合物向量表T3、子化合物转置索引In3的处理的一例进行说明。预处理部151反复执行基于基团编码文件51所包含的子化合物压缩区域的基团编码序列和子化合物辞典D3，将与子化合物对应的基团编码序列替换为子化合物的压缩码的处理，由此生成子化合物编码文件53。

[0108] 预处理部151通过对与子化合物对应的基团编码序列和基团向量表T1进行比较，来确定基团编码序列中包含的各基团的压缩码，通过对所确定的各基团的压缩码的向量进行累计，来计算与子化合物对应的向量。

[0109] 预处理部151通过将子化合物(子化合物的压缩码)与子化合物的向量对应起来，生成子化合物向量表T3。预处理部151基于子化合物的向量与子化合物编码文件53中的子化合物(子化合物的压缩码)的位置之间的关系，生成子化合物转置索引In3。

[0110] 说明预处理部151生成目标化合物编码文件54、目标化合物向量表T4和目标化合物转置索引In4的处理的一例。预处理部151反复执行基于基团编码文件51中包含的目标化合物压缩区域的基团编码序列和目标化合物辞典D4，将与目标化合物对应的基团编码序列

置换为目标化合物的压缩码的处理，由此生成目标化合物编码文件54。

[0111] 预处理部151通过对与目标化合物对应的基团编码序列和基团向量表T1进行比较，来确定基团编码序列所包含的各基团的压缩码，通过对所确定的各基团的压缩码的向量进行累计，来计算与目标化合物对应的向量。

[0112] 预处理部151通过将目标化合物(目标化合物的压缩码)与目标化合物的向量对应起来生成目标化合物向量表T4。预处理部151基于目标化合物的向量与目标化合物编码文件54中的目标化合物(目标化合物的压缩码)的位置之间的关系来生成目标化合物转置索引In4。

[0113] 预处理部151也可以生成共同结构编码文件55、共同结构向量表T5和共同结构转置索引In5。预处理部151反复执行基于基团编码文件51中包含的共同结构区域的基团编码序列和共同结构辞典D5，将与共同结构对应的基团编码序列置换为共同结构的压缩码的处理，由此生成共同结构编码文件55。

[0114] 预处理部151通过比较与共同结构对应的基团编码序列和基团向量表T1，确定基团编码序列中包含的各基团的压缩码，通过对确定出的各基团的压缩码的向量进行累计，计算与共同结构对应的向量。

[0115] 预处理部151通过将共同结构(共同结构的压缩码)与共同结构的向量对应起来而生成共同结构向量表T5。预处理部151基于共同结构的向量与共同结构编码文件55中的共同结构(共同结构的压缩码)的位置之间的关系来生成共同结构索引In5。

[0116] 对预处理部151生成学习数据65的处理的一例进行说明。预处理部151基于逆合成分析结果表60，确定目标化合物的名称与在该目标化合物的合成路径中进行反应的多个子化合物(试剂)的名称之间的关系。预处理部151基于目标化合物的名称和目标化合物向量表T4来确定目标化合物的向量。预处理部151基于各子化合物(试剂)的名称和试剂向量表T2(或者，子化合物向量表T3)，确定子化合物(试剂)的向量。预处理部151通过该处理，确定目标化合物的向量与在目标化合物的合成路径中进行反应的各子化合物(试剂)的向量之间的关系，并登记到学习数据65中。

[0117] 预处理部151通过对逆合成分析结果表60的各记录(目标化合物的名称、各子化合物(试剂)的名称)反复执行上述的处理，生成学习数据65。

[0118] 返回到图3的说明。学习部152使用学习数据65来执行学习模型70的学习。学习部152的处理对应于图1中说明的处理。学习部152从学习数据65取得目标化合物的向量和与该目标化合物的向量对应的各子化合物(试剂)的向量的组。学习部152以使得将目标化合物的向量输入到学习模型70的情况下的、学习模型70的输出的值接近各子化合物(试剂)的向量的值的方式，执行基于误差反向传播的学习，由此调整学习模型70的参数。

[0119] 学习部152通过对学习数据65的目标化合物的向量和各子化合物(试剂)的向量的组反复执行上述处理，来执行学习模型70的学习。

[0120] 计算部153在接受了分析查询80的指定的情况下，使用学习完毕的学习模型70，计算在分析查询80的目标化合物的合成路径中进行反应的各子化合物的向量。计算部153的处理对应于图2中说明的处理。计算部153可以从输入部120接受分析查询80，也可以经由通信部110从外部装置接受分析查询80。

[0121] 计算部153取得分析查询80中包含的目标化合物的示性式。计算部153将目标化合

物的示性式与基团辞典D1进行比较，确定目标化合物的示性式中包含的基团，以基团为单位将目标化合物的示性式分别变换为压缩码。

[0122] 计算部153将变换后的各基团的压缩码与基团向量表T1进行比较，确定各基团的压缩码的向量。计算部153通过对确定出的各基团的压缩码的向量进行累计，来计算与分析查询80所包含的目标化合物对应的向量Vob80。

[0123] 计算部153通过将向量Vob80输入到学习模型70，来计算与各子化合物(试剂)对应的多个向量。计算部153将计算出的各子化合物的向量输出到分析部154。

[0124] 在以下的说明中，将计算部153计算出的各子化合物(试剂)的向量分别表述为“分析向量”。

[0125] 分析部154基于分析向量，检索具有与分析向量相似的向量的试剂的信息。分析部154基于检索结果，将构成目标化合物的各子化合物的向量与相似的各试剂的向量(以下所示的相似向量)对应起来登记在子化合物・试剂表85中。

[0126] 例如，分析部154分别计算分析向量与试剂向量表T2中包含的各向量之间的距离，确定与分析向量之间的距离小于阈值的向量。试剂向量表T2中包含的、与分析向量的距离小于阈值的向量为“相似向量”。

[0127] 分析部154基于试剂向量表T2，确定与相似向量对应的试剂的压缩码，基于所确定的试剂的压缩码、试剂辞典D2、以及基团辞典D1，确定与试剂的压缩码对应的示性式。另外，试剂的特征也可以与试剂向量表T2对应起来，在该情况下，分析部154取得与相似向量对应的试剂的特征。分析部154通过执行该处理，检索与相似向量对应的试剂的示性式、试剂的特征，将检索到的结果登记在子化合物・试剂表85中。

[0128] 分析部154也可以通过对各分析向量反复执行上述处理，按每个分析向量检索与相似向量对应的试剂的示性式、试剂的特征，并登记在子化合物・试剂表85中。分析部154既可以将子化合物・试剂表85输出到显示部130进行显示，也可以发送到与网络连接的外部装置。

[0129] 接下来，说明本实施例1的信息处理装置100的处理过程的一例。图15是示出本实施例1的信息处理装置的处理过程的流程图(1)。如图15中所示，信息处理装置100的预处理部151通过执行庞加莱嵌入来计算各基团的压缩码的向量(步骤S101)。

[0130] 预处理部151基于化学结构式文件50以及基团辞典D1，生成基团编码文件51、基团向量表T1、基团转置索引In1(步骤S102)。

[0131] 预处理部151基于基团编码文件51以及子化合物辞典D3，生成子化合物编码文件53、子化合物向量表T3、子化合物转置索引In3(步骤S103)。

[0132] 预处理部151基于基团编码文件51和目标化合物辞典而生成目标化合物编码文件54、目标化合物向量表T4和目标化合物转置索引In4(步骤S104)。

[0133] 预处理部151基于逆合成分析结果表60，确定目标化合物的向量与用于制造该目标化合物的各子化合物(试剂)的向量之间的关系，生成学习数据65(步骤S105)。

[0134] 信息处理装置100的学习部152基于学习数据65来执行学习模型的学习(步骤S106)。

[0135] 图16是表示本实施例1的信息处理装置的处理步骤的流程图(2)。信息处理装置100的计算部153接受分析查询80(步骤S201)。

[0136] 计算部153基于分析查询80所包含的目标化合物的示性式，计算目标化合物的向量(步骤S202)。

[0137] 计算部153通过将计算出的目标化合物的向量输入到学习完毕的学习模型70，来计算各子化合物的向量(步骤S203)。计算部153输出各子化合物的向量和各子化合物(步骤S204)。

[0138] 分析部154使用从学习模型70输出的各子化合物的向量和试剂向量表T2，检索与构成目标化合物的各子化合物相似的各试剂的向量，生成子化合物·试剂表85(步骤S205)。

[0139] 接下来，说明本实施例1的信息处理装置100的效果。信息处理装置100在学习阶段，基于学习数据65来执行学习模型70的学习，该学习数据65定义了目标化合物的向量与基于逆合成分析的各子化合物(试剂)的向量之间的关系。信息处理装置100在分析阶段，通过向学习完毕的学习模型70输入分析查询的向量，来计算与分析查询的目标化合物对应的各子化合物(试剂)的向量。通过使用从学习模型70输出的各子化合物(试剂)的向量，能够容易地检测与在目标化合物的合成路径中定义的子化合物相似的试剂。

[0140] 作为官能团的二次结构的目标化合物由作为多个官能团的一次结构的子化合物构成。另外，构成子化合物的多个各官能团的向量的推移缓慢，但子化合物的末尾的官能团的向量与持续的子化合物的开头的官能团的向量背离的情况较多。通过基于具有实际结果的逆合成分析后的目标化合物的官能团的二次结构的向量和子化合物的官能团的一次结构的向量进行机器学习，能够提高有机化合物的逆合成分析的精度。

[0141] 实施例2

[0142] 图17是用于说明本实施例2的信息处理装置的学习阶段的处理的一例的图。如图17所示，信息处理装置使用学习数据90来执行学习模型91的学习。学习模型91对应于CNN、RNN等。

[0143] 学习数据90定义合成目标化合物的多个子化合物的向量与基于试剂在变换反应中保持的共同结构的向量之间的关系。例如，子化合物的向量对应于输入数据，多个共同结构的向量成为正解值。

[0144] 信息处理装置以使得将子化合物的向量输入到学习模型91时的输出接近各共同结构的向量的方式，执行基于误差反向传播的学习。信息处理装置基于学习数据90所包含的子化合物的向量与共同结构的向量的关系，反复执行上述处理，由此调整学习模型91的参数(执行机器学习)。

[0145] 图18是用于说明本实施例2的信息处理装置的处理的图。实施例2的信息处理装置也可以与实施例1的信息处理装置100同样地预先学习学习模型70。另外，如图17中说明的那样，信息处理装置学习与学习模型70不同的学习模型91。学习模型70在输入了分析查询(目标化合物)80的向量的情况下，输出各子化合物的向量。学习模型90在输入了分析查询(子化合物)92的向量的情况下，输出共同结构的向量。

[0146] 信息处理装置当接受了指定子化合物的分析查询92时，使用子化合物向量表T3，将分析查询92的子化合物变换为向量Vsb92-1。信息处理装置通过将子化合物的向量Vsb92-1输入到学习模型91，来计算与共同结构对应的向量Vcm92-1。

[0147] 在此，信息处理装置将子化合物的向量Vsb92-1与试剂向量表T2中包含的多个试

剂的向量进行比较。试剂向量表T2对应于实施例1中说明的试剂向量表T2。

[0148] 信息处理装置针对子化合物的向量Vsb92-1，确定相似的试剂的向量。例如，将与子化合物的向量Vsb92-1相似的试剂的向量设为Vr92-1。于是，可知在向量Vsb92-1的子化合物和向量Vr92-1的试剂中共同的共同结构的向量成为从学习模型91输出的向量Vcm92-1。另外，从试剂的向量Vr92-1减去共同结构的向量Vcm92-1的结果成为在相似的试剂和子化合物中不同的不同结构的向量(变换结构的向量)。

[0149] 信息处理装置将共同结构的向量与变换结构的向量之间的关系登记在共同结构·变换结构表93中。信息处理装置通过对各子化合物的向量反复执行上述处理，生成共同结构·变换结构表93。

[0150] 另外，信息处理装置也可以使用“子化合物的向量－共同结构的向量＝试剂的向量－共同结构的向量+变换结构向量”的关系，计算变换结构的向量。

[0151] 如上所述，本实施例2的信息处理装置向学习完毕的学习模型91输入分析查询92的向量，计算与分析查询的子化合物对应的各共同结构的向量。另外，通过从与子化合物相似的试剂的各向量中减去共同结构的向量，计算出在相似的子化合物和试剂中不同的变换结构的向量。通过使用上述共同结构的向量、变换结构的向量，能够容易地分析能够用于目标化合物的合成、制造的更好的试剂。

[0152] 下来，将说明本实施例2的信息处理装置的结构的一例。图19是示出本实施例2的信息处理装置的结构的功能框图。如图19所示，信息处理装置200包括通信部210、输入部220、显示部230、存储部240和控制部250。

[0153] 关于通信部210、输入部220、显示部230的说明与实施例1中说明的关于通信部110、输入部120、显示部130的说明相同。

[0154] 存储部240具有化学结构式文件50、基团编码文件51、试剂编码文件52、子化合物编码文件53、目标化合物编码文件54、共同结构编码文件55。存储部240具有基团辞典D1、试剂辞典D2、子化合物辞典D3、目标化合物辞典D4、共同结构辞典D5。存储部240具有基团向量表T1、试剂向量表T2、子化合物向量表T3、目标化合物表T4、共同结构向量表T5。存储部240具有基团转置索引In1、试剂转置索引In2、子化合物转置索引In3、目标化合物索引In4、共同结构索引In5。存储部240具有逆合成分析结果表60、学习数据90、学习模型91、分析查询92。存储部240具有共同结构·变换结构表93。

[0155] 存储部240例如通过RAM、闪存等半导体存储器元件、或者硬盘、光盘等存储装置来实现。

[0156] 关于化学结构式文件50、基团编码文件51、试剂编码文件52、子化合物编码文件53、目标化合物编码文件54、共同结构编码文件55的说明与实施例1中说明的内容相同。关于基团辞典D1、试剂辞典D2、子化合物辞典D3、目标化合物辞典D4、共同结构辞典D5的说明与实施例1中说明的内容相同。关于基团向量表T1、试剂向量表T2、子化合物向量表T3、目标化合物表T4、共同结构向量表T5的说明与实施例1中说明的内容相同。关于基团转置索引In1、试剂转置索引In2、子化合物转置索引In3、目标化合物索引In4、共同结构索引In5的说明与在实施例1中说明的内容相同。逆合成分析结果表60与实施例1中说明的内容相同。学习数据90与图17中说明的内容相同。关于学习模型91、分析查询92的说明与在图18中说明的内容相同。

[0157] 如在图18中说明的那样,共同结构·变换结构表93包含用于从与共同结构向量相似的试剂变换反应为子化合物的变换结构向量的信息。在图18中,例如,在共同结构·变换结构表93中包含与Vcm92-1对应的变换结构向量。将共同结构的向量和变换结构的向量累计后的向量成为与试剂的向量对应的向量。

[0158] 返回到图19的说明。控制部250具有预处理部251、学习部252、计算部253、分析部254。控制部250例如由CPU或MPU实现。另外,控制部250例如也可以通过ASIC、FPGA等集成电路来执行。

[0159] 与预处理部251有关的说明与第一实施例中说明的与预处理部151有关的处理的说明相同。由预处理部251生成基团编码文件51、试剂编码文件52、子化合物编码文件53、目标化合物编码文件54、共同结构编码文件55。由预处理部251生成基团向量表T1、试剂向量表T2、子化合物向量表T3、目标化合物表T4和共同结构向量表T5。由预处理部251生成基团转置索引In1、试剂转置索引In2、子化合物转置索引In3、目标化合物索引In4和共同结构索引In5。预处理部251可以从外部装置取得学习数据90,也可以由预处理部251生成学习数据90。

[0160] 学习部252使用学习数据90来执行学习模型91的学习。学习部252的处理对应于图17中说明的处理。学习部252从学习数据90中取得子化合物的向量和与该子化合物的向量对应的共同结构的向量的组。学习部252通过以使得将子化合物的向量输入到学习模型91的情况下的、学习模型91的输出的值接近共同结构的向量的值的方式,执行基于误差反向传播的学习,来调整学习模型91的参数。

[0161] 计算部253在接受了分析查询92的指定的情况下,使用学习完毕的学习模型91,计算在分析查询92的子化合物的合成路径中进行变换反应的各共同结构的向量。计算部253将计算出的各共同结构的向量输出到分析部254。

[0162] 在以下的说明中,将计算部253计算出的各共同结构的向量分别表述为“共同结构向量”。

[0163] 分析部254基于分析查询92的子化合物的向量、共同结构向量、试剂向量表T2,生成共同结构·变更机构表93。以下,对分析部254的处理的一例进行说明。

[0164] 分析部254分别算出子化合物的向量与试剂向量表T2中包含的各向量之间的距离,确定与子化合物的向量之间的距离小于阈值的向量。将试剂向量表T2中包含的、且与子化合物的向量之间的距离小于阈值的向量表述为“相似向量”。

[0165] 分析部254通过从相似向量中减去共同结构向量来计算变换结构的向量,并且确定共同结构向量与变换结构的向量之间的对应关系。分析部254将共同结构向量和变换结构的向量登记在共同结构·变换结构表93中。分析部245通过反复执行上述处理,生成共同结构·变换结构表93。分析部245既可以将共同结构·变换结构表93输出到显示部230进行显示,也可以发送到与网络连接的外部装置。

[0166] 接下来,说明本实施例2的信息处理装置200的处理过程的一例。图20是示出本实施例2的信息处理装置的处理过程的流程图。信息处理装置200的计算部253接受分析查询92(步骤S301)。

[0167] 计算部253基于子化合物向量表T3,将分析查询92的子化合物变换为向量(步骤S302)。

[0168] 计算部253通过将子化合物的向量输入到学习完毕的学习模型91，来计算共同结构的向量(步骤S303)。信息处理装置200的分析部254根据共同结构的向量和试剂向量表T2的各向量之间的距离，确定相似试剂向量(步骤S304)。

[0169] 分析部254通过从子化合物和相似试剂的各向量中减去共同结构的向量，计算变换结构的向量(步骤S305)。分析部254将共同结构的向量与变换结构的向量之间的关系登记在共同结构・变换结构表中(步骤S306)。分析部254输出共同结构・变换结构表的信息(步骤S307)。

[0170] 接下来，说明本实施例2的信息处理装置200的效果。信息处理装置100向学习完毕的学习模型91输入分析查询92的向量，计算与分析查询的子化合物对应的各共同结构的向量。另外，通过从与子化合物相似的试剂的向量中减去各共同结构的向量，计算出在相似的子化合物和试剂中不同的变换结构的向量。通过使用上述共同结构的向量、变换结构的向量，能够容易地分析能够用于向目标化合物的变换反应、再合成、制造的更好的试剂。

[0171] 子化合物和试剂是由多个官能团构成的一次结构。另外，通过使用官能团的分散向量，能够估计与某个官能团相邻的官能团，能够应用于各官能团的耦合度、稳定性的评价。关于具有实绩的从试剂向子化合物的变换反应，通过基于构成子化合物、试剂的一次结构的多个官能团的向量进行机器学习，能够提高从试剂开始的变换反应和再合成的分析精度。

[0172] 接着，说明实现与上述实施例所示的信息处理装置200(100)同样的功能的计算机的硬件结构的一例。图21是表示实现与实施例的信息处理装置同样的功能的计算机的硬件结构的一例的图。

[0173] 如图21所示，计算机300具有执行各种运算处理的CPU301、接受来自用户的数据的输入的输入装置302、以及显示器303。另外，计算机300具有经由有线或无线网络与外部装置等之间进行数据的收发的通信装置304和接口装置305。另外，计算机300具有暂时存储各种信息的RAM306和硬盘装置307。而且，各装置301～307与总线308连接。

[0174] 硬盘装置307包括预处理程序307a、学习程序307b、计算程序307c和分析程序307d。另外，CPU301读出各程序307a～307d并在RAM306中展开。

[0175] 预处理程序307a作为预处理进程306a发挥功能。学习程序307b作为学习进程306b发挥功能。计算程序307c作为计算进程306c发挥功能。分析程序307d作为分析进程306d发挥功能。

[0176] 预处理进程306a的处理对应于预处理部151、251的处理。学习进程306b的处理对应于学习部152、252的处理。计算进程306c的处理对应于计算部153、253的处理。分析进程306d的处理对应于分析部154、254的处理。

[0177] 另外，关于各程序307a～307d，也可以不一定从最开始就预先存储在硬盘装置307中。例如，在插入到计算机300的软盘(FD)、CD-ROM、DVD、光磁盘、IC卡等“可移动物理介质”中存储各程序。并且，计算机300也可以读出并执行各程序307a～307d。

[0178] 标记说明

[0179] 50化学结构式文件

[0180] 51基团编码文件

[0181] 52试剂编码文件

[0182] 53子化合物编码文件

[0183] 54目标化合物编码文件

[0184] 55共同结构编码文件

[0185] 60逆合成分析结果表

[0186] 65、90学习数据

[0187] 70、91学习模型

[0188] 80、92分析查询

[0189] 85子化合物·试剂表

[0190] 93共同结构·变换结构表

[0191] 100、200信息处理装置

[0192] 110、210通信部

[0193] 120、220输入部

[0194] 130、230显示部

[0195] 140、240存储部

[0196] 150、250控制部

[0197] 151、251预处理部

[0198] 152、252学习部

[0199] 153、253计算部

[0200] 154、254分析部

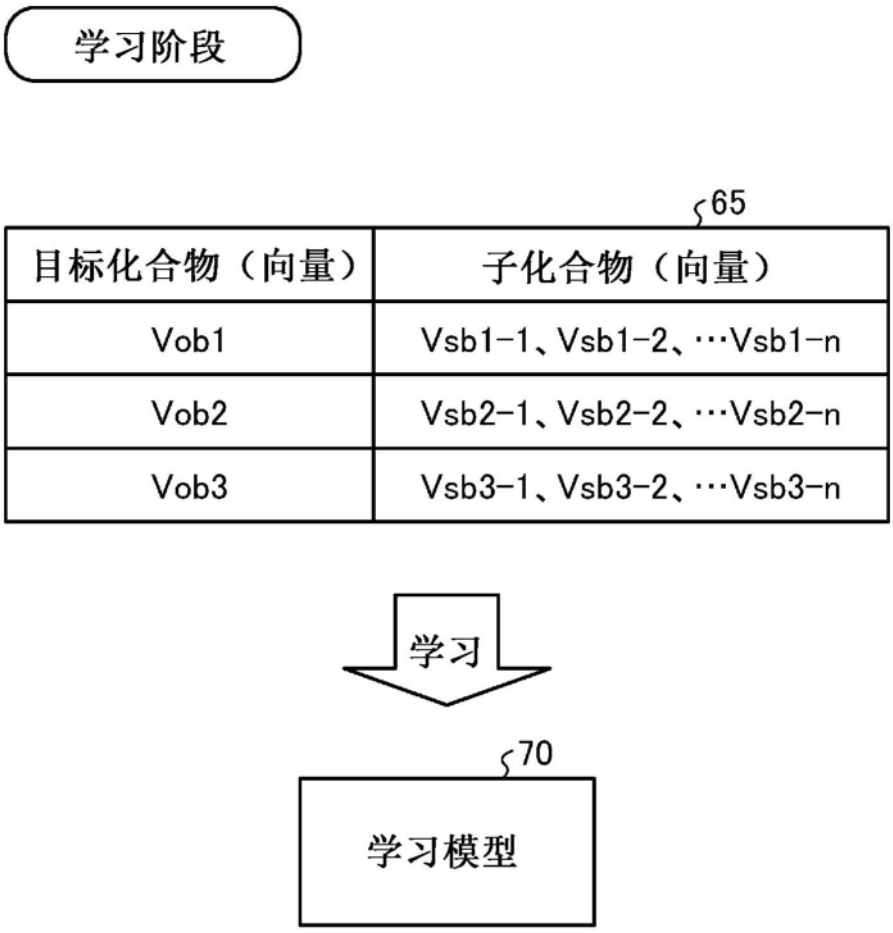

学习阶段
65
目标化合物（向量）
子化合物（向量）
Vob1
Vsb1-1、Vsb1-2、…Vsb1-n
Vob2
Vsb2-1、Vsb2-2、…Vsb2-n
Vob3
Vsb3-1、Vsb3-2、…Vsb3-n
学习
70
学习模型

图1

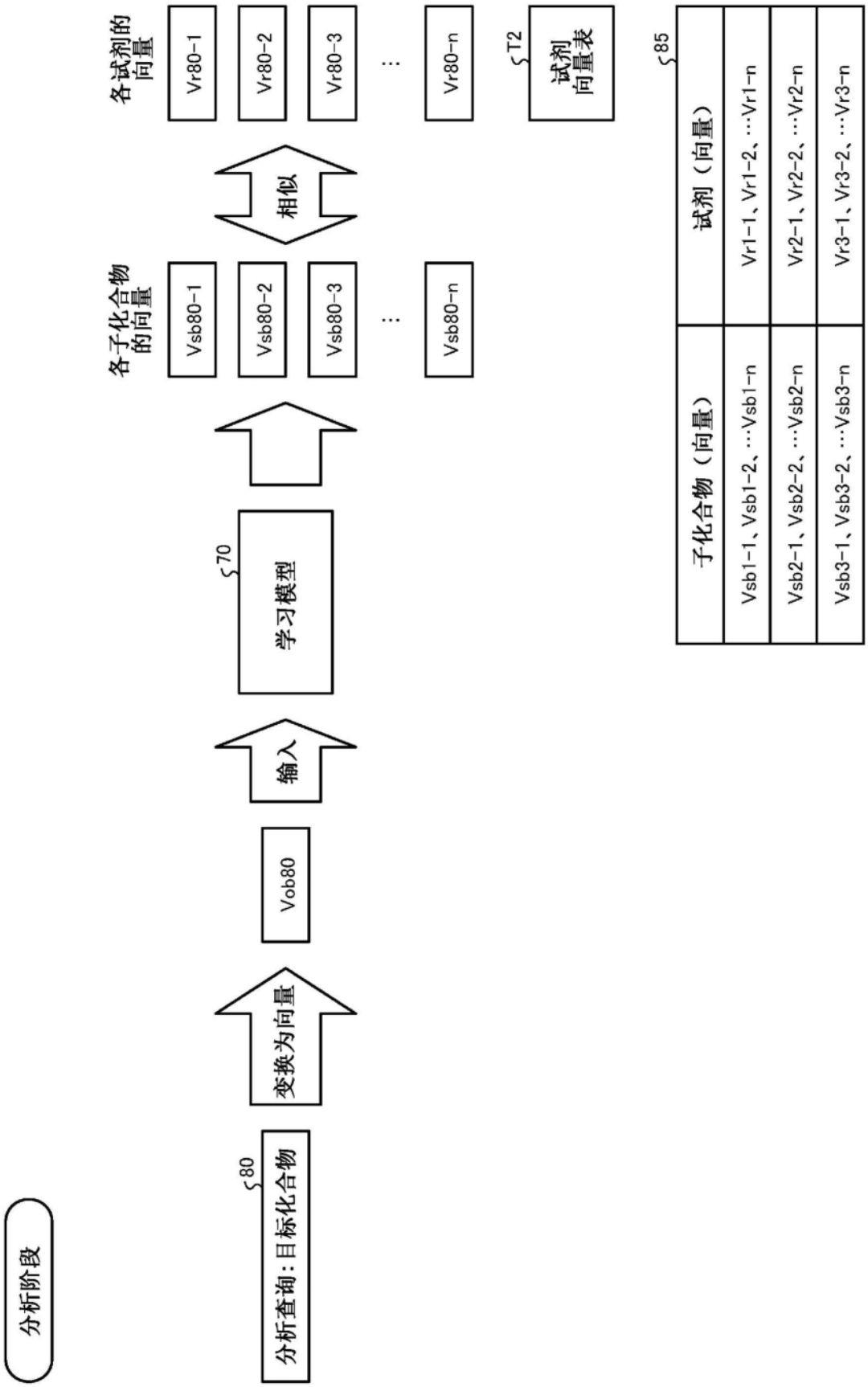

分析阶段
80
分析查询：目标化合物
变换为向量
Vob80
输入
70
学习模型
各子化合物的向量
Vsb80-1
Vsb80-2
Vsb80-3
Vsb80-n
相似
各试剂的向量
Vr80-1
Vr80-2
Vr80-3
Vr80-n
T2
试剂向量表
85
子化合物（向量）
试剂（向量）
Vsb1-1、Vsb1-2、…Vsb1-n
Vr1-1、Vr1-2、…Vr1-n
Vsb2-1、Vsb2-2、…Vsb2-n
Vr2-1、Vr2-2、…Vr2-n
Vsb3-1、Vsb3-2、…Vsb3-n
Vr3-1、Vr3-2、…Vr3-n

图2

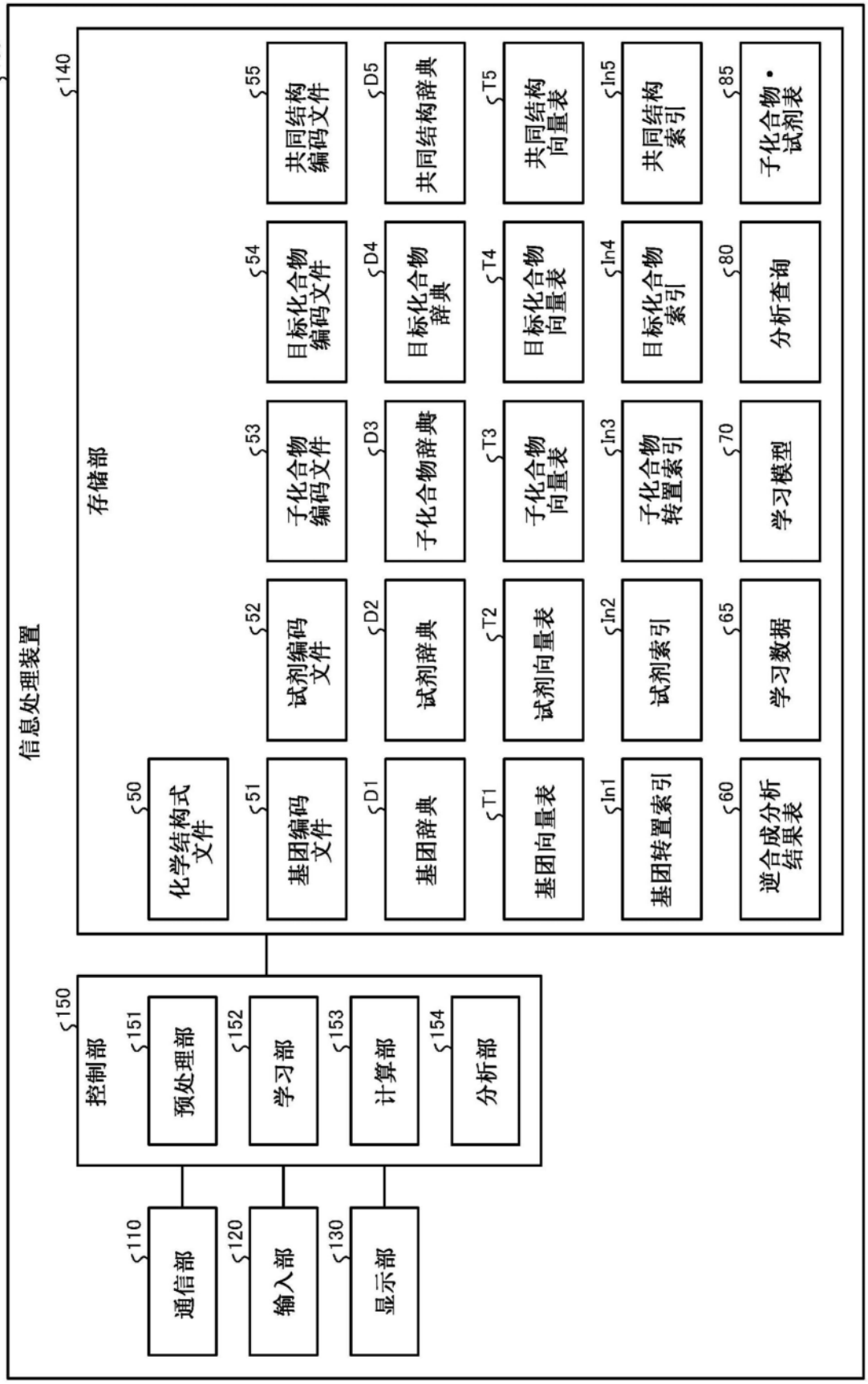

100
信息处理装置
140
存储部
50
化学结构式文件
51
基团编码文件
D1
基团辞典
T1
基团向量表
In1
基团转置索引
60
逆合成分析结果表
52
试剂编码文件
D2
试剂辞典
T2
试剂向量表
In2
试剂索引
65
学习数据
53
子化合物编码文件
D3
子化合物辞典
T3
子化合物向量表
In3
子化合物转置索引
70
学习模型
54
目标化合物编码文件
D4
目标化合物辞典
T4
目标化合物向量表
In4
目标化合物索引
80
分析查询
55
共同结构编码文件
D5
共同结构辞典
T5
共同结构向量表
In5
共同结构索引
85
子化合物·试剂表
150
控制部
151
预处理部
152
学习部
153
计算部
154
分析部
110
通信部
120
输入部
130
显示部

图3

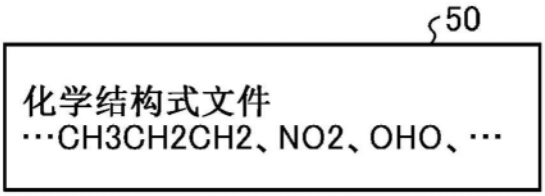


图4

D1

压缩码（基团）	名称	示性式
D0008000h	甲基	CH3
D0008001h	乙基	CH3CH2
D0008002h	丁基	CH3CH2CH2CH2
…	…	…
D0008013h	硝基	NO2
…	…	…
D000805Ah	砷氧基	OH
D000805Bh	醛基	OHO
…	…	…

图5

D2

压缩码（试剂）	名称	基团码序列
F3000000h	试剂 α	D0008001hD000822h…
F3000001h	试剂 β	D0008103hD000822h…
F3000002h	试剂 γ	D0001258hD002498h…
…	…	…

图6

D3

压缩码（子化合物）	名称	基团码序列
F1000000h	子化合物 α	D0008001hD000822h…
F1000001h	子化合物 β	D0008103hD000822h…
F1000002h	子化合物 γ	D0001258hD002498h…
…	…	…

图7A

D4

压缩码（目标化合物）	名称	基团码序列
F0000000h	目标化合物-1	D0008031hD000821h…
F0000001h	目标化合物-2	D0008103hD000681h…
F0000002h	目标化合物-3	D0331258hD002589h…
…	…	…

图7B

D5

压缩码（共同结构）	名称	基团码序列
F2000000h	共同结构 α	D0008001hD000822h…
F2000001h	共同结构 β	D0008103hD000822h…
F2000002h	共同结构 γ	D0001258hD002498h…
…	…	…

图7C

T1

压缩码（基团）	向量
D0008000h	V1-1
D0008001h	V1-2
D0008002h	V1-3
…	…

图8

T2

压缩码（试剂）	向量
F3000000h	Vsb1-1
F3000001h	Vsb1-2
F3000002h	Vsb1-3
…	…

图9

T3

压缩码（子化合物）	向量
F1000000h	Vr1-1
F1000001h	Vr1-2
F1000002h	Vr1-3
…	…

图10A

T4

压缩码（目标化合物）	向量
FA000000h	Vob1
FA000001h	Vob2
FA000002h	Vob3
…	…

图10B

T5

压缩码（共同结构）	名称
F2000000h	Vcm1-1
F2000001h	Vcm1-2
F2000002h	Vcm1-3
…	…

图10C

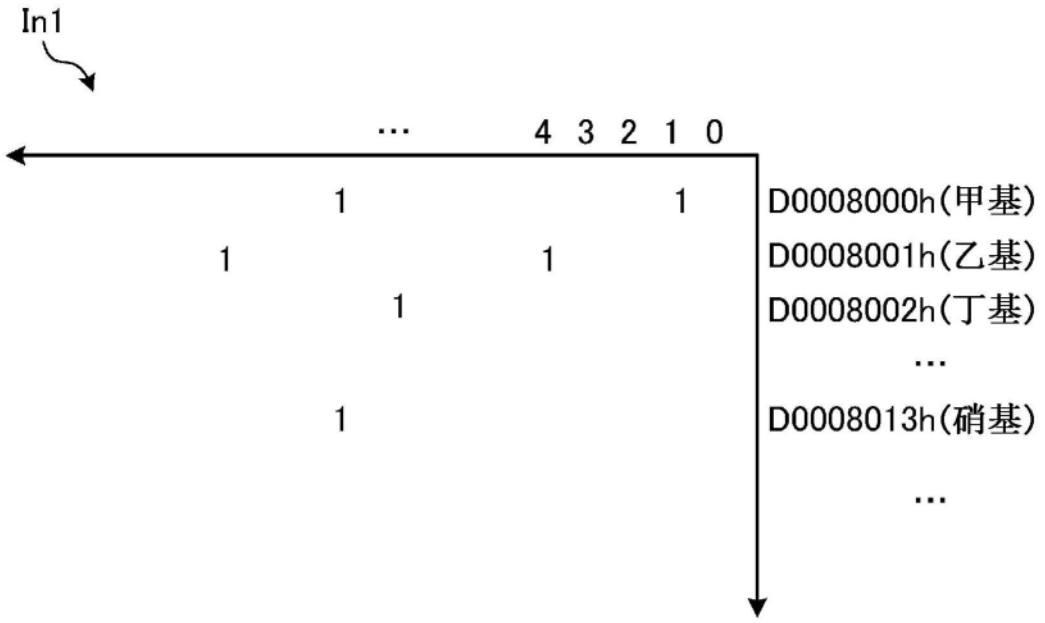

In1
… 4 3 2 1 0
1 1
1 1
1
1
D0008000h(甲基)
D0008001h(乙基)
D0008002h(丁基)
…
D0008013h(硝基)
…

图11

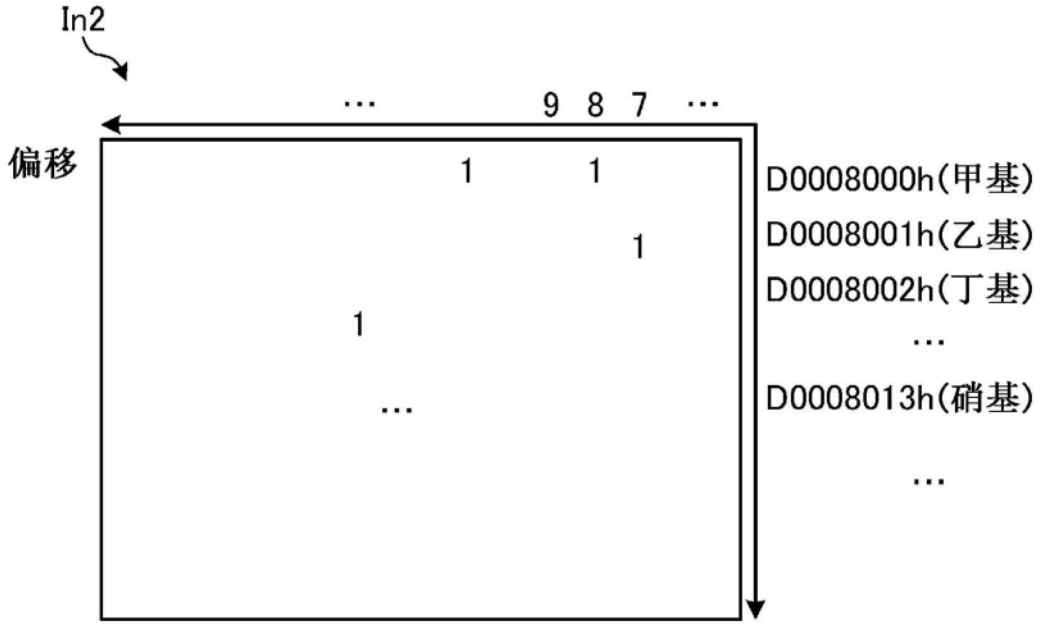

In2
… 9 8 7 …
偏移
1 1
1
1
…
D0008000h(甲基)
D0008001h(乙基)
D0008002h(丁基)
…
D0008013h(硝基)
…

图12

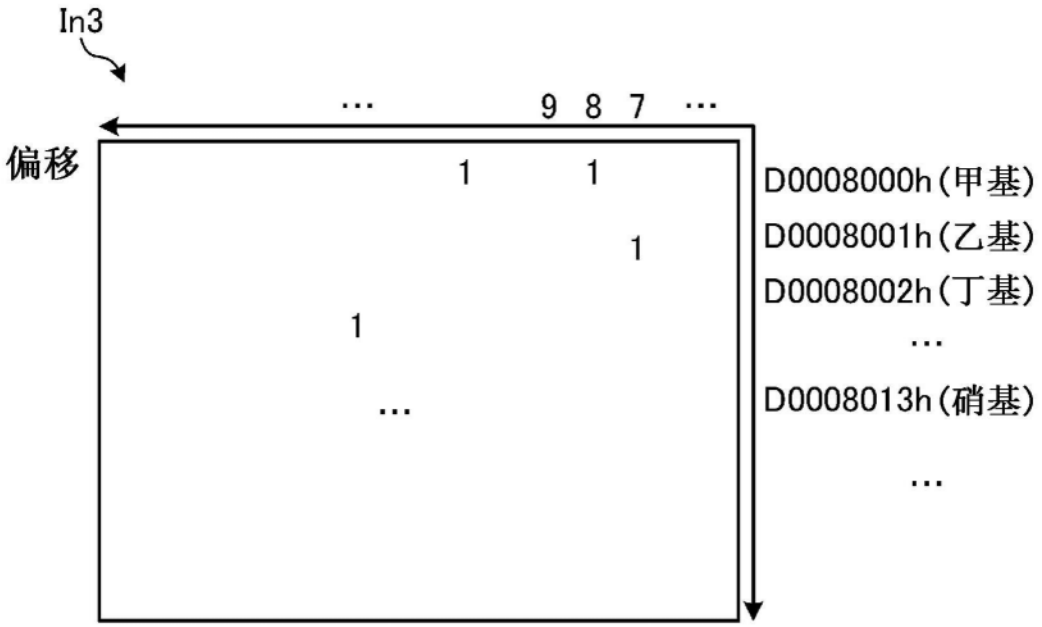

In3
… 9 8 7 …
偏移
1 1
1
1
…
D0008000h(甲基)
D0008001h(乙基)
D0008002h(丁基)
…
D0008013h(硝基)
…

图13A

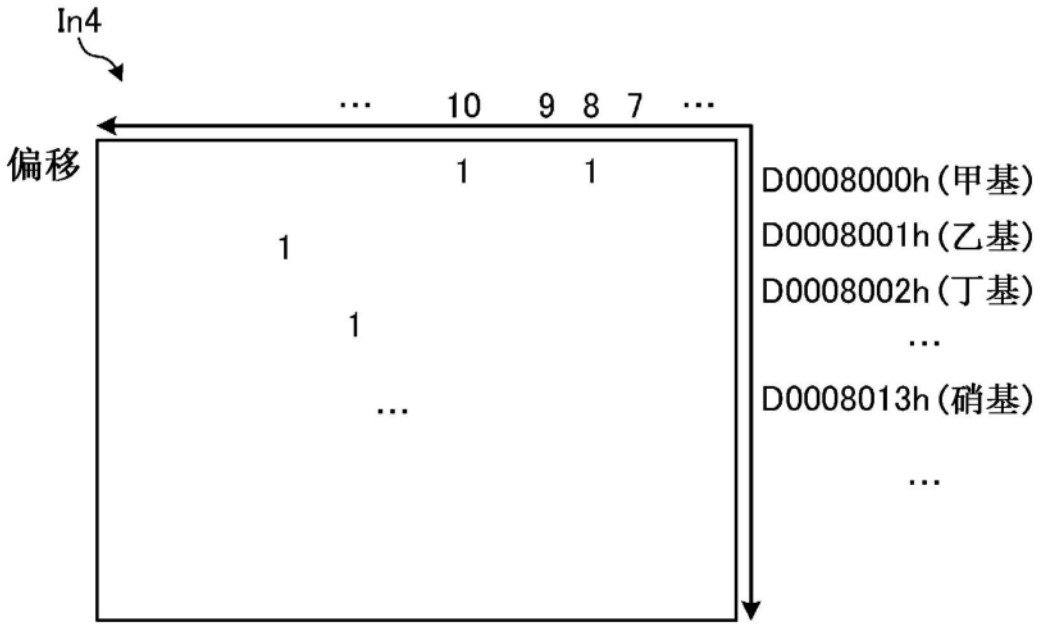

In4
… 10 9 8 7 …
偏移
1 1
1
1
…
D0008000h(甲基)
D0008001h(乙基)
D0008002h(丁基)
…
D0008013h(硝基)
…

图13B

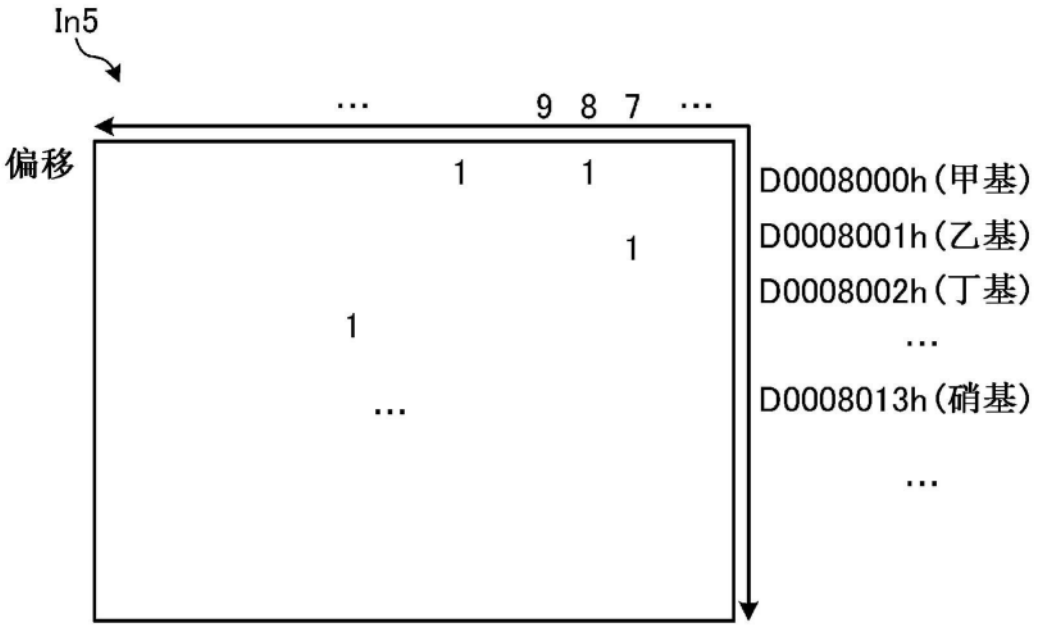


图13C

60

名称 （目标化合物）	合成路径
目标化合物-1	目标化合物-1的合成路径的信息
目标化合物-2	目标化合物-2的合成路径的信息
目标化合物-3	目标化合物-3的合成路径的信息
…	…

图14

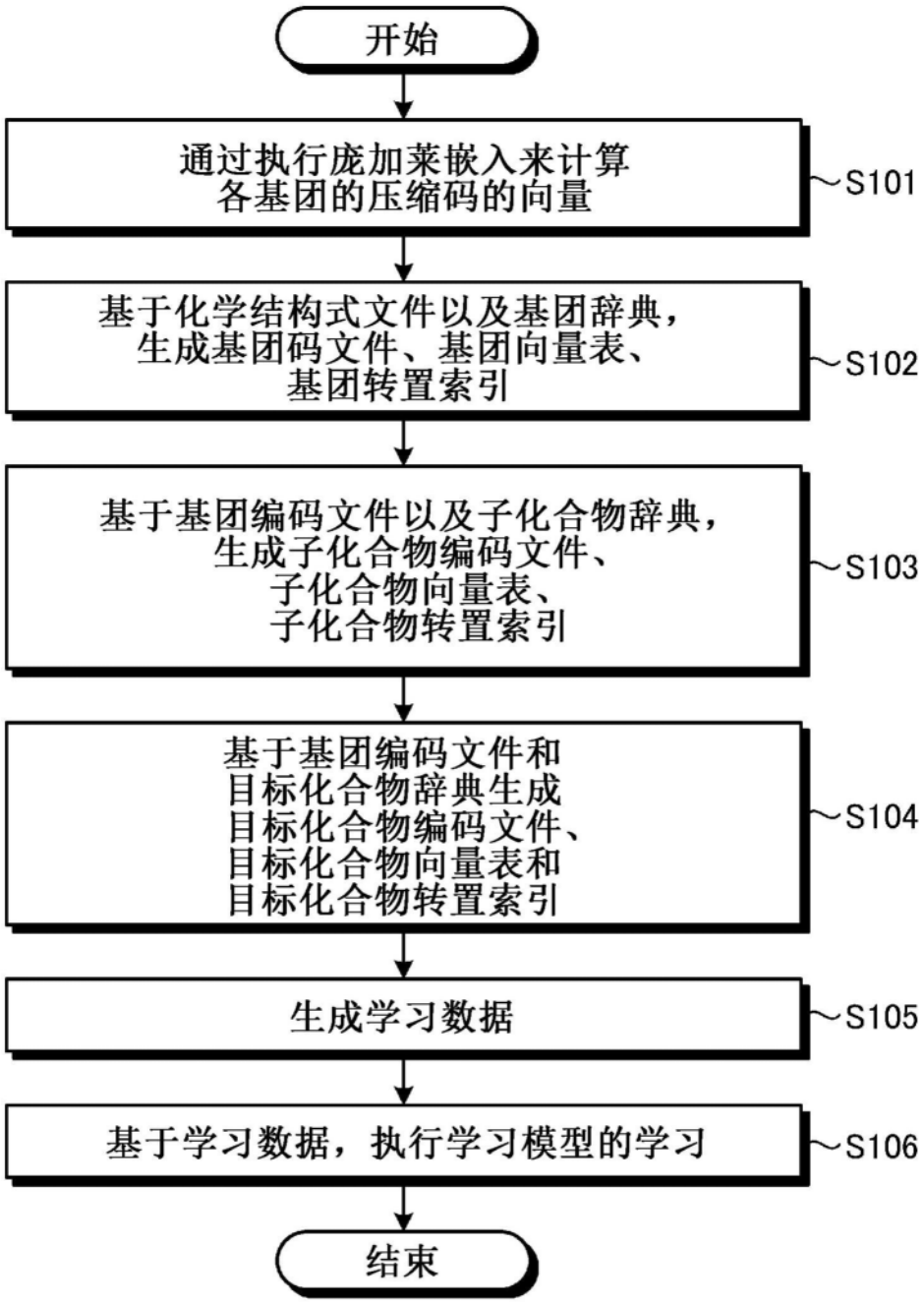

开始
通过执行庞加莱嵌入来计算
各基团的压缩码的向量
S101
基于化学结构式文件以及基团辞典，
生成基团码文件、基团向量表、
基团转置索引
S102
基于基团编码文件以及子化合物辞典，
生成子化合物编码文件、
子化合物向量表、
子化合物转置索引
S103
基于基团编码文件和
目标化合物辞典生成
目标化合物编码文件、
目标化合物向量表和
目标化合物转置索引
S104
生成学习数据
S105
基于学习数据，执行学习模型的学习
S106
结束

图15

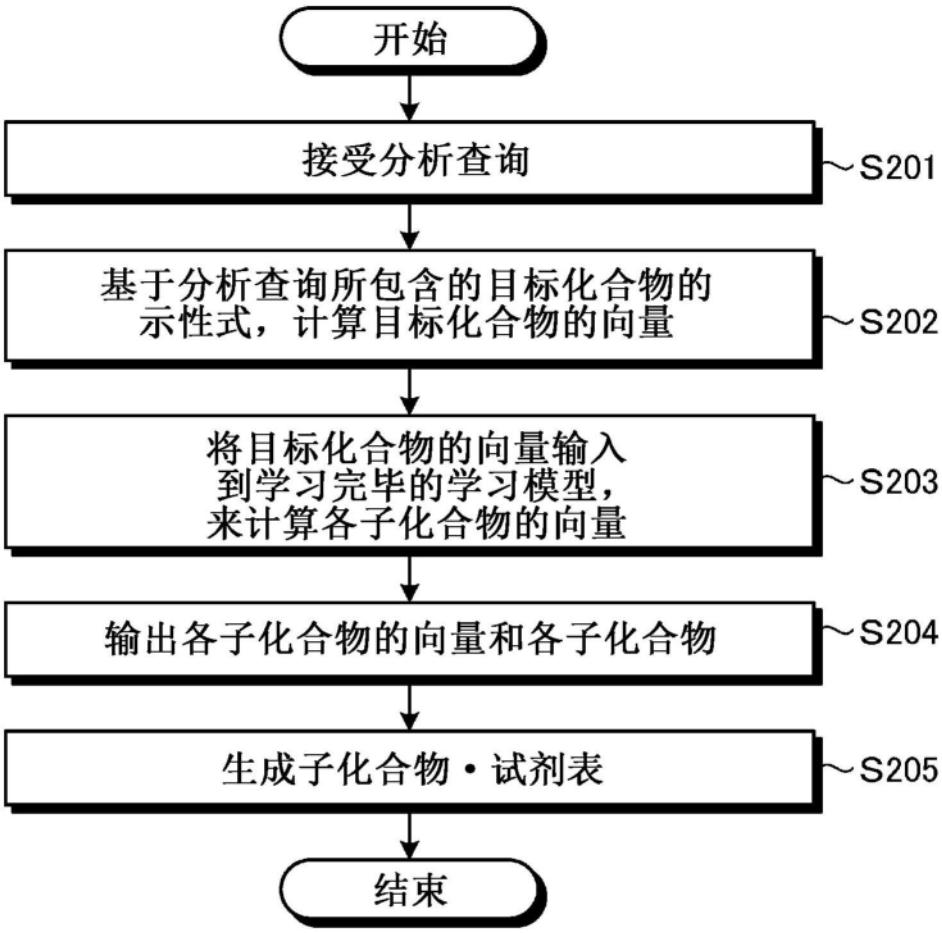

开始
接受分析查询
S201
基于分析查询所包含的目标化合物的示性式，计算目标化合物的向量
S202
将目标化合物的向量输入到学习完毕的学习模型，来计算各子化合物的向量
S203
输出各子化合物的向量和各子化合物
S204
生成子化合物·试剂表
S205
结束

图16

90

子化合物（向量）	共同结构（向量）
Vsb1-1	Vcm1-1
Vsb1-2	Vcm1-2
Vsb1-3	Vcm1-3

图17

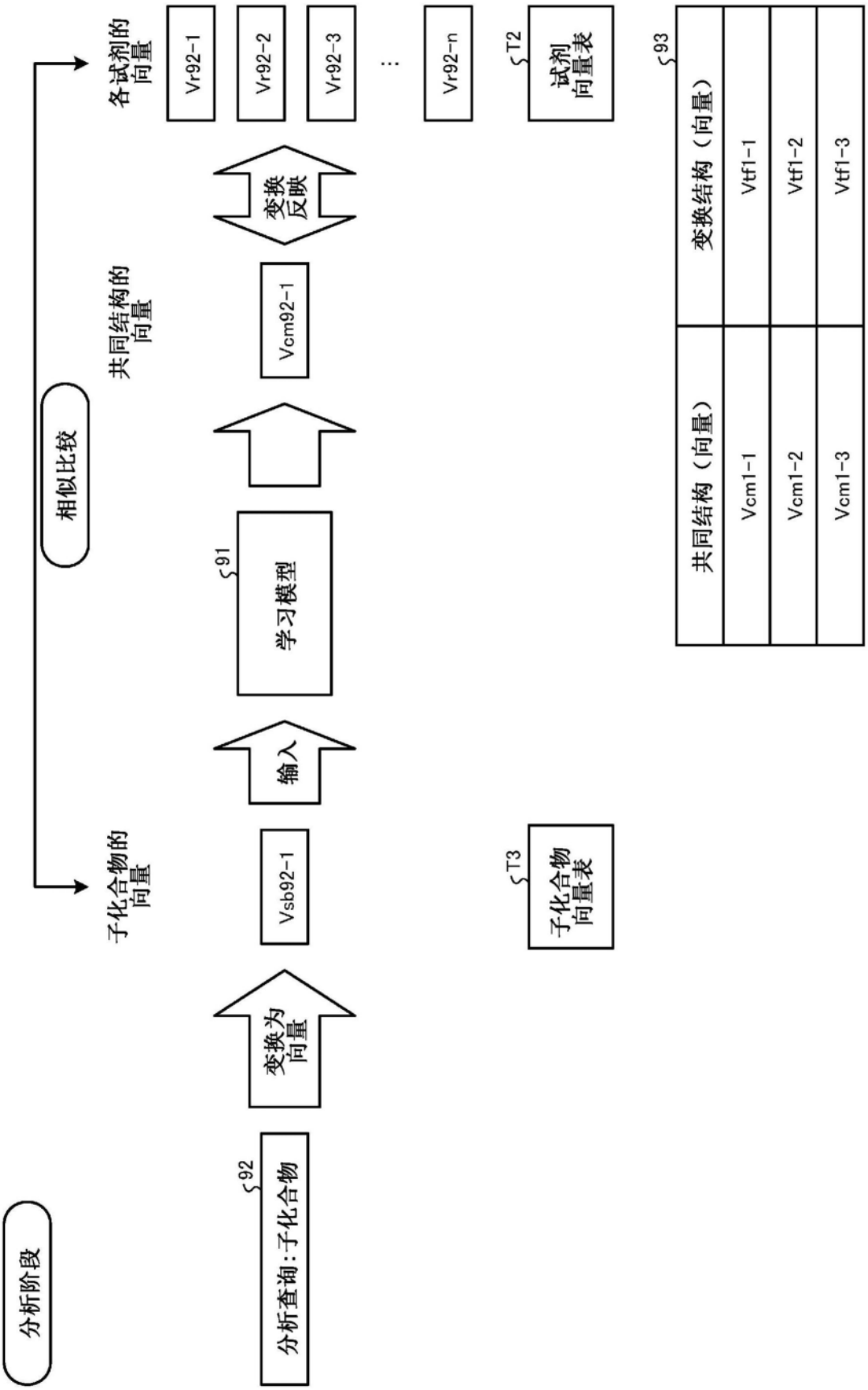

分析阶段
分析查询:子化合物
S92
变换为向量
子化合物的向量
Vsb92-1
输入
学习模型
S91
共同结构的向量
Vcm92-1
变换反映
各试剂的向量
Vr92-1
Vr92-2
Vr92-3
Vr92-n
相似比较
T3
子化合物向量表
T2
试剂向量表
S93
共同结构（向量）
Vcm1-1
Vcm1-2
Vcm1-3
变换结构（向量）
Vtf1-1
Vtf1-2
Vtf1-3

图18

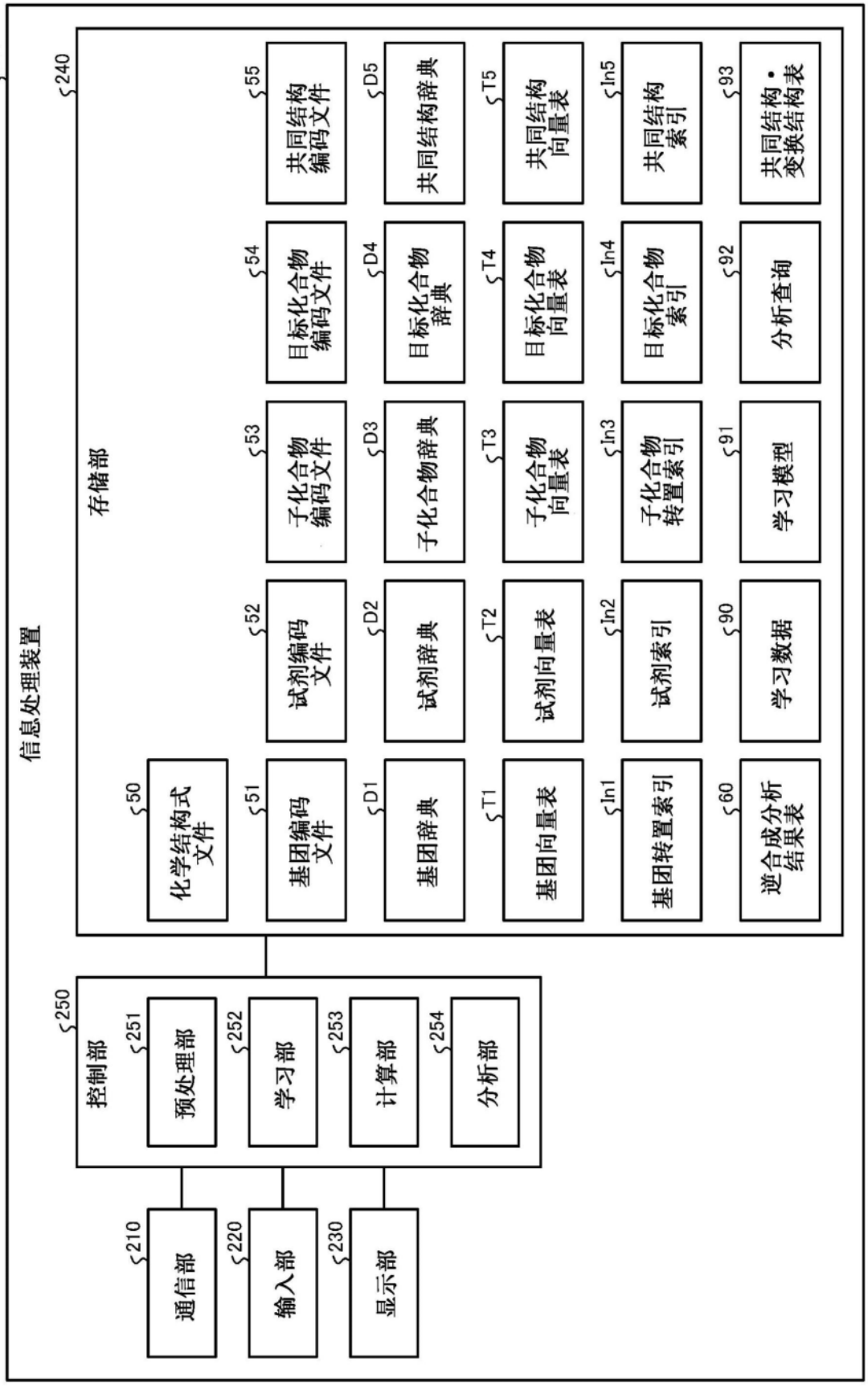

200
信息处理装置
240
存储部
50
化学结构式文件
51
基团编码文件
D1
基团辞典
T1
基团向量表
In1
基团转置索引
60
逆合成分析结果表
52
试剂编码文件
D2
试剂辞典
T2
试剂向量表
In2
试剂索引
90
学习数据
53
子化合物编码文件
D3
子化合物辞典
T3
子化合物向量表
In3
子化合物转置索引
91
学习模型
54
目标化合物编码文件
D4
目标化合物辞典
T4
目标化合物向量表
In4
目标化合物索引
92
分析查询
55
共同结构编码文件
D5
共同结构辞典
T5
共同结构向量表
In5
共同结构索引
93
共同结构·变换结构表
250
控制部
251
预处理部
252
学习部
253
计算部
254
分析部
210
通信部
220
输入部
230
显示部

图19

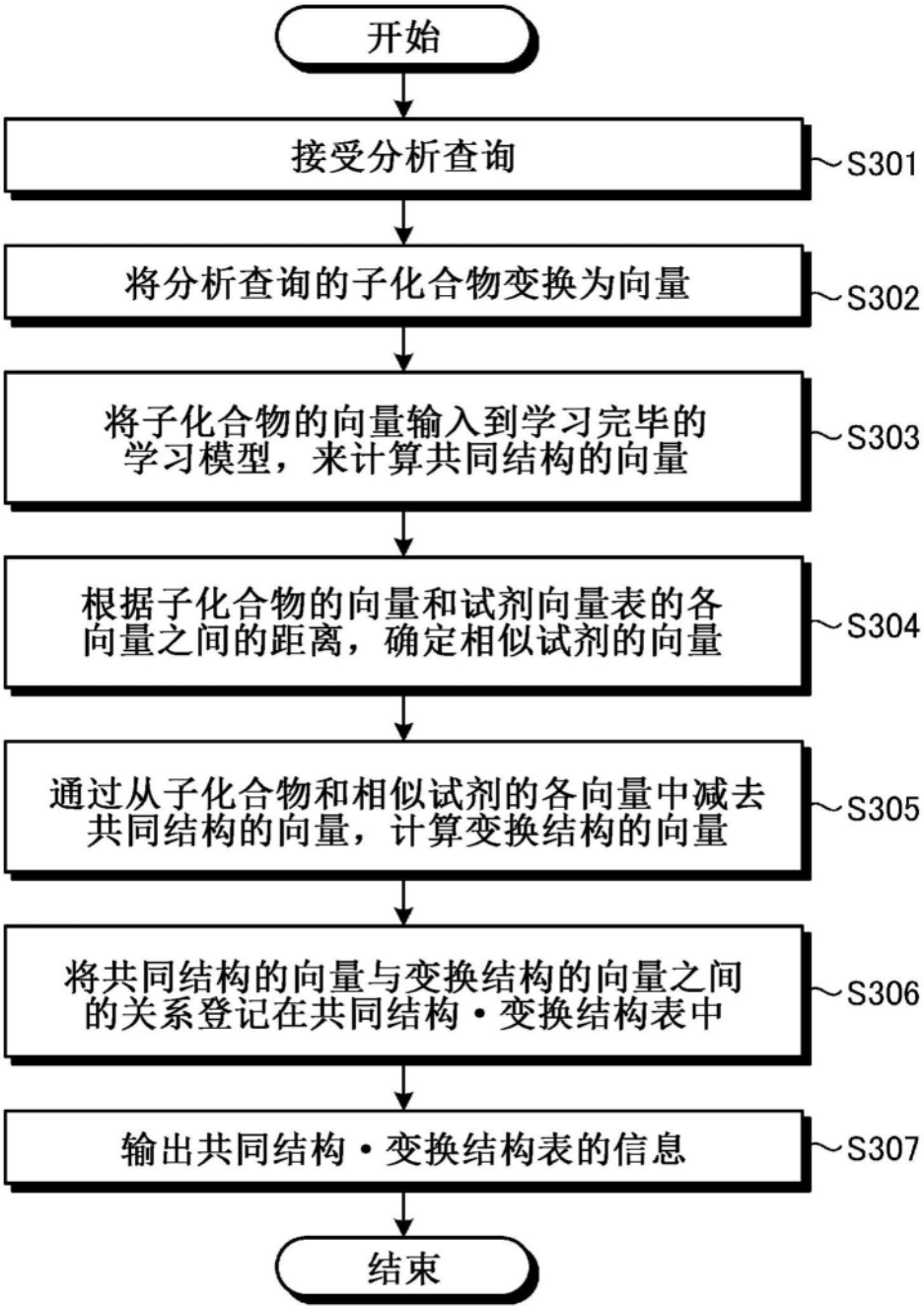

开始
接受分析查询
S301
将分析查询的子化合物变换为向量
S302
将子化合物的向量输入到学习完毕的学习模型，来计算共同结构的向量
S303
根据子化合物的向量和试剂向量表的各向量之间的距离，确定相似试剂的向量
S304
通过从子化合物和相似试剂的各向量中减去共同结构的向量，计算变换结构的向量
S305
将共同结构的向量与变换结构的向量之间的关系登记在共同结构・变换结构表中
S306
输出共同结构・变换结构表的信息
S307
结束

图20

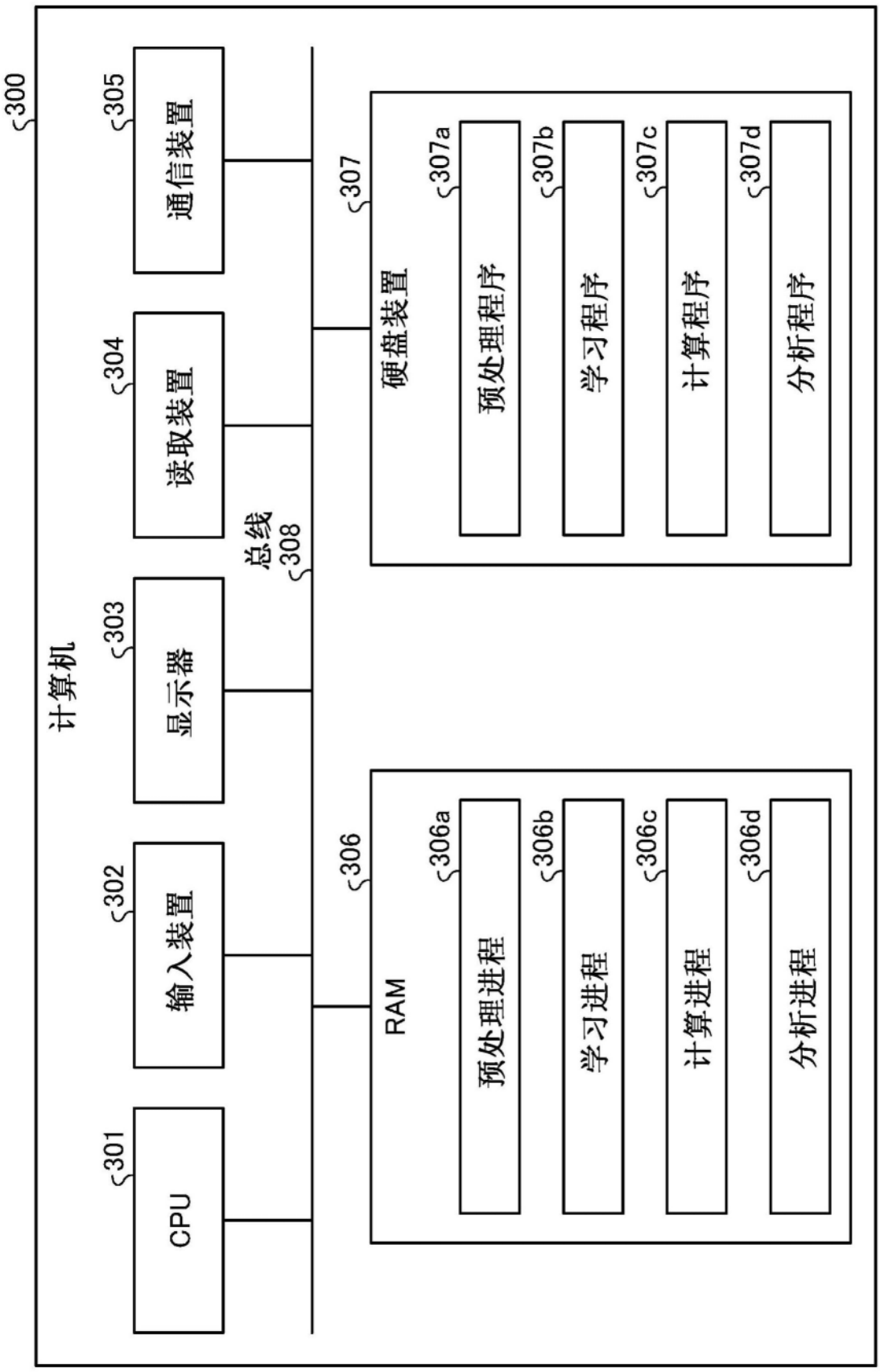

300
计算机
301
CPU
302
输入装置
303
显示器
304
读取装置
305
通信装置
306
RAM
306a
预处理进程
306b
学习进程
306c
计算进程
306d
分析进程
307
硬盘装置
307a
预处理程序
307b
学习程序
307c
计算程序
307d
分析程序
308
总线

图21

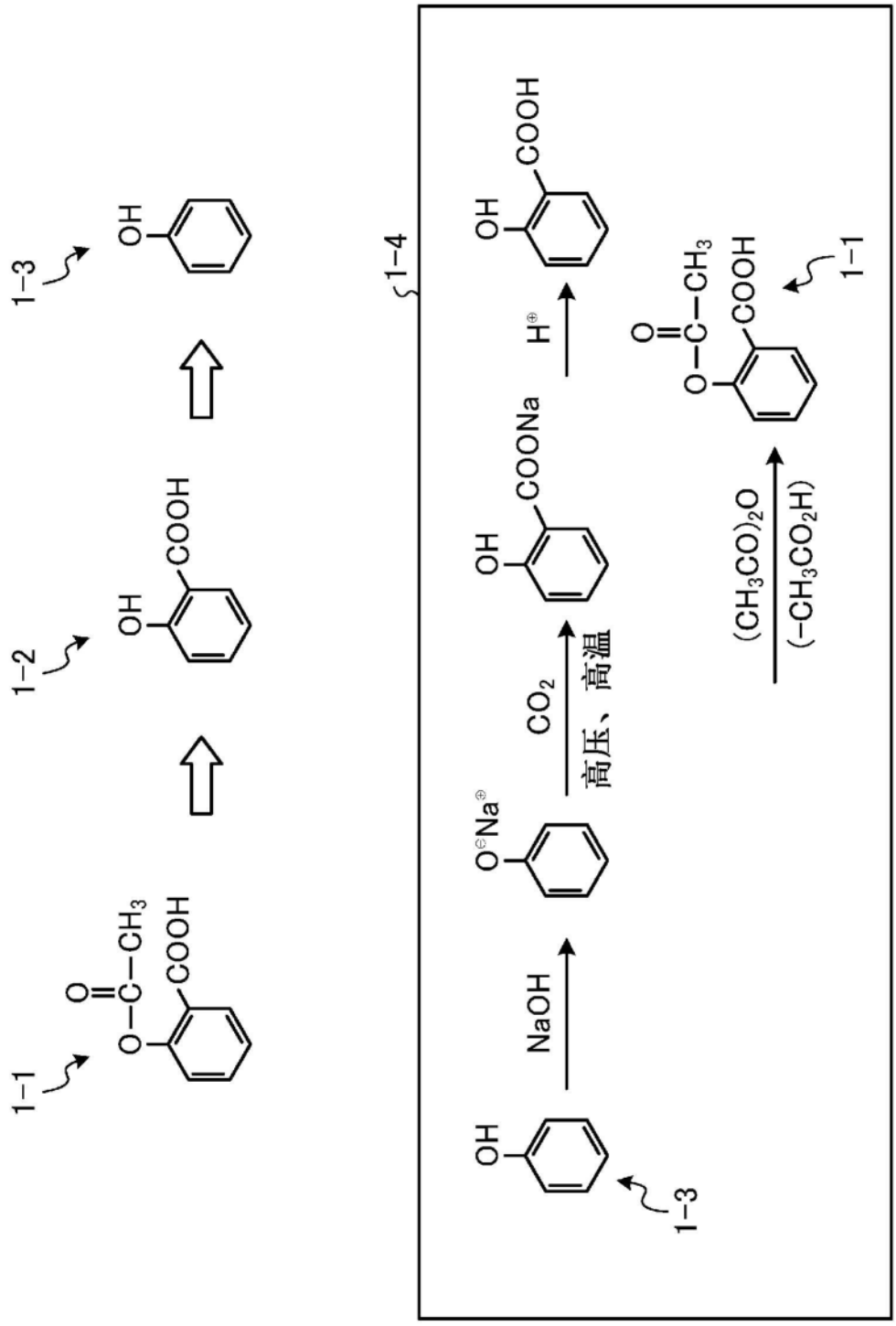


图22