(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114333997 A
(43)申请公布日 2022.04.12

(21)申请号 202111387290.6

(22)申请日 2021.11.22

(71)申请人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 王文川 杨帆 姚建华

(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
代理人 李文静

(51)Int.Cl.
G16B 25/00(2019.01)
G16B 40/00(2019.01)

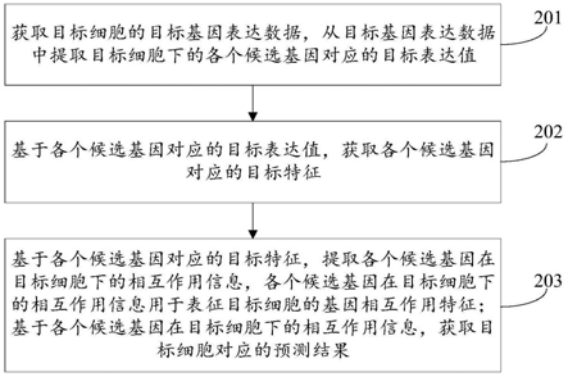
权利要求书3页 说明书26页 附图5页

(54)发明名称

数据处理、数据处理模型的训练方法、装置、设备及介质

(57)摘要

本申请公开了一种数据处理、数据处理模型的训练方法、装置、设备及介质，属于计算机技术领域。该方法包括：从目标细胞的目标基因表达数据中提取目标细胞下的各个候选基因对应的目标表达值；基于各个候选基因对应的目标表达值，获取各个候选基因对应的目标特征；基于各个候选基因对应的目标特征，提取各个候选基因在目标细胞下的相互作用信息；基于各个候选基因在目标细胞下的相互作用信息，获取目标细胞对应的预测结果。此种方式，目标细胞对应的预测结果是基于各个候选基因在目标细胞下的相互作用信息获取的，各个候选基因在目标细胞下的相互作用信息能够体现出目标细胞的功能方面的特征，获取的预测结果的准确性较高。



CN 114333997 A

1.一种数据处理方法,其特征在于,所述方法包括:

获取目标细胞的目标基因表达数据,从所述目标基因表达数据中提取所述目标细胞下的各个候选基因对应的目标表达值;

基于所述各个候选基因对应的目标表达值,获取所述各个候选基因对应的目标特征;

基于所述各个候选基因对应的目标特征,提取所述各个候选基因在所述目标细胞下的相互作用信息,所述各个候选基因在所述目标细胞下的相互作用信息用于表征所述目标细胞的基因相互作用特征;

基于所述各个候选基因在所述目标细胞下的相互作用信息,获取所述目标细胞对应的预测结果。

2.根据权利要求1所述的方法,其特征在于,所述基于所述各个候选基因对应的目标表达值,获取所述各个候选基因对应的目标特征,包括:

将第一候选基因对应的目标表达值转换成所述第一候选基因对应的表达值特征,所述第一候选基因为所述各个候选基因中的任一候选基因;

对所述第一候选基因对应的表达值特征和所述第一候选基因对应的表征特征进行融合,得到所述第一候选基因对应的目标特征。

3.根据权利要求2所述的方法,其特征在于,所述将第一候选基因对应的目标表达值转换成所述第一候选基因对应的表达值特征,包括:

对所述第一候选基因对应的目标表达值进行归一化处理,得到所述第一候选基因对应的归一化表达值;

确定所述归一化表达值对应的目标离散化表达值,将所述目标离散化表达值对应的嵌入特征作为所述第一候选基因对应的表达值特征。

4.根据权利要求3所述的方法,其特征在于,所述目标离散化表达值为参考数量个候选离散化表达值中的一个候选离散化表达值,所述将所述目标离散化表达值对应的嵌入特征作为所述第一候选基因对应的表达值特征之前,所述方法还包括:

对所述参考数量个候选离散化表达值进行向量化转换,得到所述参考数量个候选离散化表达值分别对应的嵌入特征。

5.根据权利要求1-4任一所述的方法,其特征在于,所述基于所述各个候选基因对应的目标特征,提取所述各个候选基因在所述目标细胞下的相互作用信息,包括:

调用目标数据处理模型基于所述各个候选基因对应的目标特征,提取所述各个候选基因在所述目标细胞下的相互作用信息;

所述基于所述各个候选基因在所述目标细胞下的相互作用信息,获取所述目标细胞对应的预测结果,包括:

调用所述目标数据处理模型基于所述各个候选基因在所述目标细胞下的相互作用信息,获取所述目标细胞对应的预测结果。

6.根据权利要求1-4任一所述的方法,其特征在于,所述目标细胞对应的预测结果指示所述目标细胞的类别为目标类别,所述方法还包括:

基于所述各个候选基因在所述目标细胞下的相互作用信息,在所述各个候选基因中确定属于所述目标类别的细胞对应的满足选取条件的基因。

7.一种数据处理模型的训练方法,其特征在于,所述方法包括:

获取样本细胞的样本基因表达数据和所述样本细胞对应的标准结果，从所述样本基因表达数据中提取所述样本细胞下的各个候选基因对应的样本表达值；

基于所述各个候选基因对应的样本表达值，获取所述各个候选基因对应的样本特征；

调用第一数据处理模型基于所述各个候选基因对应的样本特征，提取所述各个候选基因在所述样本细胞下的相互作用信息；基于所述各个候选基因在所述样本细胞下的相互作用信息，获取所述样本细胞对应的预测结果；

基于所述样本细胞对应的预测结果和标准结果，获取结果损失函数；利用所述结果损失函数对所述第一数据处理模型进行训练，得到目标数据处理模型。

8.根据权利要求7所述的方法，其特征在于，所述第一数据处理模型包括第一提取子模型和第一预测子模型；所述调用第一数据处理模型基于所述各个候选基因对应的样本特征，提取所述各个候选基因在所述样本细胞下的相互作用信息，包括：

调用所述第一提取子模型对所述各个候选基因对应的样本特征进行信息提取，得到所述各个候选基因在所述样本细胞下的相互作用信息；

所述基于所述各个候选基因在所述样本细胞下的相互作用信息，获取所述样本细胞对应的预测结果，包括：

调用所述第一预测子模型对所述各个候选基因在所述样本细胞下的相互作用信息进行处理，得到所述样本细胞对应的预测结果。

9.根据权利要求8所述的方法，其特征在于，所述调用所述第一提取子模型对所述各个候选基因对应的样本特征进行信息提取，得到所述各个候选基因在所述样本细胞下的相互作用信息之前，所述方法还包括：

获取训练细胞的训练基因表达数据，从所述训练基因表达数据中提取所述训练细胞下的所述各个候选基因对应的训练表达值；

基于所述各个候选基因对应的训练表达值，获取所述各个候选基因对应的训练特征，将所述各个候选基因中满足替换条件的候选基因对应的训练特征替换为参考特征；

调用初始提取子模型对所述满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征进行信息提取，得到所述各个候选基因在所述训练细胞下的相互作用信息；基于所述各个候选基因在所述训练细胞下的相互作用信息，获取所述满足替换条件的候选基因对应的预测特征；

基于所述满足替换条件的候选基因对应的预测特征和训练特征，获取特征损失函数，利用所述特征损失函数对所述初始提取子模型进行训练，得到所述第一提取子模型。

10.根据权利要求7-9任一所述的方法，其特征在于，所述样本细胞的数量为至少一个，一个样本细胞的样本基因表达数据包括所述一个样本细胞下的各个测量基因对应的样本表达值，所述从所述样本基因表达数据中提取所述样本细胞下的各个候选基因对应的样本表达值之前，所述方法还包括：

基于各个样本细胞的样本基因表达数据，统计所述各个测量基因分别命中的样本细胞的数量，一个测量基因命中一个样本细胞用于指示所述一个样本细胞下的所述一个测量基因对应的样本表达值不小于第一阈值；

将命中的样本细胞的数量不小于数量阈值的测量基因作为候选基因。

11.根据权利要求8所述的方法，其特征在于，所述第一提取子模型为由至少一个基于

注意力机制的编码器依次连接得到的语言模型。

12.一种数据处理装置,其特征在于,所述装置包括:

第一获取单元,用于获取目标细胞的目标基因表达数据,从所述目标基因表达数据中提取所述目标细胞下的各个候选基因对应的目标表达值;

第二获取单元,用于基于所述各个候选基因对应的目标表达值,获取所述各个候选基因对应的目标特征;

提取单元,用于基于所述各个候选基因对应的目标特征,提取所述各个候选基因在所述目标细胞下的相互作用信息,所述各个候选基因在所述目标细胞下的相互作用信息用于表征所述目标细胞的基因相互作用特征;

第三获取单元,用于基于所述各个候选基因在所述目标细胞下的相互作用信息,获取所述目标细胞对应的预测结果。

13.一种数据处理模型的训练装置,其特征在于,所述装置包括:

第一获取单元,用于获取样本细胞的样本基因表达数据和所述样本细胞对应的标准结果,从所述样本基因表达数据中提取所述样本细胞下的各个候选基因对应的样本表达值;

第二获取单元,用于基于所述各个候选基因对应的样本表达值,获取所述各个候选基因对应的样本特征;

提取单元,用于调用第一数据处理模型基于所述各个候选基因对应的样本特征,提取所述各个候选基因在所述样本细胞下的相互作用信息;

第三获取单元,用于基于所述各个候选基因在所述样本细胞下的相互作用信息,获取所述样本细胞对应的预测结果;

训练单元,用于基于所述样本细胞对应的预测结果和标准结果,获取结果损失函数;利用所述结果损失函数对所述第一数据处理模型进行训练,得到目标数据处理模型。

14.一种计算机设备,其特征在于,所述计算机设备包括处理器和存储器,所述存储器中存储有至少一条计算机程序,所述至少一条计算机程序由所述处理器加载并执行,以使所述计算机设备实现如权利要求1至6任一所述的数据处理方法,或者如权利要求7至11任一所述的数据处理模型的训练方法。

15.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质中存储有至少一条计算机程序,所述至少一条计算机程序由处理器加载并执行,以使计算机实现如权利要求1至6任一所述的数据处理方法,或者如权利要求7至11任一所述的数据处理模型的训练方法。

16.一种计算机程序产品,其特征在于,所述计算机程序产品包括计算机程序或计算机指令,所述计算机程序或所述计算机指令由处理器加载并执行,以使计算机实现如权利要求1至6任一所述的数据处理方法,或者如权利要求7至11任一所述的数据处理模型的训练方法。

数据处理、数据处理模型的训练方法、装置、设备及介质

技术领域

[0001] 本申请实施例涉及计算机技术领域，特别涉及一种数据处理、数据处理模型的训练方法、装置、设备及介质。

背景技术

[0002] 随着计算机技术的发展，对细胞的研究越来越广泛，例如，对单个细胞的转录组进行测序，得到该细胞的基因表达数据，进而根据该细胞的基因表达数据，获取该细胞对应的预测结果(如，分类结果、回归结果等)。

[0003] 相关技术中，由研究者根据经验确定与各种预测结果分别匹配的特定基因，在确定一个细胞对应的预测结果的过程中，从该细胞的基因表达数据中提取该细胞下某一特定基因对应的表达值，若该表达值满足高表达条件，则将与该特定基因匹配的预测结果作为该细胞对应的预测结果。

[0004] 此种数据处理过程依赖研究者的先验知识，存在较多的不稳定因素，此外，数据处理过程中依赖个别特定基因，该个别特定基因的缺失或噪声对预测结果的准确性有较大影响，难以获取较为准确的预测结果。

发明内容

[0005] 本申请实施例提供了一种数据处理、数据处理模型的训练方法、装置、设备及介质，可用于提高数据处理的稳定性以及得到的预测结果的准确性。所述技术方案如下：

[0006] 一方面，本申请实施例提供了一种数据处理方法，所述方法包括：

[0007] 获取目标细胞的目标基因表达数据，从所述目标基因表达数据中提取所述目标细胞下的各个候选基因对应的目标表达值；

[0008] 基于所述各个候选基因对应的目标表达值，获取所述各个候选基因对应的目标特征；

[0009] 基于所述各个候选基因对应的目标特征，提取所述各个候选基因在所述目标细胞下的相互作用信息，所述各个候选基因在所述目标细胞下的相互作用信息用于表征所述目标细胞的基因相互作用特征；

[0010] 基于所述各个候选基因在所述目标细胞下的相互作用信息，获取所述目标细胞对应的预测结果。

[0011] 还提供了一种数据处理模型的训练方法，所述方法包括：

[0012] 获取样本细胞的样本基因表达数据和所述样本细胞对应的标准结果，从所述样本基因表达数据中提取所述样本细胞下的各个候选基因对应的样本表达值；

[0013] 基于所述各个候选基因对应的样本表达值，获取所述各个候选基因对应的样本特征；

[0014] 调用第一数据处理模型基于所述各个候选基因对应的样本特征，提取所述各个候选基因在所述样本细胞下的相互作用信息；基于所述各个候选基因在所述样本细胞下的相

互作用信息，获取所述样本细胞对应的预测结果；

[0015] 基于所述样本细胞对应的预测结果和标准结果，获取结果损失函数；利用所述结果损失函数对所述第一数据处理模型进行训练，得到目标数据处理模型。

[0016] 另一方面，提供了一种数据处理装置，所述装置包括：

[0017] 第一获取单元，用于获取目标细胞的目标基因表达数据，从所述目标基因表达数据中提取所述目标细胞下的各个候选基因对应的目标表达值；

[0018] 第二获取单元，用于基于所述各个候选基因对应的目标表达值，获取所述各个候选基因对应的目标特征；

[0019] 提取单元，用于基于所述各个候选基因对应的目标特征，提取所述各个候选基因在所述目标细胞下的相互作用信息，所述各个候选基因在所述目标细胞下的相互作用信息用于表征所述目标细胞的基因相互作用特征；

[0020] 第三获取单元，用于基于所述各个候选基因在所述目标细胞下的相互作用信息，获取所述目标细胞对应的预测结果。

[0021] 在一种可能实现方式中，所述第二获取单元，用于将第一候选基因对应的目标表达值转换成所述第一候选基因对应的表达值特征，所述第一候选基因为所述各个候选基因中的任一候选基因；对所述第一候选基因对应的表达值特征和所述第一候选基因对应的表征特征进行融合，得到所述第一候选基因对应的目标特征。

[0022] 在一种可能实现方式中，所述第二获取单元，用于对所述第一候选基因对应的目标表达值进行归一化处理，得到所述第一候选基因对应的归一化表达值；确定所述归一化表达值对应的目标离散化表达值，将所述目标离散化表达值对应的嵌入特征作为所述第一候选基因对应的表达值特征。

[0023] 在一种可能实现方式中，所述目标离散化表达值为参考数量个候选离散化表达值中的一个候选离散化表达值，所述第二获取单元，还用于对所述参考数量个候选离散化表达值进行向量化转换，得到所述参考数量个候选离散化表达值分别对应的嵌入特征。

[0024] 在一种可能实现方式中，所述提取单元，用于调用目标数据处理模型基于所述各个候选基因对应的目标特征，提取所述各个候选基因在所述目标细胞下的相互作用信息；

[0025] 所述第三获取单元，用于调用所述目标数据处理模型基于所述各个候选基因在所述目标细胞下的相互作用信息，获取所述目标细胞对应的预测结果。

[0026] 在一种可能实现方式中，所述目标细胞对应的预测结果指示所述目标细胞的类别为目标类别，所述装置还包括：

[0027] 确定单元，用于基于所述各个候选基因在所述目标细胞下的相互作用信息，在所述各个候选基因中确定属于所述目标类别的细胞对应的满足选取条件的基因。

[0028] 还提供了一种数据处理模型的训练装置，所述装置包括：

[0029] 第一获取单元，用于获取样本细胞的样本基因表达数据和所述样本细胞对应的标准结果，从所述样本基因表达数据中提取所述样本细胞下的各个候选基因对应的样本表达值；

[0030] 第二获取单元，用于基于所述各个候选基因对应的样本表达值，获取所述各个候选基因对应的样本特征；

[0031] 提取单元，用于调用第一数据处理模型基于所述各个候选基因对应的样本特征，

提取所述各个候选基因在所述样本细胞下的相互作用信息；

[0032] 第三获取单元，用于基于所述各个候选基因在所述样本细胞下的相互作用信息，获取所述样本细胞对应的预测结果；

[0033] 训练单元，用于基于所述样本细胞对应的预测结果和标准结果，获取结果损失函数；利用所述结果损失函数对所述第一数据处理模型进行训练，得到目标数据处理模型。

[0034] 在一种可能实现方式中，所述第一数据处理模型包括第一提取子模型和第一预测子模型；所述提取单元，用于调用所述第一提取子模型对所述各个候选基因对应的样本特征进行信息提取，得到所述各个候选基因在所述样本细胞下的相互作用信息；

[0035] 所述第三获取单元，用于调用所述第一预测子模型对所述各个候选基因在所述样本细胞下的相互作用信息进行处理，得到所述样本细胞对应的预测结果。

[0036] 在一种可能实现方式中，所述第一获取单元，还用于获取训练细胞的训练基因表达数据，从所述训练基因表达数据中提取所述训练细胞下的所述各个候选基因对应的训练表达值；

[0037] 所述第二获取单元，还用于基于所述各个候选基因对应的训练表达值，获取所述各个候选基因对应的训练特征；

[0038] 所述装置还包括：

[0039] 替换单元，用于将所述各个候选基因中满足替换条件的候选基因对应的训练特征替换为参考特征；

[0040] 所述提取单元，还用于调用初始提取子模型对所述满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征进行信息提取，得到所述各个候选基因在所述训练细胞下的相互作用信息；基于所述各个候选基因在所述训练细胞下的相互作用信息，获取所述满足替换条件的候选基因对应的预测特征；

[0041] 所述训练单元，还用于基于所述满足替换条件的候选基因对应的预测特征和训练特征，获取特征损失函数，利用所述特征损失函数对所述初始提取子模型进行训练，得到所述第一提取子模型。

[0042] 在一种可能实现方式中，所述样本细胞的数量为至少一个，一个样本细胞的样本基因表达数据包括所述一个样本细胞下的各个测量基因对应的样本表达值，所述装置还包括：

[0043] 确定单元，用于基于各个样本细胞的样本基因表达数据，统计所述各个测量基因分别命中的样本细胞的数量，一个测量基因命中一个样本细胞用于指示所述一个样本细胞下的所述一个测量基因对应的样本表达值不小于第一阈值；将命中的样本细胞的数量不小于数量阈值的测量基因作为候选基因。

[0044] 在一种可能实现方式中，所述第一提取子模型为由至少一个基于注意力机制的编码器依次连接得到的语言模型。

[0045] 另一方面，提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条计算机程序，所述至少一条计算机程序由所述处理器加载并执行，以使所述计算机设备实现上述任一所述的数据处理方法或数据处理模型的训练方法。

[0046] 另一方面，还提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一条计算机程序，所述至少一条计算机程序由处理器加载并执行，以使计算机实现

上述任一所述的数据处理方法或数据处理模型的训练方法。

[0047] 另一方面，还提供了一种计算机程序产品，所述计算机程序产品包括计算机程序或计算机指令，所述计算机程序或所述计算机指令由处理器加载并执行，以使计算机实现上述任一所述的数据处理方法或数据处理模型的训练方法。

[0048] 本申请实施例提供的技术方案至少带来如下有益效果：

[0049] 本申请实施例提供的技术方案，自动根据目标细胞的目标基因表达数据获取目标细胞对应的预测结果，无需依赖研究者的先验知识，数据处理的稳定性较高。此外，目标细胞对应的预测结果是基于各个候选基因在目标细胞下的相互作用信息获取的，各个候选基因在目标细胞下的相互作用信息能够表征出目标细胞的基因相互作用特征，由于细胞在生物体中是通过基因之间的相互作用发挥功能的，所以基因相互作用特征能够体现出目标细胞的功能方面的特征，通过关注目标细胞的功能方面的特征获取的预测结果的准确性较高。

附图说明

[0050] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0051] 图1是本申请实施例提供的一种实施环境的示意图；

[0052] 图2是本申请实施例提供的一种数据处理方法的流程图；

[0053] 图3是本申请实施例提供的一种数据处理方法的流程图；

[0054] 图4是本申请实施例提供的一种数据处理模型的训练方法的流程图；

[0055] 图5是本申请实施例提供的一种对初始提取子模型进行训练的过程的示意图；

[0056] 图6是本申请实施例提供的一种数据处理模型的训练过程的示意图；

[0057] 图7是本申请实施例提供的一种数据处理装置的示意图；

[0058] 图8是本申请实施例提供的一种数据处理模型的训练装置的示意图；

[0059] 图9是本申请实施例提供的一种服务器的结构示意图；

[0060] 图10是本申请实施例提供的一种终端的结构示意图。

具体实施方式

[0061] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0062] 在示例性实施例中，本申请实施例提供的数据处理方法以及数据处理模型的训练方法可应用于各种场景，包括但不限于云技术、人工智能、智慧交通、辅助驾驶等。

[0063] 人工智能(Artificial Intelligence，AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，人工智能企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的

功能。

[0064] 人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习、自动驾驶、智慧交通等几大方向。

[0065] 机器学习(Machine Learning，ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、示教学习等技术。

[0066] 随着人工智能技术研究和进步，人工智能技术在多个领域展开研究和应用，例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服、车联网、自动驾驶、智慧交通等，相信随着技术的发展，人工智能技术将在更多的领域得到应用，并发挥越来越重要的价值。

[0067] 图1示出了本申请实施例提供的实施环境的示意图。该实施环境包括：终端11和服务器12。

[0068] 本申请实施例提供的数据处理方法可以由终端11执行，也可以由服务器12执行，还可以由终端11和服务器12共同执行，本申请实施例对此不加以限定。对于本申请实施例提供的数据处理方法由终端11和服务器12共同执行的情况，服务器12承担主要计算工作，终端11承担次要计算工作；或者，服务器12承担次要计算工作，终端11承担主要计算工作；或者，服务器12和终端11二者之间采用分布式计算架构进行协同计算。

[0069] 本申请实施例提供的数据处理模型的训练方法可以由终端11执行，也可以由服务器12执行，还可以由终端11和服务器12共同执行，本申请实施例对此不加以限定。对于本申请实施例提供的数据处理模型的训练方法由终端11和服务器12共同执行的情况，服务器12承担主要计算工作，终端11承担次要计算工作；或者，服务器12承担次要计算工作，终端11承担主要计算工作；或者，服务器12和终端11二者之间采用分布式计算架构进行协同计算。

[0070] 数据处理方法的执行设备与数据处理模型的训练方法的执行设备可以相同，也可以不同，本申请实施例对此不加以限定。

[0071] 在一种可能实现方式中，终端11可以是任何一种可与用户通过键盘、触摸板、触摸屏、遥控器、语音交互或手写设备等一种或多种方式进行人机交互的电子产品，例如PC(Personal Computer，个人计算机)、手机、智能手机、PDA(Personal Digital Assistant，个人数字助手)、可穿戴设备、PPC(Pocket PC，掌上电脑)、平板电脑、智能车机、智能电视、智能音箱、智能语音交互设备、智能家电、车载终端等。服务器12可以是一台服务器，也可以是由多台服务器组成的服务器集群，或者是一个云计算服务中心。终端11与服务器12通过有线或无线网络建立通信连接。

[0072] 本领域技术人员应能理解上述终端11和服务器12仅为举例，其他现有的或今后可能出现的终端或服务器如可适用于本申请，也应包含在本申请保护范围以内，并在此以引

用方式包含于此。

[0073] 基于上述图1所示的实施环境，本申请实施例提供一种数据处理方法，该数据处理方法由计算机设备执行，该计算机设备可以为终端11，也可以为服务器12，本申请实施例对此不加以限定。如图2所示，本申请实施例提供的数据处理方法包括如下步骤201至步骤203。

[0074] 在步骤201中，获取目标细胞的目标基因表达数据，从目标基因表达数据中提取目标细胞下的各个候选基因对应的目标表达值。

[0075] 目标细胞是指待获取预测结果的细胞，目标细胞的目标基因表达数据用于指示目标细胞的基因表达情况。在示例性实施例中，目标细胞的目标基因表达数据通过对目标细胞的转录组进行测序得到。目标细胞的转录组是指某一时刻该目标细胞内所有mRNA(信使核糖核酸)的总表达量。mRNA与基因之间存在一一对应关系，通过对目标细胞的转录组进行测序，能够确定出目标细胞的目标基因表达数据。对目标细胞的转录组进行测序所利用的测序技术可以根据实际的应用场景灵活调整，本申请实施例对此不加以限定。

[0076] 目标细胞的目标基因表达数据包括该目标细胞下的各个测量基因对应的目标表达值，任一测量基因对应的目标表达值用于指示该目标细胞的转录组中与该任一测量基因对应的mRNA的数量。测量基因是指在测序过程中关注的基因，本申请实施例对测量基因的类型以及数量不加以限定，在利用不同的测序技术对目标细胞的转录组进行测序的情况下，测量基因的类型以及数量可能相同，也可能不同。

[0077] 在示例性实施例中，目标细胞的目标基因表达数据可以预先获取并存储，此种方式下，在需要获取目标细胞对应的预测结果的情况下，直接提取目标细胞的目标基因表达数据。在示例性实施例中，计算机设备中预先存储有目标细胞的转录组，此种情况下，在需要获取目标细胞对应的预测结果的情况下，通过对目标细胞的转录组进行测序，得到目标细胞的目标基因表达数据。在示例性实施例中，计算机设备中未存储有目标细胞的转录组和目标细胞的目标基因表达数据，则获取目标细胞的目标基因表达数据的方式为：对目标细胞进行转录，得到目标细胞的转录组；对目标细胞的转录组进行测序，得到目标细胞的目标基因表达数据。

[0078] 在获取目标细胞的目标基因表达数据后，从目标基因表达数据中提取该目标细胞下的各个候选基因对应的目标表达值。候选基因是指在获取目标细胞对应的预测结果的过程中所关注的基因，示例性地，候选基因为测序基因中的部分或全部基因。由于目标基因表达数据中包括目标细胞下的各个测序基因对应的目标表达值，所以在获取目标基因表达数据后，能够从目标基因表达数据中提取出目标细胞下的各个候选基因对应的目标表达值。任一候选基因对应的目标表达值用于指示该目标细胞的转录组中与该任一候选基因对应的mRNA的数量。

[0079] 示例性地，候选基因根据经验设置，或者根据应用场景灵活调整，本申请实施例对此不加以限定。示例性地，候选基因为通过对大量细胞的基因表达数据进行分析确定的可靠性较高的基因。

[0080] 在步骤202中，基于各个候选基因对应的目标表达值，获取各个候选基因对应的目标特征。

[0081] 各个候选基因对应的目标特征是用于获取目标细胞对应的预测结果所依据的特

征，各个候选基因对应的目标特征基于各个候选基因对应的目标表达值确定。本申请实施例对候选基因对应的目标特征的形式不加以限定，示例性地，候选基因对应的目标特征的形式为多维向量，该多维向量的维度根据经验设置，或者根据应用场景灵活调整。

[0082] 一个候选基因对应的目标特征是基于该一个候选基因对应的目标表达值获取的，以第一候选基因为例，介绍基于第一候选基因对应的目标表达值，获取第一候选基因对应的目标特征的过程。其中该，第一候选基因为各个候选基因中的任一候选基因。在一种可能实现方式中，基于第一候选基因对应的目标表达值，获取第一候选基因对应的目标特征的过程包括以下步骤2021和步骤2022。

[0083] 在步骤2021中，将第一候选基因对应的目标表达值转换成第一候选基因对应的表达值特征。

[0084] 第一候选基因对应的表达值特征用于对第一候选基因对应的目标表达值进行表征。在一种可能实现方式中，将第一候选基因对应的目标表达值转换成第一候选基因对应的表达值特征的实现过程包括以下步骤1和步骤2。

[0085] 步骤1：对第一候选基因对应的目标表达值进行归一化处理，得到第一候选基因对应的归一化表达值。

[0086] 对第一候选基因对应的目标表达值进行归一化处理，有利于提高第一候选基因对应的表达值的规范性，将对第一候选基因对应的目标表达值进行归一化处理后得到的值称为第一候选基因对应的归一化表达值。

[0087] 在一种可能实现方式中，对第一候选基因对应的目标表达值进行归一化处理，得到第一候选基因对应的归一化表达值的过程包括：计算第一候选基因对应的目标表达值与第一倍数的乘积，得到第一表达值；基于第一表达值，获取第一候选基因对应的归一化表达值。

[0088] 第一倍数为参考总量和第一总量的比值，其中，参考总量为期望的某一细胞下的各个候选基因对应的表达值之和，第一总量是指目标细胞下的各个候选基因对应的表达值之和。参考总量根据经验设置，或者根据应用场景灵活调整，本申请实施例对此不加以限定。第一总量通过将目标细胞下的各个候选基因对应的目标表达值相加得到。

[0089] 需要说明的是，对于各个候选基因对应的目标表达值而言，第一倍数是相同的。示例性地，假设参考总量为10000，第一总量为2000，则第一倍数为5，通过将各个候选基因对应的目标表达值分别放大5倍，能够达到通过等比例缩放使各个候选基因对应的表达值之和达到参考总量的效果。

[0090] 在通过计算第一候选基因对应的目标表达值与第一倍数的乘积，得到第一表达值后，基于第一表达值，获取第一候选基因对应的归一化表达值。在示例性实施例中，直接将第一表达值作为第一候选基因对应的归一化表达值。在示例性实施例中，将第一表达值取对数后得到的值作为第一候选基因对应的归一化表达值。通过对第一表达值取对数，能够调整第一表达值的长尾分布，进一步提高得到的归一化表达值的规范性。

[0091] 步骤2：确定归一化表达值对应的目标离散化表达值，将目标离散化表达值对应的嵌入特征作为第一候选基因对应的表达值特征。

[0092] 在确定第一候选基因对应的归一化表达值后，进一步确定该归一化表达值对应的目标离散化表达值。在示例性实施例中，计算机设备中存储有归一化表达值和候选离散化

表达值的对应关系，根据该对应关系能够直接确定目标离散化表达值。

[0093] 在示例性实施例中，计算机设备中存储有候选离散化表达值与归一化表达值取值范围的对应关系，此种情况下，先确定归一化表达值所属的目标归一化表达值取值范围，然后将该目标归一化表达值取值范围对应的候选离散化表达值作为目标离散化表达值。

[0094] 在示例性实施例中，候选离散化表达值与归一化表达值取值范围的对应关系的获取方式包括：将候选归一化表达值根据数值从大到小或者从小到大的顺序划分到参考数量个统计桶(bin)中，为每个统计桶设置一个候选离散化表达值；将划分该一个统计桶中的各个候选的归一化表达值构成的归一化表达值取值范围作为为该一个统计桶设置的候选离散化表达值对应的归一化表达值取值范围。通过此种方式，能够设置出参考数量个候选离散化表达值，每个候选离散化表达值均对应一个归一化表达值取值范围。

[0095] 候选归一化表达值是指基因可能对应的归一化表达值，示例性地，候选归一化表达值是指连续型变量，候选归一化表达值的数量以及具体取值根据经验或者根据实际情况确定，本申请实施例对此不加以限定。在示例性实施例中，将候选归一化表达值根据数值从大到小或者从小到大的顺序划分到参考数量个统计桶中可以是指将候选归一化表达值根据数值从大到小或者从小到大的顺序均匀划分到参考数量个统计桶中，也可以是指将候选归一化表达值根据数值从大到小或者从小到大的顺序非均匀划分到参考数量个统计桶中，本申请实施例对此不加以限定。

[0096] 参考数量用于为候选离散化表达值的数量进行约束，参考数量根据经验设置，或者根据应用场景灵活调整，本申请实施例对此不加以限定。

[0097] 示例性地，参考数量为50，则候选离散化表达值的数量为50个。假设候选归一化表达值共有5000个，根据数值从小到大的顺序将该5000个候选归一化表达值均匀划分到50个统计桶中，每1000个候选归一化表达值划分到一个统计桶中，那么，任意一个候选归一化表达值被划分到第x(x为不大于50的正整数)个统计桶中，将第x个统计桶对应的候选离散化表达值设置为Bx，则被划分到该第x个统计桶中的各个候选归一化表达值对应的候选离散化表达值均为Bx。示例性地，若一个归一化表达值为零，则将该归一化表达值对应的离散化表达值记为Zero。

[0098] 在确定目标离散化表达值后，将目标离散化表达值对应的嵌入特征作为第一候选基因对应的表达值特征。目标离散化表达值为一个数值，目标离散化表达值对应的嵌入特征是指该目标离散化表达值对应的一个多维的特征。

[0099] 在一种可能实现方式中，目标离散化表达值为参考数量个候选离散化表达值中的一个候选离散化表达值，计算机设备中存储有各个候选离散化表达值分别对应的嵌入特征，通过查询可以确定目标离散化表达值对应哪个嵌入特征，从而得到第一候选基因对应的表达值特征。

[0100] 示例性地，获取参考数量个候选离散化表达值分别对应的嵌入特征的方式为：对参考数量个候选离散化表达值进行向量化转换，得到参考数量个候选离散化表达值分别对应的嵌入特征。

[0101] 参考数量个候选离散化表达值用于实现对归一化表达值的离散化，参考数量个候选离散化表达值可以根据经验设置，或者根据应用场景灵活调整，本申请实施例对此不加以限定。通过对参考数量个候选离散化表达值进行向量化转换，能够得到参考数量个候选

离散化表达值分别对应的嵌入特征。示例性地，每个候选离散化表达值对应的嵌入特征均为相同维数的向量，如，均为200维的向量。

[0102] 示例性地，利用词向量技术对参考数量个候选离散化表达值进行向量化转换，使得每个候选离散化表达值均被转换成一个多维的向量，将该多维的向量作为候选离散化表达值对应的嵌入特征。词向量技术为一种生成多个嵌入特征的技术，示例性地，词向量技术能够使用pytorch(一个开源的机器学习库)中的embedding(嵌入)函数实现。示例性地，向量化转换的原则为：将参考数量个候选离散化表达值转换成参考数量个嵌入特征，转换后得到的参考数量个嵌入特征中的每两个嵌入特征之间的距离均与其他任两个嵌入特征之间的距离相同。通过向量化转换，每个候选离散化表达值Bx均会被转换成类似[0.14,-0.33,……,0.75,0.28]的向量形式，该向量的维度可以灵活设定，示例性地，该向量的维度为200。

[0103] 需要说明的是，以上所述基于第一候选基因对应的目标表达值，获取第一候选基因对应的表达值特征的过程仅为示例性举例，本申请实施例并不局限于此。在示例性实施例中，还可以直接确定第一候选基因对应的目标表达值对应的离散化表达值，将该离散化表达值对应的嵌入特征作为第一候选基因对应的表达值特征。示例性地，不同的离散化表达值对应不同的表达值取值范围，确定第一候选基因对应的目标表达值对应的离散化表达值的方式为：将第一候选基因对应的目标表达值所属的表达值取值范围对应的离散化表达值作为第一候选基因对应的目标表达值对应的离散化表达值。

[0104] 在步骤2022中，基于第一候选基因对应的表达值特征，获取第一候选基因对应的目标特征。

[0105] 在获取第一候选基因在对应的表达值特征后，在第一候选基因对应的表达值特征的基础上，进一步获取第一候选基因对应的目标特征。

[0106] 在一种可能实现方式中，该步骤2022的实现方式包括：直接将第一候选基因对应的表达值特征作为第一候选基因对应的目标特征。

[0107] 在另一种可能实现方式中，该步骤2022的实现方式包括：对第一候选基因对应的表达值特征和第一候选基因对应的表征特征进行融合，得到第一候选基因对应的目标特征。

[0108] 第一候选基因对应的表征特征用于对该第一候选基因本身进行表征，与细胞无关。示例性地，第一候选基因对应的表征特征是基于第一候选基因的语义信息确定的。本申请实施例对第一候选基因对应的表征特征的获取方式不加以限定，示例性地，第一候选基因对应的表征特征是基于Gene2Vec(一种基因表征方式)方式提取到的特征。基于Gene2Vec方式提取到的任两个候选基因对应的表征特征之间的相似度越高，说明该任两个候选基因之间的相互作用越强。基于Gene2Vec方式提取到的特征的维度可以根据应用场景灵活调整，例如，对于基因EGFR，基于Gene2Vec方式提取到的特征为一个200维的向量[0.76,0.23,……,-0.49,0.15]。当然，在示例性实施例中，第一候选基因对应的表征特征还可以基于其他方式提取得到，只要能够对第一候选基因本身进行表征即可。

[0109] 通过综合考虑第一候选基因对应的表达值特征以及第一候选基因对应的表征特征获取第一候选基因对应的目标特征，能够利用第一候选基因对应的目标特征为后续获取目标细胞对应的预测结果的过程提供更丰富的信息，有利于提高获取的预测结果的可靠

性。

[0110] 在示例性实施例中，第一候选基因对应的表达值特征以及第一候选基因对应的表征特征的表示形式相同，对第一候选基因对应的表达值特征和第一候选基因对应的表征特征进行融合后得到的第一候选基因对应的目标特征的表示形式与第一候选基因对应的表达值特征以及第一候选基因对应的表征特征的表示形式相同。

[0111] 示例性地，第一候选基因对应的表达值特征以及第一候选基因对应的表征特征均为一个指定维数的向量，对第一候选基因对应的表达值特征和第一候选基因对应的表征特征进行融合的过程通过对两个向量中的相同位置的元素相加或者求平均实现。融合得到的第一候选基因对应的目标特征同样为一个指定维数的向量。指定维数根据经验设置，或者根据应用场景灵活调整，本申请实施例对此不加以限定。例如，指定维数为200维。

[0112] 在步骤203中，基于各个候选基因对应的目标特征，提取各个候选基因在目标细胞下的相互作用信息，各个候选基因在目标细胞下的相互作用信息用于表征目标细胞的基因相互作用特征；基于各个候选基因在目标细胞下的相互作用信息，获取目标细胞对应的预测结果。

[0113] 在获取各个候选基因对应的目标特征后，基于各个候选基因对应的目标特征，提取各个候选基因在目标细胞下的相互作用信息。各个候选基因在目标细胞下的相互作用信息不仅关注各个候选基因自身在目标细胞下的表达情况，还关注各个候选基因与其他候选基因在目标细胞下的表达情况之间的差异，各个候选基因自身在目标细胞下的表达情况以及各个候选基因与其他候选基因在目标细胞下的表达情况之间的差异能够对目标细胞的基因相互作用特征进行表征，也就是说，各个候选基因在目标细胞下的相互作用信息用于表征目标细胞基因相互作用特征。由于目标细胞在生物体中的功能利用基因之间的相互作用实现，所以目标细胞的基因相互作用特征用于描述目标细胞在生物体中的功能，目标细胞的基因相互作用特征可视为目标细胞的本质层面的特征。基于用于表征目标细胞的基因相互作用特征的相互作用信息，能够获取到更为准确且更为可靠的预测结果。

[0114] 在示例性实施例中，各个候选基因在目标细胞下的相互作用信息能够指示出每个候选基因在目标细胞下与各个候选基因的相互作用信息。在示例性实施例中，各个候选基因在目标细胞下的相互作用信息的形式为权重矩阵，该权重矩阵的维度为候选基因的数量*候选基因的数量，通过该矩阵能够确定候选基因在目标细胞下的两两相互作用信息。示例性地，两个候选基因在目标细胞下的相互作用信息利用权重矩阵中的一个权重值表示，该权重值越大，说明该两个候选基因在目标细胞下的相互作用关系越强，该权重值越小，说明该两个候选基因在目标细胞下的相互作用关系越弱。

[0115] 在示例性实施例中，基于各个候选基因对应的目标特征，提取各个候选基因在目标细胞下的相互作用信息的实现过程为：利用注意力机制对各个候选基因对应的目标特征进行处理，得到各个候选基因在目标细胞下的相互作用信息。此种方式下，各个候选基因在目标细胞下的相互作用信息还可以称为注意力权重矩阵。

[0116] 在示例性实施例中，目标细胞对应的预测结果的类型与实际的应用场景有关，本申请实施例对此不加以限定。示例性地，在根据目标细胞的基因表达数据对目标细胞进行分类的应用场景下，目标细胞对应的预测结果为目标细胞对应的分类结果，目标细胞对应的分类结果用于指示目标细胞的类别。示例性地，在根据目标细胞的基因表达数据对目标

细胞进行回归的应用场景下，目标细胞对应的预测结果为目标细胞对应的回归结果，目标细胞对应的回归结果用于指示目标细胞的回归值。回归值指示的含义与对目标细胞进行回归的研究目的有关，例如，对目标细胞进行回归的研究目标是研究目标细胞的发育时间，则该回归值用于指示目标细胞的发育时间，通过此种方式能够为下游的发育轨迹分析任务提供数据支持。示例性地，应用场景还可以为根据目标细胞的基因表达数据对目标细胞进行聚类的场景，在此种场景下，目标细胞对应的预测结果为目标细胞对应的聚类结果。

[0117] 在示例性实施例中，该步骤203可以通过运行预先编辑的计算机程序实现，也可以通过调用目标数据处理模型实现，本申请实施例对此不加以限定。本申请实施例以该步骤203通过调用目标数据处理模型实现为例进行说明。也就是说，该步骤203的实现方式包括：调用目标数据处理模型基于各个候选基因对应的目标特征，提取各个候选基因在目标细胞下的相互作用信息；调用目标数据处理模型基于各个候选基因在目标细胞下的相互作用信息，获取目标细胞对应的预测结果。

[0118] 目标数据处理模型为训练得到的能够根据各个候选基因对应的目标特征，输出较为准确的目标细胞对应的预测结果的模型。训练得到目标数据处理模型的过程参见图4所示的实施例，此处暂不赘述。

[0119] 调用目标数据处理模型获取目标细胞对应的预测结果的实现过程为目标数据处理模型的内部处理过程，与目标数据模型的结构有关，本申请实施例对目标数据处理模型的模型结构不加以限定。

[0120] 在示例性实施例中，目标数据处理模型为一个整体的模型。在此种情况下，通过将各个候选基因对应的目标特征输入目标数据处理模型，目标数据处理模型自动提取各个候选基因在目标细胞下的相互作用信息，进而直接基于候选基因在目标细胞下的相互作用信息，获取并输出目标细胞对应的预测结果。在示例性实施例中，将各个候选基因对应的目标特征输入目标数据处理模型是指将各个候选基因对应的目标特征按照各个候选基因对应的指定排列顺序依次排列，构成特征序列，将特征序列输入目标数据处理模型。指定排列根据经验设置，或者根据经验场景灵活调整，本申请实施例对此不加以限定。

[0121] 在示例性实施例中，目标数据处理模型包括目标提取子模型和目标预测子模型。此种情况下，参见图3，该步骤203的实现过程包括：301，调用目标提取子模型对各个候选基因对应的目标特征进行信息提取，得到各个候选基因在目标细胞下的相互作用信息；302，调用目标预测子模型对各个候选基因在目标细胞下的相互作用信息进行处理，得到目标细胞对应的预测结果。

[0122] 也就是说，将各个候选基因对应的目标特征输入目标提取子模型，目标提取子模型提取各个候选基因在目标细胞下的相互作用信息，然后将各个候选基因在目标细胞下的相互作用信息输入目标预测模型，由目标预测模型获取并输出目标细胞对应的预测结果。

[0123] 调用目标提取子模型对各个候选基因对应的目标特征进行信息提取，得到各个候选基因在目标细胞下的相互作用信息的过程为目标提取子模型的内部处理过程，与目标提取子模型的模型结构有关。

[0124] 示例性地，目标提取子模型为由至少一个基于注意力机制的编码器依次连接得到的语言模型，如，BERT(Bidirectional Encoder Representations from Transformers，基于转换器的双向编码表征)，或者通过对BERT中的编码器进行改进得到的深度双向语言模

型。也就是说，目标提取子模型包括依次连接的至少一个基于注意力机制的编码器。示例性地，每个基于注意力机制的编码器均具有根据输入的特征提取相互作用信息以及根据提取的相互作用信息输出与输入的特征维度相同的特征的功能。示例性地，每个基于注意力机制的编码器均包括一个注意力提取层和一个特征输出层，其中，注意力提取层用于根据输入的特征提取相互作用信息，特征输出层用于根据提取的相互作用信息输出与输入的特征维度相同的特征。

[0125] 在示例性实施例中，对于目标提取子模型包括依次连接的至少一个基于注意力机制的编码器的情况，各个候选基因在目标细胞下的相互作用信息是最后一个基于注意力机制的编码器提取到的相互作用信息。

[0126] 示例性地，基于注意力机制的编码器可以是指基于单头或多头自注意力机制的编码器。本申请实施例对基于注意力机制的编码器的类型不加以限定，示例性地，基于注意力机制的编码器可以是指Performer(一种基于注意力机制的编码器)、Transformer(一种基于注意力机制的编码器)、Reformer(一种基于注意力机制的编码器)、Linformer(一种基于注意力机制的编码器)等。示例性地，Performer是一种基于广义注意力机制的编码器，能够大大降低时间计算复杂度，实现超长序列数据的高效处理以及远距离特征关系的学习。

[0127] 本申请实施例对目标提取子模型包括的基于注意力机制的编码器的数量不加以限定，可以根据经验设置，或者根据实际的应用需求灵活调整，示例性地，目标提取子模型包括的基于注意力机制的编码器的数量为8，或者，目标提取子模型包括的基于注意力机制的编码器的数量为5。

[0128] 调用目标预测子模型对各个候选基因在目标细胞下的相互作用信息进行处理，得到目标细胞对应的预测结果的过程为目标预测子模型的内部处理过程，与目标预测子模型的结构有关。在示例性实施例中，目标预测子模型的结构根据需要获取的目标细胞对应的预测结果的类型灵活设定，本申请实施例对此不加以限定。

[0129] 示例性地，对于需要获取的目标细胞对应的预测结果为分类结果的情况，目标预测子模型包括依次连接的至少一个卷积层和至少一个全连接层。卷积层和全连接层的数量可以根据经验设置，或者根据实际的应用场景灵活调整，本申请实施例对此不加以限定。在将各个候选基因在目标细胞下的相互作用信息输入目标预测子模型后，依次经过至少一个卷积层和至少一个全连接层的处理后，得到最后一个全连接层输出的目标细胞对应的分类结果。

[0130] 最后一个全连接层可视为分类器，该分类器头的数量为各个候选类别的数量，以输出包括各个候选类别分别对应的概率的分类结果，根据该分类结果能够确定出概率最大的候选类别，将该概率最大的候选类别作为目标细胞的类别。候选类别根据实际的分类场景灵活调整，随着候选类别的数量的改变，分类器的头的数量也随之改变。

[0131] 示例性地，全连接层还可以称为前馈神经网络层，经过卷积层的卷积操作，能够有效提取关键信息，经过前馈神经网络层能够得到更准确的分类结果。

[0132] 需要说明的是，上述实施例仅以需要获取目标细胞对应的分类结果为例介绍了目标预测子模型的模型结构，随着需要获取的目标细胞对应的预测结果的类型的变化以及随着实际的应用场景的变化，目标预测子模型的结构可以灵活调整，本申请实施例对此不加以限定。

[0133] 在示例性实施例中，所以在获取目标细胞对应的预测结果之后，可以根据各个候选基因在目标细胞下的相互作用信息实现对细胞的更进一步分析。在示例性实施例中，由于在获取目标细胞对应的预测结果的过程中提取了各个候选基因在目标细胞下的相互作用信息，所以在获取目标细胞对应的预测结果之后，能够直接获取到各个候选基因在目标细胞下的相互作用信息。

[0134] 在一种可能实现方式中，目标细胞对应的预测结果为目标细胞对应的分类结果，该目标细胞对应的分类结果指示目标细胞的类别为目标类别。此种情况下，根据各个候选基因在目标细胞下的相互作用信息实现对细胞的更进一步分析的过程包括：基于各个候选基因在目标细胞下的相互作用信息，在各个候选基因中确定属于目标类别的细胞对应的满足选取条件的基因。示例性地，各个候选基因在目标细胞下的相互作用信息的形式为权重矩阵，权重矩阵的每行或每列均用于指示一个候选基因在目标细胞下与各个候选基因之间的相互作用信息。

[0135] 在示例性实施例中，基于各个候选基因在目标细胞下的相互作用信息，确定属于目标类别的细胞对应的满足选取条件的基因的过程包括：对各个候选基因在目标细胞下的相互作用信息对应的权重矩阵按列(或按行)求和，将对应的和值中前K(K为不小于1的整数)大的列(或行)对应的基因作为属于目标类别的细胞对应的满足选取条件的基因。示例性地，可以直接确定对应的和值中前K(K为不小于1的整数)大的列(或行)，也可以先按照对应的和值的大小顺序对各列(或各行)进行排序，然后再确定对应的和值中前K(K为不小于1的整数)大的列(或行)。

[0136] 对应的和值中前K(K为不小于1的整数)大的列(或行)对应的基因为对目标细胞影响比较大的基因，该对目标细胞影响比较大的基因可以直接作为属于目标类别的细胞对应的满足选取条件的基因。示例性地，满足选取条件的基因还可以称为关键基因。

[0137] 在示例性实施例中，基于各个候选基因在目标细胞下的相互作用信息，在各个候选基因中确定属于目标类别的细胞对应的满足选取条件的基因的过程包括：基于已知属于目标类别的参考细胞的基因表达数据，确定各个候选基因在每个参考细胞下的相互作用信息；将各个候选基因在每个参考细胞下的相互作用信息以及各个候选基因在目标细胞下的相互作用信息进行融合，得到融合相互作用信息；基于融合相互作用信息，从各个候选基因中确定属于目标类别的细胞对应的满足选取条件的基因。参考细胞为已知属于目标类别的细胞，参考细胞可以由研究者根据经验确定出，也可以通过调用数据处理模型的预测结果确定出，本申请实施例对此不加以限定。

[0138] 在示例性实施例中，各个候选基因在任一细胞下的相互作用信息的形式为权重矩阵，将各个候选基因在每个参考细胞下的相互作用信息以及各个候选基因在目标细胞下的相互作用信息进行融合的方式可以为将各个权重矩阵中对应位置的元素求均值。融合相互作用信息的形式与各个候选基因在任一细胞下的相互作用信息的形式相同，例如，融合相互作用信息还可以称为融合权重矩阵。示例性地，此种融合方式可以称为element-wise(逐元素)融合方式。

[0139] 基于融合相互作用信息，在各个候选基因中确定属于目标类别的细胞对应的满足选取条件的基因的原理与直接基于各个候选基因在目标细胞下的相互作用信息，在各个候选基因中确定属于目标类别的细胞对应的满足选取条件的基因的原理相同，此处不再赘

述。

[0140] 基于本申请实施例提供的方式，对于每个细胞的输入数据，可以提取用于指示各个候选基因在该细胞下的相互作用信息的权重矩阵，该权重矩阵的维度为候选基因数量*候选基因数量，该权重矩阵即代表基因之间的两两相互作用关系。通过对属于每个类别的细胞对应的权重矩阵求element-wise的均值，可以得到每个类别的细胞对应的融合权重矩阵，根据每个类别的细胞对应的融合权重矩阵能够比较各个候选基因在不同类别的细胞下的相互作用信息之间的差异。通过对某一类别的细胞对应的融合权重矩阵按列(或按行)求和，再按照数值排序取最大的一部分数值，可以得到对属于不同类别的细胞影响比较大的基因，将对属于某一类别的细胞影响比较大的基因作为属于该类别的细胞对应的关键基因。在得到属于各个类别的细胞分别对应的关键基因后，能够经过Gene Ontology(基因本体论数据库)和KEGG pathway(通路数据库)富集分析得到和细胞类别相关的功能通路。

[0141] 在示例性实施例中，获取目标细胞对应的预测结果之后，可以利用预测结果对目标细胞进行标注。本申请实施例提供的数据处理方法可以为对单细胞转录组的分析提供数据支持。转录组通过测定特定组织样本中的基因表达丰度和类型，揭示复杂生物学通路和性状调控网络分子机制，进而反映人体的临床生理状态差异。传统的转录组测序只能获取组织整体水平的基因表达信息，但对于某些特定组织尤其是肿瘤组织，细胞组成复杂且不同细胞类型可能发挥着特异的生理功能，所以了解组织样本的细胞组成及其异质性至关重要。

[0142] 近年来，各类单细胞转录组测序方法通过实现单细胞分离使得获取组织样本中单个细胞水平的具体基因表达成为常用研究手段。通过单细胞转录组测序，能够构建不同疾病发展阶段的转录组图谱，获取疾病发展的基因表达模式变化情况，进一步为胃癌等疾病的早期诊断提供分子基础。通过单细胞转录组测序，还能够分析特定疾病不同临床特征有关的免疫差异等，深入理解发病机制，从而制定更有针对性的治疗策略。

[0143] 在单细胞转录组的分析中，通过分析不同临床阶段、不同临床特征的组织基因表达模式差异，可以帮助了解组织发育、疾病发展的动态变化与方向，进而明确相关生物机理。不论是疾病的早期精确诊断，还是复杂疾病的针对性治疗策略制定，都能从单细胞转录组的分析中获得极大促进。细胞的预测结果(如，细胞类别)的精准鉴定是实现单细胞转录组分析价值的关键所在，否则可能得到错误的结论，影响疾病的诊断和治疗。本申请实施例依赖基因间相互作用信息的学习，能实现高精度且具可解释性的细胞预测结果获取方法，应用本申请实施例提供的方法对细胞进行标注，能推动单细胞转录组的研究和临床应用，对发现未知机理、治疗复杂疾病有促进作用。

[0144] 本申请实施例提供的数据处理方法，自动根据目标细胞的目标基因表达数据获取目标细胞对应的预测结果，无需依赖研究者的先验知识，数据处理的稳定性较高。此外，目标细胞对应的预测结果是基于各个候选基因在目标细胞下的相互作用信息获取的，各个候选基因在目标细胞下的相互作用信息能够表征出目标细胞的基因相互作用特征，由于细胞在生物体中是通过基因之间的相互作用发挥功能的，所以基因相互作用特征能够体现出目标细胞的功能方面的特征，通过关注目标细胞的功能方面的特征获取的预测结果的准确性较高。

[0145] 基于上述图1所示的实施环境，本申请实施例提供一种数据处理模型的训练方法，

该数据处理模型的训练方法由计算机设备执行，该计算机设备可以为终端11，也可以为服务器12，本申请实施例对此不加以限定。如图4所示，本申请实施例提供的数据处理模型的训练方法包括如下步骤401至步骤404。

[0146] 在步骤401中，获取样本细胞的样本基因表达数据和样本细胞对应的标准结果，从样本基因表达数据中提取样本细胞下的各个候选基因对应的样本表达值。

[0147] 样本细胞为具有标准结果的细胞，样本细胞对应的标准结果可以是研究者根据经验设定的结果，也可以是根据其他细胞研究方式确定的结果，本申请实施例对此不加以限定。样本细胞对应的标准结果的类型与需要利用数据处理模型输出的预测结果的类型相同，以便于利用样本细胞对应的标准结果为数据处理模型的训练过程提供监督信息。由于样本细胞对应有标准结果，所以利用样本细胞对第一数据处理模型进行训练的过程为有监督训练过程。

[0148] 示例性地，样本细胞的样本基因表达数据以及样本数据对应的标准结果存储在数据库中，从数据库中提取得到样本细胞的样本基因表达数据以及样本数据对应的标准结果。示例性地，样本细胞的样本基因表达数据以及样本数据对应的标准结果存储在PanglaoDB(一种单细胞数据库)中。示例性地，样本细胞可以是指特定组织的已经标注的细胞。

[0149] 在示例性实施例中，样本细胞的样本基因表达数据是指满足质量控制条件的细胞的基因表达数据，以保证样本细胞的基因表达数据的可靠性。示例性地，细胞是否满足质量控制条件根据细胞的基因表达数据确定，细胞的基因表达数据包括该细胞基因表达数据下的各个测量基因对应的表达值，若该细胞下对应的表达值不为零的测量基因的数量占各个测量基因的数量的比值不小于比值阈值，则确定该细胞满足质量控制条件。比值阈值根据经验设置，或者根据应用场景灵活调整，本申请实施例对此不加以限定。

[0150] 在示例性实施例中，样本细胞的数量为至少一个，一个样本细胞的样本基因表达数据包括一个样本细胞下的各个测量基因对应的样本表达值，各个候选基因是基于各个样本细胞的样本基因表达数据从各个测量基因中确定出的，在从样本基因表达数据中提取样本细胞下的各个候选基因对应的样本表达值之前，需要先确定出候选基因。示例性地，确定候选基因的过程包括：基于各个样本细胞的样本基因表达数据，统计各个测量基因分别命中的样本细胞的数量，一个测量基因命中一个样本细胞用于指示一个样本细胞下的一个测量基因对应的样本表达值不小于第一阈值；将命中的样本细胞的数量不小于数量阈值的测量基因作为候选基因。

[0151] 第一阈值根据经验设置，或者根据应用场景灵活调整，本申请实施例对此不加以限定。命中的样本细胞的数量不小于数量阈值的测量基因可认为是可靠性较高的测量基因，将可靠性较高的测量基因作为候选基因，有利于提高对数据处理模型的训练效果。

[0152] 样本细胞是对第一数据处理模型训练一次所依据的细胞，样本细胞的数量可以为一个，也可以为多个，本申请实施例对此不加以限定。本申请实施例以样本细胞的数量为一个为例进行说明，从样本基因表达数据中提取样本细胞下的各个候选基因对应的样本表达值的实现过程参见图2所示的实施例中的步骤201中的相关过程，此处不再赘述。

[0153] 在步骤402中，基于各个候选基因对应的样本表达值，获取各个候选基因对应的样本特征。

[0154] 该步骤402的实现过程参见图2所示的实施例中的步骤202，此处不再赘述。

[0155] 在步骤403中，调用第一数据处理模型基于各个候选基因对应的样本特征，提取各个候选基因在样本细胞下的相互作用信息；基于各个候选基因在样本细胞下的相互作用信息，获取样本细胞对应的预测结果。

[0156] 第一数据处理模型是指待利用样本细胞对应的样本基因表达数据以及样本细胞对应的标准结果进行训练的模型。

[0157] 在一种可能实现方式中，对于图2所示的实施例中提到的目标数据处理模型包括目标提取子模型和目标预测子模型的情况，第一数据处理模型包括第一提取子模型和第一预测子模型。此种情况下，该步骤403的实现过程包括：调用第一提取子模型对各个候选基因对应的样本特征进行信息提取，得到各个候选基因在样本细胞下的相互作用信息，调用第一预测子模型对各个候选基因在样本细胞下的相互作用信息进行处理，得到样本细胞对应的预测结果。该实现过程参见图2所示的实施例中的步骤203中的相关过程，此处不再赘述。示例性地，第一提取子模型的结构与目标提取子模型的结构相同，第一提取子模型为由至少一个基于注意力机制的编码器依次连接得到的语言模型。

[0158] 在示例性实施例中，第一数据处理模型中的第一提取子模型可以是指初始化的模型，也可以是预训练后得到的模型，本申请实施例对此不加以限定。类似地，第一数据处理模型中的第一预测子模型可以是指初始化的模型，也可以是预训练后得到的模型，本申请实施例对此不加以限定。

[0159] 本申请实施例以第一提取子模型是预训练后得到的模型，第一预测子模型是初始化的模型为例进行说明。在此种情况下，在调用第一提取子模型对各个候选基因对应的样本特征进行信息提取，得到各个候选基因在样本细胞下的相互作用信息之前，需要先训练得到第一提取子模型。

[0160] 在一种可能实现方式中，训练得到第一提取子模型的过程包括以下步骤a至步骤d。

[0161] 步骤a：获取训练细胞的训练基因表达数据，从训练基因表达数据中提取训练细胞下的各个候选基因对应的训练表达值。

[0162] 训练细胞是指无标准结果的细胞，在实际应用场景中，存在大规模的无标准结果的细胞的基因表达数据，利用大规模的无标准结果的细胞的基因表达数据对初始提取子模型进行训练，能够使初始提取子模型学习不通过数据分布的基因表达数据下的基因相互作用信息，提高提取子模型的泛化能力。在示例性实施例中，从数据库(如，PanglaoDB)中提取无标准结果的基因表达数据作为训练细胞的训练基因表达数据。从训练基因表达数据中提取训练细胞下的各个候选基因对应的训练表达值的实现过程参见图2所示的实施例中的步骤201中的相关过程，此处不再赘述。

[0163] 步骤b：基于各个候选基因对应的训练表达值，获取各个候选基因对应的训练特征，将各个候选基因中满足替换条件的候选基因对应的训练特征替换为参考特征。

[0164] 基于各个候选基因对应的训练表达值，获取各个候选基因对应的训练特征的实现过程参见图2所示的实施例中的步骤202，此处不再赘述。

[0165] 在获取各个候选基因对应的训练特征后，从各个候选基因中确定满足替换条件的候选基因。在示例性实施例中，满足替换条件的候选基因是指在各个候选基因中随机选取

的参考比例的候选基因，参考比例根据经验设置，或者根据经验场景灵活调整，例如，参考比例为15％。

[0166] 在示例性实施例中，由于训练基因表达数据的高度稀疏特征，满足替换条件的候选基因是指在对应的训练表达值不为0的候选基因中随机选取的参考比例的候选基因。例如，假设参考比例为10％，若训练细胞i下的候选基因j对应的训练表达值Aij不为零，则该候选基因j有15％的几率被确定为满足替换条件的候选基因。

[0167] 在确定出满足替换条件的候选基因后，将满足替换条件的候选基因对应的训练特征替换为参考特征。参考特征为特殊符号标识的表征特征，示例性地，特殊符号标识为[Mask]。特殊符号标识的表征特征可以根据经验设置，或者根据经验场景灵活调整，只要能够保证该特殊符号标识的表征特征与正常的训练特征能够区分即可。需要说明的是，参考特征的形式与训练特征的形式相同，例如，均为一个指定维度的向量。示例性地，将各个候选基因中满足替换条件的候选基因对应的训练特征替换为参考特征的过程可视为随机替换处理。

[0168] 步骤c：调用初始提取子模型对满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征进行信息提取，得到各个候选基因在训练细胞下的相互作用信息；基于各个候选基因在训练细胞下的相互作用信息，获取满足替换条件的候选基因对应的预测特征。

[0169] 在将满足替换条件的候选基因对应的训练特征替换为参考特征之后，将满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征输入初始提取子模型，初始提取子模型通过对满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征进行信息提取，得到各个候选基因在训练细胞下的相互作用信息。示例性地，将满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征输入初始提取子模型是指将满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征按照各个候选基因对应的指定排列顺序依次排列，构成的特征序列，将特征序列输入初始提取子模型。

[0170] 利用训练细胞的训练数据对初始提取子模型进行训练的过程可视为自监督预训练过程，在获取各个候选基因在训练细胞下的相互作用信息之后，还调用初始提取子模型基于各个候选基因在训练细胞下的相互作用信息，获取满足替换条件的候选基因对应的预测特征，以便于利用满足替换条件的候选基因对应的预测特征构建自监督预训练过程中利用的损失函数。

[0171] 在示例性实施例中，基于各个候选基因在训练细胞下的相互作用信息，获取满足替换条件的候选基因对应的预测特征的过程为：调用初始提取子模型基于各个候选基因在训练细胞下的相互作用信息，获取各个候选基因对应的预测特征；从各个候选基因对应的预测特征中提取满足替换条件的候选基因对应的预测特征。调用初始提取子模型基于各个候选基因在训练细胞下的相互作用信息，获取各个候选基因对应的预测特征的过程为初始提取子模型的内部处理过程，与初始提取子模型的模型结构有关。

[0172] 示例性地，初始提取子模型包括依次连接的至少一个基于注意力机制的编码器，每个基于注意力机制的编码器均包括一个注意力提取层和一个特征输出层，其中，注意力提取层用于根据输入的特征提取相互作用信息，特征输出层用于根据提取的相互作用信息

输出与输入的特征维度相同的特征。各个候选基因在训练细胞下的相互作用信息是最后一个基于注意力机制的编码器提取到的相互作用信息，此种情况下，通过将各个候选基因在训练细胞下的相互作用信息输入最后一个基于注意力机制的编码器中的特征输出层，即可得到特征输出层输出的各个候选基因对应的预测特征。

[0173] 步骤d：基于满足替换条件的候选基因对应的预测特征和训练特征，获取特征损失函数，利用特征损失函数对初始提取子模型进行训练，得到第一提取子模型。

[0174] 在获取满足替换条件的候选基因对应的预测特征后，基于满足替换条件的候选基因对应的预测特征和训练特征，获取用于对初始提取子模型进行训练的特征损失函数。在示例性实施例中，基于满足替换条件的候选基因对应的预测特征和训练特征获取特征损失函数的方式为：将基于满足替换条件的候选基因对应的预测特征和训练特征之间的交叉熵损失函数作为特征损失函数。当然，还可以通过其他方式获取特征损失函数，例如，将基于满足替换条件的候选基因对应的预测特征和训练特征之间的均方误差损失函数作为特征损失函数等。

[0175] 在获取特征损失函数之后，利用特征损失函数对初始提取子模型进行训练。在示例性实施例中，利用特征损失函数对初始提取子模型进行训练的过程是指利用特征损失函数更新初始提取子模型的模型参数的过程。

[0176] 在利用特征损失函数对初始提取子模型进行训练之后，得到一个训练后的提取子模型，判断该训练后的提取子模型是否满足第一训练终止条件，若该训练后的提取子模型满足第一训练终止条件，则将该训练后的提取子模型作为第一提取子模型。若该训练后的提取子模型不满足第一训练终止条件，则参考步骤a至步骤d的方式继续对该训练后得到的提取子模型进行训练，以此类推，直至得到满足第一训练终止条件的提取子模型，将该满足第一训练终止条件的提取子模型作为第一提取子模型。

[0177] 满足第一训练终止条件根据经验设置，或者根据应用场景灵活调整，本申请实施例对此不加以限定。示例性地，训练后得到的提取子模型满足第一训练终止条件包括但不限于获取该训练后得到的提取子模型时已执行的训练次数达到次数阈值、获取该训练后得到的提取子模型时的特征损失函数小于损失函数阈值或获取该训练后得到的提取子模型时的特征损失函数收敛中的任一项。

[0178] 示例性地，对提取子模型进行训练的目标是使提取子模型能够较为准确地预测出被参考特征替换的训练特征，以通过能够准确地预测出被参考特征替换的训练特征，来保证提取子模型具有较为可靠地提取相互作用信息的能力。

[0179] 在示例性实施例中，对初始提取子模型进行训练的过程如图5所示。获取各个候选基因对应的训练特征；将各个候选基因中满足替换条件的候选基因对应的训练特征替换为参考特征。将满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征输入初始提取子模型，获取初始提起子模型输出的满足替换条件的候选基因对应的预测特征；基于满足替换条件的候选基因对应的预测特征和训练特征，获取特征损失函数，利用特征损失函数对初始提取子模型进行训练。在图5中，初始提取子模型包括依次连接的8个Performer编码器。在图5中，候选基因对应的训练特征以及预测特征利用候选基因对应的离散化表达值(如，B2、B15、B7、Zero)表示，候选基因对应的参考特征利用特征符号标识[Mask]表示。

[0180] 在步骤404中，基于样本细胞对应的预测结果和标准结果，获取结果损失函数；利用结果损失函数对第一数据处理模型进行训练，得到目标数据处理模型。

[0181] 在获取样本细胞对应的预测结果后，基于样本细胞对应的预测结果和标准结果，获取结果损失函数。结果损失函数用于指示样本细胞对应的预测结果和标准结果之间的差异性。本申请实施例对基于样本细胞对应的预测结果和标准结果，获取结果损失函数的方式不加以限定，示例性地，将样本细胞对应的预测结果和标准结果之间的交叉熵损失函数或者均方误差损失函数作为结果损失函数。

[0182] 在获取结果损失函数后，利用结果损失函数对第一数据处理模型进行训练。利用结果损失函数对第一数据处理模型进行训练是指利用结果损失函数更新第一数据处理模型的参数。利用结果损失函数更新第一数据处理模型的参数可以是指利用结果损失函数更新第一数据处理模型的全部参数，也可以是指利用结果损失函数更新第一数据处理模型的部分参数，本申请实施例对此不加以限定。

[0183] 在示例性实施例中，对于第一数据处理模型中的第一提取子模型是利用训练细胞的训练基因表达数据预训练得到的模型、第一预测子模型为初始化的模型的情况，在利用结果损失函数更新第一数据处理模型的参数的过程中，可以保持第一提取子模型部分参数不变，更新第一提取子模型的其他参数以及第一预测子模型的全部参数。保持第一提取子模型的哪部分参数不变可以根据经验灵活设定，本申请实施例对此不加以限定。

[0184] 在利用结果损失函数对第一数据处理模型进行训练之后，得到一个训练后的数据处理模型，判断该训练后的数据处理模型是否满足第二训练终止条件，若该训练后的数据处理模型满足第二训练终止条件，则将该训练后的数据处理模型作为目标提取子模型。若该训练后的数据处理模型不满足第二训练终止条件，则参考步骤401至步骤404的方式继续对该训练后得到的数据处理模型进行训练，以此类推，直至得到满足第二训练终止条件的数据处理模型，将该满足第二训练终止条件的数据处理模型作为目标数据处理模型。

[0185] 满足第二训练终止条件根据经验设置，或者根据应用场景灵活调整，本申请实施例对此不加以限定。示例性地，训练后得到的数据处理模型满足第二训练终止条件包括但不限于获取该训练后得到的数据处理模型时已执行的训练次数达到次数阈值、获取该训练后得到的数据处理模型时的结果损失函数小于损失函数阈值或获取该训练后得到的数据处理模型时的结果损失函数收敛中的任一项。

[0186] 在示例性实施例中，数据处理模型的训练过程如图6所示。获取无标准结果的训练细胞的训练基因表达数据；基于训练基因表达数据，通过执行离散化处理、随机替换以及向量化转换，得到满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征，调用初始提取子模型对满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征进行处理，得到各个候选基因对应的预测特征；基于满足替换条件的候选基因对应的预测特征和训练特征，获取特征损失函数，利用特征损失函数对初始提取子模型进行训练，得到第一提取子模型。基于训练细胞的训练基因表达数据对初始提取子模型进行训练的过程可称为自监督预训练过程。

[0187] 获取有标准结果的样本细胞的样本基因表达数据，基于样本基因表达数据，通过执行离散化处理以及向量化处理，得到各个候选基因对应的样本特征；调用第一提取子模型对各个候选基因对应的样本特征进行处理，输出各个候选基因在样本细胞下的相互作用

信息;调用第一预测子模型中的卷积层对各个候选基因在样本细胞下的相互作用信息进行卷积,将卷积后得到的特征输入第一预测子模型中的全连接层,获取全连接层输出的样本细胞对应的预测结果;基于样本细胞对应的预测结果和标准结果,获取结果损失函数,利用结果损失函数对第一数据处理模型进行训练,得到目标数据处理模型。基于样本细胞的样本基因表达数据对第一数据处理模型进行训练的过程可称为有监督微调过程。

[0188] 在示例性实施例中,在获取目标数据处理模型后,可以对目标数据处理模型进行测试,以测试目标数据处理模型的模型性能。示例性地,对目标数据处理模型进行测试的过程包括:基于测试细胞的基因表达数据预,获取各个候选基因对应的测试特征;调用目标数据处理模型基于各个候选基因对应的测试特征,获取测试细胞对应的预测结果。目标数据处理模型的模型性能可以基于测试细胞对应的预测结果和测试细胞对应的标准结果计算得到。

[0189] 本申请实施例提出了一种基于注意力机制的数据处理模型,与传统的依赖人工获取细胞的预测结果的方法相比,利用本申请实施例提供的数据处理模型能够实现自动化获取细胞的预测结果的过程,大大节省了对细胞进行标注的成本,避免人为因素带来的实验误差。与其他自动化获取细胞的预测结果的方法相比,本申请实施例使用大规模无标签数据,并结合Gene2Vec的基因编码,能够有效学习不同数据分布中基因间复杂的相互作用,具有更强的泛化性能。通过有监督训练部分的卷积操作,能够有效提取关键信息,经过前馈神经网络得到更准确的细胞预测结果,进而有利于提高疾病机理研究的可靠性,推动精准医疗中疾病的早期诊断和个性化治疗。

[0190] 本申请实施例使用大规模的未标注数据结合基因相关关系进行自监督预训练,并将得到的模型使用已标注参考数据集进行微调式的有监督训练,最终可以精准的预测未标注细胞对应的结果。本申请实施例能够充分利用大规模数据进行基因间关系学习,同时结合了已有的基因关系嵌入,使得模型的关键信息获取能力大大提升。目前公开的大规模未标注单细胞转录组数据量呈指数级增长,且有各组织对应的单细胞图谱,为本申请实施例的实际应用提供了重要的先决条件。相比于相关技术,本申请实施例使用有标签的参考数据集,不需要额外提供细胞类型特异基因;大规模数据自监督预训练可以获取不同批次的数据分布,使模型具有足够的泛化性能;通过注意力机制,模型可以学习到具有生物意义的基因间相互作用信息。此方法可以将数据和知识结合,具有非常强的信息提取能力,对于单细胞转录组的分析的意义巨大。

[0191] 本申请实施例提供的数据处理模型的训练方法,基于各个候选基因在样本细胞下的相互作用信息获取样本细胞对应的预测结果,然后利用预测结果和标准结果之间的损失函数对第一数据处理模型进行训练,此种训练方式有利于提高数据处理模型提取相互作用信息的能力,以便于根据提取的相互作用信息输出更贴近标准结果的预测结果,从而保证训练得到的目标数据处理模型提取的相互作用信息的可靠性,使得目标数据处理模型能够根据提取的可靠的相互作用信息输出较为准确的预测结果。

[0192] 参见图7,本申请实施例提供了一种数据处理装置,该装置包括:

[0193] 第一获取单元701,用于获取目标细胞的目标基因表达数据,从目标基因表达数据中提取目标细胞下的各个候选基因对应的目标表达值;

[0194] 第二获取单元702,用于基于各个候选基因对应的目标表达值,获取各个候选基因

对应的目标特征；

[0195] 提取单元703，用于基于各个候选基因对应的目标特征，提取各个候选基因在目标细胞下的相互作用信息，各个候选基因在目标细胞下的相互作用信息用于表征目标细胞的基因相互作用特征；

[0196] 第三获取单元704，用于基于各个候选基因在目标细胞下的相互作用信息，获取目标细胞对应的预测结果。

[0197] 在一种可能实现方式中，第二获取单元702，用于将第一候选基因对应的目标表达值转换成第一候选基因对应的表达值特征，第一候选基因为各个候选基因中的任一候选基因；对第一候选基因对应的表达值特征和第一候选基因对应的表征特征进行融合，得到第一候选基因对应的目标特征。

[0198] 在一种可能实现方式中，第二获取单元702，用于对第一候选基因对应的目标表达值进行归一化处理，得到第一候选基因对应的归一化表达值；确定归一化表达值对应的目标离散化表达值，将目标离散化表达值对应的嵌入特征作为第一候选基因对应的表达值特征。

[0199] 在一种可能实现方式中，目标离散化表达值为参考数量个候选离散化表达值中的一个候选离散化表达值，第二获取单元702，还用于对参考数量个候选离散化表达值进行向量化转换，得到参考数量个候选离散化表达值分别对应的嵌入特征。

[0200] 在一种可能实现方式中，提取单元703，用于调用目标数据处理模型基于各个候选基因对应的目标特征，提取各个候选基因在目标细胞下的相互作用信息；

[0201] 第三获取单元704，用于调用目标数据处理模型基于各个候选基因在目标细胞下的相互作用信息，获取目标细胞对应的预测结果。

[0202] 在一种可能实现方式中，目标细胞对应的预测结果指示目标细胞的类别为目标类别，该装置还包括：

[0203] 确定单元，用于基于各个候选基因在目标细胞下的相互作用信息，在各个候选基因中确定属于目标类别的细胞对应的满足选取条件的基因。

[0204] 本申请实施例提供的数据处理装置，自动根据目标细胞的目标基因表达数据获取目标细胞对应的预测结果，无需依赖研究者的先验知识，数据处理的稳定性较高。此外，目标细胞对应的预测结果是基于各个候选基因在目标细胞下的相互作用信息获取的，各个候选基因在目标细胞下的相互作用信息能够表征出目标细胞的基因相互作用特征，由于细胞在生物体中是通过基因之间的相互作用发挥功能的，所以基因相互作用特征能够体现出目标细胞的功能方面的特征，通过关注目标细胞的功能方面的特征获取的预测结果的准确性较高。

[0205] 参见图8，本申请实施例提供了一种数据处理模型的训练装置，该装置包括：

[0206] 第一获取单元801，用于获取样本细胞的样本基因表达数据和样本细胞对应的标准结果，从样本基因表达数据中提取样本细胞下的各个候选基因对应的样本表达值；

[0207] 第二获取单元802，用于基于各个候选基因对应的样本表达值，获取各个候选基因对应的样本特征；

[0208] 提取单元803，用于调用第一数据处理模型基于各个候选基因对应的样本特征，提取各个候选基因在样本细胞下的相互作用信息；

[0209] 第三获取单元804，用于基于各个候选基因在样本细胞下的相互作用信息，获取样本细胞对应的预测结果；

[0210] 训练单元805，用于基于样本细胞对应的预测结果和标准结果，获取结果损失函数；利用结果损失函数对第一数据处理模型进行训练，得到目标数据处理模型。

[0211] 在一种可能实现方式中，第一数据处理模型包括第一提取子模型和第一预测子模型；提取单元803，用于调用第一提取子模型对各个候选基因对应的样本特征进行信息提取，得到各个候选基因在样本细胞下的相互作用信息；

[0212] 第三获取单元804，用于调用第一预测子模型对各个候选基因在样本细胞下的相互作用信息进行处理，得到样本细胞对应的预测结果。

[0213] 在一种可能实现方式中，第一获取单元801，还用于获取训练细胞的训练基因表达数据，从训练基因表达数据中提取训练细胞下的各个候选基因对应的训练表达值；

[0214] 第二获取单元802，还用于基于各个候选基因对应的训练表达值，获取各个候选基因对应的训练特征；

[0215] 该装置还包括：

[0216] 替换单元，用于将各个候选基因中满足替换条件的候选基因对应的训练特征替换为参考特征；

[0217] 提取单元803，还用于调用初始提取子模型对满足替换条件的候选基因对应的参考特征以及不满足替换条件的候选基因对应的训练特征进行信息提取，得到各个候选基因在训练细胞下的相互作用信息；基于各个候选基因在训练细胞下的相互作用信息，获取满足替换条件的候选基因对应的预测特征；

[0218] 训练单元805，还用于基于满足替换条件的候选基因对应的预测特征和训练特征，获取特征损失函数，利用特征损失函数对初始提取子模型进行训练，得到第一提取子模型。

[0219] 在一种可能实现方式中，样本细胞的数量为至少一个，一个样本细胞的样本基因表达数据包括一个样本细胞下的各个测量基因对应的样本表达值，该装置还包括：

[0220] 确定单元，用于基于各个样本细胞的样本基因表达数据，统计各个测量基因分别命中的样本细胞的数量，一个测量基因命中一个样本细胞用于指示一个样本细胞下的一个测量基因对应的样本表达值不小于第一阈值；将命中的样本细胞的数量不小于数量阈值的测量基因作为候选基因。

[0221] 在一种可能实现方式中，第一提取子模型为由至少一个基于注意力机制的编码器依次连接得到的语言模型。

[0222] 本申请实施例提供的数据处理模型的训练装置，基于各个候选基因在样本细胞下的相互作用信息获取样本细胞对应的预测结果，然后利用预测结果和标准结果之间的损失函数对第一数据处理模型进行训练，此种训练方式有利于提高数据处理模型提取相互作用信息的能力，以便于根据提取的相互作用信息输出更贴近标准结果的预测结果，从而保证训练得到的目标数据处理模型提取的相互作用信息的可靠性，使得目标数据处理模型能够根据提取的可靠的相互作用信息输出较为准确的预测结果。

[0223] 需要说明的是，上述实施例提供的装置在实现其功能时，仅以上述各功能单元的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元完成，即将设备的内部结构划分成不同的功能单元，以完成以上描述的全部或者部分功能。另外，

上述实施例提供的装置与方法实施例属于同一构思,其具体实现过程详见方法实施例,这里不再赘述。

[0224] 在示例性实施例中,还提供了一种计算机设备,该计算机设备包括处理器和存储器,该存储器中存储有至少一条计算机程序。该至少一条计算机程序由一个或者一个以上处理器加载并执行,以使该计算机设备实现上述任一种数据处理方法或数据处理模型的训练方法。该计算机设备可以为服务器,也可以为终端,本申请实施例对此不加以限定。接下来,对服务器和终端的结构分别进行介绍。

[0225] 图9是本申请实施例提供的一种服务器的结构示意图,该服务器可因配置或性能不同而产生比较大的差异,可以包括一个或多个处理器(Central Processing Units,CPU)901和一个或多个存储器902,其中,该一个或多个存储器902中存储有至少一条计算机程序,该至少一条计算机程序由该一个或多个处理器901加载并执行,以使该服务器实现上述各个方法实施例提供的数据处理方法或数据处理模型的训练方法。当然,该服务器还可以具有有线或无线网络接口、键盘以及输入输出接口等部件,以便进行输入输出,该服务器还可以包括其他用于实现设备功能的部件,在此不做赘述。

[0226] 图10是本申请实施例提供的一种终端的结构示意图。该终端例如可以是:PC、手机、智能手机、PDA、可穿戴设备、PPC、平板电脑、智能车机、智能电视、智能音箱、智能语音交互设备、智能家电、车载终端。终端还可能被称为用户设备、便携式终端、膝上型终端、台式终端等其他名称。

[0227] 通常,终端包括有:处理器1001和存储器1002。

[0228] 处理器1001可以包括一个或多个处理核心,比如4核心处理器、8核心处理器等。处理器1001可以采用DSP(Digital Signal Processing,数字信号处理)、FPGA(Field-Programmable Gate Array,现场可编程门阵列)、PLA(Programmable Logic Array,可编程逻辑阵列)中的至少一种硬件形式来实现。处理器1001也可以包括主处理器和协处理器,主处理器是用于对在唤醒状态下的数据进行处理的处理器,也称CPU(Central Processing Unit,中央处理器);协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中,处理器1001可以集成有GPU(Graphics Processing Unit,图像处理器),GPU用于负责显示屏所需要显示的内容的渲染和绘制。一些实施例中,处理器1001还可以包括AI(Artificial Intelligence,人工智能)处理器,该AI处理器用于处理有关机器学习的计算操作。

[0229] 存储器1002可以包括一个或多个计算机可读存储介质,该计算机可读存储介质可以是非暂态的。存储器1002还可包括高速随机存取存储器,以及非易失性存储器,比如一个或多个磁盘存储设备、闪存存储设备。在一些实施例中,存储器1002中的非暂态的计算机可读存储介质用于存储至少一个指令,该至少一个指令用于被处理器1001所执行,以使该终端实现本申请中方法实施例提供的数据处理方法或数据处理模型的训练方法。

[0230] 在一些实施例中,终端还可选包括有:外围设备接口1003和至少一个外围设备。处理器1001、存储器1002和外围设备接口1003之间可以通过总线或信号线相连。各个外围设备可以通过总线、信号线或电路板与外围设备接口1003相连。具体地,外围设备包括:射频电路1004、显示屏1005、摄像头组件1006、音频电路1007、定位组件1008和电源1009中的至少一种。

[0231] 外围设备接口1003可被用于将I/O(Input/Output,输入/输出)相关的至少一个外围设备连接到处理器1001和存储器1002。在一些实施例中,处理器1001、存储器1002和外围设备接口1003被集成在同一芯片或电路板上;在一些其他实施例中,处理器1001、存储器1002和外围设备接口1003中的任意一个或两个可以在单独的芯片或电路板上实现,本实施例对此不加以限定。

[0232] 射频电路1004用于接收和发射RF(Radio Frequency,射频)信号,也称电磁信号。射频电路1004通过电磁信号与通信网络以及其他通信设备进行通信。射频电路1004将电信号转换为电磁信号进行发送,或者,将接收到的电磁信号转换为电信号。可选地,射频电路1004包括:天线系统、RF收发器、一个或多个放大器、调谐器、振荡器、数字信号处理器、编解码芯片组、用户身份模块卡等等。射频电路1004可以通过至少一种无线通信协议来与其它终端进行通信。该无线通信协议包括但不限于:城域网、各代移动通信网络(2G、3G、4G及5G)、无线局域网和/或WiFi(Wireless Fidelity,无线保真)网络。在一些实施例中,射频电路1004还可以包括NFC(Near Field Communication,近距离无线通信)有关的电路,本申请对此不加以限定。

[0233] 显示屏1005用于显示UI(User Interface,用户界面)。该UI可以包括图形、文本、图标、视频及其它们的任意组合。当显示屏1005是触摸显示屏时,显示屏1005还具有采集在显示屏1005的表面或表面上方的触摸信号的能力。该触摸信号可以作为控制信号输入至处理器1001进行处理。此时,显示屏1005还可以用于提供虚拟按钮和/或虚拟键盘,也称软按钮和/或软键盘。在一些实施例中,显示屏1005可以为一个,设置在终端的前面板;在另一些实施例中,显示屏1005可以为至少两个,分别设置在终端的不同表面或呈折叠设计;在另一些实施例中,显示屏1005可以是柔性显示屏,设置在终端的弯曲表面上或折叠面上。甚至,显示屏1005还可以设置成非矩形的不规则图形,也即异形屏。显示屏1005可以采用LCD(Liquid Crystal Display,液晶显示屏)、OLED(Organic Light-Emitting Diode,有机发光二极管)等材质制备。

[0234] 摄像头组件1006用于采集图像或视频。可选地,摄像头组件1006包括前置摄像头和后置摄像头。通常,前置摄像头设置在终端的前面板,后置摄像头设置在终端的背面。在一些实施例中,后置摄像头为至少两个,分别为主摄像头、景深摄像头、广角摄像头、长焦摄像头中的任意一种,以实现主摄像头和景深摄像头融合实现背景虚化功能、主摄像头和广角摄像头融合实现全景拍摄以及VR(Virtual Reality,虚拟现实)拍摄功能或者其它融合拍摄功能。在一些实施例中,摄像头组件1006还可以包括闪光灯。闪光灯可以是单色温闪光灯,也可以是双色温闪光灯。双色温闪光灯是指暖光闪光灯和冷光闪光灯的组合,可以用于不同色温下的光线补偿。

[0235] 音频电路1007可以包括麦克风和扬声器。麦克风用于采集用户及环境的声波,并将声波转换为电信号输入至处理器1001进行处理,或者输入至射频电路1004以实现语音通信。出于立体声采集或降噪的目的,麦克风可以为多个,分别设置在终端的不同部位。麦克风还可以是阵列麦克风或全向采集型麦克风。扬声器则用于将来自处理器1001或射频电路1004的电信号转换为声波。扬声器可以是传统的薄膜扬声器,也可以是压电陶瓷扬声器。当扬声器是压电陶瓷扬声器时,不仅可以将电信号转换为人类可听见的声波,也可以将电信号转换为人类听不见的声波以进行测距等用途。在一些实施例中,音频电路1007还可以包

括耳机插孔。

[0236] 定位组件1008用于定位终端的当前地理位置，以实现导航或LBS(Location Based Service，基于位置的服务)。定位组件1008可以是基于美国的GPS(Global Positioning System，全球定位系统)、中国的北斗系统、俄罗斯的格雷纳斯系统或欧盟的伽利略系统的定位组件。

[0237] 电源1009用于为终端中的各个组件进行供电。电源1009可以是交流电、直流电、一次性电池或可充电电池。当电源1009包括可充电电池时，该可充电电池可以支持有线充电或无线充电。该可充电电池还可以用于支持快充技术。

[0238] 在一些实施例中，终端还包括有一个或多个传感器1010。该一个或多个传感器1010包括但不限于：加速度传感器1011、陀螺仪传感器1012、压力传感器1013、指纹传感器1014、光学传感器1015以及接近传感器1016。

[0239] 加速度传感器1011可以检测以终端建立的坐标系的三个坐标轴上的加速度大小。比如，加速度传感器1011可以用于检测重力加速度在三个坐标轴上的分量。处理器1001可以根据加速度传感器1011采集的重力加速度信号，控制显示屏1005以横向视图或纵向视图进行用户界面的显示。加速度传感器1011还可以用于游戏或者用户的运动数据的采集。

[0240] 陀螺仪传感器1012可以检测终端的机体方向及转动角度，陀螺仪传感器1012可以与加速度传感器1011协同采集用户对终端的3D动作。处理器1001根据陀螺仪传感器1012采集的数据，可以实现如下功能：动作感应(比如根据用户的倾斜操作来改变UI)、拍摄时的图像稳定、游戏控制以及惯性导航。

[0241] 压力传感器1013可以设置在终端的侧边框和/或显示屏1005的下层。当压力传感器1013设置在终端的侧边框时，可以检测用户对终端的握持信号，由处理器1001根据压力传感器1013采集的握持信号进行左右手识别或快捷操作。当压力传感器1013设置在显示屏1005的下层时，由处理器1001根据用户对显示屏1005的压力操作，实现对UI界面上的可操作性控件进行控制。可操作性控件包括按钮控件、滚动条控件、图标控件、菜单控件中的至少一种。

[0242] 指纹传感器1014用于采集用户的指纹，由处理器1001根据指纹传感器1014采集到的指纹识别用户的身份，或者，由指纹传感器1014根据采集到的指纹识别用户的身份。在识别出用户的身份为可信身份时，由处理器1001授权该用户执行相关的敏感操作，该敏感操作包括解锁屏幕、查看加密信息、下载软件、支付及更改设置等。指纹传感器1014可以被设置在终端的正面、背面或侧面。当终端上设置有物理按键或厂商Logo(商标)时，指纹传感器1014可以与物理按键或厂商Logo集成在一起。

[0243] 光学传感器1015用于采集环境光强度。在一个实施例中，处理器1001可以根据光学传感器1015采集的环境光强度，控制显示屏1005的显示亮度。具体地，当环境光强度较高时，调高显示屏1005的显示亮度；当环境光强度较低时，调低显示屏1005的显示亮度。在另一个实施例中，处理器1001还可以根据光学传感器1015采集的环境光强度，动态调整摄像头组件1006的拍摄参数。

[0244] 接近传感器1016，也称距离传感器，通常设置在终端的前面板。接近传感器1016用于采集用户与终端的正面之间的距离。在一个实施例中，当接近传感器1016检测到用户与终端的正面之间的距离逐渐变小时，由处理器1001控制显示屏1005从亮屏状态切换为息屏

状态;当接近传感器1016检测到用户与终端的正面之间的距离逐渐变大时,由处理器1001控制显示屏1005从息屏状态切换为亮屏状态。

[0245] 本领域技术人员可以理解,图10中示出的结构并不构成对终端的限定,可以包括比图示更多或更少的组件,或者组合某些组件,或者采用不同的组件布置。

[0246] 在示例性实施例中,还提供了一种计算机可读存储介质,该计算机可读存储介质中存储有至少一条计算机程序,该至少一条计算机程序由计算机设备的处理器加载并执行,以使计算机实现上述任一种数据处理方法或数据处理模型的训练方法。

[0247] 在一种可能实现方式中,上述计算机可读存储介质可以是只读存储器(Read-Only Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、只读光盘(Compact Disc Read-Only Memory,CD-ROM)、磁带、软盘和光数据存储设备等。

[0248] 在示例性实施例中,还提供了一种计算机程序产品,该计算机程序产品包括计算机程序或计算机指令,该计算机程序或计算机指令由处理器加载并执行,以使计算机实现上述任一种数据处理方法或数据处理模型的训练方法。

[0249] 需要说明的是,本申请中的术语“第一”、“第二”等是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。以上示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反,它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

[0250] 应当理解的是,在本文中提及的“多个”是指两个或两个以上。“和/或”,描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。

[0251] 以上所述仅为本申请的示例性实施例,并不用以限制本申请,凡在本申请的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本申请的保护范围之内。

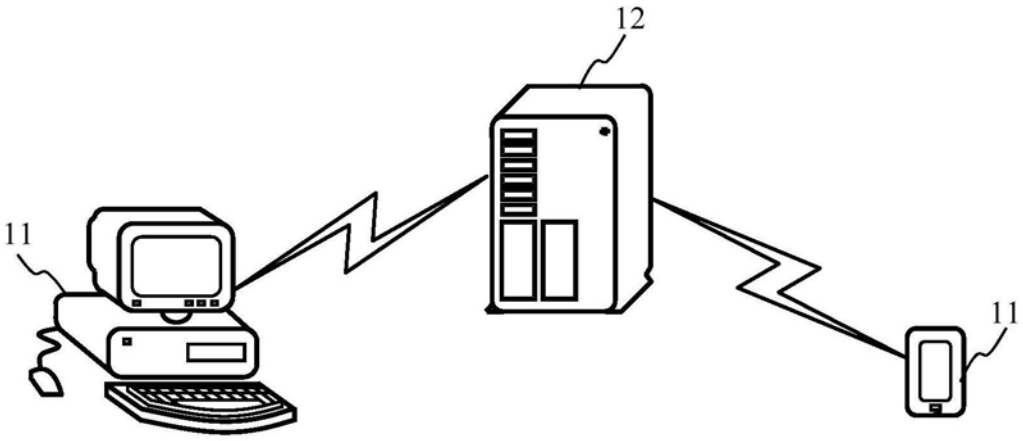

12
11
11

图1

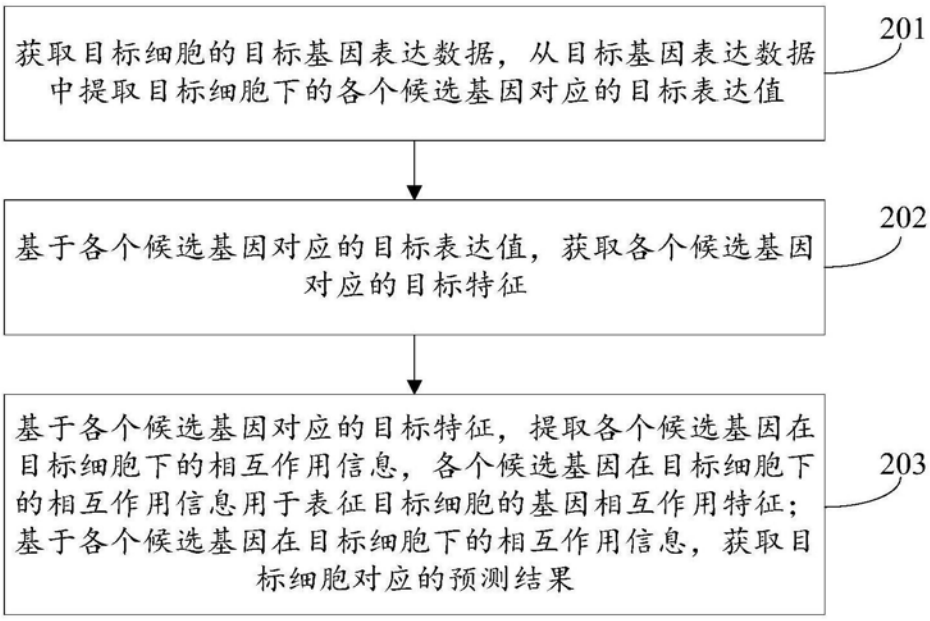

获取目标细胞的目标基因表达数据，从目标基因表达数据中提取目标细胞下的各个候选基因对应的目标表达值
201
基于各个候选基因对应的目标表达值，获取各个候选基因对应的目标特征
202
基于各个候选基因对应的目标特征，提取各个候选基因在目标细胞下的相互作用信息，各个候选基因在目标细胞下的相互作用信息用于表征目标细胞的基因相互作用特征；基于各个候选基因在目标细胞下的相互作用信息，获取目标细胞对应的预测结果
203

图2

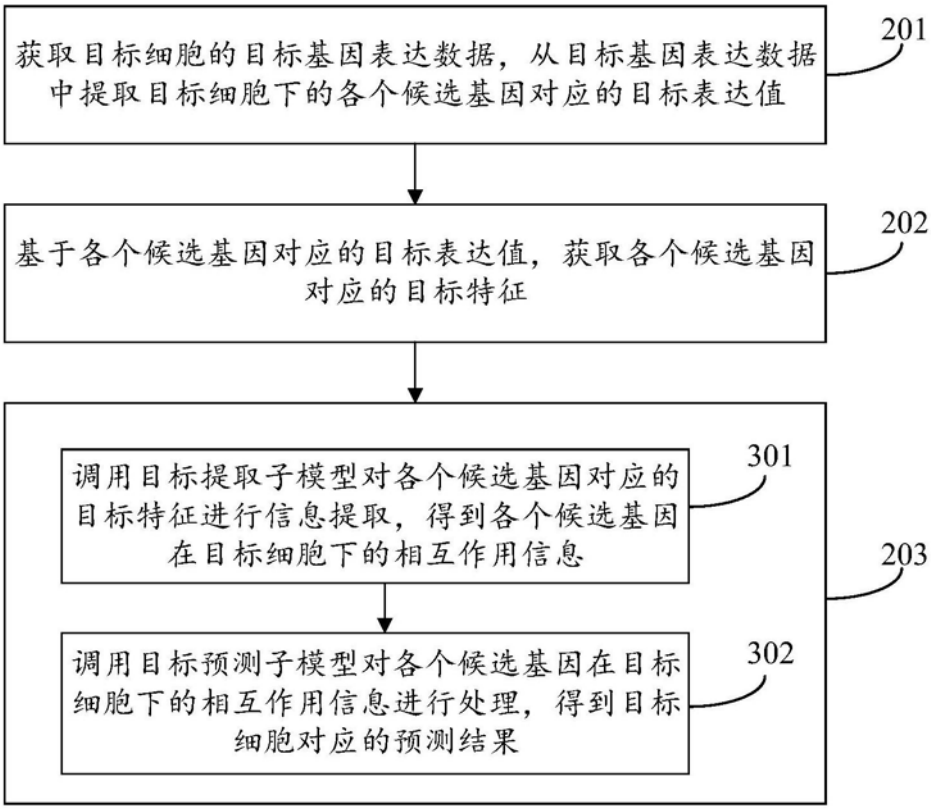

获取目标细胞的目标基因表达数据，从目标基因表达数据中提取目标细胞下的各个候选基因对应的目标表达值
201
基于各个候选基因对应的目标表达值，获取各个候选基因对应的目标特征
202
调用目标提取子模型对各个候选基因对应的目标特征进行信息提取，得到各个候选基因在目标细胞下的相互作用信息
301
调用目标预测子模型对各个候选基因在目标细胞下的相互作用信息进行处理，得到目标细胞对应的预测结果
302
203

图3

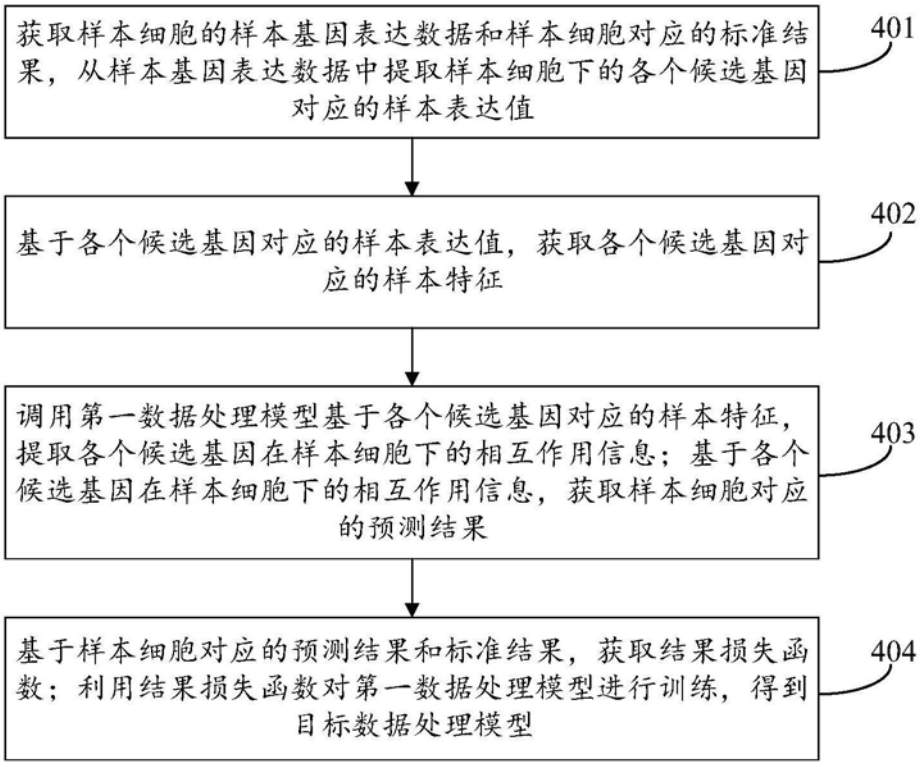

获取样本细胞的样本基因表达数据和样本细胞对应的标准结果，从样本基因表达数据中提取样本细胞下的各个候选基因对应的样本表达值
401
基于各个候选基因对应的样本表达值，获取各个候选基因对应的样本特征
402
调用第一数据处理模型基于各个候选基因对应的样本特征，提取各个候选基因在样本细胞下的相互作用信息；基于各个候选基因在样本细胞下的相互作用信息，获取样本细胞对应的预测结果
403
基于样本细胞对应的预测结果和标准结果，获取结果损失函数；利用结果损失函数对第一数据处理模型进行训练，得到目标数据处理模型
404

图4

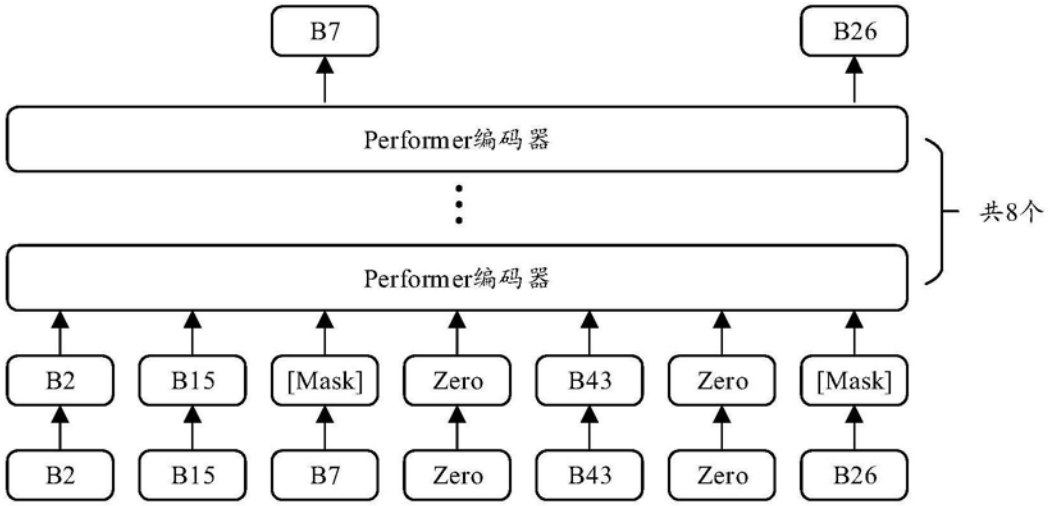

B7
B26
Performer编码器
Performer编码器
共8个
B2
B15
[Mask]
Zero
B43
Zero
[Mask]
B2
B15
B7
Zero
B43
Zero
B26

图5

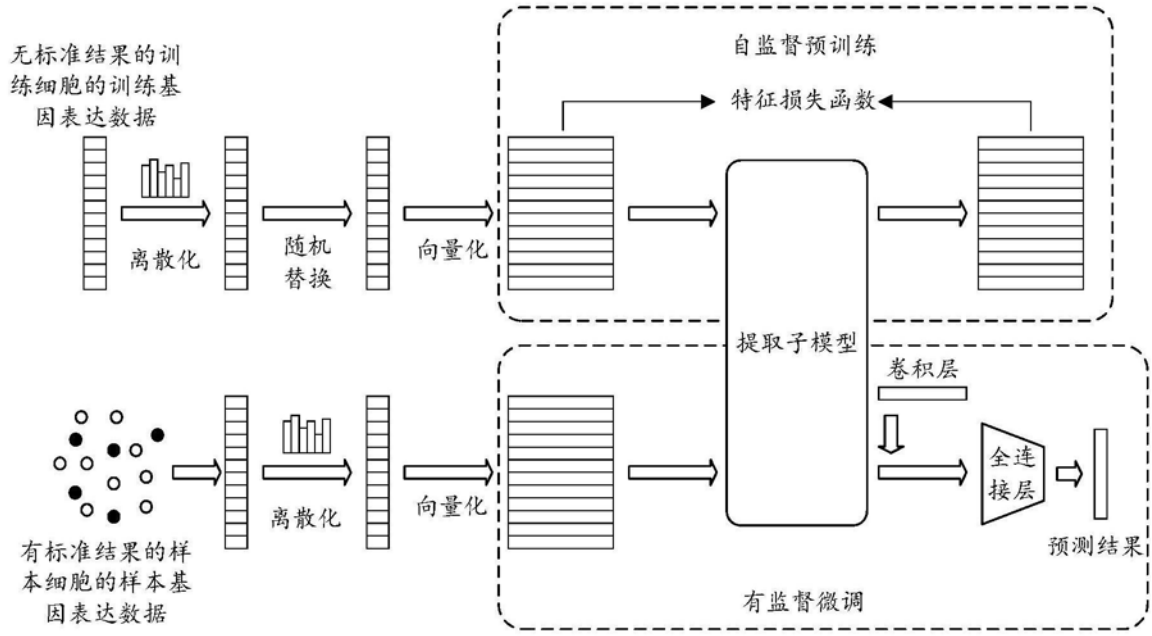

无标准结果的训练细胞的训练基因表达数据
离散化
随机替换
向量化
自监督预训练
特征损失函数
提取子模型
卷积层
全连接层
预测结果
有标准结果的样本细胞的样本基因表达数据
离散化
向量化
有监督微调

图6

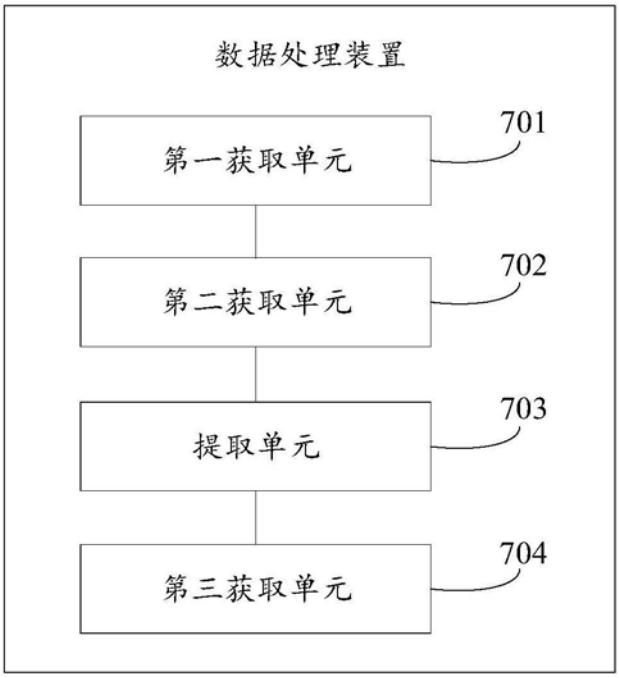

数据处理装置
第一获取单元
701
第二获取单元
702
提取单元
703
第三获取单元
704

图7

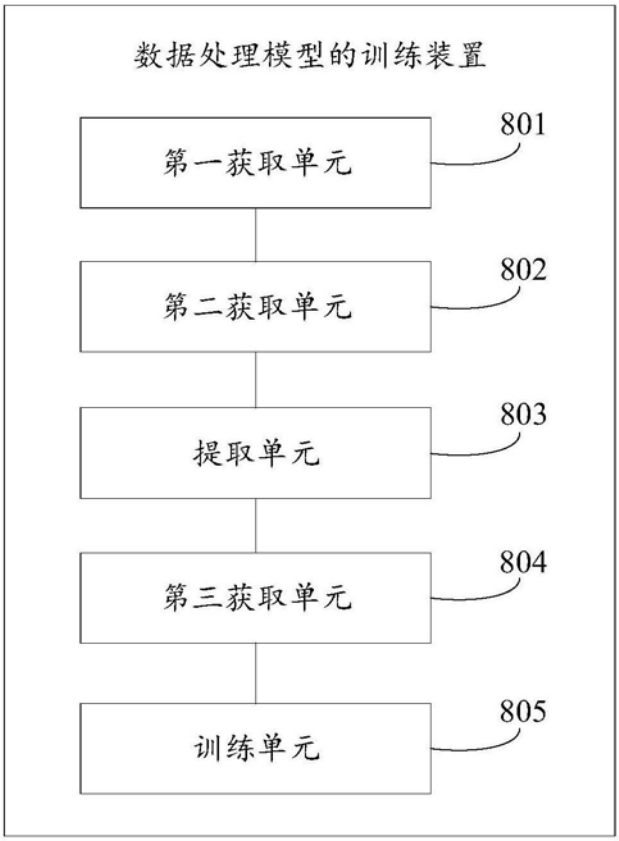

数据处理模型的训练装置
第一获取单元
801
第二获取单元
802
提取单元
803
第三获取单元
804
训练单元
805

图8

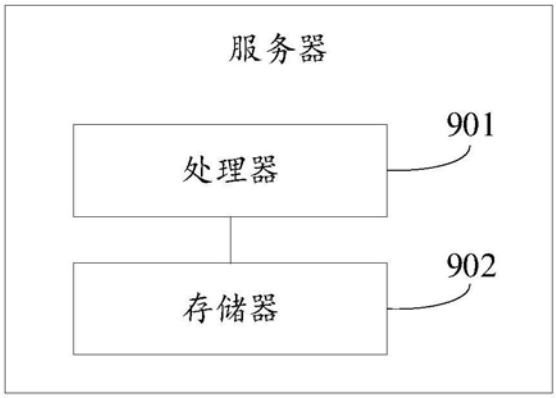

服务器
处理器
901
存储器
902

图9

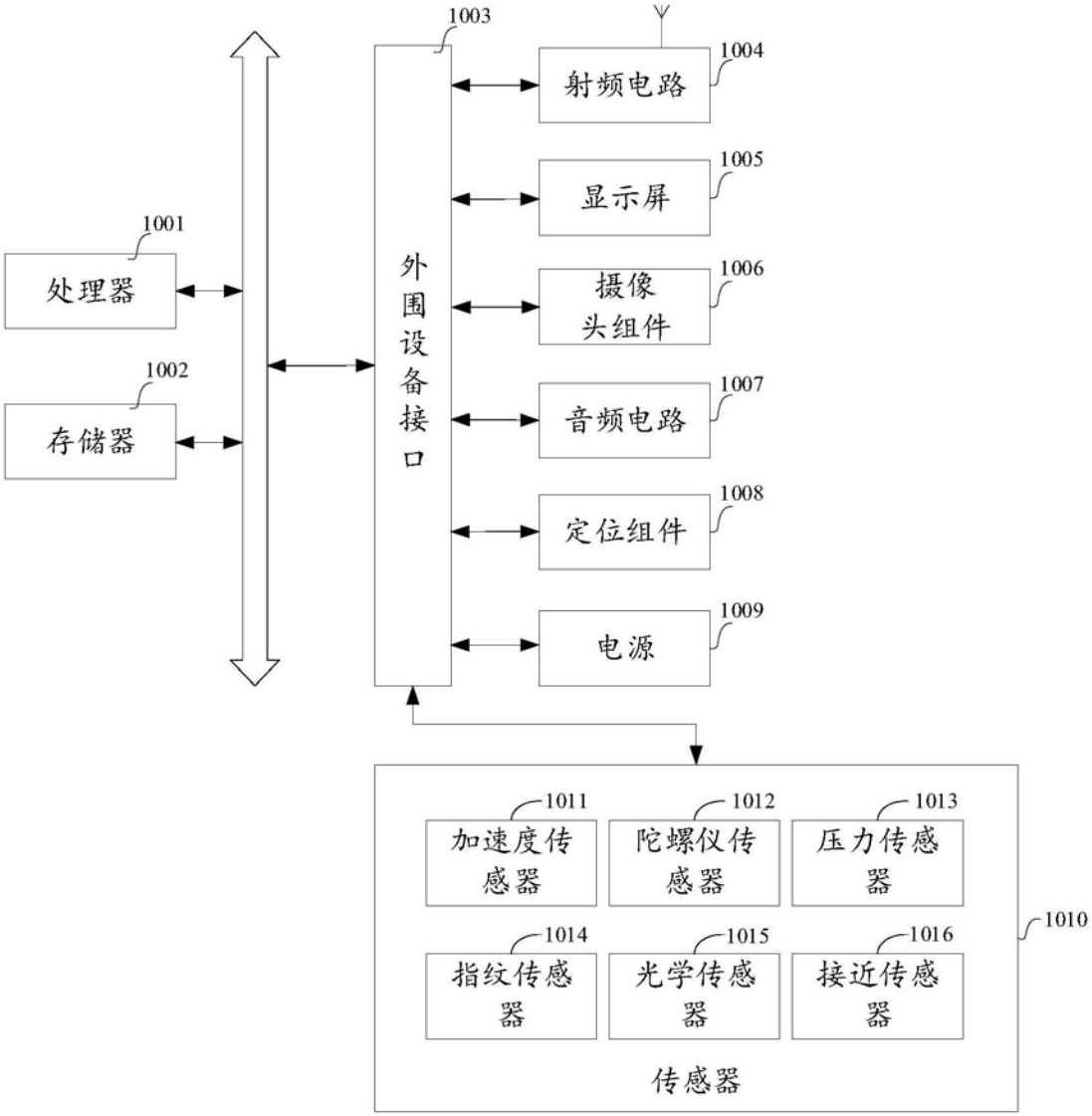

1003
射频电路
1004
显示屏
1005
摄像头组件
1006
音频电路
1007
定位组件
1008
电源
1009
外围设备接口
1001
处理器
1002
存储器
1011
加速度传感器
1012
陀螺仪传感器
1013
压力传感器
1014
指纹传感器
1015
光学传感器
1016
接近传感器
传感器
1010

图10