



República Federativa do Brasil
Ministério da Economia
Instituto Nacional da Propriedade Industrial

(21) BR 112020023429-0 A2



(22) Data do Depósito: 08/05/2019

(43) Data da Publicação Nacional: 23/02/2021

(54) Título: MÉTODOS E APARELHOS PARA A PREVISÃO MULTIMODAL USANDO UM MODELO ESTATÍSTICO TREINADO

(51) Int. Cl.: G16C 20/30; G16C 20/70; G16B 20/00; G16B 40/20; G16B 5/00.

(30) Prioridade Unionista: 30/05/2018 US 62/678,083; 30/05/2018 US 62/678,094.

(71) Depositante(es): QUANTUM-SI INCORPORATED.

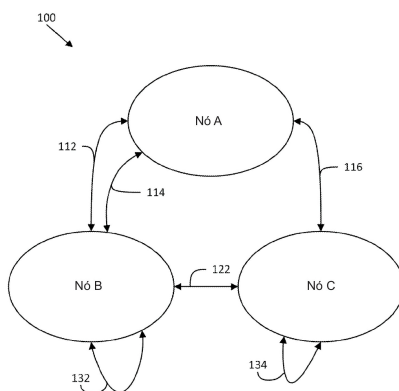
(72) Inventor(es): JONATHAN M. ROTHBERG; HENRI LICHENSTEIN; UMUT ESER; MICHAEL MEYER; MARYLENS HERNANDEZ; TIAN XU.

(86) Pedido PCT: PCT US2019031260 de 08/05/2019

(87) Publicação PCT: WO 2019/231624 de 05/12/2019

(85) Data da Fase Nacional: 17/11/2020

(57) Resumo: A presente invenção refere-se a métodos e aparelhos para prever uma associação entre dados de entrada em uma primeira modalidade e dados em uma segunda modalidade usando um modelo estatístico treinado para representar interações entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplando a pluralidade de codificadores e decodificadores. O método compreende selecionar, com base na primeira modalidade e na segunda modalidade, um par de codificador/decodificador ou um par de codificadores, a partir de entre a pluralidade de codificadores e decodificadores, e processar os dados de entrada com a representação de modalidade conjunta e o par de codificador/decodificador selecionado ou o par de codificadores para prever a associação entre os dados de entrada e os dados na segunda modalidade.



Relatório Descritivo da Patente de Invenção para
**"MÉTODOS E APARELHOS PARA A PREVISÃO MULTIMODAL
USANDO UM MODELO ESTATÍSTICO TREINADO".**

Referência Cruzada aos Pedidos Relacionados

[001] Este pedido reivindica prioridade sob 35 U.S.C. § 119 (e) do Pedido de Patente Provisório dos EUA 62/678.083, depositado em 30 de maio de 2018, e intitulado "METHODS AND APPARATUS FOR MULTI-MODAL PREDICTION USING A TRAINED STATISTICAL MODEL" e do Pedido de Patente Provisório dos EUA 62/678.094, depositado 30 de maio de 2018 e intitulado "METHODS AND APPARATUS FOR MAKING BIOLOGICAL PREDICTIONS USING A TRAINED MULTIMODAL STATISTICAL MODEL", todo o conteúdo de cada um dos quais é incorporado por referência neste documento.

Antecedentes

[002] A capacidade de reaproveitar medicamentos seguros oferece grandes vantagens para a indústria farmacêutica, incluindo economia de tempo e custo e maior taxa de sucesso na aprovação de medicamentos. A implementação de algoritmos computacionais com o objetivo de prever novas indicações de doenças para fármacos existentes ou novos tratamentos para doenças existentes surgiram recentemente com as melhorias na infraestrutura de computadores e o advento de tecnologias de alto rendimento que permitem a caracterização de doenças e fármacos em alta resolução.

[003] Algumas técnicas convencionais para descobrir novas indicações de doenças para fármacos existentes ou com o objetivo de encontrar a melhor combinação de fármacos para uma determinada doença ou paciente contam com a caracterização genômica de doenças e a caracterização molecular do mecanismo de ação do fármaco para fazer novas previsões. Essas técnicas podem ser classificadas como baseadas em fármacos ou baseadas em doenças e, embora ambas

apresentem vantagens e desafios únicos, uma abordagem computacional bem-sucedida geralmente combina aspectos de ambas as técnicas.

[004] As técnicas baseadas em fármacos drogas normalmente se concentram em semelhanças de estrutura de fármacos, similaridade de atividade molecular de fármacos ou similaridade de trajeto alvo e acoplamento molecular. Eles usam diferentes informações ou modalidades de dados, como estruturas de fármacos, alvos de fármacos, classe de fármacos e perturbação da expressão gênica no tratamento com fármacos. As técnicas com base em doenças em geral se concentram na transferência de indicação associativa, patologia molecular compartilhada ou semelhanças de efeitos colaterais. Eles incluem informações ou modalidades de dados relacionadas a mutações e vias associadas a doenças e mudanças associadas a doenças na expressão gênica, ou proteínas, ou metabólitos, ou microbioma. Exemplos de abordagens que combinam fundamentos baseados em fármacos e doenças incluem: complementaridade de assinatura de transcrição e similaridade de trajeto de doença-alvo de fármaco.

Sumário

[005] De acordo com um aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um método para treinar um modelo estatístico configurado para representar associações entre modalidades entre dados em uma rede heterogênea. O método compreende acessar dados de treinamento incluindo dados de treinamento para uma primeira modalidade e dados de treinamento para uma segunda modalidade diferente a partir da primeira modalidade, treinar o modelo estatístico, o modelo estatístico compreendendo primeiro e segundo codificadores, primeiro e segundo decodificadores, e uma representação de modalidade conjunta acoplando os primeiro e

segundo codificadores aos primeiro e segundo decodificadores. O treinamento compreende estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores usando uma técnica de aprendizado autossupervisionada, pelo menos algum de dados de treinamento, e informação descrevendo pelo menos um link entre pares de dados nos dados de treinamento, e armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores do modelo estatístico.

[006] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um método para prever uma associação entre dados de entrada em uma primeira modalidade e dados em uma segunda modalidade usando um modelo estatístico treinado para representar links entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplado a pluralidade de codificadores e decodificadores. O método compreende selecionar, com base na primeira modalidade e na segunda modalidade, um par de codificador/decodificador ou um par de codificadores, a partir de entre uma pluralidade de codificadores e decodificadores, e processar os dados de entrada com a representação de modalidade conjunta e o par de codificador/decodificador selecionado ou o par de codificadores para prever a associação entre os dados de entrada e os dados na segunda modalidade.

[007] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um método para

prever associações entre dados em uma primeira modalidade e dados em uma segunda modalidade usando um modelo estatístico treinado para representar interações entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplando a pluralidade de codificadores e decodificadores. O método compreende mapear os dados na primeira modalidade e os dados na segunda modalidade dentro de um espaço de representação comum dentro do modelo estatístico, acessar um classificador estatístico treinado usando dados rotulados, em que os dados rotulados descrevem associações entre dados nas primeira e segunda modalidades, e prever associações entre os dados na primeira modalidade e os dados na segunda modalidade mapeados dentro do espaço de representação comum usando o classificador estatístico treinado.

[008] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um sistema de computador, que compreende pelo menos um computador processador e pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador realizam um método de treinar um modelo estatístico para representar associações entre modalidades para os dados, em que os dados incluem dados para uma primeira modalidade e dados para uma segunda modalidade diferente a partir da primeira modalidade. O método compreende acessar dados de treinamento incluindo dados de treinamento para a primeira modalidade e dados de treinamento para a segunda modalidade, treinar o modelo estatístico, o modelo estatístico compreendendo primeiro e segundo

codificadores, primeiro e segundo decodificadores, e uma representação de modalidade conjunta acoplando os primeiro e segundo codificadores aos primeiro e segundo decodificadores. O treinamento compreende estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores usando uma técnica de aprendizado autossupervisionada, pelo menos alguns dos dados de treinamento, e informação descrevendo pelo menos um link entre pares de dados nos dados de treinamento, e armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores do modelo estatístico.

[009] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um sistema de computador compreendendo pelo menos um computador processador e pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador realizam um método de prever uma associação entre dados de entrada em uma primeira modalidade e dados em uma segunda modalidade usando um modelo estatístico treinado para representar interações entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplando a pluralidade de codificadores e decodificadores. O método compreende selecionar, com base na primeira modalidade e na segunda modalidade, um par de codificador/decodificador ou um par de codificadores, a partir de entre a pluralidade de codificadores e decodificadores, e processar os dados de

entrada com a representação de modalidade conjunta e o par de codificador/decodificador selecionado ou o par de codificadores para prever a associação entre os dados de entrada e os dados na segunda modalidade.

[0010] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um sistema de computador compreendendo pelo menos um computador processador e pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador, realizam um método de prever associações entre dados em uma primeira modalidade e dados em uma segunda modalidade usando um modelo estatístico treinado para representar links entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade diferente a partir da primeira modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplando a pluralidade de codificadores e decodificadores. O método compreende mapear os dados na primeira modalidade e os dados na segunda modalidade dentro de um espaço de representação comum dentro do modelo estatístico, acessar um classificador estatístico treinado usando dados rotulados, em que os dados rotulados descrevem associações entre dados nas primeira e segunda modalidades, e prever associações entre os dados na primeira modalidade e os dados na segunda modalidade mapeados dentro do espaço de representação comum usando o classificador estatístico treinado.

[0011] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um método para

treinar um modelo estatístico para representar associações entre dados de fármaco, dados de gene, e dados de doença. O método compreende acessar dados de treinamento incluindo dados de treinamento de gene, dados de treinamento de fármaco e dados de treinamento de doença, e treinar o modelo estatístico, o modelo estatístico compreendendo uma pluralidade de codificadores incluindo um codificador de gene, um codificador de fármaco e um codificador de doença, uma pluralidade de decodificadores incluindo um decodificador de gene, um decodificador de fármaco e um decodificador de doença, e uma representação conjunta acoplando a pluralidade de codificadores a uma pluralidade de decodificadores, em que a representação conjunta descreve interações entre os dados de treinamento. O treinamento compreende estimar valores para parâmetros do codificador de gene e do decodificador de gene usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene, e informação descrevendo interações entre pares de dados nos dados de treinamento de gene, estimar valores para parâmetros do codificador de gene, o decodificador de gene, o codificador de fármaco, e o decodificador de fármaco usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de fármaco, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de fármaco, estimar valores para parâmetros do codificador de gene, o decodificador de gene, o codificador de doença, e o decodificador de doença usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de doença, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de doença, e armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar

os valores estimados para parâmetros do codificador de gene, o decodificador de gene, o codificador de fármaco, o decodificador de fármaco, o codificador de doença, e o decodificador de doença do modelo estatístico.

[0012] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um sistema de computador, compreendendo pelo menos um computador processador e pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador realizam um método de treinar um modelo estatístico para representar associações entre dados de fármaco, dados de gene, e dados de doença. O método compreende acessar dados de treinamento incluindo dados de treinamento de gene, dados de treinamento de fármaco e dados de treinamento de doença, e treinar o modelo estatístico, o modelo estatístico compreendendo uma pluralidade de codificadores incluindo um codificador de gene, um codificador de fármaco e um codificador de doença, uma pluralidade de decodificadores incluindo um decodificador de gene, um decodificador de fármaco, e um decodificador de doença, e uma representação conjunta acoplando a pluralidade de codificadores a uma pluralidade de decodificadores, em que a representação conjunta descreve interações entre os dados de treinamento. O treinamento compreende estimar valores para parâmetros do codificador de gene e do decodificador de gene usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene, e informação descrevendo interações entre pares de dados nos dados de treinamento de gene, estimar valores para parâmetros do codificador de gene, o decodificador de gene, o codificador de fármaco, e o decodificador de fármaco usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de fármaco, e

informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de fármaco, e estimar valores para parâmetros do codificador de gene, o decodificador de gene, o codificador de doença, e o decodificador de doença usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de doença, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de doença, e armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de gene, o decodificador de gene, o codificador de fármaco, o decodificador de fármaco, o codificador de doença, e o decodificador de doença do modelo estatístico.

[0013] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um método para prever uma nova indicação de doença para um determinado fármaco. O método compreende projetar uma representação do determinado fármaco e representações de uma pluralidade de doenças dentro de um espaço de representação comum de um modelo estatístico treinado e prever a nova indicação de doença para o determinado fármaco com base em uma comparação da representação projetada do determinado fármaco e pelo menos uma das representações da pluralidade de doenças no espaço de representação comum.

[0014] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um sistema de computador, compreendendo pelo menos um computador processador; e pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador, realiza um método de prever a

nova indicação de doença para um determinado fármaco. O método compreende projetar a representação do determinado fármaco e representações de uma pluralidade de doenças dentro de um espaço de representação comum de um modelo estatístico treinado, e prever a nova indicação de doença para o determinado fármaco com base em uma comparação da representação projetada do determinado fármaco e pelo menos uma das representações da pluralidade de doenças no espaço de representação comum.

[0015] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um método de identificar a indicação de doenças para um determinado fármaco. O método compreende proporcionar como entrada para um modelo estatístico, representações de uma pluralidade de fármacos e uma pluralidade de doenças, e processar as representações da pluralidade de fármacos e a pluralidade de doenças usando um classificador supervisionado treinado para identificar uma probabilidade de que os fármacos na pluralidade de fármacos serão eficazes no tratamento de doenças na pluralidade de doenças, o classificador supervisionado treinado com informação sobre os pares de fármaco-doença aprovados pela Federal Drug Administration (FDA).

[0016] De acordo com outro aspecto da tecnologia descrita nesse documento, algumas modalidades são direcionadas a um sistema de computador, compreendendo pelo menos um computador processador e pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador, realiza um método de identificar indicação de doenças para um determinado fármaco. O método compreende proporcionar como entrada para um modelo estatístico, representações de uma pluralidade de fármacos e uma pluralidade de doenças, e processar as representações da pluralidade de fármacos e

a pluralidade de doenças usando um classificador supervisionado treinado para identificar uma probabilidade de que fármacos na pluralidade de fármacos serão eficazes no tratamento de doenças na pluralidade de doenças, o classificador supervisionado treinado com informações sobre pares de fármaco-doença aprovados pela Federal Drug Administration (FDA).

[0017] Deve ser observado que todas as combinações dos conceitos anteriores e conceitos adicionais discutidos em mais detalhes abaixo (desde que tais conceitos não sejam mutuamente inconsistentes) são contempladas como sendo parte do assunto inventivo divulgado neste documento.

Breve Descrição dos Desenhos

[0018] Várias modalidades não limitativas da tecnologia serão descritas com referência às Figuras a seguir. Deve ser observado que as Figuras não estão necessariamente desenhadas em escala.

[0019] A Figura 1 é um diagrama de uma rede heterogênea de acordo com algumas modalidades;

[0020] A Figura 2 é um diagrama de uma rede heterogênea de dados biológicos que podem ser representados usando um modelo estatístico multimodal de acordo com algumas modalidades;

[0021] A Figura 3 é um diagrama de um modelo de arquitetura para representar uma rede heterogênea de dados biológicos de acordo com algumas modalidades;

[0022] A Figura 4 é um gráfico de fluxo de um processo para treinar um modelo estatístico para representar uma rede heterogênea de dados biológicos de acordo com algumas modalidades;

[0023] A Figura 5 é um diagrama de um processo para realizar a incorporação de dados de acordo com algumas modalidades;

[0024] A Figura 6 é um diagrama de um processo para projetar informação de modalidade única e links de rede dentro de um espaço

latente comum de acordo com algumas modalidades;

[0025] A Figura 7 mostra exemplo de arquiteturas de redes neurais para codificadores e decodificadores usado de acordo com algumas modalidades;

[0026] A Figura 8 é um gráfico de fluxo de um processo para treinar um modelo estatístico para representar links de rede intramodalidade e intermodalidade em uma rede heterogênea de acordo com algumas modalidades;

[0027] A Figura 9 é um diagrama de um processo para treinar um modelo estatístico para representar links de rede de intramodalidade de acordo com algumas modalidades;

[0028] As Figuras 10A-10C são diagramas de processos para treinar um modelo estatístico para representar links de rede de intermodalidade de acordo com algumas modalidades;

[0029] A Figura 11 ilustra de modo esquemático a realização de uma previsão multimodal usando um modelo estatístico multimodal treinado de acordo com algumas modalidades;

[0030] A Figura 12 mostra um processo para a realização de previsões não supervisionadas em um espaço de representação específico de modalidade de acordo com algumas modalidades;

[0031] A Figura 13 ilustra de modo esquemático uma técnica para comparar posições de incorporação e vetores projetados em um espaço de representação específico de modalidade de acordo com algumas modalidades;

[0032] A Figura 14 mostra um processo para a realização de previsões não supervisionadas em um espaço de representação de modalidade conjunta de acordo com algumas modalidades;

[0033] A Figura 15 mostra um processo para a realização de previsões supervisionadas usando um modelo estatístico multimodal treinado de acordo com algumas modalidades; e

[0034] A Figura 16 mostra componentes de um sistema de computador ilustrativo no qual algumas modalidades podem ser implementadas.

Descrição Detalhada

[0035] Abordagens computacionais convencionais para prever associações entre dados biológicos (por exemplo, correspondências de fármaco-doença) usando de técnicas estatísticas ou de aprendizado de máquina normalmente emprega técnicas de aprendizado supervisionado. O conjunto de dados disponível para treinar tais técnicas é frequentemente limitado a uma quantidade relativamente pequena de dados rotulados (por exemplo, fármacos aprovados pela FDA). Essas abordagens também são normalmente focadas em uma ou duas modalidades (por exemplo, fármacos e doenças) e não consideram informações de outras modalidades durante o treinamento ou ao fazer previsões. Para este fim, algumas modalidades são direcionadas a uma técnica escalonável para integrar informações biológicas de múltiplas modalidades para incorporar informações biológicas (por exemplo, fármacos e/ou doenças) de uma ampla gama de fontes. Em particular, algumas modalidades são direcionadas para representar uma rede heterogênea de informações biológicas multimodais usando um ou mais modelos estatísticos configurados para aprender conexões entre os dados no modelo usando uma técnica de aprendizagem autossupervisionada. Um exemplo esquemático de uma rede heterogênea que pode ser representada usando um modelo estatístico de acordo com algumas modalidades é mostrado na Figura 1.

[0036] Como mostrado, a rede heterogênea 100 inclui uma pluralidade de nós e conexões entre os nós. Cada um dos nós na rede 100 está associado com dados tendo uma diferente modalidade. Por exemplo, o nó A pode representar dados associados com doenças, o

nó B pode representar dados associados com genes, e o nó C pode representar dados associados com fármacos. Os links associados com os nós na rede 100 incluem links intramodalidade (por exemplo, os links 132, 134) que descrevem interações entre dados dentro de uma única modalidade. Por exemplo, o link 132 descreve uma interação entre os dados associados com o nó B (por exemplo, genes interagindo com outros genes) e o link 134 descreve uma interação entre os dados associados com o nó C (por exemplo, fármacos tendo similaridade estrutural com outros fármacos). Cada nó na rede heterogênea pode incluir qualquer número adequado de links intramodalidade (incluindo nenhum link intramodalidade), e o número de links associados com qualquer um nó na rede pode ser dependente da modalidade dos dados associados com o nó. Por exemplo, como discutido em mais detalhes abaixo, um nó associado com a modalidade de "gene" pode ter mais links intramodalidade do que um nó associado com a modalidade de "classe de fármaco".

[0037] Cada nó na rede 100 também inclui pelo menos um link intermodalidade (por exemplo, os links 112, 114, 116 e 122) que descrevem uma interação entre dados a partir de modalidades diferentes. O link intermodalidade(s) conecta o nó a outro nó(s) na rede. Ao passo que alguns nós apenas incluem um único link intermodalidade, outros nós incluem múltiplos links intermodalidades para um ou mais outros nós o que indica associações mais complexas entre os dados na rede 100. Em virtude dos links de intermodalidade na rede 100, as associações entre dados de fontes de dados díspares na rede podem ser aprendidas em algumas modalidades para permitir previsões entre nós que estão direta ou indiretamente conectados por meio de outros nós na rede. Por exemplo, a associação entre os dados no nó A e no nó C pode ser aprendida através do link direto 116 entre esses dois nós, bem como caminhos indiretos entre o nó A e o nó C através do nó B

(por exemplo, através dos links 112, 114 e 122). A malha de conexões aprendidas entre os dados representados pelos nós na rede 100 adiciona à riqueza da representação de dados codificados usando um modelo estatístico treinado de acordo com algumas modalidades. Por exemplo, o modelo estatístico treinado pode ser usado para prever links perdidos dentro da rede heterogênea de fármaco-doença.

[0038] A Figura 2 mostra um exemplo de uma rede heterogênea de fármaco-doença que pode ser representada usando um modelo estatístico de acordo com algumas modalidades. Como mostrado, a rede inclui uma pluralidade de nós, cada um dos quais está associado com dados biológicos para uma diferente modalidade. A rede inclui links intramodalidade e intermodalidades associados com e conectando os nós na rede. Os links descrevem como pares de dados dentro de uma modalidade ou a partir de modalidades diferentes estão relacionados um ao outro. Por incluir múltiplos nós em uma rede heterogênea, relações entre fármacos e doenças podem ser estabelecidas através de múltiplas modalidades, tais como genes afetados por uma doença ou associado com a doença, genes regulados por fármacos ou objetivados por fármacos, e genes expressos em tecidos afetados pela doença. Além disso, os fármacos podem ser caracterizados pela estrutura molecular dos mesmos, pelos seus respectivos alvos proteicos, classe de fármaco e efeitos colaterais, ao passo que doenças podem também ser caracterizadas por ontologia da doença.

[0039] Na rede de fármaco-doença particular mostrada na Figura 2, o nó associado com genes representa links funcionais centrais entre fármacos e doenças por ser conectado diretamente com quatro outros nós na rede. Genes podem ser caracterizados por interações funcionais, tais como interações proteína-proteína, regulação da transcrição ou redes de co-expressão, e as suas respectivas vias biológicas ou associações de ontologia gênica. Em algumas modalidades, a rede

inclui um ou mais de metabolômicos de fármaco e doença, proteômicos, e informação de microbioma.

[0040] Na medida em que os dados biológicos adicionais se tornam disponíveis, a rede heterogênea de fármaco-doença mostrada na Figura 2 pode ser expandida para incluir nós adicionais e/ou links adicionais entre os nós. Dessa forma, a representação da rede heterogênea fármaco-doença é facilmente extensível e escalável, ao contrário de alguns modelos computacionais convencionais treinados para fazer previsões baseadas em dados de apenas uma ou duas modalidades. Novos nós ou tipos de dados representados nos nós existentes da rede heterogênea podem ser adicionados de qualquer maneira adequada. Por exemplo, em algumas modalidades, os nós dentro da rede heterogênea de fármaco-doença podem incluir dados associados a diferentes organismos (por exemplo, dados de conjuntos de dados de humanos e camundongos). As associações de fármaco-fenótipo a partir de organismos modelo, de *Saccharomyces cerevisiae* (levedura), *Caenorhabditis elegans* (verme), *Danio rerio* (peixe-zebra), *Arabidopsis thaliana* (talo ou agrião de orelha de rato) e *Drosophila melanogaster* (mosca da fruta) também podem ser incluídas. Em outro exemplo, as conexões entre organismos podem ser representadas no modelo usando associações de genes ortólogas.

[0041] Os dados associados aos nós na rede heterogênea podem ser identificados a partir de qualquer fonte de dados que forneça informações confiáveis sobre as interações entre os dados dentro de uma modalidade particular (por exemplo, interações gene-gene) ou entre dados de modalidades diferentes (por exemplo, tratamentos de fármacos para doenças). Em algumas modalidades, as informações sobre as interações de dados com a rede heterogênea são determinadas com base nas informações em bancos de dados publicamente acessíveis e/ou bancos de dados proprietários de

informações biológicas ou com base nos resultados de ensaios clínicos ou outras pesquisas médicas. Por exemplo, os dados associados a fármacos podem incluir informações relacionadas a pequenas moléculas e/ou produtos biológicos e dados associados a doenças podem incluir informações relacionadas a categorias de doenças, incluindo, mas não se limitando a neoplasias (por exemplo, leucemia, linfoma, câncer de pulmão, melanoma, câncer de tireoide, câncer hepático, câncer de próstata, câncer renal ou renal, câncer pancreático, câncer intestinal, glioblastoma, astrocitomas, câncer de mama, entre outros) e doenças não cancerosas (por exemplo, neurológicas, cardiovasculares, dermatológicas, musculoesqueléticas, urológicas, respiratórias, doenças nutricionais e metabólicas, etc.).

[0042] Uma rede heterogênea fármaco-doença usada de acordo com algumas modalidades também pode incluir informações relacionadas às interações gene-gene derivadas de telas letais sintéticas e interações gene-doença derivadas de Crispr- ou shRNA ou triagem de siRNA. Além disso, as informações sobre as interações diretas entre fármacos e doenças podem ser determinadas com base, pelo menos em parte, nas informações sobre fármacos aprovados pela FDA - indicações de doenças e experimentos de viabilidade de linha celular de câncer em vitro.

[0043] A Tabela 1 proporciona uma listagem de exemplo de conjuntos de dados e bases de dados que podem ser usados para identificar dados e interações para uma rede heterogênea de acordo com algumas modalidades. Como descrito em mais detalhes abaixo, informação sobre interações entre dados extraídos a partir das referidas fontes de dados (e outros) pode ser usada para treinar um modelo estatístico de modo que o modelo estatístico treinado é configurado para representar associações entre modalidades na rede heterogênea. O modelo estatístico treinado pode então ser usado para realizar novas previsões intermodalidades.

<u>Conjunto de dados</u>	<u>Base de dados</u>
Perfis de expressão de fármaco	CMAP-LINCS-L1000
Alvos, estrutura e classe de fármacos	ChEMBL, ChemSpider, PubChem, DrugDB, drugCentral
Perfil de expressão de doença	TCGA
Associação doença-gene (mutação)	COSMIC db, OMIM db, intogen db
Associação doença-anatomia	Medline V1.0 (Himmelstein DS. 2016)
Associações Gene-Via	KEGG, Reactome, WikiPathway, Gene Ontology
Associação/regulação Gene-Anatomia	GTEX Portal, TISSUES, Bgee
Interações proteína-proteína	StringDB, Base de dados de interação humana, e a Base de dados de referência de Proteína Humana
Interações regulatórias de gene	CMAP-LINCS-L1000

[0044] Tabela 1: Exemplo de bases de dados usado para construir a rede heterogênea de fármaco-doença.

[0045] Como discutido acima em conexão com a Figura 2, cada nó na rede heterogênea inclui pelo menos um link para um ou mais outros nós na rede. Algumas modalidades são direcionadas a codificar os referidos links entre dados na rede por treinar um modelo estatístico usando informação sobre pares de dados extraídos a partir de fontes de dados incluindo, mas não limitado às fontes de dados listadas na Tabela 1.

[0046] Cada um dos nós e seus links associados (não só intramodalidade, mas também intermodalidade) na rede da Figura 2 pode ser considerado separadamente para treinar um modelo estatístico de acordo com algumas modalidades. Cada um dos links entre dados para nós na rede pode ser representado usando características categóricas. As características categóricas permitem que os dados para cada modalidade sejam mapeadas a um vetor tendo valores contínuos usando a incorporação de dados técnicos descritos

em mais detalhes abaixo. Os vetores são então proporcionados como entrada para o modelo estatístico durante a fase de treinamento e pode ser usado para previsão em seguida do treinamento.

[0047] Em alguns casos, as interações entre dados na rede heterogênea podem ser representadas usando apenas características categóricas. Por exemplo, na interação "fármaco-trata-doença", um fármaco particular pode ou ser aprovado para tratar uma doença particular ou não aprovado. Em outras palavras, a interação "trata" é binária. Em outros casos, as interações entre dados na rede heterogênea podem além disso ser representadas usando características numéricas que indicam a intensidade da interação entre os dados vinculados. Por exemplo, na interação "fármaco-regula-gene", características categóricas podem ser usadas para representar se um determinado fármaco regula um determinado gene com base, por exemplo, em perfis de expressão de fármacos, e as características numéricas podem ser usadas para representar a extensão ou a intensidade da regulação conforme determinado, por exemplo, com base em comparações de expressão diferencial de genes.

[0048] Exemplos de interações associadas com a rede heterogênea mostrada na Figura 2 são descritos em mais detalhes abaixo incluindo uma indicação de quais dados a partir das bases de dados exemplificativas na Tabela 1 foram usados para determinar os dados de interação e se a interação foi representada na rede heterogênea usando apenas características categóricas ou características numéricas além das características categóricas. As interações na rede da Figura 2 são descritos abaixo por computar as métricas de interação de maneiras exemplares. No entanto, deve ser observado que qualquer ou todas as métricas de interação podem ser extraídas e/ou computadas a partir de fontes de dados de qualquer maneira adequada, e as modalidades não são limitadas a este respeito.

Interações centralizadas em fármaco

[0049] Como mostrado na Figura 2, o nó "fármaco" inclui seis diferentes interações centralizadas em fármaco incluindo uma interação intramodalidade (fármaco se assemelha a fármaco) e cinco interações intermodalidade que conectam o nó de fármaco a outros nós na rede. A interação intramodalidade "fármaco se assemelha a fármaco", que é definida por ambas as características numéricas e categóricas, descreve similaridades estruturais em pares de fármacos na rede. Por exemplo, a métrica "se assemelha" pode ser computada por calcular a similaridade estrutural de fármaco em pares a partir de impressões digitais que correspondem a fármacos, com base no coeficiente de Tanimoto e usando a biblioteca python RDKit (<http://www.rdkit.org>). Em uma implementação, estruturas de fármacos foram baixadas de três bancos de dados diferentes (ChEMBL, ChemSpider, PubChem) na forma de "sorrisos", seguido pela padronização de sorriso usando a biblioteca python MolVS (<https://molvs.readthedocs.io/>). Em seguida, as impressões digitais moleculares para cada fármaco foram calculadas e o coeficiente de Tanimoto de todas as comparações de impressões digitais de fármacos em pares possíveis foi calculado para determinar quais fármacos se assemelhavam a outros fármacos.

[0050] A interação "fármaco-regula-gene" é definida por ambas as características numéricas e categóricas. A referida interação pode ser determinada com base em perfis de expressão de fármaco extraídos, por exemplo, a partir da base de dados CMAP-LINCS-L1000. Em uma implementação, os dados foram baixados a partir da base de dados de Expressão gênica Omnibus (ID de Acesso = GSE92742), e continha um total de 19811 fármacos que foram rastreados em triplicata em dois pontos de tempo diferentes (6 horas e 24 horas) em um conjunto variável de 3-77 linhas de células bem anotadas. Os dados de expressão gênica usados nesta implementação incluíram dados

processados de nível 5, contendo para cada linha celular, ponto de tempo e tratamento com fármacos, os valores de expressão gênica diferencial normalizados em relação às condições de controle. Os dados podem ser representados por um vetor (por exemplo, de dimensão 1 x 12328) de genes e seus escores Z correspondentes para cada combinação de linha celular, ponto de tempo e tratamento com fármacos.

[0051] Além disso, dados de expressão gênica induzida a fármaco foram gerados para múltiplos fármacos a partir de uma base de dados proprietária. Esses perfis foram gerados em sete diferentes linhagens de células cancerígenas, em dois momentos diferentes (6 horas e 24 horas) e em duas concentrações diferentes para cada fármaco. A expressão diferencial do gene foi normalizada em relação à condição de controle e processada na forma de um escore Z. Os dados gerados para fármacos a partir do banco de dados proprietário tinham a mesma estrutura dos dados do CMAP-LINCS-L1000.

[0052] Como observado acima, a interação "fármaco-trata-doença" é categórica. A referida interação pode ser com base em uma lista de fármacos aprovados (por exemplo, aprovado pela FDA) e as indicações de doenças correspondentes dos mesmos. Em uma implementação, os dados para a referida interação foram baixados a partir da base de dados PharmacotherapyDB e continha 755 pares de doença-fármaco.

[0053] A interação "fármaco-inclui-classe de fármacos" é categórica. A referida interação descreve a correspondência entre cada fármaco e a sua classe farmacológica. Em uma implementação, os dados para a referida interação foram baixados a partir das bases de dados de drugBank (<https://www.drugbank.ca/>) e de drugCentral (<http://drugcentral.org>).

[0054] A interação "fármaco-liga-gene" é categórica. A referida interação descreve a relação entre fármacos e os alvos proteicos dos

mesmos, codificados por genes. Em uma implementação, os dados para a referida interação foram obtidos a partir das bases de dados de drugBank (<https://www.drugbank.ca/>), drugCentral (<http://drugcentral.org>), e BindingDB (<https://www.bindingdb.org>).

Interações Centradas em Doenças

[0055] Como mostrado na Figura 2, o nó de "doença" inclui cinco interações intermodalidades centradas em doenças diferentes (uma das quais é a interação "fármaco-trata-doença" descrita acima) que conectam o nó de doença a outros nós na rede. O nó da doença não está associado a nenhuma interação intramodal. A interação "doença-regula-gene" é representada usando recursos categóricos e numéricos. Em uma implementação, os dados para essa interação foram obtidos do banco de dados TCGA (<https://tcga-data.nci.nih.gov/>) e de um banco de dados proprietário. Essa interação está relacionada a genes que são regulados para mais e para menos no tecido doente, quando comparados aos tecidos de controle normais ou de indivíduos saudáveis. O banco de dados TCGA contém perfis de expressão de genes de câncer e seu perfil de tecido de controle normal correspondente para cada paciente. Em uma implementação, ambos os perfis para cada paciente foram baixados, a alteração dobrada correspondente entre o tumor e o controle foi calculada e os valores de expressão gênica foram normalizados para escores Z. Um banco de dados proprietário contendo aproximadamente 1500 perfis de expressão gênica de 575 doenças diferentes (indicações de doenças cancerígenas e não cancerígenas) também foi usado para gerar dados para a interação "doença-regula-gene". Os dados do banco de dados Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) foram baixados e processados usando as bibliotecas R GEOquery e Limma. Cada perfil de expressão da doença foi normalizado com Limma, seguido pelo cálculo da alteração da dobra do gene entre a doença e os

casos normais. Perfis de expressão de genes de doenças proprietários também foram normalizados para escores Z.

[0056] A interação "doença-associa-gene" é categórica. A referida interação se refere a mutações específicas a gene associadas a uma doença particular. Em uma implementação, as associações de mutações de gene que correspondem a Doenças de Mendel foram baixados a partir de a base de dados OMIM (<https://www.omim.org/>). As associações de mutações de gene que correspondem a cânceres específicos foram baixados a partir das bases de dados de COSMICdb (<https://cancer.sanger.ac.uk/cosmic>) e Intogen (<https://www.intogen.org/>).

[0057] A interação "doença-localiza-anatomia" é categórica. A referida interação se refere a uma associação entre doenças e tecidos humanos correspondentes afetados por doença. Em uma implementação, esses relacionamentos foram baixados a partir da associação doença Medline-tecido (Himmelstein DS. 2016). Os termos anatômicos foram mapeados para termos de ontologia de estruturas anatômicas (<http://uberon.github.io>, Mungall et al, 2012).

Interações Centradas em Gene

[0058] Como mostrado na Figura 2, o nó de "gene" inclui treze diferentes interações centradas em gene incluindo três interações intramodalidade e dez interações intermodalidade (seis das quais são descritas acima em conexão com as interações centralizadas em fármaco e doença) que conectam o nó de gene a outros nós na rede. A interação de intramodalidade "gene-interage com-gene" é categórica e se refere a interações físicas proteína-proteína baixadas, por exemplo, a partir das bases de dados de StringDB (<https://string-db.org/>), a base de dados da Humum interaction (<http://interactome.dfci.harvard.edu/>), e a base de dados da Human Protein Reference (<http://www.hprd.org>).

[0059] A interação intramodalidade "gene-regula-gene" é

representada usando ambas as características numéricas e categóricas. A referida interação se refere a níveis de expressão gênica normalizada através de diferentes linhagens de células de câncer com relação a knockdown ou superexpressão de genes específicos. Em uma implementação, os referidos dados foram baixados a partir de CMAP-LINCS-L1000, e os valores de expressão gênica foram normalizados em escores Z.

[0060] A interação intramodalidade de "gene-em covariação com-gene" é representada usando características categóricas e numéricas. Essa interação está relacionada à taxa de covariação evolutiva entre os genes. Em uma implementação, os dados para essa interação foram baixados a partir da Friedigkeit et al, 2015. O insight para incluir essa interação na rede é derivado da observação de que genes que tendem a coevoluir juntos estão geralmente envolvidos em vias biológicas similares e, portanto, podem participar de doenças similares.

[0061] A interação "gene-se expressa em-anatomia" é categórica e inclui níveis de expressão genica em tipos de tecidos humanos específicos. Em uma implementação, os dados para a referida interação foram baixados a partir da base de dados TISSUES (<https://tissues.jensenlab.org/>) e da GTEx Portal (<https://www.gtexportal.org/>). A base de dados TISSUES combina os dados a partir de expressão gênica, imuno-histoquímica, proteômicos e experimentos de mineração de texto, ao passo que a GTEx Portal contém dados de sequência de RNA a partir de múltiplos tecidos humanos.

[0062] A interação "gene regulados por anatomia" é categórica e inclui informação de regulação gênica (por exemplo, regulação para mais e para menos) em tipos de tecidos específicos. Em uma implementação, os dados para a referida interação foram extraídos a partir da base de dados de Bgee, para os humanos adultos (<https://bgee.org/>) e da GTEx Portal.

[0063] A interação "gene-participa na via" é categórica e se refere a uma associação entre genes e as suas vias celulares correspondentes. Em uma implementação, a função molecular, localização celular e processo biológico foram baixados a partir da Gene Ontology Consortium (<http://www.geneontology.org>). As associações que correspondem às vias metabólicas e de sinalização foram obtidas a partir de KEGG (www.genome.jp/kegg/), Reactome (<https://reactome.org>), e WikiPathways (<https://wikipathways.org/>).

[0064] Embora seis nós sejam mostrados na rede heterogênea ilustrativa da Figura 2, deve ser observado que uma rede heterogênea incluindo mais (ou menos) nós pode ser alternativamente representada usando um ou mais modelos estatísticos de acordo com algumas modalidades. Por exemplo, algumas modalidades são direcionadas a representar uma rede heterogênea incluindo apenas os três nós "fármaco", "gene", e "doença" e as suas ligações intramodalidades e intermodalidades correspondentes por um modelo estatístico. Em outras modalidades, uma rede heterogênea tendo pelo menos um nó que representa dados do paciente (por exemplo, a partir de um registro de saúde eletrônico) é representado usando um modelo estatístico.

[0065] Algumas modalidades são direcionadas a uma representação multimodal que integra todos os domínios e modalidades de uma rede heterogênea de dados biológicos, um exemplo do qual é descrito acima em conexão com a Figura 2. Ao contrário de algumas abordagens convencionais que dependem de aprendizagem supervisionada e de um conjunto de dados de treinamento limitado, algumas modalidades empregam técnicas de aprendizagem autossupervisionada que não requerem grandes conjuntos de dados emparelhados para treinamento. Conforme discutido em mais detalhes abaixo, o modelo estatístico é treinado em algumas modalidades para tirar proveito de conexões compartilhadas entre fármacos e doenças,

tais como genes, a fim de encontrar novas associações entre fármacos e doenças.

[0066] A Figura 3 ilustra de modo esquemático uma arquitetura de alto nível de um modelo estatístico que pode ser treinado usando técnica de aprendizado autossupervisionadas de acordo com algumas modalidades. Cada um dos nós que correspondem a uma diferente modalidade em uma rede heterogênea é representado como uma via separada a partir da entrada para a saída através da arquitetura. Apenas as modalidades "gene", "fármaco", e "doença" são representadas na arquitetura da Figura 3. No entanto, deve ser observado que outras modalidades incluindo, mas não limitadas aos outros nós na rede heterogênea da Figura 2, podem também ser incluídas no modelo de arquitetura mostrada na Figura 3.

[0067] Como mostrado, a arquitetura da Figura 3 inclui uma pluralidade de pares de codificador/decodificador, cada um dos quais é configurado para empregar uma técnica de aprendizado autossupervisionada para treinar valores para parâmetros do par de codificador/decodificador unimodal. O número de pares de codificador/decodificador incluídos na arquitetura depende do número de modalidades ou nós incluídos na rede heterogênea. Os pares de codificador/decodificador são unidos usando um espaço latente comum (também referido nesse documento como uma representação de modalidade conjunta ou representação multimodal conjunta) para formar um modelo estatístico multimodal que é capaz de aprender representação conjunta de cada nó de rede e os seus links de rede correspondentes, como descrito em mais detalhes abaixo.

[0068] Como mostrado na Figura 3, para cada par de codificador/decodificador, a arquitetura inclui uma pluralidade de representações de incorporação, que são vetores de valores contínuos que são uma transformação dos dados de entrada categóricos. Os

codificadores e decodificadores em cada par de codificador/decodificador são acoplados via uma representação de modalidade conjunta, que inclui vetores de representação conjuntos de nós de rede conectados na rede heterogênea. O número de vetores em uma representação de modalidade conjunta é igual ao número de interações na rede de modo que a representação de modalidade conjunta pode ser representada como uma matriz $N \times D$, onde N é o número de interações na rede e D é um comprimento de cada vetor de representação conjunta. Em algumas modalidades, $N > 1 \times 10^6$. Informação sobre interações entre dados na rede é codificado em uma representação de modalidade conjunta. As interações podem ser codificadas de qualquer modo adequado. Em algumas modalidades, um vetor de interação de incorporação que representa uma interação particular entre dados em um par de entrada pode ser criado e concatenado a um vetor de representação conjunta correspondente no espaço latente comum. Em outras modalidades, em vez de concatenar um vetor de interação de incorporação em um vetor de representação conjunta, o vetor de interação de incorporação pode ser concatenado para a saída a partir de dois codificadores a partir dos quais um vetor de representação conjunta é criado. Em ainda outras modalidades, a informação de interação pode ser intrinsecamente codificada em virtude de um vetor de representação conjunta sendo formado a partir da saída de dois codificadores para que dados de entrada particulares tendo uma interação particular sejam proporcionados.

[0069] Como discutido em mais detalhes abaixo, para interações intramodalidade (por exemplo, gene-gene), cada um dos pares de codificador/decodificador é treinado usando uma técnica de aprendizado autossupervisionada, pares de dados de entrada dentro da modalidade associada com um nó na rede heterogênea, e informação de interação descrevendo uma interação entre os pares de dados. Para

as interações intermodalidade (por exemplo, gene-fármaco), dois pares de codificador/decodificador são treinados usando uma técnica de aprendizado autossupervisionada, pares de dados de entrada através das duas modalidades, e informação de interação descrevendo uma interação entre os dados de entrada a partir das modalidades diferentes. Quando a interação inclui ambas as características numéricas e categóricas, as características numéricas podem ser levadas em conta, por exemplo, por multiplicar o vetor de interação de incorporação e/ou toda ou uma porção de um vetor de representação conjunta por um valor que corresponde à resistência ou grau da interação como representado nas características numéricas.

[0070] A Figura 4 ilustra um processo 400 para treinar um modelo estatístico multimodal tendo uma arquitetura mostrada na Figura 3, de acordo com algumas modalidades. No ato 410, os dados de treinamento (por exemplo, extraídos a partir de uma ou mais fontes de dados pública ou proprietária tais como as da Tabela 1) são convertidos em vetores de incorporação que têm que ser proporcionados como entrada para codificadores. Durante a incorporação de dados, variáveis categóricas relacionadas são representadas por vetores densos de números reais que capturam a relação entre os mesmos. Os vetores de incorporação representam cada variável em um espaço numérico contínuo. A criação de vetores de incorporação é descrita em mais detalhes em conexão com a Figura 5.

[0071] O processo 400 então prossegue para o ato 412, onde os vetores de incorporação são proporcionados como entrada para um codificador específico de modalidade para proporcionar um vetor de saída codificado em um espaço de representação de modalidade conjunta. O processo 400 então segue para o ato 414, onde um vetor de representação conjunta é calculado com base, pelo menos em parte, na saída dos vetores de saída codificados de dois codificadores. O vetor

de representação conjunta pode, ainda, ser calculado com base, pelo menos em parte, nas informações que descrevem uma interação entre os dados de entrada, como um vetor de interação de incorporação, conforme descrito acima. O processo 440 então prossegue para o ato 416, onde o vetor de representação conjunta é fornecido como entrada para um decodificador de modalidade específica para gerar um vetor de saída decodificado. O processo 400, então, prossegue para o ato 418, onde os pesos nos codificadores e decodificadores são atualizados com base, pelo menos em parte, em uma comparação do vetor de saída decodificado e do vetor incorporado fornecido como entrada para o codificador específico da modalidade. Por exemplo, uma técnica de aprendizagem autossupervisionada é usada para atualizar valores de parâmetros (por exemplo, pesos) no codificador e decodificador durante o treinamento. Cada um dos atos descritos no processo 400 é descrito em mais detalhes abaixo.

[0072] A Figura 5 mostra um processo para gerar vetores de incorporação para dados de entrada associados com um nó em uma rede heterogênea usando características categóricas de acordo com algumas modalidades. Uma dimensão de entrada V é definida para cada modalidade que corresponde ao tamanho do vocabulário dos dados na modalidade. No exemplo mostrado na Figura 5, a modalidade é "gene" e o tamanho do vocabulário V é 20.000, indicando que há 20.000 genes no conjunto de dados de entrada. Cada elemento da modalidade é "representado" por um vetor one-hot 510 de comprimento V , com o i^o elemento tendo um valor de 1, com todos os outros elementos no vetor sendo definidos como 0. Por exemplo, para codificar o elemento de dados de entrada "Gene A", o valor da posição 153 no vetor one-hot 510 é definido como 1, enquanto todos os outros valores no vetor são definidos como 0. Um vetor one-hot separado é criado para cada um dos elementos (por exemplo, cada um dos 20.000 genes no

exemplo da Figura 5) no conjunto de dados de entrada para a modalidade. Os vetores one-hot 510 são, então, projetados em um espaço de dimensão de incorporação inferior de tamanho $1 \times E$ que contém uma representação numérica contínua da variável de entrada, em vez de um valor binário. No exemplo mostrado na Figura 5, $E = 10$, embora deva ser observado que E pode ser definido para qualquer outro valor adequado e as modalidades não são limitadas a este respeito.

[0073] Em algumas modalidades, a incorporação de dados é realizada por transformar os vetores one-hot que correspondem a cada elemento de modalidade com uma matriz de incorporação 520 de dimensões $V \times E$ para produzir uma pluralidade de vetores de incorporação 530, cada um dos quais correspondendo a um diferente dos elementos de dados de entrada (por exemplo, Gene A no exemplo da Figura 5). Em algumas modalidades, os valores de matriz de incorporação 520 são aleatoriamente inicializados a partir de uma distribuição uniforme com uma faixa de $-1/V$ e $+1/V$. Durante o treinamento do modelo estatístico os valores para parâmetros de matriz de incorporação 520 podem permanecer fixos ou alternativamente podem ser atualizados como parte do processo de treinamento. Por atualizar os valores de parâmetro para a matriz de incorporação 520 durante o treinamento, é esperado que os vetores de incorporação 530 para os nós conectados na rede heterogênea estejam mais próximos no espaço de representação incorporada do que os nós não conectados.

[0074] Em algumas modalidades, os links de rede entre os nós na rede heterogênea são também incorporados usando um procedimento de incorporação similar como descrito acima, mas pode ter uma dimensão de incorporação mais baixa (por exemplo, 1×5) comparada à dimensão dos vetores de incorporação 530. A Figura 6 ilustra de modo esquemático um exemplo de como links de rede podem ser codificados em algumas modalidades. Em particular, a Figura 6 ilustra como vetores

de incorporação 530 produzidos como saída da arquitetura de incorporação de dados descrita em conexão com a Figura 5 são projetados dentro de um espaço latente comum 650 usando um codificador 602. O espaço latente comum 650 é também referido nesse documento como uma representação de modalidade conjunta. Como mostrado, o codificador 602 mapeia cada vetor de incorporação 530 para um vetor de representação latente de dimensão superior 604 dentro do espaço latente comum 650. No exemplo da Figura 6, o codificador 602 mapeia cada um dos vetores de incorporação a partir de uma dimensionalidade de 1×10 para uma dimensionalidade de 1×95 no espaço latente comum 650. Deve ser observado, no entanto, que a dimensionalidade de saída de codificador 602 pode adotar qualquer valor adequado. Um exemplo de arquitetura para o codificador 602 é descrito em mais detalhes abaixo em conexão com a Figura 7.

[0075] A Figura 6 também ilustra que a informação sobre os links de rede é também projetada para dentro do espaço latente comum 650 de acordo com algumas modalidades. Em um processo de incorporação similar àquele discutido de acordo com a Figura 5, a informação sobre links de rede em uma rede heterogênea pode ser incorporada por criar vetores one-hot 610 que correspondem a cada elemento de link de rede para um nó particular na rede. A Figura 6 mostra um exemplo de links de rede de incorporação para o nó de "gene" na rede heterogênea mostrada na Figura 2. O vetor one-hot 610 inclui nove elementos, cada um dos quais representa um dos nove tipos de links de rede intramodalidade ou intermodalidade associados ao nó de "gene" na Figura 2. Como mostrado, um vetor one-hot com o quinto elemento sendo definido como 1 e todos os outros elementos definidos como 0 podem ser usados, por exemplo, para incorporar o link "interage" correspondente ao link de rede "gene-interage-gene". A dimensão l do vetor one-hot 610 é baseada no número de tipos de links de rede associados a cada nó na rede.

[0076] Cada um dos vetores one-hot pode ser mapeado usando uma matriz de incorporação 620 de dimensões $I \times F$ para produzir uma pluralidade de vetores de interação de incorporação 630, cada um dos quais corresponde a um dos dados de elementos de entrada. Como descrito acima, em algumas modalidades $F < E$ de modo que a dimensionalidade dos vetores de interação de incorporação 630 é menor do que a dimensionalidade dos vetores de incorporação 530. Em algumas modalidades, os valores de matriz de incorporação 620 são aleatoriamente inicializados a partir de uma distribuição uniforme com variação de $-1/I$ e $+1/I$. Durante o treinamento do modelo estatístico os valores para parâmetros de matriz de incorporação 620 podem permanecer fixos ou alternativamente podem ser atualizados como parte do processo de treinamento. No exemplo de arquitetura da Figura 6, a informação sobre links de rede é representada no espaço latente comum 650 por concatenar um vetor de representação latente 604 e um vetor de interação de incorporação 634 emitido a partir de um processo de incorporação de link de rede, onde o vetor concatenado no espaço latente comum 650 representa não só os dados específicos de modalidade, mas também a informação de link de rede para os dados específicos de modalidade.

[0077] Como descrito acima, algumas modalidades empregam uma técnica de aprendizado autossupervisionada usando pares de codificadores/decodificadores para cada modalidade ou nó incluído na rede. Na técnica de aprendizado autossupervisionada, uma rede neural profunda é treinada para aprender ou reproduzir uma entrada X com base em o erro de reconstrução entre X e a saída X' . Treinar os parâmetros dos codificadores permite que os codificadores reconstruam representações de nível superior de vetores de entrada, ao passo que treinar os decodificadores permite que os decodificadores recuperem os vetores de entrada a partir de representações de nível superior.

[0078] Como descrito em conexão com a arquitetura da Figura 6, as entradas dos codificadores são os vetores de incorporação 530 de nós de rede, para cada variável ou elemento de cada modalidade. Os codificadores mapeiam cada vetor de incorporação em uma representação latente dimensional superior 604. Em algumas modalidades, os codificadores podem ser caracterizados por

$$Z = \alpha(W_e X + b_e) \text{ (Equação 1)}$$

onde X é o vetor de entrada de incorporação 530, Z é um vetor de saída ou representação latente 604, W_e e b_e representam pesos e orientações lineares, respectivamente, e α é uma função de ativação. Em algumas modalidades, a função de ativação é uma função de ativação não linear, por exemplo, uma Unidade Linear Retificada (ReLU), Unidade Linear Exponencial (ELU) ou função de ativação ReLU com vazamento.

[0079] A Figura 7 ilustra um exemplo de arquitetura para um codificador 620 que pode ser usado de acordo com algumas modalidades. No exemplo mostrado na Figura 7, o codificador 620 é implementado como uma rede neural amplamente conectada com uma camada oculta, e dimensões 10 (camada de entrada) -> 50 (camada oculta) -> 95 (camada de saída). A camada de saída do codificador 620 é um vetor de representação conjunta no espaço latente comum 650.

[0080] A porção de decodificador de cada par de codificador/decodificador é configurado para mapear a representação latente ou conjunta de dois nós interagindo (Z) na rede heterogênea de volta para o vetor de representação de incorporação de variáveis de entrada ou nós de rede individuais (X'). Em algumas modalidades, os decodificadores podem ser caracterizados por

$$X' = \alpha(W_d Z + b_d) \text{ (Equação 2)}$$

onde W_d e b_d representam pesos e orientações lineares, respectivamente, e α é uma função de ativação. Em algumas

modalidades, a função de ativação é uma função de ativação não linear, por exemplo, a Unidade Linear Retificada (ReLU), Unidade Linear Exponencial (ELU) ou função de ativação ReLU com vazamento.

[0081] A Figura 7 também ilustra um exemplo de arquitetura para um decodificador 720 que pode ser usado de acordo com algumas modalidades. No exemplo mostrado na Figura 7, o decodificador 620 é implementado como uma rede neural amplamente conectada com uma camada oculta, e dimensões 100 (camada de entrada) -> 50 (camada oculta) -> 10 (camada de saída). A camada de saída do decodificador 720 é um vetor decodificado X' tendo a mesma dimensionalidade que o vetor de incorporação X proporcionado como entrada para o codificador 620.

[0082] Tendo discutido uma arquitetura geral para os componentes de um modelo estatístico multimodal que pode ser usado para representar uma rede heterogênea de dados biológicos, exemplos de treinamento do modelo estatístico multimodal para aprender as associações entre dados em nós de uma rede são proporcionados abaixo.

[0083] A Figura 8 mostra um gráfico de fluxo de um processo 800 para treinar um modelo estatístico multimodal de acordo com algumas modalidades. As técnicas de treinamento particulares usadas podem depender dos tipos de interações entre os dados nos nós de uma rede heterogênea que têm que ser representadas no modelo. No ato 810, os vetores de incorporação específicos de modalidade são criados usando os processos de incorporação de dados descritos acima. Em modalidades que também criam vetores de interação de incorporação para concatenação no espaço latente comum, os referidos vetores de interação de incorporação podem também ser criados no ato 810 usando as técnicas de incorporação descritas nesse documento.

[0084] O processo 800 então prossegue para o ato 812, onde o

modelo estatístico multimodal é treinado para aprender interações intramodalidade para cada um dos nós na rede heterogênea que inclui pelo menos uma interação intramodalidade. Por exemplo, na rede heterogênea mostrada na Figura 2, apenas os nós de "gene" e "fármaco" são associados com links intramodalidade. Assim sendo, para cada um dos referidos nós, o modelo estatístico multimodal pode ser separadamente treinado para aprender os links de rede correspondentes de intramodalidade para o nó. Um exemplo de treinamento do modelo estatístico multimodal para aprender links de rede de intramodalidade é descrito em mais detalhes abaixo em conexão com a Figura 9. Deve ser observado que algumas redes heterogêneas podem não incluir quaisquer nós associados com links intramodalidade e que, para tal rede, o treinamento de links intramodalidade no ato 812 pode ser omitido.

[0085] O processo 800 então prossegue para o ato 814, onde o modelo estatístico multimodal é treinado para aprender interações intermodalidade descrevendo relações entre dados em diferentes nós conectados na rede heterogênea. Como descrito acima, cada um dos nós na rede heterogênea é conectado a pelo menos um outro nó na rede via um ou mais links de rede de intermodalidade. Para cada dos referidos links de rede, o treinamento no ato 814 é repetido até que o modelo estatístico multimodal tenha sido treinado em todos os links de rede na rede heterogênea. Um exemplo de treinamento do modelo estatístico multimodal para aprender links intermodalidade é descrito em mais detalhes abaixo em conexão com as Figuras 10A-C. Embora o ato 814 seja ilustrado em seguida do ato 812, deve ser observado que o treinamento de links intramodalidade e links intermodalidade pode ser realizado para os nós de uma rede heterogênea em qualquer ordem adequada incluindo, mas não limitado a treinamento em todos os links intramodalidade antes de treinar nos links intermodalidade, treinar em

todos os links intermodalidade antes de treinar nos links intramodalidade, e intercalar o treinamento de links intramodalidade e intermodalidades.

[0086] O processo 800 então prossegue para o ato 816, onde os parâmetros para o modelo estatístico treinado estimado durante o treinamento são armazenados para uso na realização de tarefas de previsão. Embora o ato 816 seja mostrado em seguida dos atos 812 e 814, deve ser observado que os parâmetros estimados para o modelo estatístico treinado podem ser armazenados após um ou mais iterações de treinamento em atos 812 ou 814 de modo que os parâmetros estimados determinados em uma iteração de treinamento são usados para inicializar pelo menos alguns dos parâmetros do modelo para uma iteração de treinamento subsequente. Como um exemplo, uma primeira iteração de treinamento pode ser focada no treinamento do link de rede "gene-interage-gene" com o resultado do treinamento sendo um codificador de gene e um decodificador de gene com parâmetros estimados que refletem a referida interação intramodalidade. Os parâmetros estimados para o codificadores/decodificadores de gene de gene podem ser armazenados e usados para inicializar os parâmetros de modelo para a iteração de treinamento subsequente focada no treinamento do link de rede "fármaco-liga-gene". Durante a subsequente iteração de treinamento os parâmetros estimados para o codificador/decodificador de gene são ainda refinados a partir dos valores anteriormente armazenados para refletir as associações associadas com treinamento de intermodalidade. Exemplos de propagação de parâmetros de modelo estimados a partir de uma iteração de treinamento para uma iteração de treinamento subsequente são discutidos em mais detalhes abaixo.

[0087] A Figura 9 ilustra de modo esquemático um processo para treinar um modelo estatístico multimodal para aprender um link de rede

"gene-interage-gene" de acordo com algumas modalidades. Como mostrado na Figura 9, dois pares de codificador/decodificador de gene são mostrados como sendo simultaneamente treinados. Embora mostradas como duas redes separadas para fins de ilustração, deve ser observado que cada um do par de codificador de gene e do par de decodificador de gene ilustrado na Figura 9 corresponde a uma única estrutura de rede, exemplos da qual são mostrados na Figura 7. A única estrutura de rede para o codificador de gene e o decodificador de gene inclui parâmetros (por exemplo, pesos de rede) que são estimados e atualizados durante o treinamento usando as técnicas de aprendizado autossupervisionadas descritas nesse documento.

[0088] Como mostrado, acoplar as saídas dos codificadores e entradas dos decodificadores é uma representação conjunta, que representa os links de rede de intramodalidade no qual o modelo estatístico multimodal é treinado. A Figura 9 mostra o treinamento de um link de rede que codifica uma interação entre um primeiro gene RPTOR e um segundo gene MTOR com base em fontes de dados, por exemplo, a partir de pelo menos uma das fontes de dados listadas na Tabela 1. Cada um dos genes RPTOR e MTOR é representado no modelo como vetores de incorporação (por exemplo, tendo dimensão 1×10) usando as técnicas de incorporação de dados como descrito acima. Opcionalmente, um link de rede ("interage" no exemplo da Figura 9) para ser treinado para o par gene-gene é também representado como um vetor de interação incorporado (por exemplo, tendo dimensão 1×5) como descrito acima.

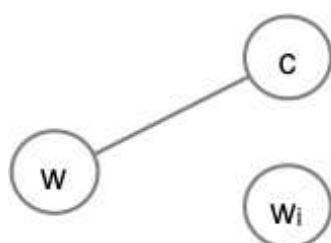
[0089] Os vetores de incorporação para RPTOR e MTOR são proporcionados como entrada para os casos do codificador de gene, que codifica a representação do vetor de incorporação para cada gene em um vetor de representação intramodalidade correspondente (por exemplo, tendo uma dimensão 1×95) no espaço latente comum. Em

modalidades em que um link de rede é também representado como um vetor de interação de incorporação, os vetores de representação intramodalidade para os dados de entrada "conectados" (isto é, os dados para genes RPTOR e MTOR na Figura 9) podem ser concatenados com o vetor de interação de incorporação no espaço latente comum como mostrado, o que resulta em dois vetores concatenados (por exemplo, tendo dimensões 1×100).

[0090] Um vetor de representação conjunta que representa os dados de entrada conectados e um link de rede que caracteriza a conexão é computado com base nos dois vetores de representação intramodalidade (opcionalmente concatenado com uma informação de link de rede) no espaço latente comum. Por exemplo, em algumas modalidades, um vetor de representação conjunta é computado por calcular a média ou produto de dois vetores de representação intramodalidade no espaço latente comum. Nessa implementação um vetor de representação conjunta tem a mesma dimensão que os vetores concatenados (isto é, 1×100 no exemplo da Figura 9). Como uma alternativa para o procedimento mostrado na Figura 9 para computar um vetor de representação conjunta, um vetor de representação conjunta pode ser computado em algumas modalidades com base em uma combinação dos dois vetores de representação intramodalidade (por exemplo, usando média ou um produto) antes da concatenação com o vetor de interação de incorporação descrevendo um link de rede, e o vetor de interação de incorporação pode ser concatenado com um vetor de representação conjunta em seguida de sua criação. Em um tal cenário um vetor de representação conjunta pode inicialmente ter uma dimensão a mesma que os vetores de representação intramodalidade individuais (por exemplo, 1×95), com a dimensão final de um vetor de representação conjunta sendo maior (por exemplo, 1×100) em seguida de concatenação.

[0091] O processo de treinamento na Figura 9 procede por proporcionar um vetor de representação conjunta (por exemplo, tendo uma dimensão 1 x 100) como entrada para o decodificador de gene (representado na Figura 9 como dois decodificadores de gene para ilustração), que é configurado para emitir vetores decodificados (por exemplo, tendo dimensão 1 x 10) para cada um dos genes de entrada RPTOR e MTOR. Um desvio entre os vetores decodificados emitidos a partir dos decodificadores e os vetores de entrada de incorporação proporcionados como entrada para os codificadores é medido e usado para atualizar os pesos no modelo estatístico de modo que o modelo aprende as associações entre os dados em um modo autossupervisionado. Em algumas modalidades, a técnica de aprendizado autossupervisionada é implementada usando uma função de perda de amostragem negativa, e o erro determinado a partir da função de perda de amostragem negativa é retropropagado através dos codificadores e decodificadores (e opcionalmente as matrizes de incorporação usadas para a incorporação de dados) para atualizar as estimativas dos parâmetros (por exemplo, pesos) para cada um dos referidos componentes do modelo.

[0092] A função de perda de amostragem negativa obriga os pares de codificador/decodificador a segregar a partir de conexões de rede aleatórias reais de acordo com a relação abaixo.



$$\log \sigma(c \cdot w) + \sum_{i=1}^n \mathbb{E}_{w_i \sim p(w)} [\log \sigma(-w_i \cdot w)]$$

onde w e c representam os nós de rede conectados, e w_i representa um nó de rede não relacionado.

[0093] Quando um link de rede sendo codificado é um link de rede de intramodalidade, como é o caso no exemplo da Figura 9, erros

determinados com base em ambos os pares de entrada/saída são considerados quando se determina como para atualizar as estimativas dos parâmetros para a representação de codificador de única modalidade. Posto de modo diferente, os parâmetros de ambas as instâncias do codificador/decodificador de gene ilustradas na Figura 9 seriam atualizadas da mesma maneira para cada ciclo de retropropagação.

[0094] Como discutido brevemente acima, algumas modalidades primeiro treinam o modelo estatístico para aprender os links de rede de intramodalidade e seguido pelo treinamento nos links de rede de intermodalidade. No caso de nós de rede já codificados em uma iteração de treinamento anterior, os parâmetros armazenados para as representações pré-treinadas dos componentes de rede (por exemplo, codificadores, decodificadores, matrizes de incorporação) podem ser usados em subseqüentes iterações de treinamento usando diferentes entradas.

[0095] A Figura 10A mostra um exemplo para treinar um modelo estatístico multimodal para aprender uma interação de intermodalidade para uma rede heterogênea de acordo com algumas modalidades. Em particular, a Figura 10A mostra como o modelo estatístico pode ser treinado para aprender o link de rede de "fármaco-liga-gene" na rede heterogênea mostrada na Figura 2. O processo de treinamento ilustrado na Figura 10A é similar àquele descrito na Figura 9 para treinar uma interação de rede de intramodalidade, com as diferenças principais sendo as entradas e os codificadores/decodificadores usados para o treinamento. Em suma, vetores de incorporação são criados para pares de dados específicos a partir de modalidades diferentes (fármacos e genes no exemplo da Figura 10A) que correspondem a diferentes nós na rede heterogênea. Os vetores de incorporação são criados usando os processos de incorporação de dados descritos acima usando vetores

one-hot e matrizes de incorporação correspondentes. No exemplo mostrado na Figura 10A, um primeiro vetor de incorporação é criado para o fármaco LAM-002 e um segundo vetor de incorporação é criado para o gene PIKFYVE. Os vetores de incorporação são proporcionados como entrada para os respectivos codificadores de fármaco e gene para mapear cada um dos vetores de incorporação em uma representação latente específica de modalidade dimensional superior no espaço de representação latente comum. A arquitetura dos codificadores de fármaco e gene pode ser similar àquela descrita acima em conexão com a Figura 7. Em algumas modalidades, a arquitetura do codificador/decodificador pode ter diferentes arquiteturas para modalidades diferentes, por exemplo, por ter um diferente número de camadas ocultas e/ou camadas com uma diferente dimensionalidade, com a representação de saída tendo a mesma dimensionalidade (por exemplo, 1 x 95) para cada um dos codificadores/decodificadores. Em outras modalidades, a arquitetura para os codificadores/decodificadores é idêntica a cada modalidade de dados representada no modelo estatístico, com as diferenças entre os codificadores/decodificadores sendo refletida nos pesos representados nas redes.

[0096] Como discutido em suma acima, um ou ambos os pares de codificador/decodificador pode ser associado com valores de parâmetros que são inicializados com base pelo menos em uma iteração de treinamento anterior. Por exemplo, em um cenário em que o treinamento de intramodalidade de um codificador/decodificador de gene como mostrado na Figura 9 foi realizado antes do treinamento de intermodalidade de codificadores/decodificadores de fármaco e gene como mostrado na Figura 10A, o par de codificador/decodificador de gene pré-treinado que resulta a partir do treinamento na Figura 9 pode ser usado para inicializar os parâmetros do par de codificador/decodificador de gene na arquitetura da Figura 10A. Desta

forma, o par codificador/decodificador para cada modalidade continua a ser treinado na medida em que novos pares de dados e interações de rede são fornecidos como entrada para o modelo estatístico multimodal.

[0097] Como mostrado na Figura 10A, as representações latentes específicas de modalidade emitidas a partir dos codificadores podem ser concatenadas a um vetor de interação de incorporação que representa um link de rede de intermodalidade particular entre os dados de entrada ("liga" no exemplo da Figura 10A). Em modalidades em que concatenação é usada para incorporar uma informação de link de rede na representação latente comum, a concatenação pode ocorrer quando se gera as representações latentes específicas de modalidade ou a concatenação pode ocorrer após as representações latentes específicas de modalidade terem sido combinadas para criar uma representação conjunta. As representações latentes específicas da modalidade podem ser combinadas, por exemplo, ao se obter uma média ou produto das duas representações latentes para calcular um vetor de representação conjunta que representa a interação de rede "gene-liga-fármaco" para o par de dados de entrada do fármaco LAM - 002 e o gene PIKFYVE. Continuando com o treinamento, a representação conjunta é fornecida como entrada para um decodificador de fármaco e um decodificador de gene para produzir vetores de saída decodificados (por exemplo, tendo dimensão 1 x 10), e os parâmetros dos codificadores e decodificadores (e opcionalmente as matrizes de incorporação) são atualizados com base em uma comparação dos vetores de saída decodificados e os vetores de incorporação fornecidos como entrada para os codificadores. Exemplos de como os pesos podem ser atualizados usando retropropagação de acordo com algumas modalidades são discutidos acima.

[0098] A Figura 10B mostra outro exemplo para treinar um modelo estatístico multimodal para aprender interações intermodalidade para

uma rede heterogênea de acordo com algumas modalidades. Em particular, a Figura 10B mostra como o modelo estatístico pode ser treinado para aprender o link de rede "doença-associa-gene" na rede heterogênea mostrada na Figura 2. O processo de treinamento ilustrado na Figura 10B é similar àquele descrito na Figura 10A, com as principais diferenças sendo as entradas e os codificadores/decodificadores usados para o treinamento. Em suma, vetores de incorporação são criados para pares de dados específicos a partir de modalidades diferentes (genes e doenças no exemplo da Figura 10B) que correspondem a diferentes nós na rede heterogênea. Os vetores de incorporação são criados usando os processos de incorporação de dados descritos acima usando vetores one-hot e matrizes de incorporação correspondentes. No exemplo mostrado na Figura 10B, um primeiro vetor de incorporação é criado para o gene BCL6 e um segundo vetor de incorporação é criado para a doença Linfoma. Os vetores de incorporação são proporcionados como entrada aos respectivos codificadores de gene e doença para mapear cada um dos vetores de incorporação em uma representação latente específica de modalidade dimensional superior na representação latente comum.

[0099] Um ou ambos os pares de codificador/decodificador pode ser associado com valores de parâmetros que são inicializados com base pelo menos em uma iteração de treinamento anterior. Por exemplo, em um cenário em que o treinamento de intermodalidade de um codificador/decodificador de gene como mostrado na Figura 10A foi realizado antes do treinamento de intermodalidade de codificadores/decodificadores de gene e doença na Figura 10B, o codificador de gene pré-treinado que resulta a partir do treinamento na Figura 10A pode ser usado para inicializar os parâmetros dos codificadores/decodificadores de gene na arquitetura da Figura 10B. Desse modo o par de codificador/decodificador para cada modalidade

continua a ser treinado como novos pares de dados e interações de rede são proporcionadas como entrada para o modelo estatístico multimodal.

[00100] Como mostrado na Figura 10B, as representações latentes específicas de modalidade emitidas a partir dos codificadores podem ser concatenadas em um vetor de interação de incorporação que representa um link de rede intermodal particular entre os dados de entrada ("associa" no exemplo da Figura 10B). Em modalidades em que a concatenação é usada para incorporar uma informação de link de rede na representação latente comum, a concatenação pode ocorrer quando se gera as representações latentes específicas de modalidade ou a concatenação pode ocorrer após as representações latentes específicas de modalidade terem sido combinadas para criar uma representação conjunta. As representações latentes específicas de modalidade podem ser combinadas, por exemplo, por se obter uma média ou produto das duas representações para computar um vetor de representação conjunta que representa a interação de rede "doença-associa-gene" para o par de dados de entrada do gene BCL6 e doença Linfoma. Continuando com o treinamento, a representação conjunta é fornecida como entrada para um decodificador de gene e um decodificador de doença para produzir vetores de saída decodificados (por exemplo, tendo dimensão 1×10), e os parâmetros dos codificadores e decodificadores (e opcionalmente as matrizes de incorporação) são atualizados com base em uma comparação dos vetores de saída decodificados e os vetores de incorporação fornecidos como entrada para os codificadores. Exemplos de como os pesos podem ser atualizados usando retropropagação de acordo com algumas modalidades são discutidos acima.

[00101] A Figura 10C mostra outro exemplo para treinar um modelo estatístico multimodal para aprender interações intermodalidade para uma rede heterogênea de acordo com algumas modalidades. Em

particular, a Figura 10C mostra como o modelo estatístico pode ser treinado para aprender o link de rede "fármaco-trata-doença" na rede heterogênea mostrada na Figura 2. O processo de treinamento ilustrado na Figura 10C é similar àquele descrito nas Figuras 10A e 10B, com as principais diferenças sendo as entradas e os codificadores/decodificadores usados para o treinamento. Em suma, vetores de incorporação são criados para os pares de dados específicos a partir de modalidades diferentes (fármacos e doenças no exemplo da Figura 10C) que correspondem a diferentes nós na rede heterogênea. Os vetores de incorporação são criados usando os processos de incorporação de dados descritos acima usando vetores one-hot e matrizes de incorporação correspondentes. No exemplo mostrado na Figura 10C, um primeiro vetor de incorporação é criado para o fármaco LAM-002 e um segundo vetor de incorporação é criado para a doença Linfoma. Os vetores de incorporação são proporcionados como entrada para os respectivos codificadores de fármaco e doença para mapear cada um dos vetores de incorporação em uma representação latente específica de modalidade dimensional superior na representação latente comum.

[00102] Um ou ambos dos pares de codificador/decodificador pode ser associado com valores de parâmetros que são inicializados com base pelo menos em uma iteração de treinamento anterior. Por exemplo, em um cenário em que o treinamento de intermodalidade de um codificador/decodificador de fármaco como mostrado na Figura 10A e o treinamento de intermodalidade de um codificador/decodificador de doença na Figura 10B foi realizado antes do treinamento de intermodalidade mostrado na Figura 10C, o par de codificador/decodificador de fármaco pré-treinado que resulta a partir do treinamento na Figura 10A pode ser usado para inicializar os parâmetros do par de codificador/decodificador de fármaco na

arquitetura da Figura 10C e o par de codificador/decodificador de doença pré-treinado que resulta a partir do treinamento na Figura 10B pode ser usado para inicializar os parâmetros para o par de codificador/decodificador de doença na arquitetura da Figura 10C. Desse modo o par de codificador/decodificador para cada modalidade continua a ser treinado na medida em que novos pares de dados e interações de rede são proporcionados como entrada para o modelo estatístico multimodal.

[00103] Como mostrado na Figura 10C, as representações latentes específicas de modalidade emitidas a partir dos codificadores podem ser concatenadas a um vetor de interação de incorporação que representa a link de rede intermodal particular entre os dados de entrada ("trata" no exemplo da Figura 10C). Em modalidades em que concatenação é usada para incorporar uma informação de link de rede na representação latente comum, a concatenação pode ocorrer quando se gera as representações latentes específicas de modalidade ou a concatenação pode ocorrer após as representações latentes específicas de modalidade terem sido combinadas para criar um vetor de representação conjunta. As representações latentes específicas de modalidade podem ser combinadas, por exemplo, por se obter uma média ou produto das duas representações para computar um vetor de representação conjunta que representa a interação de rede de "fármaco-trata-doença" para o par de dados de entrada de fármaco LAM-002 e doença Linfoma. Continuando com o treinamento, um vetor de representação conjunta é proporcionado como entrada para ambos um decodificador de fármaco e um decodificador de doença para produzir vetores de saída decodificados (por exemplo, tendo uma dimensão 1 x 10), e os parâmetros dos codificadores e decodificadores (e opcionalmente as matrizes de incorporação) são atualizados com base em uma comparação dos vetores de saída decodificados e os

vetores de incorporação proporcionado como entrada para os codificadores. Exemplos de como os pesos podem ser atualizados usando retropropagação de acordo com algumas modalidades são discutidos acima.

[00104] Todos os exemplos proporcionados acima nas Figuras 9 e 10A-C se referem ao treinamento do modelo estatístico para aprender interações de rede na rede heterogênea da Figura 2 que são categóricas apenas. Como discutido acima, algumas interações de rede podem ser ambas representadas por ambas as características numéricas e categóricas, em que as características numéricas representam a intensidade de uma interação entre dados dentro ou entre nós na rede. Para treinar o modal estatístico multimodal para aprender links de rede que são não só categóricos, mas também numéricos, a informação numérica pode ser usada para escalar os vetores de representação representados na representação de modalidade conjunta. Por exemplo, os vetores de interação de incorporação concatenados aos vetores de representação conjuntos podem ser escalados pela informação numérica.

[00105] Vários parâmetros (por exemplo, hiperparâmetros) da arquitetura estatística multimodal podem ser modificados com base na otimização para uma implementação particular. Os referidos parâmetros incluem mas não são limitados a dimensão de incorporação (exemplo, 1×10), dimensão de representação conjunta (exemplo, 1×100), dimensão de camada oculta(s) de codificadores e decodificadores (exemplo, 1×50), número de camada ocultas de codificadores e decodificadores (exemplo, 1), função de ativação para os codificadores e decodificadores, e a taxa de aprendizado.

[00106] Como discutido em conexão com a Figura 3, a arquitetura geral do modelo estatístico multimodal uma vez treinado inclui uma pluralidade de codificadores e decodificadores específicos de

modalidade treinados e uma representação de modalidade conjunta que acopla os codificadores treinados aos decodificadores treinados. Como mostrado de modo esquemático na Figura 11, o modelo estatístico multimodal treinado pode ser usado para realizar previsões entre dados de entrada tendo uma primeira modalidade e uma saída tendo uma diferente modalidade através da seleção de um par apropriado de codificadores e decodificadores treinados usados para a previsão. Especificamente, a Figura 11 mostra a habilidade do modelo estatístico multimodal treinado para realizar previsões sobre doenças que são passíveis de serem tratadas por um fármaco particular. Uma previsão é realizada, em parte, por usar um codificador de fármaco treinado e um decodificador de doença treinado, como mostrado. Múltiplos tipos de previsões podem ser feitos usando a rede estatística multimodal treinada, incluindo, mas não limitada uma nova indicação de doenças para um determinado fármaco, fármacos candidatos e terapias de combinação para uma determinada doença ou paciente, biomarcadores associado com a doença, e alvos de gene potencial para um determinado fármaco. Realizar tais previsões não é possível usando técnicas convencionais para a modelagem de dados biológicos que consideram apenas uma ou duas modalidades de dados.

[00107] Algumas modalidades são direcionadas a técnicas de previsão não supervisionadas usando um modelo estatístico multimodal treinado. A Figura 12 mostra um exemplo de uma técnica de previsão não supervisionada em que um espaço de representação para uma primeira modalidade (fármaco no exemplo da Figura 12) é mapeado em um espaço de representação para uma segunda modalidade (doença no exemplo da Figura 12) usando um decodificador para a segunda modalidade. Em uma técnica de previsão mostrada na Figura 12, indicação de doenças candidatas são previstas para um determinado fármaco proporcionado como entrada para o modelo estatístico

treinado. O codificador de fármaco treinado é usado para computar um vetor de representação latente para o fármaco de interesse em uma representação de modalidade conjunta, e o vetor de representação latente é proporcionado como entrada para o decodificador de doença treinado. A saída do decodificador de doença treinado é uma representação do fármaco de entrada projetada para dentro de um espaço de representação de doença.

[00108] Em vez de mapear o fármaco de entrada para a doença particular em um espaço de representação de doença, a saída do decodificador de doença pode ser projetada como um ponto 1310 em um espaço de representação de doença, como mostrado de modo esquemático na Figura 13. Um espaço de representação de doença mostrado na Figura 13 é uma representação de t-Distributed Stochastic Neighbor Embedding (t-SNE) do "espaço latente de doença" contendo apenas um subconjunto da base de dados de doença. Cada uma das doenças na qual o modelo estatístico multimodal foi treinado também tem uma posição intrínseca no espaço de representação de doença n-dimensional. Em algumas modalidades, uma nova indicação de doença é prevista com base, pelo menos em parte, em uma distância entre o ponto projetado 1310 e as posições de outras doenças no espaço de representação da doença. Por exemplo, novas indicações de doenças para o fármaco podem ser determinadas encontrando vizinhos mais próximos do ponto projetado 1310 e doenças candidatas projetadas dentro do espaço de representação da doença. As doenças candidatas com o maior potencial de serem tratáveis por um determinado fármaco podem incluir doenças em que a distância entre o ponto 1310 do projeto e os pontos para as doenças candidatas é pequena. Por exemplo, como mostrado na Figura 13, as doenças de gota, enxaqueca e esclerose múltipla são cada uma associada a pontos no espaço de representação da doença mais próximo do ponto projetado 1310 para um determinado

fármaco de entrada. Como tal, essas doenças podem ser bons candidatos como novos alvos de doença para o fármaco de interesse.

[00109] Em algumas modalidades, apenas a doença tendo a distância mais próxima para o ponto projetado 1310 pode ser proporcionado como uma previsão de saída. Em outras modalidades, uma lista de "n-melhor" de doenças associadas com distâncias mais próximas a o ponto projetado 1310 pode ser proporcionado como uma previsão de saída. Em ainda outras modalidades, apenas doenças tendo uma distância menor do que um valor limiar a partir do ponto projetado 1310 em um espaço de representação de doença podem ser emitidas. Outra informação além de nome(s) de doença(s) pode ser emitida incluindo, mas não limitado a uma escore de similaridade com base na distância.

[00110] Qualquer medição de distância adequada entre dois pontos no espaço de representação n-dimensional pode ser usado, e modalidades não são limitadas em relação a isso. Exemplos de medições de distância que podem ser usadas de acordo com algumas modalidades para previsão incluem, mas não são limitadas a distância Euclideana, Similaridade de cosseno, e distância de Manhattan. A fórmula para distância Euclideana entre dois vetores em um espaço de representação comum pode ser como a seguir:

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

[00111] A Figura 14 mostra um exemplo de outra técnica de previsão não supervisionada em que dados de entrada para duas modalidades diferentes (fármaco e doença no exemplo da Figura 14) é projetada para dentro de um espaço de representação de modalidade conjunta, onde comparações entre a vetores de representação conjuntos a partir das modalidades diferentes podem ser realizadas. Como mostrado, em uma técnica de previsão da Figura 14, os dados de entrada para uma

primeira modalidade (fármaco no exemplo da Figura 14) são proporcionados para um codificador treinado para a primeira modalidade. A saída do codificador treinado para a primeira modalidade é um primeiro vetor de representação conjunta para a entrada da primeira modalidade no espaço latente comum. Além disso, os dados de entrada para uma segunda modalidade (a pluralidade de doenças no exemplo da Figura 14) são proporcionados como entrada para um codificador treinado para a segunda modalidade. A saída do codificador treinado para a segunda modalidade é uma pluralidade de segundos vetores de representação conjuntos representado no espaço latente comum, cada um dos quais corresponde a dados de entrada for a segunda modalidade.

[00112] A previsão para a indicação de doenças candidatas para um determinado fármaco pode ser determinada por comparar uma distância do primeiro vetor de representação conjunta para o fármaco de entrada dentro do espaço latente comum e cada um dos segundos vetores de representação conjuntos para as doenças projetadas dentro do espaço latente comum. Por exemplo, de modo a prever a associação entre um fármaco A e quatro diferentes doenças, os codificadores de fármaco e a doença podem ser usados para computar as representações latentes correspondentes para o fármaco A e cada uma das quatro doenças. A distância entre os vetores de representação latentes para o fármaco A e os para cada doença projetada para dentro do espaço latente comum pode ser computada para prever a representação de doença mais próxima para a representação de fármaco A. As doenças candidatas com o potencial mais alto de ser tratável pelo determinado fármaco podem ser as doenças tendo posições no espaço de representação latente que são mais próximas da posição do fármaco de interesse no espaço de representação latente.

[00113] Embora as técnicas de previsão não supervisionadas como

descrito nas Figuras 12 e 14 se refiram a prever nova indicação de doenças para fármacos particulares, deve ser observado que técnicas de previsão não supervisionadas podem ser usadas para realizar previsões entre quaisquer duas modalidades representadas no modelo estatístico treinado por selecionar codificadores e/ou decodificadores treinados apropriados para permitir a previsão dentro de um espaço de representação comum dentro do modelo estatístico multimodal.

[00114] Algumas modalidades são direcionadas a técnicas de previsão supervisionadas usando um modelo estatístico multimodal treinado. A Figura 15 ilustra uma técnica de previsão supervisionada que usa um classificador supervisionado treinado com interações de rede conhecidas de duas modalidades diferentes. O classificador supervisionado pode ser implementado usando qualquer arquitetura adequada incluindo, mas não limitado a uma rede neural, um classificador de três bases, outros classificadores de aprendizado profundo ou de aprendizado de máquina, ou usando correlações técnicas estatísticas. O classificador pode ser treinado com as representações latentes dos pares de interação de rede conhecidos (por exemplo, a partir de indicação de doenças aprovadas para os fármacos aprovados pelo FDA), e as previsões sobre se há ou não uma associação verdadeira considerando o novo par podem ser feitas usando o classificador treinado.

[00115] Como mostrado, o classificador supervisionado na Figura 15 pode ser treinado com vetores de representação de FDA-aprovado fármaco-doença pares. Os vetores de entrada para os fármacos e doenças podem ter uma dimensão que corresponde à camada de incorporações de dados (por exemplo, 1×10) se usando o decodificador de doença para projetar os fármacos a um espaço de representação de doença ou o decodificador de fármaco para projetar as doenças para o espaço de representação de fármaco, ou a dimensão de um espaço de

representação latente (por exemplo, 1 x 95) se usar a representação latente de ambas as modalidades para realizar a decisão de classificação usando o classificador supervisionado treinado.

[00116] Além dos exemplos de previsão descritos acima, outros tipos de previsões são também contemplados por algumas modalidades. Por exemplo, previsões sobre novos fármacos que podem ser efetivos no tratamento de uma determinada doença podem ser realizadas. A doença de interesse e todos os fármacos podem ser projetados dentro de um espaço de representação comum (por exemplo, um espaço de representação específico de modalidade ou o espaço latente comum) no modelo estatístico multimodal e as distâncias entre vetores no espaço de representação comum podem ser usadas para prever os novos fármacos para tratar uma doença.

[00117] Pelo fato de todas as entidades na rede heterogênea representadas no modelo estatístico multimodal terem representações no mesmo espaço latente, e codificadores e decodificadores terem sido treinados para acessar o espaço latente, outras previsões de modalidade cruzada, além de novas correspondências de fármaco-doença, podem ser realizadas. Por exemplo, doenças podem ser codificadas por um codificador de doença treinado para prever genes alvos no espaço latente comum, ou por passar a representação latente de doença através do decodificador de gene e comparando a representação diretamente no espaço de gene (por exemplo, através do vizinho mais próximo e outras medições de distância ou técnicas de similaridade mencionadas acima). Dessa forma, além de prever novos fármacos associados a uma determinada doença, genes, proteínas, vias, anatomias e outras entidades biológicas também podem estar associados à doença, contextualizando a previsão de fármacos-doença. Além disso, pode ser demonstrado que uma mutação específica na rede heterogênea tem fortes associações com fármacos e doenças,

indicando biomarcadores que podem ajudar a identificar pacientes que responderão a determinados fármacos.

[00118] Em ainda outro cenário de previsão, genes alvos de um fármaco podem ser previstos de acordo com algumas modalidades. Fármacos são associado com genes, mutações, e outras entidades de rede heterogênea, que podem proporcionar percepções mecanicistas de ação de fármaco. Isso pode ser de valor, por exemplo, para o ajuste fino adicional das previsões de fármaco-doença com base em conhecimento especializado e engenharia tradicional de fármaco.

[00119] Ainda outra técnica de previsão se refere à previsão de terapias específicas do paciente. O modelo estatístico multimodal treinado pode ser usado para prever fármacos/terapias específicas para pacientes específicos. Por exemplo, conforme descrito acima, algumas modalidades são configuradas para prever biomarcadores associados a uma determinada doença. Os pacientes podem ser selecionados para esses biomarcadores, e os pacientes que possuem esses biomarcadores podem ser considerados bons candidatos para o tratamento com o determinado fármaco.

[00120] Como descrito acima, modalidades adicionais não ilustradas na Figura 2 podem também ser adicionadas à rede heterogênea representada por uma rede estatística multimodal treinada de acordo com as técnicas descritas nesse documento. Uma tal modalidade que pode ser adicionada se refere a pacientes. Por exemplo, a informação sobre o paciente pode ser incluída na rede heterogênea através de proximidade das suas propriedades dos pacientes (por exemplo, expressão gênica, mutação, variação do número de cópia, metilação do DNA) para outras entidades na rede heterogênea, ou por definir uma entidade de paciente como um novo nó na rede heterogênea (por exemplo, com um único codificador e decodificador de paciente usado para projetar informação do paciente para o espaço latente comum).

[00121] No cenário anterior, pacientes são representados no modelo estatístico multimodal com base em seus perfis de expressão gênica (ou outros atributos adquiridos de modo experimental), e essa informação pode ser ligada a outros nós (tais como por proximidade a perfis de expressão de fármacos e doenças conhecidos), e os nós ligados podem ser usados para projeção dentro do espaço latente.

[00122] No último cenário, a nova entidade de paciente ou nó pode ser adicionado à rede heterogênea, com o seu próprio codificador e decodificador incluído no modelo estatístico multimodal. Links de rede na rede heterogênea podem ser formados entre pacientes individuais (representados por um nó de paciente) e os nós de fármaco e doença na rede, por exemplo, com base em pacientes conhecidos por reagir bem a fármacos particulares ou por abrigar doenças. Além disso, os links na rede heterogênea podem ser formados entre dois pacientes que abrigam perfis similares de expressão gênica ou outra informação biológica adquirida de modo experimental ou atributos (por exemplo, DNA, RNA, Proteína, imagem médica). O codificador e decodificador de paciente pode ser treinado em um modo similar como pares de codificador/decodificador para outros nós na rede heterogênea, como descrito acima. Previsões usando o codificador/decodificador de paciente treinado pode ser realizado, por exemplo, entre um paciente de interesse e um fármaco candidato, usando uma ou mais das técnicas descritas nesse documento.

[00123] Uma implementação ilustrativa de um sistema de computador 1600 que pode ser usado em conexão com qualquer uma das modalidades da descrição proporcionadas nesse documento é mostrada na Figura 16. O sistema de computador 1600 pode incluir um ou mais processadores de hardware de computador 1600 e um ou mais artigos de fabricação que compreendem mídia de armazenamento não transitória legível por computador (por exemplo, memória 1620 e um ou

mais dispositivos de armazenamento não volátil 1630). O(s) processador(es) 1610 pode(m) controlar a gravação de dados e a leitura de dados da memória 1620 e do(s) dispositivo(s) de armazenamento não volátil 1630 de qualquer maneira adequada. Para executar qualquer uma das funcionalidades aqui descritas, o(s) processador(es) 1610 podem executar uma ou mais instruções executáveis por processador armazenadas em uma ou mais mídias de armazenamento não transitórias legíveis por computador (por exemplo, a memória 1620), que podem servir como mídia de armazenamento não transitória legível por computador que armazena instruções executáveis por processador para execução pelo(s) processador(es) 1610.

[00124] Os termos "programa" ou "software" são usados neste documento em um sentido genérico para se referir a qualquer tipo de código de computador ou conjunto de instruções executáveis por processador que podem ser empregados para programar um computador ou outro processador (físico ou virtual) para implementar vários aspectos das modalidades, conforme discutido acima. Além disso, de acordo com um aspecto, um ou mais programas de computador que, quando executados, executam métodos da divulgação fornecida neste documento, não precisam residir em um único computador ou processador, mas podem ser distribuídos de forma modular entre diferentes computadores ou processadores para implementar vários aspectos da divulgação fornecida nesse documento.

[00125] As instruções executáveis por processador podem estar em muitas formas, tais como módulos de programa, executados por um ou mais computadores ou outros dispositivos. Geralmente, os módulos do programa incluem rotinas, programas, objetos, componentes, estruturas de dados, etc. que executam tarefas específicas ou implementam tipos de dados abstratos específicos. Normalmente, a funcionalidade dos módulos do programa pode ser combinada ou distribuída.

[00126] Além disso, as estruturas de dados podem ser armazenadas em uma ou mais mídias de armazenamento não transitórias legíveis por computador em qualquer forma adequada. Para simplificar a ilustração, as estruturas de dados podem ter campos que estão relacionados por meio da localização na estrutura de dados. Tais relações podem ser alcançadas da mesma forma atribuindo armazenamento para os campos com localizações em um meio não transitório legível por computador que transmite a relação entre os campos. No entanto, qualquer mecanismo adequado pode ser usado para estabelecer relações entre informações em campos de uma estrutura de dados, incluindo através do uso de ponteiros, marcações ou outros mecanismos que estabelecem relações entre os elementos de dados.

[00127] Vários conceitos inventivos podem ser incorporados como um ou mais processos, dos quais exemplos foram fornecidos. Os atos realizados como parte de cada processo podem ser ordenados de qualquer forma adequada. Assim, as modalidades podem ser construídas nas quais os atos são realizados em uma ordem diferente da ilustrada, o que pode incluir a execução de alguns atos simultaneamente, mesmo que mostrados como atos sequenciais em modalidades ilustrativas.

[00128] Tal como aqui utilizado no relatório descritivo e nas reivindicações, a frase "pelo menos um", em referência a uma lista de um ou mais elementos, deve ser entendida como significando pelo menos um elemento selecionado a partir de qualquer um ou mais dos elementos na lista de elementos, mas não necessariamente incluindo pelo menos um de cada um dos elementos especificamente listados na lista de elementos e não excluindo quaisquer combinações de elementos na lista de elementos. Esta definição também permite que elementos possam estar opcionalmente presentes diferentes dos elementos especificamente identificados na lista de elementos aos

quais a frase "pelo menos um" se refere, sejam eles relacionados ou não àqueles elementos especificamente identificados. Assim, por exemplo, "pelo menos um de A e B" (ou, equivalentemente, "pelo menos um de A ou B" ou, equivalentemente "pelo menos um de A e/ou B") pode se referir, em uma modalidade, para pelo menos um, opcionalmente incluindo mais de um, A, sem B presente (e opcionalmente incluindo elementos diferentes de B); em outra modalidade, para pelo menos um, opcionalmente incluindo mais de um, B, sem A presente (e opcionalmente incluindo elementos diferentes de A); em ainda outra modalidade, para pelo menos um, opcionalmente incluindo mais de um, A, e pelo menos um, opcionalmente incluindo mais de um B (e opcionalmente incluindo outros elementos); etc.

[00129] A frase "e/ou", tal como aqui utilizada no relatório descritivo e nas reivindicações, deve ser entendida como significando "um ou ambos" os elementos assim conjugados, ou seja, elementos que estão conjuntivamente presentes em alguns casos e de forma disjuntiva presentes em outros casos. Vários elementos listados com "e/ou" devem ser interpretados da mesma maneira, ou seja, "um ou mais" dos elementos assim unidos. Outros elementos podem opcionalmente estar presentes além dos elementos especificamente identificados pela cláusula "e/ou", sejam eles relacionados ou não a esses elementos especificamente identificados. Assim, como um exemplo não limitativo, uma referência a "A e/ou B", quando usada em conjunto com a linguagem aberta, tal como "compreendendo" pode se referir, em uma modalidade, apenas a A (opcionalmente incluindo elementos diferentes B); em outra modalidade a B apenas (opcionalmente incluindo elementos diferentes de A); em ainda outra modalidade, para A e B (opcionalmente incluindo outros elementos); etc.

[00130] O uso de termos ordinais, como "primeiro", "segundo", "terceiro", etc., nas reivindicações para modificar um elemento de

reivindicação não conota por si só qualquer prioridade, precedência ou ordem de um elemento de reivindicação sobre outro ou a ordem temporal em que os atos de um método são executados. Esses termos são usados apenas como rótulos para distinguir um elemento de reivindicação com um determinado nome de outro elemento com o mesmo nome (mas para uso do termo ordinal). A fraseologia e a terminologia usadas aqui são para fins de descrição e não devem ser consideradas como limitantes. O uso de "incluindo", "compreendendo", "tendo", "contendo", "envolvendo" e variações dos mesmos, se destina a abranger os itens listados a seguir e itens adicionais.

[00131] Tendo descrito várias modalidades das técnicas aqui descritas em detalhes, várias modificações e melhorias ocorrerão prontamente para aqueles versados na técnica. Tais modificações e melhorias se destinam a estar dentro do espírito e escopo da divulgação. Consequentemente, a descrição anterior é apenas a título de exemplo e não se destina a ser limitante. As técnicas são limitadas apenas conforme definido pelas reivindicações a seguir e seus equivalentes.

REIVINDICAÇÕES

1. Método para treinar um modelo estatístico configurado para representar associações de intramodalidade entre dados em uma rede heterogênea, caracterizado pelo fato de que o método compreende:

acessar dados de treinamento incluindo dados de treinamento para uma primeira modalidade e dados de treinamento para uma segunda modalidade diferente a partir da primeira modalidade;

treinar o modelo estatístico, o modelo estatístico compreendendo primeiro e segundo codificadores, primeiro e segundo decodificadores, e uma representação de modalidade conjunta acoplando os primeiro e segundo codificadores aos primeiro e segundo decodificadores, o treinamento compreendendo:

estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores usando uma técnica de aprendizado autossupervisionada, pelo menos alguns dos dados de treinamento, e informação descrevendo pelo menos um link entre pares de dados nos dados de treinamento; e

armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores do modelo estatístico.

2. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que ainda compreende:

criar uma primeira modalidade de vetores de incorporação com base nos dados de treinamento para a primeira modalidade;

criar uma segunda modalidade de vetores de incorporação com base nos dados de treinamento para a segunda modalidade, em que o treinamento ainda compreende:

proporcionar como entrada aos primeiro e segundo

codificadores, as primeira e segunda modalidades de vetores de incorporação, respectivamente.

3. Método, de acordo com a reivindicação 2, caracterizado pelo fato de que o modelo estatístico ainda compreende primeira e segunda camadas de incorporação, e em que o treinamento ainda compreende estimar valores para parâmetros das primeira e segunda camadas de incorporação.

4. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que ainda compreende:

criar um vetor de intramodalidade descrevendo um link entre pares de dados nos dados de treinamento para a primeira modalidade, e

em que a informação na representação de modalidade conjunta é determinada com base, pelo menos em parte, no vetor de intramodalidade.

5. Método, de acordo com a reivindicação 4, caracterizado pelo fato de que ainda compreende:

dimensionar o vetor de intramodalidade com base em uma característica numérica indicando a intensidade do link entre os pares de dados nos dados de treinamento para a primeira modalidade, e

em que a informação na representação de modalidade conjunta é determinada com base, pelo menos em parte, no vetor de intramodalidade dimensionado.

6. Método, de acordo com a reivindicação 4, caracterizado pelo fato de que ainda compreende:

concatenar cada uma de um primeiro vetor de característica de saída a partir do primeiro codificador e um segundo vetor de característica de saída a partir do segundo codificador com o vetor de intramodalidade para produzir o primeiro e o segundo vetores de características concatenadas; e

computar um vetor de representação conjunta dentro de uma representação de modalidade conjunta usando o primeiro e o segundo vetores de características concatenadas.

7. Método, de acordo com a reivindicação 4, caracterizado pelo fato de que ainda compreende:

computar um vetor de característica conjunta usando um primeiro vetor de característica de saída a partir do primeiro codificador e um segundo vetor de característica de saída a partir do segundo codificador; e

concatenar o vetor de característica conjunta com o vetor de intramodalidade para produzir um vetor de representação conjunta dentro de uma representação de modalidade conjunta.

8. Método, de acordo com a reivindicação 2, caracterizado pelo fato de que os primeiro e segundo codificadores e os primeiro e segundo decodificadores são configurados para processar dados a partir da primeira modalidade, e em que o treinamento ainda compreende:

proporcionar como entrada para o primeiro codificador um primeiro dos vetores de incorporação de primeira modalidade;

proporcionar como entrada para o segundo codificador um segundo dos vetores de incorporação de primeira modalidade; e

computar um vetor de representação conjunta na representação de modalidade conjunta com base em um primeiro vetor de característica de saída a partir do primeiro codificador, um segundo vetor de característica de saída a partir do segundo codificador, e o vetor de intramodalidade;

proporcionar um vetor de representação conjunta como entrada aos primeiro e segundo decodificadores para produzir primeiro e segundo vetores decodificados; e

estimar valores para parâmetros dos primeiro e segundo

codificadores e dos primeiro e segundo decodificadores com base em um de primeiro e de segundo dos vetores de incorporação de primeira modalidade e de primeiro e segundo vetores decodificados.

9. Método, de acordo com a reivindicação 8, caracterizado pelo fato de que estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores compreende usar uma função de perda de amostragem negativa.

10. Método, de acordo com a reivindicação 9, caracterizado pelo fato de que o modelo estatístico ainda compreende primeira e segunda camadas de incorporação, e em que treinar ainda compreende estimar valores para parâmetros das primeira e segunda camadas de incorporação usando a função de perda de amostragem negativa.

11. Método, de acordo com a reivindicação 9 ou 10, caracterizado pelo fato de que ainda compreende:

repetir o treinamento do modelo estatístico para cada um de uma pluralidade de links entre pares de dados nos dados de treinamento para a primeira modalidade.

12. Método, de acordo com a reivindicação 2, caracterizado pelo fato de que ainda compreende:

criar um vetor de intermodalidade descrevendo um link entre os dados de treinamento para a primeira modalidade e os dados de treinamento para a segunda modalidade, e

em que a informação na representação de modalidade conjunta é determinada com base, pelo menos em parte, no vetor de intermodalidade.

13. Método, de acordo com a reivindicação 12, caracterizado pelo fato de que ainda compreende:

dimensionar o vetor de intermodalidade com base em uma característica numérica indicando a intensidade do link entre os pares de dados nos dados de treinamento para a primeira modalidade, e

em que a informação na representação de modalidade conjunta é determinada com base, pelo menos em parte, no vetor de intermodalidade dimensionado.

14. Método, de acordo com a reivindicação 12, caracterizado pelo fato de que ainda compreende:

concatenar cada um de um primeiro vetor de característica de saída a partir do primeiro codificador e um segundo vetor de característica de saída a partir do segundo codificador com o vetor de intermodalidade para produzir primeiro e segundo vetores de características concatenadas; e

computar um vetor de representação conjunta dentro da representação de modalidade conjunta usando o primeiro e o segundo vetores de características concatenadas.

15. Método, de acordo com a reivindicação 12, caracterizado pelo fato de que ainda compreende:

computar um vetor de característica conjunta usando um primeiro vetor de característica de saída a partir do primeiro codificador e um segundo vetor de característica de saída a partir do segundo codificador; e

concatenar o vetor de característica conjunta com o vetor de intermodalidade para produzir um vetor de representação conjunta dentro da representação de modalidade conjunta.

16. Método, de acordo com a reivindicação 12, caracterizado pelo fato de que o primeiro codificador e o primeiro decodificador são configurados para processar dados a partir da primeira modalidade e o segundo codificador e o segundo decodificador são configurados para processar dados a partir da segunda modalidade, e em que o treinamento ainda compreende:

proporcionar como entrada para o primeiro codificador dos um dos vetores de incorporação de primeira modalidade;

proporcionar como entrada para o segundo codificador um dos vetores de incorporação de segunda modalidade; e

computar um vetor de representação conjunta na representação de modalidade conjunta com base em um primeiro vetor de característica de saída a partir do primeiro codificador, um segundo vetor de característica de saída a partir do segundo codificador, e o vetor de intermodalidade;

proporcionar um vetor de representação conjunta como entrada aos primeiro e segundo decodificadores para produzir primeiro e segundo vetores decodificados; e

estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores com base em um dos vetores de incorporação de primeira modalidade e um dos vetores de incorporação de segunda modalidade e os primeiro e segundo vetores decodificados.

17. Método, de acordo com a reivindicação 16, caracterizado pelo fato de que estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores compreende usar a função de perda de amostragem negativa.

18. Método, de acordo com a reivindicação 16, caracterizado pelo fato de que ainda compreende:

repetir o treinamento do modelo estatístico para cada um de uma pluralidade de links entre os dados de treinamento para a primeira modalidade e os dados de treinamento para a segunda modalidade.

19. Método, de acordo com a reivindicação 16, caracterizado pelo fato de que ainda compreende:

inicializar, antes do treinamento, os valores dos parâmetros para o primeiro codificador, em que a inicialização é realizada com base em resultados de treinamento do primeiro codificador usando uma

técnica de aprendizado autossupervisionada.

20. Método, de acordo com a reivindicação 19, caracterizado pelo fato de que os resultados de treinamento do primeiro codificador usando uma técnica de aprendizado autossupervisionada compreende os resultados de treinamento do primeiro codificador usando dados de treinamento a partir de apenas a primeira modalidade.

21. Método, de acordo com a reivindicação 19, caracterizado pelo fato de que os resultados de treinamento do primeiro codificador usando uma técnica de aprendizado autossupervisionada compreende os resultados de treinamento do primeiro codificador com dados de treinamento tendo uma terceira modalidade diferente a partir da segunda modalidade.

22. Método, de acordo com a reivindicação 19, caracterizado pelo fato de que ainda compreende:

inicializar, antes do treinamento, os valores dos parâmetros para o segundo codificador, em que a inicialização é realizada com base em resultados de treinamento do segundo codificador usando uma técnica de aprendizado autossupervisionada.

23. Método, de acordo com a reivindicação 2, caracterizado pelo fato de que criar a primeira modalidade de vetores de incorporação compreende:

definir, para cada dado nos dados de treinamento para a primeira modalidade, um vetor one-hot tendo um comprimento V_1 ; e

multiplicar cada um dos vetores one-hot de comprimento V_1 por uma primeira matriz de incorporação tendo dimensionalidade $V_1 \times E$, onde $E < V_1$, e em que E é um comprimento de cada uma da primeira modalidade de vetores de incorporação.

24. Método, de acordo com a reivindicação 23, caracterizado pelo fato de que criar a segunda modalidade de vetores de incorporação compreende:

definir, para cada dado nos dados de treinamento para a segunda modalidade, um vetor one-hot tendo um comprimento V_2 ; e

multiplicar cada um dos vetores one-hot de comprimento V_2 por uma matriz de incorporação tendo dimensionalidade $V_2 \times E$, onde $E < V_2$, e em que E é um comprimento de cada vetores de incorporação de segunda modalidade.

25. Método, de acordo com a reivindicação 24, caracterizado pelo fato de que cada um dos primeiro e segundo codificadores compreende uma camada de entrada tendo E entradas e uma camada de saída tendo R saídas, onde $R > E$.

26. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que cada um dos primeiro e segundo codificadores compreende pelo menos uma camada oculta.

27. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que cada um dos primeiro e segundo codificadores compreende uma rede neural.

28. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que os dados de treinamento ainda incluem dados de treinamento para uma terceira modalidade diferente a partir da primeira modalidade e da segunda modalidade, em que o modelo estatístico ainda compreende um terceiro codificador e um terceiro decodificador, e

em que treinar o modelo estatístico ainda compreende estimar valores para parâmetros do terceiro codificador e do terceiro decodificador usando uma técnica de aprendizado autossupervisionada, a terceira modalidade de vetores de entrada, e informação descrevendo pelo menos um link entre dados de treinamento para a terceira modalidade e dados de treinamento para a primeira ou segunda modalidades.

29. Método, de acordo com a reivindicação 28,

caracterizado pelo fato de que ainda compreende criar a terceira modalidade de vetores de incorporação com base nos dados de treinamento para a terceira modalidade, e em que treinar o modelo estatístico ainda compreende proporcionar como entrada para o terceiro codificador, uma primeira da terceira modalidade de vetores de incorporação.

30. Método para prever uma associação entre dados de entrada em uma primeira modalidade e dados em uma segunda modalidade usando um modelo estatístico treinado para representar links entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplando a pluralidade de codificadores e decodificadores, caracterizado pelo fato de que o método compreende:

selecionar, com base na primeira modalidade e na segunda modalidade, um par de codificador/decodificador ou um par de codificadores, a partir de entre a pluralidade de codificadores e decodificadores; e

processar os dados de entrada com a representação de modalidade conjunta e o par de codificador/decodificador selecionado ou o par de codificadores para prever a associação entre os dados de entrada e os dados na segunda modalidade.

31. Método, de acordo com a reivindicação 30, caracterizado pelo fato de que ainda compreende:

selecionar um codificador treinado para processar dados para a primeira modalidade e um decodificador treinado para processar dados para a segunda modalidade.

32. Método, de acordo com a reivindicação 31,

caracterizado pelo fato de que ainda compreende:

prever a associação entre os dados de entrada e os dados na segunda modalidade em um espaço de representação para a segunda modalidade.

33. Método, de acordo com a reivindicação 32, caracterizado pelo fato de que ainda compreende:

emitir a partir do decodificador treinado para processar dados para a segunda modalidade, um vetor de saída no espaço de representação para a segunda modalidade;

projetar os dados a partir da segunda modalidade em um espaço de representação para a segunda modalidade para produzir uma pluralidade de vetores projetados; e

prever a associação entre os dados de entrada e os dados na segunda modalidade com base em uma comparação do vetor de saída e os vetores projetados no espaço de representação para a segunda modalidade.

34. Método, de acordo com a reivindicação 33, caracterizado pelo fato de que ainda compreende:

calcular uma distância entre o vetor de saída e cada um da pluralidade de vetores projetados; e

prever a associação com base nas distâncias calculadas.

35. Método, de acordo com a reivindicação 34, caracterizado pelo fato de que calcular uma distância compreende calcular a distância Euclideana.

36. Método, de acordo com a reivindicação 30, caracterizado pelo fato de que ainda compreende:

selecionar um primeiro codificador treinado para processar dados para a primeira modalidade e um segundo codificador treinado para processar dados para a segunda modalidade.

37. Método, de acordo com a reivindicação 36,

caracterizado pelo fato de que ainda compreende:

prever a associação entre os dados de entrada e os dados na segunda modalidade em um espaço de representação latente associado com a representação de modalidade conjunta.

38. Método, de acordo com a reivindicação 37, caracterizado pelo fato de que ainda compreende:

proporcionar como entrada para o primeiro codificador, os dados de entrada para produzir um vetor de característica de primeira modalidade no espaço de representação latente;

proporcionar como entrada para o segundo codificador, os dados para a segunda modalidade para produzir uma pluralidade de vetores de característica de segunda modalidade no espaço de representação latente; e

prever a associação entre os dados de entrada e os dados na segunda modalidade com base em uma comparação de um vetor de característica de primeira modalidade e uma pluralidade de vetores de característica de segunda modalidade no espaço de representação latente.

39. Método, de acordo com a reivindicação 38, caracterizado pelo fato de que ainda compreende:

calcular uma distância entre um vetor de característica de primeira modalidade e cada um da pluralidade de vetores de característica de segunda modalidade no espaço de representação latente; e

prever a associação com base nas distâncias calculadas.

40. Método, de acordo com a reivindicação 39, caracterizado pelo fato de que calcular uma distância compreende calcular a distância Euclideana.

41. Método para prever associações entre dados em uma primeira modalidade e dados em uma segunda modalidade usando um

modelo estatístico treinado para representar interações entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplando a pluralidade de codificadores e decodificadores, caracterizado pelo fato de que o método compreende:

- mapear os dados na primeira modalidade e os dados na segunda modalidade dentro de um espaço de representação comum dentro do modelo estatístico;

- acessar um classificador estatístico treinado usando dados rotulados, em que os dados rotulados descrevem associações entre dados nas primeira e segunda modalidades; e

- prever associações entre os dados na primeira modalidade e os dados na segunda modalidade mapeados dentro do espaço de representação comum usando o classificador estatístico treinado.

42. Método, de acordo com a reivindicação 41, caracterizado pelo fato de que mapear os dados na primeira modalidade e os dados na segunda modalidade em um espaço de representação comum compreende mapear os dados em um espaço de representação de modalidade conjunta do modelo estatístico.

43. Método, de acordo com a reivindicação 41, caracterizado pelo fato de que mapear os dados na primeira modalidade e os dados na segunda modalidade em um espaço de representação comum compreende mapear os dados em um espaço de representação específico de modalidade para a primeira modalidade ou a segunda modalidade.

44. Sistema de computador, caracterizado pelo fato de que compreende:

pelo menos um computador processador; e

pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador realizam um método de treinar um modelo estatístico para representar associações de intermodalidades para os dados, em que os dados incluem dados para uma primeira modalidade e dados para uma segunda modalidade diferente a partir da primeira modalidade, o método compreendendo:

acessar dados de treinamento incluindo dados de treinamento para a primeira modalidade e dados de treinamento para a segunda modalidade;

treinar o modelo estatístico, o modelo estatístico compreendendo primeiro e segundo codificadores, primeiro e segundo decodificadores, e uma representação de modalidade conjunta acoplando os primeiro e segundo codificadores aos primeiro e segundo decodificadores, o treinamento compreendendo:

estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores usando uma técnica de aprendizado autossupervisionada, pelo menos alguns dos dados de treinamento, e informação descrevendo pelo menos um link entre pares de dados nos dados de treinamento; e

armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores do modelo estatístico.

45. Sistema de computador, de acordo com a reivindicação 44, caracterizado pelo fato de que o método ainda compreende:

criar vetores de incorporação de primeira modalidade com base nos dados de treinamento para a primeira modalidade;

criar vetores de incorporação de segunda modalidade com

base nos dados de treinamento para a segunda modalidade, em que o treinamento ainda compreende:

proporcionar como entrada aos primeiro e segundo codificadores, os vetores de incorporação de primeira e segunda modalidade, respectivamente.

46. Sistema de computador, de acordo com a reivindicação 45, caracterizado pelo fato de que o modelo estatístico ainda compreende primeira e segunda camadas de incorporação, e em que o treinamento ainda compreende estimar valores para parâmetros das primeira e segunda camadas de incorporação.

47. Sistema de computador, de acordo com a reivindicação 44, caracterizado pelo fato de que o método ainda compreende:

criar um vetor de intramodalidade descrevendo um link entre pares de dados nos dados de treinamento para a primeira modalidade, e

em que a informação na representação de modalidade conjunta é determinada com base, pelo menos em parte, no vetor de intramodalidade.

48. Sistema de computador, de acordo com a reivindicação 47, caracterizado pelo fato de que o método ainda compreende:

dimensionar o vetor de intramodalidade com base em uma característica numérica associada com a interação entre os pares de dados nos dados de treinamento para a primeira modalidade, e

em que a informação na representação de modalidade conjunta é determinada com base, pelo menos em parte, no vetor de intramodalidade dimensionado.

49. Sistema de computador, de acordo com a reivindicação 47, caracterizado pelo fato de que o método ainda compreende:

concatenar cada um de um primeiro vetor de característica de saída a partir do primeiro codificador e um segundo vetor de

característica de saída a partir do segundo codificador com o vetor de intramodalidade para produzir primeiro e segundo vetores de características concatenadas; e

computar um vetor de representação conjunta dentro da representação de modalidade conjunta usando o primeiro e o segundo vetores de características concatenadas.

50. Sistema de computador, de acordo com a reivindicação 47, caracterizado pelo fato de que o método ainda compreende:

computar um vetor de característica conjunta usando um primeiro vetor de característica de saída a partir do primeiro codificador e um segundo vetor de característica de saída a partir do segundo codificador; e

concatenar o vetor de característica conjunta com o vetor de intramodalidade para produzir um vetor de representação conjunta dentro da representação de modalidade conjunta.

51. Sistema de computador, de acordo com a reivindicação 45, caracterizado pelo fato de que os primeiro e segundo codificadores e os primeiro e segundo decodificadores são configurados para processar dados a partir da primeira modalidade, e em que o treinamento ainda compreende:

proporcionar como entrada para o primeiro codificador um primeiro dos vetores de incorporação de primeira modalidade;

proporcionar como entrada para o segundo codificador um segundo dos vetores de incorporação de primeira modalidade; e

computar um vetor de representação conjunta na representação de modalidade conjunta com base em um primeiro vetor de característica de saída a partir do primeiro codificador, um segundo vetor de característica de saída a partir do segundo codificador, e o vetor de intramodalidade;

proporcionar um vetor de representação conjunta como

entrada aos primeiro e segundo decodificadores para produzir primeiro e segundo vetores decodificados; e

estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores com base em um de primeiro e de segundo dos vetores de incorporação de primeira modalidade e os primeiro e segundo vetores decodificados.

52. Sistema de computador, de acordo com a reivindicação 51, caracterizado pelo fato de que estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores compreende usar a função de perda de amostragem negativa.

53. Sistema de computador, de acordo com a reivindicação 52, caracterizado pelo fato de que o modelo estatístico ainda compreende primeira e segunda camadas de incorporação, e em que treinar ainda compreende estimar valores para parâmetros das primeira e segunda camadas de incorporação usando a função de perda de amostragem negativa.

54. Sistema de computador, de acordo com a reivindicação 52, caracterizado pelo fato de que o método ainda compreende:

repetir o treinamento do modelo estatístico para cada um de uma pluralidade de links entre pares de dados nos dados de treinamento para a primeira modalidade.

55. Sistema de computador, de acordo com a reivindicação 45, caracterizado pelo fato de que o método ainda compreende:

criar um vetor de intermodalidade descrevendo um link entre os dados de treinamento para a primeira modalidade e os dados de treinamento para a segunda modalidade, e

em que a informação na representação de modalidade conjunta é determinada com base, pelo menos em parte, no vetor de intermodalidade.

56. Sistema de computador, de acordo com a reivindicação 55, caracterizado pelo fato de que ainda compreende:

dimensionar o vetor de intermodalidade com base em uma característica numérica associada com o link entre os pares de dados nos dados de treinamento para a primeira modalidade, e

em que a informação na representação de modalidade conjunta é determinada com base, pelo menos em parte, no vetor de intermodalidade dimensionado.

57. Sistema de computador, de acordo com a reivindicação 55, caracterizado pelo fato de que o método ainda compreende:

concatenar cada um de um primeiro vetor de característica de saída a partir do primeiro codificador e um segundo vetor de característica de saída a partir do segundo codificador com o vetor de intermodalidade para produzir primeiro e segundo vetores de características concatenadas; e

computar um vetor de representação conjunta dentro da representação de modalidade conjunta usando o primeiro e o segundo vetores de características concatenadas.

58. Sistema de computador, de acordo com a reivindicação 55, caracterizado pelo fato de que o método ainda compreende:

computar um vetor de característica conjunta usando um primeiro vetor de característica de saída a partir do primeiro codificador e um segundo vetor de característica de saída a partir do segundo codificador; e

concatenar o vetor de característica conjunta com o vetor de intermodalidade para produzir um vetor de representação conjunta dentro da representação de modalidade conjunta.

59. Sistema de computador, de acordo com a reivindicação 55, caracterizado pelo fato de que o primeiro codificador e o primeiro decodificador são configurados para processar dados a partir da

primeira modalidade e o segundo codificador e o segundo decodificador são configurados para processar dados a partir da segunda modalidade, e em que o treinamento ainda compreende:

- proporcionar como entrada para o primeiro codificador dos vetores de incorporação de primeira modalidade;

- proporcionar como entrada para o segundo codificador vetores de incorporação de segunda modalidade; e

- computar um vetor de representação conjunta na representação de modalidade conjunta com base em um primeiro vetor de característica de saída a partir do primeiro codificador, um segundo vetor de característica de saída a partir do segundo codificador, e o vetor de intermodalidade;

- proporcionar um vetor de representação conjunta como entrada aos primeiro e segundo decodificadores para produzir primeiro e segundo vetores decodificados; e

- estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores com base nos vetores de incorporação de primeira modalidade e um dos vetores de incorporação de segunda modalidade e os primeiro e segundo vetores decodificados.

60. Sistema de computador, de acordo com a reivindicação 59, caracterizado pelo fato de que estimar valores para parâmetros dos primeiro e segundo codificadores e dos primeiro e segundo decodificadores compreende usar a função de perda de amostragem negativa.

61. Sistema de computador, de acordo com a reivindicação 59, caracterizado pelo fato de que o método ainda compreende:

- repetir o treinamento do modelo estatístico para cada um de uma pluralidade de interações entre os dados de treinamento para a primeira modalidade e os dados de treinamento para a segunda modalidade.

62. Sistema de computador, de acordo com a reivindicação 59, caracterizado pelo fato de que o método ainda compreende:

inicializar, antes do treinamento, os valores dos parâmetros para o primeiro codificador, em que a inicialização é realizada com base em resultados de treinamento do primeiro codificador usando uma técnica de aprendizado autossupervisionada.

63. Sistema de computador, de acordo com a reivindicação 62, caracterizado pelo fato de que os resultados de treinamento do primeiro codificador usando uma técnica de aprendizado autossupervisionada compreende os resultados de treinamento do primeiro codificador usando dados de treinamento a partir de apenas a primeira modalidade.

64. Sistema de computador, de acordo com a reivindicação 62, caracterizado pelo fato de que os resultados de treinamento do primeiro codificador usando uma técnica de aprendizado autossupervisionada compreende os resultados de treinamento do primeiro codificador com dados de treinamento tendo uma terceira modalidade diferente a partir da segunda modalidade.

65. Sistema de computador, de acordo com a reivindicação 62, caracterizado pelo fato de que o método ainda compreende:

inicializar, antes do treinamento, os valores dos parâmetros para o segundo codificador, em que a inicialização é realizada com base em resultados de treinamento do segundo codificador usando uma técnica de aprendizado autossupervisionada.

66. Sistema de computador, de acordo com a reivindicação 45, caracterizado pelo fato de que criar a primeira modalidade de vetores de incorporação compreende:

definir, para cada dado nos dados de treinamento para a primeira modalidade, um vetor one-hot tendo um comprimento V_1 ; e

multiplicar cada um dos vetores one-hot de comprimento V_1

por uma primeira matriz de incorporação tendo dimensionalidade $V_1 \times E$, onde $E < V_1$, e em que E é um comprimento de cada um dos vetores de incorporação de primeira modalidade.

67. Sistema de computador, de acordo com a reivindicação 66, caracterizado pelo fato de que criar os vetores de incorporação de segunda modalidade compreende:

definir, para cada dado nos dados de treinamento para a segunda modalidade, um vetor one-hot tendo um comprimento V_2 ; e

multiplicar cada um dos vetores one-hot de comprimento V_2 por uma matriz de incorporação tendo dimensionalidade $V_2 \times E$, onde $E < V_2$, e em que E é um comprimento de cada dos vetores de incorporação de segunda modalidade.

68. Sistema de computador, de acordo com a reivindicação 67, caracterizado pelo fato de que cada um dos primeiro e segundo codificadores compreende uma camada de entrada tendo E entradas e uma camada de saída tendo R saídas, onde $R > E$.

69. Sistema de computador, de acordo com a reivindicação 68, caracterizado pelo fato de que cada um dos primeiro e segundo codificadores compreende pelo menos uma camada oculta.

70. Sistema de computador, de acordo com a reivindicação 44, caracterizado pelo fato de que cada um dos primeiro e segundo codificadores compreende uma rede neural.

71. Sistema de computador, de acordo com a reivindicação 44, caracterizado pelo fato de que os dados de treinamento ainda incluem dados de treinamento para uma terceira modalidade diferente a partir da primeira modalidade e da segunda modalidade, em que o modelo estatístico ainda compreende um terceiro codificador e um terceiro decodificador, e

em que treinar o modelo estatístico ainda compreende estimar valores para parâmetros do terceiro codificador e do terceiro

decodificador usando uma técnica de aprendizado autossupervisionada, a terceira modalidade de vetores de entrada, e informação descrevendo pelo menos um link entre dados de treinamento para a terceira modalidade e dados de treinamento para a primeira ou segunda modalidades.

72. Sistema de computador, de acordo com a reivindicação 71, caracterizado pelo fato de que ainda compreende criar terceira modalidade de vetores de incorporação com base nos dados de treinamento para a terceira modalidade, e em que treinar o modelo estatístico ainda compreende proporcionar como entrada para o terceiro codificador, uma primeira da terceira modalidade de vetores de incorporação.

73. Sistema de computador, caracterizado pelo fato de que compreende:

pelo menos um computador processador; e

pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador realizam um método de prever uma associação entre dados de entrada em uma primeira modalidade e dados em uma segunda modalidade usando um modelo estatístico treinado para representar interações entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplando a pluralidade de codificadores e decodificadores, o método compreendendo:

selecionar, com base na primeira modalidade e na segunda modalidade, um par de codificador/decodificador ou um par de codificadores, a partir de entre a pluralidade de codificadores e decodificadores; e

processar os dados de entrada com a representação de modalidade conjunta e o par de codificador/decodificador selecionado ou o par de codificadores para prever a associação entre os dados de entrada e os dados na segunda modalidade.

74. Sistema de computador, de acordo com a reivindicação 73, caracterizado pelo fato de que o método ainda compreende selecionar um codificador treinado para processar dados para a primeira modalidade e um decodificador treinado para processar dados para a segunda modalidade.

75. Sistema de computador, de acordo com a reivindicação 74, caracterizado pelo fato de que o método ainda compreende prever a associação entre os dados de entrada e os dados na segunda modalidade em um espaço de representação para a segunda modalidade.

76. Sistema de computador, de acordo com a reivindicação 75, caracterizado pelo fato de que o método ainda compreende:

emitir a partir do decodificador treinado para processar dados para a segunda modalidade, um vetor de saída no espaço de representação para a segunda modalidade;

projetar os dados a partir da segunda modalidade em um espaço de representação para a segunda modalidade para produzir uma pluralidade de vetores projetados; e

prever a associação entre os dados de entrada e os dados na segunda modalidade com base em uma comparação do vetor de saída e os vetores projetados no espaço de representação para a segunda modalidade.

77. Sistema de computador, de acordo com a reivindicação 76, caracterizado pelo fato de que o método ainda compreende:

calcular uma distância entre o vetor de saída e cada da pluralidade de vetores projetados; e

prever a associação com base nas distâncias calculadas.

78. Sistema de computador, de acordo com a reivindicação 77, caracterizado pelo fato de que calcular uma distância compreende calcular a distância Euclideana.

79. Sistema de computador, de acordo com a reivindicação 73, caracterizado pelo fato de que o método ainda compreende selecionar um primeiro codificador treinado para processar dados para a primeira modalidade e um segundo codificador treinado para processar dados para a segunda modalidade.

80. Sistema de computador, de acordo com a reivindicação 79, caracterizado pelo fato de que o método ainda compreende prever a associação entre os dados de entrada e os dados na segunda modalidade em um espaço de representação latente associado com a representação de modalidade conjunta.

81. Sistema de computador, de acordo com a reivindicação 80, caracterizado pelo fato de que ainda compreende:

proporcionar como entrada para o primeiro codificador, os dados de entrada para produzir um vetor de característica de primeira modalidade no espaço de representação latente;

proporcionar como entrada para o segundo codificador, os dados para a segunda modalidade para produzir uma pluralidade de vetores de característica de segunda modalidade no espaço de representação latente; e

prever a associação entre os dados de entrada e os dados na segunda modalidade com base em uma comparação de um vetor de característica de primeira modalidade e uma pluralidade de vetores de característica de segunda modalidade no espaço de representação latente.

82. Sistema de computador, de acordo com a reivindicação 81, caracterizado pelo fato de que ainda compreende:

calcular uma distância entre um vetor de característica de primeira modalidade e cada um da pluralidade de vetores de característica de segunda modalidade; e

prever a associação com base nas distâncias calculadas.

83. Sistema de computador, de acordo com a reivindicação 82, caracterizado pelo fato de que calcular uma distância compreende calcular a distância Euclideana.

84. Sistema de computador, caracterizado pelo fato de que compreende:

pelo menos um computador processador; e

pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador, realizam um método de prever associações entre dados em uma primeira modalidade e dados em uma segunda modalidade usando um modelo estatístico treinado para representar links entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade diferente a partir da primeira modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplando a pluralidade de codificadores e decodificadores, o método compreendendo:

mapear os dados na primeira modalidade e os dados na segunda modalidade dentro de um espaço de representação comum dentro do modelo estatístico;

acessar um classificador estatístico treinado usando dados rotulados, em que os dados rotulados descrevem associações entre dados nas primeira e segunda modalidades; e

prever associações entre os dados na primeira modalidade

e os dados na segunda modalidade mapeados dentro do espaço de representação comum usando o classificador estatístico treinado.

85. Sistema de computador, de acordo com a reivindicação 84, caracterizado pelo fato de que mapear os dados na primeira modalidade e os dados na segunda modalidade em um espaço de representação comum compreende mapear os dados em um espaço de representação de modalidade conjunta do modelo estatístico.

86. Sistema de computador, de acordo com a reivindicação 84, caracterizado pelo fato de que mapear os dados na primeira modalidade e os dados na segunda modalidade em um espaço de representação comum compreende mapear os dados em um espaço de representação específico de modalidade for a primeira modalidade ou a segunda modalidade.

87. Método para treinar um modelo estatístico para representar associações entre dados de fármaco, dados de gene, e dados de doença, caracterizado pelo fato de que o método compreende:

acessar dados de treinamento incluindo dados de treinamento de gene, dados de treinamento de fármaco e dados de treinamento de doença;

treinar o modelo estatístico, o modelo estatístico compreendendo uma pluralidade de codificadores incluindo um codificador de gene, um codificador de fármaco e um codificador de doença, uma pluralidade de decodificadores incluindo um decodificador de gene, um decodificador de fármaco e um decodificador de doença, e uma representação conjunta acoplando a pluralidade de codificadores a uma pluralidade de decodificadores, em que a representação conjunta descreve interações entre os dados de treinamento, o treinamento compreendendo:

estimar valores para parâmetros do codificador de gene e do decodificador de gene usando uma técnica de aprendizado

autossupervisionada, os dados de treinamento de gene, e informação descrevendo interações entre pares de dados nos dados de treinamento de gene;

estimar valores para parâmetros do codificador de gene, o decodificador de gene, o codificador de fármaco, e o decodificador de fármaco usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de fármaco, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de fármaco; e

estimar valores para parâmetros do codificador de gene, o decodificador de gene, o codificador de doença, e o decodificador de doença usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de doença, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de doença; e

armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de gene, o decodificador de gene, o codificador de fármaco, o decodificador de fármaco, o codificador de doença, e o decodificador de doença do modelo estatístico.

88. Método, de acordo com a reivindicação 87, caracterizado pelo fato de que ainda compreende:

criar vetores de incorporação de modalidade de gene com base nos dados de treinamento de gene, em que o treinamento ainda compreende proporcionar como entrada para os codificadores de gene, os vetores de incorporação de gene.

89. Método, de acordo com a reivindicação 87, caracterizado pelo fato de que a informação descrevendo interações

entre pares de dados nos dados de treinamento de gene compreende informação de múltiplos tipos de interações incluindo informação sobre interações gene-gene, informação sobre covariações de gene-gene, e informação sobre regulação de gene-gene, e em que o treinamento ainda compreende estimar valores para parâmetros do codificador de gene e do decodificador de gene separadamente para cada um dos múltiplos tipos de interações.

90. Método, de acordo com a reivindicação 87, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de fármaco compreende informação sobre múltiplos tipos de interações incluindo informação sobre regulação para mais de fármaco-gene, informação sobre regulação para menos de fármaco-gene, e informação sobre ligação de fármaco-gene, e em que o treinamento ainda compreende estimar valores para parâmetros dos codificadores de gene e fármaco e dos decodificadores de gene e fármaco separadamente para cada um dos múltiplos tipos de interações.

91. Método, de acordo com a reivindicação 87, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de doença compreende informação sobre múltiplos tipos de interações incluindo informação sobre regulação para mais de gene-doença, informação sobre regulação para menos de gene-doença, e informação sobre associações de gene-doença, e em que o treinamento ainda compreende estimar valores para parâmetros dos codificadores de gene e doença e decodificadores de gene e doença separadamente para cada um dos múltiplos tipos de interações.

92. Método, de acordo com a reivindicação 87,

caracterizado pelo fato de que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de fármaco, do decodificador de fármaco, do codificador de doença e do decodificador de doença usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de fármaco e os dados de treinamento de doença, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de fármaco e os elementos de dados nos dados de treinamento de doença; e

armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de fármaco, do decodificador de fármaco, do codificador de doença e do decodificador de doença do modelo estatístico.

93. Método, de acordo com a reivindicação 92, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de fármaco e os elementos de dados nos dados de treinamento de doença compreende informação sobre tratamento de fármaco-doença.

94. Método, de acordo com a reivindicação 87, caracterizado pelo fato de que os dados de treinamento ainda compreendem dados de treinamento de classe de fármaco,

em que a pluralidade de codificadores ainda compreende um codificador de classe de fármaco,

em que a pluralidade de decodificadores ainda compreende um decodificador de classe de fármaco; e

em que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de fármaco, do decodificador de fármaco, do codificador de classe de fármaco e do decodificador de classe de fármaco usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de fármaco e os dados

de treinamento de classe de fármaco, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de fármaco e os elementos de dados nos dados de treinamento de classe de fármaco, e

em que o método ainda compreende armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de fármaco, do decodificador de fármaco, do codificador de classe de fármaco e do decodificador de classe de fármaco do modelo estatístico.

95. Método, de acordo com a reivindicação 94, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de fármaco e os elementos de dados nos dados de treinamento de classe de fármaco compreende informação sobre inclusão de classe fármaco-fármaco.

96. Método, de acordo com a reivindicação 87, caracterizado pelo fato de que os dados de treinamento ainda compreendem dados de treinamento de via biológica,

em que a pluralidade de codificadores ainda compreende um codificador de via,

em que a pluralidade de decodificadores ainda compreende um decodificador de via; e

em que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de gene, do decodificador de gene, do codificador de via e do decodificador de via usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de via biológica, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de via biológica, e

em que o método ainda compreende armazenar informação

especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de gene, do decodificador de gene, do codificador de via e do decodificador de via do modelo estatístico.

97. Método, de acordo com a reivindicação 96, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de via biológica compreende informação sobre participação gene-via.

98. Método, de acordo com a reivindicação 87, caracterizado pelo fato de que os dados de treinamento ainda compreendem dados de treinamento de anatomia,

em que a pluralidade de codificadores ainda compreende um codificador de anatomia,

em que a pluralidade de decodificadores ainda compreende um decodificador de anatomia; e

em que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de doença, do decodificador de doença, do codificador de anatomia e do decodificador de anatomia usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de doença e os dados de treinamento de anatomia, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de doença e os elementos de dados nos dados de treinamento de anatomia, e

em que o método ainda compreende armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de doença, do decodificador de doença, do codificador de anatomia e do decodificador de anatomia do modelo estatístico.

99. Método, de acordo com a reivindicação 98,

caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de doença e os elementos de dados nos dados de treinamento de anatomia compreende informação sobre a localização de doença-anatomia.

100. Método, de acordo com a reivindicação 98, caracterizado pelo fato de que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de gene, do decodificador de gene, do codificador de anatomia e do decodificador de anatomia usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de anatomia, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de anatomia, e

em que o método ainda compreende armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de gene, do decodificador de gene, do codificador de anatomia e do decodificador de anatomia do modelo estatístico.

101. Método, de acordo com a reivindicação 100, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de anatomia compreende informação sobre múltiplos tipos de interações incluindo informação sobre regulação para mais de gene-anatomia, informação sobre regulação para menos de gene-anatomia, e informação sobre expressão de gene-anatomia, e em que o treinamento ainda compreende estimar valores para parâmetros dos codificadores de gene e anatomia e dos decodificadores de gene e anatomia separadamente para cada um dos múltiplos tipos de interações.

102. Sistema de computador, caracterizado pelo fato de que compreende:

pelo menos um computador processador; e

pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador realizam um método de treinar um modelo estatístico para representar associações entre dados de fármaco, dados de gene, e dados de doença, o método compreendendo:

acessar dados de treinamento incluindo dados de treinamento de gene, dados de treinamento de fármaco e dados de treinamento de doença;

treinar o modelo estatístico, o modelo estatístico compreendendo uma pluralidade de codificadores incluindo um codificador de gene, um codificador de fármaco e um codificador de doença, uma pluralidade de decodificadores incluindo um decodificador de gene, um decodificador de fármaco, e um decodificador de doença, e uma representação conjunta acoplando a pluralidade de codificadores a uma pluralidade de decodificadores, em que a representação conjunta descreve interações entre os dados de treinamento, o treinamento compreendendo:

estimar valores para parâmetros do codificador de gene e do decodificador de gene usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene, e informação descrevendo interações entre pares de dados nos dados de treinamento de gene;

estimar valores para parâmetros do codificador de gene, o decodificador de gene, o codificador de fármaco, e o decodificador de fármaco usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de fármaco, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de fármaco; e

estimar valores para parâmetros do codificador de gene, o decodificador de gene, o codificador de doença, e o decodificador de doença usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de doença, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de doença; e

armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de gene, o decodificador de gene, o codificador de fármaco, o decodificador de fármaco, o codificador de doença, e o decodificador de doença do modelo estatístico.

103. Sistema de computador, de acordo com a reivindicação 102, caracterizado pelo fato de que o método ainda compreende:

criar de vetores de incorporação de modalidade de gene com base nos dados de treinamento de gene, em que o treinamento ainda compreende proporcionar como entrada para o codificadores de gene, o vetores de incorporação de gene.

104. Sistema de computador, de acordo com a reivindicação 102, caracterizado pelo fato de que a informação descrevendo interações entre pares de dados nos dados de treinamento de gene compreende informação sobre múltiplos tipos de interações incluindo informação sobre interações gene-gene, informação sobre covariações de gene-gene, e informação sobre regulação de gene-gene, e em que o treinamento ainda compreende estimar valores para parâmetros do codificador de gene e do decodificador de gene separadamente para cada um dos múltiplos tipos de interações.

105. Sistema de computador, de acordo com a reivindicação 102, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de

gene e os elementos de dados nos dados de treinamento de fármaco compreende informação sobre múltiplos tipos de interações incluindo informação sobre regulação para mais de fármaco-gene, informação sobre regulação para menos de fármaco-gene, e informação sobre ligação de fármaco-gene, e em que o treinamento ainda compreende estimar valores para parâmetros dos codificadores de gene e fármaco e dos decodificadores de gene e fármaco separadamente para cada um dos múltiplos tipos de interações.

106. Sistema de computador, de acordo com a reivindicação 102, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de doença compreende informação sobre múltiplos tipos de interações incluindo informação sobre regulação para mais de gene-doença, informação sobre regulação para menos de gene-doença, e informação sobre associações de gene-doença, e em que o treinamento ainda compreende estimar valores para parâmetros dos codificadores de gene e doença e decodificadores de gene e doença separadamente para cada um dos múltiplos tipos de interações.

107. Sistema de computador, de acordo com a reivindicação 102, caracterizado pelo fato de que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de fármaco, do decodificador de fármaco, do codificador de doença e do decodificador de doença usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de fármaco e os dados de treinamento de doença, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de fármaco e os elementos de dados nos dados de treinamento de doença; e

armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para

parâmetros do codificador de fármaco, do decodificador de fármaco, do codificador de doença e do decodificador de doença do modelo estatístico.

108. Sistema de computador, de acordo com a reivindicação 107, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de fármaco e os elementos de dados nos dados de treinamento de doença compreende informação sobre tratamento de fármaco-doença.

109. Sistema de computador, de acordo com a reivindicação 102, caracterizado pelo fato de que os dados de treinamento ainda compreendem dados de treinamento de classe de fármaco,

em que a pluralidade de codificadores ainda compreende um codificador de classe de fármaco,

em que a pluralidade de decodificadores ainda compreende um decodificador de classe de fármaco; e

em que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de fármaco, do decodificador de fármaco, do codificador de classe de fármaco e do decodificador de classe de fármaco usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de fármaco e os dados de treinamento de classe de fármaco, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de fármaco e os elementos de dados nos dados de treinamento de classe de fármaco, e

em que o método ainda compreende armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de fármaco, do decodificador de fármaco, do codificador de classe de fármaco e do decodificador de classe de fármaco do modelo estatístico.

110. Sistema de computador, de acordo com a reivindicação

109, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de fármaco e os elementos de dados nos dados de treinamento de classe de fármaco compreende informação sobre inclusão de classe fármaco-fármaco.

111. Sistema de computador, de acordo com a reivindicação 102, caracterizado pelo fato de que os dados de treinamento ainda compreendem dados de treinamento de via biológica,

em que a pluralidade de codificadores ainda compreende um codificador de via,

em que a pluralidade de decodificadores ainda compreende um decodificador de via; e

em que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de gene, do decodificador de gene, do codificador de via e do decodificador de via usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de gene e os dados de treinamento de via biológica, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de via biológica, e

em que o método ainda compreende armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de gene, do decodificador de gene, do codificador de via e do decodificador de via do modelo estatístico.

112. Sistema de computador, de acordo com a reivindicação 111, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de via biológica compreende informação sobre participação gene-via.

113. Sistema de computador, de acordo com a reivindicação 102, caracterizado pelo fato de que os dados de treinamento ainda compreendem dados de treinamento de anatomia,

em que a pluralidade de codificadores ainda compreende um codificador de anatomia,

em que a pluralidade de decodificadores ainda compreende um decodificador de anatomia; e

em que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de doença, do decodificador de doença, do codificador de anatomia e do decodificador de anatomia usando uma técnica de aprendizado autossupervisionada, os dados de treinamento de doença e os dados de treinamento de anatomia, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de doença e os elementos de dados nos dados de treinamento de anatomia, e

em que o método ainda compreende armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de doença, do decodificador de doença, do codificador de anatomia e do decodificador de anatomia do modelo estatístico.

114. Sistema de computador, de acordo com a reivindicação 113, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de doença e os elementos de dados nos dados de treinamento de anatomia compreende informação sobre a localização de doença-anatomia.

115. Sistema de computador, de acordo com a reivindicação 113, caracterizado pelo fato de que o treinamento ainda compreende:

estimar valores para parâmetros do codificador de gene, do decodificador de gene, do codificador de anatomia e do decodificador de anatomia usando uma técnica de aprendizado autossupervisionada,

os dados de treinamento de gene e os dados de treinamento de anatomia, e informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de anatomia, e

em que o método ainda compreende armazenar informação especificando o modelo estatístico pelo menos em parte por armazenar os valores estimados para parâmetros do codificador de gene, do decodificador de gene, do codificador de anatomia e do decodificador de anatomia do modelo estatístico.

116. Sistema de computador, de acordo com a reivindicação 115, caracterizado pelo fato de que a informação descrevendo interações entre os elementos de dados nos dados de treinamento de gene e os elementos de dados nos dados de treinamento de anatomia compreende informação sobre múltiplos tipos de interações incluindo informação sobre regulação para mais de gene-anatomia, informação sobre regulação para menos de gene-anatomia, e informação sobre expressão de gene-anatomia, e em que o treinamento ainda compreende estimar valores para parâmetros dos codificadores de gene e anatomia e dos decodificadores de gene e anatomia separadamente para cada um dos múltiplos tipos de interações.

117. Método para prever a nova indicação de doença para um determinado fármaco, caracterizado pelo fato de que o método compreende:

projetar a representação do determinado fármaco e representações de uma pluralidade de doenças dentro de um espaço de representação comum de um modelo estatístico treinado; e

prever a nova indicação de doença para o determinado fármaco com base em uma comparação da representação projetada do determinado fármaco e pelo menos uma das representações da pluralidade de doenças no espaço de representação comum.

118. Método, de acordo com a reivindicação 117, caracterizado pelo fato de que prever a nova indicação de doença compreende calcular uma distância entre a representação projetada do determinado fármaco e pelo menos uma das representações da pluralidade de doenças no espaço de representação comum e prever a nova indicação de doença com base na distância calculada.

119. Sistema de computador, caracterizado pelo fato de que compreende:

pelo menos um computador processador; e

pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador, realiza um método de prever a nova indicação de doença para um determinado fármaco, o método compreendendo:

projetar a representação do determinado fármaco e representações de uma pluralidade de doenças dentro de um espaço de representação comum de um modelo estatístico treinado; e

prever a nova indicação de doença para o determinado fármaco com base em uma comparação da representação projetada do determinado fármaco e pelo menos uma das representações da pluralidade de doenças no espaço de representação comum.

120. Sistema de computador, de acordo com a reivindicação 119, caracterizado pelo fato de que prever a nova indicação de doença compreende calcular uma distância entre a representação projetada do determinado fármaco e pelo menos uma das representações da pluralidade de doenças no espaço de representação comum e prever a nova indicação de doença com base na distância calculada.

121. Método de identificar indicação de doenças para um determinado fármaco, caracterizado pelo fato de que o método compreende:

proporcionar como entrada para um modelo estatístico, representações de uma pluralidade de fármacos e uma pluralidade de doenças; e

processar as representações da pluralidade de fármacos e a pluralidade de doenças usando um classificador supervisionado treinado para identificar uma probabilidade de que fármacos na pluralidade de fármacos serão eficazes no tratamento de doenças na pluralidade de doenças, o classificador supervisionado treinado com informação sobre pares de fármaco-doença aprovados pela Federal Drug Administration (FDA).

122. Sistema de computador, caracterizado pelo fato de que compreende:

pelo menos um computador processador; e

pelo menos um dispositivo de armazenamento codificado com uma pluralidade de instruções as quais, quando executadas por o pelo menos um computador processador, realiza um método de identificar indicação de doenças para um determinado fármaco, o método compreendendo:

proporcionar como entrada para um modelo estatístico, representações de uma pluralidade de fármacos e uma pluralidade de doenças; e

processar as representações da pluralidade de fármacos e a pluralidade de doenças usando um classificador supervisionado treinado para identificar uma probabilidade de que fármacos na pluralidade de fármacos serão eficazes no tratamento de doenças na pluralidade de doenças, o classificador supervisionado treinado com informação sobre pares de fármaco-doença aprovados pela Federal Drug Administration (FDA).

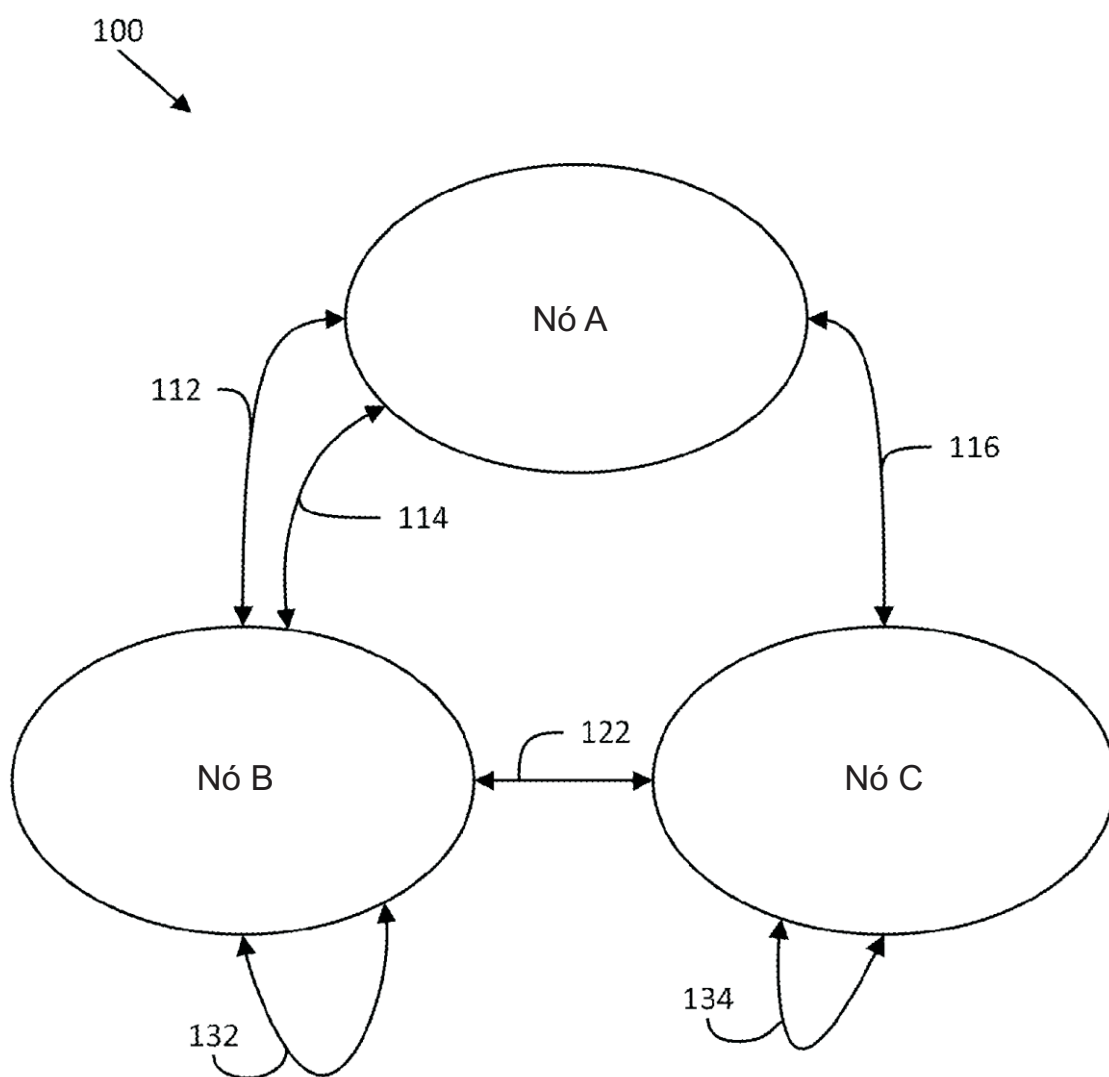


FIG. 1

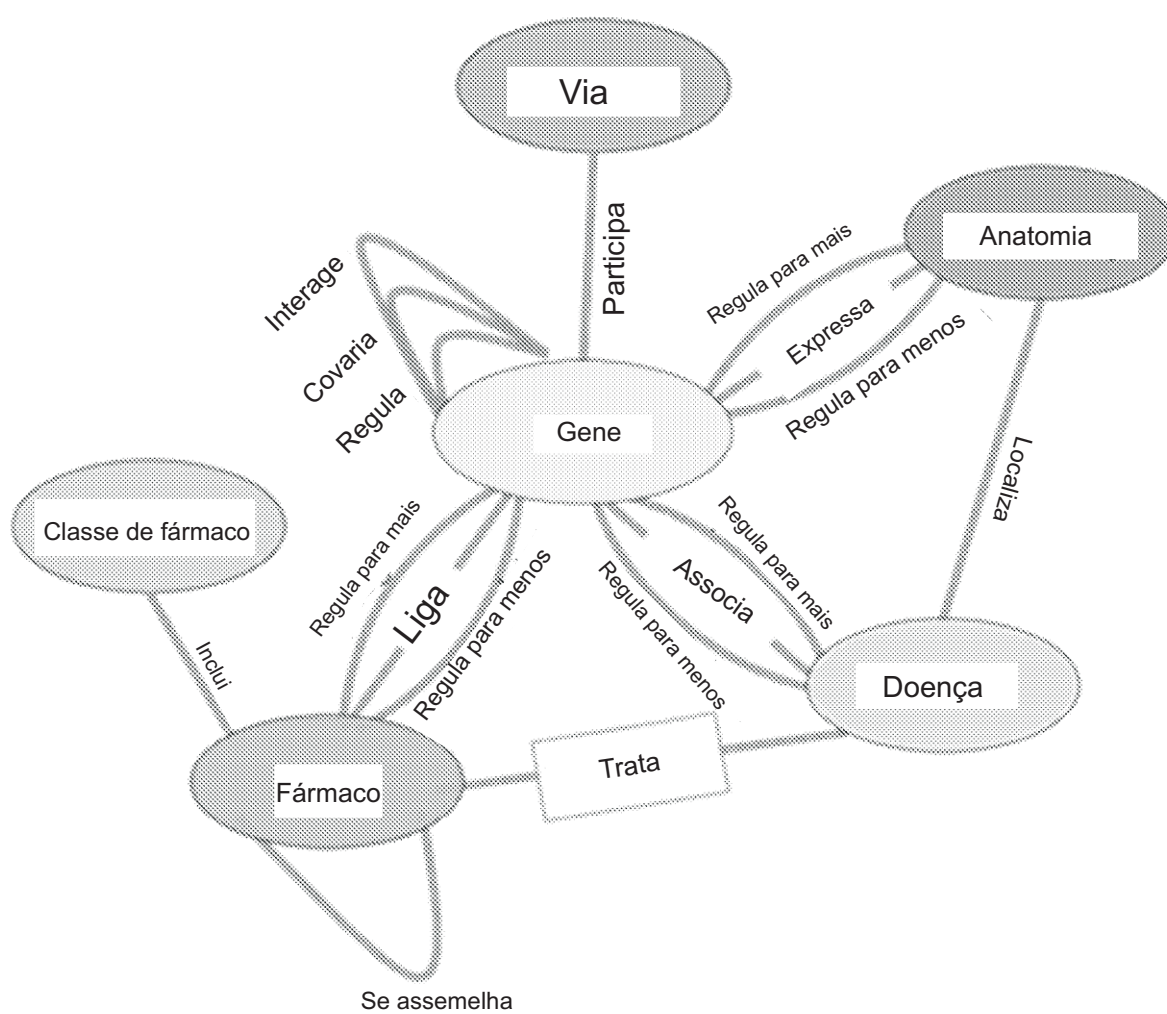


FIG. 2

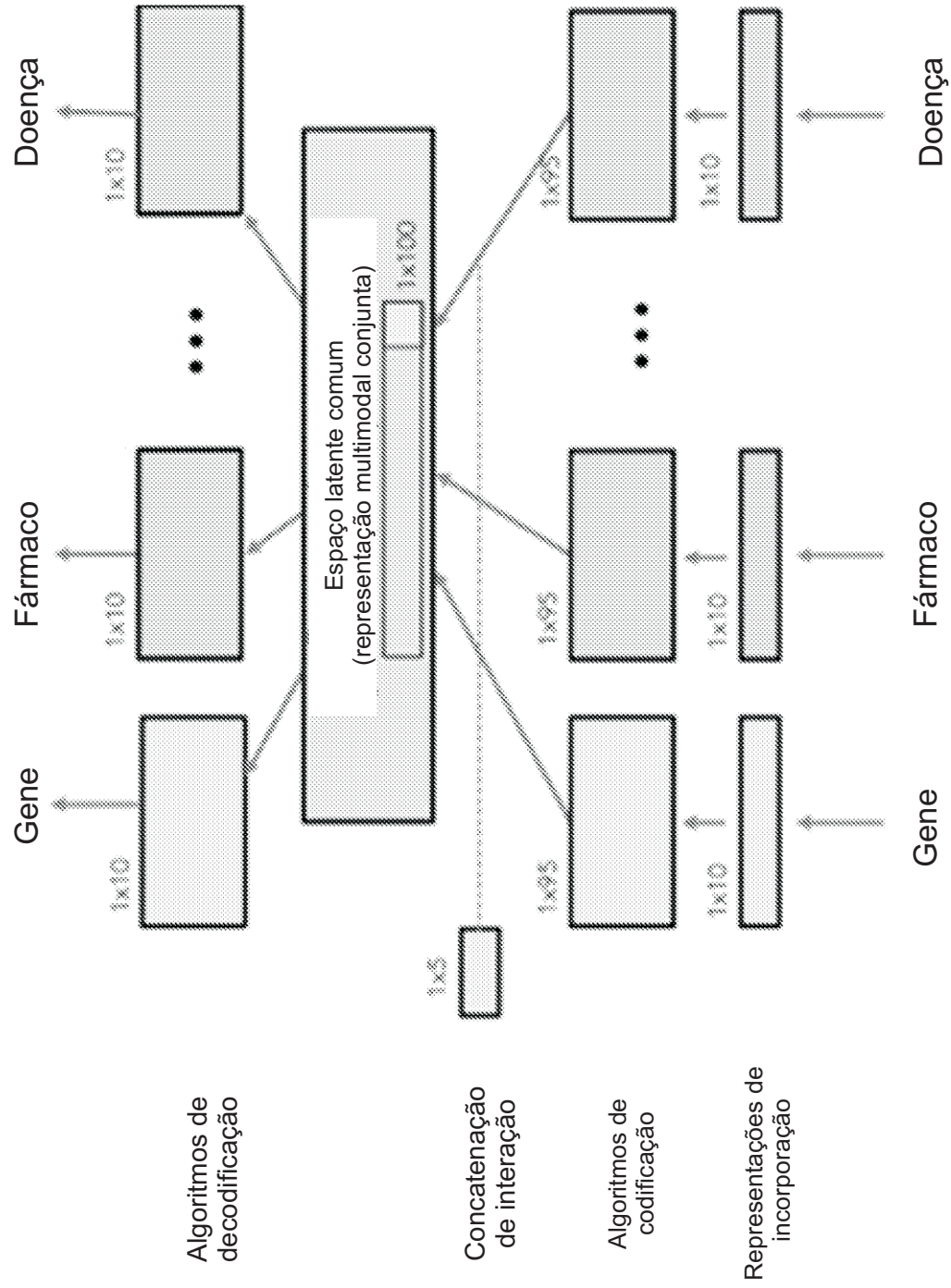


FIG. 3

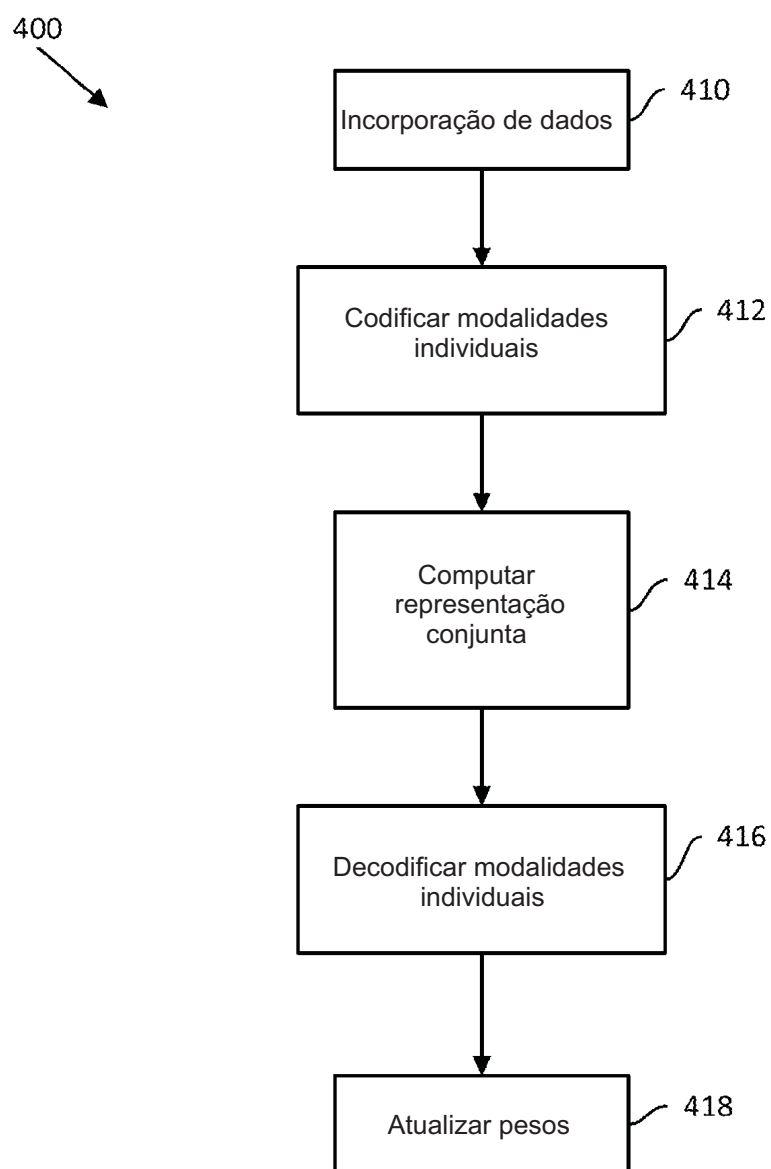


FIG. 4

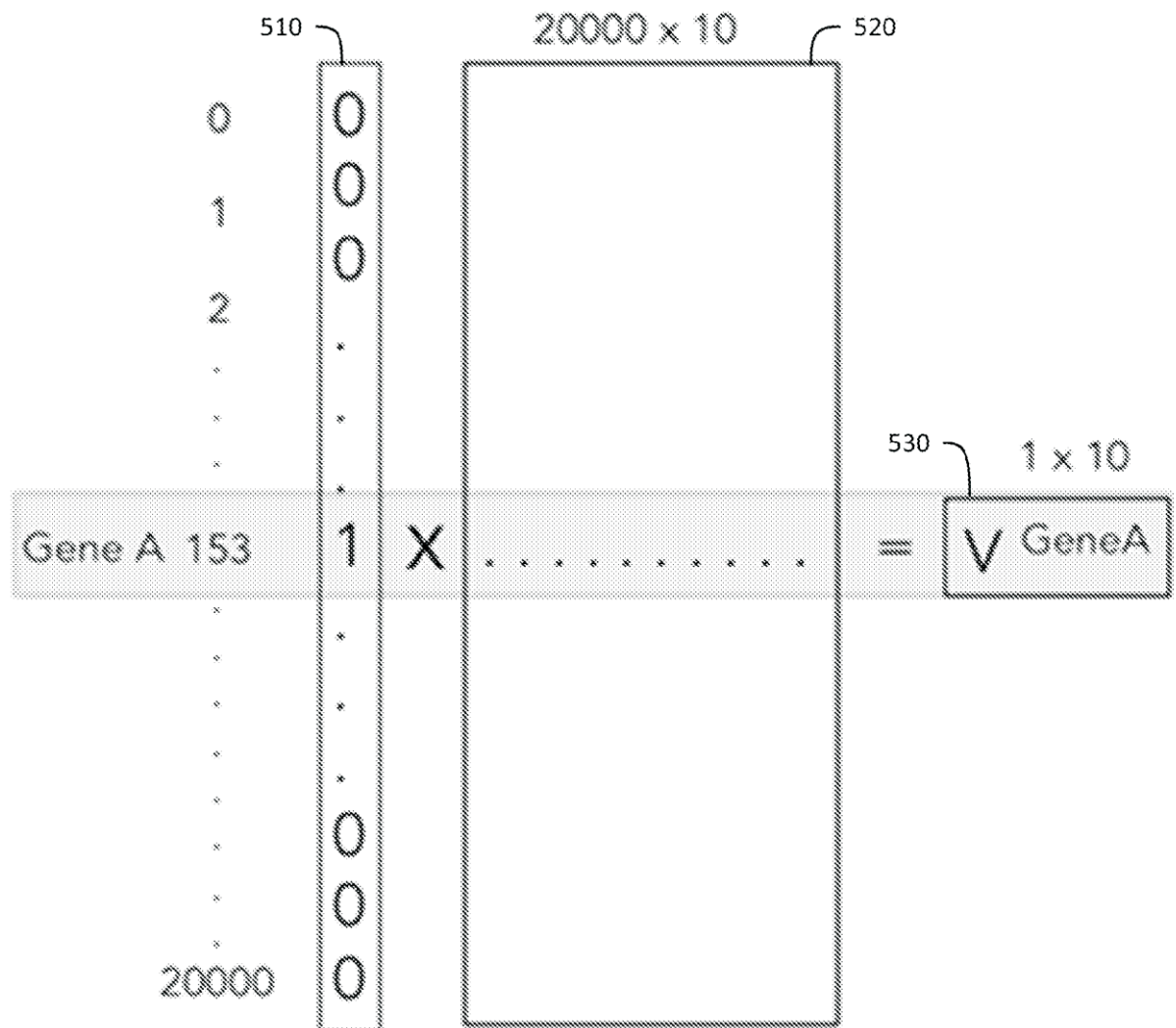


FIG. 5

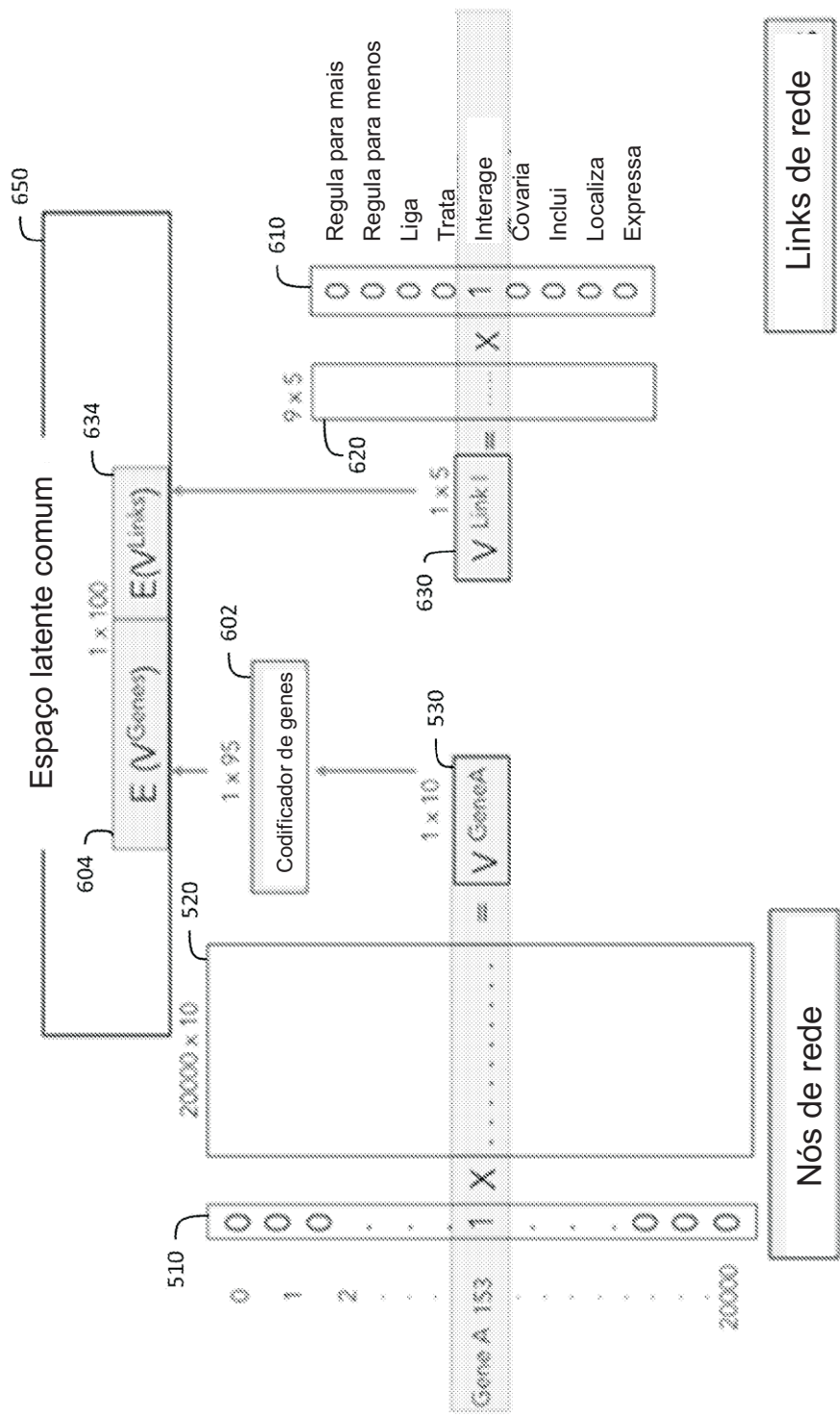


FIG. 6

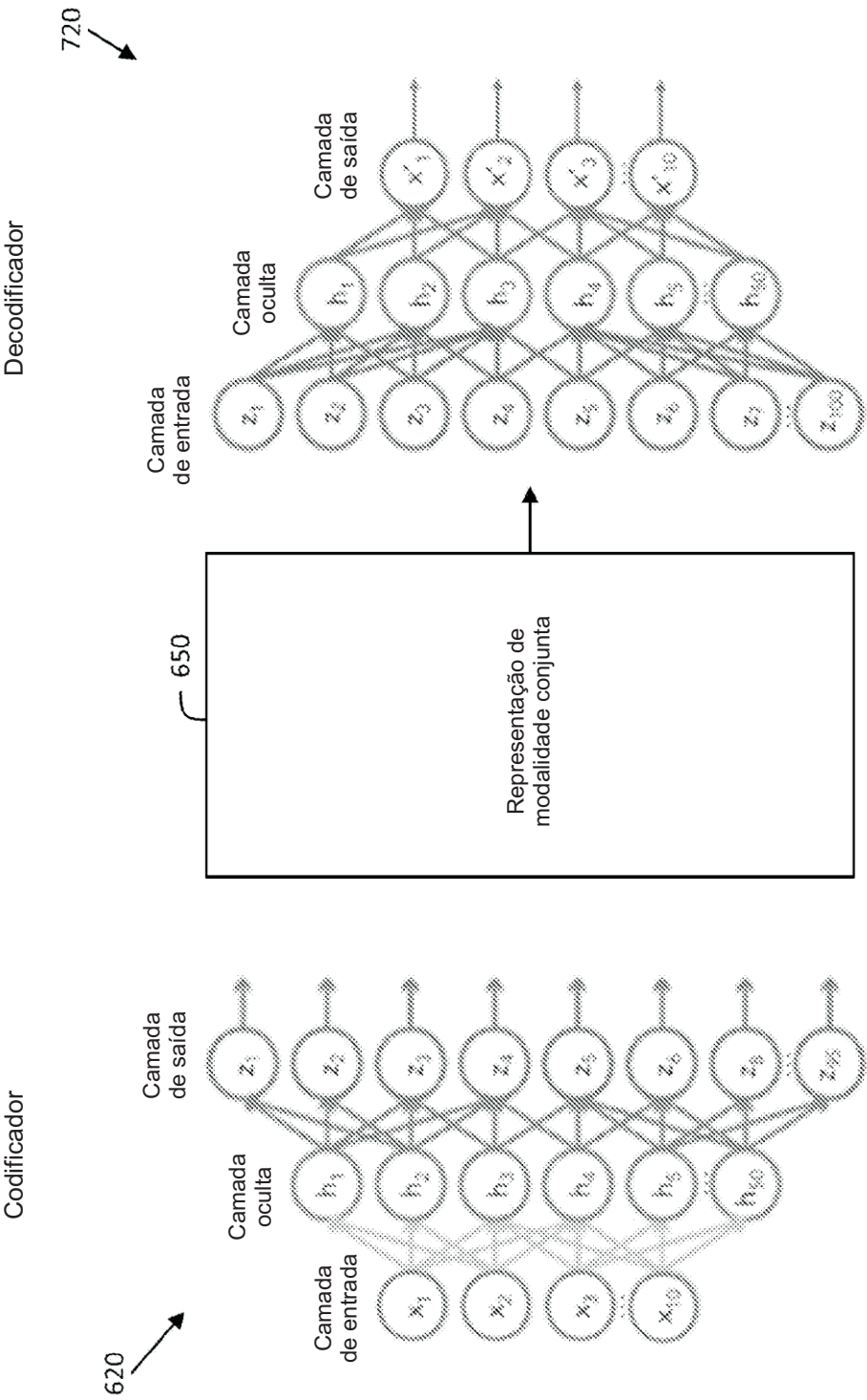


FIG. 7

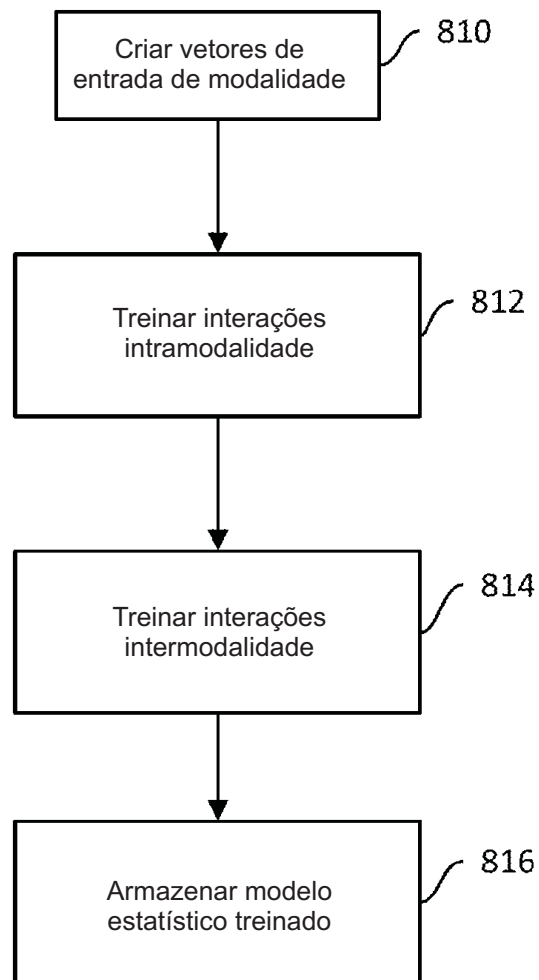
800


FIG. 8

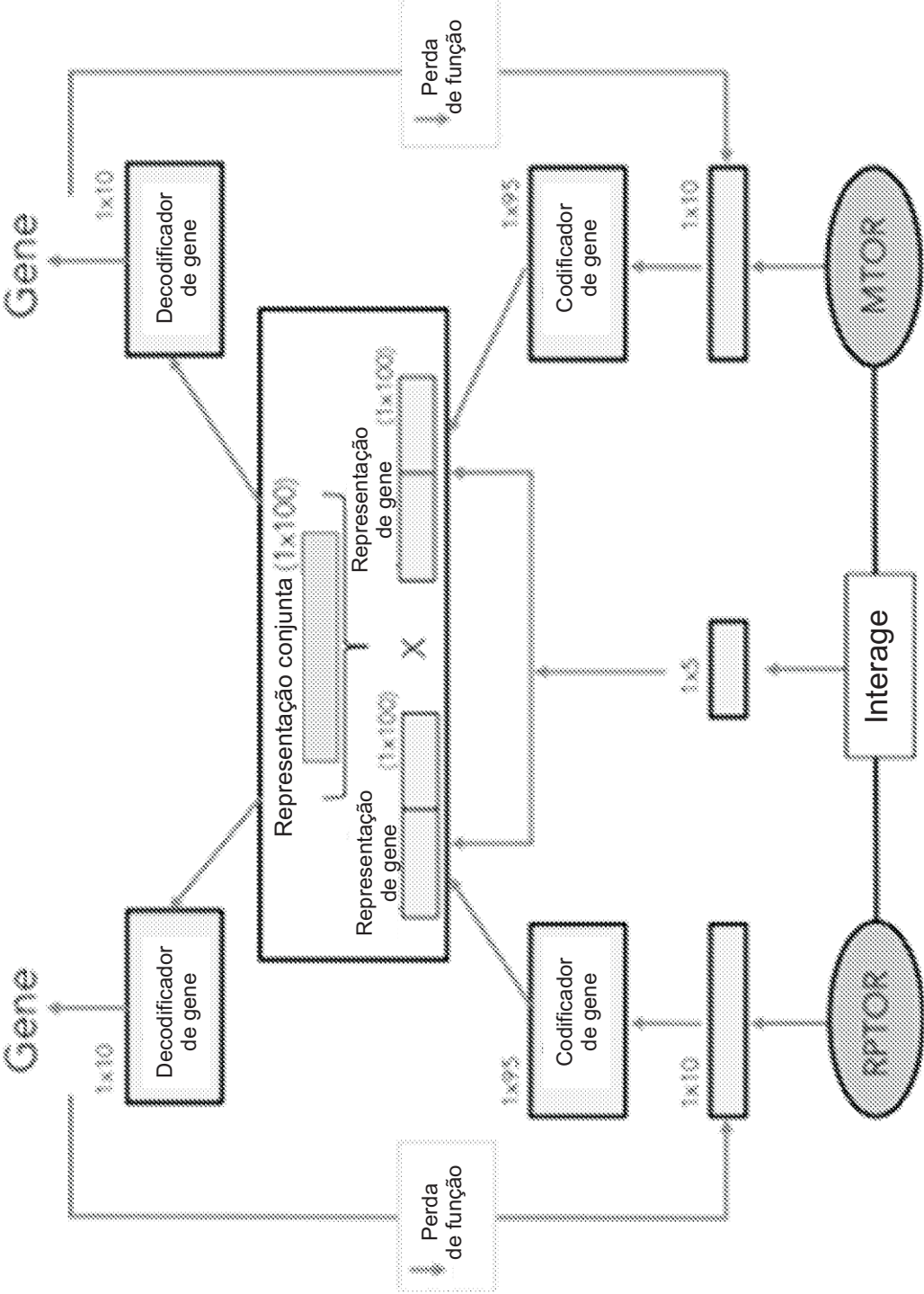


FIG. 9

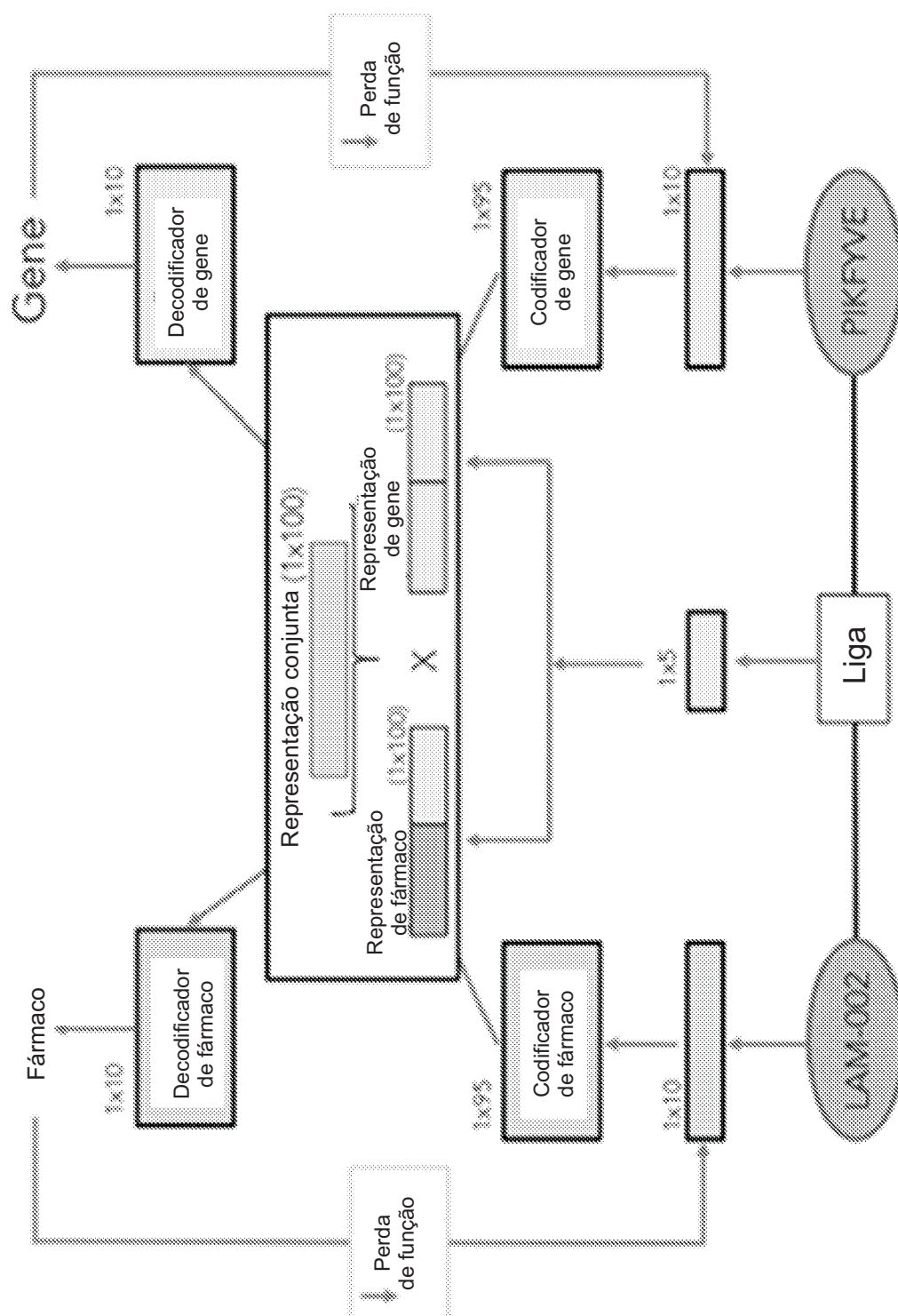


FIG. 10A

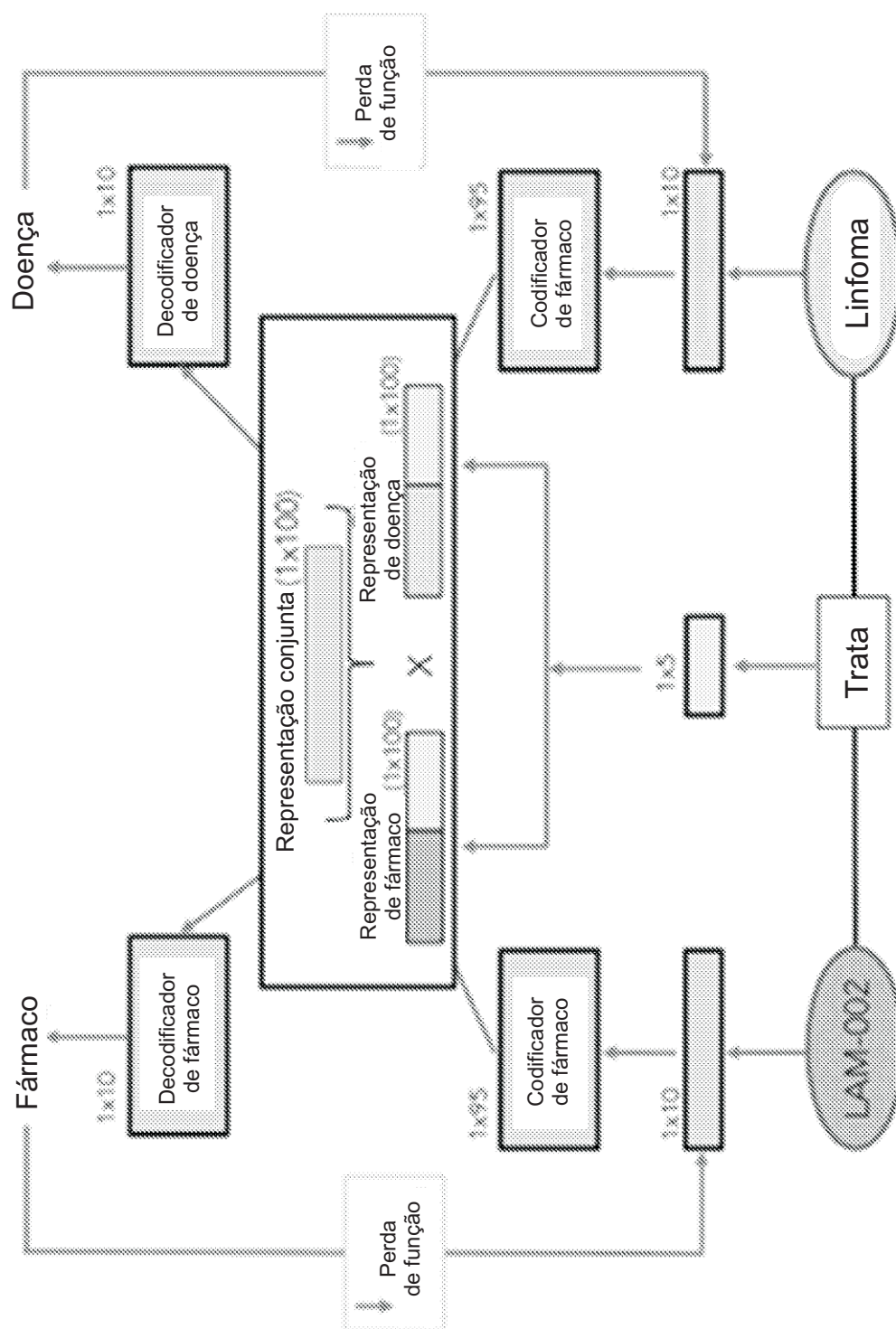


FIG. 10C

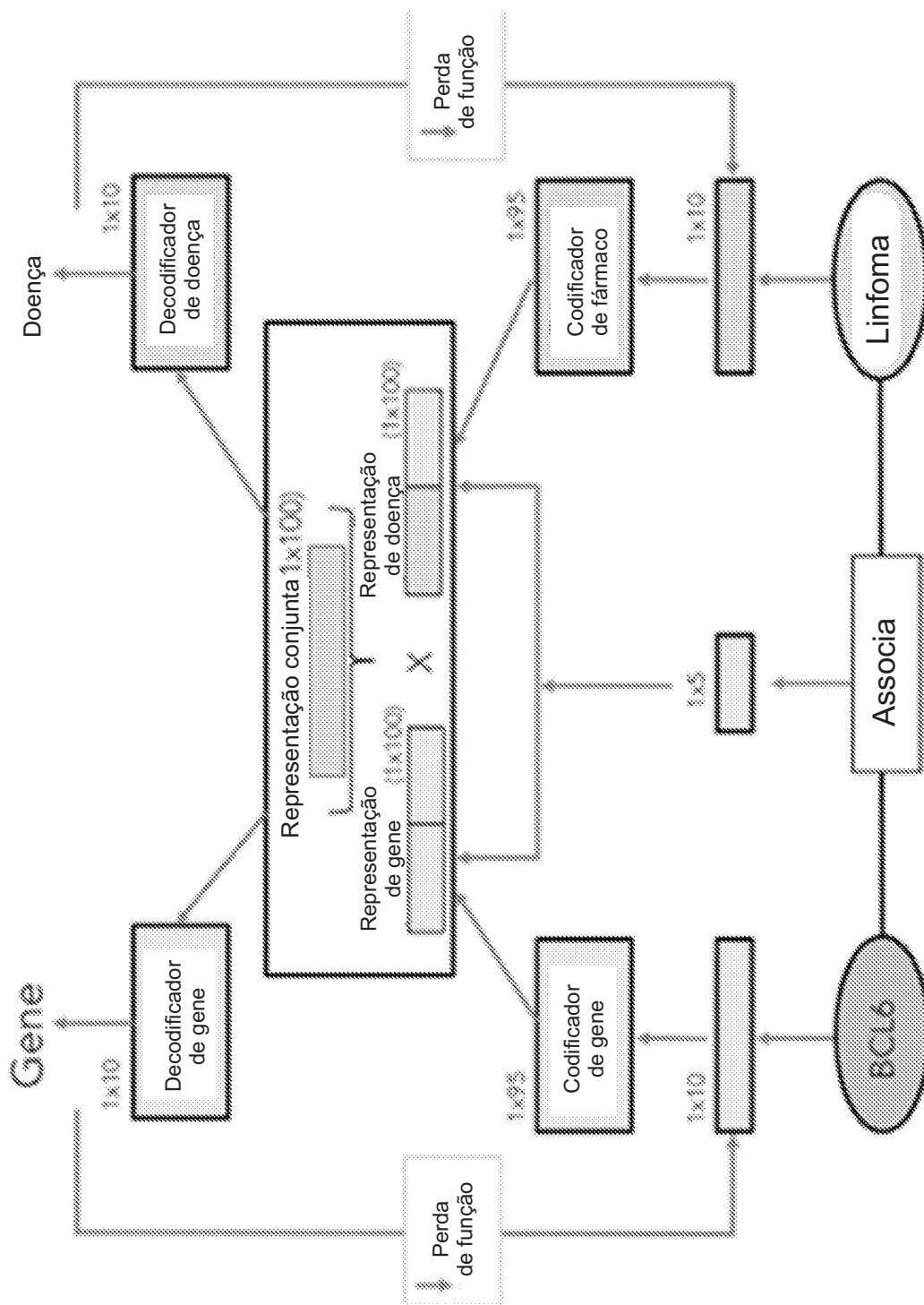


FIG. 10B

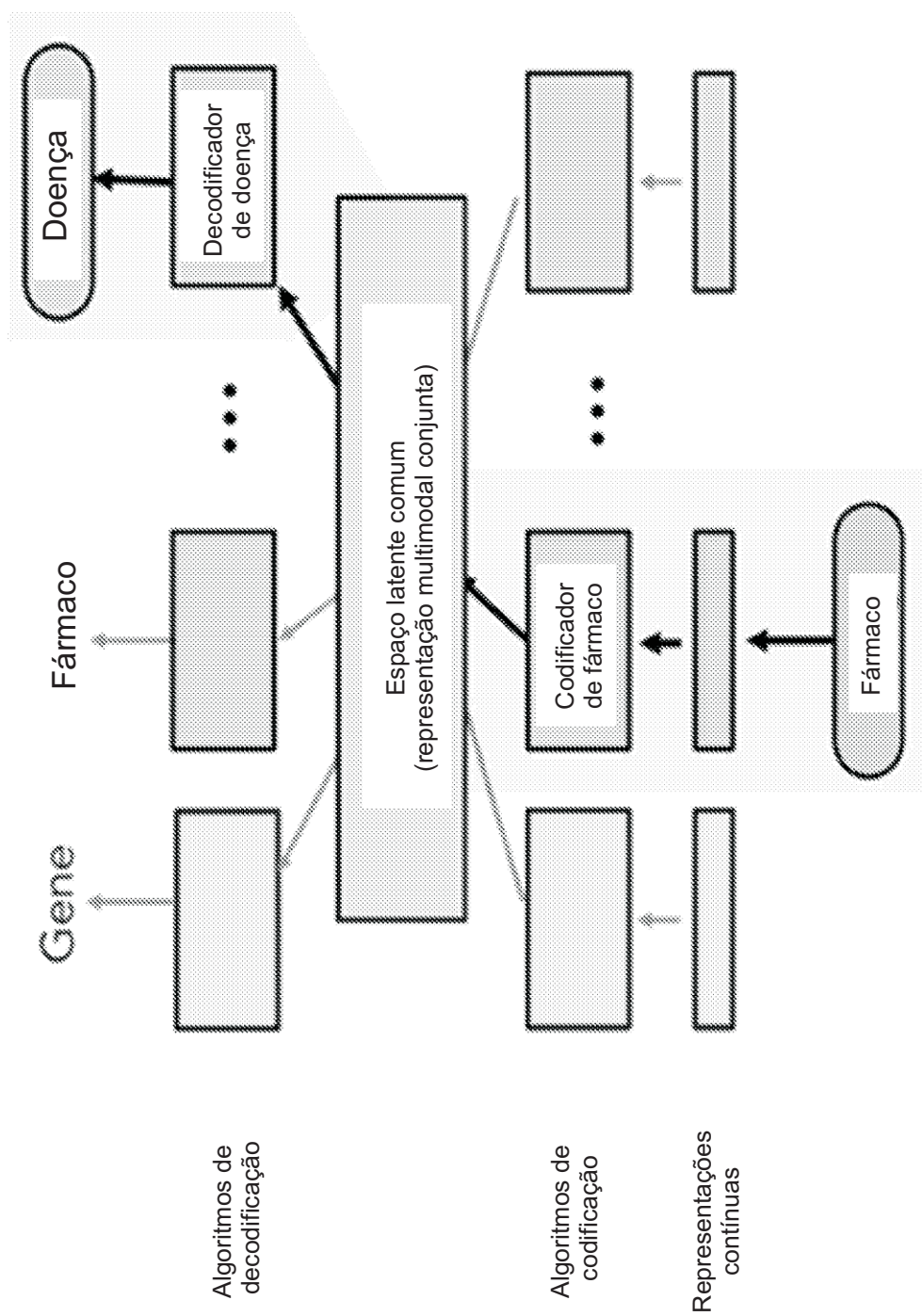


FIG. 11

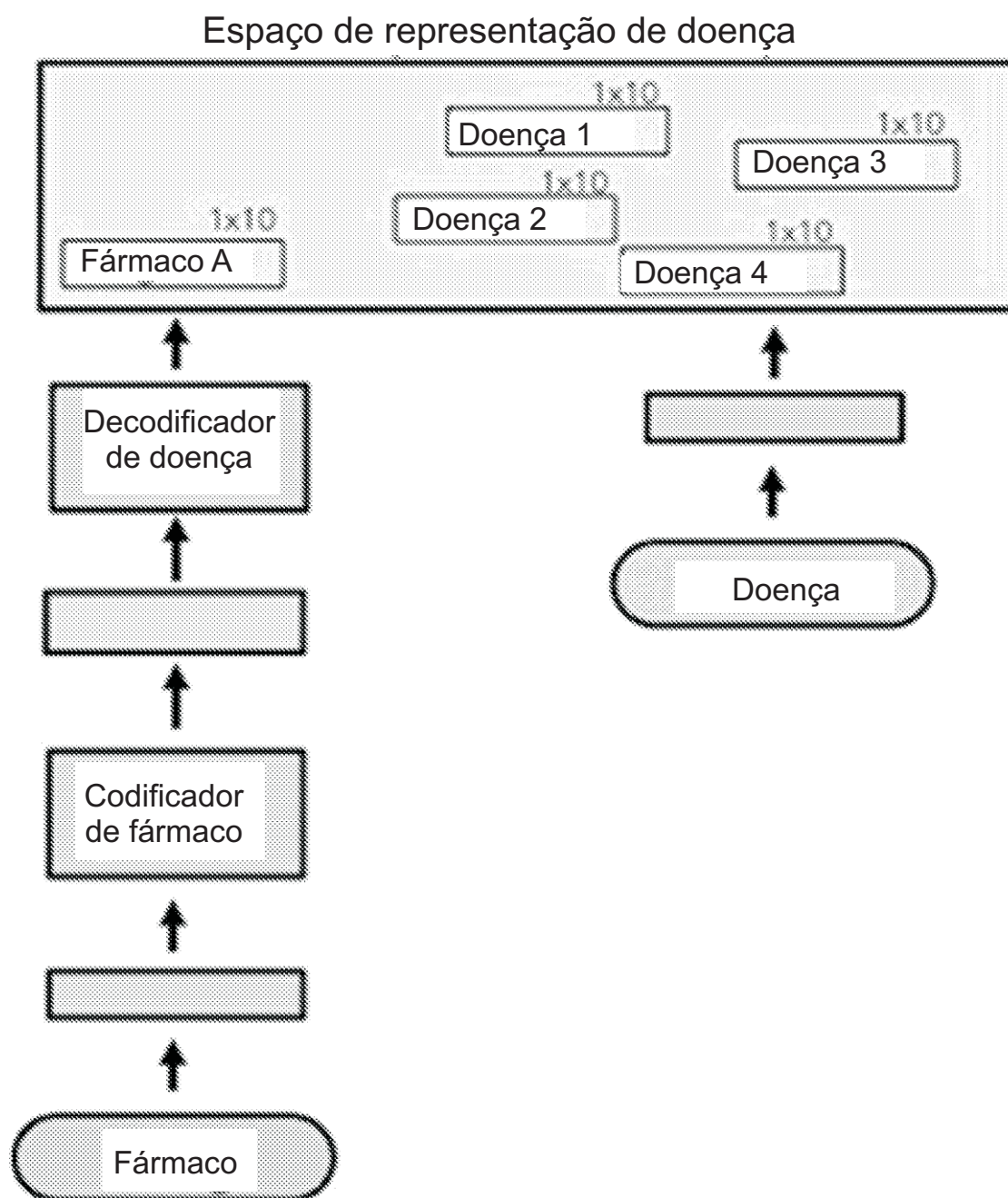


FIG. 12

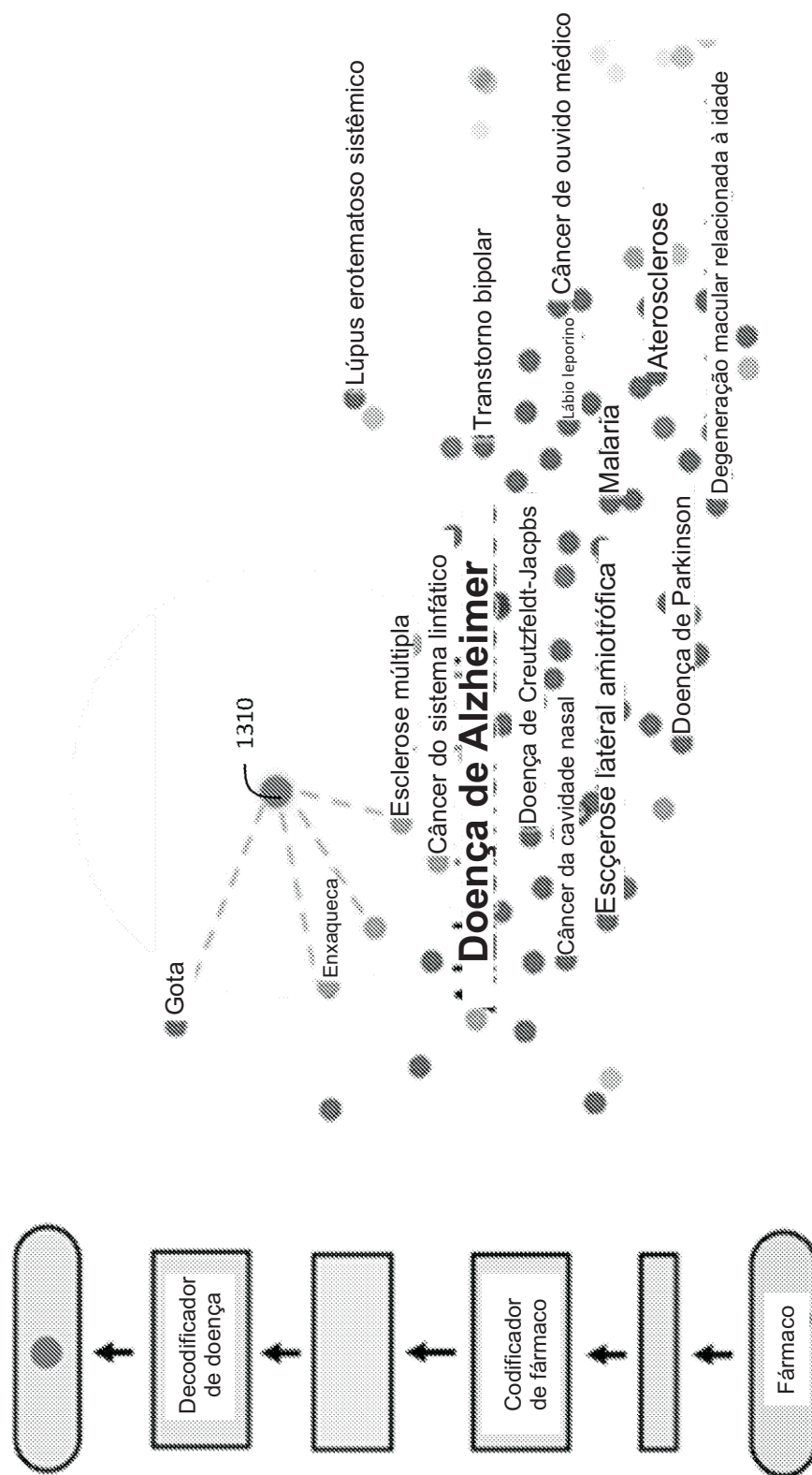


FIG. 13

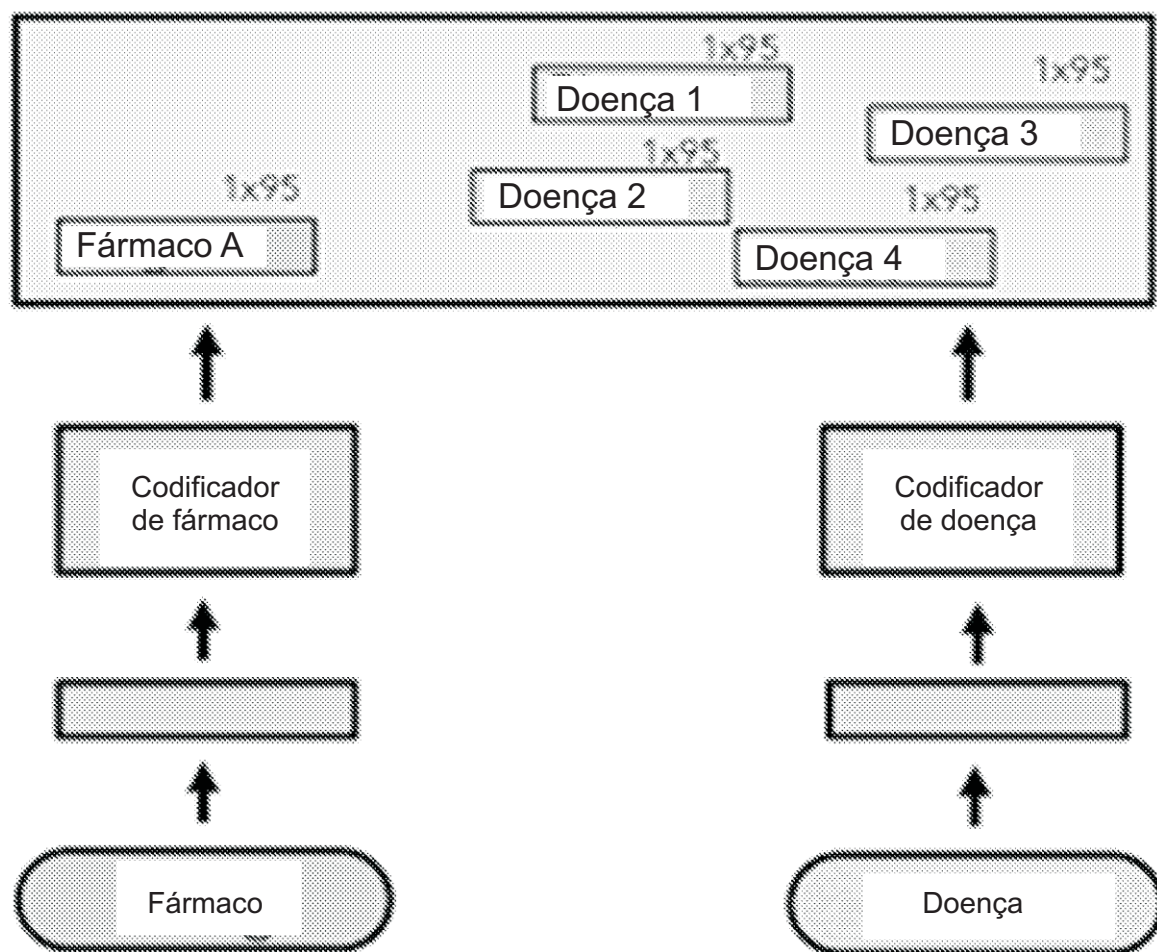


FIG. 14

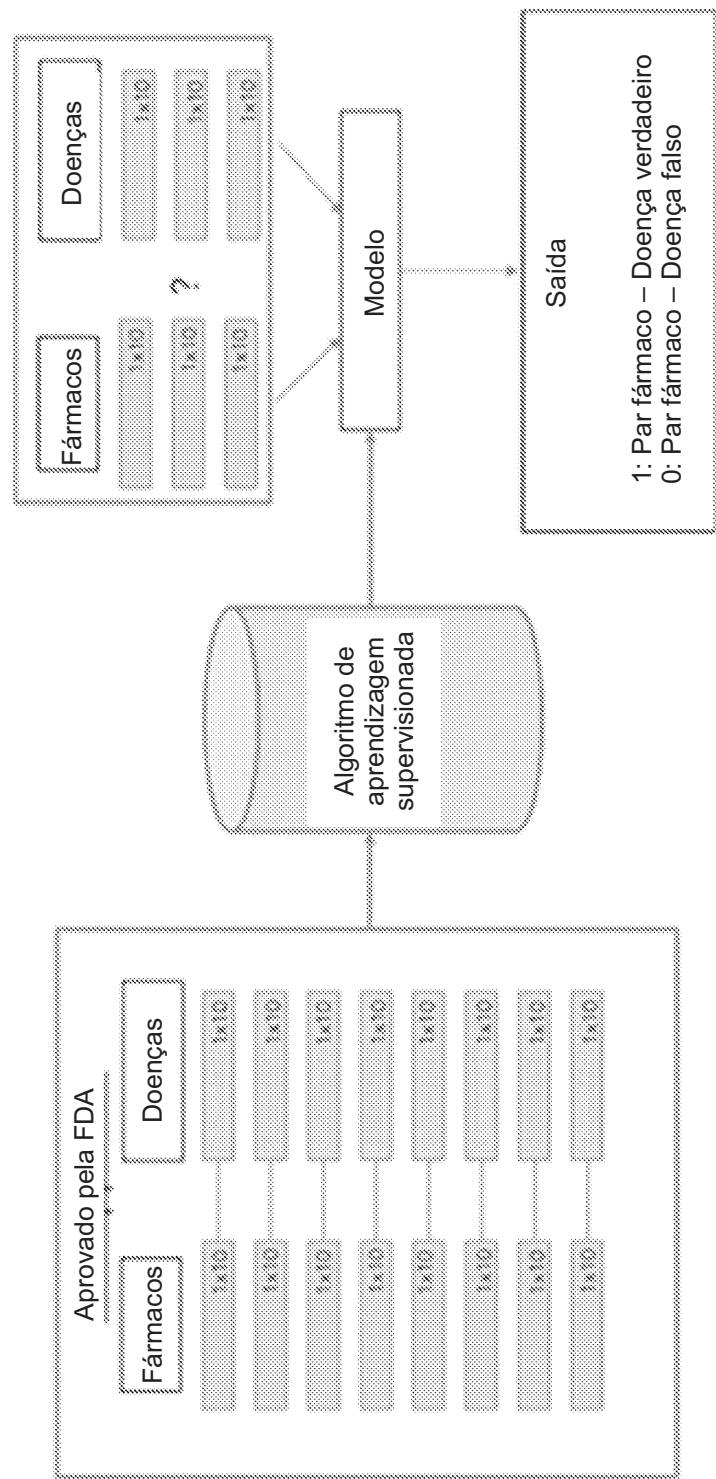


FIG. 15

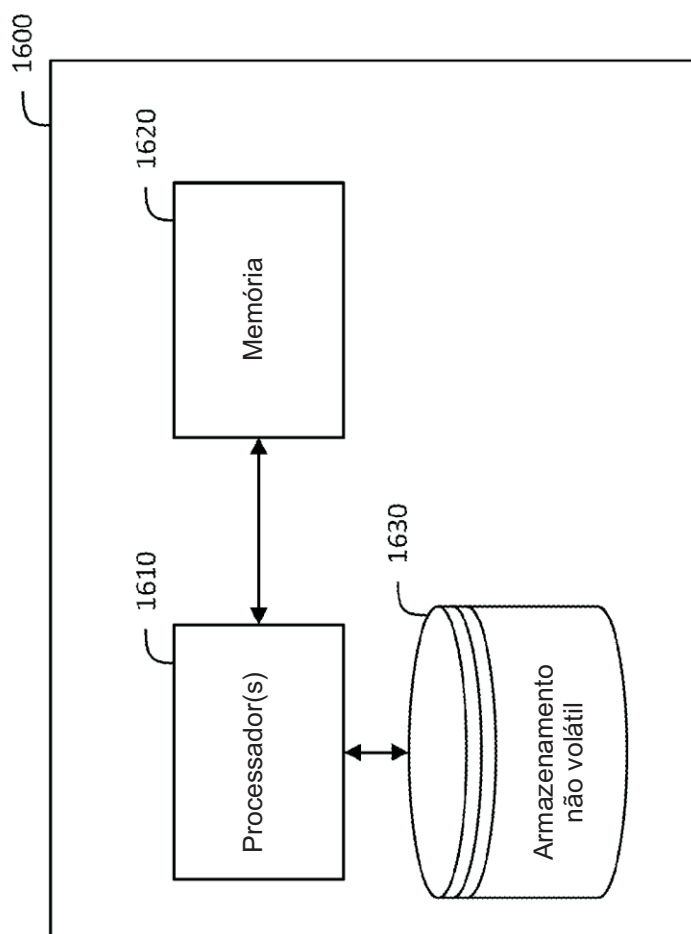


FIG. 16

RESUMO

Patente de Invenção: **"MÉTODOS E APARELHOS PARA A PREVISÃO MULTIMODAL USANDO UM MODELO ESTATÍSTICO TREINADO"**.

A presente invenção refere-se a métodos e aparelhos para prever uma associação entre dados de entrada em uma primeira modalidade e dados em uma segunda modalidade usando um modelo estatístico treinado para representar interações entre dados tendo uma pluralidade de modalidades incluindo a primeira modalidade e a segunda modalidade, o modelo estatístico compreendendo uma pluralidade de codificadores e decodificadores, cada um dos quais é treinado para processar dados para uma da pluralidade de modalidades, e uma representação de modalidade conjunta acoplando a pluralidade de codificadores e decodificadores. O método compreende selecionar, com base na primeira modalidade e na segunda modalidade, um par de codificador/decodificador ou um par de codificadores, a partir de entre a pluralidade de codificadores e decodificadores, e processar os dados de entrada com a representação de modalidade conjunta e o par de codificador/decodificador selecionado ou o par de codificadores para prever a associação entre os dados de entrada e os dados na segunda modalidade.