(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 772 687**

(51) Int. CI.:

G06K 9/46 (2006.01)
G06K 9/62 (2006.01)
G06N 3/04 (2006.01)
G06N 3/08 (2006.01)
G06N 20/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **04.10.2016 PCT/US2016/055369**

(87) Fecha y número de publicación internacional: **13.04.2017 WO17062382**

(96) Fecha de presentación y número de la solicitud europea: **04.10.2016 E 16854180 (3)**

(97) Fecha y número de publicación de la concesión europea: **27.11.2019 EP 3356999**

(54) Título: **Sistema para aplicar una red convolucional a datos espaciales**

(30) Prioridad:

04.10.2015 US 201562236962 P
23.02.2016 US 201615050983
20.06.2016 US 201615187018

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
08.07.2020

(73) Titular/es:

ATOMWISE INC. (100.0%)
717 Market Street, Suite 800
San Francisco, CA 94103, US

(72) Inventor/es:

HEIFETS, ABRAHAM, SAMUEL;
WALLACH, IZHAR y
DZAMBA, MICHAEL

(74) Agente/Representante:

VALLEJO LÓPEZ, Juan Pedro

DESCRIPCIÓN

Sistema para aplicar una red convolucional a datos espaciales

Referencia cruzada a solicitudes relacionadas

Esta solicitud reivindica prioridad sobre la Solicitud de Estados Unidos N.º 15/187.018 titulada "Systems and Methods for Applying a Convolutional Network to Spatial Data", presentada el 20 de junio de 2016, que es una continuación de la Patente de Estados Unidos N.º 9.373.059 titulada "Systems and Methods for Applying a Convolutional Network to Spatial Data", presentada el 23 de febrero de 2016. Esta solicitud reivindica también prioridad sobre la Solicitud de Patente Provisional de Estados Unidos N.º 62/236.962 titulada "Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Discovery", presentada el 4 de octubre de 2015.

Campo técnico

Lo siguiente se refiere en general a aplicar redes neurales convolucionales a datos espaciales.

Antecedentes

Detectar motivos estructurales dentro de datos espaciales tridimensionales de objetos de prueba anclados en objetos objetivos (complejos) que influencian tal anclaje es una tarea de reconocimiento de patrones importante, y tiene amplias aplicaciones, incluyendo, pero sin limitación, la predicción de la afinidad de objetos de prueba a objetos objetivos. Al mismo tiempo, la detección de tales motivos estructurales está dificultada por incertidumbres en la precisión de los datos espaciales tridimensionales e incertidumbre en la forma en que objetos de prueba se unen a objetos objetivos. Por ejemplo, una interacción dada entre un objeto de prueba y un objeto objetivo puede verse afectada por la distancia, ángulo, tipo de átomo, carga y polarización, y factores ambientales estabilizantes o desestabilizantes implicados.

La técnica anterior incluye sistemas y métodos que (i) se basan en conocimiento, (ii) son empíricos o (iii) incluyen funciones de puntuación basadas en campo de fuerza. Características basadas en conocimiento habitualmente constan de recuentos del número de veces que pares de átomos o grupos funcionales se separan por una distancia dada en complejos. Porque estas características son simples (dos puntos separados por una distancia), son incapaces de capturar el conjunto de complejos de factores de influencia descritos anteriormente. Funciones de puntuación empíricas ajustan un conjunto de pesos de relativa importancia a un número pequeño (pocas docenas) de características diseñadas a mano, tal como el número de enlaces rotativos, pares de donador-aceptor de enlace de hidrógeno, apilamientos aromáticos, electrostática, complementariedad estérica o cepa, o área hidrofóbica accesible por solvente.
El desarrollo de estas características requiere conocimiento experto y ajuste manual extensivo, incluso cualquiera de tales características será necesariamente una aproximación limitada ya que, como se ha analizado anteriormente, las fuerzas que gobiernan interacciones entre objetos de prueba y objetos objetivos no pueden separarse de forma consistente. Funciones basadas en campo de fuerza de puntuación se diseñan para ser computacionalmente eficientes, que requiere aproximaciones a resultados teóricos a partir de predicciones de fase gaseosa. Por ejemplo, tales sistemas ignoran o aproximan toscamente la mediación importante de fuerza de campo mediante solvente. Una técnica anterior de antecedente es el documento US 2015/0238148 A1.

Teniendo en cuenta los antecedentes anteriores, existe una necesidad para soluciones que proporcionan detección más precisa y/o más eficiente de motivos estructurales dentro de datos espaciales tridimensionales de complejos que influencian el anclaje de objetos de prueba en objetos objetivos.

Sumario

Se proporcionan sistemas y métodos para clasificación de objeto de prueba en los que el objeto de prueba se modeliza con un objeto objetivo en una pluralidad de diferentes poses para formar mapas de vóxel. Los mapas de vóxel se vectorizan y proporcionan secuencialmente a una red neural convolucional. La red neural convolucional comprende una capa de entrada, una pluralidad de capas convolucionales conectadas secuencialmente y ponderadas individualmente, y un calificador de salida. Las capas convolucionales incluyen una capa inicial y una capa final. En respuesta a entrada vectorizada, la capa de entrada proporciona valores en la capa convolucional inicial. Cada capa convolucional respectiva, distinta de la capa convolucional final, proporciona valores intermedios como una función de los pesos de la capa convolucional respectiva y valores de entrada de la capa convolucional respectiva en otra de las capas convolucionales. La capa convolucional final proporciona valores en el calificador como una función de los pesos de capa final y valores de entrada. De esta manera, el calificador puntúa cada uno del vector de entrada y estas puntuaciones se usan colectivamente para caracterizar el objeto de prueba.

Un aspecto de la presente divulgación proporciona un sistema informático para predecir afinidad de unión de un compuesto químico a un polímero objetivo usando datos espaciales. El sistema informático comprende al menos un procesador general y memoria general accesible por el al menos un procesador general. La memoria general

almacena al menos un programa para ejecución por el al menos un procesador general. El al menos un programa comprende instrucciones para obtener coordenadas espaciales para el polímero objetivo. El al menos un programa comprende además instrucciones para modelizar el compuesto químico con el polímero objetivo en cada pose de una pluralidad de diferentes poses, creando de este modo una pluralidad de mapas de vóxel. Cada respectivo mapa de vóxel en la pluralidad de mapas de vóxel comprende el compuesto químico en una respectiva pose en la pluralidad de diferentes poses.

El al menos un programa comprende además instrucciones para desplegar cada mapa de vóxel en la pluralidad de mapas de vóxel en un correspondiente vector, creando de este modo una pluralidad de vectores. Cada vector en la pluralidad de vectores es del mismo tamaño.

Cada vector respectivo en la pluralidad de vectores se introduce en una arquitectura de red que incluye (i) una capa de entrada para recibir secuencialmente la pluralidad de vectores, (ii) una pluralidad de capas convolucionales y (iii) un calificador. La pluralidad de capas convolucionales incluye una capa convolucional inicial y una capa convolucional final. Cada capa convolucional en la pluralidad de capas convolucionales se asocia con un conjunto diferente de pesos. En respuesta a entrada de un vector respectivo en la pluralidad de vectores, la capa de entrada proporciona una primera pluralidad de valores en la capa convolucional inicial como una primera función de valores en el vector respectivo. Cada capa convolucional respectiva, distinta de la capa convolucional final, proporciona valores intermedios, como una segunda función respectiva de (i) el conjunto diferente de pesos asociados a la capa convolucional respectiva y (ii) valores de entrada recibidos por la capa convolucional respectiva, en otra capa convolucional en la pluralidad de capas convolucionales. La capa convolucional final proporciona valores finales, como una tercera función de (i) el conjunto diferente de pesos asociados a la capa convolucional final y (ii) valores de entrada recibidos por la capa convolucional final, en el calificador.

El al menos un programa comprende además instrucciones para obtener una pluralidad de puntuaciones desde el calificador, en el que cada puntuación en la pluralidad de puntuaciones corresponde a la entrada de un vector en la pluralidad de vectores en la capa de entrada. La pluralidad de puntuaciones se usa para predecir la afinidad de unión del compuesto químico con el polímero objetivo.

En algunas realizaciones, el calificador comprende una pluralidad de capas totalmente conectadas y una capa de evaluación. Además, una capa totalmente conectada en la pluralidad de capas totalmente conectadas se proporciona a la capa de evaluación.

En algunas realizaciones, el calificador comprende un árbol de decisión, un árbol de regresión aditiva múltiple, un algoritmo de agrupación, análisis de componente principal, un análisis de vecino más cercano, un análisis discriminante lineal, un análisis discriminante cuadrático, una máquina de vector de soporte, un método evolucionario, una búsqueda de proyección y conjuntos de los mismos.

En algunas realizaciones, cada vector en la pluralidad de vectores es un vector unidimensional.

En algunas realizaciones, la pluralidad de diferentes poses comprende 2 o más poses, 10 o más poses, 100 o más poses o 1000 o más poses.

En algunas realizaciones, la pluralidad de diferentes poses se obtiene usando una función de puntuación de anclaje en uno de muestreo Monte Carlo de cadena de marcas, recocido simulado, Algoritmos Genéticos de Lamarckian o algoritmos genéticos.

En algunas realizaciones, la pluralidad de diferentes poses se obtiene mediante búsqueda incremental usando un algoritmo voraz.

En algunas realizaciones, el objeto objetivo es un polímero (por ejemplo, una proteína, un polipéptido, un ácido polinucleico, un ácido polirribonucleico, un polisacárido o un conjunto de cualquier combinación de los mismos).

En algunas realizaciones, el objeto objetivo es un polímero y las coordenadas espaciales son un conjunto de coordenadas tridimensionales $\{x_1,..., x_N\}$ para una estructura cristalina del polímero resuelto a una resolución de 2,5 Å o mejor o una resolución de 3,3 Å o mejor.

En algunas realizaciones, el objeto objetivo es un polímero y las coordenadas espaciales son un conjunto de coordinadas tridimensionales para el polímero determinado mediante resonancia magnética nuclear, difracción de neutrones o criomicroscopía electrónica.

En algunas realizaciones, el objeto de prueba es un compuesto químico y usar la pluralidad de puntuaciones para caracterizar el objeto de prueba comprende tomar una medida de tendencia central de la pluralidad de puntuaciones. En algunas de tales realizaciones: cuando la medida de tendencia central satisface un valor umbral predeterminado o intervalo de valores umbral predeterminados, la caracterización comprende considerar que el objeto de prueba tiene una primera clasificación, y cuando la medida de tendencia central no satisface el valor umbral predeterminado

o intervalo de valores umbral predeterminados, la caracterización comprende considerar que el objeto de prueba tiene una segunda clasificación. En algunas de tales realizaciones, la primera clasificación es una determinación de que el objeto de prueba no es tóxico para un organismo huésped, y la segunda clasificación es una determinación de que el objeto de prueba es tóxico para el organismo huésped. En algunas de tales realizaciones, la primera clasificación es una predicción de que el objeto de prueba se une a un objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por debajo de un primer valor de unión, y la segunda clasificación es una predicción de que el objeto de prueba se une al objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por encima del primer valor de unión. En algunas de tales realizaciones, la primera clasificación es una predicción de que el objeto de prueba se une a un primer objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por debajo de un primer valor de unión y que el objeto de prueba se une a un segundo objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por encima del primer valor de unión, y la segunda clasificación es una predicción de que el objeto de prueba se une a un primer objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por debajo de un primer valor de unión y que el objeto de prueba se une a un segundo objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por debajo del primer valor de unión. En algunas de tales realizaciones, el primer valor de unión es uno micromolar. En algunas de tales realizaciones, el primer valor de unión es diez micromolar.

En algunas realizaciones, el uso de la pluralidad de puntuaciones para caracterizar el objeto de prueba comprende tomar un promedio ponderado de la pluralidad de puntuaciones, en la que, cuando el promedio ponderado satisface un valor umbral predeterminado o intervalo de valores umbral predeterminados, el objeto de prueba se considera que tiene una primera clasificación, y cuando el promedio ponderado no satisface el valor umbral predeterminado o intervalo de valores umbral predeterminados, el objeto de prueba se considera que tiene una segunda clasificación. En algunas de tales realizaciones, el promedio ponderado es un promedio de Boltzman de la pluralidad de puntuaciones. En algunas de tales realizaciones, la primera clasificación es una IC_{50}, EC_{50}, Kd o KI para el objeto de prueba con respecto al objeto objetivo que está por encima de un primer valor de unión, y la segunda clasificación es una IC_{50}, EC_{50}, Kd o KI para el objeto de prueba con respecto al objeto objetivo que está por debajo del primer valor de unión. En algunas de tales realizaciones, el primer valor de unión es uno micromolar. En algunas de tales realizaciones, el primer valor de unión es diez micromolar. En algunas de tales realizaciones, el primer valor de unión es la IC50, EC50, Kd o KI predicha para un objeto de prueba diferente con respecto al objeto objetivo. En algunas de tales realizaciones, la primera clasificación es una determinación de que el objeto de prueba no es tóxico para un organismo huésped, y la segunda clasificación es una determinación de que el objeto de prueba es tóxico para el organismo huésped. En algunas de tales realizaciones, la primera clasificación es una predicción de que el objeto de prueba se une a un primer objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por debajo de un primer valor de unión y que el objeto de prueba se une a un segundo objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por encima del primer valor de unión, y la segunda clasificación es una predicción de que el objeto de prueba se une a un primer objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por debajo de un primer valor de unión y que el objeto de prueba se une a un segundo objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o KI que está por debajo del primer valor de unión (por ejemplo, el primer valor de unión es uno micromolar o diez micromolar, etc.).

En algunas realizaciones, el uso de la pluralidad de puntuaciones para caracterizar el objeto de prueba comprende tomar un promedio ponderado de la pluralidad de puntuaciones, en la que cuando el promedio ponderado satisface un intervalo respectivo de valores umbral en una pluralidad de intervalos de valores umbral, el objeto de prueba se considera que tiene una clasificación respectiva en una pluralidad de una clasificación respectiva que inequívocamente corresponde al intervalo respectivo de valores umbral. En algunas de tales realizaciones, cada clasificación respectiva en la pluralidad de clasificaciones es un intervalo de IC_{50}, EC_{50}, Kd o KI para el objeto de prueba con respecto al objeto objetivo. En algunas de tales realizaciones, una primera clasificación en la pluralidad de clasificaciones está entre uno micromolar y diez micromolar. En algunas de tales realizaciones, una primera clasificación en la pluralidad de clasificaciones está entre uno nanomolar y 100 nanomolar.

En algunas realizaciones, el objeto objetivo es un polímero con un sitio activo, el objeto de prueba es una composición química y la modelización comprende anclar el objeto de prueba en el sitio activo del polímero.

En algunas realizaciones, una capa convolucional en la pluralidad de capas convolucionales tiene una pluralidad de filtros y en el que cada filtro en la pluralidad de filtros convoluciona un espacio de entrada cúbico de N^3 con zancada Y, donde N es un número entero de dos o mayor e Y es un número entero positivo. En algunas de tales realizaciones, el conjunto diferente de pesos asociados a la capa convolucional se asocian con respectivos filtros en la pluralidad de filtros.

En algunas realizaciones, el calificador comprende una pluralidad de capas totalmente conectadas y una capa de coste de regresión logística en el que una capa totalmente conectada en la pluralidad de capas totalmente conectadas se proporciona a la capa de coste de regresión logística.

En algunas realizaciones, el objeto objetivo es un polímero con un sitio activo, el objeto de prueba es una composición química, la modelización comprende realizar una ejecución de dinámica molecular del objeto objetivo y el objeto de prueba, formando de este modo una trayectoria del objeto objetivo y el objeto de prueba juntos con el

paso del tiempo, y se obtiene un subconjunto de la pluralidad de diferentes poses tomando instantáneas de la trayectoria en un periodo de tiempo.

En algunas realizaciones, el sistema informático comprende además una unidad de procesamiento gráfico que tiene una memoria de procesamiento gráfico, en el que la memoria de procesamiento gráfico comprende la arquitectura de red y realiza la provisión (D) y el al menos un procesador general realiza el uso (F).

En algunas realizaciones, la caracterización del objeto de prueba es una predicción de toxicidad del objeto de prueba.

En algunas realizaciones, la caracterización del objeto de prueba es una predicción de la potencia del objeto de prueba contra un objetivo de enfermedad molecular. Por ejemplo, en algunas de tales realizaciones, la potencia es una predicción de afinidad de unión del objeto de prueba contra el objetivo de enfermedad molecular.

En algunas realizaciones, la caracterización del objeto de prueba es una predicción de la selectividad del objeto de prueba contra un primer objetivo molecular frente a un segundo objeto molecular, en el que el primer objetivo molecular se vincula a una enfermedad.

En algunas realizaciones, el objeto objetivo se vincula con una enfermedad y se predice la inhibición del objeto objetivo mediante la unión del objeto de prueba al objeto objetivo para aliviar la enfermedad.

Breve descripción de los dibujos

En los dibujos, realizaciones de los sistemas y método de la presente divulgación se ilustran a modo de ejemplo. Debe entenderse expresamente que la descripción y dibujos son únicamente para el propósito de ilustración y como una ayuda para el entendimiento, y no se conciben como una definición de los límites de los sistemas y métodos de la presente divulgación.

La Figura 1 ilustra un sistema informático que aplica una red neural convolucional a datos espaciales de acuerdo con algunas realizaciones.

Las Figuras 2A, 2B, 2C, 2D, 2E y 2F ilustran sistemas informáticos y métodos para aplicar una red neural convolucional a datos espaciales de acuerdo con algunas realizaciones.

La Figura 3 es una vista esquemática de un objeto de prueba de ejemplo en dos diferentes poses en relación con un objeto objetivo, de acuerdo con una realización.

La Figura 4 es una vista esquemática de una representación geométrica de características de entrada en forma de una cuadrícula tridimensional de vóxeles, de acuerdo con una realización.

La Figura 5 y la Figura 6 son vistas de dos objetos codificados en una cuadrícula bidimensional de vóxeles, de acuerdo con una realización.

La Figura 7 es la vista de la visualización de la Figura 6, en la que los vóxeles se han numerado, de acuerdo con una realización.

La Figura 8 es una vista esquemática de representación geométrica de características de entrada en forma de ubicaciones de coordenadas de centros de átomo, de acuerdo con una realización.

La Figura 9 es una vista esquemática de las ubicaciones de coordenadas de la Figura 8 con un intervalo de ubicaciones, de acuerdo con una realización.

La Figura 10 ilustra una distribución de valores de AUC y de logAUC de 50 objetivos de PMD de ChEMBL-20 para AtomNet y Smina de acuerdo con una realización.

La Figura 11 ilustra la distribución de valores de AUC y de logAUC de 102 objetivos de DUDE para AtomNet y Smina de acuerdo con una realización.

La Figura 12 ilustra la distribución de valores de AUC y de logAUC de 149 objetivos de inactivos de ChEMBL-20 para AtomNet y Smina de acuerdo con una realización.

Las Figuras 13A y 13B ilustran las diferencias entre las mediciones de AUC y de logAUC con respecto al enriquecimiento temprano de acuerdo con una realización.

La Figura 14 es una representación de la aplicación de elementos de cálculo de funciones múltiples (g1, g2,...) a las entradas de vóxel (x1, x2,... , x100) y composición de las salidas de elementos de cálculo de función juntas

usando g(), de acuerdo con una realización.

Las Figuras 15A y 15B ilustran las ubicaciones tridimensionales en un objeto objetivo al que dispara un filtro particular de una primera capa convolucional de acuerdo con algunas realizaciones.

Números de referencia similares hacen referencia a partes correspondientes a través de las varias vistas de los dibujos.

Descripción detallada

Se hará ahora referencia en detalle a las realizaciones, ejemplos de las cuales se ilustran en los dibujos adjuntos. En la siguiente descripción detallada, se exponen numerosos detalles específicos para proporcionar un completo entendimiento de la presente divulgación. Sin embargo, será evidente para un experto en la materia que la presente divulgación puede practicarse sin estos detalles específicos. En otros casos, no se han descrito en detalle métodos bien conocidos, procedimientos, componentes, circuitos y redes para no obstaculizar innecesariamente aspectos de las realizaciones.

Se entenderá también que, aunque los términos primero, segundo, etc. pueden usarse en este documento para describir diversos elementos, estos elementos no deberían limitarse por estos términos. Estos términos se usan únicamente para distinguir un elemento de otro. Por ejemplo, un primer sujeto podría denominarse un segundo sujeto y, de manera similar, un segundo sujeto podría denominarse un primer sujeto, sin alejarse del alcance de la presente divulgación. El primer sujeto y el segundo sujeto son ambos sujetos, pero no son el mismo sujeto.

La terminología usada en la presente divulgación es para el propósito de descripción de realizaciones particulares únicamente y no pretende ser una limitación de la invención. Como se usa en la descripción de la invención y las reivindicaciones adjuntas, las formas singulares "un", "una", "el" y "la" se conciben para incluir también las formas plurales, a menos que el contexto lo indique claramente de otra manera. Se entenderá también que el término "y/o" como se usa en este documento se refiere a e incluye cualquiera y todas las posibles combinaciones de uno o más de los artículos listados asociados. Se entenderá adicionalmente que los términos "comprende" y/o "que comprende", cuando se usan en esta memoria descriptiva, especifican la presencia de características indicadas, elementos integrantes, etapas, operaciones, elementos y/o componentes, pero no impiden la presencia o adición de una o más otras características, elementos integrantes, etapas, operaciones, elementos, componentes y/o grupos de los mismos.

Como se usa en este documento, el término "si" puede interpretarse para significar "cuando" o "tras" o "en respuesta a la determinación" o "en respuesta a la detección", dependiendo del contexto. De manera similar, la frase "si se determina" o "si se detecta [una condición o evento indicado]" puede interpretarse para significar "tras la determinación" o "en respuesta a la determinación" o "tras la detección de [la condición o evento indicado]" o "en respuesta a la detección de [la condición o evento indicado]", dependiendo del contexto.

La presente divulgación proporciona sistemas y métodos para clasificación de objeto de prueba. Un objeto de prueba se ancla con un objeto objetivo en una pluralidad de modos de unión aceptables energéticamente diferentes, denominados poses, para formar una correspondiente pluralidad de mapas de vóxel. Un ejemplo de un objeto objetivo es un polímero con un sitio activo y un ejemplo de un objeto de prueba es un compuesto que puede o puede no unirse al sitio activo con afinidad apreciativa. En algunas realizaciones, los mapas de vóxel se vectorizan y proporcionan secuencialmente en una red neural convolucional. En algunas realizaciones, los mapas de vóxel se proporcionan secuencialmente directamente en una red neural convolucional sin vectorización. En algunas realizaciones, cada tal mapa de vóxel representa la pose del objeto de prueba en relación con el objeto objetivo. Por ejemplo, en algunas realizaciones, cada mapa de vóxel representa un compuesto que se une en una orientación diferente en el sitio activo del polímero. La red neural convolucional comprende una capa de entrada, una pluralidad de capas convolucionales ponderadas individualmente y un calificador de salida. Las capas convolucionales incluyen una capa inicial y una capa final. En respuesta a la entrada, la capa de entrada proporciona valores en la capa convolucional inicial. Cada capa convolucional respectiva, distinta de la capa convolucional final, proporciona valores intermedios como una función de los pesos de la capa convolucional respectiva y valores de entrada de la capa convolucional respectiva en otra de las capas convolucionales. La capa convolucional final proporciona valores en el calificador como una función de los pesos de capa final y valores de entrada. De esta manera, el calificador puntúa cada uno de los vectores de entrada (o mapas de vóxel de entrada) y estas puntuaciones se usan colectivamente para clasificar el objeto de prueba. En algunas realizaciones, el calificador proporciona una única puntuación para cada uno de los vectores de entrada (o mapas de vóxel de entrada) y se usa el promedio ponderado de estas puntuaciones para clasificar el objeto de prueba.

La Figura 1 ilustra un sistema informático 100 que aplica la red neural convolucional descrita anteriormente a datos espaciales. Por ejemplo, puede usarse como un sistema de predicción de afinidad de unión para generar predicciones precisas con respecto a la afinidad de unión de uno o más objetos de prueba (por ejemplo, compuestos químicos) con un conjunto de uno o más objetos objetivos (por ejemplo, polímeros).

Haciendo referencia a la Figura 1, en realizaciones típicas, el sistema informático de análisis 100 comprende uno o más componentes. Para propósitos de ilustración en la Figura 1, el sistema informático de análisis 100 se representa como un único ordenador que incluye toda la funcionalidad del sistema informático de análisis 100 divulgado. Sin embargo, la divulgación no se limita de esta forma. La funcionalidad del sistema informático de análisis 100 puede propagarse a través de cualquier número de ordenadores en red y/o residir en cada uno de varios ordenadores en red. Un experto en la materia apreciará que son posibles una amplia gama de diferentes topologías de ordenadores para el sistema informático de análisis 100 y todas tales tecnologías están dentro del alcance de la presente divulgación.

Volviendo a la Figura 1 con lo anterior en mente, un sistema informático de análisis 100 comprende una o más unidades de procesamiento (CPU) 74, una red u otra interfaz de comunicaciones 84, una interfaz de usuario (por ejemplo, incluyendo un visualizador 82 y teclado 80 u otra forma de dispositivo de entrada), una memoria 92 (por ejemplo, memoria de acceso aleatorio), uno o más dispositivos persistentes y/o de almacenamiento de disco magnético 90 opcionalmente accedidos por uno o más controladores 88, uno o más buses de comunicación 12 para interconectar los componentes anteriormente mencionados, y una fuente de alimentación 76 para alimentar los componentes anteriormente mencionados. Los datos en memoria 92 pueden compartirse sin discontinuidades con la memoria no volátil 90 usando técnicas informáticas conocidas tal como almacenamiento en memoria caché. La memoria 92 y/o memoria 90 pueden incluir almacenamiento masivo que se ubica remotamente con respecto a la unidad o unidades de procesamiento central 74. En otras palabras, algunos datos almacenados en la memoria 92 y/o memoria 90 pueden alojarse de hecho en ordenadores que son externos al sistema informático de análisis 100, pero que pueden accederse electrónicamente por el sistema informático de análisis a través de una Internet, intranet u otra forma de red o cable electrónico usando la interfaz de red 84. En algunas realizaciones, el sistema informático de análisis 100 hace uso de una red neural convolucional que se ejecuta desde la memoria 52 asociada con una o más unidades de procesamiento gráfico 50 para mejorar la velocidad y rendimiento del sistema. En algunas realizaciones alternativas, el sistema informático de análisis 100 hace uso de una red neural convolucional que se ejecuta desde la memoria 92 en lugar de la memoria asociada con una unidad de procesamiento gráfico 50.

La memoria 92 del sistema informático de análisis 100 almacena:

- un sistema operativo 54 que incluye procedimientos para tratar diversos servicios de sistema básicos;
- un módulo de evaluación de datos espaciales 56 para evaluar datos espaciales tal como la unión de objetos de prueba (u objetos de entrenamiento) a objetos objetivos;
- datos para uno o más objetos objetivos 58, incluyendo datos estructurales 60 y opcionalmente información de sitio activo 62;
- una librería de entrenamiento de objeto 64 que incluye datos de unión 68 contra objetos objetivos 58 para cada uno de una pluralidad de objetos de entrenamiento 66;
- una librería de evaluación de objetos de prueba 70 que comprende información para una pluralidad de objetos de prueba 72; y
- una pluralidad de mapas de vóxel 40, representando cada mapa de vóxel la pose de un objeto de entrenamiento 66 u objeto de prueba 72 contra un objeto objetivo 58;

La memoria 52, u opcionalmente la memoria 92, del sistema informático de análisis 100 almacena:

- un módulo de evaluación convolucional 20 para aplicar redes neurales convolucionales a datos espaciales (por ejemplo, para aplicar una red neural convolucional a objeto de prueba o de entrenamiento anclado en objetos objetivos);
- una o más representaciones vectorizadas 22 (opcionalmente) de mapas de vóxel 40; y
- una red neural convolucional 24 que incluye una capa de entrada 26, una o más capas convolucionales 28 y un calificador terminal 30.

En algunas implementaciones, uno o más de los módulos o elementos de datos anteriormente identificados del sistema informático de análisis 100 se almacenan en uno o más de los dispositivos de memoria anteriormente mencionados, y corresponden a un conjunto de instrucciones para realizar una función descrita anteriormente. Los datos, módulo o programas identificados anteriormente (por ejemplo, conjuntos de instrucciones) no necesitan implementarse como programas de software, procedimientos o módulos separados y, por lo tanto, diversos subconjuntos de estos módulos pueden combinarse o de otra manera redisponerse en diversas implementaciones. En algunas implementaciones, la memoria 92 y/o 90 (y opcionalmente 52) opcionalmente almacena un subconjunto de los módulos y estructuras de datos identificados anteriormente. Adicionalmente, en algunas realizaciones la memoria 92 y/o 90 (y opcionalmente 52) almacena módulos y estructuras de datos adicionales no descritos anteriormente.

Ahora que se ha divulgado un sistema para evaluación del anclaje de objetos de prueba o de entrenamiento en objeto objetivo usando datos espaciales, se detallan y analizan a continuación métodos para realizar tal evaluación con referencia a la Figura 2.

Obtener coordenadas espaciales para un objeto objetivo 202. De acuerdo con la Figura 2 se realizan métodos en o con un sistema informático 100 para clasificación de un objeto de prueba 72 (u objeto de entrenamiento) que usa datos espaciales. El sistema informático 100 opcionalmente comprende una unidad de procesamiento gráfico 50 que tiene una memoria de procesamiento gráfico 52. El sistema informático 100 comprende un procesador general 74 y memoria general 90 / 92 accesible por la unidad de procesamiento general. La memoria general almacena al menos un programa 56 para ejecución por el al menos un procesador general. El al menos un programa obtiene coordenadas espaciales 60 para un objeto objetivo 58.

En algunas realizaciones, el objeto objetivo 58 es un polímero (204). Ejemplos de polímeros incluyen, pero sin limitación proteínas, polipéptidos, ácidos polinucleicos, ácidos polirribonucleicos, polisacáridos o conjuntos de cualquier combinación de los mismos (206). Un polímero, tal como los estudiados que usan algunas realizaciones de los sistemas y métodos divulgados, es una molécula grande compuesta de residuos repetitivos. En algunas realizaciones, el polímero es un material natural. En algunas realizaciones, el polímero es un material sintético. En algunas realizaciones, el polímero es un elastómero, laca, ámbar, caucho natural o sintético, celulosa, baquelita, nailon, poliestireno, polietileno, polipropileno, poliacrilonitrilo, polietilenglicol o un polisacárido.

En algunas realizaciones, el objeto objetivo 58 es un heteropolímero (copolímero). Un copolímero es un polímero derivado a partir de dos (o más) especies monoméricas, a diferencia de un homopolímero en el que únicamente se usa un monómero. Copolimerización se refiere a métodos usados para sintetizar químicamente un copolímero. Ejemplos de copolímeros incluyen, pero sin limitación, plástico ABS, SBR, caucho nitrílico, estireno-acrilonitrilo, estireno-isopreno-estireno (SIS) y acetato de etilenvinilo. Ya que copolímero consta de al menos dos tipos de unidades constituyentes (también unidades estructurales o partículas), copolímeros pueden clasificarse basándose en cómo se disponen estas unidades a lo largo de la cadena. Estos incluyen copolímeros alternos con unidades A y B alternas regulares. Véase, por ejemplo, Jenkins, 1996, "Glossary of Basic Terms in Polymer Science", Pure Appl. Chem. 68 (12): 2287-2311. Ejemplos adicionales de copolímeros son copolímeros periódicos con unidades A y B dispuestas en una secuencia repetitiva (por ejemplo, $(A\text{-}B\text{-}A\text{-}B\text{-}B\text{-}A\text{-}A\text{-}A\text{-}A\text{-}B\text{-}B\text{-}B)_n$). Ejemplos adicionales de copolímeros son copolímeros estadísticos en los que la secuencia de residuos de monómeros en el copolímero siguen una regla estadística. Véase, por ejemplo, Painter, 1997, Fundamentals of Polymer Science, CRC Press, 1997, p 14. Aún otros ejemplos de copolímeros que pueden evaluarse usando los sistemas y métodos divulgados son copolímeros en bloque que comprenden dos o más subunidades de homopolímeros unidas mediante enlaces covalentes. La unión de las subunidades de homopolímeros puede requerir una subunidad no repetitiva intermedia, conocida como un bloque de unión. Copolímeros en bloque con dos o tres bloques distintos se llaman copolímeros dibloque y copolímeros tribloque, respectivamente.

En algunas realizaciones, el objeto objetivo 58 es de hecho una pluralidad de polímeros, en la que los polímeros respectivos en la pluralidad de polímeros no tienen todos el mismo peso molecular. En algunas de tales realizaciones, los polímeros en la pluralidad de polímeros se encuadran dentro de un intervalo de pesos con una correspondiente distribución de longitudes de cadena. En algunas realizaciones, el polímero es una molécula de polímero ramificado que comprende una cadena principal con una o más cadenas o ramificaciones secundarias sustituyentes. Tipos de polímeros ramificados incluyen, pero sin limitación, polímeros de estrella, polímeros de peine, polímeros de cepillo, polímeros dendronizados, escaleras y dendrímeros. Véase, por ejemplo, Rubinstein et al., 2003, Polymer physics, Oxford; Nueva York: Oxford University Press, p. 6.

En algunas realizaciones, el objeto objetivo 58 es un polipéptido. Como se usa en este documento, el término "polipéptido" significa dos o más aminoácidos o residuos unidos mediante un enlace peptídico. Los términos "polipéptido" y "proteína" se usan indistintamente en este documento e incluyen oligopéptidos y péptidos. Un "aminoácido", "residuo" o "péptido" se refiere a cualquiera de las veinte unidades estructurales estándar de proteínas como se conoce en la técnica, que incluyen iminoácidos, tal como prolina e hidroxiprolina. La designación de un isómero de aminoácido puede incluir D, L, R y S. La definición de aminoácido incluye aminoácidos no naturales. Por lo tanto, selenocisteína, pirrolisina, lantionina, ácido 2-aminoisobutírico, ácido gama-aminobutírico, dehidroalanina, ornitina, citrulina, homocisteína se consideran todos aminoácidos. En la técnica se conocen otras variantes o análogos de los aminoácidos. Por lo tanto, un polipéptido puede incluir estructuras peptidomiméticas sintéticas tal como peptoides. Véase Simon et al., 1992, Proceedings of the National Academy of Sciences USA, 89, 9367. Véase también Chin et al., 2003, Science 301, 964; y Chin et al., 2003, Chemistry & Biology 10.511.

Los objetos objetivos 58 evaluados de acuerdo con algunas realizaciones de los sistemas y métodos divulgados también pueden tener cualquier número de modificaciones postraduccionales. Por lo tanto, un objeto objetivo incluye aquellos polímeros que se modifican mediante acilación, alquilación, amidación, biotinilación, formilación, γ-carboxilación, glutamilación, glucosilación, glucilación, hidroxilación, yodación, isoprenilación, lipoilación, adición de cofactor (por ejemplo, de un hemo, flavina, metal, etc.), adición de nucleósidos y sus derivados, oxidación, reducción, pegilación, adición del fosfatidilinositol, fosfopantoetinilación, fosforilación, formación de piroglutamato, racemización, adición de aminoácidos por ARNt (por ejemplo, arginilación), sulfatación, selenoilación, ISGilación, SUMOilación, ubiquitinación, modificaciones químicas (por ejemplo, citrulinación y deamidación) y tratamiento con otras enzimas (por ejemplo, proteasas, fosfatasas y quinasas). En la técnica se conocen otros tipos de modificaciones postraduccionales y también se incluyen.

En algunas realizaciones, el objeto objetivo 58 es un complejo organometálico. Un complejo organometálico es un compuesto químico que contiene enlaces entre carbono y metal. En algunos casos, compuestos organometálicos se distinguen mediante el prefijo "organo-" por ejemplo compuestos de organopaladio.

En algunas realizaciones, el objeto objetivo 58 es un tensioactivo. Los tensioactivos son compuestos que disminuyen la tensión de un líquido, la tensión interfacial entre dos líquidos, o la de entre un líquido y un sólido. Los tensioactivos pueden actuar como detergentes, agentes humectantes, emulsionantes, agentes espumantes y dispersantes. Los tensioactivos normalmente son compuestos orgánicos que son anfifílicos, significando que contienen tanto grupos hidrofóbicos (sus colas) como grupos hidrofílicos (sus cabezas). Por lo tanto, una molécula tensioactiva contiene tanto un componente insoluble en agua (o soluble en aceite) y un componente soluble en agua. Las moléculas tensioactivas se disiparán en agua y absorberán en interfaces entre aire y agua o en la interfaz entre aceite y agua, en el caso en el que el agua esté mezclada con aceite. El grupo hidrofóbico insoluble puede entenderse fuera de la fase de agua a granel, en la fase de aire o de aceite, mientras el grupo de cabeza soluble en agua permanece en la fase de agua. Este alineamiento de moléculas tensioactivas en la superficie modifica las propiedades de superficie de agua en la interfaz de agua/aire o agua/aceite.

Ejemplos de tensioactivos iónicos incluyen tensioactivos iónicos tales como tensioactivos aniónicos, catiónicos o zwitteriónicos (anfóteros). En algunas realizaciones, el objeto objetivo 58 es una micela inversa o liposoma.

En algunas realizaciones, el objeto objetivo 58 es un fullereno. Un fullereno es cualquier molécula compuesta en su totalidad de carbono, en forma de una esfera, elipsoide o tubo hueco. Fullerenos esféricos también se llaman buckybolas, y se parecen a las bolas usadas en la asociación de fútbol. Los cilíndricos se llaman nanotubos de carbono o buckytubos. Los fullerenos son similares en estructura al grafito, que se compone de láminas de grafeno apiladas de anillos hexagonales enlazados; pero también pueden contener anillos pentagonales (o en ocasiones heptagonales).

En algunas realizaciones, el objeto objetivo es un polímero y las coordenadas espaciales son un conjunto de coordenadas tridimensionales {x1,..., xN} para una estructura cristalina del polímero resuelto a una resolución de 2,5 Å o mejor (208). En algunas realizaciones, el objeto objetivo es un polímero y las coordenadas espaciales son un conjunto de coordenadas tridimensionales {x1,..., xN} para una estructura cristalina del polímero resuelto a una resolución de 3,3 Å o mejor (210). En algunas realizaciones, el objeto objetivo es un polímero y las coordenadas espaciales son un conjunto de coordenadas tridimensionales {x1,..., xN} para una estructura cristalina del polímero resuelto (por ejemplo, mediante técnicas cristalográficas de rayos X) a una resolución de 3,3 Å o mejor, 3.2 Å o mejor, 3,1 Å o mejor, 3,0 Å o mejor, 2,5 Å o mejor, 2,2 Å o mejor, 2,0 Å o mejor, 1,9 Å o mejor, 1,85 Å o mejor, 1,80 Å o mejor, 1,75 Å o mejor o 1,70 Å o mejor.

En algunas realizaciones, el objeto objetivo 58 es un polímero y las coordenadas espaciales son un conjunto de diez o más, veinte o más o treinta o más coordenadas tridimensionales para el polímero determinado mediante resonancia magnética nuclear en el que el conjunto tiene una RMSD de esqueleto de 1.0 Å o mejor, 0,9 Å o mejor, 0,8 Å o mejor, 0,7 Å o mejor, 0,6 Å o mejor, 0,5 Å o mejor, 0,4 Å o mejor, 0,3 Å o mejor, o 0,2 Å o mejor. En algunas realizaciones las coordenadas espaciales se determinan mediante difracción de neutrones o criomicroscopía electrónica (212).

En algunas realizaciones, el objeto objetivo 58 incluye dos diferentes tipos de polímeros, tal como un ácido nucleico unido a un polipéptido. En algunas realizaciones, el polímero nativo incluye dos polipéptidos unidos entre sí. En algunas realizaciones, el polímero nativo en estudio incluye uno o más iones de metal (por ejemplo, una metaloproteinasa con uno o más átomos de cinc). En tales casos, los iones de metal y/o las moléculas pequeñas orgánicas pueden incluirse en las coordenadas espaciales 60 para el objeto objetivo 58.

En algunas realizaciones el objeto objetivo es un polímero y hay diez o más, veinte o más, treinta o más, cincuenta o más, cien o más, entre cien y mil o menos de 500 residuos en el polímero.

En algunas realizaciones, las coordenadas espaciales del objeto objetivo 58 se determinan usando métodos de modelización tal como métodos ab initio, métodos funcionales de densidad, métodos empíricos y semiempíricos, mecánicas moleculares, dinámicas químicas o dinámicas moleculares.

En una realización, las coordenadas espaciales se representan mediante las coordinadas cartesianas de los centros de los átomos que comprenden el objeto objetivo. En algunas realizaciones alternativas, las coordenadas espaciales 60 para un objeto objetivo 58 se representan mediante la densidad de electrones del objeto objetivo según se miden, por ejemplo, mediante cristalografía de rayos X. Por ejemplo, en algunas realizaciones, las coordenadas espaciales 60 comprenden un mapa de densidad de electrones $2F_{observada}$-$F_{calculada}$ calculado usando las coordinadas atómicas calculadas del objeto objetivo 58, donde $F_{observada}$ es las amplitudes de factor de estructura observadas del objeto objetivo y Fc es las amplitudes de factor de estructura calculadas a partir de las coordinadas atómicas calculadas del objeto objetivo 58.

Por lo tanto, las coordenadas espaciales 60 para el objeto objetivo pueden recibirse como datos de entrada desde diversas fuentes, tal como, pero sin limitación, conjuntos de estructuras generados mediante solución de RMN, cocomplejos según se interpretan a partir de cristalografía de rayos X, difracción de neutrones o criomicroscopía electrónica, muestreo a partir de simulaciones computacionales, modelización de homología o muestreo de librería de rotámeros, y combinaciones de estas técnicas.

Modelizar un objeto de prueba con el objeto objetivo (214). En la etapa 214, el objeto de prueba 72 (u objeto de entrenamiento) se modeliza con el objeto objetivo 58 en cada una pluralidad de diferentes poses. En este documento, se describen primero objetos de prueba 72 representativos (y objetos de entrenamiento 66) de acuerdo con la presente divulgación. A continuación, se describen técnicas de modelización y modelización representativa.

Objetos de prueba 72 representativos (y objetos de entrenamiento 66). La diferencia significativa entre objetos de prueba 72 y objetos de entrenamiento 66 es que los objetos de entrenamiento 66 se etiquetan (por ejemplo, con datos de unión complementarios obtenidos a través de ensayos de unión de laboratorio húmedo, etc.) y tal etiquetado se usa para entrenar a la red neural convolucional, mientras que los objetos de prueba 72 no se etiquetan y la red neural convolucional se usa para clasificar objetos de prueba 72. En otras palabras, los objetos de entrenamiento ya están clasificados mediante etiquetas, y tal clasificación se usa para entrenar a la red neural convolucional de modo que la red neural convolucional puede clasificar a continuación los objetos de prueba. Los objetos de prueba habitualmente no se clasifican antes de aplicación de la red neural convolucional. En realizaciones típicas, las clasificaciones asociadas con los objetos de entrenamiento 66 son datos de unión contra cada uno de los objetos objetivos 58 obtenidos mediante ensayos de unión de laboratorio húmedo. Como tal, en algunas realizaciones, cada objeto de entrenamiento 58 se etiqueta potencialmente contra varios diferentes objetos objetivos 58. Por ejemplo, considérese el caso en el que existen dos objetos objetivos 58, una primera enzima A (para la que se muestran inhibidores) y una segunda enzima B (para la que se no muestran inhibidores y para la que no es deseable inhibir para minimizar efectos secundarios dañinos). Cada objeto de entrenamiento 58 recibirá una primera etiqueta contra enzima A y una segunda etiqueta contra enzima B. Estas primera y segunda etiquetas pueden ser la misma o diferentes, por ejemplo, serán diferentes si el objeto de entrenamiento 58 es un mejor inhibidor de enzima A que es de enzima B.

En algunas realizaciones, objetos de prueba 72 y objetos de entrenamiento 66 son compuestos orgánicos que satisfacen dos o más reglas, tres o más reglas o las cuatro reglas de la Regla de Cinco de Lipinski: (i) no más de cinco donadores de enlace de hidrógeno (por ejemplo, grupos OH y NH), (ii) no más de diez aceptores de enlace de hidrógeno (por ejemplo N y O), (iii) un peso molecular inferior a 500 dalton, y (iv) un LogP inferior a 5. La "Regla de Cinco" se llama así porque tres de los cuatro criterios implican el número cinco. Véase, Lipinski, 1997, Adv. Drug Del. Rev. 23, 3.

En algunas realizaciones, un objeto de prueba 72 u objeto de entrenamiento 66 satisface uno o más criterios además de la Regla de Cinco de Lipinski. Por ejemplo, en algunas realizaciones, el objeto de prueba 72 u objeto de entrenamiento 66 tiene cinco o menos anillos aromáticos, cuatro o menos anillos aromáticos, tres o menos anillos aromáticos o dos o menos anillos aromáticos. En algunas realizaciones, un objeto de prueba 72 u objeto de entrenamiento 66 es cualquier compuesto orgánico que tiene un peso molecular de menos de 2000 dalton, de menos de 4000 dalton, de menos de 6000 dalton, de menos de 8000 dalton, de menos de 10000 dalton o menos de 20000 dalton.

Sin embargo, los sistemas y métodos de la presente divulgación no tienen limitación en el tamaño de los objetos de prueba 72 u objetos de entrenamiento 66. Por ejemplo, en algunas realizaciones, tales objetos son polímeros grandes, tales como anticuerpos.

Modelización. Volviendo al elemento 214 de la Figura 2A, los objetos de prueba 72 y/u objetos de entrenamiento 66 se modelizan con los objetos objetivos 58 en cada pose de una pluralidad de diferentes poses. En algunas realizaciones, el objeto objetivo 58 es un polímero con un sitio activo, el objeto de prueba (u objeto de entrenamiento) es un compuesto químico, y la modelización comprende anclar el objeto de prueba en el sitio activo del polímero (216). En algunas realizaciones, el objeto de prueba 72 u objeto de entrenamiento 66 se ancla en el objeto objetivo 58 una pluralidad de veces para formar una pluralidad de poses. En algunas realizaciones, el objeto de prueba 72 u objeto de entrenamiento 66 se ancla en el objeto objetivo 58 dos veces, tres veces, cuatro veces, cinco o más veces, diez o más veces, cincuenta o más veces, 100 o más veces o 1000 o más veces (218). Cada tal anclaje representa una pose diferente del objeto de prueba 72 u objeto de entrenamiento 66 anclado en el objeto objetivo 58. En algunas realizaciones, el objeto objetivo 58 es un polímero con un sitio activo y el objeto de prueba 72 u objeto de entrenamiento 66 se ancla en el sitio activo en cada una de pluralidad de diferentes formas, representando cada forma una pose diferente. Se espera que muchas de estas poses no sean correctas, significando que tales poses no representan verdaderas interacciones entre el objeto de prueba 72 (u objeto de entrenamiento 66) y el objeto objetivo 58 que surge en la naturaleza. Ventajosamente, durante el entrenamiento con los objetos de entrenamiento 66, la red neural convolucional será capaz de filtrar (reducir el peso) poses incorrectas porque no surgirán patrones consistentes entre las poses incorrectas y las etiquetas de objeto de entrenamiento. Sin pretender limitarse por cualquier teoría particular, se espera que interacciones entre objetos (por ejemplo, intermolecular) observadas entre poses incorrectas se cancelarán entre sí como ruido blanco mientras las

interacciones entre objetos formadas por poses correctas formadas por objetos de entrenamiento 66 se reforzarán entre sí y, por lo tanto, entrenarán los pesos de la red con el paso del tiempo. Por lo tanto, durante modo de entrenamiento con respecto a poses incorrectas, la red neural no encontrará patrones que expliquen la diferencia entre los objetos de entrenamiento 66 activos y los objetos de entrenamiento 66 inactivos (por ejemplo, para discriminar entre datos de etiquetado de los objetos de entrenamiento). Con respecto a poses incorrectas, la red aprendería el peso de los objetos de entrenamiento 66, su tamaño y descriptores de resumen global similares, pero ninguna de las interacciones intermoleculares reales que se forman entre los objetos de entrenamiento y el objeto de prueba en la naturaleza. Por lo tanto, ventajosamente, los sistemas y métodos divulgados no son sensibles a poses incorrectas, particularmente cuando se toman más de 10 poses por objeto de entrenamiento 66, más de cien poses por objeto de entrenamiento 66 o más de mil poses por objeto de entrenamiento 66. Análogamente, cuando se muestrean objetos de prueba 72, también se toman una pluralidad de poses. Por lo tanto, incluso dentro de un objeto de prueba o de entrenamiento, se espera que las poses equivocadas se cancelarán entre sí, y las poses que están lo suficientemente próximas para implicar algo próximo a la clase de interacciones entre objetos (por ejemplo, unión intermolecular) que surge en la naturaleza, que tales poses serán las que contribuirán a la señal final generada por la pluralidad de poses para un único objeto de prueba o de entrenamiento.

En algunas realizaciones, objetos de entrenamiento 66 y objetos de prueba 72 se anclan mediante o bien técnicas de generación de poses aleatorias o bien mediante generación de poses sesgadas. En algunas realizaciones, objetos de entrenamiento 66 y/u objetos de prueba 72 se anclan mediante muestreo Monte Carlo de cadena de Markov. En algunas realizaciones, tal muestreo permite la total flexibilidad de objetos de entrenamiento y/u objetos de prueba en los cálculos de anclaje y una función de puntuación que es la suma de la energía de interacción entre el objeto de entrenamiento (o prueba) y el objeto objetivo 58 así como la energía conformacional del objeto de entrenamiento (o prueba). Véase, por ejemplo, Liu y Wang, 1999, "MCDOCK: A Monte Carlo simulation approach to the molecular docking problem", Journal of Computer-Aided Molecular Design 13, 435-451.

En algunas realizaciones, se usan algoritmos tal como DOCK (Shoichet, Bodian, y Kuntz, 1992, "Molecular docking using shape descriptors", Journal of Computational Chemistry 13(3), páginas 380-397; y Knegtel, Kuntz, y Oshiro, 1997 "Molecular docking to ensembles of protein structures", Journal of Molecular Biology 266, páginas 424-440) para encontrar una pluralidad de poses para cada uno de los objetos de prueba 72 y/u objetos de entrenamiento 66 contra cada uno de los objetos objetivos 58. Tales algoritmos modelan el objeto objetivo y el objeto de prueba (o entrenamiento) como cuerpos rígidos. La conformación anclada se busca usando superficie complementaria para encontrar poses.

En algunas realizaciones, se usan algoritmos tal como AutoDOCK (Morris et al., 2009, "AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility", J. Comput. Chem. 30(16), páginas 2785-2791; Sotriffer et al., 2000, "Automated docking of ligands to antibodies: methods and applications", Methods: A Companion to Methods in Enzymology 20, páginas 280-291; y "Morris et al., 1998, "Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function", Journal of Computational Chemistry 19: páginas 1639-1662) para encontrar una pluralidad de poses para cada uno de los objetos de prueba 72 y/u objetos de entrenamiento 66 contra cada uno de los objetos objetivos 58. AutoDOCK usa a modelo cinemático del ligando y soporta Monte Carlo, recocido simulado, el Algoritmo Genético de Lamarckian y algoritmos genéticos. Por consiguiente, en algunas realizaciones la pluralidad de diferentes poses (para un par de objeto de prueba - objeto objetivo dado o par de objeto de entrenamiento - objeto de prueba dado) se obtienen mediante muestreo Monte Carlo de cadena de Markov, recocido simulado, Algoritmos Genéticos de Lamarckian o algoritmos genéticos, usando una función de puntuación de anclaje (220).

En algunas realizaciones, se usan algoritmos tal como FlexX (Rarey et al., 1996, "A Fast Flexible Docking Method Using an Incremental Construction Algorithm", Journal of Molecular Biology 261, páginas 470-489) para encontrar una pluralidad de poses para cada uno de los objetos de prueba 72 y/u objetos de entrenamiento 66 contra cada uno de los objetos objetivos 58. FlexX hace una construcción incremental del objeto de prueba 72 y/u objeto de entrenamiento 66 en el sitio activo de un objeto objetivo 58 usando un algoritmo voraz. Por consiguiente, en algunas realizaciones la pluralidad de diferentes poses (para un par de objeto de prueba - objeto objetivo dado o par de objeto de entrenamiento - objeto de prueba dado) se obtienen mediante un algoritmo voraz (222).

En algunas realizaciones, se usan algoritmos tal como GOLD (Jones et al., 1997, "Development and Validation of a Genetic Algorithm for flexible Docking", Journal Molecular Biology 267, páginas 727-748) para encontrar una pluralidad de poses para cada uno de los objetos de prueba 72 y/u objetos de entrenamiento 66 contra cada uno de los objetos objetivos 58. GOLD significa Optimización Genética para Anclaje de Ligando. GOLD construye una red de enlace de hidrógeno genéticamente optimizada entre el objeto de prueba 72 y/u objeto de entrenamiento 66 y el objeto objetivo 58.

En algunas realizaciones, la modelización comprende realizar una ejecución de dinámica molecular del objeto objetivo y el objeto de prueba. Durante la ejecución de dinámica molecular, se permite que los átomos del objeto objetivo y el objeto de prueba interactúen durante un periodo de tiempo fijo, proporcionando una vista de la evolución dinámica del sistema. La trayectoria de átomos en el objeto objetivo y el objeto de prueba (u objeto de entrenamiento) se determina resolviendo numéricamente las ecuaciones de movimiento de Newton para un sistema

de partículas interactivas, en la que fuerzas entre las partículas y sus energías potenciales se calculan usando potenciales interatómicos o campos de fuerzas de mecánica molecular. Véase Alder y Wainwright, 1959, "Studies in Molecular Dynamics. I. General Method", J. Chem. Phys. 31 (2): 459; y Bibcode, 1959, J.Ch.Ph. 31, 459A, doi:10,1063/1.1730376. Por lo tanto, de esta manera, la ejecución de dinámica molecular produce una trayectoria del objeto objetivo y el objeto de prueba juntos con el paso del tiempo. Esta trayectoria comprende la trayectoria de los átomos en el objeto objetivo y el objeto de prueba. En algunas realizaciones, se obtiene un subconjunto de la pluralidad de diferentes poses tomando instantáneas de esta trayectoria en un periodo de tiempo. En algunas realizaciones, poses se obtienen a partir de instantáneas de varias trayectorias diferentes, en las que cada trayectoria comprende una ejecución de dinámica molecular diferente del objeto objetivo que interactúa con el objeto de prueba. En algunas realizaciones, antes de una ejecución de dinámica molecular, un objeto de prueba (o un objeto de entrenamiento) se ancla primero en un sitio activo del objeto objetivo usando una técnica de anclaje.

Independientemente de qué método de modelización se use, lo que se consigue para cualquier par dado de objeto de prueba 72 / objeto de entrenamiento 66 - objeto objetivo 58 es un conjunto diverso de poses del objeto de prueba / entrenamiento con el objeto objetivo con la expectativa de que una o más de las poses esté lo suficientemente próxima a la pose producida naturalmente para demostrar alguna de las interacciones intermoleculares relevantes entre el par dado de objeto de prueba 72 / objeto de entrenamiento 66 - objeto objetivo 58.

En algunas realizaciones se genera una pose inicial del objeto de prueba u objeto de entrenamiento en el sitio activo de un objeto objetivo 58 usando cualquiera de las técnicas anteriormente descritas y se generan poses adicionales a través de la aplicación de alguna combinación de rotación, traslación, y operadores de espejo en cualquier combinación de los tres planos X, Y y Z. Rotación y traslación del objeto de prueba o de entrenamiento puede seleccionarse aleatoriamente (dentro de algún intervalo, por ejemplo más o menos 5 Å del origen) o generarse de forma uniformemente en algún incremento especificado previamente (por ejemplo, todos incrementos de 5 grados alrededor del círculo). La Figura 3 proporciona una ilustración de muestra de un objeto de prueba 72 en dos diferentes poses 302 en el sitio activo de un objeto objetivo 58.

Crear un mapa de vóxel. Haciendo referencia al elemento 224 de la Figura 2B, después de la generación de cada una de las poses para cada de objetos de objetivo y/o de prueba, se crea un mapa de vóxel 40 de cada pose. En algunas realizaciones, cada respectivo mapa de vóxel 40 en la pluralidad de mapas de vóxel se crea mediante un método que comprende: (i) muestrear el objeto de prueba 72 (u objeto de entrenamiento 68), en una respectiva pose en la pluralidad de diferentes poses, y el objeto objetivo 58 sobre una base de cuadrícula tridimensional, formando de este modo un correspondiente panal de relleno de espacio uniforme tridimensional que comprende una correspondiente pluralidad de células poliédricas (tridimensionales) de relleno de espacio y (ii) rellenar, para cada respectiva célula poliédrica tridimensional en la correspondiente pluralidad de células tridimensionales, un vóxel (conjunto discreto de células poliédricas espaciadas de forma regular) en el respectivo mapa de vóxel 40 basándose en una propiedad (por ejemplo, propiedad química) de la respectiva célula poliédrica tridimensional (226). Por lo tanto, si un objeto de prueba particular tiene diez poses en relación con un objeto objetivo, se crean diez correspondientes mapas de vóxel, si un objeto de prueba particular tiene cien poses en relación con un objeto objetivo, se crean cien correspondientes mapas de vóxel, y así sucesivamente. Ejemplos de panales de relleno de espacio incluyen panales cúbicos con células de paralelepípedo, panales prismáticos hexagonales con células de prisma hexagonales, dodecaedros rómbicos con células de dodecaedro rómbico, dodecaedros alargados con células de dodecaedro alargados y octaedros truncados con células de octaedro truncado.

En algunas realizaciones, el panal de relleno de espacio es un panal cúbico con células cúbicas y las dimensiones de tales vóxeles determinan su resolución. Por ejemplo, puede elegirse una resolución de 1 Å que significa que cada vóxel, en tales realizaciones, representa un correspondiente cubo de los datos geométricos con dimensiones de 1 Å (por ejemplo, 1 Å x 1 Å x 1 Å en la respectiva altura, anchura y profundidad de las respectivas células). Sin embargo, en algunas realizaciones, se usa espaciado de cuadrícula más fino (por ejemplo, 0,1 Å o incluso 0,01 Å) o espaciado de cuadrícula más grueso (por ejemplo, 4 Å), en las que el espaciado produce un número entero de vóxeles para cubrir los datos geométricos de entrada. En algunas realizaciones, el muestreo se produce a una resolución que está entre 0,1 Å y 10 Å (227). Como una ilustración, para un cubo de entrada de 40 Å, con una resolución de 1 Å, una disposición de este tipo produciría 40 * 40 * 40 = 64.000 vóxeles de entrada.

En algunas realizaciones, el objeto de prueba 72 (u objeto de entrenamiento 66) es un primer compuesto y el objeto objetivo 58 es un segundo compuesto, una característica de un átomo generado en el muestreo (i) se sitúa en un único vóxel en el respectivo mapa de vóxel mediante el rellenado (ii), y cada vóxel en la pluralidad de vóxeles representa una característica de un máximo de un átomo (228). En algunas realizaciones, la característica del átomo consta de una enumeración del tipo de átomo (230). Como un ejemplo, para datos biológicos, algunas realizaciones de los sistemas y métodos divulgados se configuran para representar la presencia de cada átomo en un vóxel dado del mapa de vóxel 40 como un número diferente para esa entrada, por ejemplo, si un carbono está en un vóxel, se asigna un valor de 6 a ese vóxel porque el número atómico del carbono es 6. Sin embargo, una codificación de este tipo podría implicar que átomos con número atómicos próximos se comportarán de manera similar, que puede no ser particularmente útil dependiendo de la aplicación. Además, el comportamiento de elemento puede ser más similar dentro de grupos (columnas en la tabla periódica) y, por lo tanto, una codificación de este tipo plantea trabajo adicional para que decodifique la red neural convolucional 24.

En algunas realizaciones, la característica del átomo se codifica en el vóxel como una variable categórica binaria (232). En tales realizaciones, tipos de átomo se codifican en lo que se denomina como una codificación "*one-hot*": cada tipo de átomo tiene un canal separado. Por lo tanto, en tales realizaciones, cada vóxel tiene una pluralidad de canales y al menos un subconjunto de la pluralidad de canales representan tipos de átomo. Por ejemplo, un canal dentro de cada vóxel puede representar carbono mientras que otro canal dentro de cada vóxel puede representar oxígeno. Cuando se encuentra un tipo de átomo dado en el elemento de cuadrícula tridimensional que corresponde a un vóxel dado, el canal para ese tipo de átomo dentro del vóxel dado se asigna un primer valor de la variable categórica binaria, tal como "1", y cuando el tipo de átomo no se encuentra en el elemento de cuadrícula tridimensional que corresponde al vóxel dado, el canal para ese tipo de átomo se asigna un segundo valor de la variable categórica binaria, tal como "0" dentro del vóxel dado.

Mientras existen más de 100 elementos, la mayoría no se encuentran en biología. Sin embargo, incluso representando los elementos biológicos más comunes (es decir, H, C, N, O, F, P, S, Cl, Br, I, Li, Na, Mg, K, Ca, Mn, Fe, Co, Zn) puede producir 18 canales por vóxel o 10.483 * 18 = 188.694 entradas en el campo receptor. Como tal, en algunas realizaciones, cada respectivo vóxel en un mapa de vóxel 40 en la pluralidad de mapas de vóxel comprende una pluralidad de canales, y cada canal en la pluralidad de canales representa una propiedad diferente que puede surgir en la célula poliédrica de relleno de espacio tridimensional que corresponde al respectivo vóxel (233). El número de posibles canales para un vóxel dado es incluso mayor en aquellas realizaciones en las que características adicionales de los átomos (por ejemplo, carga parcial, presencia en ligando frente a objetivo de proteína, electronegatividad, o tipo de átomo SYBYL) se presentan adicionalmente como canales independientes para cada vóxel, necesitando más canales de entrada para diferenciar entre átomos de otra manera equivalentes.

En algunas realizaciones, cada vóxel tiene cinco o más canales de entrada (234). En algunas realizaciones, cada vóxel tiene quince o más canales de entrada (236). En algunas realizaciones, cada vóxel tiene veinte o más canales de entrada, veinticinco o más canales de entrada, treinta o más canales de entrada, cincuenta o más canales de entrada o cien o más canales de entrada. En algunas realizaciones, cada vóxel tiene cinco o más canales de entrada seleccionados a partir de los descriptores encontrados en la Tabla 1 a continuación (240). Por ejemplo, en algunas realizaciones, cada vóxel tiene cinco o más canales, cada uno codificado como una variable categórica binaria en la que cada tal canal representa un tipo de átomo SYBYL seleccionado a partir de la Tabla 1 a continuación. Por ejemplo, en algunas realizaciones, cada respectivo vóxel en un mapa de vóxel 40 incluye un canal para el tipo de átomo C.3 (carbono sp3) significando que si la cuadrícula en espacio para un complejo de objeto de prueba - objeto objetivo dado (u objeto de entrenamiento - objeto objetivo) representado por el respectivo vóxel incluye un carbono sp3, el canal adopta un primer valor (por ejemplo, "1") y de lo contrario es un segundo valor (por ejemplo, "0").

Tabla 1 - tipos de átomo SYBYL

TIPO DE ÁTOMO SYBYL	DESCRIPCIÓN
C.3	carbono sp3
C.2	carbono sp2
C.ar	carbono aromático
C.1	carbono sp
N.3	nitrógeno sp3
N.2	nitrógeno sp2
N.1	nitrógeno sp
O.3	oxígeno sp3
O.2	oxígeno sp2
S.3	azufre sp3
N.ar	nitrógeno aromático
P.3	fósforo sp3
H	hidrógeno
Br	bromo
Cl	cloro
F	flúor
I	yodo
S.2	azufre sp2
N.pl3	nitrógeno plano trigonal pl3
LP	par único
Na	sodio
K	potasio
Ca	calcio
Li	litio
Al aluminio	aluminio
Si	silicio

(continuación)

TIPO DE ÁTOMO SYBYL	DESCRIPCIÓN
N.am	nitrógeno de amida
S.o	azufre de sulfóxido
S.o2	azufre de sulfona
N.4	cargado positivamente
	nitrógeno
O.co2	oxígeno en grupo carboxilato o fosfato
C.cat	carbocatión, usado únicamente en un grupo guanidina
H.spc	hidrógeno en modelo de agua SPC
O.spc	oxígeno en modelo de agua SPC
H.t3p	hidrógeno en modelo de agua TIP3P
O.t3p	oxígeno en modelo de agua TIP3P
ANY	cualquier átomo
HEV	átomo (no H) pesado
HET	heteroátomo (N, O, S, P)
HAL	halógeno
Mg	magnesio
Cr.oh	hidroxi cromo
Cr.th	cromo
Se	selenio
Fe	hierro
Cu	cobre
Zn	cinc
Sn	estaño
Mo	molibdeno
Mn	manganeso
Co.oh	hidroxi cobalto

En algunas realizaciones, cada vóxel comprende diez o más canales de entrada, quince o más canales de entrada o veinte o más canales de entrada seleccionados a partir de los descriptores encontrados en la Tabla 1 anterior. En algunas realizaciones, cada vóxel incluye un canal para halógenos.

En algunas realizaciones, se genera una puntuación de huella de interacción de proteína-ligando estructural (SPLIF) para cada pose de un objeto de prueba dado (u objeto de entrenamiento) a un objeto objetivo y esta puntuación de SPLIF se usa como entrada adicional en la red neural subyacente o se codifica individualmente en el mapa de vóxel. Para una descripción de SPLIF, séase Da y Kireev, 2014, J. Chem. Inf. Model. 54, páginas 2555-2561, "Structural Protein-Ligand Interaction Fingerprints (SPLIF) for Structure- Based Virtual Screening: Method and Benchmark Study". Una SPLIF codifica implícitamente todos los posibles tipos de interacciones que pueden producirse entre fragmentos interactivos del objeto de prueba (o entrenamiento) y el objeto objetivo (por ejemplo, π-π, CH-π, etc.). En la primera etapa, se inspecciona un complejo (pose) de objeto de prueba (o entrenamiento) - objeto objetivo para contactos intermoleculares. Dos átomos se consideran que están en un contacto si la distancia entre ellos está dentro de un umbral especificado (por ejemplo, dentro de 4,5 Å). Para cada tal par de átomos intermoleculares, el respectivo átomo de prueba (o entrenamiento) y átomos de objeto objetivo se expanden a fragmentos circulares, por ejemplo, fragmentos que incluyen los átomos en cuestión y sus vecindades sucesivas hasta una cierta distancia. Cada tipo de fragmento circular se asigna un identificador. En algunas realizaciones, tales identificadores se codifican en canales individuales en los respectivos vóxeles. En algunas realizaciones, pueden usarse las Huellas de Conectividad Extendida hasta el primer vecino más próximo (ECFP2) según se definen en el software Pipeline Pilot. Véase, Pipeline Pilot, ver. 8.5, Accelrys Software Inc., 2009. ECFP retiene información acerca de todos los tipos de átomos/uniones y usa un identificador de número entero único para representar una subestructura (es decir, fragmento circular). La huella SPLIF codifica todos los identificadores de fragmento circular encontrados. En algunas realizaciones, la huella SPLIF no codifica vóxeles individuales, sino que sirve como una entrada independiente separada en la red neural convolucional 24 analizada a continuación.

En algunas realizaciones, en lugar de o además de SPLIF, se calculan huellas de interacción estructural (SIFt) para cada pose de un objeto de prueba dado (u objeto de entrenamiento) a un objeto objetivo y proporcionan independientemente como entrada en la red neural convolucional 24 analizada a continuación o se codifican en el mapa de vóxel. Para un cálculo de SIFt, véase Deng et al., 2003, "Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein- Ligand Binding Interactions", J. Med. Chem. 47 (2), páginas 337-344.

En algunas realizaciones, en lugar de o además de SPLIF y SIFT, se calculan fragmentos de interacción basada en pares de átomos (APIF) para cada pose de un objeto de prueba dado (u objeto de entrenamiento) a un objeto objetivo y proporcionan independientemente como entrada en la red neural convolucional 24 analizada a continuación o se codifica individualmente en el mapa de vóxel. Para un cálculo de APIF, véase Pérez-Nueno et al., 2009, "APIF: a new interaction fingerprint based on atom pairs and its application to virtual screening", J. Chem. Inf. Model. 49(5), páginas 1245-1260.

La representación de datos puede codificarse con los datos biológicos en una forma que habilita la expresión de diversas relaciones estructurales asociadas con moléculas/proteínas por ejemplo. La representación geométrica puede implementarse en diversas formas y topografías, de acuerdo con diversas realizaciones. La representación geométrica se usa para la visualización y análisis de datos. Por ejemplo, en una realización, geometrías pueden representarse usando vóxeles dispuestos en diversas topografías, tal como espacio cartesiano / euclidiano 2-D, 3-D, espacio no euclidiano 3-D, distribuidores, etc. Por ejemplo, la Figura 4 ilustra una estructura de cuadrícula tridimensional 400 de muestra que incluye una serie de subcontenedores, de acuerdo con una realización. Cada subcontenedor 402 puede corresponder a un vóxel. Puede definirse un sistema de coordenadas para la cuadrícula, de tal forma que cada subcontenedor tiene un identificador. En algunas realizaciones de los sistemas y métodos divulgados, el sistema de coordenadas es un sistema cartesiano en espacio 3-D, pero en otras realizaciones del sistema, el sistema de coordenadas puede ser cualquier otro tipo de sistema de coordenadas, tal como un esferoide achatado, sistemas de coordenadas cilíndricas o esféricas, sistemas de coordenadas polares, otros sistemas de coordenadas diseñados para diversos espacios de distribuidores y vectores, entre otros. En algunas realizaciones, los vóxeles pueden tener valores particulares asociados a los mismos, que pueden, por ejemplo, representarse aplicando etiquetas, y/o determinando su posicionamiento, entre otros.

Porque redes neurales requieren un tamaño de entrada fijo, algunas realizaciones de los sistemas y métodos divulgados recortan los datos geométricos (el complejo de objeto objetivo-de prueba u objetivo-de entrenamiento) para ajustarse dentro de un cuadro delimitador apropiado. Por ejemplo, puede usarse un cubo de 25 - 40Å a un lado. En algunas realizaciones en las que objetos de objetivo y/o de prueba se han anclado en el sitio activo de objetos objetivos 58, el centro del sitio activo sirve como el centro del cubo.

Mientras en algunas realizaciones se usa un cubo cuadrado de dimensiones fijas centrado en el sitio activo del objeto objetivo para dividir el espacio en la cuadrícula de vóxel, los sistemas divulgados no se limitan de esta forma. En algunas realizaciones, se usa cualquiera de diversas formas para dividir el espacio en la cuadrícula de vóxel. En algunas realizaciones, se usan poliedros, tal como primas rectangulares, formas de poliedros, etc. para dividir el espacio.

En una realización, la estructura de cuadrícula puede configurarse para ser similar a una disposición de vóxeles. Por ejemplo, cada subestructura puede asociarse con un canal para cada átomo que se analiza. También, puede proporcionarse un método de codificación para representar cada átomo numéricamente.

En algunas realizaciones, el mapa de vóxel tiene en cuenta el factor de tiempo y puede ser, por lo tanto, en cuatro dimensiones (X, Y, Z y tiempo).

En algunas realizaciones, en lugar de vóxeles pueden usarse otras implementaciones tal como píxeles, puntos, formas poligonales, poliedros, o cualquier otro tipo de forma en múltiples dimensiones (por ejemplo, formas en 3D, 4D, y así sucesivamente).

En algunas realizaciones, los datos geométricos se normalizan eligiendo que el origen de las coordenadas X, Y y Z sea el centro de masa de un sitio de unión del objeto objetivo según se determina mediante un algoritmo de inundación de cavidad (256). Para detalles representativos de tales algoritmos, véase Ho y Marshall, 1990, "Cavity search: An algorithm for the isolation and display of cavity-like binding regions", Journal of Computer-Aided Molecular Design 4, páginas 337-354; y Hendlich et al., 1997, "Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins", J. Mol. Graph. Model 15, n.º 6. Como alternativa, en algunas realizaciones, el origen del mapa de vóxel se centra en el centro de masa de todo el cocomplejo (del objeto de prueba unido al objeto objetivo u objeto de entrenamiento unido al objeto objetivo, de solo el objeto objetivo, o de solo el objeto de prueba u objeto de entrenamiento). Los vectores de base pueden elegirse opcionalmente que sean los momentos principales de inercia de todo el cocomplejo, de solo el objetivo, o de solo el objeto de prueba / objeto de entrenamiento. En algunas realizaciones, el objeto objetivo 58 es un polímero que tiene un sitio activo, y el muestreo muestrea el objeto de prueba 72 (u objeto de entrenamiento 66), en cada una de las respectivas poses en la pluralidad de diferentes poses para el objeto de prueba 72 (u objeto de entrenamiento 66), y el sitio activo sobre la base de cuadrícula tridimensional en la que se toma un centro de masa del sitio activo como el origen y el correspondiente panal uniforme tridimensional para el muestreo representa una porción del polímero y el objeto de prueba 72 (u objeto de entrenamiento 66) centrado en el centro de masa (248). En algunas realizaciones, el panal uniforme es un panal cúbico regular y la porción del polímero y el objeto de prueba es un cubo de dimensiones fijas predeterminadas. El uso de un cubo de dimensiones fijas predeterminadas, en tales realizaciones, asegura que se usa una porción relevante de los datos geométricos y que cada mapa de vóxel es del mismo tamaño. En algunas realizaciones, las dimensiones fijas predeterminadas del cubo son N Å x N Å x N Å, donde N es un número entero o valor real entre 5

y 100, un número entero entre 8 y 50 o un número entero entre 15 y 40 (250, 252). En algunas realizaciones, el panal uniforme es un panal de prisma rectangular y la porción del polímero y el objeto de prueba es un prisma rectangular de dimensiones fijas predeterminadas Q Å x R Å x S Å, en el que Q es un primer número entero entre 5 y 100, R es un segundo número entero entre 5 y 100, S es un tercer número entero o valor real entre 5 y 100, y al menos un número en el conjunto {Q, R, S} no es igual a otro valor en el conjunto {Q, R, S}.

En una realización, cada vóxel tiene uno o más canales de entrada, que pueden tener diversos valores asociados a los mismos, que en una implementación simple podrían ser encendido/apagado, y pueden configurarse para codificar para un tipo de átomo. Tipos de átomo pueden indicar el elemento del átomo, o tipos de átomo pueden refinarse adicionalmente para distinguir entre otras características de átomo. Átomos presentes pueden codificarse a continuación en cada vóxel. Diversos tipos de codificación pueden utilizarse usando diversas técnicas y/o metodologías. Como un método de codificación de ejemplo, puede utilizarse el número atómico del átomo, produciendo un valor por vóxel que oscila desde uno para hidrógeno hasta 118 para oganesón (o cualquier otro elemento).

Sin embargo, como se ha analizado anteriormente, pueden utilizarse otros métodos de codificación, tal como "codificación *one-hot*", en la que cada vóxel tiene muchos canales de entrada paralelos, cada uno de los cuales está o bien apagado o encendido y codifica para un tipo de átomo. Tipos de átomo pueden indicar el elemento del átomo, o tipos de átomo pueden refinarse adicionalmente para distinguir entre otras características de átomo. Por ejemplo, tipos de átomo SYBYL distinguen carbonos de un enlace sencillo de los de un enlace doble, enlace triple o carbonos aromáticos. Para tipos de átomo SYBYL, véase Clark et al., 1989, "Validation of the General Purpose Tripos Force Field, 1989, J. Comput. Chem. 10, páginas 982-1012.

En algunas realizaciones, cada vóxel incluye adicionalmente uno o más canales para distinguir entre átomos que son parte del objeto objetivo 58 o cofactores frente a parte del objeto de prueba 72 u objeto de entrenamiento 66. Por ejemplo, en una realización, cada vóxel incluye adicionalmente un primer canal para el objeto objetivo 58 y un segundo canal para el objeto de prueba 72 u objeto de entrenamiento 66 (238). Cuando un átomo en la porción de espacio representado por el vóxel es del objeto objetivo 58, el primer canal se establece a un valor, tal como "1", y de lo contrario es cero (por ejemplo, porque la porción de espacio representado por el vóxel no incluye ningún átomo o uno o más átomos del objeto de prueba 72 u objeto de entrenamiento 66). Además, cuando un átomo en la porción de espacio representado por el vóxel es del objeto de prueba 72 u objeto de entrenamiento 66, el segundo canal se establece a un valor, tal como "1", y de lo contrario es cero (por ejemplo, porque la porción de espacio representado por el vóxel no incluye ningún átomo o uno o más átomos del objeto objetivo 58). Análogamente, otros canales pueden especificar adicionalmente (o como alternativa) información adicional tal como carga parcial, polarización, electronegatividad, espacio solvente accesible y densidad de electrones. Por ejemplo, en algunas realizaciones, un mapa de densidad de electrones para el objeto objetivo solapa el conjunto de coordinadas tridimensionales, y la creación del mapa de vóxel adicionalmente muestrea el mapa de densidad de electrones (258). Ejemplos de mapas de densidad de electrones adecuados incluyen, pero sin limitación, mapas de sustitución isomorfa múltiple, mapas de sustitución isomorfa única con señal anómala, mapas de dispersión anómala de longitud de onda única, mapas de dispersión anómala de longitud de onda múltiple y mapas 2Fo-Fc (260). Véase McRee, 1993, Practical Protein Crystallography, Academic Press.

En algunas realizaciones, codificación de vóxel de acuerdo con los sistemas y métodos divulgados puede incluir refinamientos de codificación opcional adicionales. Los siguientes dos se proporcionan como ejemplos.

En un primer refinamiento de codificación, la memoria requerida puede reducirse reduciendo el conjunto de átomos representados por un vóxel (por ejemplo, reduciendo el número de canales representados por un vóxel) sobre la base de que la mayoría de elementos se producen raramente en sistemas biológicos. Átomos pueden correlacionarse para compartir el mismo canal en un vóxel, o bien combinando átomos raros (que pueden impactar raramente, por lo tanto, en el rendimiento del sistema) o combinando átomos con propiedades similares (que, por lo tanto, podrían minimizar la imprecisión a partir de la combinación).

Un refinamiento de codificación es tener vóxeles representan posiciones de átomo activando parcialmente vóxeles vecinos. Esto resulta en activación parcial de neuronas vecinas en la red neural posterior y se aleja desde codificación *one-hot* a codificación "*several-warm*". Por ejemplo, puede ser ilustrativo considerar un átomo de cloro, que tiene un diámetro de van der Waals de 3,5 Å y, por lo tanto, un volumen de 22,4 $Å^3$ cuando se sitúa una cuadrícula de 1 $Å^3$, vóxeles dentro del átomo de cloro se llenarán completamente y vóxeles en el borde del átomo únicamente se rellenarán parcialmente. Por lo tanto, el canal que representa cloro en los vóxeles parcialmente rellenos se encenderán en proporción a la cantidad de tales vóxeles que se encuentran dentro del átomo de cloro. Por ejemplo, si el cincuenta por ciento del volumen de vóxel se encuentra dentro del átomo de cloro, el canal en el vóxel que representa cloro se activará al cincuenta por ciento. Esto puede resultar en una representación "suavizada" y más precisa en relación con la codificación *one-hot* discreta. Por lo tanto, en algunas realizaciones, el objeto de prueba es un primer compuesto y el objeto objetivo es un segundo compuesto, una característica de un átomo generado en el muestreo se extiende a través de un subconjunto de vóxeles en el respectivo mapa de vóxel 40 y este subconjunto de vóxeles comprende dos o más vóxeles, tres o más vóxeles, cinco o más vóxeles, diez o más vóxeles o veinte-cinco o más vóxeles (242). En algunas realizaciones, la característica del átomo consta de una

enumeración del tipo de átomo (244) (por ejemplo, uno de los tipos de átomo SYBYL).

Por lo tanto, voxelación (rasterización) de los datos geométricos (el anclaje de un objeto de prueba o de entrenamiento en un objeto objetivo) que se ha codificado se basa en diversas reglas aplicadas a los datos de entrada.

La Figura 5 y la Figura 6 proporcionan vistas de dos moléculas 502 codificadas en una cuadrícula bidimensional 500 de vóxeles, de acuerdo con algunas realizaciones. La Figura 5 proporciona las dos moléculas superpuestas en la cuadrícula bidimensional. La Figura 6 proporciona la codificación *one-hot*, usando los diferentes patrones de sombreado para codificar respectivamente la presencia de oxígeno, nitrógeno, carbono y espacio vacío. Como se ha indicado anteriormente, tal codificación puede denominarse como codificación "*one-hot*". La Figura 6 muestra la cuadrícula 500 de la Figura 5 con las moléculas 502 omitidas. La Figura 7 proporciona una vista de la cuadrícula bidimensional de vóxeles de la Figura 6, en la que los vóxeles se han numerado.

En algunas realizaciones, geometría característica se representa en formas distintas de vóxeles. La Figura 8 proporciona una vista de diversas representaciones en las que características (por ejemplo, centros de átomo) se representan como puntos 0-D (representación 802), puntos 1-D (representación 804), puntos 2-D (representación 806) o puntos 3-D (representación 808). Inicialmente, el espaciado entre los puntos puede elegirse aleatoriamente. Sin embargo, a medida que el modelo predictivo se entrena, los puntos pueden moverse más cerda juntos, o más separados. La Figura 9 ilustra un intervalo de posibles posiciones para cada punto.

Desplegar un mapa de vóxel en un correspondiente vector. Haciendo referencia al elemento 262, cada mapa de vóxel 40 se despliega opcionalmente en un correspondiente vector, creando de este modo una pluralidad de vectores, en la que cada vector en la pluralidad de vectores es del mismo tamaño. En algunas realizaciones, cada vector en la pluralidad de vectores es un vector unidimensional (264). Por ejemplo, en algunas realizaciones, un cubo de 20 Å en cada lado se centra en el sitio activo del objeto objetivo 58 y se muestrea con un espaciado de cuadrícula fijada tridimensional de 1 Å para formar correspondientes vóxeles de un mapa de vóxel que mantienen en respectivos canales básicos de las características estructurales de vóxel tal como tipos de átomo así como, opcionalmente, descriptores de objeto de prueba - objeto objetivo más complejos, como se ha analizado anteriormente. En algunas realizaciones, los vóxeles de este mapa de vóxel tridimensional se despliegan en un vector de punto de flotación unidimensional.

Someter un vector 22 a una red neural convolucional. Haciendo referencia al elemento 266 de la Figura 2, la representación vectorizada de mapas de vóxel 22 se someten a una red convolucional 24. En algunas realizaciones, como se ilustra en la Figura 2, la representación vectorizada de mapas de vóxel 22 se almacenan en la memoria 52 junto con un módulo de evaluación convolucional 20 y una red neural convolucional 24. Esto proporciona la ventaja de procesar la representación vectorizada de mapas de vóxel 22 a través de la red neural convolucional 24 a velocidades más rápidas. Sin embargo, en otras realizaciones, cualquiera o todas las representaciones vectorizadas de mapas de vóxel 22, el módulo de evaluación convolucional 20 y la red neural convolucional 24 están en la memoria 92 del sistema 100 o simplemente son accesibles por el sistema 92 a través de una red. En algunas realizaciones cualquiera o todas las representaciones vectorizadas de mapas de vóxel 22, el módulo de evaluación convolucional 20 y la red neural convolucional 24 están en un entorno informático en la nube.

En algunas realizaciones, la pluralidad de vectores 22 se proporciona a la memoria de unidad de procesamiento gráfico 52, en la que la memoria de unidad de procesamiento gráfico incluye una arquitectura de red que incluye una red neural convolucional 24 que comprende una capa de entrada 26 para recibir secuencialmente la pluralidad de vectores, una pluralidad de capas convolucionales 28 y un calificador 30. La pluralidad de capas convolucionales incluye una capa convolucional inicial y una capa convolucional final. En algunas realizaciones, la red neural convolucional 24 no está en la memoria de GPU, sino que está en la memoria de fin general del sistema 100. En algunas realizaciones, los mapas de vóxel no se vectorizan antes de introducirse en la red 24.

En algunas realizaciones, una capa convolucional 28 en la pluralidad de capas convolucionales comprende un conjunto de filtros asimilables (también denominados núcleos). Cada filtro tiene tamaño tridimensional fijo que convoluciona (escalona a una tasa de escalonado predeterminada) a través de la profundidad, altura y anchura del volumen de entrada de la capa convolucional, que calcula un producto escalar (u otras funciones) entre entradas (pesos) del filtro y la entrada creando de este modo un mapa de activación de múltiples dimensiones de ese filtro. En algunas realizaciones, la tasa de escalonado de filtro es un elemento, dos elementos, tres elementos, cuatro elementos, cinco elementos, seis elementos, siete elementos, ocho elementos, nueve elementos, diez elementos o más de diez elementos del espacio de entrada. Por lo tanto, considérese el caso en el que a filtro tiene tamaño 5^3. En algunas realizaciones, este filtro calculará el producto escalar (u otra función matemática) entre un cubo contiguo de espacio de entrada que tiene una profundidad de cinco elementos, una anchura de cinco elementos y una altura de cinco elementos, para un número total de valores de espacio de entrada de 125 por vóxel canal.

El espacio de entrada a la capa convolucional inicial (por ejemplo, la salida desde la capa de entrada 26) se forma o bien desde un mapa de vóxel 40 o bien una representación vectorizada del mapa de vóxel 22. En algunas realizaciones, la representación vectorizada del mapa de vóxel es una representación vectorizada unidimensional del

mapa de vóxel que sirve como el espacio de entrada a la capa convolucional inicial. Sin embargo, cuando un filtro convoluciona su espacio de entrada y el espacio de entrada es una representación vectorizada unidimensional del mapa de vóxel, el filtro aún obtiene a partir de la representación vectorizada unidimensional esos elementos que representan un correspondiente cubo contiguo de espacio fijo en el complejo de objeto objetivo - objeto de prueba (o entrenamiento). En algunas realizaciones, el filtro usa técnicas de contabilidad estándar para seleccionar esos elementos de dentro de la representación vectorizada unidimensional que forman el correspondiente cubo contiguo de espacio fijo en el complejo de objeto objetivo - objeto de prueba (o entrenamiento). Por lo tanto, en algunos casos, esto implica necesariamente tomar un subconjunto no contiguo de elemento en la representación vectorizada unidimensional para obtener los valores de elemento del correspondiente cubo contiguo de espacio fijo en el complejo de objeto objetivo - objeto de prueba (o entrenamiento).

En algunas realizaciones, el filtro se inicializa (por ejemplo, a ruido gaussiano) o entrena para tener 125 correspondientes pesos (por canal de entrada) en los que tomar el producto escalar (o alguna otra forma de operación matemática tal como la función divulgada en la Figura 14) de los 125 valores de espacio de entrada para calcular un primer único valor (o conjunto de valores) de la capa de activación que corresponde al filtro. En alguna realización los valores calculados por el filtro se suman, ponderan y/o sesgan. Para calcular valores adicionales de la capa de activación que corresponde al filtro, el filtro se escalona (convoluciona) a continuación en una de las tres dimensiones del volumen de entrada mediante la tasa de escalonado (zancada) asociada con el filtro, en cuyo punto el producto escalar (o alguna otra forma de operación matemática tal como la función matemática divulgada en la Figura 14) entre los pesos de filtro y los 125 valores de espacio de entrada (por canal) se toma como la nueva ubicación se toma en el volumen de entrada. Este escalonamiento (convolución) se repite hasta que el filtro ha muestreo todo el espacio de entrada de acuerdo con la tasa de escalonado. En algunas realizaciones, el borde del espacio de entrada se rellena con ceros para controlar el volumen espacial del espacio de salida producido por la capa convolucional. En realizaciones típicas, cada uno de los filtros de la capa convolucional cubre toda la entrada tridimensional formando de esta manera de este modo un correspondiente mapa de activación. La colección de mapas de activación de los filtros de la capa convolucional forman colectivamente el volumen de salida tridimensional de una capa convolucional, y sirve de este modo como la entrada tridimensional (tres dimensiones espaciales) de una capa convolucional posterior. Cada entrada en el volumen de salida, por lo tanto, también puede interpretarse como una salida de una única neurona (o un conjunto de neuronas) que mira a una pequeña región en el espacio de entrada a la capa convolucional y comparte parámetros con neuronas en el mismo mapa de activación. Por consiguiente, en algunas realizaciones, una capa convolucional en la pluralidad de capas convolucionales tiene una pluralidad de filtros y cada filtro en la pluralidad de filtros convoluciona (en tres dimensiones espaciales) un espacio de entrada cúbico de N^3 con zancada Y, donde N es un número entero de dos o mayor (por ejemplo, 2, 3, 4, 5, 6, 7, 8, 9, 10 o mayor de 10) e Y es un número entero positivo (por ejemplo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o mayor de 10) (268).

Cada capa en la pluralidad de capas convolucionales se asocia con un conjunto diferente de pesos. Con más particularidad, cada capa en la pluralidad de capas convolucionales incluye una pluralidad de filtros y cada filtro comprende una pluralidad independiente de pesos (270). En algunas realizaciones, una capa convolucional tiene 128 filtros de dimensión 5^3 y, por lo tanto, la capa convolucional tiene 128 x 5 x 5 x 5 o 16.000 pesos por canal en el mapa de vóxel. Por lo tanto, si existen cinco canales en el mapa de vóxel, la capa convolucional tendrá 16.000 x 5 pesos, u 80.000 pesos. En algunas realizaciones algunas o todas de tales pesos (y, opcionalmente, sesgos) de cada filtro en una capa convolucional dada pueden atarse juntos, es decir, limitarse para ser idénticos.

En respuesta a entrada de un vector respectivo 22 en la pluralidad de vectores, la capa de entrada 26 proporciona una primera pluralidad de valores en la capa convolucional inicial como una primera función de valores en el vector respectivo, en la que la primera función se calcula opcionalmente usando la unidad de procesamiento gráfico 50.

Cada capa convolucional respectiva 28, distinta de la capa convolucional final, proporciona valores intermedios, como una segunda función respectiva de (i) el conjunto diferente de pesos asociados a la capa convolucional respectiva y (ii) valores de entrada recibidos por la capa convolucional respectiva, en otra capa convolucional en la pluralidad de capas convolucionales, en la que la segunda función se calcula usando la unidad de procesamiento gráfico 50. Por ejemplo, cada respectivo filtro de la capa convolucional respectiva 28 cubre el volumen de entrada (en tres dimensiones espaciales) a la capa convolucional de acuerdo con la zancada tridimensional característica de la capa convolucional y en cada respectiva posición de filtro, toma el producto escalar (o alguna otra función matemática) de los pesos de filtro del respectivo filtro y los valores del volumen de entrada (cubo contiguo que es un subconjunto del espacio de entrada total) en la respectiva posición de filtro, produciendo de este modo un punto calculado (o un conjunto de puntos) en la capa de activación que corresponde a la respectiva posición de filtro. Las capas de activación de los filtros de la capa convolucional respectiva representan colectivamente los valores intermedios de la capa convolucional respectiva.

La capa convolucional final proporciona valores finales, como una tercera función de (i) el conjunto diferente de pesos asociados a la capa convolucional final y (ii) valores de entrada recibidos por la capa convolucional final que se calcula opcionalmente usando la unidad de procesamiento gráfico 50, en el calificador. Por ejemplo, cada respectivo filtro de la capa convolucional final 28 cubre el volumen de entrada (en tres dimensiones espaciales) a la capa convolucional final de acuerdo con la zancada tridimensional característica de la capa convolucional y en cada

respectiva posición de filtro, toma el producto escalar (o alguna otra función matemática) de los pesos de filtro del filtro y los valores del volumen de entrada en la respectiva posición de filtro, calculando de este modo un punto (o un conjunto de puntos) en la capa de activación que corresponde a la respectiva posición de filtro. Las capas de activación de los filtros de la capa convolucional final representan colectivamente los valores finales que se proporcionan al clasificador 30.

En algunas realizaciones, la red neural convolucional tiene una o más capas de activación. En algunas realizaciones, la capa de activación es una capa de neuronas que aplica la función de activación de no saturación f(x) = max(0, x). aumenta las propiedades no lineales de la función de decisión y de la red general sin afectar a los campos receptivos de la capa de convolución. En otras realizaciones, la capa de activación tiene otras funciones para aumentar la no linealidad, por ejemplo, la función de tangente hiperbólica de saturación f(x) = tanh, f(x) = |tanh(x)|, y la función sigmoide $f(x) = (1 + e^{-x})^{-1}$. Ejemplos no limitantes de otras funciones de activación encontradas en otras capas de activación en algunas realizaciones para la red neuronal pueden incluir, pero sin limitación, logística (o sigmoide), softmax, gaussiana, promedio ponderado de Boltzmann, valor absoluto, lineal, lineal rectificada, lineal rectificada delimitada, lineal rectificada suave, lineal rectificada parametrizada, promedio, máximo, mínimo, alguna norma de vector LP (para p=1, 2, 3,... ,∞), signo, cuadrado, raíz cuadrada, multicuadrática, cuadrática inversa, multicuadrática inversa, spline poliarmónica y spline de placa delgada.

La red 24 aprende filtros dentro de las capas convolucionales 28 que se activan cuando ven algún tipo específico de característica en alguna posición espacial en la entrada. Como se analiza en la sección de entrenamiento de red a continuación, en algunas realizaciones, los pesos iniciales de cada filtro en una capa convolucional se obtienen mediante entrenamiento de la red neural convolucional contra la librería de entrenamiento de objeto 64 como se analiza a continuación. Por consiguiente, la operación de la red neural convolucional 24 puede producir características más complejas que las características usadas históricamente para llevar a cabo predicción de afinidad de unión. Por ejemplo, un filtro en una capa convolucional dada de la red 24 que sirve como un detector de enlace de hidrógeno puede ser capaz de reconocer no únicamente que un donador y aceptor de enlace de hidrógeno están a una distancia y ángulos dados, sino que también reconocen que el entorno bioquímico alrededor del donador y aceptor refuerza o debilita el enlace. Adicionalmente, los filtros dentro de la red 24 pueden entrenarse para discriminar de forma efectiva aglutinantes de no aglutinantes de los datos subyacentes.

En algunas realizaciones, la red neural convolucional 24 se configura para adaptar sistemas dinámicos, tal como las posiciones alternativas que pueden encontrarse tanto como el movimiento del objeto objetivo como de objeto de prueba. En un complejo de objeto objetivo -objeto de prueba de este tipo, pueden adoptarse un número de diferentes configuraciones, con la proporción relativa basándose en la distribución de Boltzmann de la energía libre de cada forma. Tanto los componentes entálpicos y entrópicos de la energía libre del complejo de objeto objetivo -objeto de prueba puede depender de las poses adoptadas por el objeto ($\Delta G = \Delta H - T\Delta S$). La afinidad de unión final puede encontrarse para ser una función del promedio ponderado de las energías del conjunto de poses disponibles para el complejo de objeto objetivo -objeto de prueba. Para modelar este fenómeno físico, la red neural convolucional 24 puede configurarse para muestrear un gran número de posiciones alternativas debido un movimiento de objeto objetivo y objeto de prueba y para basar sus predicciones de afinidad de unión en este conjunto muestreo de configuraciones del complejo (por ejemplo, tomando el promedio ponderado de todas las puntuaciones de la red 24 de estas diversas posiciones alternativas).

Como se describe anteriormente, en algunas realizaciones la red neuronal 24 se configura para desarrollar capas convolucionales tridimensionales. La región de entrada a la capa convolucional 28 de menor nivel puede ser un cubo (u otra región contigua) de canales de vóxel desde el campo receptivo. Las capas convolucionales 28 superiores evalúan la salida de las capas convolucionales inferiores, mientras aún tiene su salida que es una función de una región delimitada de vóxeles que están juntos (en distancia euclidiana 3D).

Actividad biológica puede ser invariante en rotación, así como traslación, de forma que la red 24 puede configurarse opcionalmente para generar mapas de característica rotados que se aprovechan de las simetrías rotacionales de división de espacio. Por ejemplo, si el sistema se configuró para usar cubos para dividir los datos de entrada, el sistema podría configurarse para generar mapas de característica rotados atando los pesos de los cálculos de función juntos después de una rotación de 90 grados.

Puede ser ilustrativo considerar un cubo que se rota según las agujas del reloj: los pesos en la cara superior de un filtro se atan a los pesos en la cara derecha de un filtro diferente; en otras palabras, los pesos pueden limitarse para ser idénticos. La rotación puede generar 24 mapas de característica rotando según las agujas del reloj 90 grados, 180 grados, 270 grados, para cada uno de los tres planos XY / XZ / YZ. Esta disposición reduce el número de parámetros a 1/24º de sin atadura de peso rotacional, ya que sin atadura de peso cada filtro tiene sus propios pesos.

Como un ejemplo alternativo, si el sistema se configurase para usar otros poliedros para dividir los datos de entrada, el sistema puede configurarse para usar otras rotaciones para acceder a las isometrías apropiadas a sus grupos de simetría. Por ejemplo, donde el espacio se ha dividido usando octaedros truncados, habría 3 ejes de simetría rotacional de 90 grados, 4 ejes de simetría rotacional de 120 grados y seis ejes de simetría de 180 grados.

En una realización, la red 24 se configura para aplicar técnicas de regularización para reducir la tendencia de los modelos para sobreajustar los objetos de entrenamiento 66 y datos de unión de entrenamiento 68.

Cero o más de las capas de red en la red 24 pueden constar de capas de agrupamiento. Como en una capa convolucional, una capa de agrupamiento es un conjunto de cálculos de funciones que aplican la misma función a través de diferentes parches espacialmente locales de entrada. Para capas de agrupamiento, la salida se proporciona mediante operadores de agrupamiento, por ejemplo, alguna norma de vector LP para p=1, 2, 3,..., ∞, en varios vóxeles. El agrupamiento se hace habitualmente por canal, en lugar de a través de canales. El agrupamiento divide el espacio de entrada en un conjunto de cajas tridimensionales y, para cada tal subregión, emite el máximo. La operación de agrupamiento proporciona una forma de invariancia de traslación. La función de la capa de agrupamiento es reducir progresivamente el tamaño espacial de la representación para reducir la cantidad de parámetros y cálculos en la red y, por lo tanto, también para controlar sobreajuste. En algunas realizaciones se inserta una capa de agrupamiento entre capas convolucionales 28 sucesivas en la red 24. Una capa de agrupamiento de este tipo opera independientemente en cada segmento de profundidad de la entrada y redimensiona el mismo espacialmente. Además de agrupamiento máximo, las unidades de agrupamiento también pueden realizar otras funciones, tal como agrupamiento promedio o incluso agrupamiento de norma L2.

Cero o más de las capas en la red 24 pueden constar de capas de normalización, tal como normalización de respuesta local o normalización de contraste local, que pueden aplicarse a través de canales en la misma posición o para un canal particular a través de varias posiciones. Estas capas de normalización pueden fomentar la variedad en la respuesta de varios cálculos de funciones a la misma entrada.

En algunas realizaciones, el calificador 30 comprende una pluralidad de capas totalmente conectadas y una capa de evaluación en la que una capa totalmente conectada en la pluralidad de capas totalmente conectadas se proporciona a la capa de evaluación (272). Neuronas en una capa totalmente conectada tienen conexiones totales a todas las activaciones en la capa previa, como se ve en redes neurales regulares. Sus activaciones pueden calcularse, por lo tanto, con una multiplicación de matriz seguida por una compensación de sesgo. En algunas realizaciones, cada capa totalmente conectada tiene 512 unidades escondidas, 1024 unidades escondidas o 2048 unidades escondidas. En algunas realizaciones no existen capas totalmente conectadas, una capa totalmente conectada, dos capas totalmente conectadas, tres capas totalmente conectadas, cuatro capas totalmente conectadas, cinco capas totalmente conectadas, seis o más capas totalmente conectadas o diez o más capas totalmente conectadas en el calificador.

En algunas realizaciones, la capa de evaluación discrimina entre una pluralidad de clases de actividad. En algunas realizaciones, la capa de evaluación comprende una capa de coste de regresión logística a través de dos clases de actividad, tres clases de actividad, cuatro clases de actividad, cinco clases de actividad o seis o más clases de actividad.

En algunas realizaciones, la capa de evaluación comprende una capa de coste de regresión logística a través de una pluralidad de clases de actividad. En algunas realizaciones, la capa de evaluación comprende una capa de coste de regresión logística a través de dos clases de actividad, tres clases de actividad, cuatro clases de actividad, cinco clases de actividad o seis o más clases de actividad.

En algunas realizaciones, la capa de evaluación discrimina entre dos clases de actividad y la primera clase de actividad (primera clasificación) representa una IC_{50}, EC_{50} o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está por encima de un primer valor de unión, y la segunda clase de actividad (segunda clasificación) es una IC_{50}, EC_{50}, o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está por debajo del primer valor de unión. En algunas realizaciones, el primer valor de unión es uno nanomolar, diez nanomolar, cien nanomolar, uno micromolar, diez micromolar, cien micromolar o uno milimolar.

En algunas realizaciones, la capa de evaluación comprende una capa de coste de regresión logística en dos clases de actividad y la primera clase de actividad (primera clasificación) representa una IC_{50}, EC_{50} o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está por encima de un primer valor de unión, y la segunda clase de actividad (segunda clasificación) es una IC_{50}, EC_{50}, o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está por debajo del primer valor de unión. En algunas realizaciones, el primer valor de unión es uno nanomolar, diez nanomolar, cien nanomolar, uno micromolar, diez micromolar, cien micromolar o milimolar.

En algunas realizaciones, la capa de evaluación discrimina entre tres clases de actividad y la primera clase de actividad (primera clasificación) representa una IC_{50}, EC_{50} o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está por encima de un primer valor de unión, la segunda clase de actividad (segunda clasificación) es una IC_{50}, EC_{50}, o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está entre el primer valor de unión y un segundo valor de unión, y la tercera clase de actividad (tercera clasificación) es una IC_{50}, EC_{50}, o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está por debajo del segundo valor de unión, en la que el primer valor de unión es distinto del segundo valor de unión.

En algunas realizaciones, la capa de evaluación comprende una capa de coste de regresión logística en tres clases de actividad y la primera clase de actividad (primera clasificación) representa una IC_{50}, EC_{50} o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está por encima de un primer valor de unión, la segunda clase de actividad (segunda clasificación) es una IC_{50}, EC_{50}, o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está entre el primer valor de unión y un segundo valor de unión, y la tercera clase de actividad (tercera clasificación) es una IC_{50}, EC_{50}, o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está por debajo del segundo valor de unión, en la que el primer valor de unión es distinto del segundo valor de unión.

En algunas realizaciones, el calificador 30 comprende una única capa totalmente conectada o perceptrón de múltiples capas. En algunas realizaciones el calificador comprende una máquina de vector de soporte, bosques aleatorios, vecino más cercano. En algunas realizaciones, el calificador 30 asigna una puntuación numérica que indica la fuerza (o confianza o probabilidad) de clasificación de la entrada en las diversas categorías de salida. En algunos casos, las categorías son aglutinantes y no aglutinantes o, como alternativa, el nivel de potencia (potencias de IC_{50}, EC_{50} o KI de por ejemplo, < 1 molar, < 1 milimolar, < 100 micromolar, < 10 micromolar, < 1 micromolar, < 100 nanomolar, < 10 nanomolar, < 1 nanomolar).

Obtener una pluralidad de puntuaciones desde el calificador* (276) *y usar una puntuación de la red neural convolucional para caracterizar el objeto de prueba (278). Detalles para obtener una puntuación de calificador desde la red neuronal 24 para un complejo entre un objeto de prueba 72 (u objeto de entrenamiento 66) y un objeto objetivo 58 se han descrito anteriormente. Como se ha analizado anteriormente, cada objeto de prueba 72 (u objeto de entrenamiento 66) se ancla en una pluralidad de poses con respecto al objeto objetivo. Para presentar todas tales poses a la vez a la red neural convolucional 24 puede requerir un campo de entrada prohibitivamente grande (por ejemplo, un campo de entrada de tamaño igual al número de vóxeles * número de canales * número de poses). Mientras en algunas realizaciones todas las poses se presentan simultáneamente a la red 24, en realizaciones preferidas cada tal pose se procesa en un mapa de vóxel, vectorizado, y sirve como entrada secuencial en red neural convolucional 24. Haciendo referencia a la Figura 2E, de esta manera, una pluralidad de puntuaciones se obtienen a partir del calificador 30, en el que cada puntuación en la pluralidad de puntuaciones corresponde a la entrada de un vector en la pluralidad de vectores en la capa de entrada 26 del calificador 30 (276). En algunas realizaciones, las puntuaciones para cada una de las poses de un objeto de prueba dado 72 (u objeto de entrenamiento 66) con un objeto objetivo 58 dado se combinan juntos para producir una puntuación final para todo el objeto de prueba 72 (u objeto de entrenamiento 66).

En una realización en la que la salida de calificador es numérica, las salidas pueden combinarse usando cualquiera de las funciones de activación descritas en este documento que se conocen o desarrollan. Ejemplos incluyen, pero sin limitación, una función de activación de no saturación f(x) = max(0,x), una función de tangente hiperbólica de saturación f(x) = tanh, f(x) = |tanh(x)|, la función sigmoide $f(x) = (1 + e^{-x})^{-1}$, logística (o sigmoide), softmax, gaussiana, promedio ponderado de Boltzmann, valor absoluto, lineal, lineal rectificada, lineal rectificada delimitada, lineal rectificada suave, lineal rectificada parametrizada, promedio, máximo, mínimo, alguna norma de vector LP (para p=1, 2, 3,... ,∞), signo, cuadrado, raíz cuadrada, multicuadrática, cuadrática inversa, multicuadrática inversa, spline poliarmónica y spline de placa delgada.

En algunas realizaciones de la presente divulgación, el sistema puede configurarse para utilizar la distribución de Boltzmann para combinar salidas, ya que esta coincide con la probabilidad física de poses si las salidas se interpretan como indicativas de energías de unión. En otras realizaciones de la invención, la función max() también puede proporcionar una aproximación razonable a la Boltzmann y es computacionalmente eficiente.

En una realización en la que la salida de calificador no es numérica, el clasificador 30 puede configurarse para combinar las salidas que usan utilizar diversos esquemas de votación de conjunto, que pueden incluir, como ejemplos ilustrativos y no limitantes, mayoría, promedio ponderado, métodos de Condorcet, recuento de Borda, entre otros.

En una realización, el sistema puede configurarse para aplicar un conjunto de clasificadores 30, por ejemplo, para generar indicadores de afinidad de unión.

Haciendo referencia al elemento 280 de la Figura 2E, en algunas realizaciones, el objeto de prueba 72 (u objeto de entrenamiento 66) es un compuesto químico y usar la pluralidad de puntuaciones (de la pluralidad de poses para el objeto de prueba o de entrenamiento) para caracterizar (por ejemplo, determinar una clasificación) el objeto de prueba (o entrenamiento) comprende tomar una medida de tendencia central de la pluralidad de puntuaciones. Cuando la medida de tendencia central satisface un valor umbral predeterminado o intervalo de valores umbral predeterminados, el objeto de prueba se considera que tiene una primera clasificación. Cuando la medida de tendencia central no satisface el valor umbral predeterminado o intervalo de valores umbral predeterminados, el objeto de prueba se considera que tiene una segunda clasificación (280).

Haciendo referencia al elemento 282 de la Figura 2E, en algunas realizaciones el uso de la pluralidad de puntuaciones para caracterizar el objeto de prueba 72 (u objeto de entrenamiento 66) comprende tomar un promedio

ponderado de la pluralidad de puntuaciones (de la pluralidad de poses para el objeto de prueba o de entrenamiento). Cuando el promedio ponderado satisface un valor umbral predeterminado o intervalo de valores umbral predeterminados, el objeto de prueba se considera que tiene una primera clasificación. Cuando el promedio ponderado no satisface el valor umbral predeterminado o intervalo de valores umbral predeterminados, el objeto de prueba se considera que tiene una segunda clasificación. En algunas realizaciones, el promedio ponderado es un promedio de Boltzman de la pluralidad de puntuaciones (284). En algunas realizaciones, la primera clasificación es una IC50, EC50 o KI para el objeto de prueba (u objeto de entrenamiento) con respecto al objeto objetivo que está por encima de un primer valor de unión (por ejemplo, uno nanomolar, diez nanomolar, cien nanomolar, uno micromolar, diez micromolar, cien micromolar o uno milimolar) y la segunda clasificación es una IC50, EC50, Kd o KI para el objeto de prueba con respecto al objeto objetivo que está por debajo del primer valor de unión (286).

Haciendo referencia al elemento 288 de la Figura 2E, en algunas realizaciones el uso de la pluralidad de puntuaciones para caracterizar el objeto de prueba 72 (u objeto de entrenamiento 66) comprende tomar un promedio ponderado de la pluralidad de puntuaciones (de la pluralidad de poses para el objeto de prueba o de entrenamiento). Cuando el promedio ponderado satisface un intervalo respectivo de valores umbral en una pluralidad de intervalos de valores umbral, el objeto de prueba (o entrenamiento) se considera que tiene una clasificación respectiva en una pluralidad de una clasificación respectiva que inequívocamente corresponde al intervalo respectivo de valores umbral. En algunas realizaciones, cada clasificación respectiva en la pluralidad de clasificaciones es un intervalo de IC_{50}, EC_{50}, Kd o KI (por ejemplo, entre uno micromolar y diez micromolar, entre uno nanomolar y 100 nanomolar) para el objeto de prueba con respecto al objeto objetivo (290).

En algunas realizaciones, se usa una única pose para cada respectivo objeto de prueba contra un objeto objetivo dado se ejecuta a través de la red neuronal 24 y la respectiva puntuación asignada por la red neuronal 24 para cada uno de los respectivos objetos de prueba sobre esta base para clasificar los objetos de prueba.

En algunas realizaciones, el promedio ponderado de la red 24 puntuaciones de una o más poses de un objeto de prueba contra cada uno de una pluralidad de objetos objetivos 58 evaluados por la red neuronal 24 usando las técnicas divulgadas en este documento se usa para clasificar el objeto de prueba. Por ejemplo, en algunas realizaciones, la pluralidad de objetos objetivos 58 se toman a partir de una ejecución de dinámica molecular en la que cada objeto objetivo en la pluralidad de objetos objetivos representa el mismo polímero en una etapa de tiempo diferente durante la ejecución de dinámica molecular. Un mapa de vóxel de cada una de una o más poses del objeto de prueba contra cada uno de estos objetos objetivos se evalúa contra la red 24 para obtener una puntuación para cada par independiente de pose - objeto objetivo y el promedio ponderado de estas puntuaciones se usa para clasificar el objeto objetivo.

Entrenar el modelo predictivo. En algunas realizaciones, en las que se implementa una red neural profunda (por ejemplo, la red neural convolucional 24), el módulo de evaluación convolucional 20 se configura para entrenar a la red 24 para recibir los datos geométricos introducidos y para emitir una predicción (probabilidad) de si un objeto de prueba dado se une o no a un objeto objetivo. Por ejemplo, en algunas realizaciones, los objetos de entrenamiento 66, que tienen datos de unión conocidos contra los objetos objetivos (debido a sus datos de unión 68 asociados) se ejecutan secuencialmente a través de la red neuronal 24 usando las técnicas analizadas anteriormente en relación con la Figura 2 y la red neuronal proporciona un único valor para cada respectivo objeto de entrenamiento.

En algunas de tales realizaciones, la red neuronal emite una de dos posibles clases de actividad para cada objeto de entrenamiento contra un objeto objetivo dado. Por ejemplo, el único valor proporcionado para cada respectivo objeto de entrenamiento por la red neuronal 24 está en una primera clase de actividad (por ejemplo, aglutinantes) cuando está por debajo de un valor umbral predeterminado y está en una segunda clase de actividad (por ejemplo, no aglutinantes) cuando el número está por encima del valor umbral predeterminado. Las clases de actividad asignadas por la red neuronal 24 se comparan con las clases de actividad reales según se representan por los datos de unión 68 de objeto de entrenamiento. En realizaciones no limitantes típicas, tales datos de unión 68 de objeto de entrenamiento son de ensayos de unión de laboratorio húmedo independientes. Errores en asignaciones de clase de actividad hechas por la red neuronal, según se verifican contra los datos de unión 68, se propagan hacia atrás a continuación a través de los pesos de la red neuronal para entrenar la red neuronal 24. Por ejemplo, los pesos de filtro de respectivos filtros en las capas convolucionales 28 de la red se ajustan en tal propagación hacia atrás. En una realización ilustrativa, la red neuronal 24 se entrena contra los errores en las asignaciones de clase de actividad hechas por la red 24, en vista de los datos de unión 68, mediante descenso de gradiente estocástico con el método de aprendizaje adaptativo AdaDelta (Zeiler, 2012 "ADADELTA: an adaptive learning rate method", CoRR, vol. abs/1212.5701), y el algoritmo de propagación hacia atrás proporcionado en Rumelhart et al., 1988, "Neurocomputing: Foundations of research", ch. Learning Representations by Back-propagating Errors, páginas 696-699, Cambridge, MA, Estados Unidos: MIT Press, que se incorpora por la presente por referencia. En algunas de tales realizaciones las dos posibles clases de actividad son respectivamente una constante de unión mayor que una cantidad de umbral dada (por ejemplo, una IC50, EC50, o KI para el objeto de entrenamiento con respecto al objeto objetivo que es mayor que uno nanomolar, diez nanomolar, cien nanomolar, uno micromolar, diez micromolar, cien micromolar o uno milimolar) y una constante de unión que está por debajo de la cantidad de umbral dada (por ejemplo, una IC50, EC50, o KI para el objeto de entrenamiento con respecto al objeto objetivo que es menos de uno nanomolar, diez nanomolar, cien nanomolar, uno micromolar, diez micromolar, cien micromolar o uno milimolar). En

algunas de tales realizaciones, una pluralidad de poses para cada objeto de entrenamiento contra un objeto objetivo dado se ejecutan secuencialmente a través de la red neuronal y el promedio ponderado de las puntuaciones para estas poses según se calculan por la red neuronal 24 se compara con datos de unión 68 que se adquieren mediante ensayos de unión de laboratorio húmedo.

En algunas de tales realizaciones, la red neuronal emite una de una pluralidad de posibles clases de actividad (por ejemplo, tres o más clases de actividad, cuatro o más clases de actividad, cinco o más clases de actividad) para cada objeto de entrenamiento contra un objeto objetivo dado. Por ejemplo, el único valor proporcionado para cada respectivo objeto de entrenamiento por la red neuronal 24 (por ejemplo, el promedio ponderado de una pluralidad de poses o un único valor de una única pose) está en una primera clase de actividad cuando el número se encuadra dentro de un primer intervalo, está en una segunda clase de actividad cuando el número se encuadra dentro de un segundo intervalo, está en una tercera clase de actividad cuando el número se encuadra dentro de un tercer intervalo, y así sucesivamente. Las clases de actividad asignadas por la red neuronal 24 se comparan con las clases de actividad reales según se representan por los datos de unión 68 de objeto de entrenamiento. Errores en asignaciones de clase de actividad hechas por la red neuronal, según se verifican contra los datos de unión 68, se usan a continuación para entrenar a la red neuronal 24 usando las técnicas analizadas anteriormente. En algunas realizaciones, cada clasificación respectiva en la pluralidad de clasificaciones es un intervalo de IC50, EC50 o KI para el objeto de entrenamiento con respecto al objeto objetivo.

En algunas realizaciones, una única pose para cada respectivo objeto de entrenamiento contra un objeto objetivo dado se ejecuta a través de la red neuronal y la respectiva puntuación resultante asignada por la red neuronal 24 para cada respectivo objeto de entrenamiento se compara con datos de unión 68 para el respectivo objeto de entrenamiento que se ha adquirido de forma separada por una o más técnicas de ensayo de unión de laboratorio húmedo. A continuación, errores en asignaciones de clase de actividad hechas por la red neuronal 24 para los objetos de entrenamiento, según se verifican contra los datos de unión 68 para los objetos de entrenamiento, se usan para entrenar a la red neuronal 24 usando las técnicas analizadas anteriormente.

En algunas realizaciones, el promedio ponderado de una o más poses de un objeto de entrenamiento contra cada uno de una pluralidad de objetos objetivos 58 evaluados por la red neuronal 24 usando las técnicas divulgadas en este documento se compara con los datos de unión 68 para los respectivos objetos de entrenamiento que se adquieren de forma separada por una o más técnicas de ensayo de unión de laboratorio húmedo. Por ejemplo, en algunas realizaciones, la pluralidad de objetos objetivos 58 se toman a partir de una ejecución de dinámica molecular en la que cada objeto objetivo en la pluralidad de objetos objetivos representa el mismo polímero en una etapa de tiempo diferente durante la ejecución de dinámica molecular. Discrepancias entre objeto objetivo clasificación por la red neuronal 24 y el objeto clasificación por los ensayos de unión de laboratorio húmedo se usan a continuación para entrenar a la red neuronal 24 usando las técnicas analizadas anteriormente.

En algunas realizaciones, clasificación de la red neural 24 de una pluralidad de objetos de entrenamiento se compara con los datos de unión 68 usando técnicas no paramétricas. Por ejemplo, la red neuronal 24 se usa para ordenar por clasificación la pluralidad de objetos de entrenamiento con respecto a una propiedad dada (por ejemplo, unión contra un objeto objetivo dado) y este orden de clasificación se compara con el orden de clasificación proporcionado por los datos de unión 68 que se adquieren mediante ensayos de unión de laboratorio húmedo para la pluralidad de objetos de entrenamiento. Esto hacer surgir la capacidad de entrenar a la red 24 en los errores en el orden de clasificación calculado usando las técnicas de corrección de errores de la red 24 analizadas anteriormente. En algunas realizaciones, el error (diferencias) entre la clasificación por los objetos de entrenamiento por la red neuronal 24 y la clasificación de los objetos de entrenamiento según se determinan por los datos de unión 68 se calcula usando una función de Wilcoxon Mann Whitney (prueba de rangos con signo de Wilcoxon) u otra prueba no paramétrica y este error se propaga hacia atrás a través de la red neuronal 24 para entrenar adicionalmente la red usando las técnicas de corrección de la red neuronal 24 analizadas anteriormente.

En una realización en la que las técnicas de aprendizaje profundo utilizan una red neural 24 como se ha descrito anteriormente, el módulo de evaluación convolucional 20 puede configurarse para entrenar a la red 24 para mejorar la precisión de su predicción modificando los pesos en los filtros en las capas convolucionales 28 así como los sesgos en las capas de red. Los pesos y sesgos pueden limitarse adicionalmente con diversas formas de regularización tal como L1, L2, degradación de pesos y abandono.

En una realización, la red neuronal 24 puede configurarse opcionalmente para ajustar los pesos de la red para modelizar la entrada distribución de los datos de entrenamiento a través de entrenamiento previo generativo, de capa y voraz contra los objetos de entrenamiento usando el algoritmo de divergencia contrastante.

En una realización, la red neuronal 24 puede ajustar opcionalmente, en la que datos de entrenamiento se etiquetan (por ejemplo, con los datos de unión 68), los pesos dentro de la red 24 para minimizar potencialmente el error entre las afinidades y/o categorizaciones de unión predichas de la red neuronal y las afinidades y/o categorizaciones de unión notificadas de los datos de entrenamiento. Pueden usarse diversos métodos para minimizar función de error, tal como métodos de descenso de gradiente, que pueden incluir, pero sin limitación, métodos de pérdida logística, suma de errores cuadrados, pérdida de bisagra. Estos métodos pueden incluir métodos de segundo orden o

aproximaciones tales como momento, estimación libre hessiana, gradiente acelerado de Nesterov, adagrad, etc. También pueden combinarse entrenamiento previo generativo sin etiquetar y entrenamiento discriminativo etiquetado.

Datos geométricos de entrada pueden agruparse en ejemplos de entrenamiento. Por ejemplo, a menudo es el caso de que un único conjunto de moléculas, cofactores y proteína tienen múltiples mediciones geométricas, en el que cada "instantánea" describe conformaciones alternativas y poses que el objeto objetivo y los objetos de entrenamiento (u objetos de prueba) pueden adoptar. De manera similar, en casos en los que el objeto objetivo es una proteína, también pueden muestrearse diferentes tautómeros para las cadenas secundarias de proteínas, cofactores, y los objetos de entrenamiento (o prueba). Porque estos estados contribuyen todos al comportamiento del sistema biológico, de conformidad con la distribución de Boltzmann, puede configurarse un sistema para predecir afinidad de unión para considerar estos estados juntos (por ejemplo, tomando el promedio ponderado de estas muestras). Opcionalmente, estos ejemplos de entrenamiento pueden etiquetarse con información de unión. Si información de unión cuantitativa está disponible (por ejemplo, datos de unión 68), las etiquetas pueden ser las afinidades de unión numéricas. Como alternativa, los ejemplos de entrenamiento pueden ser etiquetas asignadas desde un conjunto de dos o más categorías ordenadas (por ejemplo, dos categorías de aglutinantes y no aglutinantes, o varias categorías posiblemente solapantes que describen los ligandos como aglutinantes de potencias < 1 molar, < 1 milimolar, < 100 micromolar, < 10 micromolar, < 1 micromolar, < 100 nanomolar, < 10 nanomolar, < 1 nanomolar). Los datos de unión 68 pueden derivarse o recibirse desde diversas fuentes, tal como mediciones experimentales, estimadas calculadas, percepción experta o presunción (por ejemplo, un par aleatorio de molécula y proteína son altamente improbables que se unan).

Ejemplo 1 - Construcción de puntos de referencia experimentales.

La aplicación de sistemas y métodos divulgados se demuestra en tres puntos de referencia: el punto de referencia de Directorio de Señuelos Útiles Mejorado (DUDE) (véase Mysinger et al., 2012, "Directory of useful decoys, enhanced (dude): Better ligands and decoys for better benchmarking", Journal of Medicinal Chemistry 55, n.º 14, páginas 6582-6594, PMID: 22716043), un punto de referencia de tipo DUDE interno; y un punto de referencia con moléculas inactivas experimentalmente verificadas. Cada uno de estos puntos de referencia proporciona una evaluación diferente y complementaria del rendimiento de los sistemas y métodos divulgados. Como el punto de referencia estándar, DUDE permite dirigir comparaciones con otros sistemas de predicción de afinidad de unión basados en estructura. Desafortunadamente, DUDE es únicamente específico un conjunto de prueba, sin especificar un conjunto de entrenamiento separado. Construyendo nuestro propio punto de referencia de tipo DUDE, garantizamos que no hay solapamiento entre las moléculas de entrenamiento y de prueba. Clasificar correctamente moléculas activas e inactivas experimentalmente verificadas es una prueba desafiante porque moléculas estructuralmente similares pueden tener diferentes etiquetas. Véase Hu et al., "Systematic identification and classification of three-dimensional activity cliffs", Journal of Chemical Information and Modeling 52, n.º 6, páginas 11490-1498. Tales casos se excluyen de puntos de referencia usando señuelos de propiedad coincidente debido a el requisito de disimilitud para presumir que los señuelos están inactivos.

La metodología del punto de referencia DUDE se describe completamente por Mysinger et al., 2012, "Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking", Journal of Medicinal Chemistry 55, n.º 14, páginas 6582-6594, PMID: 22716043. Se construyó un punto de referencia interno de forma similar para este ejemplo. Brevemente, ambos puntos de referencia se construyen recopilando primero varios conjuntos de moléculas activas para un conjunto de proteínas objetivo. Sesgo analógico se mitiga eliminando activos similares; activos similares se eliminan agrupando primero los activos basándose en similitud de soporte, a continuación seleccionando activos ilustrativos de cada agrupamiento. A continuación, cada molécula activa se empareja con un conjunto de señuelos de propiedad coincidente (PMD). Véase Wallach y Lilien, 2011, "Virtual Decoy Sets for Molecular Docking Benchmarks", J Chem. Inf. and Model., 51, n.º 2, páginas 196-202; y Wallach et al., 2011 "Normalizing molecular clocking rankings using virtually generated decoys", J. Chem. Inf. and Model., 51, n.º 8, páginas 1817-1830. PMD se seleccionan para ser similares entre sí y a activos conocidos con respecto a algunos descriptores físicoquímicos unidimensionales (por ejemplo, peso molecular) mientras son disimilares topológicamente basándose en algunas huellas 2D (por ejemplo, ECFP que se describen en Rogers y Hahn, "Extended-connectivity fingerprints", 2010, Journal of Chemical Information and Modeling 50, n.º 5, páginas 742-754). El cumplimiento de la disimilitud topológica soporta la suposición de que es probable que los señuelos estén inactivos porque son químicamente diferentes de cualquier activo conocido.

DUDE. El DUDE es un punto de referencia bien conocido para métodos de cribado virtuales basados en estructura del laboratorio Shoichet en UCSF. Véase Mysinger et al., 2012, "Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking", Journal of Medicinal Chemistry, 55, n.º 14, páginas 6582-6594, PMID: 22716043. Consta de 102 objetivos, 22.886 activos (un promedio de 224 activos por objetivo) y 50 PMD por activo. Treinta objetivos se seleccionaron aleatoriamente como el conjunto de prueba y los restantes 72 objetivos se designaron como el conjunto de entrenamiento.

PMD de ChEMBL-20. Se construyó un conjunto de datos de tipo DUDE derivado a partir de ChEMBL versión 20 (Bento et al., 2014, "The chembl bioactivity database: an update: an update", Nucleic Acids Research 42, n.º D1,

páginas D1083-D1090). Se consideraron todas las mediciones de actividad que pasaron los siguientes filtros: (i) unidades de afinidad medidas en IC50 o Ki y menores de 1 μM, (ii) confianza objetivo mayor o igual que 6, (iii) objetivo tiene un sitio de unión anotado en la base de datos scPDB (Desaphy et al. 2014, "sc-pclb: a 3d-database of ligandable binding sites 10 years on", Nucleic Acids Research D3 99-404) y resolución < 2,5 Å, y (iv) ligandos pasaron filtros PAINS (Baell y Holloway, 2010, "New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays", Journal of Medicinal Chemistry 53, n.º 7, páginas 2719-2740) y reglas de promiscuidad (Bruns y Watson, 2012, "Rules for identifying potentially reactive or promiscuous compounds", Journal of Medicinal Chemistry 55, n.º 22, páginas 9763-9772). Después de *Mysinger et al.,* afinidades objetivo se agruparon por su prefijo de nombre de gen (Bruns y Watson, 2012 "Rules for identifying potentially reactive or promiscuous compounds", Journal of Medicinal Chemistry 55, n.º 22, páginas 9763-9772) y se eliminaron objetivos para los que había menos de 10 ligandos activos. Este proceso de filtrado proporcionó un conjunto de 123.102 activos y 348 objetivos. Segundo, cada activo se emparejó con un conjunto de 30 PMD seleccionados a partir de la base de datos CINC (Irwin y Shoichet, 2005, "ZINC-a free database of commercially available compounds for virtual screening", J. Chem. Inf. Model. 45, n.º 1, páginas 177-182) de forma similar a Mysinger et al., 2012, "Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking", Journal of Medicinal Chemistry 55, n.º 14, páginas 6582-6594, PMID: 22716043. Tercero, los datos se dividieron en conjuntos de entrenamiento, validación y prueba agrupando primero los ligandos activos para cada objetivo basándose en sus soportes Bemis-Murcko (Bemis y Murcko, 1996, "The properties of known drugs. I. molecular frameworks", Journal of Medicinal Chemistry 39, n.º 15, páginas 2887-2893) y eligiendo ligandos que estuvieran al menos 3 μM alejados como los ejemplares de agrupamiento. Se descartaron agrupaciones con menos de 10 ejemplares. Cuarto, el conjunto de prueba se definió seleccionado aleatoriamente 50 objetivos con sus correspondientes activos y señuelos. Por último, el conjunto de entrenamiento se dividió adicionalmente en las agrupaciones en conjuntos de validación cruzada de 5 pliegues. El conjunto de datos final consta de 78.904 activos, 2.367.120 señuelos y 290 objetivos.

Inactivos verificados experimentalmente. Una limitación de puntos de referencia basándose en PMD es que excluyen señuelos que son similares a moléculas activas. La decisión de diseño está en lugar para soportar la suposición de que los señuelos seleccionados es probable que estén inactivos sin validación experimental. Esta disimilitud forzada entre activos y señuelos significa que puntos de referencia PMD carecen de algunos casos desafiantes en los que moléculas activas e inactivas son altamente similares (Hu et al., 2012, "Systematic identification and classification of three-dimensional activity cliffs", Journal of Chemical Information and Modeling 52, n.º 6, páginas 1490-1498). Tales casos desafiantes se incluyeron sustituyendo señuelos con moléculas que se han validado experimentalmente para estar inactivas. Se construyó un punto de referencia similar al PMD de ChEMBL-20, pero PMD se sustituyó con moléculas inactivas. Una molécula se define en este punto como inactiva si su actividad medida es mayor de 30 μM. Esto resultó en un conjunto de 78.904 activos, 363.187 inactivos y 290 objetivos que se dividieron en conjuntos de validación cruzada de 3 pliegues en agrupaciones Bemis-Murcko. Objetivos con menos agrupaciones nunca se asignaron en un conjunto de validación. Por lo tanto, el número de objetivos en los conjuntos de validación era 149.

Red neural convolucional profunda basada en estructura. La topología de red para la red neural convolucional 24 en este experimento (AtomNet) constaba de una capa de entrada 26 seguida por convolucional 3D múltiple 28 y un calificador 30 que constaba de capas totalmente conectadas encabezadas por una capa de coste logístico que asigna probabilidades sobre las clases activas e inactivas. Todas las unidades en capas se implementan con la función de activación ReLU (Nair y Hinton, 2010 "Rectified linear units improve restricted Boltzmann machines" en Proceedings of the 27th International Conference on Machine Learning (ICML-10), 21-24 de junio de 2010, Haifa, Israel, páginas 807-814).

Representación de entrada. La capa de entrada 26 recibe versiones vectorizadas de cuadrículas 3D de 1 Å situadas sobre cocomplejos de las proteínas objetivo (objetos objetivos 58) y moléculas pequeñas (objetos de entrenamiento / objetos de prueba) que se muestrean dentro del sitio de unión del objetivo. Primero, el sitio de unión se define usando a algoritmo de inundación *(*Véase Hendlich et al., 1997, "Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins", J. Mol. Graph. Model 15, n.º 6) provisto por un ligando unido anotado en la base de datos scPDB *(*Véase Desaphy et al. 2014, "sc-pdb: a 3d-database of ligandable binding sites 10 years on", Nucleic Acids Research D3 99-404). Segundo, las coordenadas de los cocomplejos se desplazan a un sistema cartesiano tridimensional originado en el centro de masa del sitio de unión. Tercero, se muestrean múltiples poses dentro de la cavidad de sitio de unión. Cuarto, los datos geométricos se recortan para caber dentro de un cuadro delimitador. En este estudio se usa un cubo de 20 Å, centrado en el origen. Quinto, los datos de entrada se trasladan en una cuadrícula de tamaño fijo con espaciado de 1 Å. Cada célula de cuadrícula mantiene un valor que representa la presencia de algunas características estructurales básicas en esa ubicación. Las características estructurales básicas pueden variar desde una simple enumeración de tipos de átomo hasta descriptores de proteína-ligando más complejos tal como SPLIF (Da y Kireev, "Structural protein ligand interaction fingerprints (splif) for structure-based virtual screening: Method and benchmark study", 2014, Journal of Chemical Information and Modeling 54, n.º 9, páginas 2555-2561), SIFt (Deng et al., 2004, "Structural interaction fingerprint (SIFt): a novel method for analyzing threedimensional protein-ligand binding interactions", J. Med. Chem. 47, n.º 2, páginas 337-344) o APIF (Pérez-Nueno, 2009 "Apif: A new interaction fingerprint based on atom pairs and its application to virtual screening", Journal of Chemical Information and Modeling 49, n.º 5, páginas 1245-1260). Por último, la cuadrícula 3D se despliega en un vector de punto flotante 1D.

Arquitectura de red. Las capas convolucionales 3D 28 se implementaron para soportar parámetros tal como filtro tamaño, zancada y relleno de una forma similar a la implementación de Krizhevsky et al., 2012, "Imagenet classification with deep convolutional neural networks", en Advances in Neural Information Processing Systems 2, Pereira, Burges, Bottou, Weinberger, ed., páginas 1097-1105, Curran Associates, Inc. La arquitectura de red la red neural convolucional 24 constaba de una capa de entrada 26 como se ha descrito anteriormente, seguida por cuatro capas convolucionales 28 de 128 x 5^3, 256 x 3^3, 256 x 3^3, 256 x 3^3 (número de filtros x dimensión de filtro), y dos capas totalmente conectadas cada una con 1024 unidades escondidas, encabezadas por una capa de coste de regresión logística sobre dos clases de actividad.

Entrenamiento de red neural convolucional 24. Entrenar la red neural convolucional 24 se hizo usando descenso de gradiente estocástico con el método de aprendizaje adaptativo AdaDelta (Zeiler, 2012 "ADADELTA: an adaptive learning rate method", CoRR, vol. abs/1212.5701), el algoritmo de propagación hacia atrás (Rumelhart et al., 1988, "Neurocomputing: Foundations of research", ch. Learning Representations by Back-propagating Errors, páginas 696-699, Cambridge, MA, Estados Unidos: MIT Press), y mini lotes de 768 ejemplos por escalón de gradiente. No se hizo ningún intento de optimizar metaparámetros excepto la limitación de ajustar el modelo a una memoria de GPU. El tiempo de entrenamiento fue de aproximadamente una semana en seis GPU Nvidia-K10.

Método de línea de base para comparación. Se usó Smina *(*Véase Koes et al., 2013 "Lessons learned in empirical scoring with smina from the csar2011 benchmarking exercise", Journal of Chemical Information and Modeling 53, n.º 8, páginas 1893-1904, 2013), una ramificación de AutoDock Vina (Trott y Olson, 2010 "Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading", Journal of Computational Chemistry 31, n.º 2, páginas 455-461), como una línea de base para la evaluación basada en estructura. Smina implementa una función de puntuación empírica mejorada y rutinas minimización sobre su predecesor y está disponible libremente en la licencia GPLv2.

Resultados. El área bajo la característica de operación de receptor (AUC) y logAUC se usaron para notificar resultados en los tres puntos de referencia. La AUC indica rendimiento de clasificación (u orden clasificado) midiendo el área bajo la curva de tasa de positivo verdadero frente a la tasa positivo falso. Un valor de AUC de 1,0 significa separación perfecta mientras que un valor de 0,5 implica separación aleatoria. LogAUC es una medición similar a AUC que enfatiza rendimiento de enriquecimiento temprano poniendo más peso en el comienzo de la curva de forma que casos clasificados correctamente en la parte superior de la lista ordenada por clasificación contribuyen más a la puntuación que los posteriores. En este documento, se usó una base logarítmica de 10, que significa que el peso del primer 1 % de los resultados clasificados es igual al peso del siguiente 10 %. Porque la no linealidad de un valor de logAUC hace difícil interpretar el mismos, el área bajo la curva aleatoria escalada logarítmicamente (0,14462) se restó del logAUC para conseguir un logAUC ajustado *(*Véase Mysinger y Shoichet, 2010, "Rapid context-dependent ligand desolvation in molecular docking", Journal of Chemical Information and Modeling 50, n.º 9, páginas 1561-1573). Por lo tanto, valores de logAUC ajustados positivos implican rendimiento mejor que aleatorio, mientras que los negativos implican rendimiento peor que aleatorio. Por brevedad, el logAUC ajustado y logAUC se usan indistintamente en este documento.

La Tabla 2 y Figuras 10 a 13 resumen los resultados a través de los tres diferentes puntos de referencia. El sistema ilustrativo y método de la presente divulgación se comportan de forma similar en los puntos de referencia de PMD de ChEMBL-20 y de DUDE. El sistema ilustrativo y método de la presente divulgación consigue una AUC media de 0,78 y un logAUC medio de 0,32 en PMD de ChEMBL-20 y 0,8 y 0,33 en DUDE respectivamente. Este rendimiento similar no es sorprendente porque los dos puntos de referencia se construyeron de forma similar.

La Figura 10 proporciona la distribución de valores de AUC y de logAUC de 50 objetivos de PMD de ChEMBL-20 para AtomNet y Smina. La Figura 11 proporciona la distribución de valores de AUC y de logAUC de 102 objetivos de DUDE para AtomNet y Smina. La Figura 12 proporciona la distribución de valores de AUC y de logAUC de 149 objetivos de inactivos de ChEMBL-20 para AtomNet y Smina. Las Figuras 13A y 13B proporcionan una ilustración de las diferencias entre las mediciones de AUC y de logAUC con respecto al enriquecimiento temprano.

Tabla 2

		AUC		logAUC ajustado	
		Media	Mediana	Media	Mediana
PMD de ChEMBL-20	AtomNet	0,781	0,792	0,317	0,328
	Smina	0,552	0,544	0,04	0,021
DUDE-30	AtomNet	0,855	0,875	0,321	0,355
	Smina	0,7	0,694	0,153	0,139
DUDE-102	AtomNet	0,895	0,915	0,385	0,38
	Smina	0,696	0,707	0,138	0,132
inactivos de ChEMBL-20	AtomNet	0,745	0,737	0,145	0,133
	Smina	0,607	0,607	0,054	0,044

Tabla 2: comparaciones de AtomNet y Smina en los puntos de referencia de DUDE, PMD de ChEMBL-20 e inactivos de ChEMBL-20. DUDE-30 se refiere al conjunto retenido de 30 objetivos mientras que DUDE-102 se refiere a al conjunto de datos completo.

Tabla 3

AUC		> 0,5	> 0,6	> 0,7	> 0,8	> 0,9
PMD de ChEMBL-20	AtomNet	49	44	36	24	10
	Smina	38	10	4	1	0
DUDE-30	AtomNet	30	29	27	22	14
	Smina	29	25	14	5	1
DUDE-102	AtomNet	102	101	99	88	59
	Smina	96	84	53	17	1
inactivos de ChEMBL-20	AtomNet	149	136	105	45	10
	Smina	129	81	31	4	0

Tabla 3: el número de objetivos en el que AtomNet y Smina exceden umbrales AUC dados. Por ejemplo, en el conjunto de PMD de ChEMBL-20, AtomNet consigue una AUC de 0,8 o mejor para 24 objetivos (de entre 50 posibles objetivos). ChEMBL-20 PMD contiene 50 objetivos, DUDE-30 contiene 30 objetivos, DUDE-102 contiene 102 objetivos e inactivos de ChEMBL-20 contiene 149 objetivos.

Tabla 4

Log AUC ajustado		> 0,0	> 0,1	> 0,2	> 0,3	> 0,4
PMD de ChEMBL-20	AtomNet	49	44	36	24	10
	Smina	35	8	2	1	0
DUDE-30	AtomNet	30	27	22	17	10
	Smina	29	19	8	2	1
DUDE-102	AtomNet	102	99	88	69	43
	Smina	94	65	28	5	1
inactivos de ChEMBL-20	AtomNet	147	107	36	10	2
	Smina	123	35	5	0	0

Tabla 4: el número de objetivos en los que AtomNet y Smina exceden umbrales de logAUC ajustado dados. Por ejemplo, el en conjunto PMD de ChEMBL-20, AtomNet consigue un logAUC ajustado de 0,3 o mejor para 27 objetivos (de entre 50 posibles objetivos). PMD de ChEMBL-20 contiene 50 objetivos, DUDE-30 contiene 30 objetivos, DUDE-102 contiene 102 objetivos, e inactivos de ChEMBL-20 contiene 149 objetivos.

En cada uno de nuestros cuatro conjuntos de datos de evaluación, el sistema y método divulgados (AtomNet) consiguen una mejora de orden de magnitud sobre Smina en un nivel de precisión útil para descubrimiento de fármacos. En el conjunto de DUDE completo, AtomNet consigue o excede 0,9 AUC en 59 objetivos (o 57,8 %). Smina únicamente consigue 0,9 AUC para un único objetivo (wee1), aproximadamente un uno por ciento del punto de referencia. AtomNet consigue 0,8 o mejor AUC para 88 objetivos (86,3 %), mientras Smina lo consigue para 17 objetivos (16,7 %). Cuando la evaluación se restringe al subconjunto de 30 objetivos retenidos de DUDE, AtomNet excede una AUC de 0,9 y 0,8 para 14 objetivos (46,7 %) y 22 objetivos (73,3 %). respectivamente. Smina consigue la misma precisión para un objetivo (3,3 %) y 5 objetivos (16,7 %), respectivamente. AtomNet consigue AUC media y mediana de 0,855 y 0,875 en el conjunto retenido en comparación con 0,7 y 0,694 conseguido por Smina, reduciendo el error medio disponible en un 51,6 %. Como se esperaba, el rendimiento de AtomNet cae ligeramente para sus ejemplos retenidos mientras que el rendimiento de Smina no lo hace.

En el conjunto de datos de PMD, AtomNet consigue una AUC de 0,9 o mejor para 10 objetivos retenidos (20 % del conjunto), mientras Smina los consigue en cero objetivos. Cuando el estándar de precisión se reduce a una AUC de 0,8 o mejor, AtomNet tiene éxito en 25 objetivos (50 %) mientras Smina únicamente tiene éxito en 1 objetivo (2 %).

El tercer punto de referencia, que usa inactivos en lugar de señuelos de propiedad coincidente, parece ser más desafiante que los otros dos. AtomNet predice con una AUC en o mejor de 0,9 para 10 objetivos (6,7 %), mientras Smina tiene éxito en cero. Para cumplir o exceder 0,8 AUC, AtomNet tiene éxito para 45 objetivos (30,2 %) y Smina tiene éxito para 4 (2,70 %). Aunque tanto Atomnet como Smina se comportan peor que los puntos de referencia anteriores, AtomNet aún supera significativamente a Smina con respecto rendimientos de enriquecimiento general y temprano. Porque este punto de referencia usa inactivos, incluye casos clasificación desafiantes de moléculas estructuralmente similares con diferentes etiquetas (Hu et al., "Systematic identification and classification of three-dimensional activity cliffs", 2012, Journal of Chemical Information and Modeling 52, n.º 6, páginas 1490-1498). Estos casos se excluyen de puntos de referencia usando PMD porque señuelos deben ser estructuralmente disimilares para suponer que pueden etiquetarse como inactivos.

Adicionalmente AtomNet muestra buen rendimiento de enriquecimiento temprano como se indica por los valores logAUC altamente positivos. AtomNet supera a Smina con respecto a su enriquecimiento temprano, consiguiendo un logAUC medio de 0,321 en comparación con 0,153 de Smina. Visualizando las curvas ROC ilustran la diferencia entre las mediciones de AUC y de logAUC con respecto al enriquecimiento temprano. Por ejemplo, la Figura 13A muestra que el valor de AUC para el objetivo *1m9m* es 0,66 que puede implicar un rendimiento mediocre. Sin embargo, el enriquecimiento temprano indicado por el logAUC para ese objetivo es 0,25 que sugiere que se concentran muchos activos en la parte superior de los resultados ordenados por clasificación. De manera similar, el objetivo *lqzy* tiene un valor de AUC de 0,76 pero la representación de log-BLscale sugiere que el 35 % de sus activos se concentran en la parte superior de la lista ordenada por clasificación con logAUC de 0,44.

Descripción - visualización de filtro. Capas convolucionales 28 constan de múltiples diferentes filtros que aprenden a identificar características localmente relacionadas específicas aplicando repetidamente estos filtros a través del campo receptivo. Cuando trata con imágenes, se pueden visualizar estos filtros para verificar que el modelo es capaz de aprender características relevantes. Por ejemplo, Krizhevsky et al., 2012, "Imagenet classification with deep convolutional neural networks", en Advances in Neural Information Processing Systems 2, Pereira, Burges, Bottou, Weinberger, ed., páginas 1097-1105, Curran Associates, Inc., demostró que filtros en la primera capa convolucional de su modelo podría detectar líneas, bordes y gradientes de color. En nuestro caso, sin embargo, los filtros no se visualizan fácilmente porque: (i) los filtros son tridimensionales y (ii) los canales de entrada son discretos. Por ejemplo, dos valores RGB cercanos resultarán con dos colores similares pero carbono no está más próximo a nitrógeno que a oxígeno. Es decir, valores similares no implican funcionalidades similares. Para superar estas limitaciones se toma un enfoque indirecto. En lugar de visualizar directamente filtros para entender su especialización, se aplican filtros a datos de entrada y se examina la ubicación en la que disparan máximamente. Usando esta técnica, los filtros se correlacionaron con función química. Por ejemplo, inspección visual de las ubicaciones tridimensionales en el objeto objetivo en el que dispara un filtro particular de la primera capa convolucional 28 revela que este filtro se especializa como un detector de sulfonilo/sulfonamida. Véase, por ejemplos, las Figuras 15A y 15B que ilustran una interacción de este tipo. Esto demuestra la capacidad del modelo para aprender características químicas complejas a partir de unas más simples. En este caso, el filtro ha inferido una disposición espacial significativa de tipos de átomos de entrada sin ningún conocimiento químico anterior.

Comparación con otros métodos basados en estructura. Este ejemplo proporciona una realización del sistema y método divulgados para aplicar una red neural convolucional profunda 24 a predicciones de bioactividad en lugar de notificar comparaciones cabeza a cabeza con otros métodos basados en estructura. Para situar los resultados en contexto, se usó el programa popular Smina como un punto de referencia de línea de base. Smina tiene ventajas prácticas: es rápido, gratis y en desarrollo activo, por tanto, es adecuado para analizar grandes puntos de referencia de una manera oportuna y rentable. Sin embargo, usando trabajo publicado, se proporciona un contexto más amplio comparando AtomNet con otros algoritmos de anclaje comerciales notificados en la bibliografía. Como Smina, DUDE está disponible públicamente y se usa ampliamente. DUDE tiene ciertas limitaciones: por ejemplo, puntos de referencia de DUDE y otros PMD no son apropiados para la evaluación de modelos basados en ligandos (véase Irwin, "Community benchmarks for virtual screening", 2008, J. Comput.-Aided Mol. Des. 22, n.º 3-4, páginas 193-199), porque se usan los mismos descriptores usados para hacer cumplir la diversidad entre activos y señuelos para entrenar clasificadores basados en ligandos. Además, como se ha analizado anteriormente, no puede garantizarse la ausencia de contaminación entre entrenamiento y prueba cuando se evalúa en DUDE, que fue la principal motivación para construir el punto de referencia de PMD de ChEMBL-20 divulgado. Sin embargo, el rendimiento similar en estos dos puntos de referencia sugiere que los resultados son robustos. Por lo tanto, se presentan las siguientes comparaciones con resultados anteriormente descritos: Gabel *et al. (*Véase Gabel et al., 2014 "Beware of machine learning-based scoring functions on the danger of developing black boxes", Journal of Chemical Information and Modeling 54, n.º 10, páginas 2807-2815) evaluó Surflex-Dock *(*Véase Spitzer y Jain, 2012 "Surftex-dock: Docking benchmarks and real-world application", Journal of Computer-Aided Molecular Design 26, n.º 6, páginas 687-699) en un conjunto representativo de 10 objetivos del DUDE. La AUC mediana de Surflex-Dock fue 0,76 en comparación con 0,83 conseguida por AtomNet. Coleman *et al.* (Véase Coleman et al., 2014, "Samp14 & dock3.7: lessons for automated docking procedures", Journal of Computer-Aided Molecular Design 28, n.º 3, páginas 201-209, que se incorpora por la presente por referencia) evaluó DOCK-3.7 (Coleman et al., "Ligand pose and orientational sampling in molecular docking", PLoS ONE 8, p. e75992) de una manera totalmente automatizada sobre todo el punto de referencia de DUDE. Consiguieron una AUC media de 0,674 y logAUC de 0,164 en comparación con nuestra AUC de 0,792 y logAUC de 0,306.

Conclusión. El sistema y método divulgados presentados en este ejemplo (AtomNet) es la primera red neural convolucional profunda basada en estructura diseñada para predecir la bioactividad de pequeñas moléculas para aplicaciones de descubrimiento de fármacos. La arquitectura convolucional profunda localmente limitada permite que el sistema modelice el fenómeno no lineal complejo de unión molecular mediante la composición jerárquica de características químicas básicas próximas en unas más intrincadas. Incorporando información de objetivo estructural, AtomNet puede predecir nuevas moléculas activas incluso para objetivos sin moduladores anteriormente conocidos. AtomNet muestra excelentes resultados en un punto de referencia basado en estructura ampliamente usado consiguiendo una AUC mayor de 0,9 en el 57,8 % de los objetivos, 59 veces más objetivos que el método de anclaje ampliamente usado.

Ejemplo 2 - Casos de uso.

Lo siguiente son casos de uso de ejemplo proporcionados para propósitos de ilustración únicamente que describen algunas aplicaciones de algunas realizaciones de la invención. Pueden considerarse otros usos, y los ejemplos proporcionados a continuación no son limitantes y pueden someterse a variaciones, omisiones o pueden contener elementos adicionales.

Aunque cada ejemplo a continuación ilustra predicción de afinidad de unión, puede encontrarse que los ejemplos difieren en si las predicciones se hacen sobre una única molécula, un conjunto o una serie de moléculas modificadas iterativamente; si las predicciones se hacen para un único objetivo o muchos, si se desea o debe evitarse actividad contra los objetivos, y si la cantidad importante es actividad absoluta o relativa; o, si los conjuntos de moléculas u objetivos se eligen específicamente (por ejemplo, para moléculas, para que sean fármacos o pesticidas existentes; para proteínas, para tener toxicidades o efectos secundarios conocidos).

Descubrimiento exitoso. Las compañías farmacéuticas gastan millones de dólares en el cribado de compuestos para descubrir nuevos avances potenciales de fármacos. Se prueban grandes colecciones de compuestos para encontrar el pequeño número de compuestos que tienen interacción con la enfermedad objetivo de interés. Desafortunadamente, cribado de laboratorio húmedo sufre de errores experimentales y, además del coste y tiempo para realizar los experimentos de ensayos, la recopilación de grandes colecciones de cribado impone desafíos significativos a través de limitaciones de almacenamiento, estabilidad o costes químicos. Incluso las compañías farmacéuticas más grandes tienen únicamente entre cientos de miles a unos pocos millones de compuestos, frente a las decenas de millones de moléculas comercialmente disponibles y los cientos de millones de moléculas con capacidad de simulación.

Una alternativa potencialmente más eficiente a experimentación física es cribado virtual de alto rendimiento. De la misma manera que las simulaciones físicas pueden ayudar a ingenieros aeroespaciales a evaluar posibles diseños de alas antes de que un modelo se pruebe físicamente, el cribado computacional de moléculas puede centrar la prueba experimental en un subconjunto pequeño de moléculas de alta probabilidad. Esto puede reducir el tiempo y coste de cribado, reducir falsos negativos, mejorar tasas de éxito y/o cubrir una franja más amplia del espacio químico.

En esta aplicación, puede proporcionarse un objetivo de proteína como entrada al sistema. También puede proporcionarse un conjunto grande de moléculas. Para cada molécula, se predice una afinidad de unión frente al objetivo de proteína. Las puntuaciones resultantes pueden usarse para clasificar las moléculas, siendo más probable que las moléculas con mejor clasificación se unan a la proteína objetivo. Opcionalmente, la lista de moléculas clasificadas puede analizarse para agrupaciones de moléculas similares; una agrupación grande puede usarse como una predicción fuerte de unión de moléculas, o pueden seleccionarse moléculas a través de agrupaciones para garantizar diversidad en los experimentos confirmatorios.

Predicción de efectos secundarios fuera de objetivo. Puede encontrarse que muchos fármacos tienen efectos secundarios. A menudo, estos efectos secundarios se deben a interacciones con trayectorias biológicas distintas de la responsable del efecto terapéutico del fármaco. Estos efectos secundarios fuera de objetivo pueden ser incómodos o peligrosos y restringen la población de pacientes en los que el uso del fármaco es seguro. Efectos secundarios fuera de objetivo son, por lo tanto, un criterio importante con el que evaluar qué fármacos candidatos desarrollar adicionalmente. Mientras es importante caracterizar las interacciones de un fármaco con muchos objetivos biológicos alternativos, tales pruebas pueden ser caras y llevar mucho tiempo en desarrollar y ejecutar. La predicción computacional puede hacer este proceso más eficiente.

En aplicar una realización de la invención, puede construirse un panel de objetivos biológicos que se asocian con respuestas biológicas y/o efectos secundarios significativos. El sistema puede configurarse a continuación para predecir unión contra cada proteína en el panel a su vez. Fuerte actividad (es decir, actividad tan potente como componentes que se conocen que activan la proteína fuera de objetivo) contra un objetivo particular puede implicar la molécula en efectos secundarios debido a efectos fuera de objetivo.

Predicción de toxicidad. Predicción de toxicidad es un caso especial de predicción de efectos secundarios fuera de objetivo particularmente importante. Aproximadamente la mitad de los fármacos candidatos en ensayos clínicos de fase tardía fallan debido a una toxicidad inaceptable. Como parte del nuevo proceso de aprobación de fármaco (y antes de que un fármaco candidato pueda probarse en humanos), la FDA requiere datos de prueba de toxicidad contra un conjunto de objetivos que incluyen las enzimas hepáticas de citocromo P450 (cuya inhibición puede conducir a toxicidad de interacciones de fármaco-fármaco) o el canal hERG (cuya unión puede conducir a prolongación de QT que conduce a arritmias ventriculares y otros efectos cardiacos adversos).

En predicción de toxicidad, el sistema puede configurarse para limitar las proteínas fuera de objetivo para ser antiobjetivos clave (por ejemplo, CYP450, hERG, o receptor 5-HT_{2B}). La afinidad de unión para un fármaco candidato puede predecirse a continuación frente a estas proteínas. Opcionalmente, la molécula puede analizarse

para predecir un conjunto de metabolitos (moléculas posteriores generadas por el cuerpo durante metabolismo/degradación de la molécula original), que también puede analizarse para unión contra los antiobjetivos. Pueden identificarse moléculas problemáticas y modificarse para evitar la toxicidad o puede detenerse el desarrollo en las series moleculares para evitar desperdiciar recursos adicionales.

Optimización de potencia. Uno de los requisitos clave de un fármaco candidato es unión fuerte contra su enfermedad objetivo. Es raro que una criba encuentre compuestos que se unen lo suficientemente fuerte para ser clínicamente efectivos. Por lo tanto, los componentes iniciales proveen un largo proceso de optimización, en el que químicos medicinales modifican iterativamente la estructura molecular para proponer nuevas moléculas con fuerza aumentada de unión objetivo. Cada nueva molécula se sintetiza y prueba, para determinar si los cambios mejoraron satisfactoriamente la unión. El sistema puede configurarse para facilitar este proceso sustituyendo la prueba física por predicción computacional.

En esta aplicación, la enfermedad objetivo y un conjunto de moléculas líder pueden introducirse en el sistema. El sistema puede configurarse para producir predicciones de afinidad de unión para el conjunto de líderes. Opcionalmente, el sistema podría resaltar diferencias entre las moléculas candidatas que podrían ayudar a informar de las razones para las diferencias predichas en afinidad de unión. El usuario químico medicinal puede usar esta información para proponer un nuevo conjunto de moléculas, con suerte, con actividad mejorada contra el objetivo. Estas nuevas moléculas alternativas pueden analizarse de la misma manera.

Optimización de selectividad. Como se ha analizado anteriormente, moléculas tienden a unirse a un huésped de proteínas en diversas fuerzas. Por ejemplo, las cavidades de unión de quinasas de proteína (que son objetivos de quimioterapia populares) son muy similares y la mayoría de inhibidores de quinasa afectan a muchas quinasas diferentes. Esto significa que diversas trayectorias biológicas se modifican simultáneamente, lo que proporciona un perfil medicinal "sucio" y muchos efectos secundarios. El desafío crítico en el diseño de muchos fármacos, por lo tanto, no es actividad *per se* sino especificidad: la capacidad de apuntar selectivamente a una proteína (o un subconjunto de proteínas) de entre un conjunto de proteínas relacionadas posiblemente próximas.

Nuestro sistema puede reducir el tiempo y coste de optimización de la selectividad de un fármaco candidato. En esta aplicación, un usuario puede introducir dos conjuntos de proteínas. Un conjunto describe proteínas contra las que el compuesto debería estar activo, mientras que el otro conjunto describe proteínas contra las que el compuesto debería estar inactivo. El sistema puede configurarse para hacer predicciones para la molécula contra todas las proteínas en ambos conjuntos, estableciendo un perfil de fuerzas de interacción. Opcionalmente, estos perfiles podrían analizarse para sugerir patrones explicativos en las proteínas. El usuario puede usar la información generada por el sistema para considerar modificaciones estructurales a una molécula que mejoraría la unión relativa a los diferentes conjuntos de proteínas, y para diseñar nuevas moléculas candidatas con mejor especificidad. Opcionalmente, el sistema podría configurarse para resaltar diferencias entre las moléculas candidatas que podría ayudar a informar de las razones para las diferencias predichas en selectividad. Los candidatos propuestos pueden analizarse iterativamente, para refinar adicionalmente la especificidad de sus perfiles de actividad.

Función de adecuación para diseño molecular automatizado: son valiosas las herramientas automatizadas para realizar las optimizaciones anteriores. Una molécula exitosa requiere optimización y equilibrio entre potencia, selectividad y toxicidad. "Salto de soporte" (cuando la actividad de un componente líder se preserva pero la estructura química se altera significativamente) puede proporcionar perfiles mejorados de farmacocinética, farmacodinámica, toxicidad o propiedad intelectual. Algoritmos existes para sugerir iterativamente nuevas moléculas, tal como generación aleatoria de moléculas, crecimiento de fragmentos moleculares para rellenar un sitio de unión dado, algoritmos genéticos para "mutar" y "cruzar" una población de moléculas, e intercambio de piezas de una molécula con sustituciones bioisostéricas. Los fármacos candidatos generador por cada uno de estos métodos deben evaluarse contra los múltiples objetivos descritos anteriormente (potencia, selectividad, toxicidad) y, de la misma forma que la tecnología puede ser informativa sobre cada uno de los ajustes manuales anteriores (predicción de unión, selectividad, efecto secundario y predicción de toxicidad), puede incorporarse en un sistema de diseño molecular automatizado.

Reorientación de fármacos. Todos los fármacos tienen efectos secundarios y, de vez en cuando, estos efectos secundarios son beneficiosos. El ejemplo más conocido podría ser la aspirina, que se usa generalmente como un tratamiento contra el dolor de cabeza pero también se toma para la salud cardiovascular. Reorientación de fármacos puede reducir significativamente el coste, tiempo y riesgo de descubrimiento de fármacos porque los fármacos ya se han probado que son seguros en humanos y se han optimizado para una rápida absorción y estabilidad favorable en pacientes. Desafortunadamente, la reorientación de fármacos ha sido principalmente fortuita. Por ejemplo, sildenafilo (Viagra), se desarrolló como un fármaco para la hipertensión y se observó inesperadamente que era un tratamiento efectivo para la disfunción eréctil. La predicción computacional de efectos fuera de objetivo puede usarse en el contexto de reorientación de fármacos para identificar compuestos que podrían usarse para tratar enfermedades alternativas.

En esta aplicación, como en predicción de efectos secundarios fuera de objetivo, el usuario puede reunir un conjunto de posibles proteínas objetivo, en el que cada proteína se vincula a una enfermedad. Es decir, inhibición de cada

proteína trataría una enfermedad (posiblemente diferente); por ejemplo, inhibidores de ciclooxigenasa-2 puede proporcionar alivio para inflamación, mientras que inhibidores de Factor Xa pueden usarse como anticoagulantes. Estas proteínas se anotan con la afinidad de unión de fármacos aprobados, si existe alguno. A continuación reunimos un conjunto de moléculas, restringiendo el conjunto a moléculas que se han aprobado o investigado para uso en humanos. Finalmente, para cada par de proteína y molécula, el usuario puede usar el sistema para predecir la afinidad de unión. Pueden identificarse candidatos para reorientación de fármacos si la afinidad de unión predicha de la molécula está próxima a la afinidad de unión de fármacos efectivos para la proteína.

Predicción de resistencia de fármaco. La resistencia de fármaco es un resultado inevitable de uso farmacéutico, que pone presión de selección en poblaciones de patógenos que se dividen y mutan rápidamente. La resistencia de fármaco se ven en tales diversos agentes de enfermedades como virus (VIH), microorganismos exógenos (MRSA) y células huésped desreguladas (cánceres). Con el paso del tiempo, una medicina se volverá inefectiva, independientemente de si la medicina es antibiótico o quimioterapia. En ese punto, la intervención puede desplazarse a una medicina diferente que es, con suerte, aún potente. En VIH, existen trayectorias de progresión de enfermedad bien conocidas que se definen por qué mutaciones acumulará el virus mientras el paciente se está tratando.

Existe un considerable interés en *predecir* cómo los agentes de la enfermedad se adaptar a la intervención médica. Un enfoque es caracterizar qué mutaciones se producirán en el agente de la enfermedad mientras está en tratamiento. Específicamente, el objetivo de proteína de una medicine necesita mutar para evitar la unión del fármaco mientras simultáneamente continúa uniendo su sustrato natural.

En esta aplicación, puede proponerse un conjunto de posibles mutaciones en la proteína objetivo. Para cada mutación, puede predecirse la forma de proteína resultante. Para cada una de estas formas de proteína mutante, el sistema puede configurarse para predecir una afinidad de unión tanto para el sustrato natural como el fármaco. Las mutaciones, que provocan que la proteína ya no se una al fármaco pero también continúan uniendo al sustrato natural, son candidatas para conferir resistencia de fármaco. Estas proteínas mutadas pueden usarse como objetivos contra los que diseñar fármacos, por ejemplo usando estas proteínas como entradas a uno de estos otros casos de uso de predicción.

Medicina personalizada. Medicinas no efectivas no deberían administrarse. Además del coste e inconvenientes, todas las medicinas tienen efectos secundarios. Consideraciones económicas y morales hacen imperativo proporcionar medicinas únicamente cuando los beneficios superan estos daños. Puede ser importante poder predecir cuándo una medicina será útil. La gente difiere entre sí por un puñado de mutaciones. Sin embargo, pequeñas mutaciones pueden tener efectos profundos. Cuando estas mutaciones se producen en los sitios activos (ortostéricos) o regulatorios (alostéricos) de la enfermedad objetivo, pueden evitar que el fármaco se una y, por lo tanto, bloquee la actividad de la medicina. Cuando se conoce (o predice) la estructura de proteína de una persona particular, el sistema puede configurarse para predecir si un fármaco será efectivo o el sistema puede configurarse para predecir cuándo el fármaco *no* funcionará.

Para esta aplicación, el sistema puede configurarse para recibir como entrada la estructura química del fármaco y la proteína expresada particular del paciente específico. El sistema puede configurarse para predecir unión entre el fármaco y la proteína y, si la afinidad de unión predicha del fármaco que la estructura de proteína del paciente particular es demasiado débil para ser clínicamente efectiva, médicos o facultativos pueden evitar que el fármaco se prescriba inútilmente para ese paciente.

Diseño de ensayo de fármaco. Esta aplicación generaliza el caso de uso de medicina personalizada anterior al caso de poblaciones de pacientes. Cuando el sistema puede predecir si un fármaco será efectivo para un fenotipo de paciente particular, esta información puede usarse para ayudar a diseñar ensayos clínicos. Excluyendo pacientes cuyos objetivos de enfermedad particulares no se verán los suficientemente afectados por un fármaco, un ensayo clínico puede conseguir potencia estadística usando menos pacientes. Menos pacientes reduce directamente el coste y complejidad de ensayos clínicos.

Para esta aplicación, un usuario puede segmentar la posible población de pacientes en subpoblaciones que se caracterizan mediante la expresión de diferentes proteínas (debido a, por ejemplo, mutaciones o isoformas). El sistema puede configurarse para predecir la fuerza de unión del fármaco candidato contra los diferentes tipos de proteína. Si la fuerza de unión predicha contra un tipo de proteína particular indica una concentración de fármaco necesaria que se encuentra por debajo de la concentración hospitalaria clínicamente conseguible (como se basa en, por ejemplo, caracterización física en tubos de ensayo, modelos de animales o voluntarios sanitarios), a continuación el fármaco candidato se predice para que falle para esa subpoblación de proteína. Pacientes con esa proteína pueden excluirse a continuación de un ensayo de fármaco.

Diseño agroquímico. Además de aplicaciones farmacéuticas, la industria agroquímica usa predicción de unión en el diseño de nuevos pesticidas. Por ejemplo, un objetivo esencial para pesticidas es que detengan una única especia de interés, sin impactar de forma adversa en cualquier otra especie. Por seguridad ecológica, una persona podría desear matar un gorgojo sin matar un abejorro.

Para esta aplicación, el usuario podría introducir un conjunto de estructuras de proteína, de las diferentes especies en consideración, en el sistema. Un subconjunto de proteínas podría especificarse como las proteínas contra las que estar activa, mientras que el resto se especificaría como proteínas contra las que las moléculas deberían estar inactivas. Como con casos de uso anteriores, algunos conjuntos de moléculas (ya sea en bases de datos existentes o generadas de nuevo) se considerarían contra cada objetivo, y el sistema devolvería las moléculas con máxima efectividad contra el primer grupo de proteínas mientras evita el segundo.

Ciencia de materiales. Para predecir el comportamiento y propiedades de nuevos materiales, puede ser útil analizar interacciones moleculares. Por ejemplo, para estudiar la solvatación, el usuario puede introducir una estructura cristalina repetida de una molécula pequeña dada y evaluar la afinidad de unión de otra instancia de la molécula pequeña en la superficie del cristal. Para estudiar la resistencia del polímero, puede introducirse un conjunto de hebras de polímero de forma análoga a una estructura de objetivo de proteína, y un oligómero del polímero puede introducirse como una molécula pequeña. El sistema puede predecir, por lo tanto, la afinidad de unión entre las hebras de polímero.

En un ejemplo específico, el sistema puede usarse para predecir la resistencia de un material tal como Kevlar, por ejemplo, prediciendo la resistencia de los enlaces de hidrógeno y los apilamientos de enlaces pi. Por tanto, puede usarse la predicción de afinidad de unión como se divulga en este documento para facilitar el desarrollo de materiales mejorados tales como KEVLAR.

Simulación. Los simuladores a menudo miden la afinidad de unión de una molécula a una proteína, porque la propensión de una molécula para permanecer en una región de la proteína se correlaciona con su afinidad de unión ahí. Una descripción precisa de las características que gobiernan la unión podría usarse para identificar regiones y poses que tienen energía de unión particularmente alta o baja. La descripción energética puede plegarse en simulaciones Monte Carlo para describir el movimiento de una molécula y la ocupación de la región de unión de proteína. De manera similar, simuladores estocásticos para estudiar y modelizar biología de sistemas podrían beneficiarse de predicción precisa de cómo pequeños cambios en concentraciones de moléculas impactan en redes biológicas.

CONCLUSIÓN

La descripción anterior, para propósito de explicación, se ha descrito con referencia a implementaciones específicas. Sin embargo, las descripciones ilustrativas anteriores no se conciben para ser exhaustivas o para limitar las implementaciones a las formas precisas divulgadas. Son posibles muchas modificaciones y variaciones en vista de las anteriores descripciones. Las implementaciones se eligieron y describieron para explicar mejor los principios y sus aplicaciones prácticas, para habilitar de este modo que otros expertos en la materia utilicen las implementaciones y diversas implementaciones con diversas modificaciones según sea adecuado para el uso particular contemplado. El alcance de la presente invención se define mediante los términos de las reivindicaciones adjuntas.

REIVINDICACIONES

1. Un sistema informático para predecir afinidad de unión de un compuesto químico a un polímero objetivo usando datos espaciales, comprendiendo el sistema informático:

al menos un procesador general (74); y
memoria general (90/92) accesible por el al menos un procesador general, almacenando la memoria general al menos un programa (56) para ejecución por el al menos un procesador general, comprendiendo el al menos un programa instrucciones para:

(A) obtener coordenadas espaciales (60) para el polímero objetivo;
(B) modelizar el compuesto químico con el polímero objetivo en cada pose de una pluralidad de diferentes poses, creando de este modo una pluralidad de mapas de vóxel, en donde cada respectivo mapa de vóxel (40) en la pluralidad de mapas de vóxel comprende el compuesto químico en una respectiva pose en la pluralidad de diferentes poses;
(C) desplegar cada mapa de vóxel en la pluralidad de mapas de vóxel en un correspondiente vector, creando de este modo una pluralidad de vectores, en el que cada vector en la pluralidad de vectores es del mismo tamaño;
(D) introducir cada vector respectivo en la pluralidad de vectores en una arquitectura de red que incluye (i) una capa de entrada (26) para recibir secuencialmente la pluralidad de vectores, (ii) una pluralidad de capas convolucionales (28) y (iii) un calificador (30), en donde
la pluralidad de capas convolucionales incluye una capa convolucional inicial y una capa convolucional final, cada capa convolucional en la pluralidad de capas convolucionales está asociada a un conjunto diferente de pesos, en respuesta a entrada de un vector respectivo en la pluralidad de vectores, la capa de entrada proporciona una primera pluralidad de valores en la capa convolucional inicial como una primera función de valores en el vector respectivo,
cada capa convolucional respectiva, distinta de la capa convolucional final, proporciona valores intermedios, como una segunda función respectiva de (i) el conjunto diferente de pesos asociados a la capa convolucional respectiva y (ii) valores de entrada recibidos por la capa convolucional respectiva, en otra capa convolucional en la pluralidad de capas convolucionales, y
la capa convolucional final proporciona valores finales, como una tercera función de (i) el conjunto diferente de pesos asociados a la capa convolucional final y (ii) valores de entrada recibidos por la capa convolucional final, en el calificador;
(E) obtener una pluralidad de puntuaciones desde el calificador, en donde cada puntuación en la pluralidad de puntuaciones corresponde a la entrada de un vector en la pluralidad de vectores en la capa de entrada; y
(F) usar la pluralidad de puntuaciones para predecir la afinidad de unión del compuesto químico con el polímero objetivo.

2. El sistema informático de la reivindicación 1, en el que el polímero es una proteína, un polipéptido, un ácido polinucleico, un ácido polirribonucleico, un polisacárido o un conjunto de cualquier combinación de los mismos.

3. El sistema informático de las reivindicaciones 1 o 2, en el que el uso de la pluralidad de puntuaciones para predecir la afinidad de unión del compuesto químico con el polímero objetivo comprende tomar una medida de tendencia central de la pluralidad de puntuaciones, en donde
cuando la medida de tendencia central satisface un valor umbral predeterminado o intervalo de valores umbral predeterminados, el uso (F) comprende considerar que el compuesto químico tiene una primera clasificación de afinidad de unión, y cuando la medida de tendencia central no satisface el valor umbral predeterminado o intervalo de valores umbral predeterminados, el uso (F) comprende considerar que el compuesto químico tiene una segunda clasificación de afinidad de unión.

4. El sistema informático de una cualquiera de las reivindicaciones 1-3, en el que el uso (F) comprende tomar un promedio ponderado de la pluralidad de puntuaciones, en donde
cuando el promedio ponderado satisface un valor umbral predeterminado o intervalo de valores umbral predeterminados, se considera que el compuesto químico tiene una primera clasificación de afinidad de unión, y
cuando el promedio ponderado no satisface el valor umbral predeterminado o intervalo de valores umbral predeterminados, se considera que el compuesto químico tiene una segunda clasificación de afinidad de unión.

5. El sistema informático de la reivindicación 4, en el que el promedio ponderado es un promedio de Boltzman de la pluralidad de puntuaciones.

6. El sistema informático de la reivindicación 4, en el que
la primera clasificación de afinidad de unión es una concentración inhibitoria media máxima (IC_{50}), concentración efectiva media máxima (EC_{50}), constante de disociación (Kd) o constante de inhibición (KI) para el compuesto químico con respecto al polímero objetivo que está por encima de un primer valor de unión, y
la segunda clasificación es una IC_{50}, EC_{50}, Kd o KI para el compuesto químico con respecto al polímero objetivo que está por debajo del primer valor de unión.

7. El sistema informático de la reivindicación 6, en el que el primer valor de unión es uno micromolar.

8. El sistema informático de la reivindicación 6, en el que el primer valor de unión es diez micromolar.

9. El sistema informático de una cualquiera de las reivindicaciones 1-8, en el que el uso de la pluralidad de puntuaciones para predecir la afinidad de unión del compuesto químico con el polímero objetivo comprende tomar un promedio ponderado de la pluralidad de puntuaciones, en donde cuando el promedio ponderado satisface un intervalo respectivo de valores umbral en una pluralidad de intervalos de valores umbral, el uso (F) comprende considerar que el compuesto químico tiene una clasificación respectiva de afinidad de unión en una pluralidad de clasificaciones de afinidad de unión que inequívocamente corresponde al intervalo respectivo de valores umbral.

10. El sistema informático de la reivindicación 9, en el que cada clasificación respectiva de afinidad de unión en la pluralidad de clasificaciones de afinidad de unión es un intervalo de IC_{50}, EC_{50}, Kd o Kl para el compuesto químico con respecto al polímero objetivo.

11. El sistema informático de la reivindicación 10, en el que una primera clasificación de afinidad de unión en la pluralidad de clasificaciones de afinidad de unión está entre uno micromolar y diez micromolar.

12. El sistema informático de la reivindicación 11, en el que una primera clasificación de afinidad de unión en la pluralidad de clasificaciones de afinidad de unión está entre uno nanomolar y 100 nanomolar.

13. El sistema informático de la reivindicación 1, en el que el polímero objetivo tiene un sitio activo, y la modelización comprende anclar el compuesto químico al sitio activo del polímero objetivo.

14. El sistema informático de la reivindicación 3, en el que
la primera clasificación de afinidad de unión indica que el compuesto químico no es tóxico para un organismo huésped, y la segunda clasificación de afinidad de unión indica que el compuesto químico es tóxico para el organismo huésped.

15. El sistema informático de la reivindicación 3, en el que
la primera clasificación de afinidad de unión indica que la composición química se une a un objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o Kl que está por debajo de un primer valor de unión, y
la segunda clasificación de afinidad de unión indica que la composición química se une a un objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o Kl que está por encima del primer valor de unión.

16. El sistema informático de la reivindicación 3, en el que
la primera clasificación de afinidad de unión indica que la composición química se une a un primer objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o Kl que está por debajo de un primer valor de unión y que la composición química se une a un segundo objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o Kl que está por encima del primer valor de unión, y
la segunda clasificación de afinidad de unión indica que la composición química se une a un primer objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o Kl que está por debajo del primer valor de unión y que la composición química se une a un segundo objetivo de enfermedad molecular con una IC_{50}, EC_{50}, Kd o Kl que está por debajo del primer valor de unión.

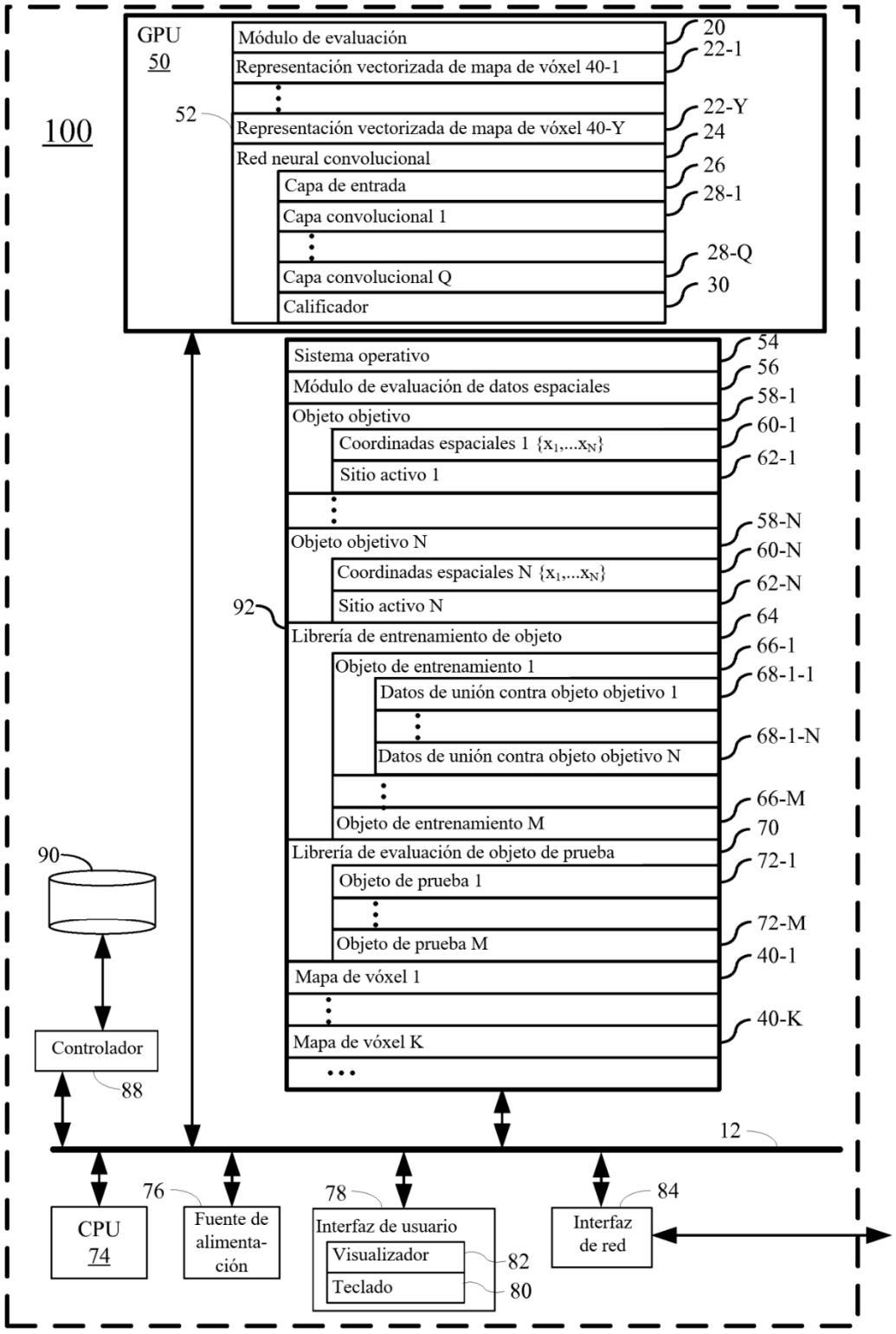
100
GPU
50
52
Módulo de evaluación
Representación vectorizada de mapa de vóxel 40-1
Representación vectorizada de mapa de vóxel 40-Y
Red neural convolucional
Capa de entrada
Capa convolucional 1
Capa convolucional Q
Calificador
20
22-1
22-Y
24
26
28-1
28-Q
30
Sistema operativo
Módulo de evaluación de datos espaciales
Objeto objetivo
Coordinadas espaciales 1 {x1,...xN}
Sitio activo 1
Objeto objetivo N
Coordinadas espaciales N {x1,...xN}
Sitio activo N
92
Librería de entrenamiento de objeto
Objeto de entrenamiento 1
Datos de unión contra objeto objetivo 1
Datos de unión contra objeto objetivo N
Objeto de entrenamiento M
Librería de evaluación de objeto de prueba
Objeto de prueba 1
Objeto de prueba M
Mapa de vóxel 1
Mapa de vóxel K
54
56
58-1
60-1
62-1
58-N
60-N
62-N
64
66-1
68-1-1
68-1-N
66-M
70
72-1
72-M
40-1
40-K
90
Controlador
88
12
CPU
74
76
Fuente de alimenta-ción
78
Interfaz de usuario
Visualizador
Teclado
82
80
84
Interfaz de red

Fig. 1

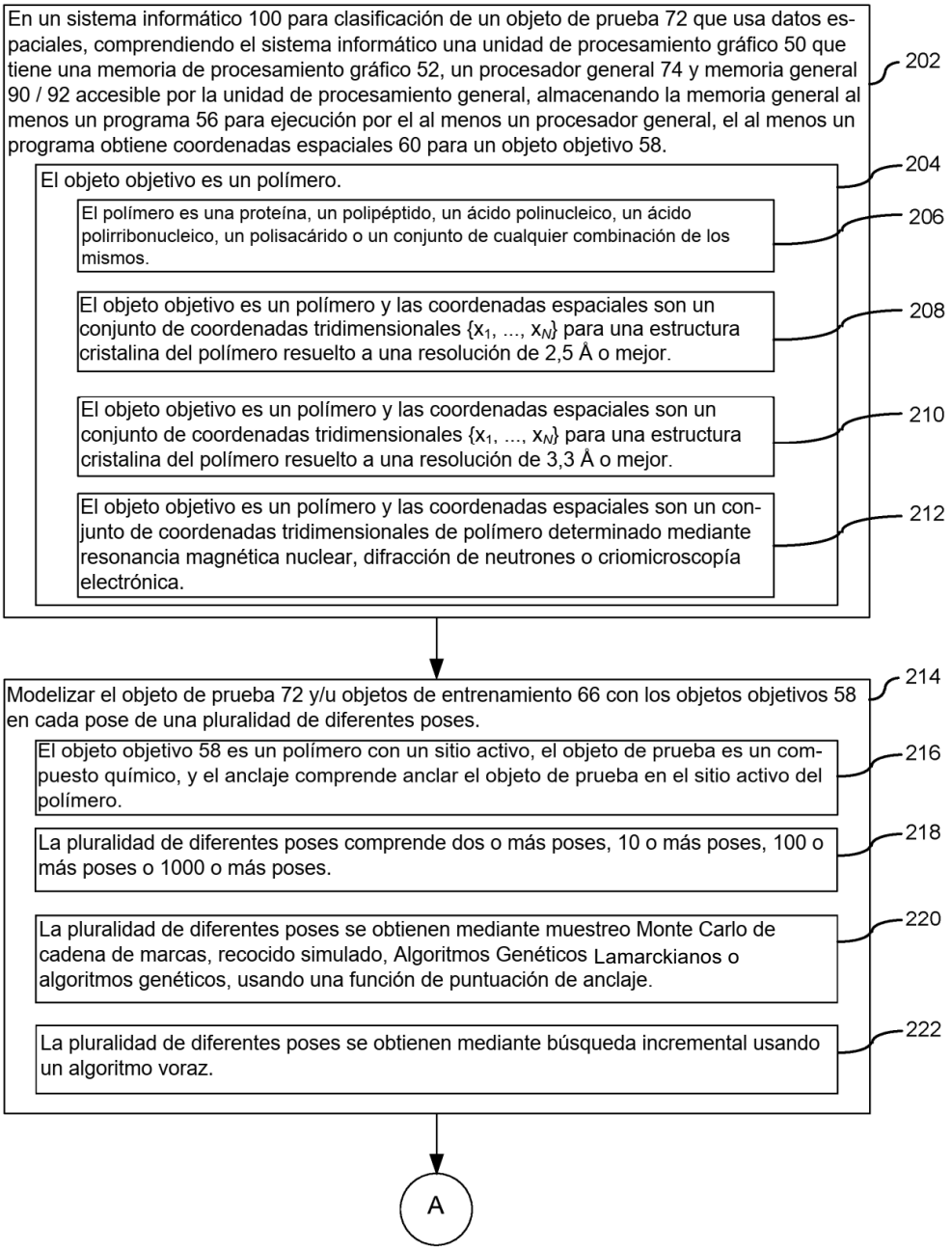
En un sistema informático 100 para clasificación de un objeto de prueba 72 que usa datos espaciales, comprendiendo el sistema informático una unidad de procesamiento gráfico 50 que tiene una memoria de procesamiento gráfico 52, un procesador general 74 y memoria general 90 / 92 accesible por la unidad de procesamiento general, almacenando la memoria general al menos un programa 56 para ejecución por el al menos un procesador general, el al menos un programa obtiene coordenadas espaciales 60 para un objeto objetivo 58.
202
El objeto objetivo es un polímero.
204
El polímero es una proteína, un polipéptido, un ácido polinucleico, un ácido polirribonucleico, un polisacárido o un conjunto de cualquier combinación de los mismos.
206
El objeto objetivo es un polímero y las coordenadas espaciales son un conjunto de coordenadas tridimensionales {x1, ..., xN} para una estructura cristalina del polímero resuelto a una resolución de 2,5 Å o mejor.
208
El objeto objetivo es un polímero y las coordenadas espaciales son un conjunto de coordenadas tridimensionales {x1, ..., xN} para una estructura cristalina del polímero resuelto a una resolución de 3,3 Å o mejor.
210
El objeto objetivo es un polímero y las coordenadas espaciales son un conjunto de coordenadas tridimensionales de polímero determinado mediante resonancia magnética nuclear, difracción de neutrones o criomicroscopía electrónica.
212
Modelizar el objeto de prueba 72 y/u objetos de entrenamiento 66 con los objetos objetivos 58 en cada pose de una pluralidad de diferentes poses.
214
El objeto objetivo 58 es un polímero con un sitio activo, el objeto de prueba es un compuesto químico, y el anclaje comprende anclar el objeto de prueba en el sitio activo del polímero.
216
La pluralidad de diferentes poses comprende dos o más poses, 10 o más poses, 100 o más poses o 1000 o más poses.
218
La pluralidad de diferentes poses se obtienen mediante muestreo Monte Carlo de cadena de marcas, recocido simulado, Algoritmos Genéticos Lamarckianos o algoritmos genéticos, usando una función de puntuación de anclaje.
220
La pluralidad de diferentes poses se obtienen mediante búsqueda incremental usando un algoritmo voraz.
222
A

Fig. 2A

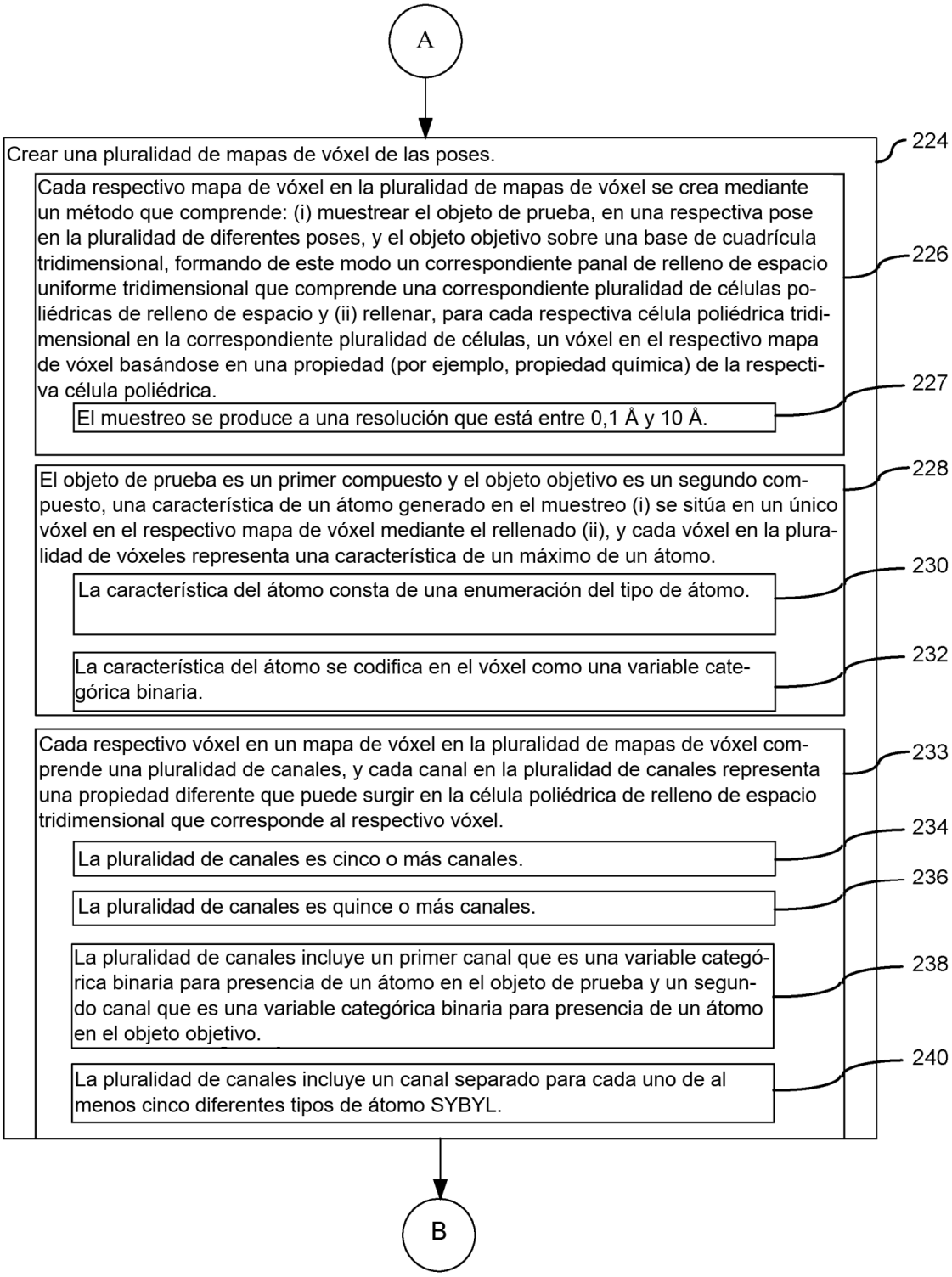
A
Crear una pluralidad de mapas de vóxel de las poses.
224
Cada respectivo mapa de vóxel en la pluralidad de mapas de vóxel se crea mediante un método que comprende: (i) muestrear el objeto de prueba, en una respectiva pose en la pluralidad de diferentes poses, y el objeto objetivo sobre una base de cuadrícula tridimensional, formando de este modo un correspondiente panal de relleno de espacio uniforme tridimensional que comprende una correspondiente pluralidad de células poliédricas de relleno de espacio y (ii) rellenar, para cada respectiva célula poliédrica tridimensional en la correspondiente pluralidad de células, un vóxel en el respectivo mapa de vóxel basándose en una propiedad (por ejemplo, propiedad química) de la respectiva célula poliédrica.
226
El muestreo se produce a una resolución que está entre 0,1 Å y 10 Å.
227
El objeto de prueba es un primer compuesto y el objeto objetivo es un segundo compuesto, una característica de un átomo generado en el muestreo (i) se sitúa en un único vóxel en el respectivo mapa de vóxel mediante el rellenado (ii), y cada vóxel en la pluralidad de vóxeles representa una característica de un máximo de un átomo.
228
La característica del átomo consta de una enumeración del tipo de átomo.
230
La característica del átomo se codifica en el vóxel como una variable categórica binaria.
232
Cada respectivo vóxel en un mapa de vóxel en la pluralidad de mapas de vóxel comprende una pluralidad de canales, y cada canal en la pluralidad de canales representa una propiedad diferente que puede surgir en la célula poliédrica de relleno de espacio tridimensional que corresponde al respectivo vóxel.
233
La pluralidad de canales es cinco o más canales.
234
La pluralidad de canales es quince o más canales.
236
La pluralidad de canales incluye un primer canal que es una variable categórica binaria para presencia de un átomo en el objeto de prueba y un segundo canal que es una variable categórica binaria para presencia de un átomo en el objeto objetivo.
238
La pluralidad de canales incluye un canal separado para cada uno de al menos cinco diferentes tipos de átomo SYBYL.
240
B

Fig. 2B

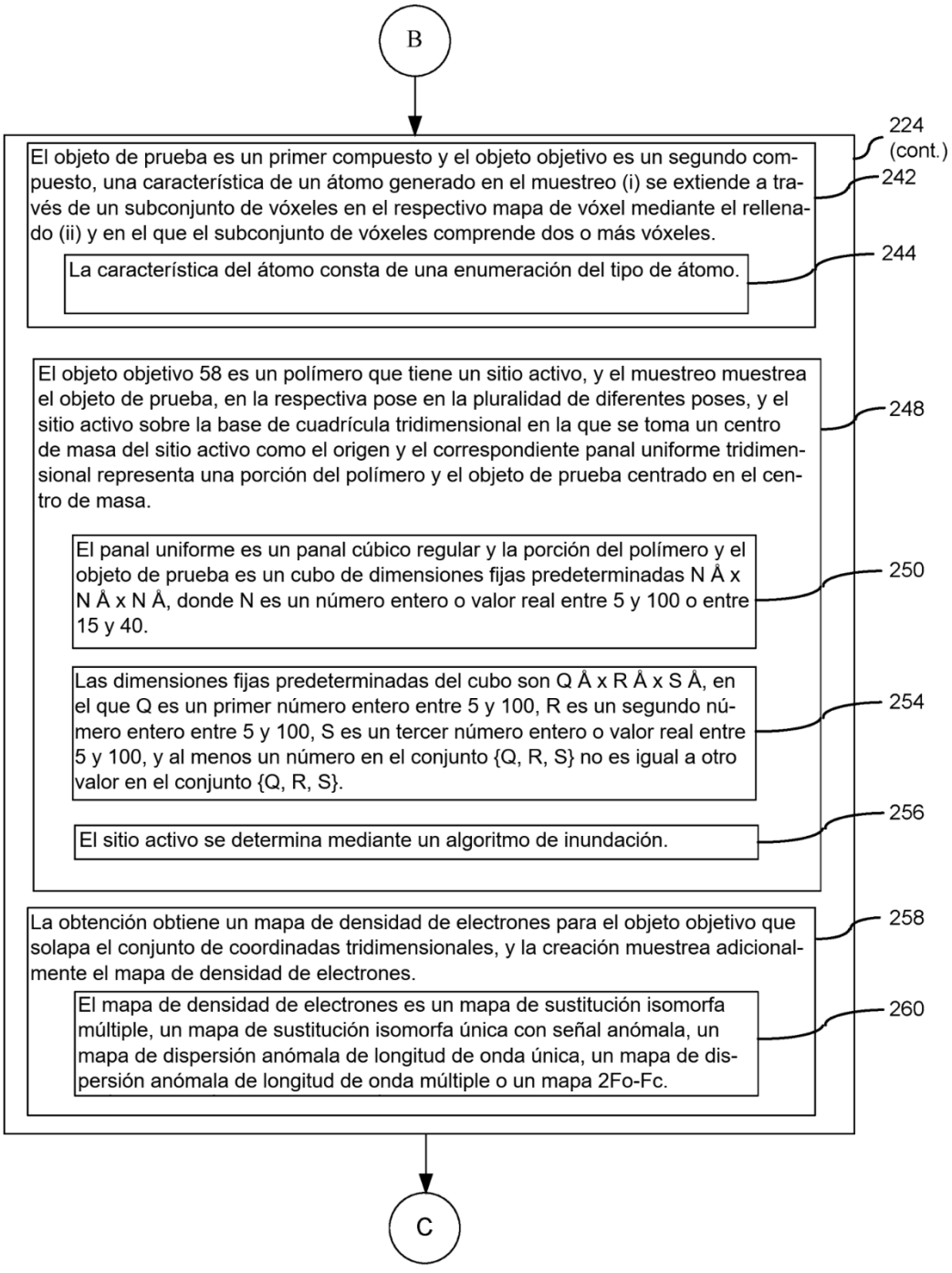
B
224
(cont.)
El objeto de prueba es un primer compuesto y el objeto objetivo es un segundo compuesto, una característica de un átomo generado en el muestreo (i) se extiende a través de un subconjunto de vóxeles en el respectivo mapa de vóxel mediante el rellenado (ii) y en el que el subconjunto de vóxeles comprende dos o más vóxeles.
242
La característica del átomo consta de una enumeración del tipo de átomo.
244
El objeto objetivo 58 es un polímero que tiene un sitio activo, y el muestreo muestrea el objeto de prueba, en la respectiva pose en la pluralidad de diferentes poses, y el sitio activo sobre la base de cuadrícula tridimensional en la que se toma un centro de masa del sitio activo como el origen y el correspondiente panal uniforme tridimensional representa una porción del polímero y el objeto de prueba centrado en el centro de masa.
248
El panal uniforme es un panal cúbico regular y la porción del polímero y el objeto de prueba es un cubo de dimensiones fijas predeterminadas N Å x N Å x N Å, donde N es un número entero o valor real entre 5 y 100 o entre 15 y 40.
250
Las dimensiones fijas predeterminadas del cubo son Q Å x R Å x S Å, en el que Q es un primer número entero entre 5 y 100, R es un segundo número entero entre 5 y 100, S es un tercer número entero o valor real entre 5 y 100, y al menos un número en el conjunto {Q, R, S} no es igual a otro valor en el conjunto {Q, R, S}.
254
El sitio activo se determina mediante un algoritmo de inundación.
256
La obtención obtiene un mapa de densidad de electrones para el objeto objetivo que solapa el conjunto de coordinadas tridimensionales, y la creación muestrea adicionalmente el mapa de densidad de electrones.
258
El mapa de densidad de electrones es un mapa de sustitución isomorfa múltiple, un mapa de sustitución isomorfa única con señal anómala, un mapa de dispersión anómala de longitud de onda única, un mapa de dispersión anómala de longitud de onda múltiple o un mapa 2Fo-Fc.
260
C

Fig. 2C

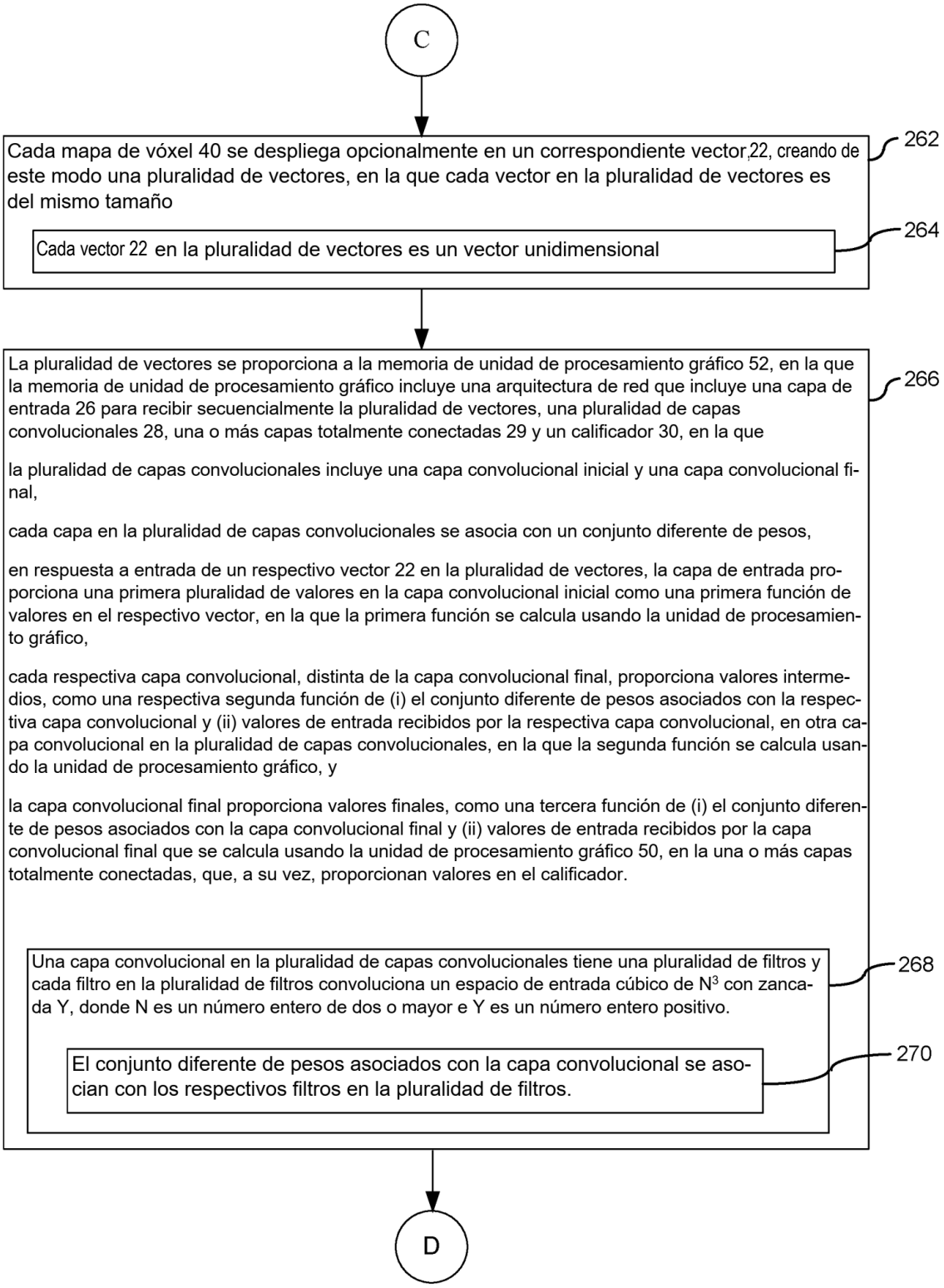
C
Cada mapa de vóxel 40 se despliega opcionalmente en un correspondiente vector 22, creando de este modo una pluralidad de vectores, en la que cada vector en la pluralidad de vectores es del mismo tamaño
262
Cada vector 22 en la pluralidad de vectores es un vector unidimensional
264
La pluralidad de vectores se proporciona a la memoria de unidad de procesamiento gráfico 52, en la que la memoria de unidad de procesamiento gráfico incluye una arquitectura de red que incluye una capa de entrada 26 para recibir secuencialmente la pluralidad de vectores, una pluralidad de capas convolucionales 28, una o más capas totalmente conectadas 29 y un calificador 30, en la que
266
la pluralidad de capas convolucionales incluye una capa convolucional inicial y una capa convolucional final,
cada capa en la pluralidad de capas convolucionales se asocia con un conjunto diferente de pesos,
en respuesta a entrada de un respectivo vector 22 en la pluralidad de vectores, la capa de entrada proporciona una primera pluralidad de valores en la capa convolucional inicial como una primera función de valores en el respectivo vector, en la que la primera función se calcula usando la unidad de procesamiento gráfico,
cada respectiva capa convolucional, distinta de la capa convolucional final, proporciona valores intermedios, como una respectiva segunda función de (i) el conjunto diferente de pesos asociados con la respectiva capa convolucional y (ii) valores de entrada recibidos por la respectiva capa convolucional, en otra capa convolucional en la pluralidad de capas convolucionales, en la que la segunda función se calcula usando la unidad de procesamiento gráfico, y
la capa convolucional final proporciona valores finales, como una tercera función de (i) el conjunto diferente de pesos asociados con la capa convolucional final y (ii) valores de entrada recibidos por la capa convolucional final que se calcula usando la unidad de procesamiento gráfico 50, en la una o más capas totalmente conectadas, que, a su vez, proporcionan valores en el calificador.
Una capa convolucional en la pluralidad de capas convolucionales tiene una pluralidad de filtros y cada filtro en la pluralidad de filtros convoluciona un espacio de entrada cúbico de N^3 con zancada Y, donde N es un número entero de dos o mayor e Y es un número entero positivo.
268
El conjunto diferente de pesos asociados con la capa convolucional se asocian con los respectivos filtros en la pluralidad de filtros.
270
D

Fig. 2D

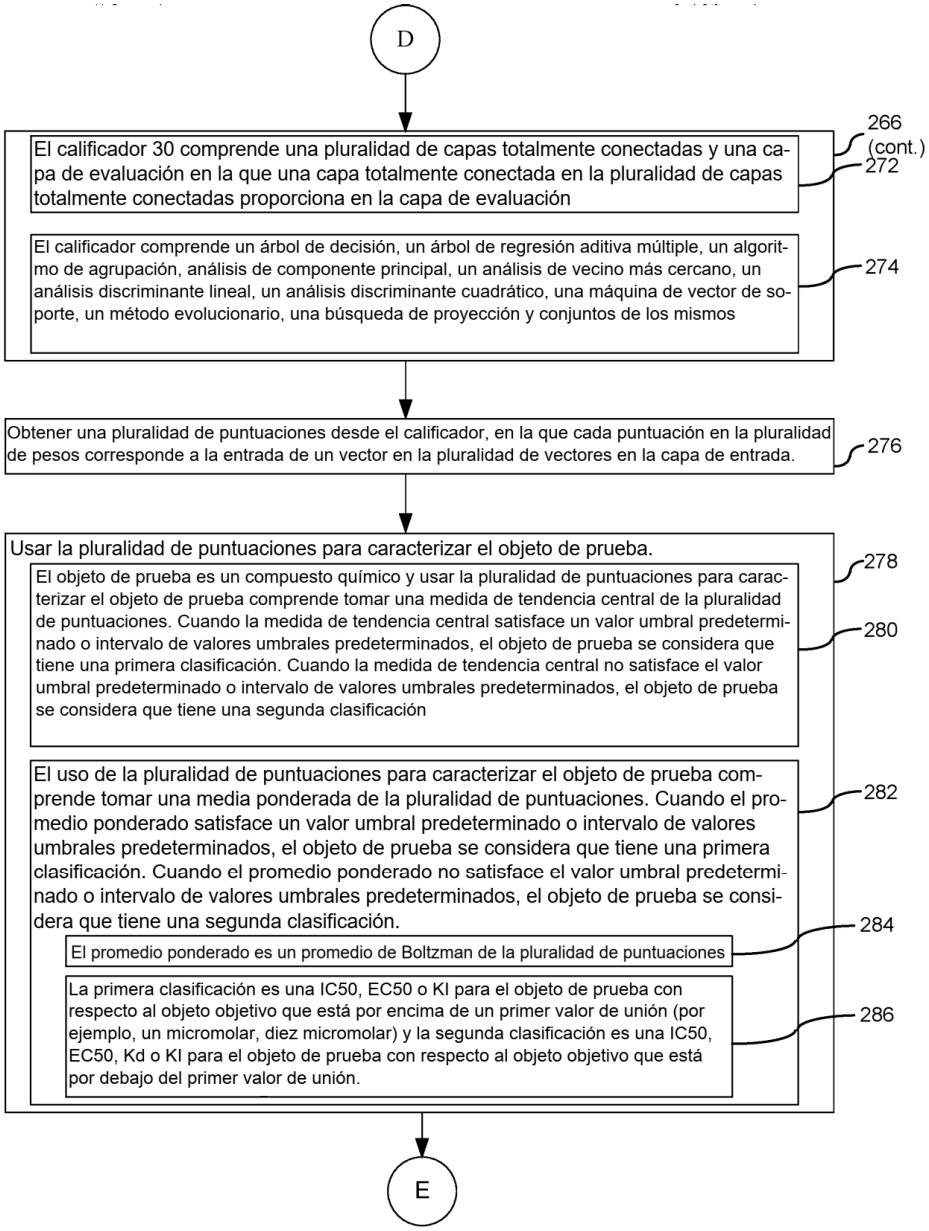
D
266 (cont.)
El calificador 30 comprende una pluralidad de capas totalmente conectadas y una capa de evaluación en la que una capa totalmente conectada en la pluralidad de capas totalmente conectadas proporciona en la capa de evaluación
272
El calificador comprende un árbol de decisión, un árbol de regresión aditiva múltiple, un algoritmo de agrupación, análisis de componente principal, un análisis de vecino más cercano, un análisis discriminante lineal, un análisis discriminante cuadrático, una máquina de vector de soporte, un método evolucionario, una búsqueda de proyección y conjuntos de los mismos
274
Obtener una pluralidad de puntuaciones desde el calificador, en la que cada puntuación en la pluralidad de pesos corresponde a la entrada de un vector en la pluralidad de vectores en la capa de entrada.
276
Usar la pluralidad de puntuaciones para caracterizar el objeto de prueba.
278
El objeto de prueba es un compuesto químico y usar la pluralidad de puntuaciones para caracterizar el objeto de prueba comprende tomar una medida de tendencia central de la pluralidad de puntuaciones. Cuando la medida de tendencia central satisface un valor umbral predeterminado o intervalo de valores umbrales predeterminados, el objeto de prueba se considera que tiene una primera clasificación. Cuando la medida de tendencia central no satisface el valor umbral predeterminado o intervalo de valores umbrales predeterminados, el objeto de prueba se considera que tiene una segunda clasificación
280
El uso de la pluralidad de puntuaciones para caracterizar el objeto de prueba comprende tomar una media ponderada de la pluralidad de puntuaciones. Cuando el promedio ponderado satisface un valor umbral predeterminado o intervalo de valores umbrales predeterminados, el objeto de prueba se considera que tiene una primera clasificación. Cuando el promedio ponderado no satisface el valor umbral predeterminado o intervalo de valores umbrales predeterminados, el objeto de prueba se considera que tiene una segunda clasificación.
282
El promedio ponderado es un promedio de Boltzman de la pluralidad de puntuaciones
284
La primera clasificación es una IC50, EC50 o KI para el objeto de prueba con respecto al objeto objetivo que está por encima de un primer valor de unión (por ejemplo, un micromolar, diez micromolar) y la segunda clasificación es una IC50, EC50, Kd o KI para el objeto de prueba con respecto al objeto objetivo que está por debajo del primer valor de unión.
286
E

Fig. 2E

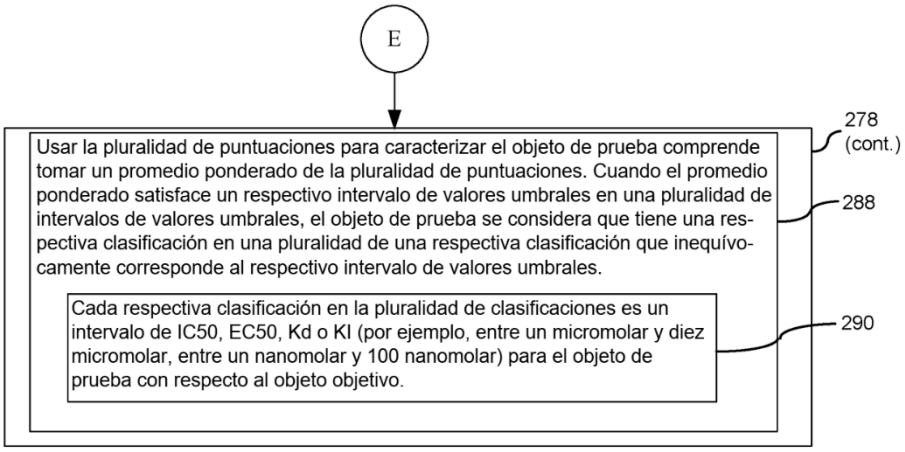
E
278
(cont.)
Usar la pluralidad de puntuaciones para caracterizar el objeto de prueba comprende tomar un promedio ponderado de la pluralidad de puntuaciones. Cuando el promedio ponderado satisface un respectivo intervalo de valores umbrales en una pluralidad de intervalos de valores umbrales, el objeto de prueba se considera que tiene una respectiva clasificación en una pluralidad de una respectiva clasificación que inequívocamente corresponde al respectivo intervalo de valores umbrales.
288
Cada respectiva clasificación en la pluralidad de clasificaciones es un intervalo de IC50, EC50, Kd o KI (por ejemplo, entre un micromolar y diez micromolar, entre un nanomolar y 100 nanomolar) para el objeto de prueba con respecto al objeto objetivo.
290

Fig. 2F

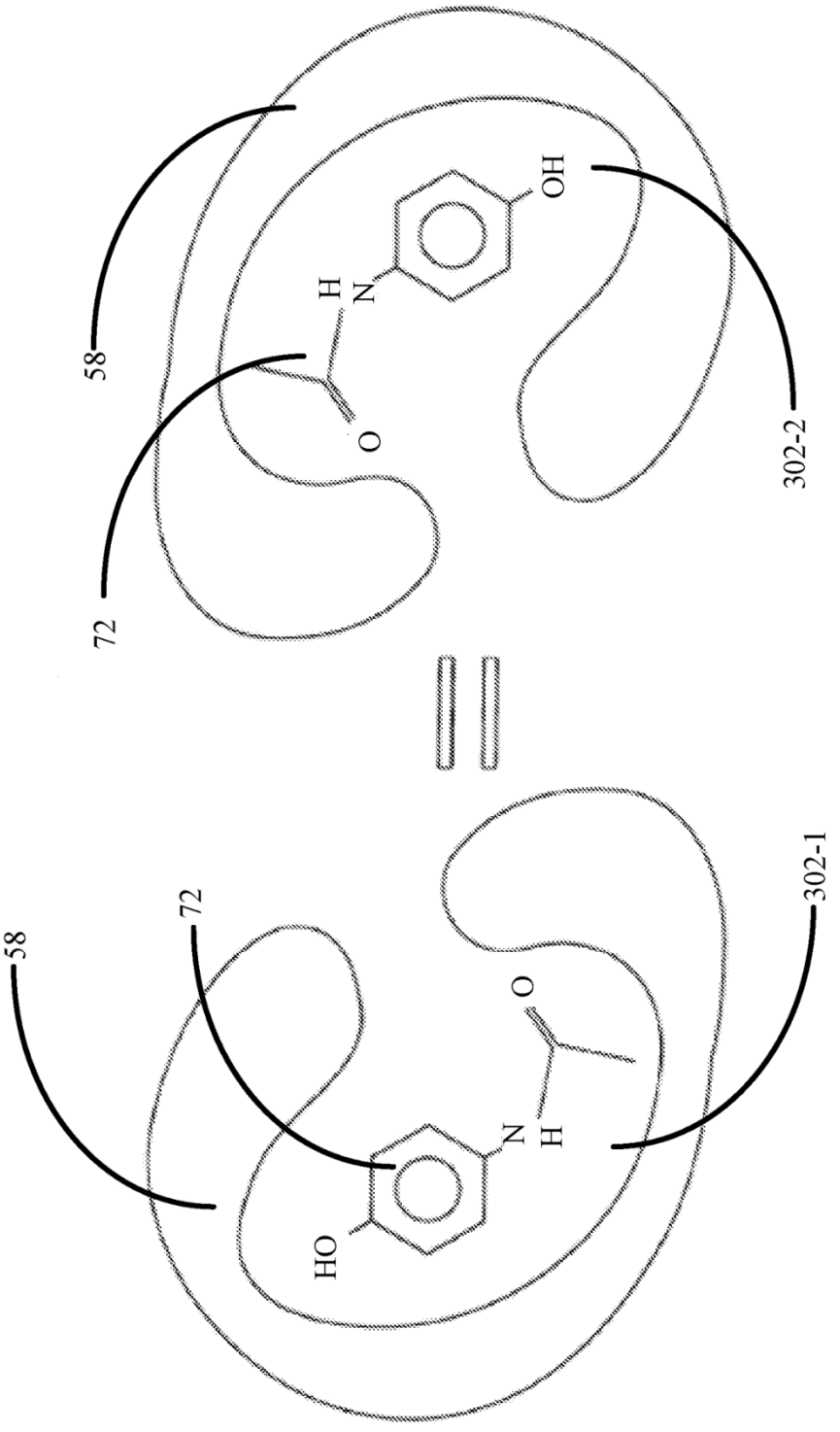

Fig. 3

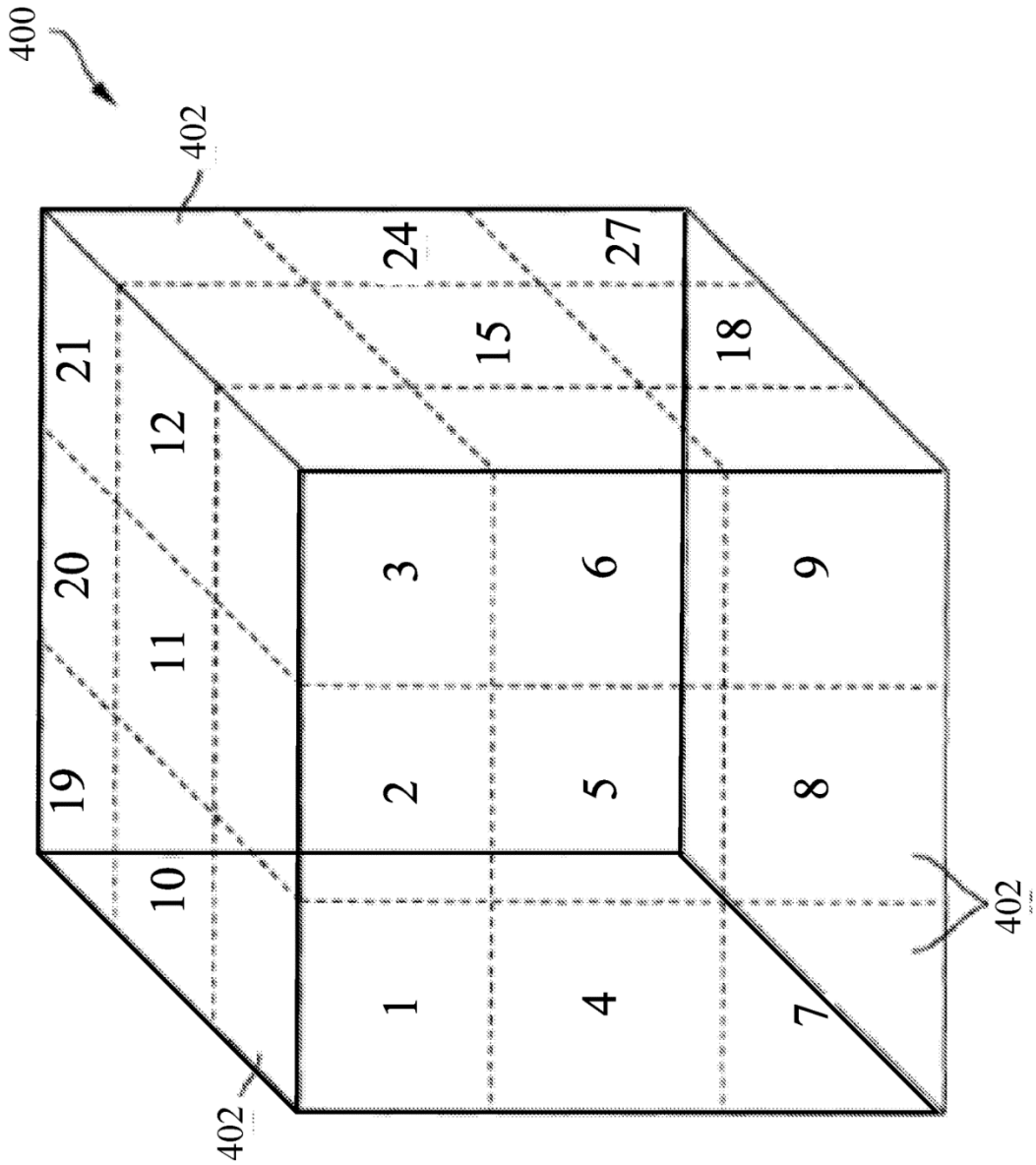

Fig. 4

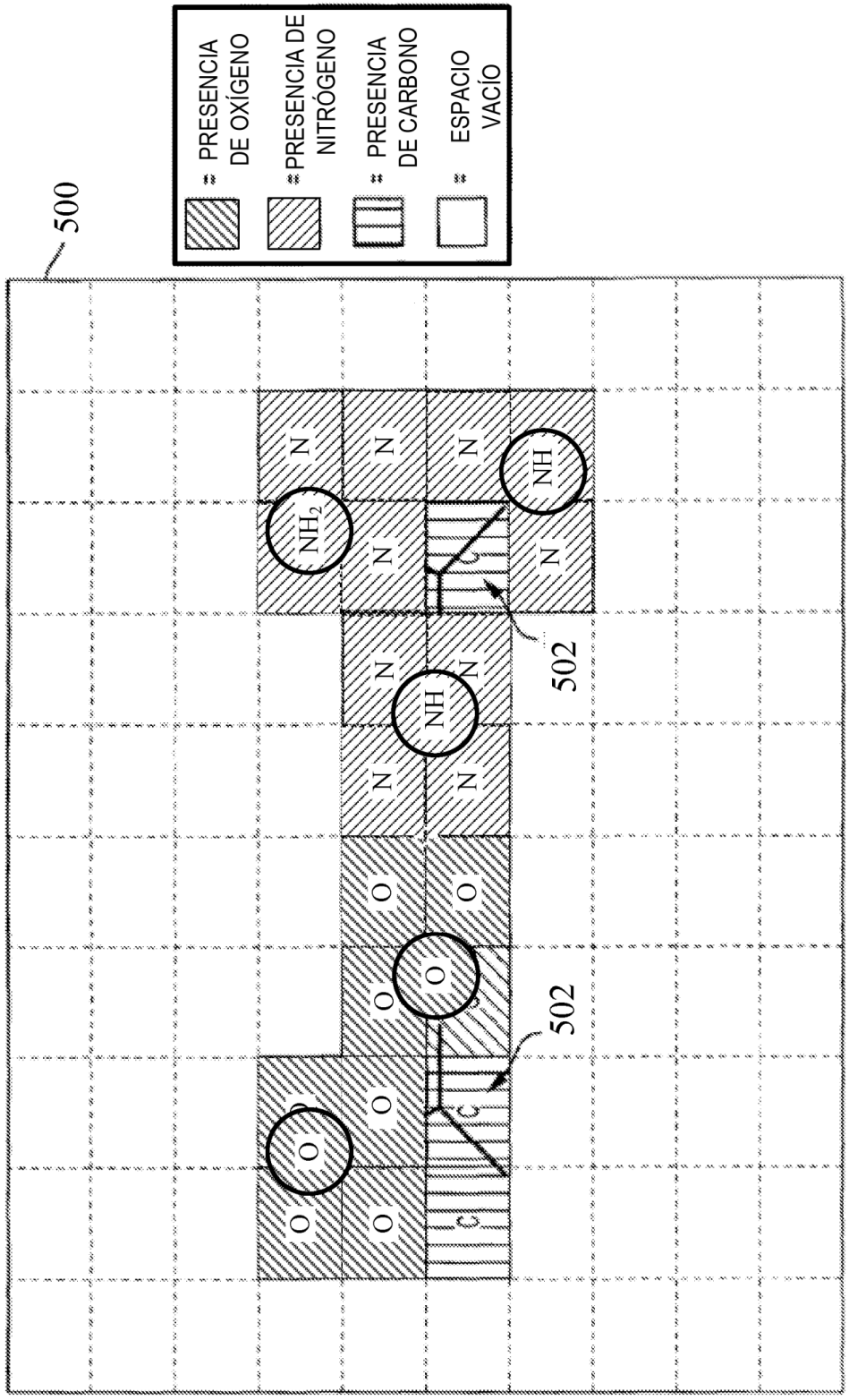
500
PRESENCIA DE OXÍGENO
PRESENCIA DE NITRÓGENO
PRESENCIA DE CARBONO
ESPACIO VACÍO
502
502

Fig. 5

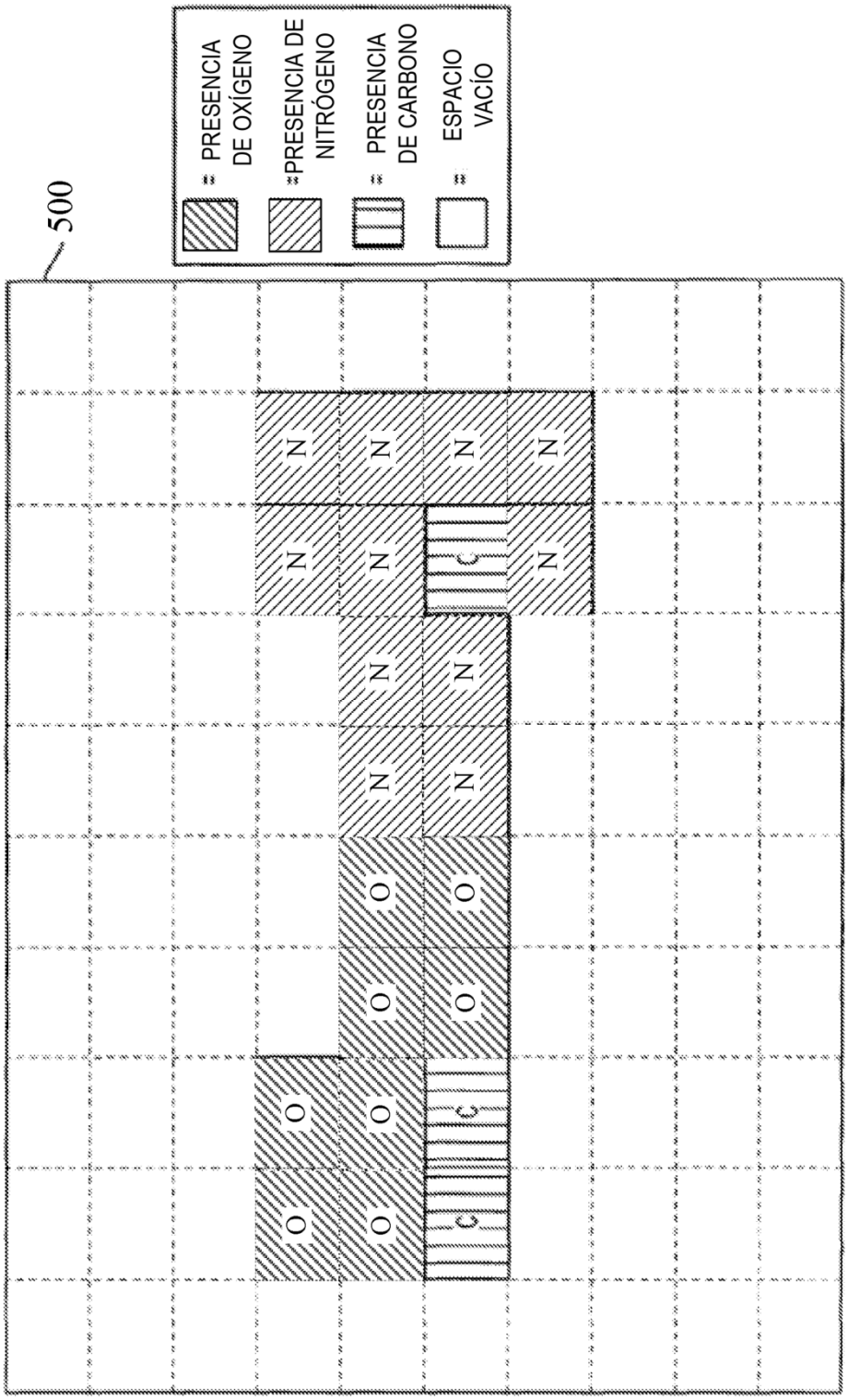
500
= PRESENCIA DE OXÍGENO
= PRESENCIA DE NITRÓGENO
= PRESENCIA DE CARBONO
= ESPACIO VACÍO

Fig. 6

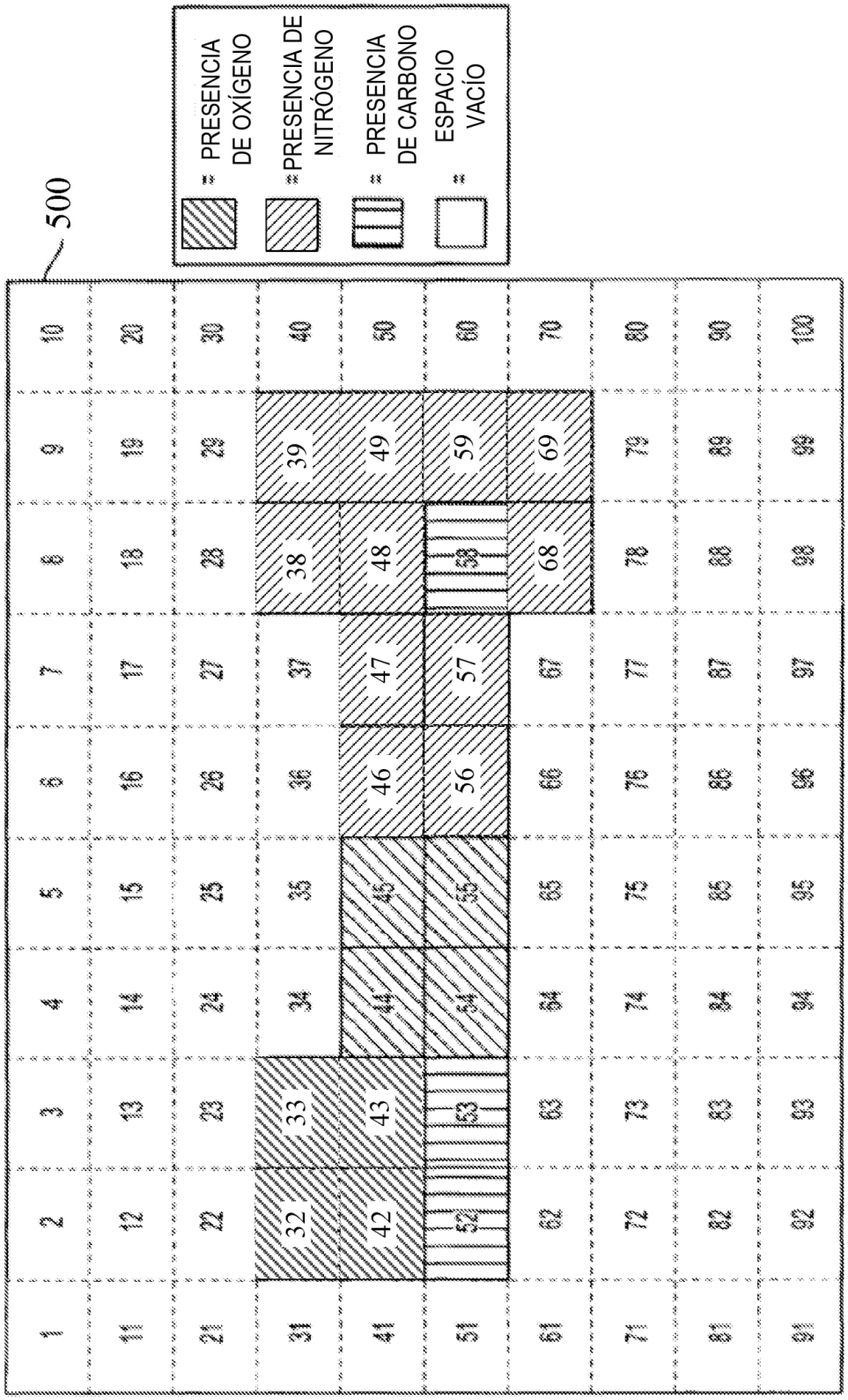
500
= PRESENCIA DE OXÍGENO
= PRESENCIA DE NITRÓGENO
= PRESENCIA DE CARBONO
= ESPACIO VACÍO

Fig. 7

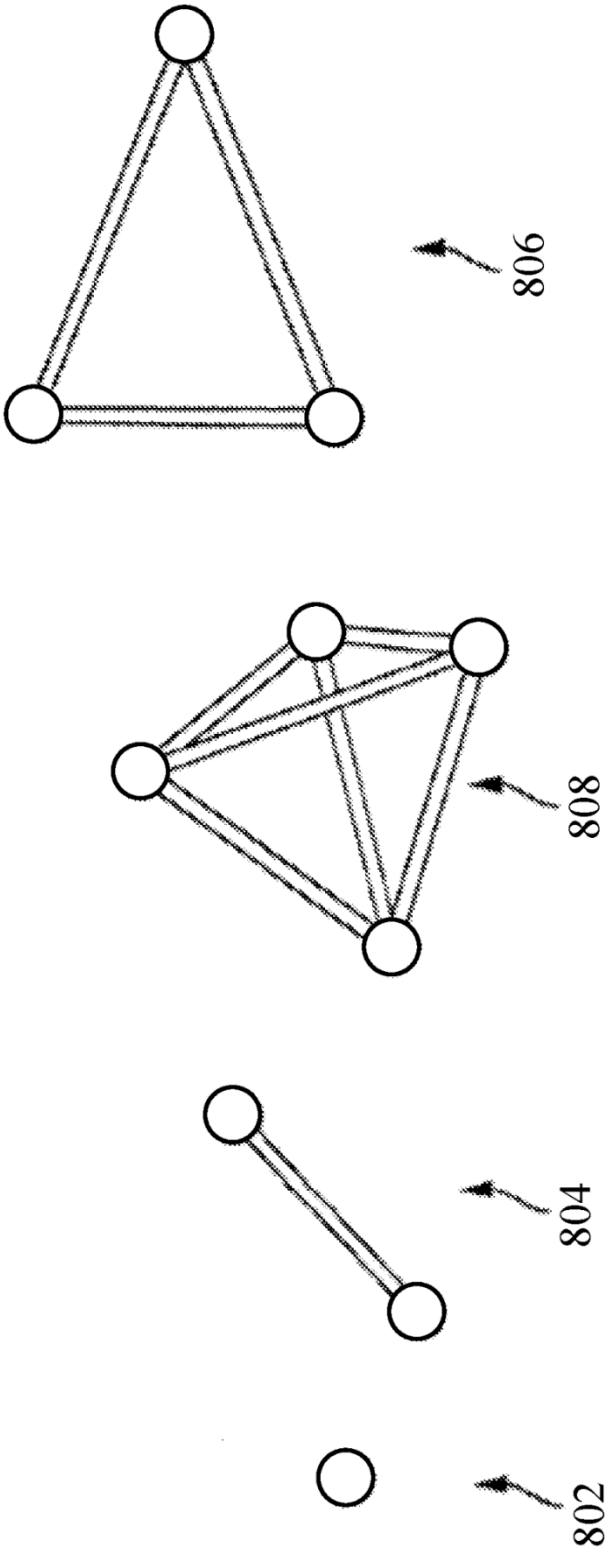

Fig. 8

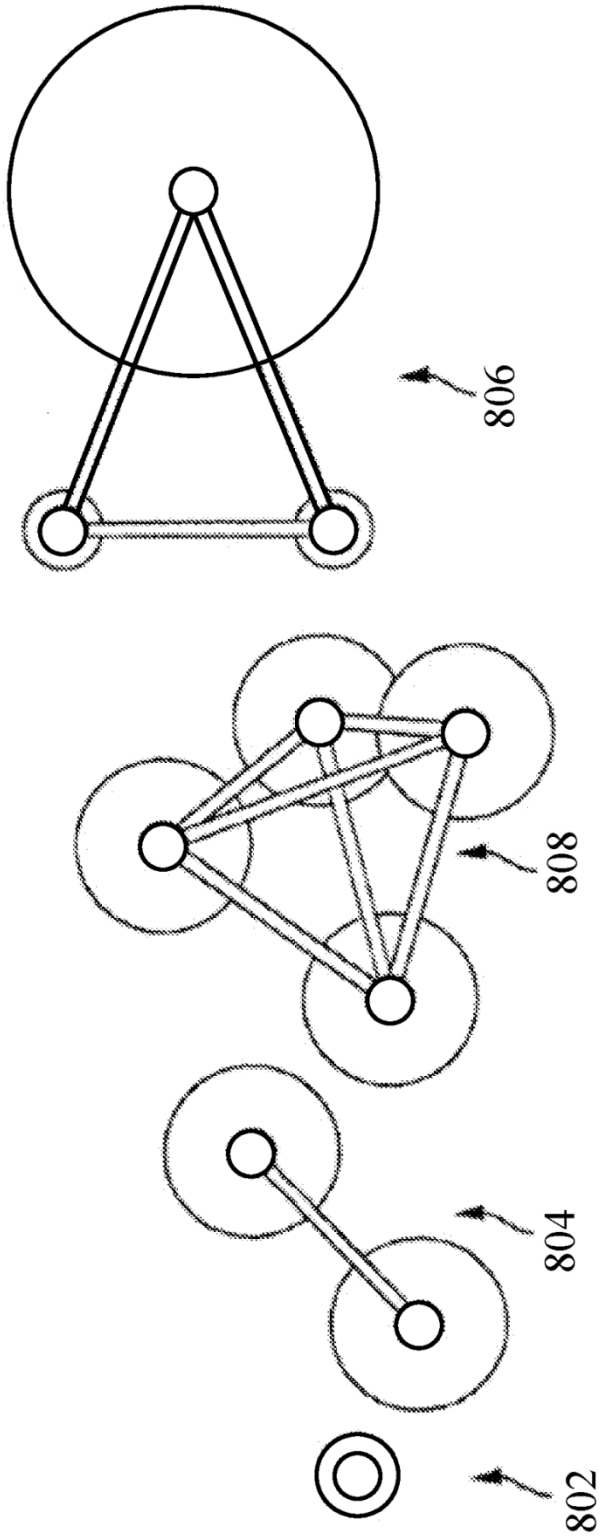

Fig. 9

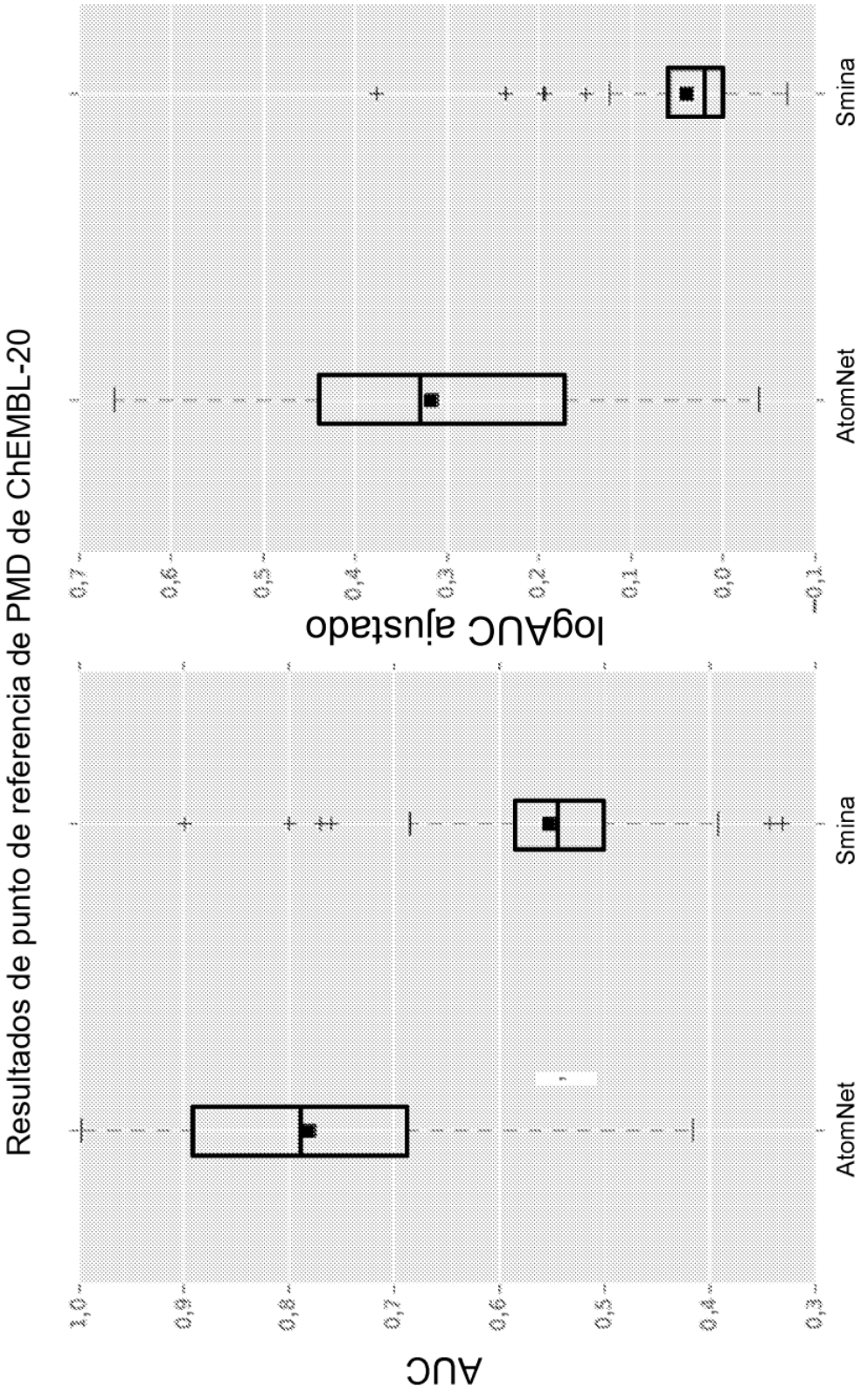
Resultados de punto de referencia de PMD de ChEMBL-20
AUC
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
AtomNet
Smina
logAUC ajustado
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
−0,1
AtomNet
Smina

Fig. 10

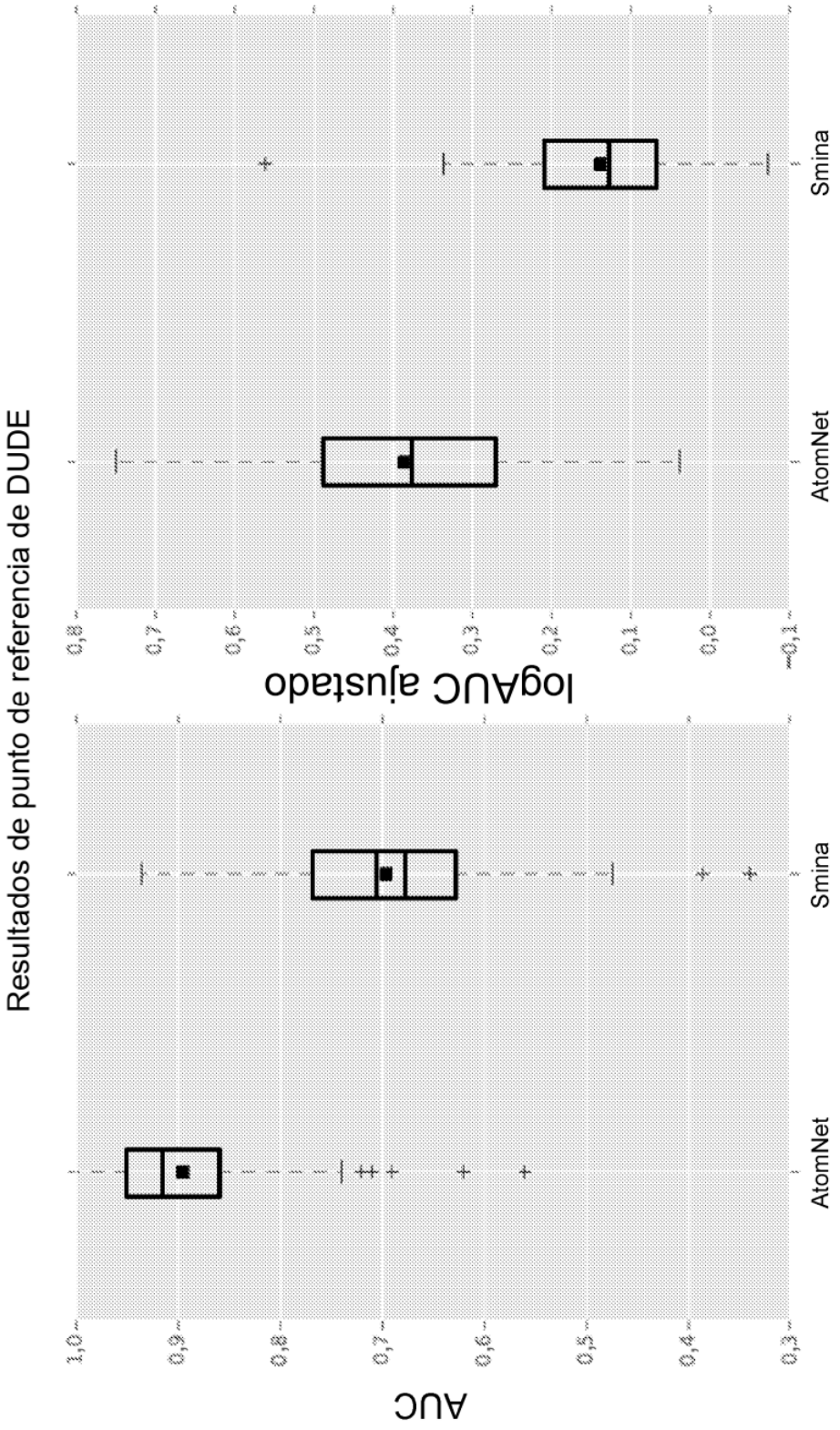
Resultados de punto de referencia de DUDE
AUC
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
AtomNet
Smina
logAUC ajustado
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
−0,1
AtomNet
Smina

Fig. 11

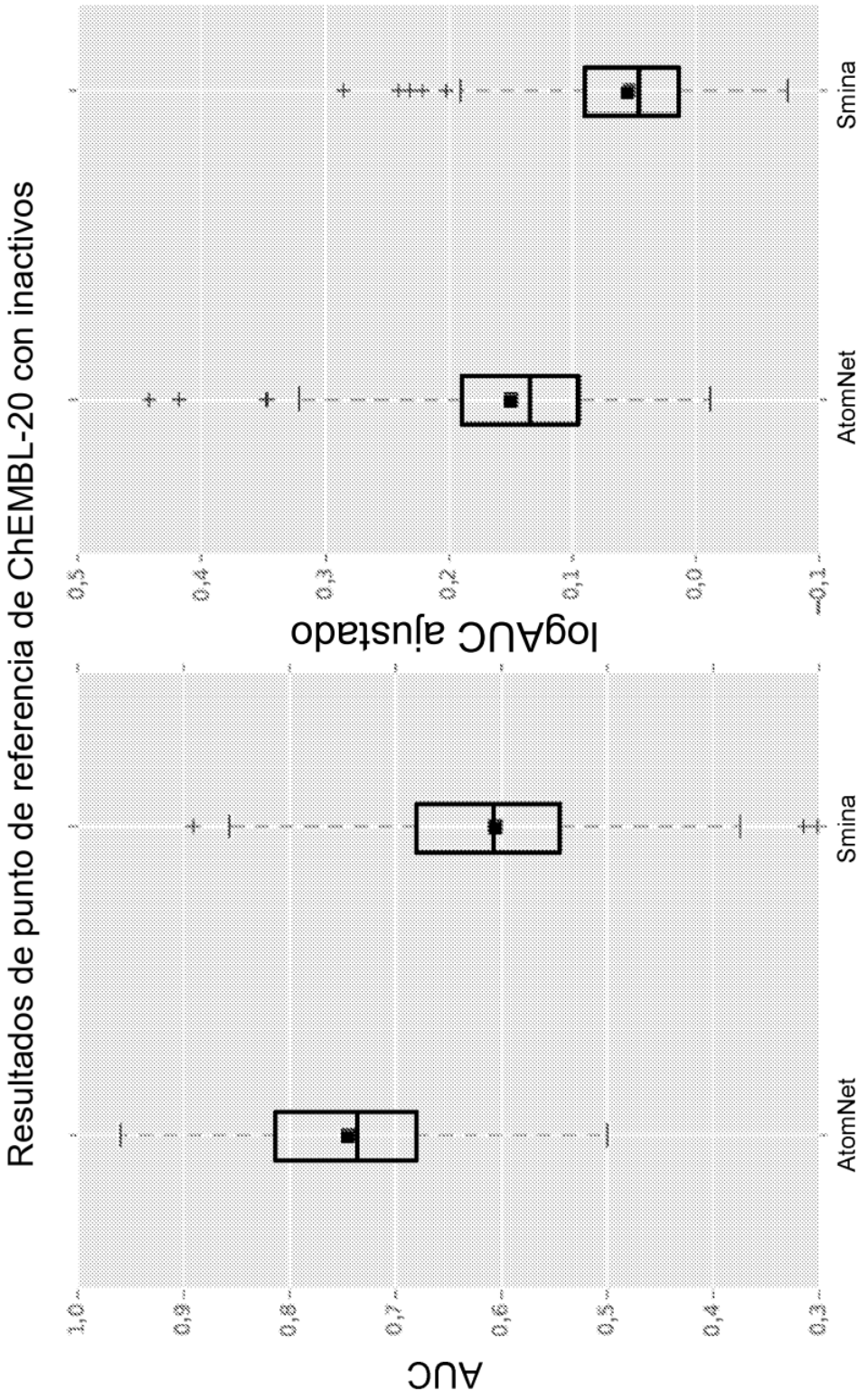
Resultados de punto de referencia de ChEMBL-20 con inactivos
AUC
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
AtomNet
Smina
logAUC ajustado
0,5
0,4
0,3
0,2
0,1
0,0
−0,1
AtomNet
Smina

Fig. 12

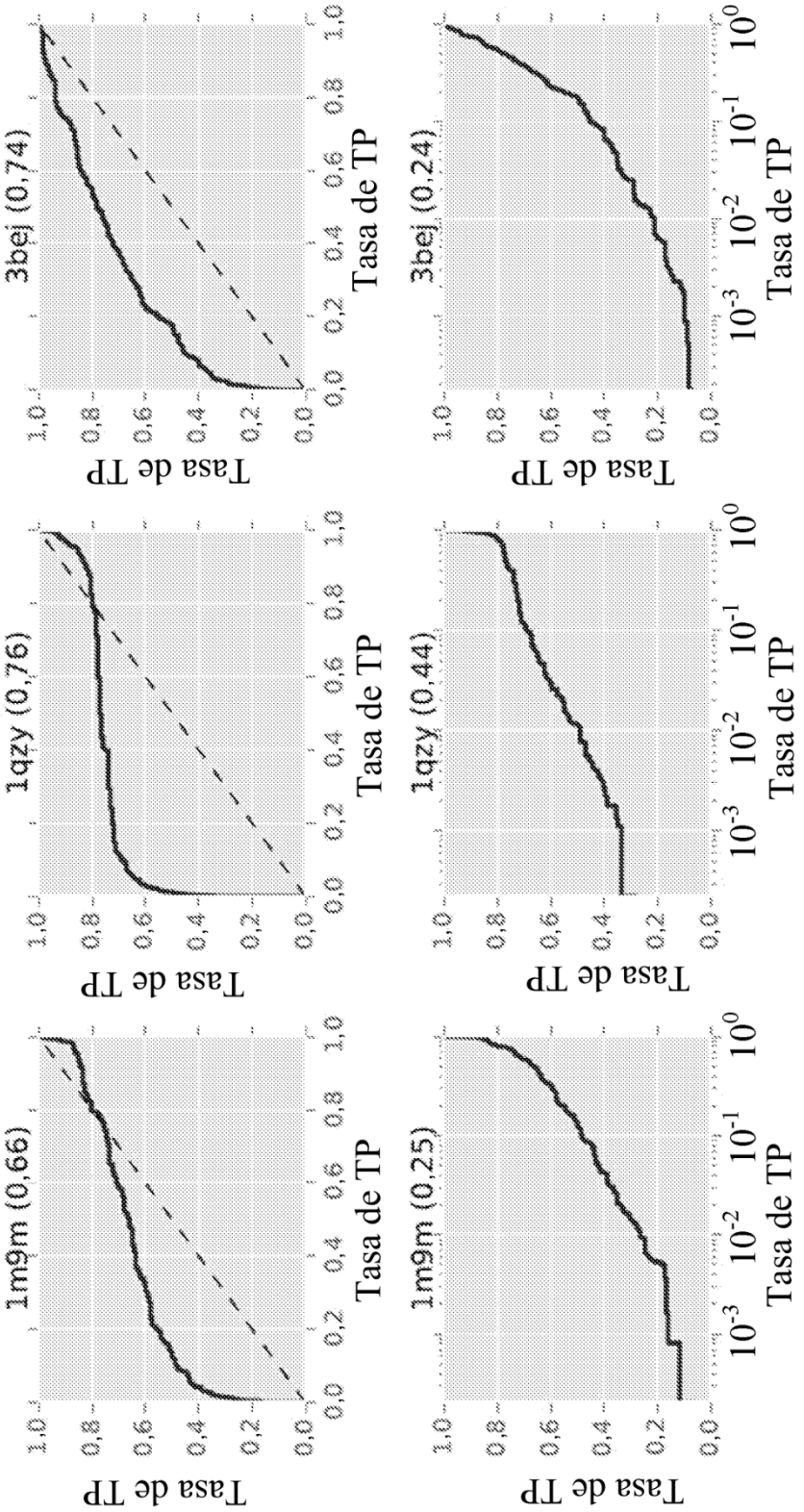
3bej (0,74)
3bej (0,24)
1qzy (0,76)
1qzy (0,44)
1m9m (0,66)
1m9m (0,25)
Tasa de TP
Tasa de TP

Fig. 13A

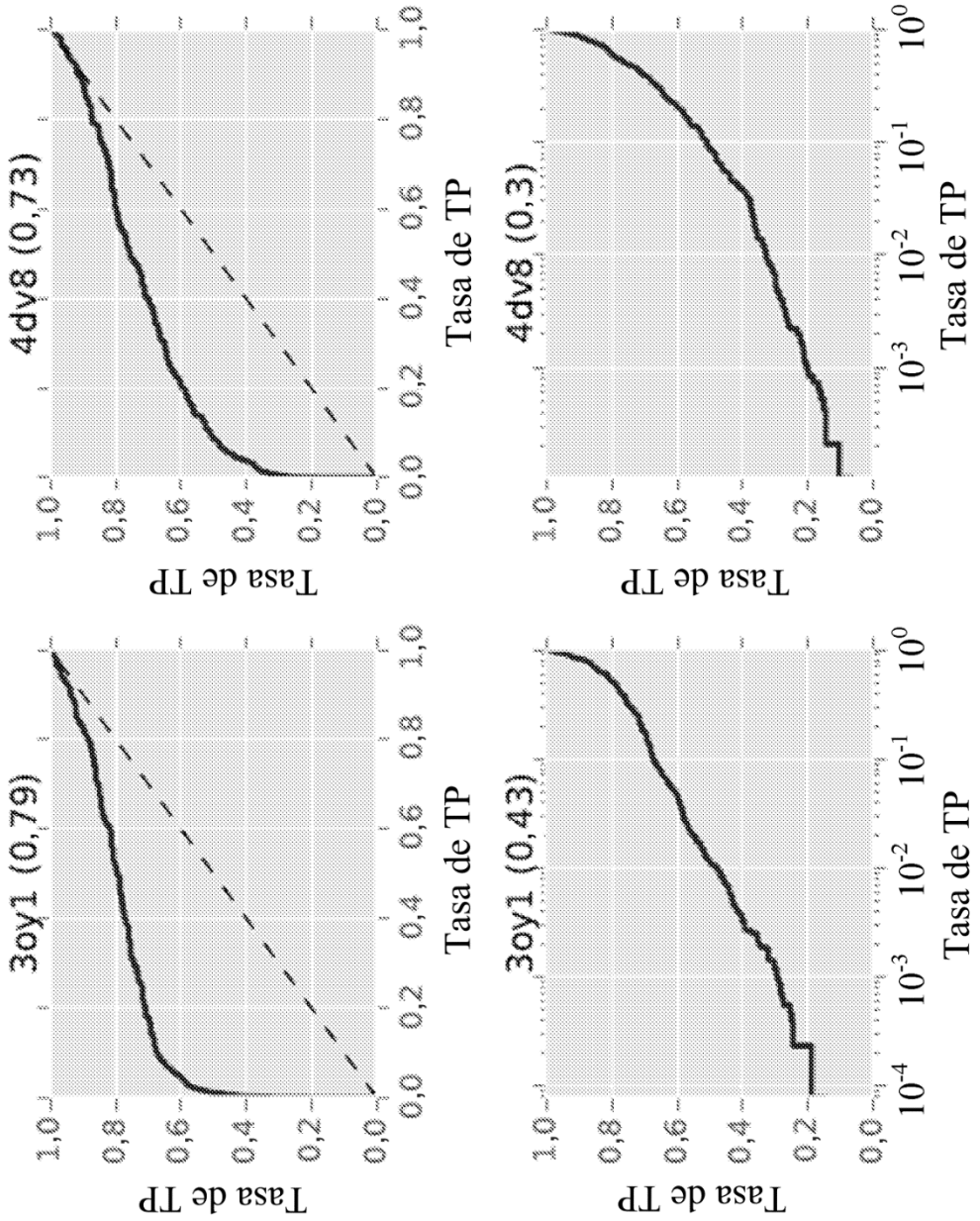

Fig. 13B

$$f(\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_{100}}) = g\big(g_0(\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_{100}}), g_1(\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_{100}}), \ldots, g_n(\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_{100}})\big)$$

$$\overrightarrow{x_1} = (x_1^1,\ x_1^2, \ldots,\ x_1^{40})$$

$$\overrightarrow{x_2} = (x_2^1,\ x_2^2, \ldots,\ x_2^{40})$$

Fig. 14

Fig. 15A

Fig. 15B