Implementación de redes neuronales convolucionales para el meta-análisis de acoplamientos moleculares de complejos proteína-ligando

Adrián Antonio Rodríguez Pié

21 de noviembre de 2019

Universidad Nacional Autónoma de México

Outline

Sobre proteínas

Sobre inteligencia artificial

Sobre redes y neuronas

Sobre meta-análisis del acoplamiento

Deep-pose

Entrenamiento y resultados

¿Preguntas?

Sobre proteínas

Proteínas

Orígen

Originado del griego *proteios* que significa "primario" o "de primer orden".

Proteínas

Orígen

Originado del griego *proteios* que significa "primario" o "de primer orden".

Definición (según la RAE)

Sustancia constitutiva de la materia viva, formada por una o varias cadenas de aminoácidos.

Ligandos

- Un ligando es una molécula que se une a otra molécula específica, en algunos casos mandando una señal en el proceso.

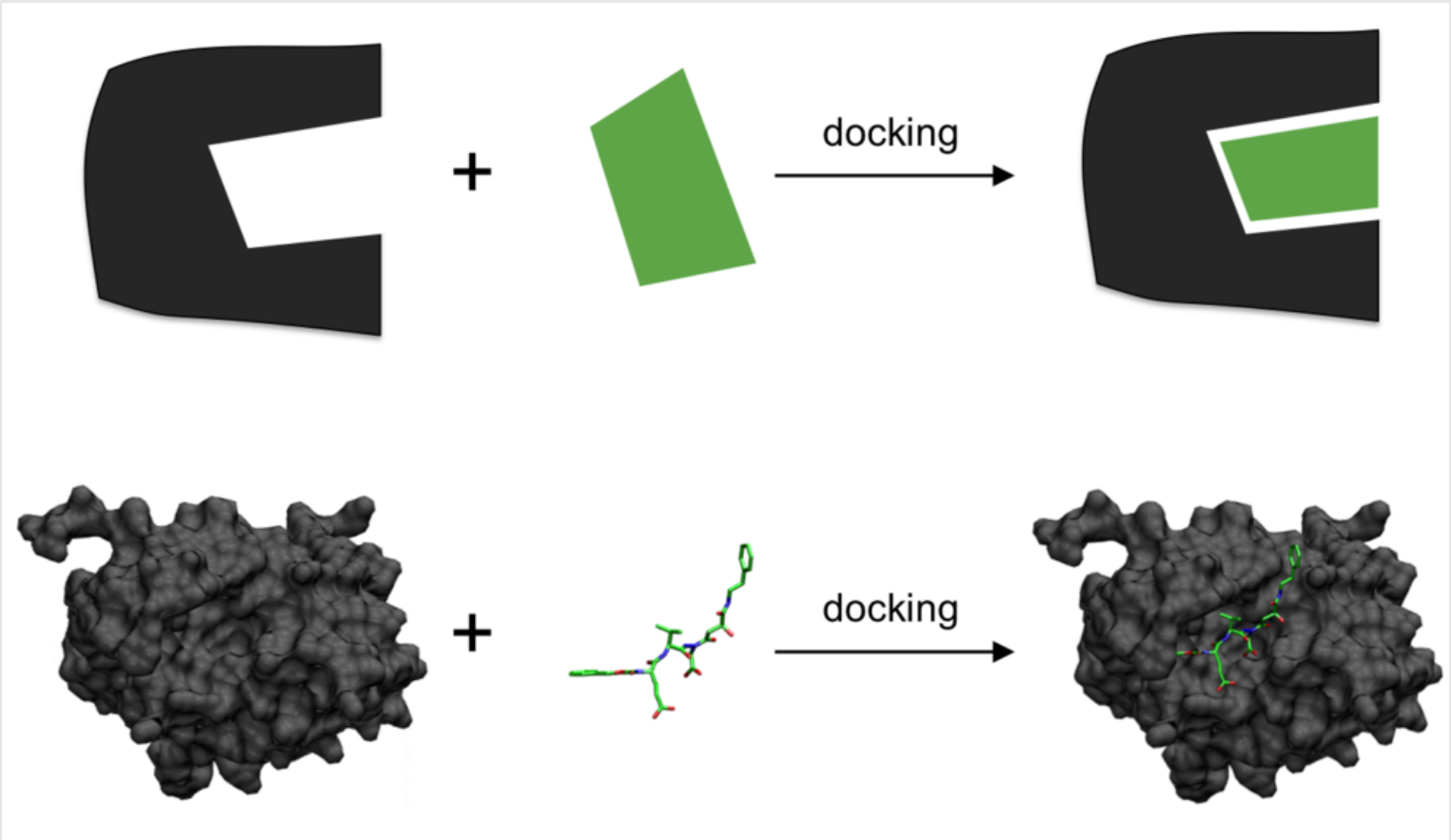
Ligandos

- Un ligando es una molécula que se une a otra molécula específica, en algunos casos mandando una señal en el proceso.
- Estos ligandos interactuan con moleculas objetivo (usualmente otras proteínas). Son a estas proteínas a las que llamamos receptores o residuos.

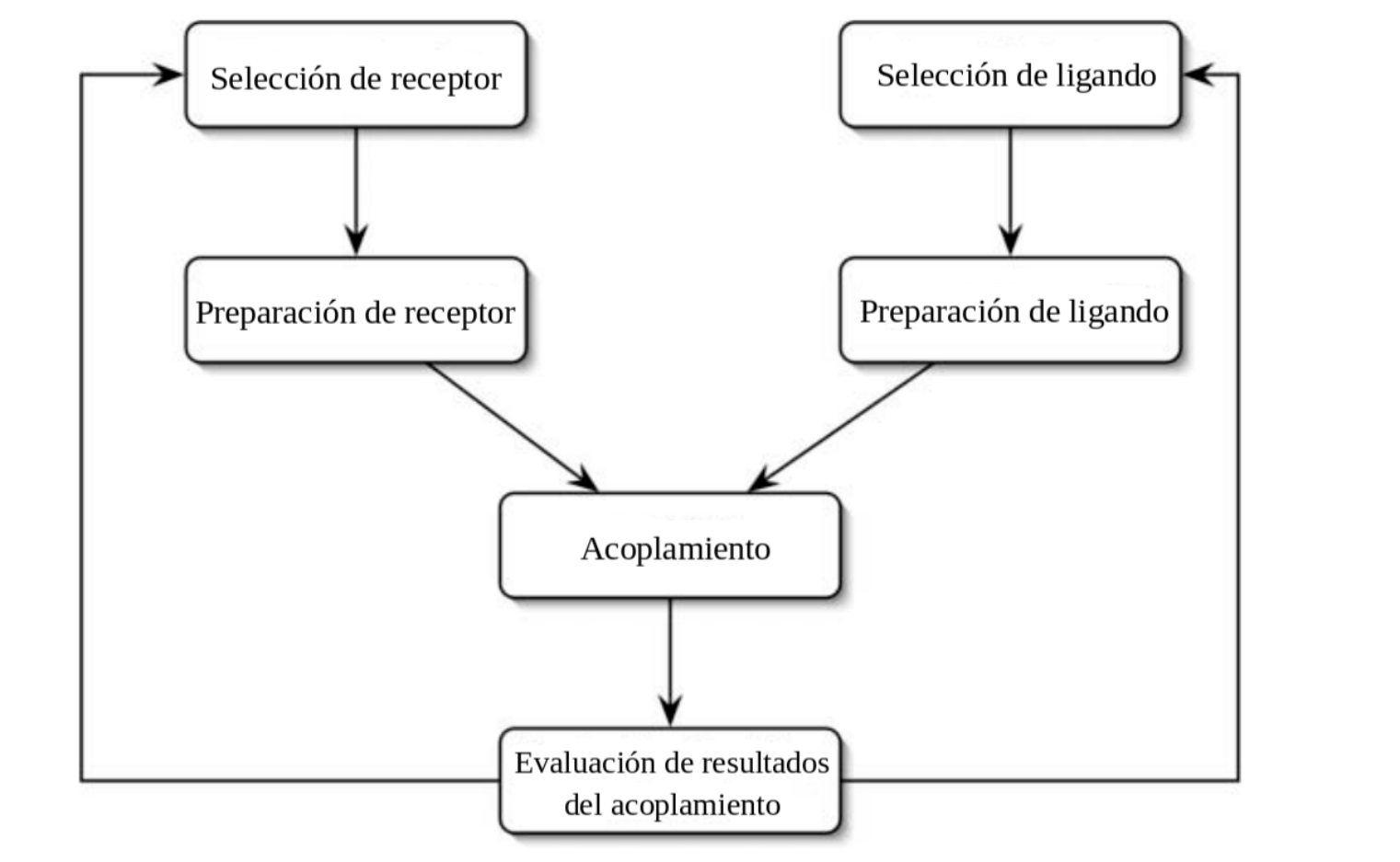
Docking

Acoplamiento molecular

Método cuyo objetivo es predecir los estados tanto estructurales, llamadas poses, como energéticos, prediciendo la afinidad del enlace entre moléculas.

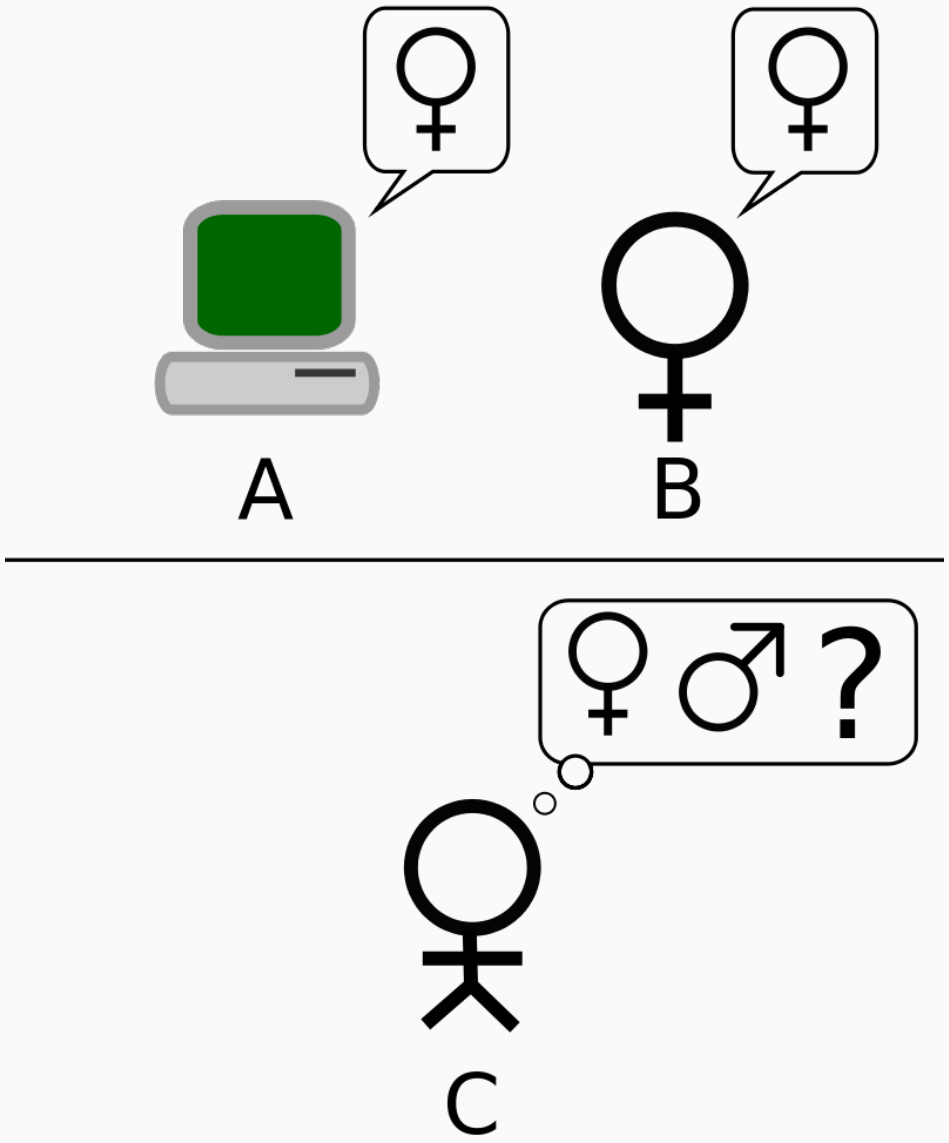


Pasos del docking



Sobre inteligencia artificial

La prueba de Turing



IA y agentes

Inteligencia artrificial

Agentes racionales que, mediante sensores, pueden percibir su entorno y actuar sobre él a partir de un sistema de decisión.

IA y agentes

Inteligencia artrificial

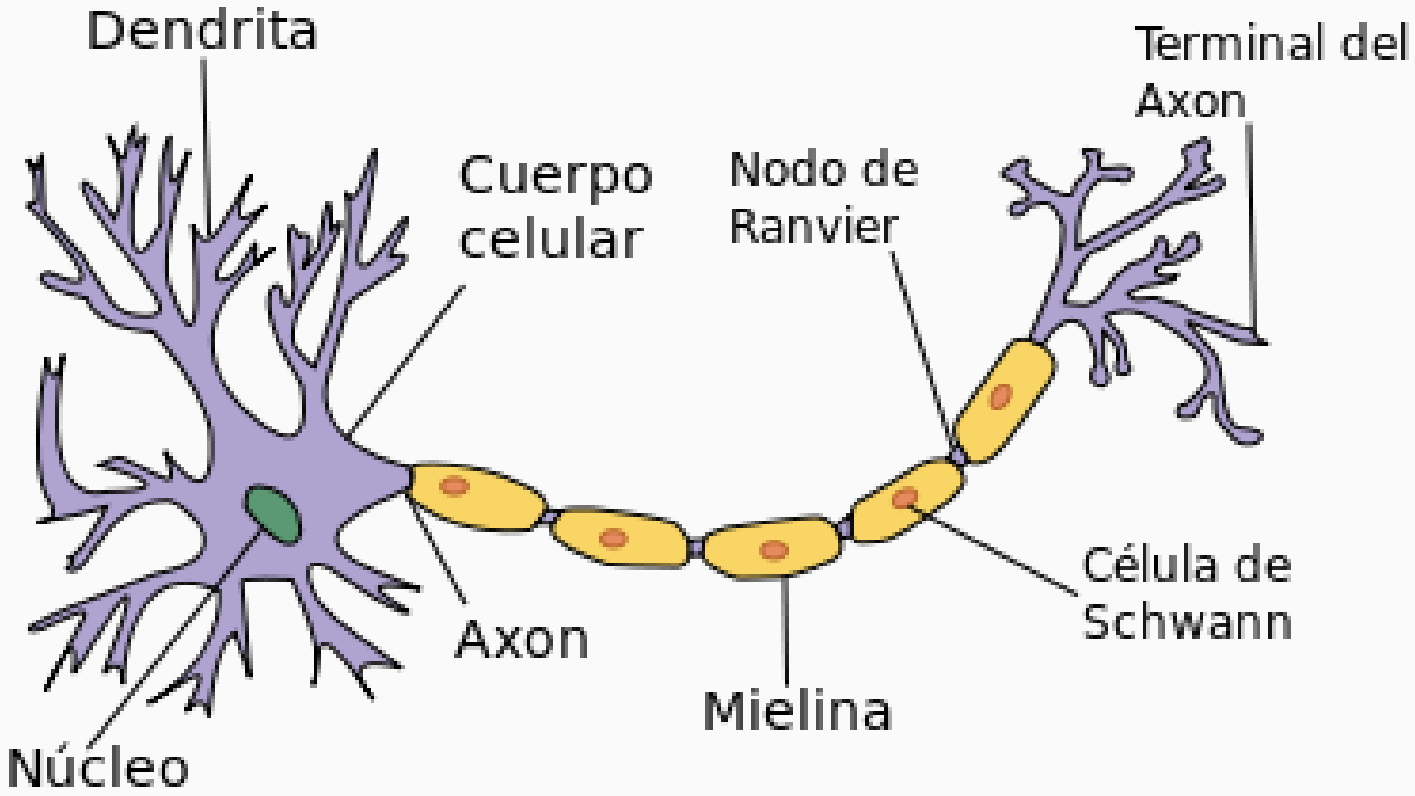
Agentes racionales que, mediante sensores, pueden percibir su entorno y actuar sobre él a partir de un sistema de decisión.

Agentes

Máquina compuesta por un conjunto finito de estados, cuyas transiciones están dadas por reglas de inferencias.

Sobre redes y neuronas

Inspiración en la biología



El perceptrón

Definiciónes

- $x \in \mathbb{R}^n$ (muestra)
- $w \in \mathbb{R}^n$ (vector de pesos)
- $\theta \in \mathbb{R}^n$ (umbral de activación)
- $y \in \{0, 1\}$ (valor real de la muestra)
- $\hat{y} \in \{0, 1\}$ (valor predicho de la muestra)

Por último, definimos z como una combinación líneal de x y w

$z = w_1 x_1 + ... w_n x_n$

Llamamos a z la *entrada de la red*.

El perceptrón

Definiciónes

- $x \in \mathbb{R}^n$ (muestra)
- $w \in \mathbb{R}^n$ (vector de pesos)
- $\theta \in \mathbb{R}^n$ (umbral de activación)
- $y \in \{0, 1\}$ (valor real de la muestra)
- $\hat{y} \in \{0, 1\}$ (valor predicho de la muestra)

Por último, definimos z como una combinación líneal de x y w

$z = w_1 x_1 + ... w_n x_n$

Llamamos a z la *entrada de la red*.

Función de activación

Definimos

$$\phi(z) = \begin{cases} 1 & \text{si } z \geq \theta \\ -1 & \text{en otro caso} \end{cases}$$

Pasos del perceptrón

1. Inicializar los pesos en cero o en números aleatorios cercanos a cero.

Pasos del perceptrón

1. Inicializar los pesos en cero o en números aleatorios cercanos a cero.

1. Para cada muestra de entrenamiento x, realizar lo siguiente:

Pasos del perceptrón

1. Inicializar los pesos en cero o en números aleatorios cercanos a cero.

1. Para cada muestra de entrenamiento x, realizar lo siguiente:

a) Calcular el valor de salida $\hat{y}$ ($\hat{y} = \phi(z)$).

Pasos del perceptrón

1. Inicializar los pesos en cero o en números aleatorios cercanos a cero.

1. Para cada muestra de entrenamiento x, realizar lo siguiente:

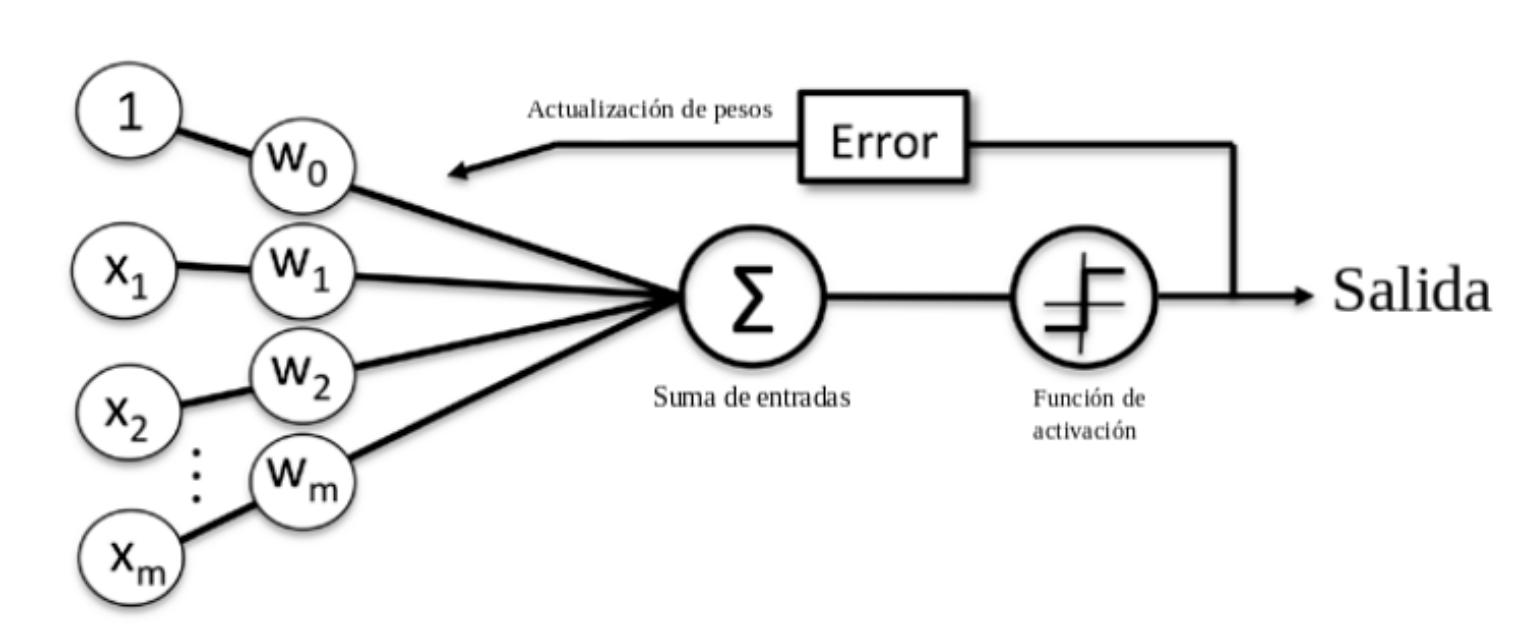
a) Calcular el valor de salida $\hat{y}$ ($\hat{y} = \phi(z)$).
b) Actualizar los pesos en w a partir del error Δw.

Con Δw dado por:

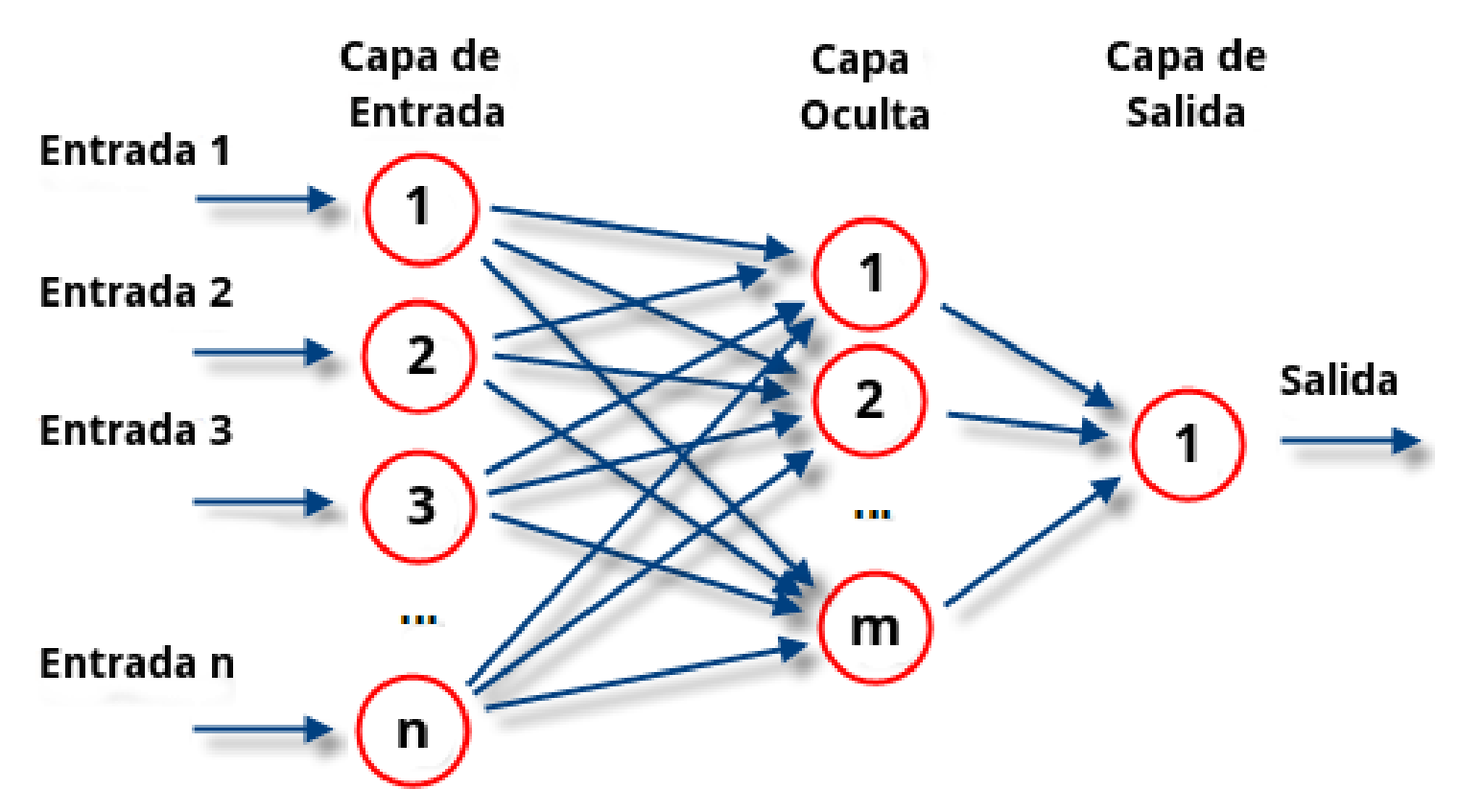
$$\Delta w = \eta(y - \hat{y})x$$

Donde $\eta \in [0, 1]$ es el *índice de aprendizaje*.

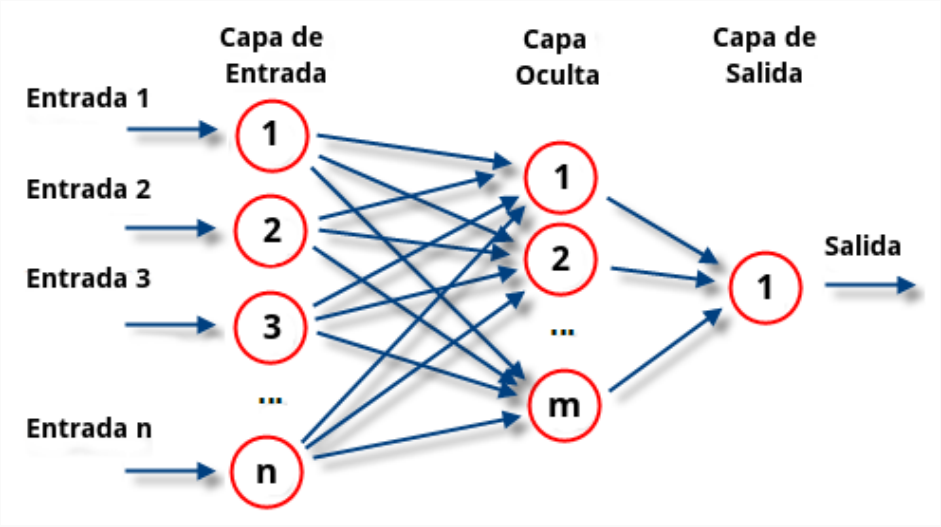
Diagrama del perceptrón



El perceptrón multicapa



El perceptrón multicapa

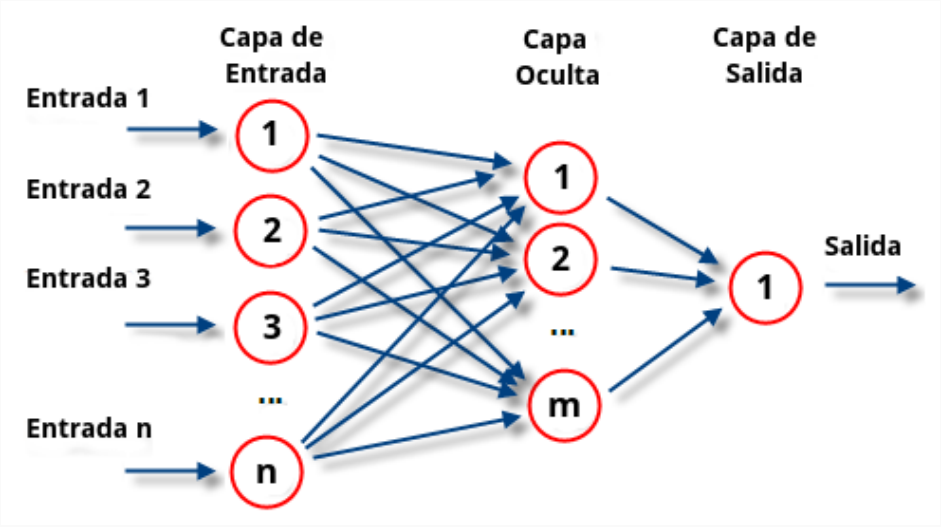


Función de costo o error

Definimos la función de costo J para el perceptrón multicapa como la suma de los errores cuadrados entre la salida calculada y el valor real:

$$J(w) = 1/2n \sum_{i=1}^{n} (\hat{y}_i - y_i^2)$$

El perceptrón multicapa



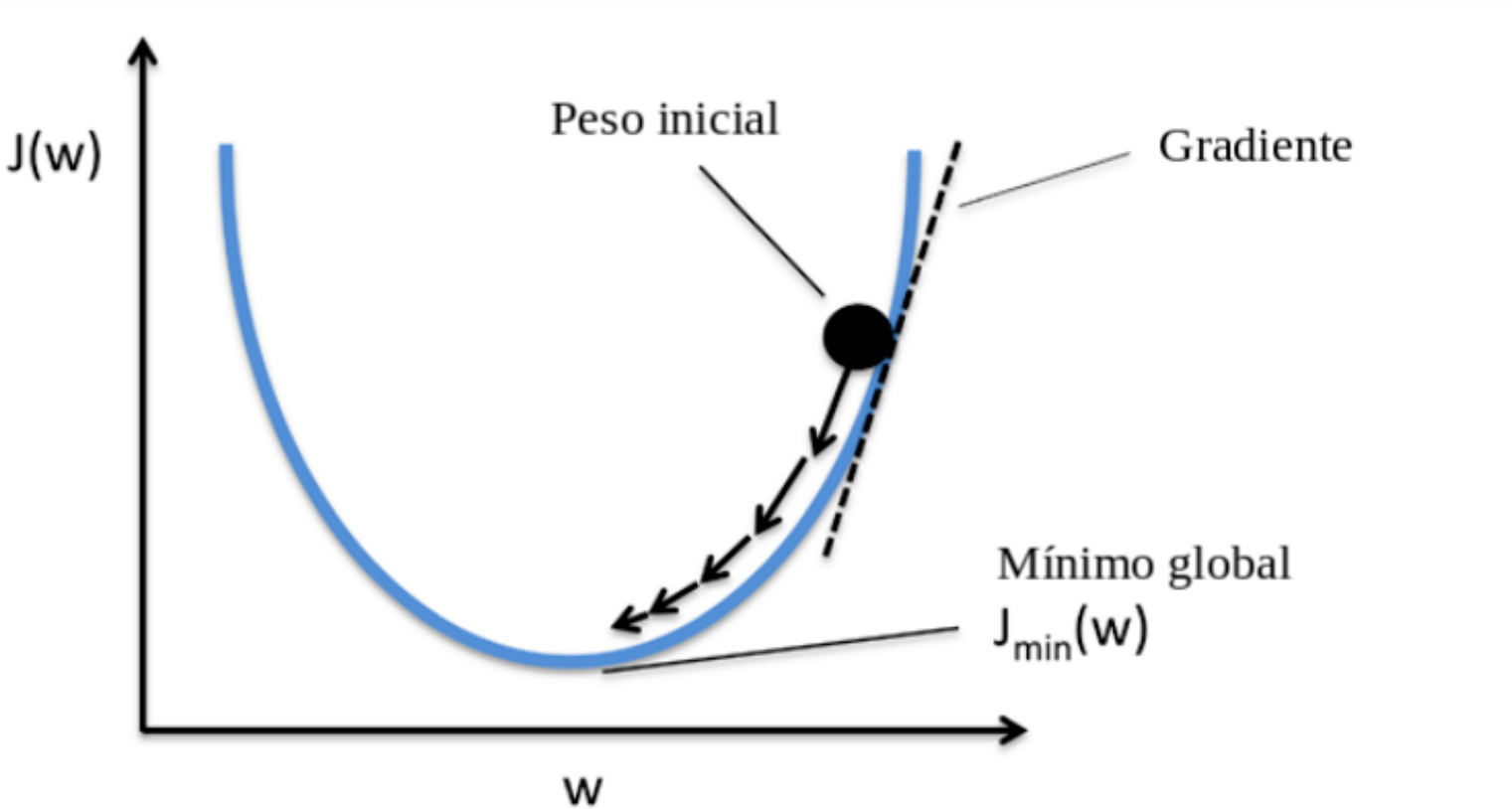
Función de costo o error

Definimos la función de costo J para el perceptrón multicapa como la suma de los errores cuadrados entre la salida calculada y el valor real:

$$J(w) = 1/2n \sum_{i=1}^{n} (\hat{y}_i - y_i^2)$$

¡Es diferenciable!

Descenso por el gradiente



Sobre meta-análisis del acoplamiento

Preparación de la base de datos

Preparación de la base de datos

1. Se filtran las proteínas que no contengan ligandos.

Preparación de la base de datos

1. Se filtran las proteínas que no contengan ligandos.
2. Se eliminan todas las proteínas con peso molecular menor a 300 Da.

Preparación de la base de datos

1. Se filtran las proteínas que no contengan ligandos.
2. Se eliminan todas las proteínas con peso molecular menor a 300 Da.
3. Se filtran de las proteínas las cadenas de ADN y ARN.

Preparación de la base de datos

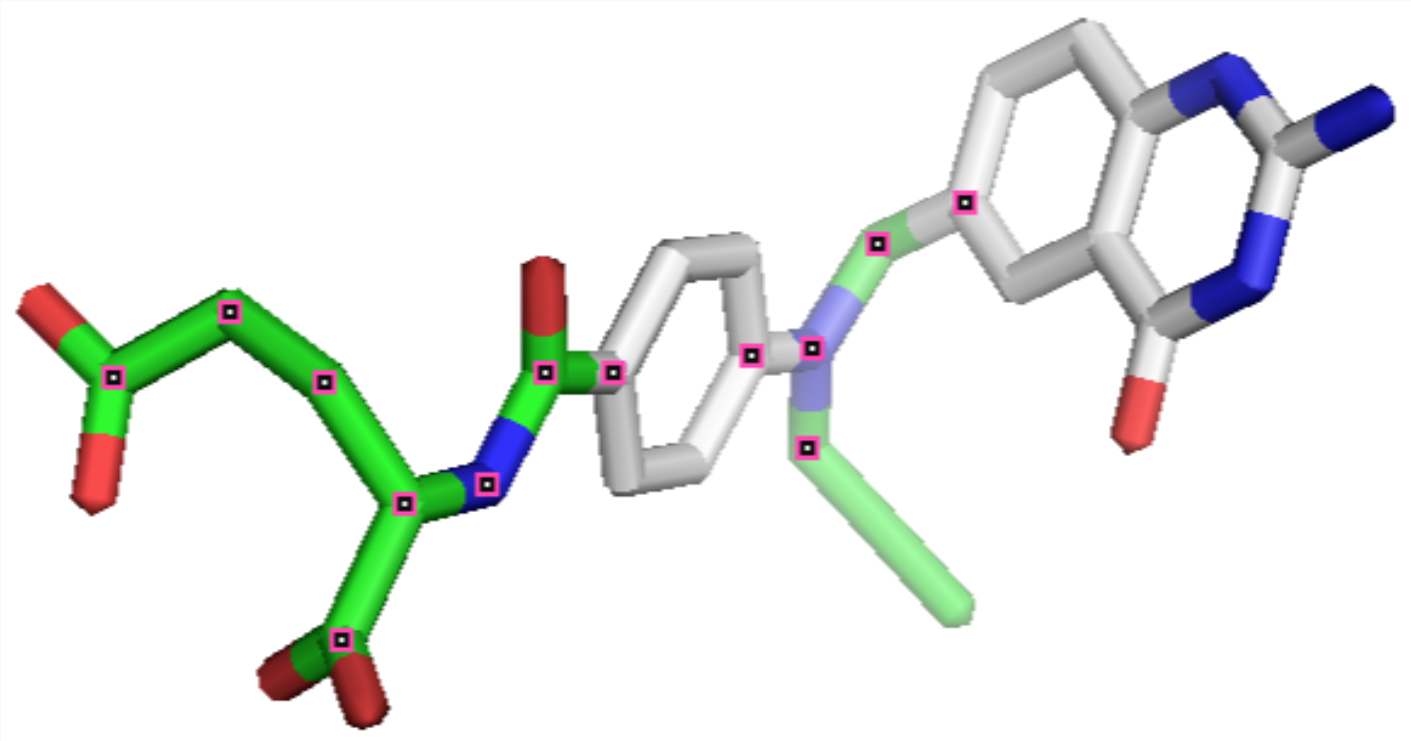
1. Se filtran las proteínas que no contengan ligandos.
2. Se eliminan todas las proteínas con peso molecular menor a 300 Da.
3. Se filtran de las proteínas las cadenas de ADN y ARN.
4. Se quitan también los metales pesados de las proteínas.

Preparación de la base de datos

5. Se definen las cargas de la proteína, así como sus libertades de torsión (ramas).

Preparación de la base de datos

5. Se definen las cargas de la proteína, así como sus libertades de torsión (ramas).



Preparación de la base de datos

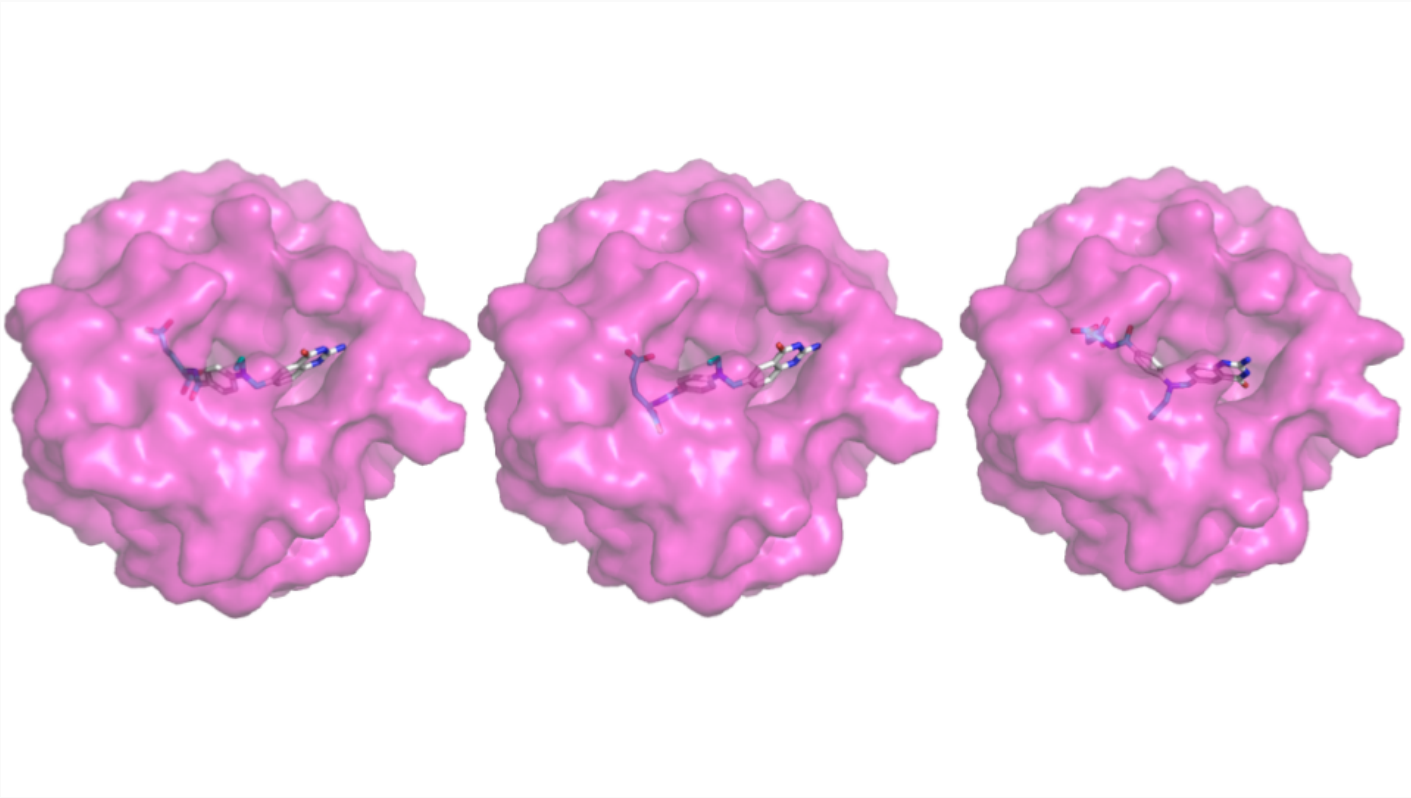
6. Se separa cada par proteína-ligando.

Preparación de la base de datos

6. Se separa cada par proteína-ligando.
7. Se hace el acoplamiento virtual de cada par.

Preparación de la base de datos

6. Se separa cada par proteína-ligando.
7. Se hace el acoplamiento virtual de cada par.



Preparación de la base de datos

Se genera entonces un listado de poses con una calificación asociada, que se compara con el RMSD del compuesto cristalográfico original.

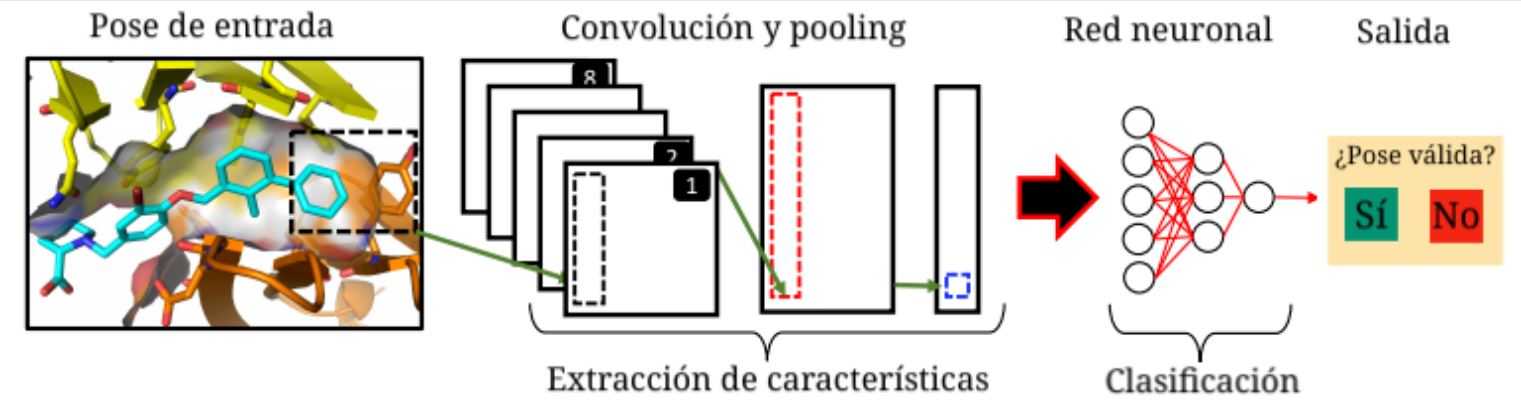
Preparación de la base de datos

Se genera entonces un listado de poses con una calificación asociada, que se compara con el RMSD del compuesto cristalográfico original.

Pose	Clasificación (según AutoDock Vina)	Calificación	RMSD
4EIL_CB3_A_1	1	-10.2	3.08
4EIL_CB3_A_2	2	-10.0	3.02
4EIL_CB3_A_3	3	-9.8	3.02
4EIL_CB3_A_4	4	-9.5	1.31
4EIL_CB3_A_5	5	-9.3	3.0

Deep-pose

Una red neuronal convolucional profunda que toma la información de un acomplamiento en un complejo proteína-ligando como entrada y produce una calificación de qué tan viable es dicha pose.



Deep-pose

Contexto de la rama

Contexto de la rama

Codificación del contexto de la rama

SMILES (*Simple Molecular Input Line Entry System*) es un sencillo lenguaje químico que permite describir moléculas utilizando únicamente caracteres ASCII.

Contexto de la rama

Codificación del contexto de la rama

SMILES (*Simple Molecular Input Line Entry System*) es un sencillo lenguaje químico que permite describir moléculas utilizando únicamente caracteres ASCII.

- Es sumamente compacta.

Contexto de la rama

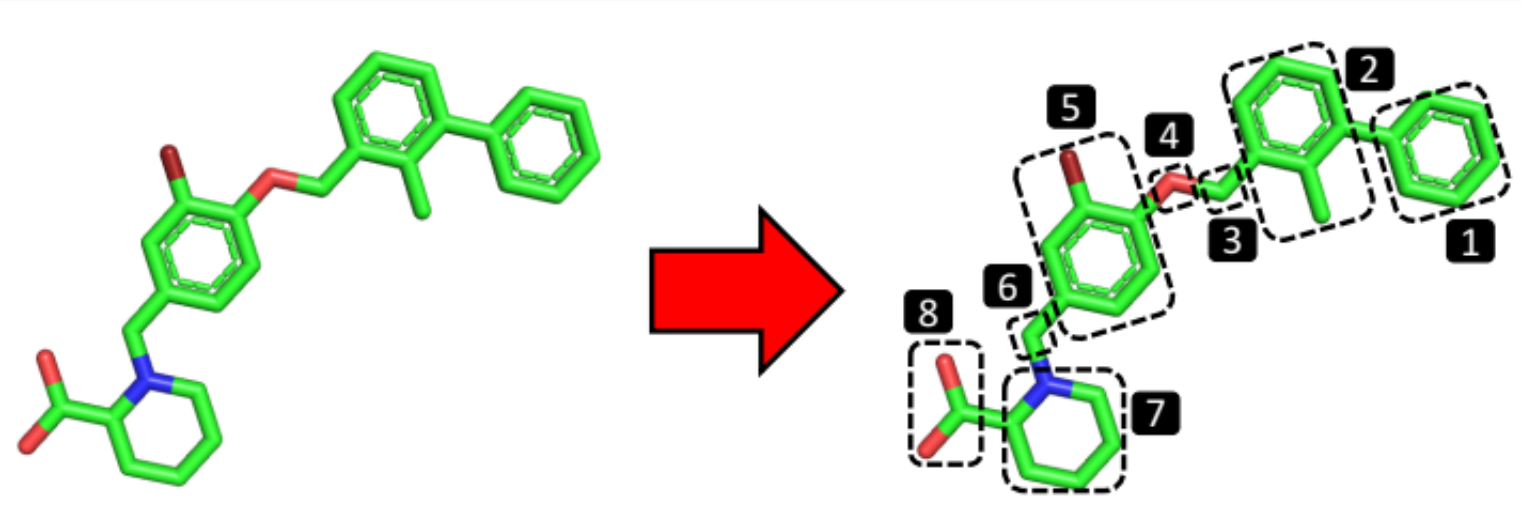
Codificación del contexto de la rama

SMILES (*Simple Molecular Input Line Entry System*) es un sencillo lenguaje químico que permite describir moléculas utilizando únicamente caracteres ASCII.

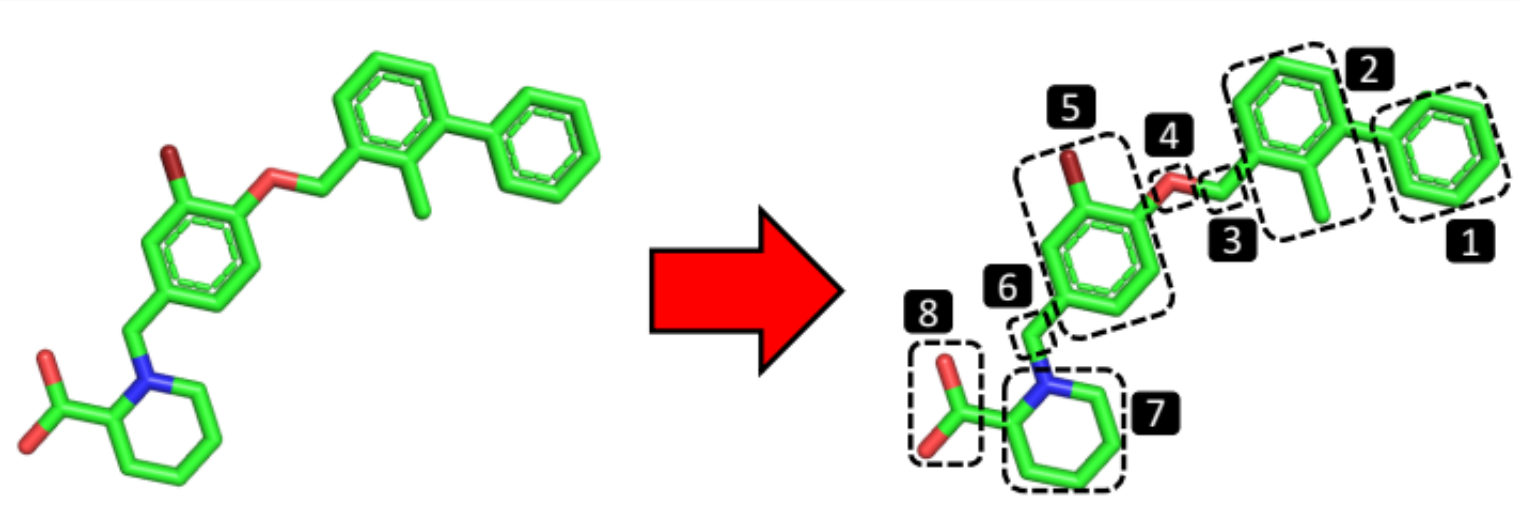
- Es sumamente compacta.
- Es canónica.

Diccionarios de ramas

Diccionarios de ramas



Diccionarios de ramas



Diccionarios de ramas

- Se enlistan todas las ramas codificadas distintas y a cada una se le asigna un índice (diccionario de ramas).
- Se segmentan los rangos de distancias entre ramas encontrados en compartimentos, y a cada uno de estos se le asigna también un índice (diccionario de distancias).

Fragmento de los diccionarios de ramas y de distancias

SMILES	Idx	Rango de distancia (Å)	Idx
NC1=N[C](=NC=C1)=O	93	3.0526 - 3.2631	6
C1CCCCC1	94	3.2632 - 3.4736	7
CNC=O	95	3.4737 - 3.6842	8
NC=N	96	3.6843 - 3.8947	9
CC=C	97	3.8948 - 4.1052	10

Codificación del contexto de la rama

Para cada rama del ligando, se codifican entonces las cinco ramas más cercanas del receptor a través de sus tipos y sus distancias.

Codificación del contexto de la rama

Para cada rama del ligando, se codifican entonces las cinco ramas más cercanas del receptor a través de sus tipos y sus distancias.

Traducción de la rama OP(O)O en una tupla.

Codificación del contexto de la rama

Para cada rama del ligando, se codifican entonces las cinco ramas más cercanas del receptor a través de sus tipos y sus distancias.

Traducción de la rama OP(O)O en una tupla.

Ramas cercanas a OP(O)O	Distancia en Å
N	5.794664
C1CC1	5.691862
NC1=N[C](=NC=C1)=O	4.449922
NC=N	3.785496
O	3.747894

Codificación del contexto de la rama

Para cada rama del ligando, se codifican entonces las cinco ramas más cercanas del receptor a través de sus tipos y sus distancias.

Traducción de la rama OP(O)O en una tupla.

Ramas cercanas a OP(O)O	Distancia en Å
N	5.794664
C1CC1	5.691862
NC1=N[C](=NC=C1)=O	4.449922
NC=N	3.785496
O	3.747894

$$\downarrow$$

$$OP(O)O = \Big[(2, 13, 93, 96, 4) \quad (11, 10, 7, 4, 4)\Big]$$

Representación vectorial del contexto de la rama

Definiciones

- B El conjunto de tipos de ramas.
- N Dimensión de los vectores característicos (hiperparámetro).
- $W^{b\_type} \in \mathbb{R}^{N \times |B|}$.

Representación vectorial del contexto de la rama

$$Rama = \Big[OP = O, \quad OC = O, \quad C = O, \quad OPO, \quad OP = O\Big]$$

W_{b_type}

OP=O	OC=O	OPO	C=O	NC=O	OPO	OP=O

$\downarrow$

$z^T_{b\_type} =$ OP=O | OC=O | C=O | OPO

Representación vectorial del contexto de la rama

Analogamente se genera el vector z_{b_dist}.

Representación vectorial del contexto de la rama

Analogamente se genera el vector z_{b_dist}.

Representación del contexto de la rama

Finalmente, la representación del contexto de la rama b se define como

$$z_b = z_{b\_type} \bullet z_{b\_dist}$$

Representación de la pose de un complejo proteina-ligando

La entrada de la capa convolucional es una lista de vectores $\{z_1, z_2, ..., z_n\}$ donde z_i es la representación vectorial del contexto de la i-ésima rama del ligando.

Representación de la pose de un complejo proteina-ligando

La entrada de la capa convolucional es una lista de vectores $\{z_1, z_2, ..., z_n\}$ donde z_i es la representación vectorial del contexto de la i-ésima rama del ligando.

Primera etapa de la capa convolucional (extracción de características)

$$u_i = f(z_i + b^{conv})$$

donde:

- f es la función tangente hiperbólica.
- b^{conv} es el sesgo.

Representación de la pose de un complejo proteina-ligando

La entrada de la capa convolucional es una lista de vectores $\{z_1, z_2, ..., z_n\}$ donde z_i es la representación vectorial del contexto de la i-ésima rama del ligando.

Primera etapa de la capa convolucional (extracción de características)

$$u_i = f(z_i + b^{conv})$$

donde:

- f es la función tangente hiperbólica.
- b^{conv} es el sesgo.

Segunda etapa de la capa convolucional *(max-pooling)*

$$[r]_j = \max_{1 \leq i \leq n} [u_i]_j$$

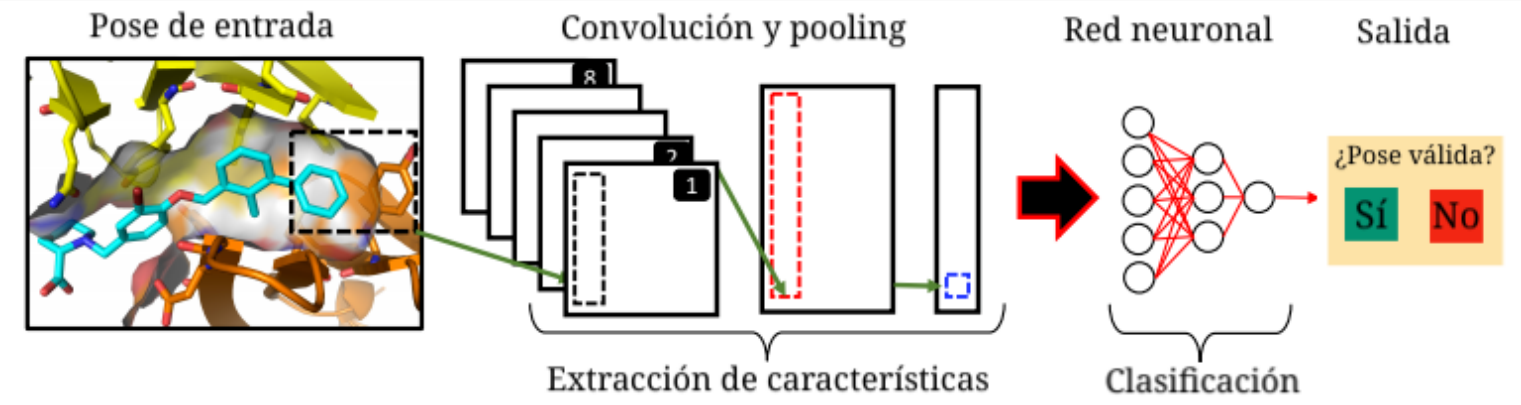
Clasificación de la pose

Finalmente el vector r es procesado por dos capas neuronales más:

1. Una tercera capa oculta que reprsenta un nivel más de abstracción.
2. Una última capa de salida donde se dá la clasificación.

Es en esta última capa donde se computa una calificación para cada una de las posibles clasificaciónes de la pose: (0) pose señuelo y (1) pose válida.

Arquitectura de la red

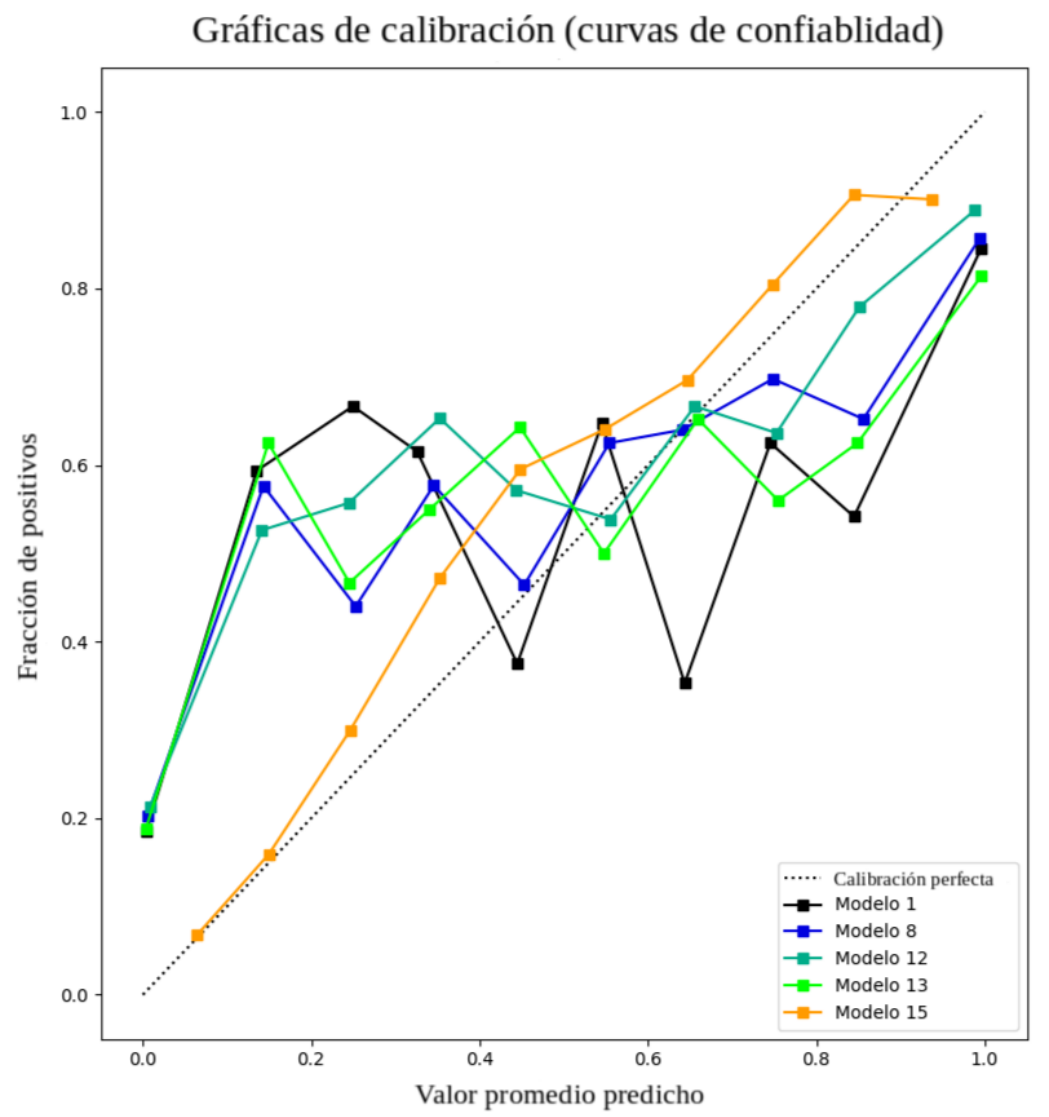


Entrenamiento y resultados

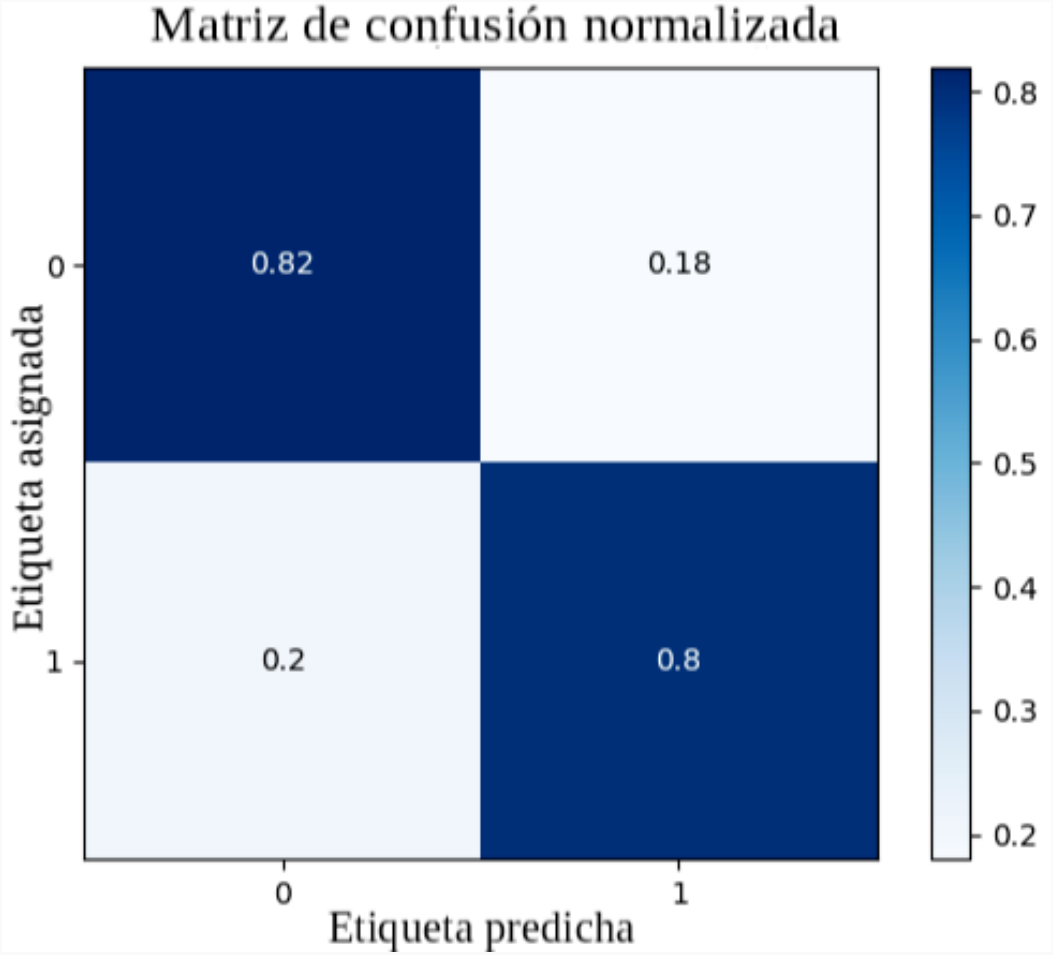
Hiperparámetros

Hiperparámetro	Descripción	Valor
N	Dimensión del vector característico	80
cf	Unidades en la capa convolucional	150
h	Unidades en la capa oculta	60
bs	*Tamaño de los minilotes*	20
λ	Índice de aprendizaje	0.1

Calibración de hiperparámetros



Resultados


Matriz de confusión normalizada
0.82
0.18
0.2
0.8
0
1
Etiqueta asignada
0
1
Etiqueta predicha
0.8
0.7
0.6
0.5
0.4
0.3
0.2

¿Preguntas?