# Protocolo seguido para generar el modelo de CY2C19

## Endpoint

Sustratos o no sustratos para CYP2C19

## Origen de los datos

Los datos vienen de diversas bases de datos:

* **CYP2C19\_CYProduct**
* **CYP2C19\_CYPstrate**
* **CYP2C19\_Deep-PK**
* **CYP2C19\_Magna**
* **CYP2C19\_Terri**

## Tratamiento de los datos

* **CYP2C19\_CYProduct:**

Los datos para crear los modelos de sustratos de las enzimas CYP se han obtenido a partir del siguiente artículo: https://doi.org/10.1021/acs.jcim.1c00144

Se ha descargado la base de datos de EBoMD.sdf y EBoMD2.sdf, compuesta por dos archivos en formato sdf y que contenía las columnas: Molecule Name, Name, InChiKey, PubChem, CYP (1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1 y 3A4), References y Comments. Se utilizó el script "sdf\_to\_csv.py" para añadir la columna con los SMILES y transformar el archivo sdf a csv. A continuación se modificaron las columnas de las CYP, mediant el script "Zaretzki\_dataset\_modf", para que apareciese 1 o 0, en función de si aparecía alguna información o no en la celda de cada CYP. Finalmente, mediante un script (CYPstrate\_DB\_processing), he creado las bases de datos de sustratos/no-sustratos para cada enzima y he añadido los compuestos de la base de datos de Hunt a cada una (ya que los sustratos para cada enzima están separados en hojas de Excel).

* **CYP2C19\_CYPstrate**

Los datos para crear los modelos de sustratos de las enzimas CYP se han obtenido a partir del siguiente artículo: 10.3390/molecules26154678

Para el desarrollo del modelo, se compiló un "conjunto de datos principal" de sustratos y no sustratos de nueve isoenzimas humanas del CYP (1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1 y 3A4) a partir de los trabajos de Hunt et al. y Tian et al. El conjunto de datos de Hunt es una colección bien curada de un total de 484 sustratos conocidos de siete de los nueve CYP investigados (excluyendo CYPs 2A6 y 2B6). El conjunto de datos de Tian contiene un total de 1800 sustratos y no sustratos conocidos para los mismos nueve CYPs. Los sustratos incluidos en este conjunto de datos provienen, en gran parte, del conjunto de datos de sitios de metabolismo de Zaretzki et al. Dado que el conjunto de datos de Zaretzki ha sido revisado desde entonces, la parte de los datos que proviene del conjunto original de Zaretzki fue reemplazada por los datos del conjunto de datos revisado de Zaretzki para los fines de este estudio.

En primer lugar se ha descargado la base de datos de Tian, compuesta por dos archivos (Train y Test) en formato sdf y que contenía las columnas: Molecule Name, Name, InChiKey, CYP (1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1 y 3A4), References y Comments. Se utilizó el script "sdf\_to\_csv.py" para añadir la columna con los SMILES y transformar el archivo sdf a csv. A continuación se eliminaron todos aquellos compuestos marcados con una R, que hacían referencia a los sustratos.

En segundo lugar se descargó la base de datos de sustratos actualizada de Zaretzki. Se utilizó el script "sdf\_to\_csv.py" para añadir la columna con los SMILES y transformar el archivo sdf a csv. A continuación se modificaron las columnas de las CYP, mediant el script "Zaretzki\_dataset\_modf", para que apareciese 1 o 0, en función de si aparecía algún numero o no en la celda de cada CYP

En tercer lugar se creó una hoja de excel llamada CYP\_generalDB.csv. En este archivo se juntaron manualmente la información de las tres bases de datos.

Finalmente, mediante un script (CYPstrate\_DB\_processing), he creado las bases de datos de sustratos/no-sustratos para cada enzima y he añadido los compuestos de la base de datos de Hunt a cada una (ya que los sustratos para cada enzima están separados en hojas de Excel).

* **CYP2C19\_Deep-PK**

Se obtienen un total de tres archivos csv, separados en test, train y validation.

* **CYP2C19\_Metabase**

Se descargaron los datos disponibles en formato .txt para las CYP (1A2, 2C9, 2C19, 2D6, 3A4, 2C8, 2E1) del repositorio de Metrabase, en el siguiente link: http://www-metrabase.ch.cam.ac.uk.

El archivo de texto fue procesado para generar un archivo csv que guardase las columnas cmpd\_id cmpd\_name, y action\_type. Mediante un script en python se transformo la información de action\_type (inhibitor, substrate, non-substrate) en números (-1, 0, 1). Además se obtuvieron los smiles también por el script. Aquellos smiles que no fueron obtenidos por el script se comprobaron manualmente.

* **CYP2C19\_Terri**

Los datos para crear los modelos de sustratos de las enzimas CYP se han obtenido a partir del siguiente artículo: https://s3-us-west-2.amazonaws.com/drugbank/cite\_this/attachments/files/000/001/582/original/cyp450\_drug\_interactions.pdf?1537393944

El nombre de los compuestos y sus smiles aparecen en el propio artículo, se creó un excel con el nombre del compuesto, el smiles y se etiquetaron todos como 1 al ser sustratos.

A continuación, se eliminaron los Outliers de todas las bases de datos, mediante el script “Elbow\_Undersampling.py”. Las bases de datos finales se concatenaron mediante un script de Python y se procesaron mediante HYGIEIA.

Posteriormente se llevó a cabo un Undersampling con el script “Elbow\_Undersampling.py”.

## Transformación de la “y”

No hay transformación

## Train/test ratio

Se ha realizado una partición del 75% para el train, 10% para el test y 15% para el validation. Durante el proceso de partición se mantienen todos los compuestos señalados como *quite dissimilar*..

## Scaler

Se ha usado el *Standard Scaler*