## **bioavailability20 & 30**

* El paper de donde vienen los datos es correcto
* **Description\_extra2:** Pequeño parrafito incluido
* **Experimental protocol:** En el artículo de referencia no hay ninguna información del proceso experimental, solo que los datos vienen, principalmente, de un libro de referencia. Así que por no dejar vacío el campo he puesto una pequeña descripción genérica de cómo se calcula.
* **Dependent\_variable\_class:** Pequeña descripción de la transformación de los datos incluida.
* **Criteria:** Pequeña explicación de que significa una predicción positiva o negativa.

El paper de donde vienen los datos es correcto. El número de compuestos en la base de datos original es de 995, que se reducen a 988 seguramente por los métodos en el curado y selección de los datos. Los datos en el paper original se dividen en a partir de una biodisponibilidad de 50, por lo que existen una menor cantidad de valores con %F positivo = 1 (509). Esto es debido al límite establecido, ya que para %F de 20 y 30 estamos cogiendo un límite menor al del artículo original, por lo que el número de compuestos positivos en nuestra base de datos es mayor (733 para el caso de %F20 y 652 en %F30).

El número de compuestos que coinciden entre la base de datos original y la empleada para el modelo es de 988

## **caco-2\_permeability**

* El paper de donde vienen los datos es correcto. El dataset usado tiene algunos compuestos eliminados, supongo que outliers.
* **Description:** Pequeña ampliación del párrafo.
* **Experimental protocol:** En el artículo de referencia no hay ninguna información del proceso experimental, solo de donde han sacado los datos que han curado. Así que por no dejar vacío el campo he puesto una pequeña descripción genérica de cómo se calcula.

El paper de donde vienen los datos es correcto. En el QMRF indica que hay un total de 887 compuestos empleados en la base de datos, no obstante, el archivo .csv empleado tiene un total de 895 compuestos. Todos los compuestos se encuentran en el rango establecido (entre -7.8 and -3.5), por lo que habría que saber si esos compuestos se eliminaron y no está indicado o hay un error al poner la información en el QMRF.

El número de compuestos en la base de datos original es de 1272, que se reducen a 887 seguramente por los métodos en el curado y selección de los datos. Todos los compuestos en la base de datos original se encuentran en el rango establecido (entre -7.8 and -3.5)

El número de compuestos que coinciden entre la base de datos original y la empleada para el modelo es de 851

## **p-gp\_inhibitor**

* El paper de donde vienen los datos es correcto
* **Experimental protocol:** Pequeño párrafo proveniente del artículo incluido.
* **Dependent\_variable\_reg:** No hay transformación porque los datos los obtenemos ya transformados.
* **Criteria:** Pequeña explicación de que significa una predicción positiva o negativa.

El paper de donde vienen los datos es correcto. El número de compuestos en la base de datos original es de 1275, que se reducen a 1216 seguramente por los métodos en el curado y selección de los datos. Los datos en el QMRF coinciden con la base de datos.

## **p-gp\_ substrate**

* El paper de donde vienen los datos parece que no se corresponde. El artículo suministra ~350 compuestos, y en el dataset usado como input (después de HYGEIA) figuran 818 compuestos
* **Experimental protocol:** Pequeño párrafo proveniente del json que rellenábamos antes para dar la información de los modelos.
* **Criteria:** Pequeña frase sobre que significa una predicción positiva.

No he encontrado el dataset original. En los datos suplementarios hay información sobre los descriptores empelados y los datasets utilizados para el train y el test. No obstante el número de compuestos en ellos es muy pequeño (+-350) en comparación con los datos utilizados en comparación al utilizado 818. . En el QMRF indica que hay un total de 818 compuestos empleados en la base de datos, no obstante, el archivo .csv empleado tiene un total de 817 compuestos. Lo mismo ocurre con el número de compuestos positivos, en el QMRF pone 418 mientras que en el dataset hay 419.

## **skin\_permeability**

* El paper de donde vienen los datos parece correcto (Cuando se rehaga el QMRF hay que mirar cantidad de compuestos, ya que no cuadra del todo de momento)
* En “species” pone “Human”, pero en el artículo de referencia no aparece la palabra *vivo* en ningún momento, pero sí que aparece *vitro* en dos ocasiones (Una de ellas en el título de uno de los múltiples papers de los que sacaron los datos que curaron para su base de datos propia)
* **Experimental protocol:** En el artículo de referencia no hay ninguna información del proceso experimental, solo de donde vienen los datos vienen. Así que por no dejar vacío el campo he puesto una pequeña descripción de cómo se calcula (sacada de un OECD Guideline, pero intentando generalizar un poco comparando el proceso *in vitro* con el *in vivo*).

El paper de donde vienen los datos es correcto. El número de compuestos en la base de datos original es de 283, que se reducen a 185 seguramente por los métodos en el curado y selección de los datos. En el QMRF indica que hay un total de 267 compuestos empleados en la base de datos, no obstante, el archivo .csv empleado tiene un total de 185 compuestos. Lo mismo ocurre con el rango de distribución de valores ya que en el QMRF pone que el rango es de -6.1 y -0.8, mientras que en el dataset el rango va entre -5.52 y -0.69.

## **Human intestinal absorption**

* Revisar SMILES porque la cantidad no coincide.
* No encuentro información de como se obtiene las categorías. No está siquiera explicado en el paper de referencia.
* Tampoco encuentro información del protocolo experimental.
* **Description\_extra2:** Pequeño parrafito incluido.

El número de compuesto de la base de datos original es de 970, tras la limpieza de datos el archivo contiene un total de 943 compuestos. La comparación con el archivo utilizado para entrenar el modelo ha mostrado un total de 898 coincidencias. Estos compuestos son pocos en comparación con la base de datos del modelo, ya que esta se compone de un total de 1245. En el QMRF indica que hay un total de 1239 compuestos, mientras que en la base de datos hay 1245 .

## **blood-brain\_barrier**

* El paper de dónde vienen los datos parece correcto. Hay menos compuestos (habrán sido eliminados en Hygeia). He comprobado un par a mano y parece que coincide, pero no he podido comparar todo porque la Supporting Information del paper viene en formato .pdf y no en .csv/.xlsx
* **Description\_extra:** He pasado el contenido a “description extra 2” y he añadido aquí un poco más de información.
* **Criteria:** Pequeña explicación de que significa una predicción positiva o negativa.

Nada que añadir

## **plasma-protein\_binding**

* En el dataset usado como input hay más compuestos que juntando los 4 documentos de la Suplementary Information del artículo de referencia (1804 vs 1569 después de Hygeia/1727 antes de Hygeia). Además, los datos de origen están en %, mientras que los de ProtoADME en tanto por 1. Coinciden en ambos datasets unos 570 compuestos, pero no en todos el valor experimental es el mismo, aunque en la mayoría de los que no coinciden la diferencia es muy pequeña)
* **References:** Se ha incluido una referencia para el “experiemntal protocol” (sacada del paper de referencia)
* **Experimental protocol:** Se ha añadido un pequeño párrafo.

## **volume\_of\_distribution**

* El paper de donde se supone que vienen los datos cuenta con 120 compuestos de training set y 18 en el testing set. El juego de datos usado por nosotros cuenta con más de 2500 compuestos.
* **Experimental protocol:** Pequeño párrafo proveniente del artículo incluido.

Nada que añadir, parece que tiene otros artículos más recientes con bases de datos extensas, pero no aparece el SMILE de los compuestos, únicamente su nombre

## **CYP450 family inhibitor/substrate**

* En el artículo de donde se supone que vienen los datos solo he encontrado información sobre “substrate”, pero no sobre inhibidores. Además, en todos los casos tenemos más datos en el database de entrada para nuestros modelos de “substrate” que lo que aparece en el artículo. (**1A2**: 651 vs 271; **2C9:** 641 vs 226; **2C19:** 655 vs 218; **2D6:** 661 vs 270; **3A4:** 647 vs 475)
* **Experimental protocol:** Pequeño párrafo sacado del artículo incluido.
* **Criteria:** Pequeña explicación de que significa una predicción positiva o negativa (Usando lo incluido en p-gp inhibitor/substrate).

Nada que añadir

## **Half life**

* El paper de dónde vienen los datos parece correcto. La cantidad de datos es muy parecida, pero no podemos asegurar si los compuestos coinciden, ya que los compuestos en el artículo de referencia están dados en un archivo .pdf como imagen escaneada y en como nombre, no en SMILES. He comparado 4 compuestos y los 4 estaban en el dataset de entrada. 2 de ellos con el mismo valor, los otros dos ligeramente diferentes (-0.40 vs -0.35 y 1.08 vs 1.28).
* **Criteria:** Pequeña explicación de que significa un valor alto o bajo (explicación muy simple, principalmente por no dejar la sección vacía).

El número de compuesto de la base de datos original es de 1105, tras la limpieza de datos el archivo contiene un total de 1100 compuestos. La comparación con el archivo utilizado para entrenar el modelo ha mostrado un total de 259 coincidencias. Estos compuestos son pocos en comparación con la base de datos del modelo, ya que esta se compone de un total de 650.

En el QMRF indica que la base de datos tiene un total de 929 compuestos, pero en el conjunto de datos hay un total de 650

## **human\_liver\_microsomal**

* El artículo de referencia parece que es de dónde vienen los datos. Pero:
  + Si comparo SMILES en ambos lados se pierden unos 100-150 compuestos de los que se usaron en el modelo de ProtoADME
  + Hay 59 compuestos cuya categoría es “1” en el artículo, pero “0” en el dataset usado en ProtoADME
* **Experimental protocol:** Pequeño párrafo sacado del artículo incluido.
* **Criteria:** Pequeña explicación de que significa un valor bajo (Sacada del paper de referencia).

El número de compuesto de la base de datos original es de 3654, tras la limpieza de datos el archivo contiene un total de 3616 compuestos. La comparación con el archivo utilizado para entrenar el modelo ha mostrado un total de 2996 coincidencias. La comparación de los datos varía, ya que la base de datos empleada para entrenar el modelo tiene un total de 3199, esta diferencia se podría deber a la manera de tratar los datos

## **OATP1B1/OATP1B3**

* Se supone que en el artículo hablan de 225 compuestos estudiados. Nuestro modelo se ha hecho con un juego de datos que contiene más de 1100 compuestos.
* **Description\_extra2:** Pequeño parrafito incluido.
* **Experimental protocol:** Pequeño párrafo sacado del artículo incluido.
* **Criteria:** Pequeña explicación de que significa un positivo y/o negativo.

Nada que añadir

## **BSEP**

* En el artículo se habla de un dataset de 408 compuestos (aunque no veo que lo incluyan, sino que hacen referencia a otro artículo del cual proceden) y nuestro modelo está realizado con un juego de datos de 595 compuestos.
* **Experimental protocol:** Pequeño párrafo sacado del artículo incluido.
* **Criteria:** Pequeña explicación de que significa un positivo y/o negativo.

Nada que añadir