**MATERIALES Y MÉTODOS**:

**Esquema de trabajo**:

Imagen que contiene Forma

Descripción generada automáticamente

1. **Obtención de los datos**:

Para este estudio, se utilizó un fichero que contiene 71,494 compuestos, divididos en tres grupos: Drug (1.419), FoodFL (47.779) y FoodnoFL (22.296). El grupo Drug son drogas obtenidas del DrugBank, mientras que los grupos FoodFL (compuestos lipídicos) y FoodnoFL (compuestos no lipídicos) son compuestos bioactivos derivados de alimentos obtenidos de FooDB. Todas las moléculas han sido recopiladas y utilizadas en estudios previos (47, 40).

Asimismo, se ha utilizado otro fichero en el cual venían asignadas las clases químicas de los compuestos. Estas clases químicas vienen derivadas del algoritmo ClassyFire (43), que es un algoritmo y programa informático que asigna cada molécula a una taxonomía jerárquica basada en reglas estructurales computacionales claras. La taxonomía incluye hasta 11 niveles diferentes (Reino, Superclase, Subclase, etc.) y más de 4800 categorías (40).

Los ficheros anteriores fueron proporcionados por el tutor Gonzalo Colmenarejo.

Aparte de los ficheros mencionados, se utilizó otro que contenía dianas y clases de dianas farmacológicas asociadas a los compuestos, adquirido del material suplementario de uno de los artículos del grupo del tutor (47).

1. **Entorno y herramientas informáticas**:

Se utilizó Miniconda configurado en Windows 11 (10.0.22631-SP0) para gestión de paquetes y entornos virtuales. La versión de Miniconda utilizada es la 24.9.2.

Se usó el Entorno de Desarrollo Integrado (IDE) Spyder (versión 6.0.1) en Python 3.12.7.

Algunas de las bibliotecas utilizadas incluyen:

* Manipulación de datos: pandas y numpy.
* Visualización de datos: matplotlib y seaborn.
* Análisis estadístico: scipy.stats y statsmodels.
* Análisis químico: RDKit y chembl\_structure\_pipeline.
* Otras herramientas: IPython e itertools.

1. **Procesamiento de las estructuras moleculares**:

El hilo de procesamiento de datos que se sigue es principalmente a partir de las librerías RDKit (versión 2024.09.1) (69) y ChEMBL (versión 1.2.0) (30,50), ya que se utilizan para todos los análisis y preprocesamientos de estructuras moleculares.

En un primer momento, se transforma el Identificador Químico Internacional (InChI) al objeto de clase Mol, el cual contiene estructuras y distintos datos relativos a la molécula que se genera del compuesto; más tarde a partir de éstos, se extrae la estructura padre del compuesto (pMol), y una vez obtenidas todas las estructuras (moléculas) necesarias se estandarizan para garantizar la coherencia y la calidad de la representación molecular.

Además de estandarizar las moléculas, se eliminan las no válidas y se busca que no queden moléculas compuestas para evitar posibles errores en posteriores etapas del trabajo.

Una vez limpias y extraídas las moléculas necesarias se sigue con la generación de sus grupos funcionales.

1. **Generación de grupos funcionales**:

Para la generación de grupos funcionales (FGs) se ha usado una implementación completa y precisa del algoritmo de Ertl (48) en RDKit desarrollada por el grupo de Gonzalo Colmenarejo (1).

1. **Propiedades físicoquímicas**:
   1. Obtención propiedades:

Para la correlación de FGs con propiedades fisicoquímicas, primero se ha realizado la sanitización de las moléculas para evitar errores como ya se ha hecho en apartados anteriores, utilizando la función Chem.SanitizeMol() de RDKit, que permite verificar y corregir posibles problemas estructurales en cada molécula representada como un objeto Mol. A partir de esas moléculas se extrajeron los 15 FGs más frecuentes por compuesto y a partir de ahí, los FGs comunes entre los 3 tipos de compuestos.

Una vez obtenidos los FGs, se calculan las propiedades usando RDKit (46).

Las propiedades calculadas son:

* Área de superficie del polo topológico (TPSA).
* Coeficiente de partición octanol/agua (LogP).
* Número de juntas giratorias (Rotatable Bonds).
* Número de donantes de enlaces de hidrógeno (HBD).
* Número de aceptores de enlaces de hidrógeno (HBA).
* Peso molecular (MW).
* Número de anillos (Number of Rings) y número de anillos aromáticos (Aromatic Rings).
* Evaluación cuantitativa de la afinidad por estructuras similares a fármacos (QED).
* Fracción de carbono hibridada sp3 (Fsp3).

Se realizan gráficos boxplot/violinplot sin outliers para ver las distribuciones de las propiedades por cada FGs y compuesto.

* 1. Análisis estadístico de las propiedades:

Se utilizó la prueba no paramétrica de Kruskal-Wallis para analizar diferencias en las distribuciones de propiedades fisicoquímicas de los FGs entre grupos de compuestos. Una vez obtenidos los valores, se aplicó la corrección de FDR de Benjamini-Hochberg para ajustar los valores p y controlar la tasa de falsos descubrimientos. (48)

Cuando se encontró que las distribuciones corregidas eran significativamente diferentes (umbral de significancia de 0,05), se utilizó el análisis post-hoc de Conover con la corrección del valor p de Holm para identificar pares de grupos significativos. (48)

Finalmente, se calculó el tamaño del efecto en lenguaje común (CLES) entre pares para estimar la probabilidad de que un valor aleatorio en un grupo sea mayor que un valor aleatorio en el otro grupo. Valores de CLES mayores a 0.5 indican una distribución desplazada hacia valores más altos en el primer grupo, mientras que valores menores a 0.5 indican un desplazamiento hacia valores más bajos. (48, 40)

1. **Análisis de interacciones con dianas farmacológicas**:

Para el análisis de interacciones con dianas, se contó el número de interacciones únicas por tipo de compuesto y clase de diana, utilizando las moléculas que contenían cada grupo funcional específico. Se hicieron gráficos de lo anterior para visualizar por cada grupo de compuesto qué FGs (dentro de los comunes entre los 3 tipos de compuesto) están asociados a ciertas clases de dianas farmacológicas.