

# Material suplementario

## S1. Detalle del filtrado de datos para la extracción de datos de ChEMBL 34.

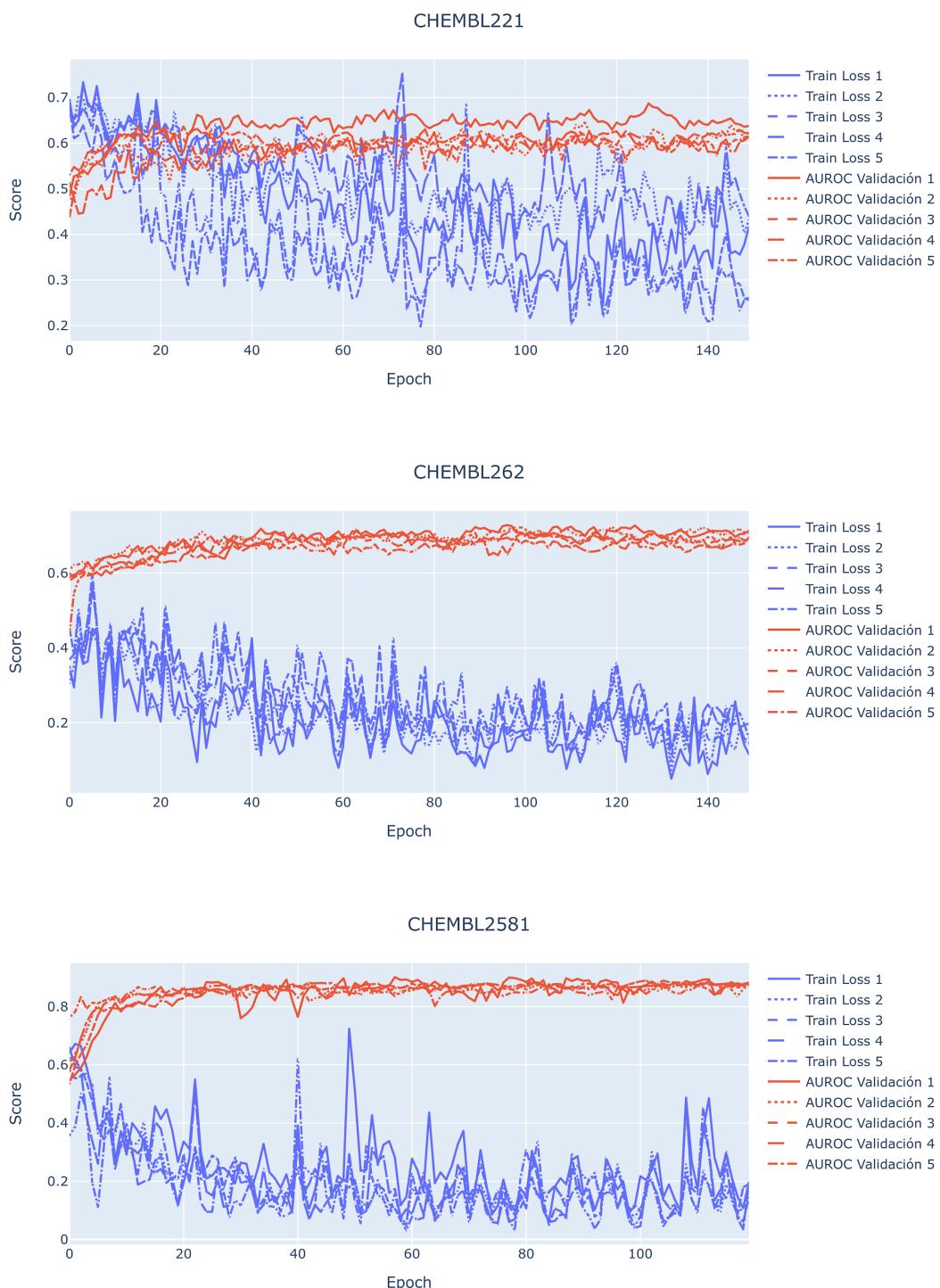
A continuación se presenta la *query* inicial utilizada para extraer los datos crudos de la base de datos de MySQL de CHEMBL.

```
SELECT act.molregno,
       md.chembl_id AS comp_id,
       ass.tid AS tid,
       ass.assay_id AS assay_id,
       trgd.chembl_id AS target,
       act.standard_relation AS relation,
       act.standard_value AS bioactivity,
       act.standard_units AS units,
       act.standard_type AS type,
       act.potential_duplicate,
       act.pchembl_value,
       trgd.organism,
       trgd.target_type,
       trgd.pref_name,
       ass.assay_type,
       cmpstc.canonical_smiles AS smiles,
       cmpstc.standard_inchi_key AS inchi_key,
       cmpprop.*,
       cs.sequence AS sequence,
       cs.organism AS sequence_organism
  FROM activities AS act
 JOIN assays AS ass ON act.assay_id = ass.assay_id
 JOIN target_dictionary AS trgd ON ass.tid = trgd.tid
 JOIN compound_structures AS cmpstc ON act.molregno = cmpstc.molregno
 JOIN molecule_dictionary AS md ON act.molregno = md.molregno
 JOIN compound_properties AS cmpprop ON md.molregno = cmpprop.molregno
 JOIN target_components AS tc ON ass.tid = tc.tid
 JOIN component_sequences AS cs ON tc.component_id = cs.component_id
 WHERE trgd.target_type='SINGLE PROTEIN'
   AND ass.assay_type='B'
   AND cmpprop.heavy_atoms>3
   AND act.standard_value IS NOT NULL
   AND trgd.organism IS NOT NULL
;
```

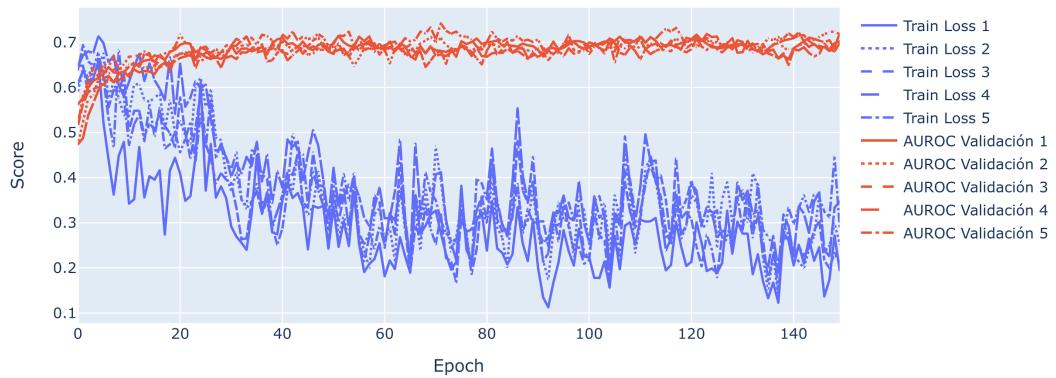
Se extrajo el resultado de esta *query* en un archivo .csv, y se continuó el curado que involucra arreglos más complejos, y análisis puntuales, utilizando la librería Pandas. Dicho proceso de filtrado se encuentra en el repositorio de la tesina: [link al notebook de filtrado y curado](#)

## S2. Curvas de entrenamiento de los modelos entrenados.

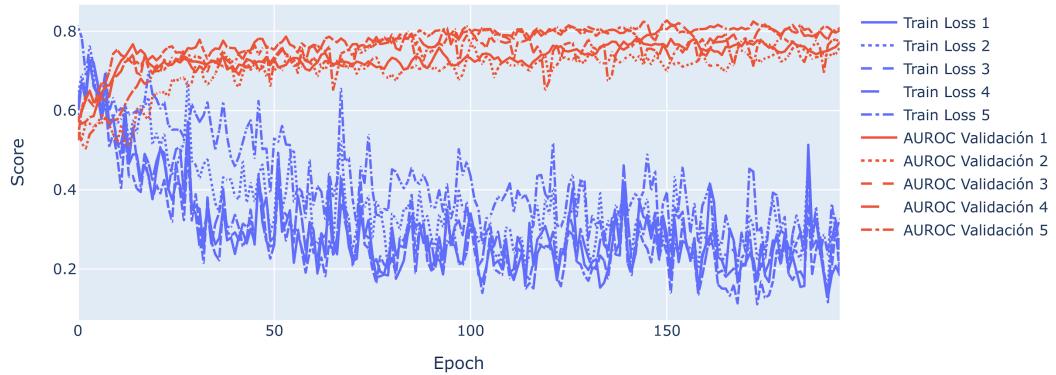
Los modelos se encuentran ordenados según AUROC máximo en el *set* de validación. Se utiliza un suavizado de media móvil exponencial con una ventana de 2, para facilitar la visualización. Se grafica el descenso del resultado de la función de coste, entrenando el modelo junto con el aumento del área bajo la curva ROC en el *set* de validación, implicando un aprendizaje de los compuestos activos e inactivos frente al *target*.



CHEMBL4072

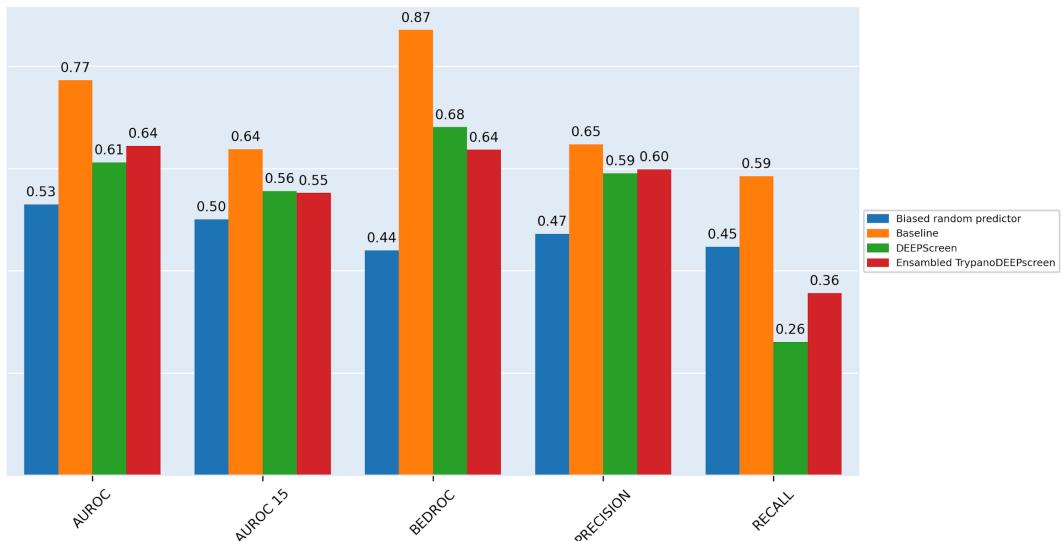


CHEMBL4567

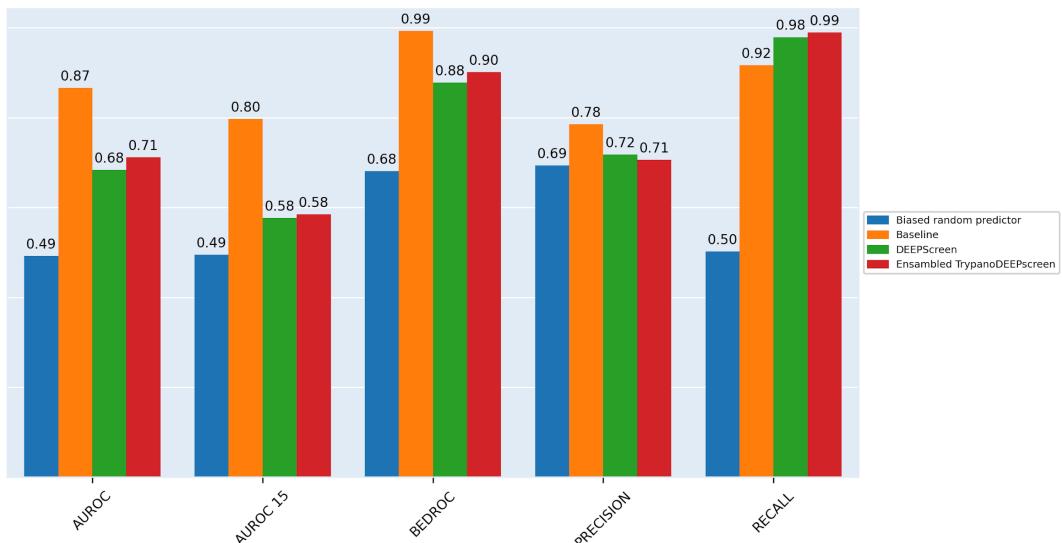


### S3. Performance de todos los modelos para cada uno de los targets.

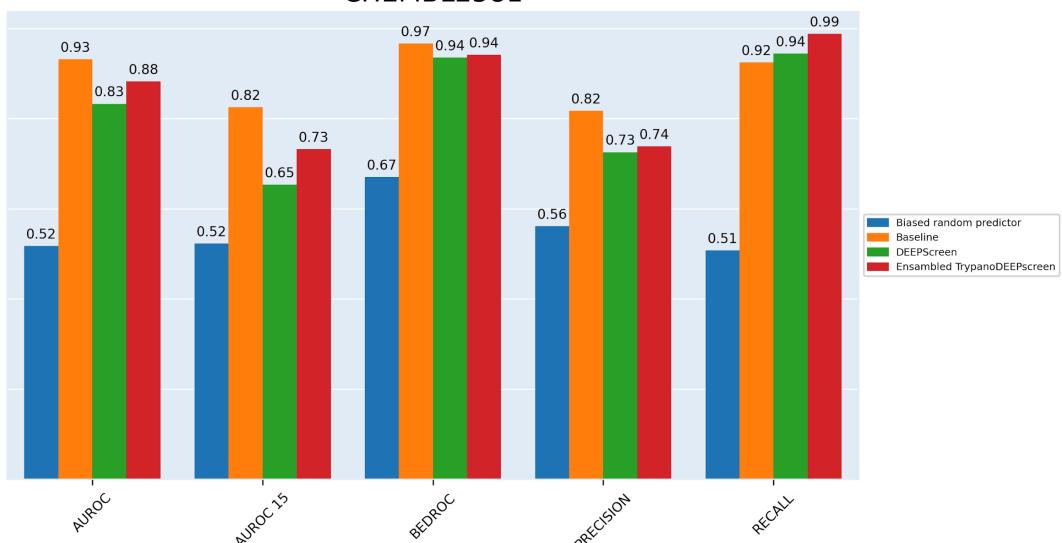
CHEMBL221



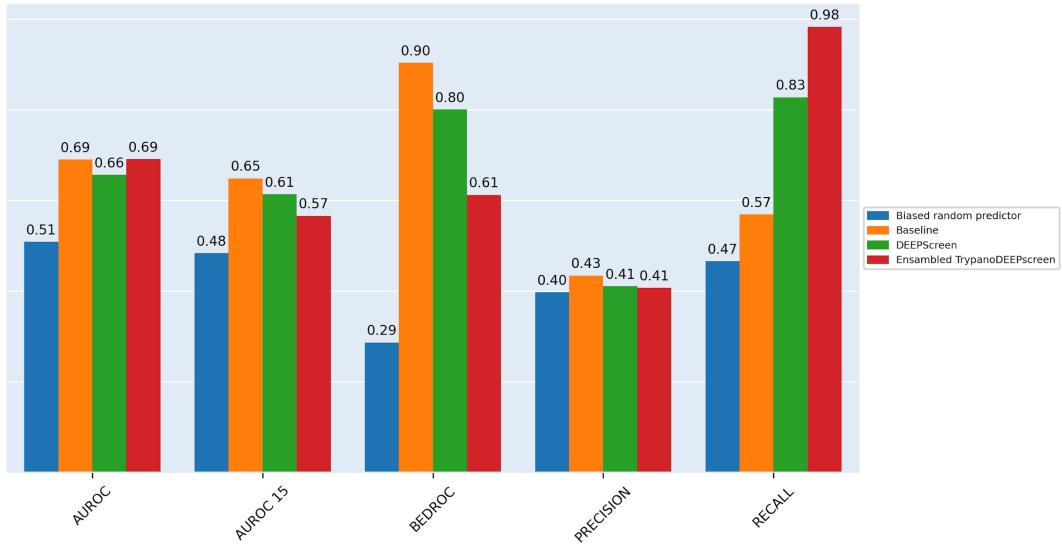
CHEMBL262



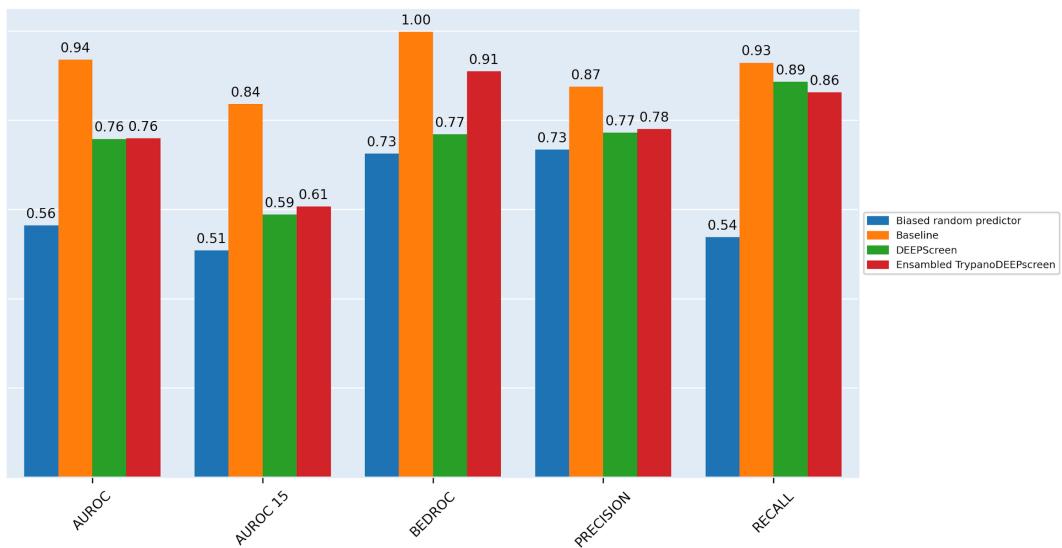
CHEMBL2581



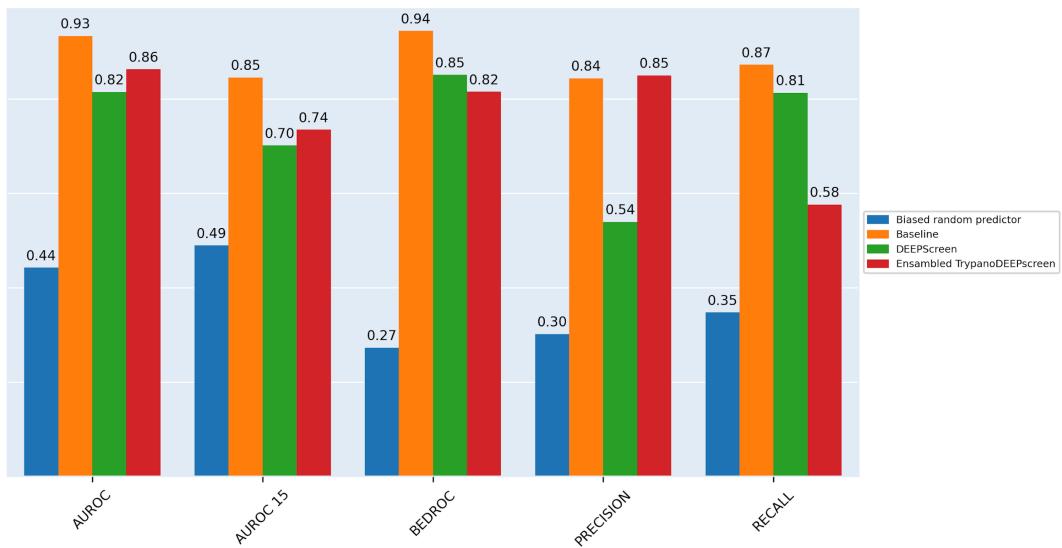
CHEMBL2850



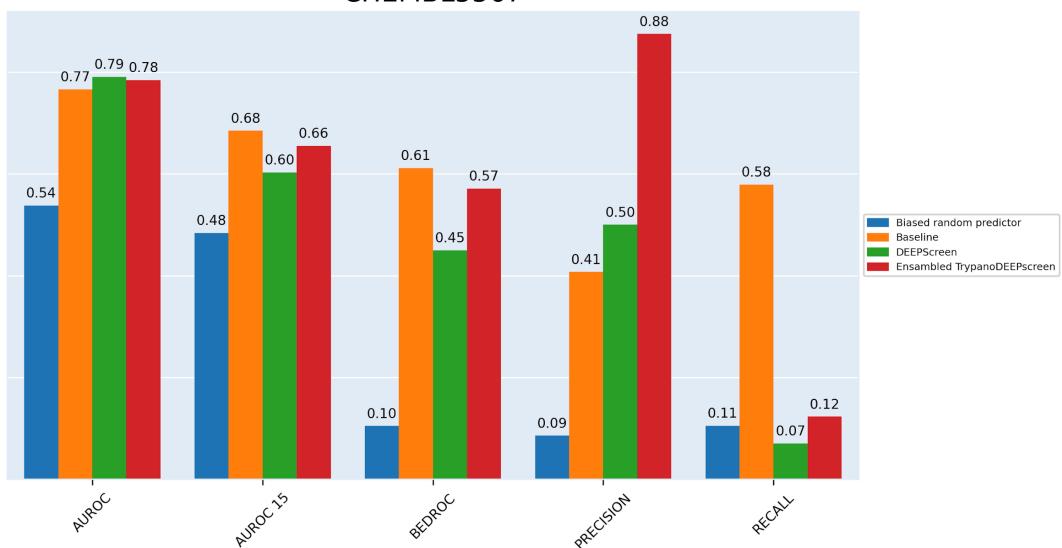
CHEMBL4072



CHEMBL4657



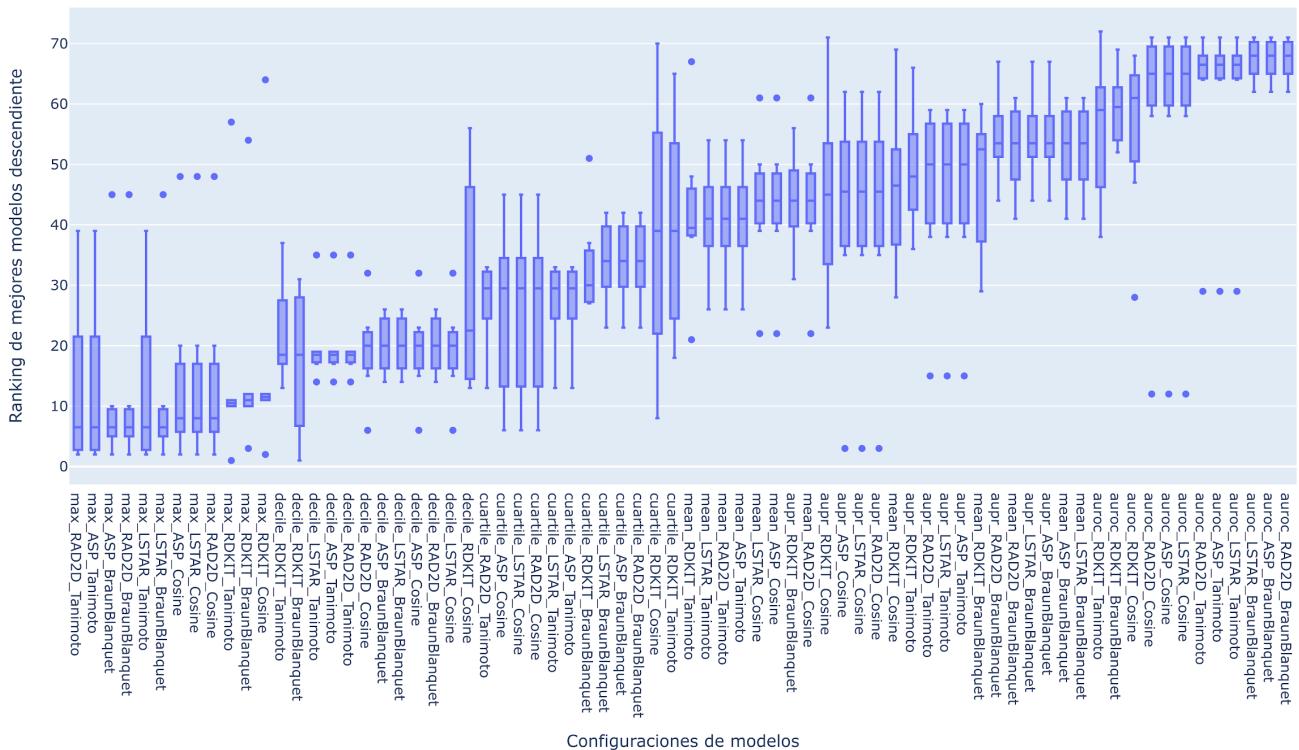
CHEMBL5567



#### S4. Análisis del grid search de configuraciones de baseline.

Se observó la *performance* de todas las configuraciones del algoritmo en los 7 targets de evaluación de la tesis. Luego se rankeo de mejor a peor *performance* todas las configuraciones en los distintos targets. Dichos rankings se presentan en la siguiente figura donde se graficaron boxplots del *ranking* descendente, de las distintas configuraciones de *baseline* en todos los targets. O sea, que la mejor configuración sería aquella que consistentemente resultó entre las mejores configuraciones en la mayor cantidad de targets.

Se observa como la configuración de la *fingerprint* ASP, con la similitud de Braun-Blanquet y la selección por la similitud máxima, fue consistente en estar entre los 10 mejores modelos para cada *target*. Por ese motivo se eligió dicha configuración. Los modelos están ordenados de derecha a izquierda por mediana del *ranking* creciente, siendo el *ranking* 1 el mejor.

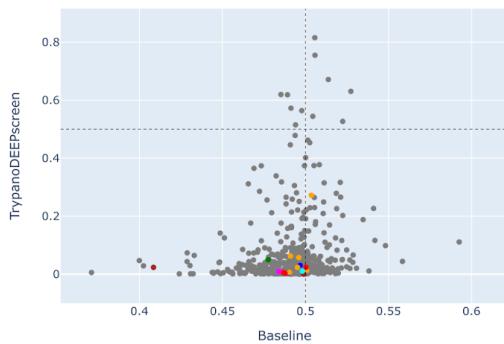


#### Aclaración

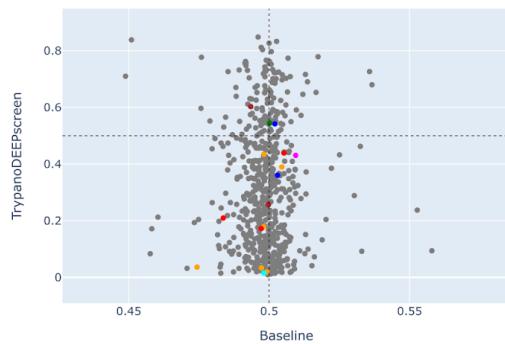
*Las configuraciones cuyo nombre comienza con AUROC o AUPR fueron intentos fallidos del algoritmo. En estas no se dividía el dataset de referencia en compuestos activos e inactivos, sino que se calculaba directamente el área bajo la curva ROC o PR a partir de la similitud del compuesto evaluado frente a la totalidad del conjunto de referencia. De este modo, si el compuesto era similar a moléculas activas, estas tendrían una señal elevada y el área bajo la curva se acercaría a 1; por el contrario, si era similar a los compuestos inactivos, el área tendía a 0. Dado que estas configuraciones fueron descartadas inicialmente y tenían una definición compleja, no fueron descritas en la metodología.*

#### S5. Evaluación de los modelos entrenados con conjuntos de datos con actividad Tripanocida.

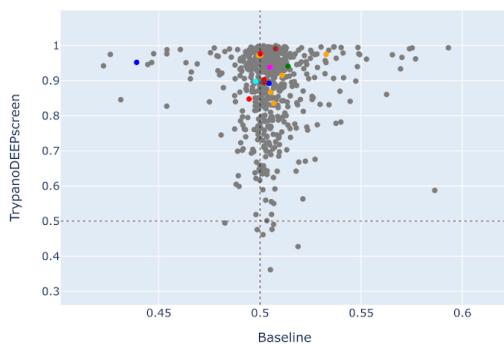
Luciferin 4-monoxygenase (CHEMBL5567)



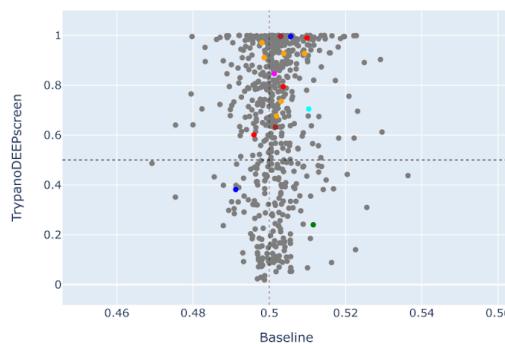
Cyclooxygenase-1 (CHEMBL221)



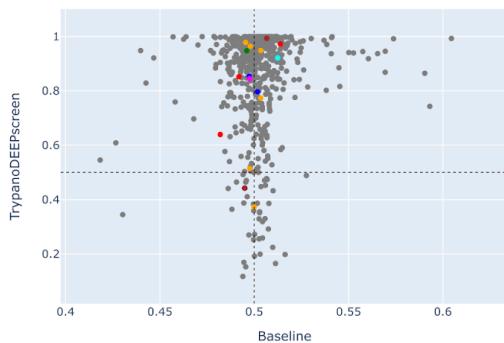
Glycogen synthase kinase-3 beta (CHEMBL262)



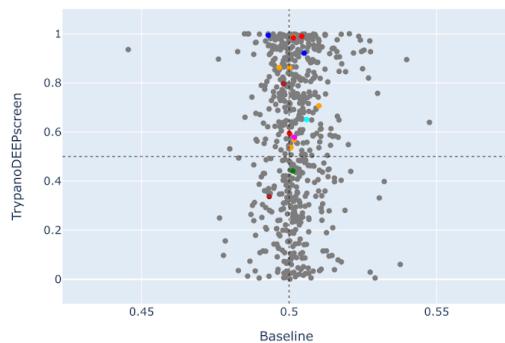
Cathepsin D (CHEMBL2581)



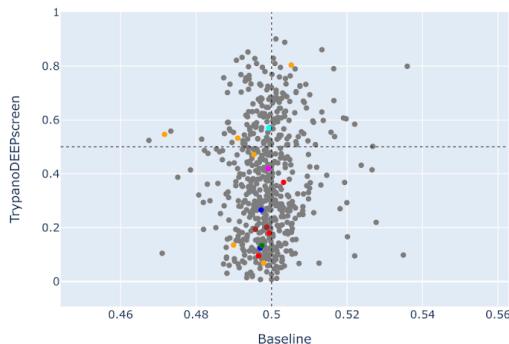
Glycogen synthase kinase-3 alfa (CHEMBL2850)



Cathepsin B (CHEMBL4072)



Dipeptidyl peptidase VIII (CHEMBL4657)



## Targets putativos

- Sin asignar
- serine/threonine\_protein\_kinase
- isocitrate\_dehydrogenase
- lanosterol\_14-alpha-demethylase
- mitogen-activated\_protein\_kinase
- phosphatidylinositol-kinase\_domain\_protein
- silent\_information\_regulator\_2
- phosphodiesterase
- NAD(P)-dependent\_oxidoreductase