Negative dataset (non-bioavailable)	Positive dataset (bioavailable)
CASE 1. NIHMS negative, %F<50 **486** compounds Both datasets are downlodable from the article https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3955412/	NIHMS positive, %F>=50 **509** compounds
CASE 2. HSDB dataset, 3551 minus substances that are in approved drugs, **3530** substances *Data retrieval from pubchem API: pubchem_retrieval.ipynb* *Data FILE with data:hsdb-smiles-no-repeat.csv*	ChEMBL approved and oral drugs, not withdrawn or discontinued + all drugs in trials **3842** compounds *Query FILE: chembl_ba_descriptors.sql*
CASE 3. non-BA from curated data (DrugBank, CentralDrug combined), **1451** compounds *Query FILE: curated_non_ba_descriptors.sql*	BA AND ORAL from curated data (DrugBank, CentralDrug combined), **889** compounds *Query FILE: curated_ba_descriptors.sql*
CASE 4. Chembl compounds that are NOT in the clinical trials (and have descriptors) **5590** compounds *Query FILE: chembl_not_in_trials_descriptors.sql*	ChEMBL approved and oral drugs, not withdrawn or discontinued + all drugs in trials **3842** compounds *Query FILE: chembl_ba_descriptors.sql*
CASE 5. Chembl compounds with QED <0.3 and ro5_violations>=2 **1997** compounds *Query FILE: chembl_qed_ro5.sql*	ChEMBL approved and oral drugs, not withdrawn or discontinued + all drugs in trials **3842** compounds *Query FILE: chembl_ba_descriptors.sql*

Figure 1. Datasets used in training the model to predict on structure, descriptors, both

A dataset of around 13,000 compounds from curated data mart were selected to predict on (all_curated.sql query selects the data). The prediction was made on the Case 1 and Case 5 model by eliminating insignificant number of duplicates between the datasets to ensure that the prediction on the data of which of the model was not trained on. The predicted two probabilities (one probability per prediction) of the compound (to be bioavailable) are meant to be outputted to the user. This would happen in the application simultaneously with the bioavailability class and/or percent, if such existed in the data.

We also recorded the x and y coordinates of the Uniform Manifold Approximation & Projection (UMAP) from the data set on Case 1 model. The coordinates are meant to be used in the application to find the 10 closest compounds to the one that the user is interested in. Case1 dataset was obtained from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3955412/ . Case 2 dataset was selected from the curated data mart.