Human intestinal absorption and bioavailability datasets

QSAR and Molecular Modelling Department, Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences, Sofia, Bulgaria

Data collected for the European Community’s 7th Framework Program project "Integrated In Silico Models for the Prediction of Human Repeated Dose Toxicity of Cosmetics to Optimise Safety (COSMOS)"

2011-2015
Data compilation

The data were compiled from 28 research papers containing human gastro-intestinal absorption (GIA) and bioavailability data (consisting of 15 – 550 compounds), both collected from the literature or measured by the authors. The GIA data compiled were absorbed fractions (FA, %) or jejunal permeability coefficients (Peff, cm/s); bioavailability data were systemically available fractions after oral ingestion (BA, %).

Structural information

In the processed papers, the compounds were presented mainly by commercial or trivial names, thus the structural information and other important identifiers were collected from the NCI/CADD Chemical Identifier Resolver service (https://cactus.nci.nih.gov/chemical/structure/) and from the NCBI PubChem project (https://pubchem.ncbi.nlm.nih.gov/). The the SMILES codes retrieved were converted to canonical "inchified" form by means of OpenBabel software (http://openbabel.org/), mixtures and compounds with uncertain structure (e.g., polymers) were omitted but compounds were not desalted (some BA data differ significantly for acids and their salts).

Data reduction

Since most of the data in the source papers were collected from the literature, it was assumed that identical values could represent a single datapoint and thus, only a single instance of these datapoints was left, retaining the information for all sources where the value was found. After the procedure, the total datapoints were reduced from 3255 to 1973 for total of 853 compounds.

Datasets content

The data are presented in three comma-separated-values files, one for each of the collected parameters (FA, Peff, BA):

GIA_FA_data.csv (1227 values for 783 compounds)
GIA_Peff_data.csv (93 values for 57 compounds)
Oral_BA_data.csv (653 values for 543 compounds)

The compound name found in the source paper, InChI key, canonical "inchified" SMILES, CAS registry number, PubChem CID, collected value, and the references it was found in, are provided for each datapoint.

Note

Data from ref. 28 are just summarized in the GIA FA dataset. Complete set of data from this paper are available from the authors' website (http://modem.ucsd.edu/adme/).
Data sources


