Main Compound-Target Pair Information

Column Name	Type	Information about	Based on	Comments
parent_molregno	Int	Compound	ChEMBL: molecule_dictionary	Mapped from activities through molecule_hierarchy
parent_chemblid	String			
parent_pref_name	String			
max_phase	Float			
first_approval	Int			
usan_year	Int			
black_box_warning	Int			
prodrug	Int			
oral	Int			
parenteral	Int			
topical	Int			
tid	Int	Target	ChEMBL: assays	
mutation	String		ChEMBL: variant_sequences	
target_chembl_id	String		ChEMBL: target_dictionary	
target_pref_name	String			
target_type	String			
organism	String			
tid_mutation	String		<tid>_<mutation>	
cpd_target_pair	String	Compound-Target Pair	<parent_molregno>_<tid>	
cpd_target_pair_mutation	String		<parent_molregno>_ <tid_mutation>	

Aggregated Values

Aggregated per compound-target pair using parent_molregno and tid_mutation.
_BF: based on pchembl_value_mean_BF (based on binding + functional assays)
_B: based on pchembl_value_mean_B (based on binding assays)

Column Name	Type	Information about	Based on	Comments
pchembl_value_mean_BF	Float	Compound-Target Pair	ChEMBL: pchembl values in activities in initial query (dropped after aggregation)	Mean pchemb_value for the compound-target pair
pchembl_value_mean_B	Float			
pchembl_value_max_BF	Float			Maximum pchemb_value for the compound-target pair
pchembl_value_max_B	Float			
pchembl_value_median_BF	Float			Median pchemb_value for the compound-target pair
pchembl_value_median_B	Float			
first_publication_cpd_target_pair_BF	Int		ChEMBL: year values in docs in initial query (dropped after aggregation)	first publication in ChEMBL with this compound-target pair
first_publication_cpd_target_pair_B	Int			
first_publication_cpd_target_pair_w_pchembl_BF	Int			first publication in ChEMBL with this compound-target pair and an associated pchembl value
first_publication_cpd_target_pair_w_pchembl_B	Int			

DTI (Drug-Target Interaction) Annotations

Assigned based on cpd_target_pair, does not include mutation information.

Column Name	Type	Information about	Based on	Comments
therapeutic_target	Bool	Target	ChEMBL: drug_mechanism table	Is the target in the drug mechanism table?
DTI	String	Compound-Target Pair	Assigned as per rules layed out in code / comments / below	Drug target interaction annotation

Mechanism to assign the DTI annotations

in drug_mechanism table?	max_phase?	therapeutic target?	DTI annotation	Explanation
Yes	4	–	D_DT	drug - drug target
	3	–	C3_DT	clinical candidate in phase 3 - drug target
	2	–	C2_DT	clinical candidate in phase 2 - drug target
	1	–	C1_DT	clinical candidate in phase 1 - drug target
	< 1 [1]	–	C0_DT	compound in unknown clinical phase - drug target
No	–	Yes	DT	drug target
	–	No	NDT	not drug target

Is the compound-target pair in the drug_mechanisms table? = Is it a known relevant compound-target interaction?
What is the max_phase of the compound? = Is it a drug / clinical compound?
Is the target in the drug_mechanisms table? = Is it a therapeutic target?

[1] There are three possible annotations in ChEMBL with max_phase not between 1 and 4:
0.5 = early phase 1 clinical trials
-1 = clinical phase unknown for drug or clinical candidate drug, i.e., where ChEMBL cannot assign a clinical phase
NULL = preclinical compounds with bioactivity data
All three are grouped together into the annotation C0_DT.

Compound and Target Properties Based on ChEMBL Data

Column Name	Type	Information about	Based on	Comments
first_publication_cpd	Int	Compound	ChEMBL: docs	first appearance of the compound in the literature
mw_freebase	Float		ChEMBL: compound_properties	
alogp	Float			
hba	Int			
hbd	Int			
psa	Float			
rtb	Int			
ro3_pass	String			
num_ro5_violations	Int			
cx_most_apka	Float			
cx_most_bpka	Float			
cx_logp	Float			
cx_logd	Float			
molecular_species	String			
full_mwt	Float			
aromatic_rings	Int			
heavy_atoms	Int			
qed_weighted	Float			
mw_monoisotopic	Float			
full_molformula	String			
hba_lipinski	Int			
hbd_lipinski	Int			
num_lipinski_ro5_violations	Int			
standard_inchi	String		ChEMBL: compound_structures	
standard_inchi_key	String			
canonical_smiles	String			

Compound and Target Properties Based on ChEMBL Data cont.

_BF: based on pchembl_value_mean_BF (based on binding + functional assays)
_B: based on pchembl_value_mean_B (based on binding assays)

Column Name	Type	Information about	Based on	Comments
LE_BF / LE_B	Float	Compound	Calculated based on pchembl_value_mean	Ligand efficiency metric
BEI_BF / BEI_B	Float			
SEI_BF / SEI_B	Float			
LLE_BF / LLE_B	Float			
atc_level1	String		ChEMBL: atc_classification, molecule_atc_classification	Anatomical Therapeutic Chemical (ATC) classification, level 1
target_class_l1	String	Target	ChEMBL: protein_classification, protein_family_classification	Target class, level 1
target_class_l2	String			Target class, level 2

RDKit-Based Compound Descriptors

Column Name	Type	Information about	Based on	Comments
fraction_csp3	Float	Compound	canonical_smiles + built-in RDKit methods	
num_heteroatoms	Int			
num_stereocentres	Int			
ring_count	Int			
num_aliphatic_rings	Int			
num_aliphatic_carbocycles	Int			
num_aliphatic_heterocycles	Int			
num_aromatic_rings	Int			
num_aromatic_carbocycles	Int			
num_aromatic_heterocycles	Int			
num_saturated_rings	Int			
num_saturated_carbocycles	Int			
num_saturated_heterocycles	Int			
aromatic_atoms	Int		canonical_smiles + RDKit-based methods	
aromatic_c	Int			
aromatic_n	Int			
aromatic_hetero	Int			
scaffold_w_stereo	String		canonical_smiles + built-in RDKit methods	
scaffold_wo_stereo	String			

Annotations for Filtering

The columns based on the calculated subsets are only available in the full dataset to facilitate the filtering into subsets.

Column Name	Type	Information about	Based on	Comments
in_dm_table	Bool	Compound-Target Pair	ChEMBL: drug_ mechanism	Is the compound-target pair (cpd_target_pair) in the drug mechanism table?
keep_for_ binding	Bool			Rows to keep if interested in information based only on binding assays + the drug_mechanism table. True if pchembl_value_mean_B (based on binding assays) exists or if in_dm_table == True, i.e., the pair is in the drug mechanism table.
BF_100	Bool		calculated subsets	binding + functional data; at least 100 comparator compounds* per target
BF_100_ c_dt_d_dt	Bool			binding + functional data; at least 100 comparator compounds* per target; at least one compound annotated with D_DT or C<p>_DT (C0_DT, C1_DT, C2_DT, C3_DT) per target
BF_100_d_dt	Bool			binding + functional data; at least 100 comparator compounds* per target; at least one compound annotated with D_DT per target
B_100	Bool			binding data only; at least 100 comparator compounds* per target
B_100_ c_dt_d_dt	Bool			binding data only; at least 100 comparator compounds* per target; at least one compound annotated with D_DT or C<p>_DT (C0_DT, C1_DT, C2_DT, C3_DT) per target
B_100_d_dt	Bool			binding data only; at least 100 comparator compounds* per target; at least one compound annotated with D_DT per target

* comparator compounds must have a pchembl value but don't have a specified max_phase or DTI annotation, i.e., drugs and clinical candidates are counted as comparator compounds as well