Compound-protein interaction prediction for SARS-CoV-2

In this homework, you will predict the potential drug candidates targeting the SARS-CoV-2 proteins, using the information from the given compound-protein interactions (CPIs).

1. Develop and evaluate your model:

You need to develop **an algorithm to predict virus-related CPIs, using either a regression or a classification model.** The data for building a regression model are in "chembl_reg_virus_all.tsv", in which labels are continuous (i.e., p-bioactivity values), and data for classification are in "chembl_cls_virus_all.tsv", in which labels are binary. Data were downloaded from the ChEMBL database (https://www.ebi.ac.uk/chembl/). The columns of these tab-separated files are ChEMBL IDs of the compounds, UniProt IDs of the proteins (except for "SARS-3CLpro", "SARS-PLpro" and "SARS-helicase"), InChIs of the compounds, sequences of the proteins, and labels.

Please use two different 10-fold cross-validation procedures to evaluate performance of your algorithm. One is randomly splitting the interactions into 10 folds. The other is randomly splitting the proteins into 10 folds and assign the corresponding interactions into either training or test data. That is, in each fold, the CPIs involving the same proteins cannot appear in both training and test data. This new-protein setting is close to the real situation, in which the test proteins are not seen in the training data. **Please report the AUC and AUPR scores for the regression task, or RMSE and Pearson correlation for the classification task**.

2. Apply your model to predict drugs for SARS-CoV-2 proteins

The drug candidates should be predicted from the 6255 existing drugs, provided in file "drug_info.tsv". **Please list top 10 predicted drugs for SARS-CoV-2 3CLpro and PLpro.** The protein sequences for these SARS-CoV-2 proteins are provided below. It would be nice if you can provide more information about the drugs in your report. According to their DrugBank IDs, you can find the chemical structures (images), original indications and targets of these drugs in https://www.drugbank.ca/.

> Sequence of SARS-CoV-2 3CLpro:
SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLL
IRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLAC
YNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAG
TDLEGNFYGPFVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVA
MKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQNGMNGRTILGSALLEDEFTPF
DVVRQCSGVTFQ
> Sequence of SARS-CoV-2 PLpro:
APTKVTFGDDTVIEVQGYKSVNITFELDERIDKVLNEKCSAYTVELGTEVNEFACVVADAV
IKTLQPVSELLTPLGIDLDEWSMATYYLFDESGEFKLASHMYCSFYPPDEDEEEGDCEEE

EFEPSTQYEYGTEDDYQGKPLEFGATSAALQPEEEQEEDWLDDDSQQTVGQQDGSED
NQTTTIQTIVEVQPQLEMELTPVVQTIEVNSFSGYLKLTDNVYIKNADIVEEAKKVKPTVVV
NAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVGGSCVLSGHNLAKHCLH
VVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFGADPIHSLRVCVDTVRTNVYLA
VFDKNLYDKLVSSFLEMKSEKQVEQKIAEIPKEEVKPFITESKPSVEQRKQDDKKIKACVE
EVTTTLEETKFLTENLLLYIDINGNLHPDSATLVSDIDITFLKKDAPYIVGDVVQEGVLTAVVI
PTKKAGGTTEMLAKALRKVPTDNYITTYPGQGLNGYTVEEAKTVLKKCKSAFYILPSIISN
EKQEILGTVSWNLREMLAHAEETRKLMPVCVETKAIVSTIQRKYKGIKIQEGVVDYGARF
YFYTSKTTVASLINTLNDLNETLVTMPLGYVTHGLNLEEAARYMRSLKVPATVSVSSPDA
VTAYNGYLTSSSKTPEEHFIETISLAGSYKDWSYSGQSTQLGIEFLKRGDKSVYYTSNPT
TFHLDGEVITFDNLKTLLSLREVRTIKVFTTVDNINLHTQVVDMSMTYGQQFGPTYLDGA
DVTKIKPHNSHEGKTFYVLPNDDTLRVEAFEYYHTTDPSFLGRYMSALNHTKKWKYPQV
NGLTSIKWADNNCYLATALLTLQQIELKFNPPALQDAYYRARAGEAANFCALILAYCNKTV
GELGDVRETMSYLFQHANLDSCKRVLNVVCKTCGQQQTTLKGVEAVMYMGTLSYEQF
KKGVQIPCTCGKQATKYLVQQESPFVMMSAPPAQYELKHGTFTCASEYTGNYQCGHYK
HITSKETLYCIDGALLTKSSEYKGPITDVFYKENSYTTTIKPVTYKLDGVVCTEIDPKLDNY
YKKDNSYFTEQPIDLVPNQPYPNASFDNFKFVCDNIKFADDLNQLTGYKKPASRELKVTF
FPDLNGDVVAIDYKHYTPSFKKGAKLLHKPIVWHVNNATNKATYKPNTWCIRCLWSTKPV
ETSNSFDVLKSEDAQGMDNLACEDLKPVSEEVVENPTIQKDVLECNVKTTEVVGDIILKP
ANNSLKITEEVGHTDLMAAYVDNSSLTIKKPNELSRVLGLKTLATHGLAAVNSVPWDTIAN
YAKPFLNKVVSTTTNIVTRCLNRVCTNYMPYFFTLLLQLCTFTRSTNSRIKASMPTTIAKN
TVKSVGKFCLEASFNYLKSPNFSKLINIIIWFLLLSVCLGSLIYSTAALGVLMSNLGMPSYC
TGYREGYLNSTNVTIATYCTGSIPCSVCLSGLDSLDTYPSLETIQITISSFKWDLTAFGLVA
EWFLAYILFTRFFYVLGLAAIMQLFFSYFAVHFISNSWLMWLIINLVQMAPISAMVRMYIFF
ASFYYVWKSYVHVVDGCNSSTCMMCYKRNRATRVECTTIVNGVRRSFYVYANGGKGF
CKLHNWNCVNCDTFCAGSTFISDEVARDLSLQFKRPINPTDQSSYIVDSVTVKNGSIHLY
FDKAGQKTYERHSLSHFVNLDNLRANNTKGSLPINVIVFDGKSKCEESSAKSASVYYSQ
LMCQPILLLDQALVSDVGDSAEVAVKMFDAYVNTFSSTFNVPMEKLKTLVATAEAELAKN
VSLDNVLSTFISAARQGFVDSDVETKDVVECLKLSHQSDIEVTGDSCNNYMLTYNKVEN
MTPRDLGACIDCSARHINAQVAKSHNIALIWNVKDFMSLSEQLRKQIRSAAKKNNLPFKL
TCATTRQVVNVVTTKIALKGG