Instructions & Guidelines

This document is not intended to provide a detailed step-by-step guideline for different molecular modelling, chemometric and cheminformatic techniques. It lists only the key points and specifications to reflect the expectations of different PS creators from the received entries. This document along with the collections of resources has been compiled based on the inputs received from different PS creators, several ambassadors and experts of the Drug Discovery Hackathon.

Software tools and databases mentioned in this document and the collection of resources are only suggestions to the participants. Several of the tools and databases may not be made available to the participants. Participants are free to choose any free or proprietary software tools available to them for solving problems. It is also recommended to mention in the Input Forms the software used for different molecular modelling, chemometric and cheminformatics procedures while solving the problems.

Table of Contents

Track 1-Drug Design for anti-COVID19 hits/leads

General Instructions/ Specifications/ Software/ Databases

General instructions/specifications along with suggested/indicative software tools and different databases for different chemometric/cheminformatic procedures are listed pointwise.

Methods - Benchmarks

Homology Modelling

- Identity not less than 25%
- Sequence vs Template PDB should have satisfactory coverage in the binding site region
- RMSD and DOPE score w.r.t selected template
- Models should pass all the validation criteria using WHATIF, PROCHECK, etc.
- MD validation is mandatory
- Modelled protein should be justified before using it for molecular docking studies. Any expert should approve the homology model before taking up further research.

References:

1. http://dx.doi.org/10.1111/cbdd.13388
2. https://doi.org/10.2174/1570180814666170110122027

Molecular Docking

- Submit at least top 25% (max. 100) or top 100 hits
- If co-crystal ligands available and used as reference, compare the hits
- Benchmark with known inhibitors
- Diversity to be considered
- Key AA Active site residues with catalytic amino acid as the grid centre. Whole protein centre docking is also recommended in order to examine other sites of binding.
- Flexible or Rigid Ligand Flexible preferred
- LE - frag & lig - MW
- Docking score (best within your screening set)
- Binding Energy (kcal/mol)
- Binding Affinity
- Selection criteria for doing MD post docking top 5-10 docking score or high energy complexes would go for MD simulations. Better to drop unfavourable conformations of ligands, namely those with eclipsed conformations or steric clashes.
- Enrichment factor calculations

References:

1. https://dx.doi.org/10.3390%2Fijms20184331
2. http://dx.doi.org/10.4018/978-1-5225-0115-2.ch003
3. https://doi.org/10.1021/jm0608356

Molecular Dynamics

- Time range minimum 20-100ns
- RMSD, RMSF, Radius of Gyration, Pot. Energy, PCA
- T-temperature (300K)
- Implicit or explicit conditions - Solvation
- To save the computational resources, attempts should be made so that the Expert and the participant might come to an agreement on the chosen molecules to be taken up for molecular dynamics simulations.

References:

1. https://doi.org/10.1080/17460441.2018.1403419
2. https://doi.org/10.1007/3-540-29623-9_0820

QSAR: Detailed Instructions

Instructions for developing QSAR models are given in the following link:

- https://drive.google.com/file/d/1VxDXkyUalwbqdO9m4-dc_1WZlPOeLeYk/view?usp=sharing

References:

1. http://dx.doi.org/10.4018/IJQSPR.20200701.oa1
2. J Indian Chem Soc, 95 (2018) 1497-1502.
3. http://dx.doi.org/10.4018/IJQSPR.2016010101
4. http://dx.doi.org/10.1007/978-3-319-17281-1
5. https://doi.org/10.1021/acs.jcim.6b00088
6. https://doi.org/10.1002/cplu.201200038

Pharmacophore Screening

- Use X-Ray crystal geometry as reference if available
- Similarity criteria to be checked
- Validation using known databases of actives and inactives and a decoy set

References:
1. http://dx.doi.org/10.1016/B978-0-12-801505-6.00010-7
2. https://doi.org/10.1021/ci2005274

List of databases / datasets

Drug Target Protein - PDB IDs

PDBe-KB COVID-19

RCSB COVID19 - Resources

ORF1a - This looks like a polyprotein. Structures are available for specific domains only

Papain-like protease (6WZU) (4MM3, 3MJ5, 5E6J, 6YVA, 3E9S)

PLpro inhibitors (GRL0617)
3CL-protease - 6Y2E, 6LU7 (PDB ID)

3CLpro inhibitors (lopinavir)
NSP 1-16 2′-O-methyltransferase (NSP16) - PDB ID: 6WKQ

CYP inhibitors (Alisporivir)

RNA dependent RNA polymerase: 6M71 (PDB ID : 6M71, 7BV2)

RdRp inhibitors (ribavirin, BCX4430)
Inhibit RNA replication mechanism (Remdesivir, GS-5734)

7BV2 (RdRp co-crystallized with Remdesivir)
7BV4
7BTF (RNA-dependent RNA polymerase (RdRp, also named nsp12))

Helicase 6JYT

Helicase inhibitors (SSYA10-001)
Endribonuclease - 6VWW (PDB ID)

Structural Proteins
Spike Protein - 6VSB (PDB ID)

S protein bound to ACE2 - 6VW1 (PDB ID)
S protein bound to ACE2- 6M0J

Home-pentameric membrane E protein - 5X29 (PDB ID)

Human coronavirus spike protein - 5I08 (PDB ID)

Closed conformation - 6VXX (PDB ID)

Open conformation - 6VYB (PDB ID)

S1 and S2 subunits contain the key amino acids.

Human Transmembrane Protease serine-2 (TMPRSS2) T99908,
http://db.idrblab.net/ttd/data/covid19-target/details/t99908

Lipid Databases

- LipidBank
- LIPID MAPS structure database

Small Molecule Datasets

- FDA approved drugs
- DrugBank
- DrugCentral
- ZINC15
- Chemspider
- Asinex
- ChEMBL
- PubChem
- Maybridge
- Selleckchem
- MolPort
- Enamine
- ChemDiv
- CAS Antiviral
- Therapeutic Target database
- GHDDI-Data
- TIMBAL

Phytochemical Datasets

- IMPPAT
- TIPdb
- COCONUT
- Phytochem Lib
- HerbMedPro
- Herbal constituents SWSBM
- ESCOP

Fragment Datasets

- Maybridge
- ChemBridge
- Asinex fragments
- ZINC fragment
- FCH fragment library
- Otava
- ChemDiv fragments
- Glide fragments
- REAL fragment library
- MolMatInf

Protein-ligand Databases

- STITCH
- Bindingdb
- Matador
- Binding MOAD
- BioLiP
- Protein ligand interaction cluster
- Protein drug interaction database

Helminths Database

- Wormbase
- Uniport

Virtual Screened Datasets

- Screened Hits on 6LU7

Comprehensive Data for COVID19

- PubChem (Compounds, Assay, Proteins, Literatures)
- DrugBank COVID-19
- WHO COVID-19
- COVID19: Drug Candidates

MD Datasets

- SARS CoV2 MD Data

Software Tools

Cheminformatics/Ligand Preparation

- DataWarrior
- Instant JChem
- StarDrop
- Dotmatics
- KNIME

Docking

- AutoDock Vina / rDock / Online Servers (HADdock)
- OEDocking (Fred, Openeye)
- Glide
- SeeSAR-FlexX
- MOE

- Ligandfit
- PyRx
- PLANTS
- Virtual flow
- SANJEEVINI (online tool)
- COVID19 Dock Server
- Molegro Virtual Docker
- Virtual Screening Drug Design Bruno Villoutreix (3300 tools (online & standalone)
- D3Targets-2019-nCoV

Homology/Ab initio Modelling

- Online Servers (MobWeb, SwissModel, iTASSER, Phyre3D)
- MODELLER
- Rosetta
- Robetta
- MOE

Protein Model Validations

- Rosetta
- RAMPAGE: Ramachandran Plot Assessment
- SAVES v5.0 (WHATCHECK, PROCHECK, ERRAT, Verify3D, PROVE, CRYST)
- ProSA-web - Protein Structure Analysis

Molecular Dynamics

- Gromacs
- AMBER
- NAMD
- Desmond

Pharmacophore mapping

- Phase
- Openeye
- Discovery Studio

QSAR

- Descriptor computation tools -
  - PaDel-Descriptor - Free
  - CDK - Free
  - Dragon - Proprietary
  - AlvaDescriptor, , Proprietary
  - OCHEM; Free
  - Introduction-ChemDes-Molecular descriptors computing platform (Free)

- Statistical (regression/classification) modelling tools-
  - DTC Lab softwares - Link and Link - Free
  - QSARINS - Free
  - InfoStat - Free for students
  - RStudio with R - Free
  - MINITAB - Proprietary
  - STATISTICA - Proprietary
  - SIMCA - Proprietary

- ML QSAR model
  - StarDrop - Auto Modeller

Chemical Space Search

- InfiniSee

ADME/T

- Online Servers (Swiss ADME)
- QikProp
- StarDrop
- MOE
- KNIME Nodes
- DSSTox

Molecular Visualization Software

Following software may be used to visualize and prepare ligand-molecular interactions, dynamics, large molecules and complexes

- UCSD Chimera
- Discovery Studio Visualizer
- PyMol
- VMD

Track 2- Machine Learning/Models/Tools

General Guidelines

Obtaining a good machine learning model with excellent generalization is an art and requires a multitude of algorithms, features, parameter optimization and several iterations. Participants are thus encouraged to think about the underlying problem, required features and assess different approaches in Machine learning.

Consensus, ensemble, bagging , boosting are just ways in which better solutions can be arrived at. Classification problems can be converted into regression and vice-versa when desirable. SImilarly feature encoding and selection are very important steps that need to be fully explored and documented.

Constructing re-usable, modular, parametrized pipelines for each of the stages in Machine learning is highly recommended. For example each stage from data preparation, feature engineering, feature, selection, model building, optimization & prediction should be clearly separated into groups of nodes or python programs, as the case might be.

Various cut-offs like confidence levels while selecting models, methods for missing value treatment etc can be read from parameter files to be read and used during the modelling phase. File input/output, nomenclature of files should also be parameterized and not hard-coded. In general, given particular data the final model should be reproducible given a certain set of parameters and configuration.

In the case of chemometric, QSAR, QSPR models, particular libraries used to generate specific features should be clearly mentioned, since there are liable to be significant differences in how things are calculated.

For classification it is recommended to have a 70:30 training, validation split and a separate blinded test dataset.

Exploratory analysis

Exploratory analysis is extremely important in ascertaining data quality, biases/skew and need to engineer new features. This step is crucial to further modelling efforts. Any package, library in R, python or visual tools may be used to perform this and prepare a report or a visualization.

Measures to account for class imbalance, normalization/standardization of numerical features, binning are some strategies that can help post this exploratory analysis. These need to be articulated.

Model Evaluation

Overfitting and sensitivity to data, parameters is of utmost concern to machine learning. Other than this the domain of the problem and its applicability to similar domains is also a big consideration. For example models built for a particular class of drugs might or might not be easily transferable or usable for another class of molecules. Thus the choice of and explanation of the Applicability domain in terms of original data or features needs to be clearly explained.

Most often the Control class or negative class is very hard to define for many classification problems. The approach to define this and the accuracy variability, based on changes in how negative classes are defined should be documented.

Some well known guidelines are available at
https://scikit-learn.org/stable/modules/model_evaluation.html

Summary

As specified in input form the differentiator in approach based on above should be clearly articulated. Similarly the *initial* architecture of the pipeline or deep learning models to be explored should be mentioned in the proposal. While its clear that the final architecture, particularly in Deep learning may not be known in advance, some general ideas for exploration need to be documented in the proposal.

Classification Metrics

AUC, confusion metric, log loss and Precision , Recall, F1 score and support.

Regression Metrics

MAE, MAPE, R^2 , adjusted R^2s can be used as metrics. However, the following guide will help in choosing the right metric. The proposals and final reports should describe which metric was chosen and why.
https://www.h2o.ai/blog/regression-metrics-guide/

F1 Score, Sensitivity, selectivity should be > 0.8 and >0.9 if possible in general. MAPE should also be >80% on test sets. Other regression metrics will be proportional. Great care needs to be taken to check Y/response variables are not used as independent variables. Anytime accuracies reach 99% , substantial checks should be performed for overtraining and spurious results.

For deep learning Inception score, Fechet Inception distance and others can be explored.

Software Links

Reinforcement learning libraries

This link describes commonly used python packages or Reinforcement learning.
References:
https://analyticsindiamag.com/python-libraries-reinforcement-learning-dqn-rl-ai/

Some of the same backend libraries are also used for Deep learning
References:
https://towardsdatascience.com/best-python-libraries-for-machine-learning-and-deep-learning-b0bd40c7e8c
https://www.kdnuggets.com/2018/11/top-python-deep-learning-libraries.html
https://towardsdatascience.com/top-10-best-deep-learning-frameworks-in-2019-5ccb90ea6de

Some libraries like Keras provide a higher level abstraction to others like Tensorflow. Some of these deep learning libraries will also provide support for auto-encoders.e.g. Keras.
References:
https://www.tensorflow.org/

Conventional Machine learning modelling (Non-DL)

References:
https://scikit-learn.org/stable/
https://www.ubuntupit.com/best-r-machine-learning-packages/
https://www.h2o.ai/

Other supporting Recommended libraries for ML
Python NumPy, Scipy, Pandas, Matplotlib,

Biomedical text data querying, scraping , parsing, text mining

While not all biomedical text repositories might be supported, these libraries and packages are quite handy for literature mining

References:
https://pypi.org/project/pubmed-lookup/
https://pypi.org/project/entrezpy/
https://github.com/titipata/pubmed_parser
https://biopython.org/DIST/docs/api/Bio.Entrez-module.html
https://cran.r-project.org/web/packages/easyPubMed/index.html
https://cran.r-project.org/web/packages/pubmed.mineR/index.html
https://amunategui.github.io/pubmed-query/
https://rdrr.io/cran/biorxivr/

For Bioarxiv.org
References:
https://sandbox.idre.ucla.edu/sandbox/uncategorized/web-scraping-example-scrape-article-search-pages-iteratively
https://pypi.org/project/biorxiv-cli/
https://predictablynoisy.com/scrape-biorxiv
https://github.com/blekhmanlab/rxivist

NLP might be necessary for mining text and named entity recognition
References:
https://hub.packtpub.com/9-useful-r-packages-for-nlp-text-mining/
https://elitedatascience.com/python-nlp-libraries

Cheminformatics Libraries

SMARTS chemical string language patterns for chemical pattern searching
Many toolkits like RDKIT, Indigo and commercial software provide a means to perform SMARTS pattern searches. There are several examples of frequently used smarts patterns.
https://www.zbh.uni-hamburg.de/en/forschung/amd/datasets/smarts-dataset.html

Some time-limited editors are also available
https://www.biosolveit.de/SMARTStools/

However, it's harder to encode a query pattern without understanding the underlying logic.
References:
https://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html
https://www.ics.uci.edu/~dock/manuals/DaylightTheoryManual/theory.smarts.html
https://www.molsoft.com/icm/smiles.html

Python Libraries
Starting python code : https://github.com/rdkit/mmpdb, python library: www.rdkit.org
Other visual programming tools that have MMP implementations : https://hub.knime.com/knime/spaces/Examples/latest/99_Community/04_Vernalis/04_Database d_MMP_Example

Software : www.rdkit.org,
http://rdkit.blogspot.com/2018/02/simple-isostere-replacement.html

BioisostereNodeFactory,
https://www.myexperiment.org/workflows/2683.html
https://nodepit.com/node/com.schrodinger.knime.node.bioisostere.BioisostereNodeFactory

Links: https://www.hiv.lanl.gov/content/sequence/PEPTGEN/Explanation.html .

Rules to generate potential bioactive peptides. Generating an original set of peptides negating these rules will give a starting adversarial set.

References:

https://www.researchgate.net/publication/335705093_GANDALF_A_Prototype_of_a_GAN-based_Peptide_Design_Method, https://medium.com/@devnag/generative-adversarial-networks-gans-in-50-lines-of-code-pytorch-e81b79659e3f , https://medium.com/ai-society/gans-from-scratch-1-a-deep-introduction-with-code-in-pytorch-and-tensorflow-cb03cdcdba0f

Python library :https://pypi.org/project/pygan/

Data: http://crdd.osdd.net/servers/avpdb/ , http://pepbank.mgh.harvard.edu/,

Data: http://stitch.embl.de/cgi/network.pl?taskId=MH3cT2DN94Jx,
http://ctdbase.org/detail.go?type=disease&acc=MESH%3aC000657245

Software: Data (csv) graph format import export https://gephi.org/, https://neo4j.com/,
www.cytoscape.org

References: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6249928/,
https://snap.stanford.edu/biodata/datasets/10004/10004-DCh-Miner.html ,
https://www.clinicaltrialsarena.com/analysis/coronavirus-mers-cov-drugs/

References:

J. Wang, J. Chem. Inf. Model. (2020).

D.S. Paul and N. Gautham, Journal of Molecular Modeling **22**, 1 (2016).

K. Vengadesan and N. Gautham, Biophysical Journal **84**, 2897 (2003).

D. Sam Paul and N. Gautham, Journal of Computer-Aided Molecular Design (2018).

Track 3- COVID19 - Moonshot

General Guidelines

- These should be based on non-traditional and non-conventional ideas in search of therapeutics for COVID-19, not covered in other two tracks
- Innovative, out-of the box, non-conventional ideas should be submitted with the Proof of concept (PoC)
- The ideas should be shared as PPT and XLS (strictly in the prescribed format)
- There are no input forms provided for Track 3.

Computing Resources

DD Virtual Tool Room will be made available to the participants requiring computational resources using the resources of CDAC. However, the participants are free to work using their local resources, if they opt for it. They may use CDAC resources only when they need sufficient computational resources for solving problems of the DD Hackathon. The instructions for accessing DD Virtual Tool Room will be provided separately. The usage of CDAC resources is restricted to solving DD Hackathon Problems only.

Other Relevant Resources

Additional resources like databases, CoV2 related datasets, key literature references, screened data for analysis and many more will be provided in the MyGov site.

Submission Instructions

- Online Meeting & Q&A Session
  There will be provisions for Online Meeting and Q&A sessions between the PS Creators/Experts and participants via, e.g.,Facebook Live sessions. Such sessions will be notified later on.

- Submission forms - Common & PS specific Input forms
  The participants will be required to submit their entries using Common and PS specific input Format through Online Forms available in the MyGov site and/or uploading of specifically formatted files as instructed in the Website.