MATH1309/2142 Assessment Task 3

5 Questions, Total Marks = 315, Worth = 40% of final course grade

Assessing Druggability in Drug Discovery: A Bioinformatics Study

(Dataset: Refer to the “Drugbank dataset” excel file, N=400)

Project Description:

- Drug-likeness is not a precisely defined concept in drug discovery. Predicting druggability is of high practical relevance in pharmaceutical research. In vitro absorption, distribution, metabolism, and elimination (ADME) assays are now being conducted throughout the drug discovery process, but there is still a need to develop faster and better analytic methods to enhance the 'developability' of drug leads, and to formalise strategies for ADME assessment of good molecular candidates in the drug discovery and pre-clinical stages.
- This study involves 400 small molecules data retrieved from the DrugBank 3.0 database a unique chem-informatics resource analysed by Hudson et al., (2014, 2017, 2019, 2020).
- The data set contains 9 physico-chemical variables (MW, PSA, log P, Log D, …), and the molecule’s mode of delivery (oral versus non-oral). See Table 1 below.

Table 1:

Molecular Weight (MW)
LogP
HB donors
HB acceptors
Polar Surface Area (PSA)
ROT BONDS
Number of N,O atoms (NATOM)
Rings number (NRING)
Log D

In addition, the data set contains new druggability rules (score functions counting the number of violations for each molecule on each of the 9 variables) developed by Hudson et al. These account for the molecule’s size, permeability etc., but use new cutpoints for each of 9 molecular parameters (Table 2), different to those conventionally used by the Food and Drug Advisory group (FDA) (Lipinski’s rule Ro5, Table 2).

Hudson et al based on the 9 molecular variables (ADME variables) found distinct clusters of the molecules identified as “poor” versus “good” druggables. The data set contains the 9 ADME variables (Table 1), and a scoring function (score9_LogD) along with the molecule’s mode of delivery (oral versus non-oral).

The score is denoted as **score9_ LogD**. Note that the function **score9_LogD** is a continuous variable of range 0 to 9 - comprised of the 4 traditional parameters of the rule of five (Ro5) (Lipinski, 2016) (Table 1) plus 4 extra parameters (PSA, number of rotatable bonds, rings, N and O atoms) with 2 extra candidates lipophicility, log P or logD, the latter is the distribution coefficient, recently suggested as a possible preferable predictor for permeation, preferable to Lipinski’s traditional partition coefficient, Log P, an often used predictor for permeation.

We also dichotomise the score9_LogD_ into 2 groups based on the cutpoint of 4 violations:

- Cutpoint <=4 is a non-violator molecule
- Cutpoint >4 is violator (non-druggable) molecule.

This is equivalent to: Score9_Log D_group <=4 (non-violators) versus Score9_log D_group >4 (violators)

Table 2: Values above the cutpoints score a 1.0

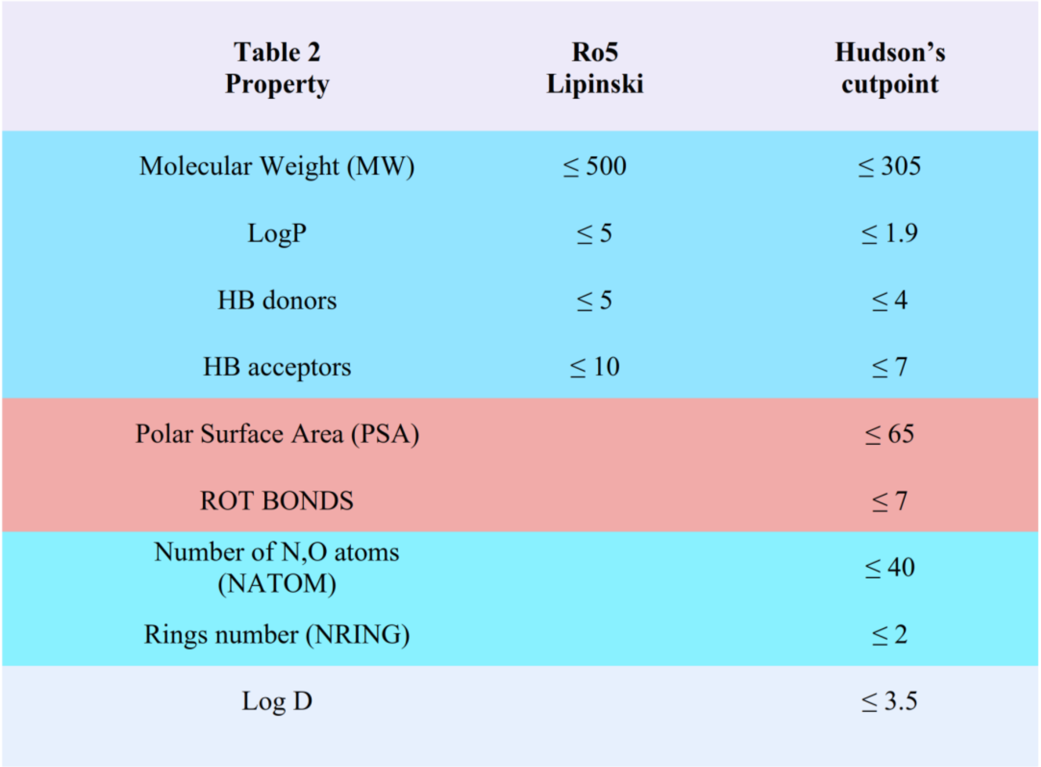
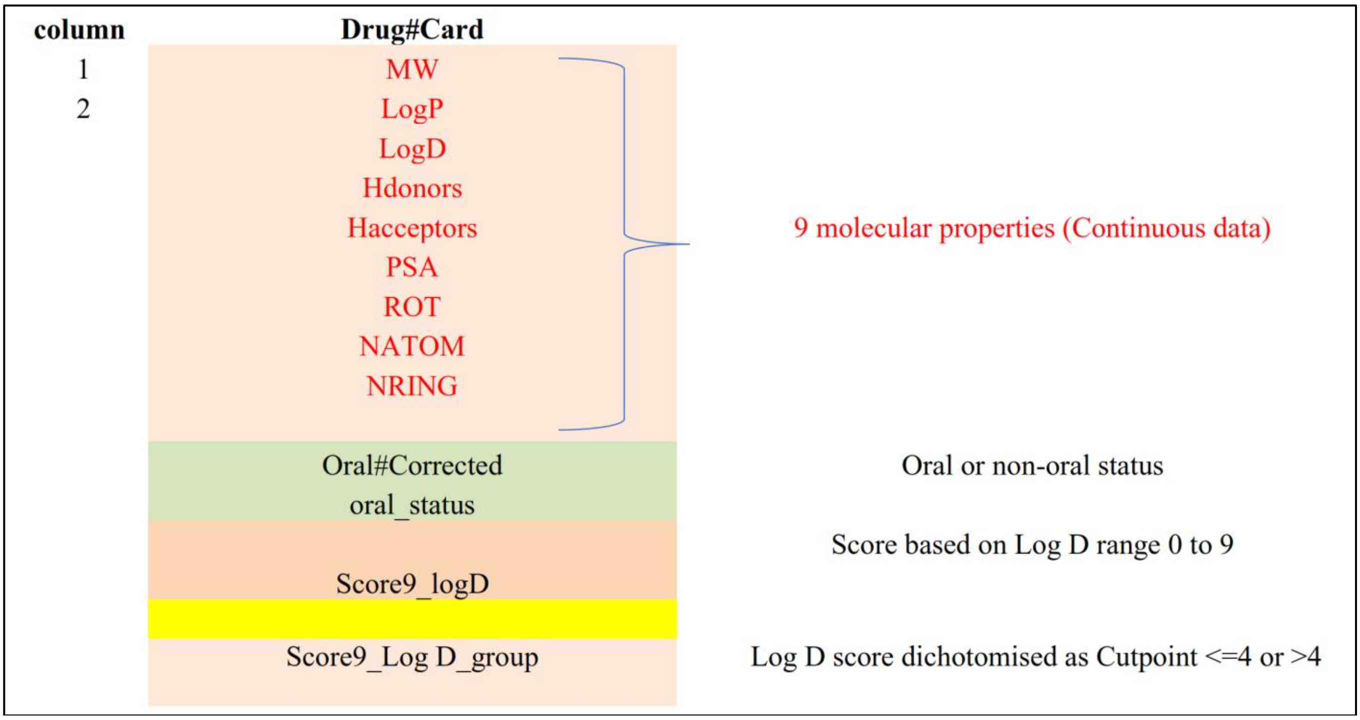
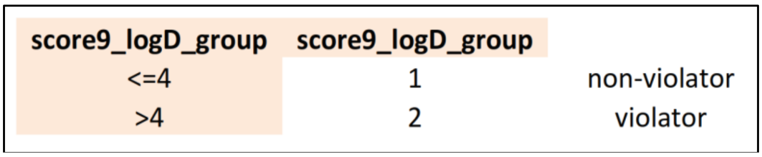


Table 2 Property	Ro5 Lipinski	Hudson's cutpoint
Molecular Weight (MW)	≤ 500	≤ 305
LogP	≤ 5	≤ 1.9
HB donors	≤ 5	≤ 4
HB acceptors	≤ 10	≤ 7
Polar Surface Area (PSA)		≤ 65
ROT BONDS		≤ 7
Number of N,O atoms (NATOM)		≤ 40
Rings number (NRING)		≤ 2
Log D		≤ 3.5

Description of the drugbank dataset of N = 400 molecules:

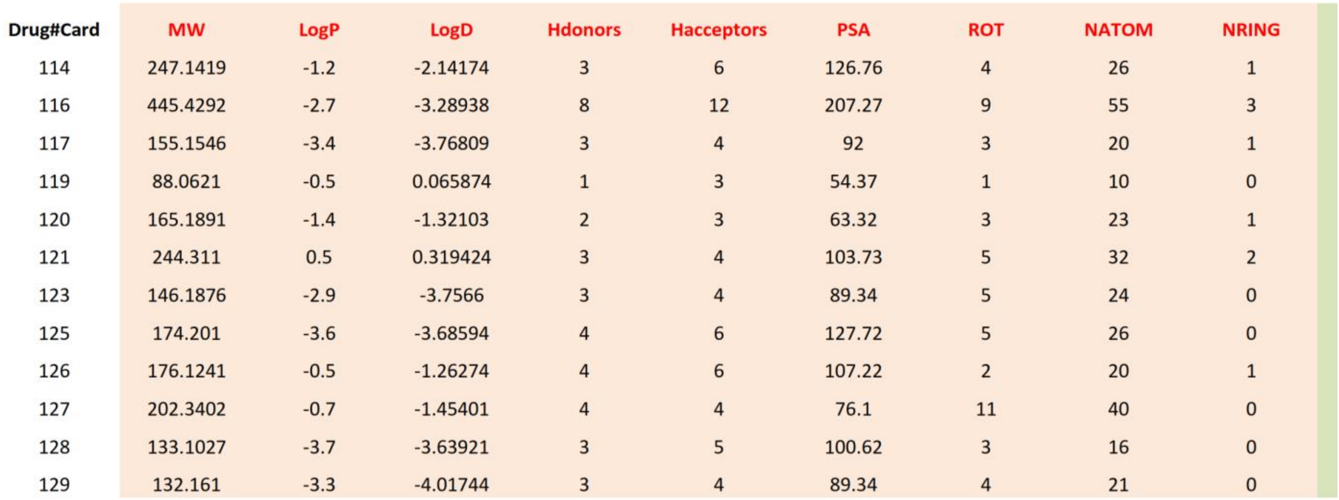


column	Drug#Card	
1	MW	9 molecular properties (Continuous data)
2	LogP	
	LogD	
	Hdonors	
	Hacceptors	
	PSA	
	ROT	
	NATOM	
	NRING	
	Oral#Corrected oral_status	Oral or non-oral status
	Score9_logD	Score based on Log D range 0 to 9
	Score9_Log D_group	Log D score dichotomised as Cutpoint <=4 or >4



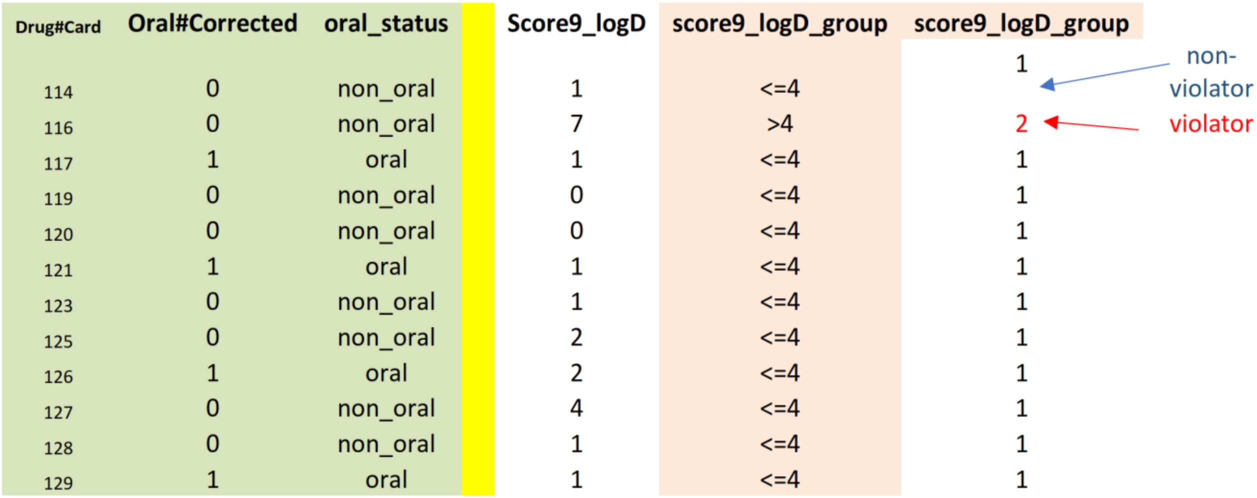
score9_logD_group	score9_logD_group	
<=4	1	non-violator
>4	2	violator

A random sample of 12 molecules' data is shown below as an example (this is not the full N=400 dataset):



Drug#Card	MW	LogP	LogD	Hdonors	Hacceptors	PSA	ROT	NATOM	NRING
114	247.1419	-1.2	-2.14174	3	6	126.76	4	26	1
116	445.4292	-2.7	-3.28938	8	12	207.27	9	55	3
117	155.1546	-3.4	-3.76809	3	4	92	3	20	1
119	88.0621	-0.5	0.065874	1	3	54.37	1	10	0
120	165.1891	-1.4	-1.32103	2	3	63.32	3	23	1
121	244.311	0.5	0.319424	3	4	103.73	5	32	2
123	146.1876	-2.9	-3.7566	3	4	89.34	5	24	0
125	174.201	-3.6	-3.68594	4	6	127.72	5	26	0
126	176.1241	-0.5	-1.26274	4	6	107.22	2	20	1
127	202.3402	-0.7	-1.45401	4	4	76.1	11	40	0
128	133.1027	-3.7	-3.63921	3	5	100.62	3	16	0
129	132.161	-3.3	-4.01744	3	4	89.34	4	21	0

data columns continued...



Drug#Card	Oral#Corrected	oral_status	Score9_logD	score9_logD_group	score9_logD_group	
114	0	non_oral	1	<=4	1	non-violator
116	0	non_oral	7	>4	2	violator
117	1	oral	1	<=4	1	
119	0	non_oral	0	<=4	1	
120	0	non_oral	0	<=4	1	
121	1	oral	1	<=4	1	
123	0	non_oral	1	<=4	1	
125	0	non_oral	2	<=4	1	
126	1	oral	2	<=4	1	
127	0	non_oral	4	<=4	1	
128	0	non_oral	1	<=4	1	
129	1	oral	1	<=4	1	

Question 1: PCA analysis [85 marks]

Perform the following in SAS (ensure to include your code and outputs and interpretations):

a) Perform a principal component analysis using SAS on the **correlation** matrix for the 9 ADME variables. Show your full SAS code and output. (10 marks)

b) Ensure you obtain the following 5 types of plots related to PROC PCA. **(All plots should be placed in clearly labelled Appendices)**. (10 marks)

- Scree plot
- Profile plot
- Component Pattern plots
- Score plots
- Loading Plots

c) Report the eigenvalues and the eigenvectors. (5 marks)

d) What percentage of the total sample variation and cumulative variation is accounted for by **each** of the PCs? (5 marks)

e) Write out the formulation for the PCs. (10 marks)

f) Interpret the PCs via eigenvalues, your component pattern profiles AND your loading plots from SAS. (10 marks)

g) Label your score plot for PC2 versus PC1 by violator and non-violator status and summarise any trends and findings. (5 marks)

h) Label your score plot for PC2 versus PC1 by oral status and summarise any trends and findings. (5 marks)

i) Label your score plot for PC3 versus PC2 by violator and non-violator status and summarise any trends and findings. (5 marks)

j) Label your score plot for PC3 versus PC2 by oral status and interpret any trends and findings. (5 marks)

k) Using BOTH a **formal test of hypothesis** and relevant plots can the data be effectively summarized in fewer than 9 dimensions, k< p? Report k and justify your answer and establish what your k is via the relevant hypothesis test. Show your SAS code and formula. (15 marks)

Question 2: PCA with reduced k < p for plots [40 marks]

Using your reduced dimensionality k determined in Question 1 (k), rerun the PCA on the 9 ADME variables for the violators and the non-violators groups **separately** (where violatory status is delineated by score_9 log D).

a) Recreate the 5 plots related to PROC PCA for your given k.
(All plots should be placed in clearly labelled Appendices) (10 marks)

b) Interpret the PCs via eigenvalues, your component pattern profiles AND your loading plots from SAS based on your reduced dimensionality k and k PC's. (15 marks)

c) Label your score plot for PC2 versus PC1 by oral status and summarise any trends and findings. (5 marks)

d) Label your score plot for PC2 versus PC1 by violatory status, summarise any trends and findings. (5 marks)

e) Which of the k PCs are skewed? Use matrix plots of the PC scores to answer this. (5 marks)

Question 3: DISCRIMINANT ANALYSIS ON 9 ADME VARIABLES BY 2 GROUPS OF MOLECULES [55 marks]

Aim: to run PROC DISCRIM to investigate how the 9 ADME variables discriminate the violators from the non-violators.

a) Generate the means, standard deviations, and variance-covariance matrix of the data for the violators. (5 marks)

b) Generate the means, standard deviations, and variance-covariance matrix of the data for the non-violators (5 marks)

c) Produce the correlation matrix with associated p values, and a matrix scatterplot of the inputted data for the violators. (5 marks)

d) Produce the correlation matrix with associated p values, and a matrix scatterplot of the inputted data for the non-violators. (5 marks)

e) Run SAS DISCRIM and from your resultant outputs answer the following questions. HINT; Use priors: "violators"=0.30 "non-violators"=0.70. Ensure your output is clearly labelled in an Appendix. (10 marks)

f) Is $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ justify your answer based on the appropriate test statistic and output from SAS. (5 marks)

g) How is a molecule with X_0^T = (MW, LogP, LogD, Hdonors, Hacceptors, PSA, ROT, NATOM, NRING) = (445.429, -2.7, -3.28938, 8, 12, 207.27, 9, 55, 3) allocated? (10 marks)

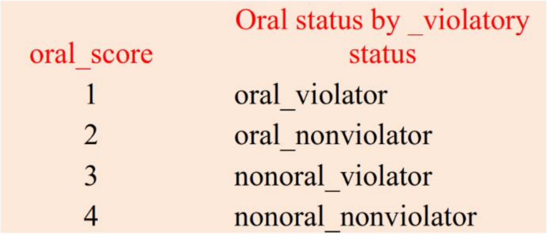
h) Report the LDFs obtained from the output and describe what they mean? (5 marks)

i) Show the resultant confusion matrix **and** interpret it. (5 marks)

Question 4: STEPWISE DISCRIM ON 4 GROUPS OF MOLECULES [90 marks]

Now perform a stepwise DISCRIM using oral by violatory status groups defined below.

a) Create the following variable i.e., an interaction term between oral status and score 9_ Log D violation status at 4 levels as defined below: (5 marks)



oral_score	Oral status by _violatory status
1	oral_violator
2	oral_nonviolator
3	nonoral_violator
4	nonoral_nonviolator

b) Obtain a cross-table in SAS or otherwise of oral by violatory status for the whole data set. How many molecules and percentages are in each of these 4 levels? Along with the table create an appropriate histogram. Interpret your results (10 marks)

c) Generate the means, standard deviations, and the variance-covariance matrix and correlation matrices of the ADME data for each of the 4 levels defined by the Oral status by_violatory status variable. Interpret your descriptive profiles in terms of how the variables differ across the 4 levels. (20 marks)

d) Generate matrix plots of the 9 ADME variables for the 4 levels defined by the Oral status by_ violatory status variable. Interpret how the variables differ in distribution, correlation, across the 4 levels. (15 marks)

e) Run a STEPWISE DISCRIM analysis using the 9 ADME variables (Table 1) as the input and the above 4 level grouping variable, Oral status by_violatory status. (25 marks)

f) Which variables best discriminate the 4 Oral status by_violatory status classes? (5 marks)

g) Give the mean, variances and correlations between these best discriminating variables across the 4 level Oral status by_violatory status variable and interpret trends. (10 marks)

Question 5: [45 marks]

a) Run a STEPWISE DISCRIM analysis using your subset of k **PCs** from Question 2, now as the input variables and the above 4 level grouping variable, Oral status by _violatory status. (20 marks)

b) Which **PC variables** best discriminate between the 4 oral by violatory groups/classes? (5 marks)

c) Give the mean vector, variance-covariance matrix and correlations between these chosen PCs variables for each of the 4 oral by violatory groups/classes and interpret trends. (10 marks)

d) For the PC variables selected by the stepwise discriminant analysis determine the correlation between them and the original data (i.e., the 9 ADME variables in Table 1). (10 marks)

THE END