Assignment 02

Assignment Name:	**Week 2 Assignment - Descriptor analysis of problem sets**
Weight:	6.25% of final grade
Due Date:	Tuesday, Midnight
Associated Learning Outcomes:	6. Recognize and calculate basic physicochemical properties of small molecule compounds. 7. Describe the underlying basis for cheminformatics techniques, and apply them to compare small molecule compounds. Applications for drug discovery.
Assignment Prompt:	**Problem 1 - Data Analysis.** Download the fish toxicity dataset from the course, which contains 87 structures in smiles format and toxicity values for a particular fish. Higher LC50's indicate less toxicity. Investigate if there is any relationship between the toxicity observed and the following descriptors: molecular weight, number of rotateable bonds, number of aromatic bonds, ClogP (octanol/water partition coefficient), and TPSA (topological polar surface area). Other descriptors should be explored and discussed if they are relevant. **Hint:** LC50 is a continuous variable, but it may be more informative to "bin" continuous variables into categories such as "non-toxic vs toxic, or high, medium and low toxicity. Then you can graph the groups of compounds separately, and compare properties between groups. Turn in A) Your analysis and conclusions. (**6 points**) Are there any hypotheses you can make about the descriptors as to how they pertain to physical phenomenon? (i.e. are "greasy" compounds more toxic?) B) An image of your knime workflows and supporting graphs and graphics (**4 points**), with captions. A complete analysis is likely to include box plots, and histograms. Use different color fills to illustrate toxic vs non-toxic so that we can see the differences between groups.

Problem 2 - Simple Model Building: The blood-brain barrier (BBB) penetration ability of a compound is an important consideration in drug design. Drugs that act on targets in the central nervous system need to pass through the BBB, while a compound that acts on a peripheral target is generally safer if it does not pass the BBB. In this exercise, we'll look at how physiochemical properties and descriptors correlate with the BBB penatration ability of a compound.

Download the blood-brain barrier compound dataset on the course site. The BBB penetration ability is indicated as positive (p) or negative (n). Convert the smiles file to an SDF file using OpenBabel, either in Knime or separately. You can obtain OpenBabel from http://openbabel.org. Construct a report on this data that includes the following sections:

1) Exploratory data analysis (**5 points**)
    a) Summarize the properties you feel relevant to achieving either penetrant (p) or nonpenetrant (n) compounds. Are there any relevant cutoffs that seem to materialize in the properties to discriminate?
    b) Provide histograms or density plots relevant to the discussion, in a similar manner to above
    c) Provide any relevant code or workflow screenshots as captioned figures
2) Simple Model Building (**10 points**)
    a) Discuss 3 (or more) properties from above of your choice, and describe how they discriminate from penetrant or non-penetrant, and any caveats that you can think of. Pick a cutoff (decision boundary) for these 3 properties. Generate a simple model using one, two, three or more properties to separate permeable from non-permeable compounds (e.g. Property A > 4, Property B < 3, Property C > 0.5)
    b) Run your model against the dataset (**Hint**: Use the rule engine node). Provide a Confusion Matrix for your model and analyze performance.
    c) How can you make your model better? What happens to performance if you change the model? Are there any specific examples of false positives or false negatives that are instructive?

	Note: Please **<u>do not</u>** try to build flexible function (machine learning) models in this section. The idea here is to understand simple correlations first. Remember, the simple model is often the best model.
Context/Purpose:	In this exercise we introduce numerical interpretation of small molecules relative to a property we care about. The concept of model-building and evaluation is introduced, and we apply it to a practical example.
Requirements and Logistics:	Students should turn in a typed word or google docs document converted to PDF. All figures should be pasted into the document to generate a one-file submission. Handwritten materials will not be accepted.

Assignment Rubric

	Excellent	**Acceptable**	**Incorrect / Absent**
Assignment Questions	Questions are answered and workflows screenshots are provided. **Full point value**	Partially correct, or supporting workflows are not provided. **Partial Point Value**	Incorrect, or not answered **No Point Value**

Late policy - **5 point penalty per day late - More than 3 days late: No credit**