# 1 By Noureddin

1. I have split the dataset from Dundee into 4565 datasets using the column TID (Target ID)

2. Some datasets are too small ($< 10$ instances) and some have thousands of instances!

3. Some stats:

   (a) No of DS with >= 2000 instances = 74

   (b) No of DS with >= 1500 instances = 126

   (c) No of DS with >= 1000 instances = 234

   (d) No of DS with >= 500 instances = 506

4. We have a large list of metrics from WEKA. However, not all are suitable for QSAR studies

5. There are some domain specific metrics (specific for QSAR) and they're not in WEKA

6. Some of these domain specific metrics include: pROC, BEDROC (for ROC curve, they focus on what they denote *early recognition problem*) and RIE

7. Another metric is called SLR from this paper[1]

8. Matthew's Correlation Coefficient is also popular[2]. It exists in WEKA

9. There is a paper about *Evaluating Virtual Screening Methods*[3]. It looks VERY interesting. Maybe I will read it after I come back. Please go through it if you have time!

10. Some Metrics from[4]:

    (a) the area under the receiver operating characteristic curve (ROC)

    (b) the area under the accumulation curve (AUAC)

    (c) the average rank of actives

    (d) the enrichment factor (EF)

    (e) the robust initial enhancement (RIE)

    (f) Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC)

[1] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2722655/?tool=pmcentrez&report=abstract

[2] http://journal.chemistrycentral.com/content/2/1/3

[3] http://www.ncbi.nlm.nih.gov/pubmed/17288412

[4] http://www.ncbi.nlm.nih.gov/pubmed/17288412

11. [Noureddin: Interesting Presentation] here[5], NS $---$ > MUST go through it to develop better understanding of Virtual Screening/QSAR

12. [Noureddin: Structural Variability of Datasets.] This was suggested by Ross during online meeting on Fri 9 May 2014. He mentioned using Manhattan distance (Crina mentioned Hamming distance)

## 2 By Larisa

1. We have analysed available descriptors from various sources (bottom-up approach), and there are too many of them. We now will try top-down approach. We will consider several datasets, from Dundee and publicly available ones, e.g. from ChEMBL [Noureddin: I think people from Dundee always use ChEMBL] and also several QSAR studies. We will try to annotate them with some descriptors that would convey what is important to record about those datasets and QSAR- specific studies. We then will come up with a set of descriptors that are suited for our task. We will check if what existing resources have the required descriptors, and if necessary we will define new ones (see the next point).

2. Crina suggested that we can develop new descriptors/ measures that fit better to support our task.

3. We will target to come up with ˜50 descriptors and we will evaluate them with experts through questionary and select  20 the most popular descriptors.

4. We will also put our questionary to a public domain and give an opportunity to everyone to comment on descriptors. Such an approach would ensure that we identify a set of useful descriptors [Noureddin: I emailed Egon Willighagen[6] but he has not responded yet].

## 3 Crina

My two suggestions during the meeting where:

- Find the characteristics of the data set (something which defines very well our data and makes the difference between our data and other data sets). For instance, if we say 10 features, this will not make a difference. But if we say 10 features, first 3 are real values, 4 and 5 are binary and 6-10 are real values between 0-1, this will be more specific.

[5] https://www.ebi.ac.uk/training/sites/ebi.ac.uk.training/files/materials/2013/131209DrugDiscovery/1_-_val_gillet_-_ligand-based_and_structure-based_virtual_screening.pdf

[6] http://chem-bla-ics.blogspot.co.uk/

- define the problem output (or scope): in our case will be binary classification and not regression?[Noureddin: they usually use both but at the end they use a threshold to decide whether the real value resulting from regression indicates compound is active or inactive - and this makes it binary classification]
- define how we are going to achieve this scope: this may involve more criteria in which case we will need to look at the performance of the methods we apply from various angles
- define a set of machine learning methods suitable for this scope [Noureddin: popular ones are Naive Bayes and Random Forests, I need to investigate further]
- define a set of measures or metrics which evaluate these machine learning methods for this scope. Some of the metrics may be standard, but we might need some other aspects and for this we will have to define ourself some measures

[Crina: Noureddin, put together what you have so far and what you find in the literature strictly for this type of problem and I can also look after that and complement with what I know.]