# REPORT ASSIGNMENT 2

CHEMINFORMATICS WINTER SEMESTER 2022-23

**Stephan Liu**
Universität Tübingen
stephan.liu@student.uni-tuebingen.de

November 7, 2022

**ABSTRACT**

KNIME, also called Konstanz Information Miner, is an open-source data analytics program with the goal to advance the impact and understanding of data science through visual programming. [1]Thus, this program makes abstract tasks easier to understand as the problem is visualized. The graphical user interface is simple and clear and makes operations fairly user friendly.

## 1 Introduction

The importance of data and data sets cannot be neglected in science. The goal is to process these data in such a way that noise is minimized and still representative enough to be useable. The sheer volume, variance, and dimensionality of the data available make it difficult for the individual to identify, extract and interpret them correctly, which increases the importance of data analysis and statistical learning significantly. [2] Data analytics and machine learning simplify and advance our understanding in our endeavor to work with data.

Natural to say as the amount of data increases it gets gradually harder to keep clarity over the data and harder to reprocess the data. Accordingly, there are data processing programs that are used to maintain better uniformity. Besides well-known software like Rapidminer, IBM SPSS statistics and SAP analytics cloud, there are free open-source data analytics like KNIME that provide easy access for university purposes. In the field of science KNIME can be used to process large amounts of raw data such as a high-throughput-screening (HTS) produces.

## 2 Material and Methods

For the first part of the task three chemical structures were provided and their free ion pairs, as well as the hybridizations of substructures a) and c), were displayed.
To analyze the HTS data a KNIME workflow was created.
The 500 CSV files produced by the HTS contains raw data that are all read in into one single table. All screening plates are then visualized using heat maps and checked for their quality using the Z-prime factor considering the positive and negative controls.
The plates then were divided into plates of good quality plates with a Z-score of 0.5 to 1.0 and plates of bad quality with a Z-score of 0 to 0.5. Hit picking using the normalized percent inhibition (NPI) normalization with a $NPI \geq 70\%$ was used for the good plates. Hit picking using percent of control (POC) with a $POC \geq 60\%$ was used for the bad plates.

The data of the picked hits then was merged again and joined with the screening library file. The matches between the compound data and screening library file then were identified and its properties filled into a new CSV file.

[1] https://en.wikipedia.org/wiki/KNIME
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8511823/

## 3 Results

Figure 1 shows the chemical structures with free ione pairs and their hybridizations.
As for structure a) the upper all C-atoms have a sp2-hybridization. Both N-atoms have one free ione pair and therefore a sp2-hybridizations. The lower ring has the same properties as the upper structure with the difference that the NH+ has no free ione pairs but an sp3-hybridization.
Structure b) has one free ione pair at the central C-atom as it has only three bonds to its neighbors.
Structure c) shows the upper molecule with two O-atoms that have their two free ione pairs drawn in. They both have an sp2-hybridization.
Meanwhile the lower substructure contains an O-atom with a sp2-hybridization and two free ione pairs and another O-atom with a sp3-hybridization and three free ione pairs.

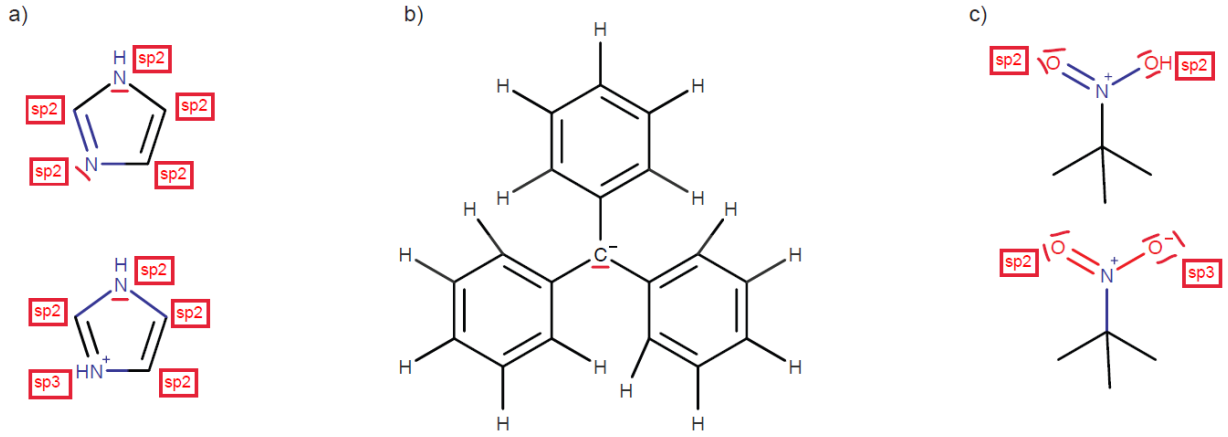


Figure 1: Chemical structures with free ione pairs and hybridizations.

Figure 1 shows the heat map that is generated using HCS tools plate heat map viewer therefore every plate shows what can be considered good or not within the plates itself in column one.

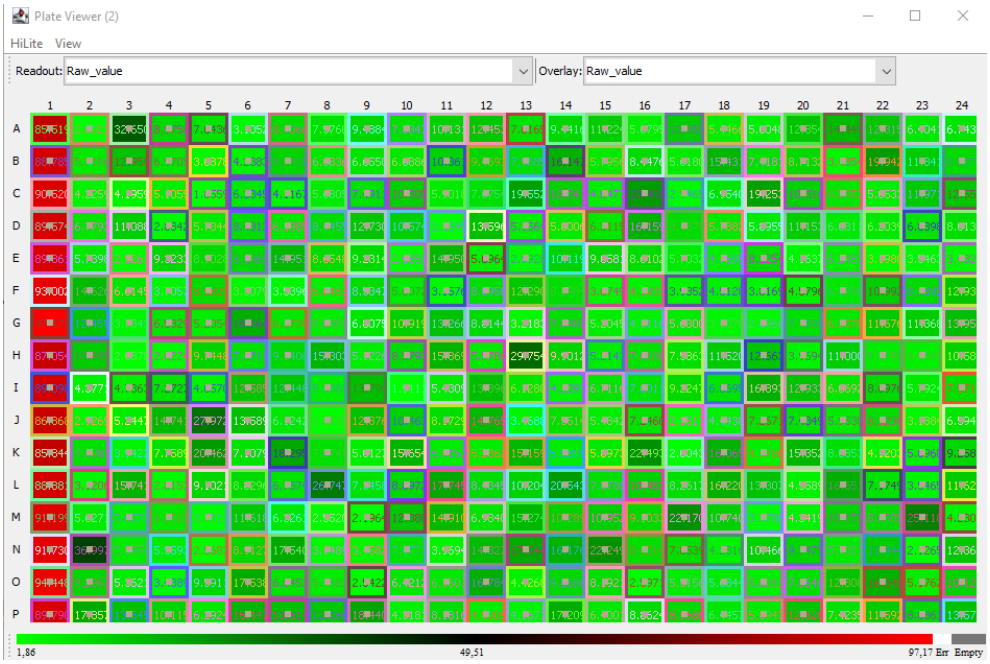


Figure 2: Heat map of plate 2 shown exemplary.

Figure 3 shows the output after merging the compound data with the screening library file and running through the node "Molecule to CDK" which allows to convert the SMILES properties to visualized molecular structures.

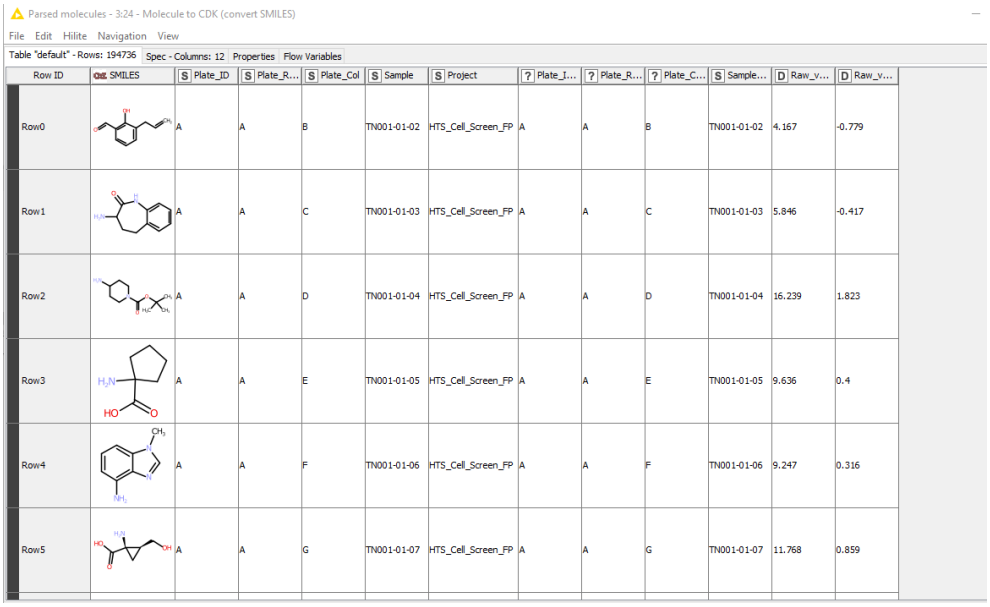


Figure 3: Table with converted SMILES structures.

Figure 4 shows the output that is finally saved into the CSV file. It contains the header (1)sample ID, (2)molecular weight, (3)number of h-bond donors, (4) number of h-bond acceptors and (5)SlogP. It reports the chemical properties of the compounds of the screening library that are matching with the sample ID of the hits of the plates respectively.

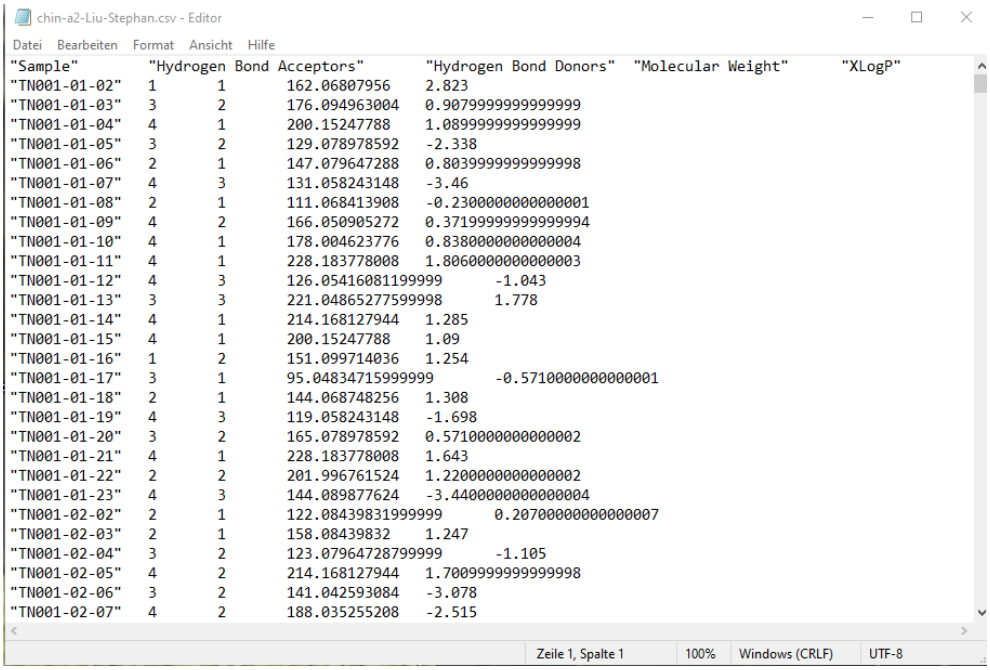


Figure 4: CSV file with the chemical properties of the matching compounds.

## References

[1] Berthold, Michael R. and Cebron, Nicolas and Dill, Fabian and Gabriel, Thomas R. and Kötter, Tobias and Meinl, Thorsten and Ohl, Peter and Thiel, Kilian and Wiswedel, Bernd. KNIME - the Konstanz Information Miner: Version 2.0 and Beyond *SIGKDD Explor. Newsl.*, 2009.