

Example Application of the MSC for Machine Learning Evaluation

Jan-Mathis Hein



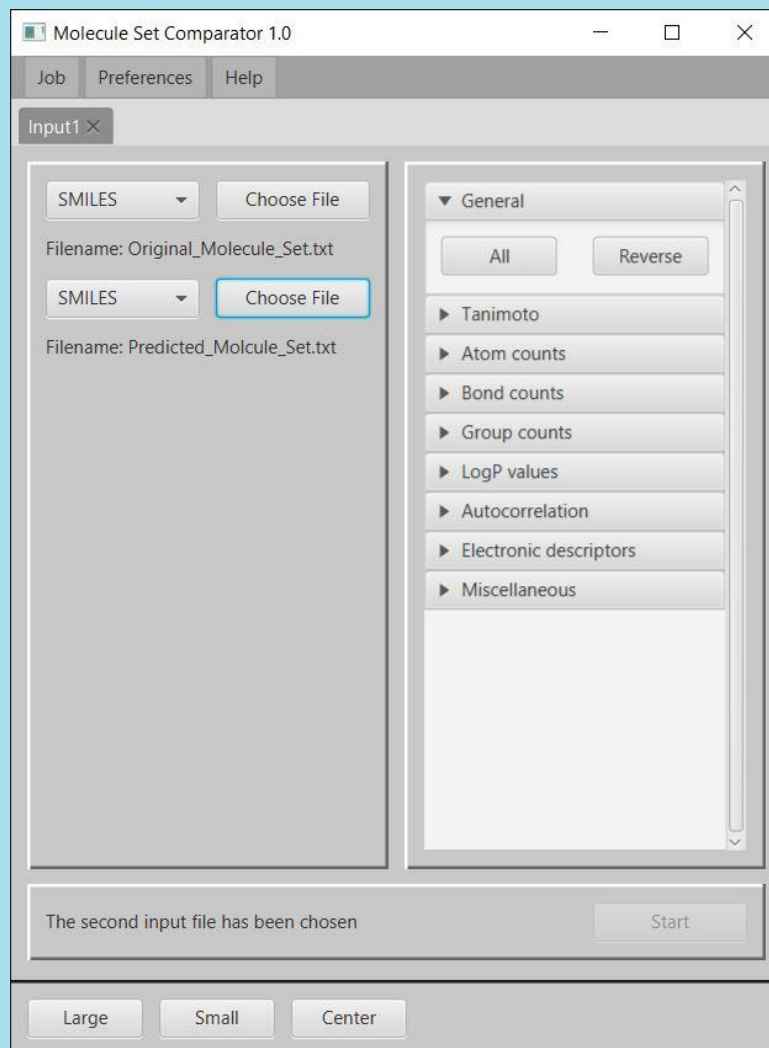
**Westfälische
Hochschule**



**FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA**

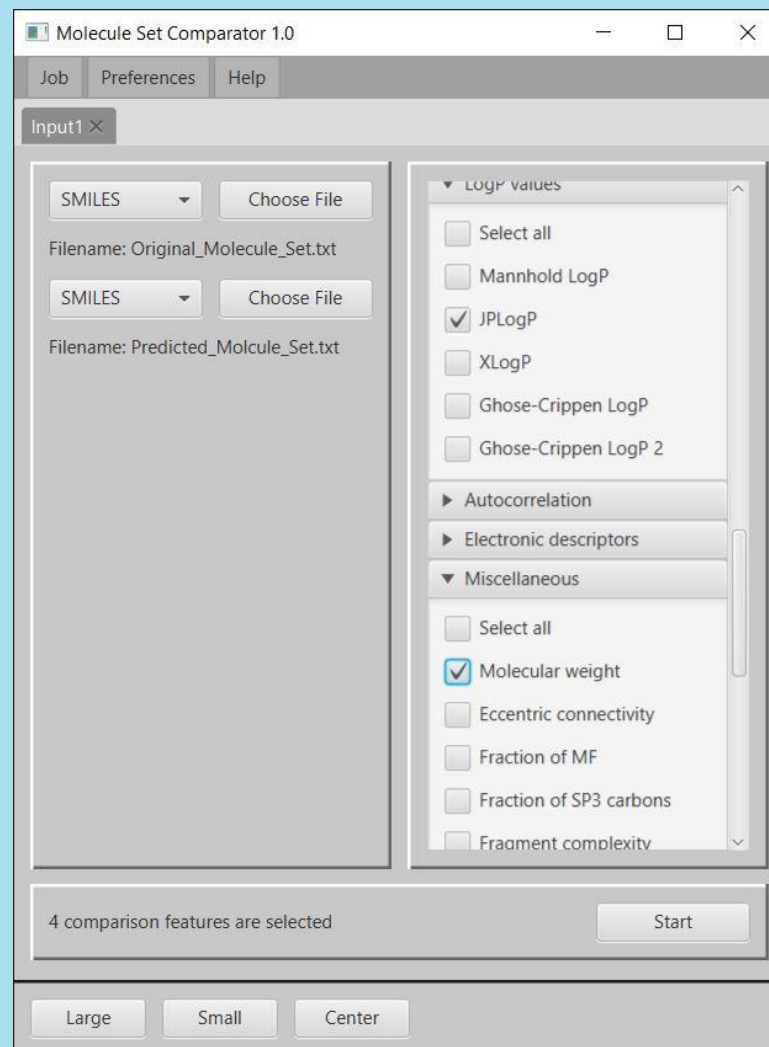
Initializing the job

- In this tutorial an exemplary application of the MSC for the evaluation of a machine learning algorithm is showcased
- First the application is started, and a new job is created
- Then the original and predicted molecule sets are loaded into the MSC. The original set was presented in some form to the machine learning algorithm. The algorithm then outputted the predicted set.
- Both sets are coded in the SMILES format



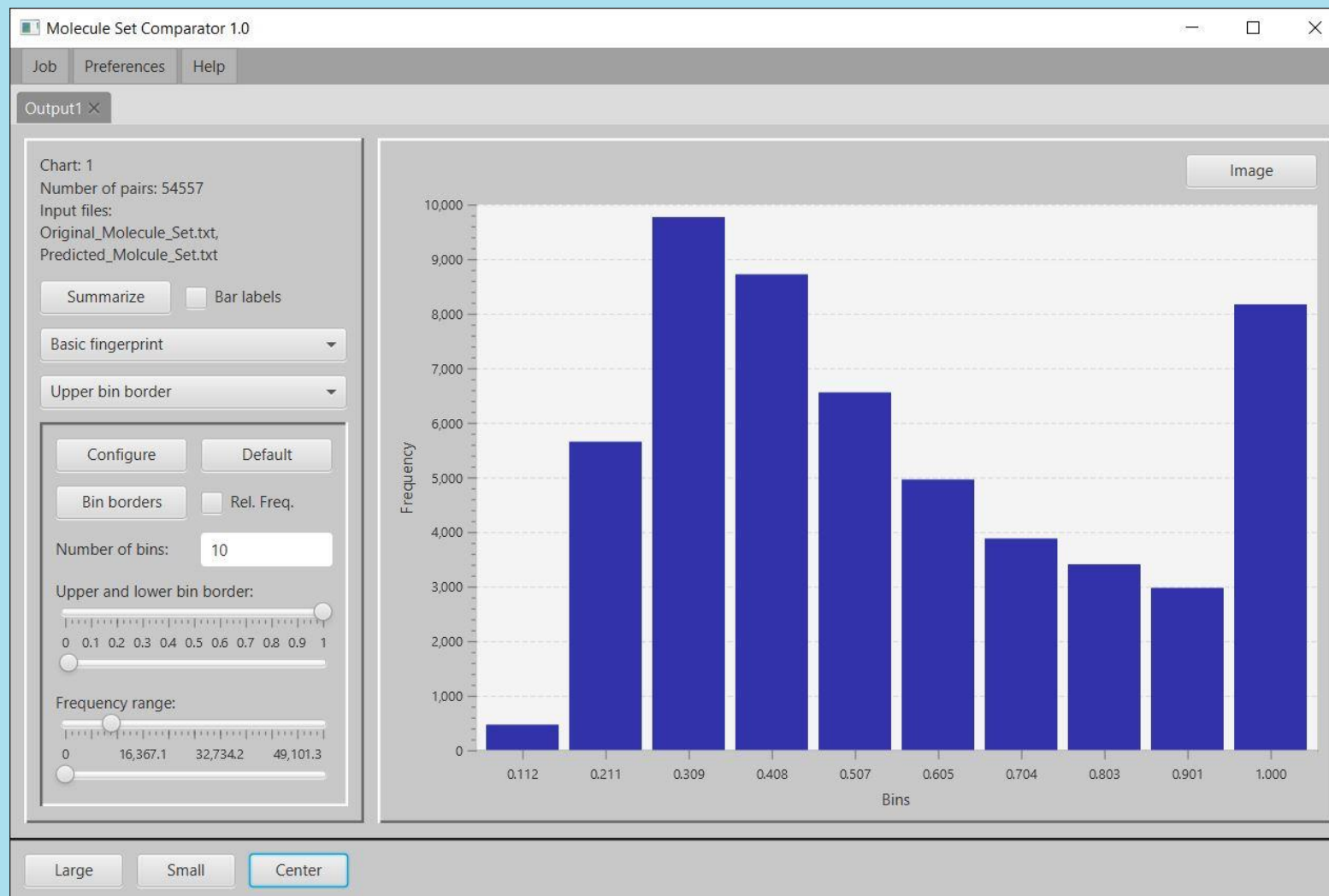
Initializing the job

- Next the molecular descriptors are selected based on which the two sets will be compared.
- These are the Tanimoto coefficient with the basic CDK fingerprint, the atom count, the JPLogP value and the molecular weight. For the last three each pair is compared based on the difference of the two descriptors.
- Then the job is started, and the two sets are compared pairwise (first original molecule to first predicted molecule and so on)



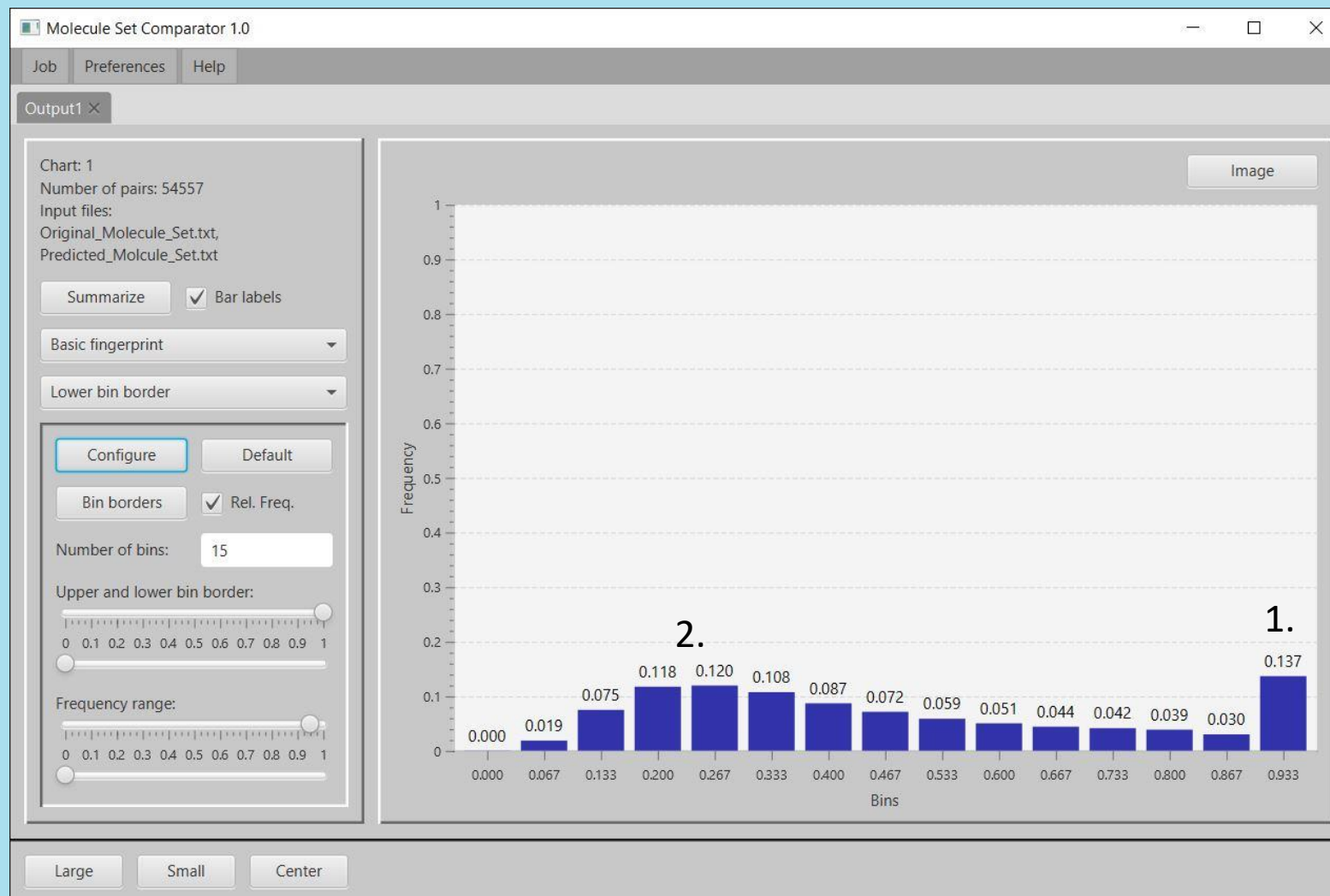
Analyzing the results

- After a few seconds, the execution of the job has finished and the output tab with the results is displayed
- The histogram shows the absolute frequencies of molecule pairs whose Tanimoto coefficient lies in a specific interval. On the x-axis the upper border of each interval is shown.
- The first step in analyzing these results is to configure the histogram for an easier comprehension



Analyzing the results

- Instead of absolute frequencies relative ones are now used
- Every bar is labeled with its frequency
- The number of intervals and the y-axis range were increased
- The x-axis labels now display the lower border of each interval
- Now the data can easily be analyzed. Around 13.7% of the original set were predicted very well by the algorithm (1.). But many molecules were predicted poorly (2.)



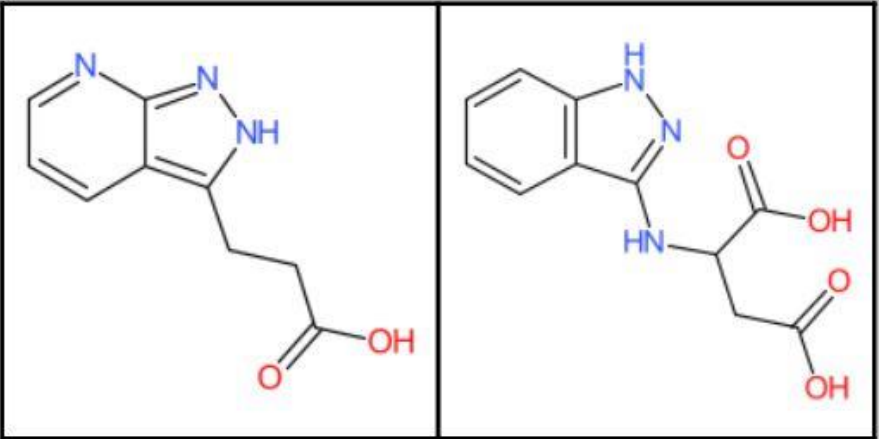
The detail window

- The poorly predicted molecules can be examined more closely by clicking on one of the bars
- On the right the detail window for the fifth bar is shown. Here one can browse through the list of input-output-pairs of this bin
- It is immediately visible that the first molecule pair doesn't seem that different. This intuitive rating contradicts the low Tanimoto coefficient of 0.29

Molecule pairs of bin 5

Used ComparisonFeature: Basic fingerprint
Number of displayed pairs: 6534
Frequency value of this bin: 0.1198
Lower bin border: 0.2667
Upper bin border: 0.3333
SMILES of the first molecule: OC(=O)CCC1NN=C2N=CC=CC=12
SMILES of the second molecule: OC(=O)C(NC1=NNC2=C1C=CC=C2)CC(O)=O
Comparison result: 0.293956

☒ Basic fingerprint
☐ LINGO fingerprint
☐ Extended fingerprint
☐ E-State fingerprint
☐ PubChem fingerprint
☐ Shortest path fingerprint
☐ Substructure fingerprint



|< < 1 > >|

L list L image
R list R image

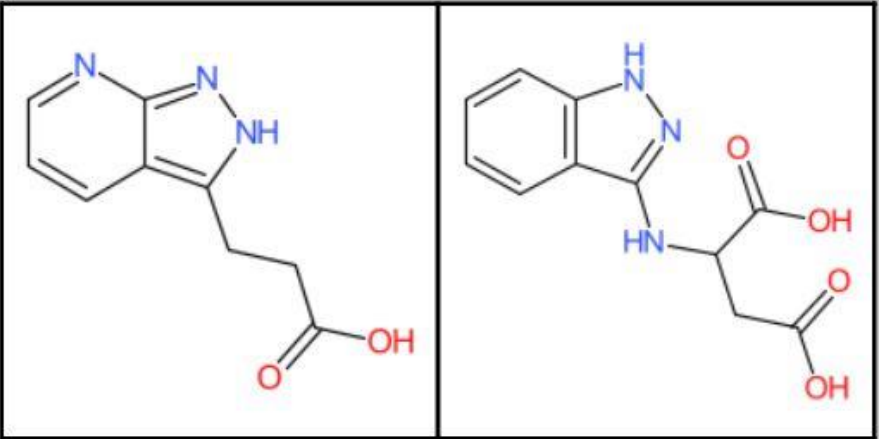
The detail window

- This perceived contradiction can be solved by calculating additional Tanimoto coefficients with different fingerprints
- For example, the PubChem fingerprint yields a value of 0.72 which seems a lot more reasonable.
- The Tanimoto coefficient is very dependent on the fingerprint and in another job another fingerprint could be used to get a more realistic result

Molecule pairs of bin 5

Used ComparisonFeature: Basic fingerprint
Number of displayed pairs: 6534
Frequency value of this bin: 0.1198
Lower bin border: 0.2667
Upper bin border: 0.3333
SMILES of the first molecule: OC(=O)CCC1NN=C2N=CC=CC=12
SMILES of the second molecule: OC(=O)C(NC1=NNC2=C1C=CC=C2)CC(O)=O
Comparison result: 0.293956

<input checked="" type="checkbox"/> LINGO fingerprint	0.56
<input checked="" type="checkbox"/> Extended fingerprint	0.27
<input checked="" type="checkbox"/> E-State fingerprint	0.83
<input checked="" type="checkbox"/> PubChem fingerprint	0.72
<input checked="" type="checkbox"/> Shortest path fingerprint	0.32
<input checked="" type="checkbox"/> Substructure fingerprint	0.93

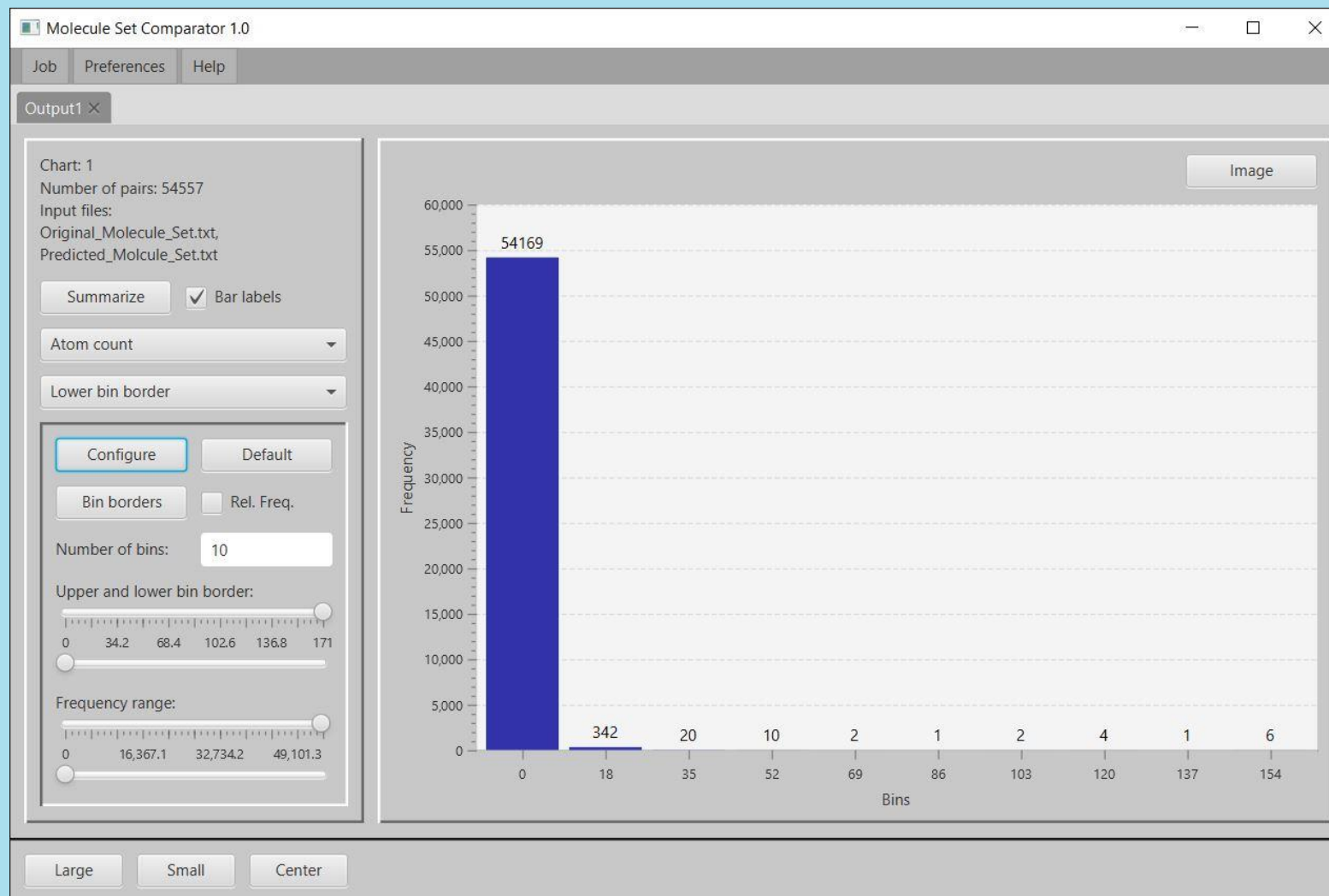


Navigation: |< < 1 > >|

Buttons: L list, L image, R list, R image

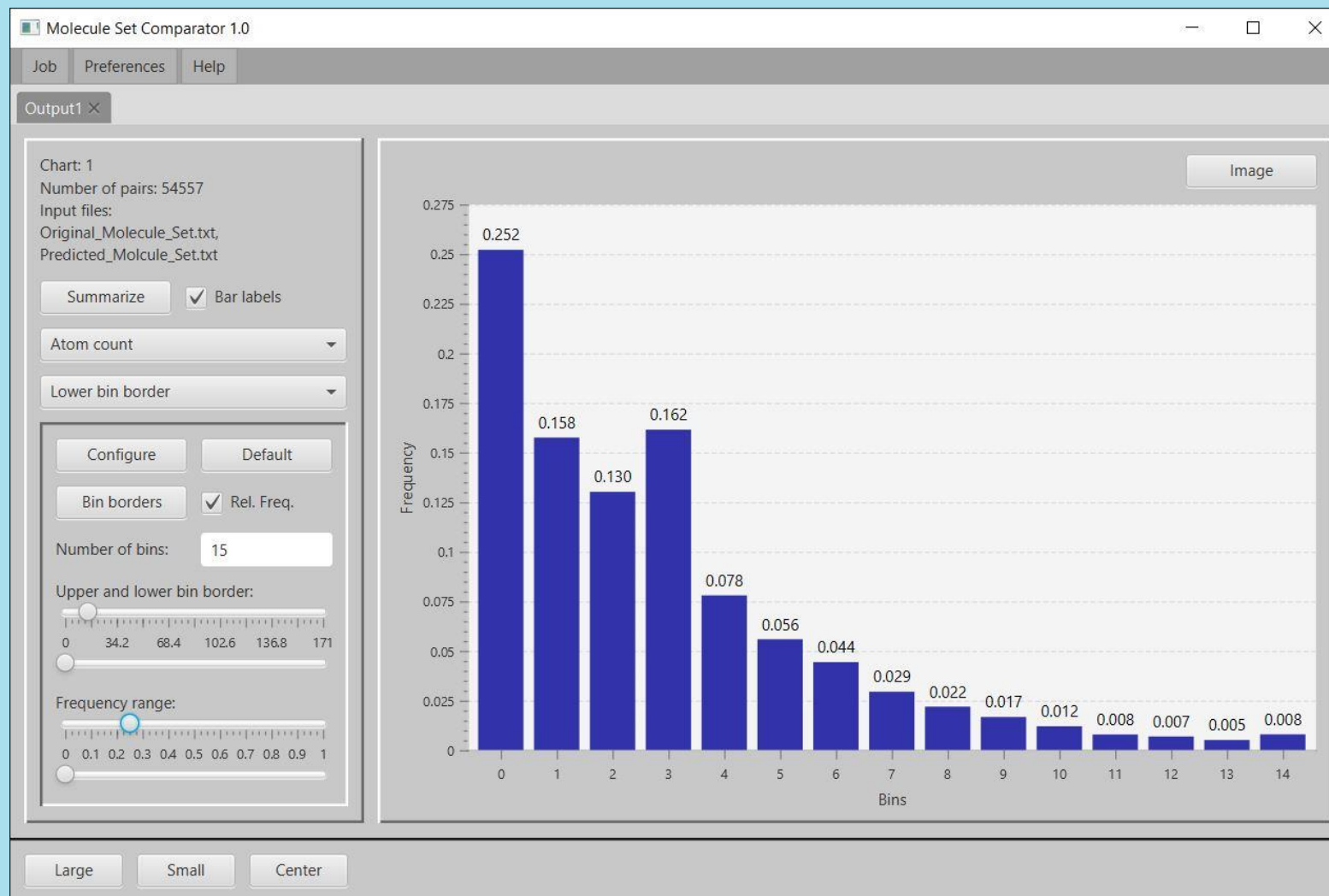
Other descriptors

- In the output tab a fast switching between the different calculated descriptors is possible
- Here the histogram of the atom counts descriptor is shown.
- But without configuration the histogram is very uninformative.



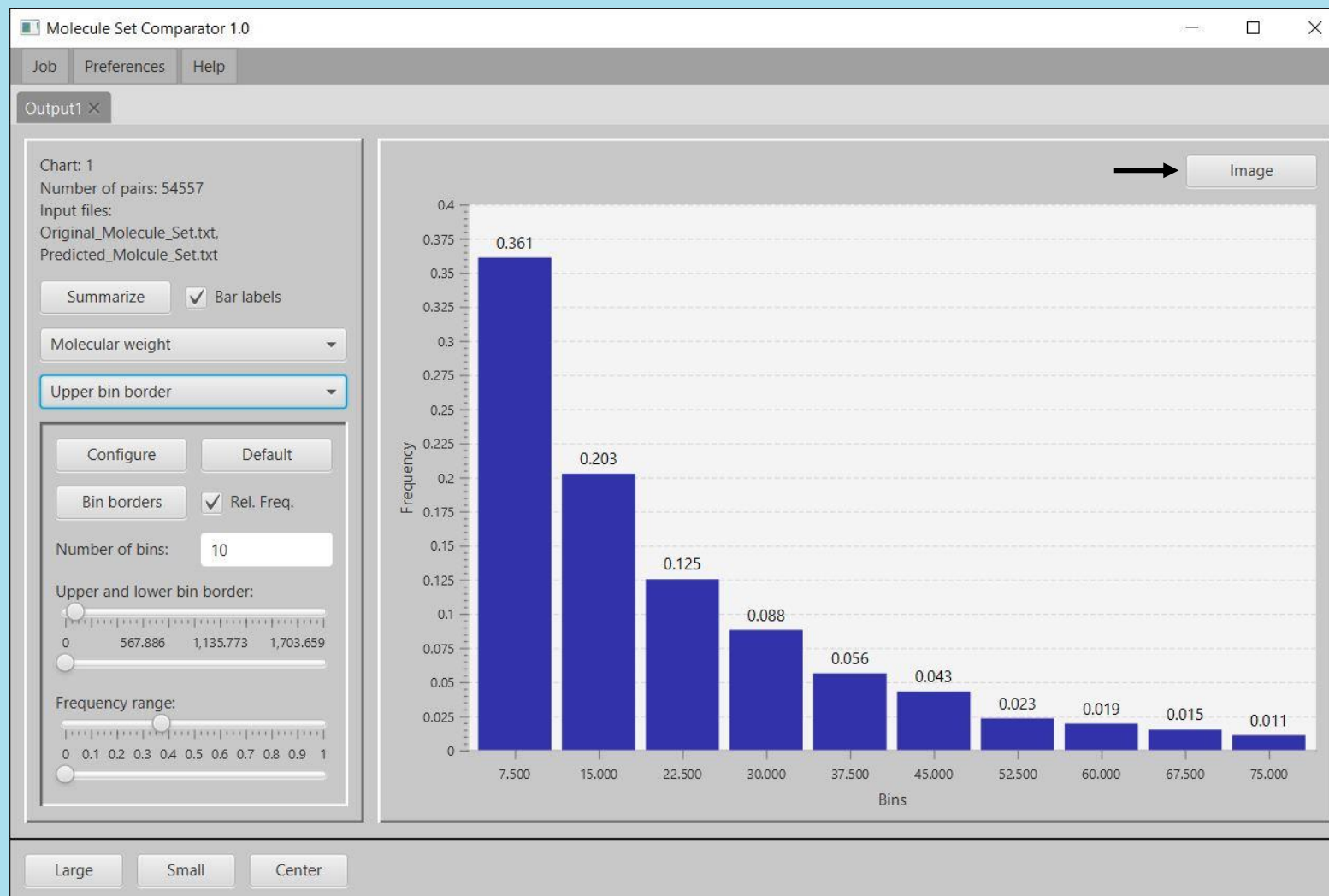
Other descriptors

- Now the histogram conveys a lot more information.
- Around 25% of the molecules were predicted with the right number of atoms
- The histogram shows a “nice” distribution except for the 4th bar.
- But with the help of the detail window it can be concluded that this anomaly is likely caused by the addition of one C atom which also leads to two additional H atoms



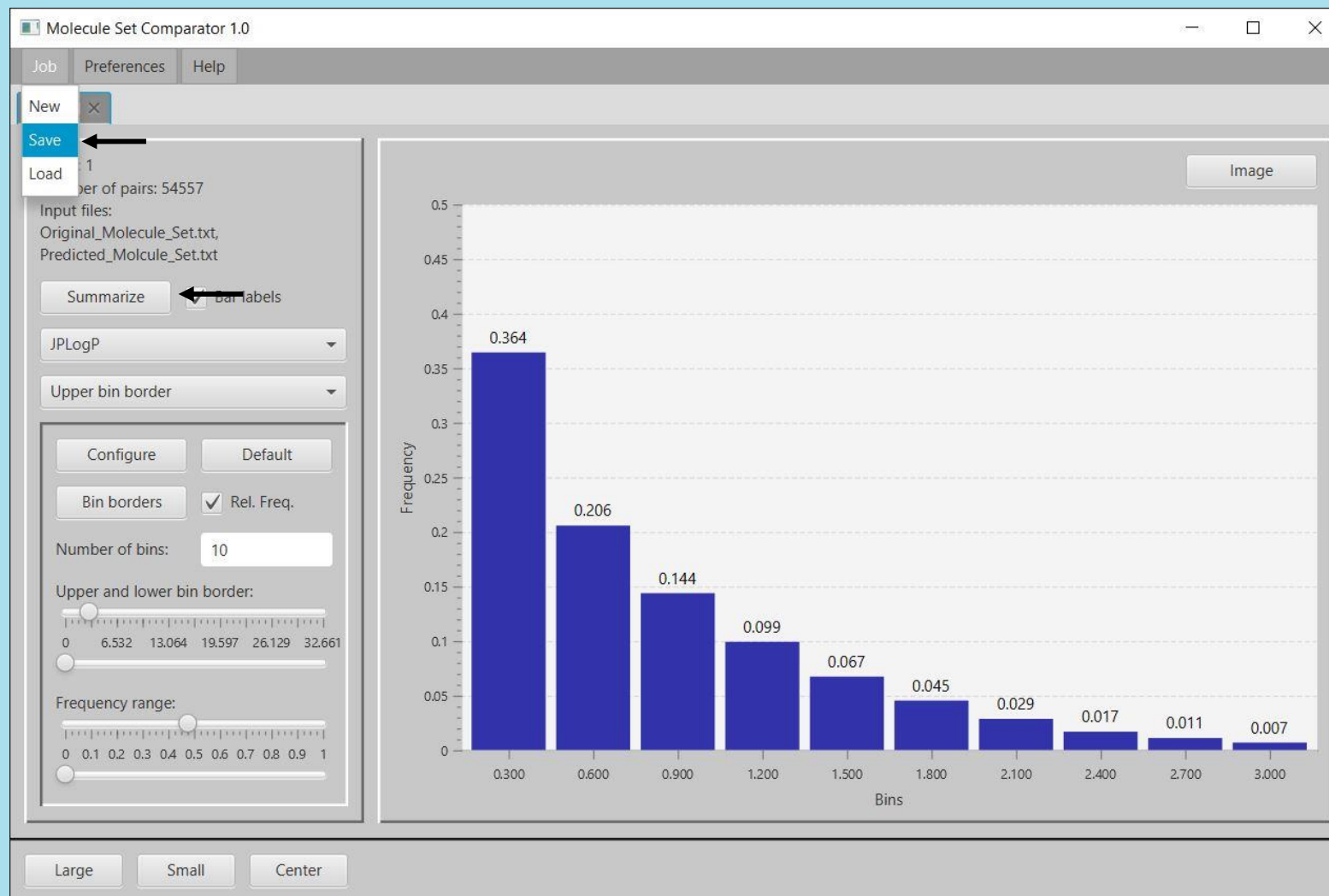
Other descriptors

- The configured histogram of the molecular weight descriptor also exhibits a “nice” distribution without anomalies
- Roughly 36% of the predicted molecules have a similar molecular weight that only differs by about 0 to 7.5 daltons
- With the **Image** button the histogram can be exported into various file formats. This can be used for the sharing of results



Other descriptors

- The configured histogram of the JPLogP value can be seen on the right
- It shows a distribution that is very similar to the one of the molecular weight histogram
- This connection could be worthy of a more thorough exploration with e.g. the detail window
- With the **Summarize** button a documentation text file can be generated
- Also the whole job output can be saved
- Both these options enable an easy sharing of results



Summary

- The goal was to evaluate a machine learning algorithm that got the original molecule set (in some form of representation) as input and predicted for each input a molecule (the predicted set)
- For the evaluation four descriptors were consulted (Tanimoto coefficient, atom count, JPLogP value, molecular weight)
- Based on the histograms a fast assessment of the predictive power of the algorithm was possible. Some molecules were predicted pretty good, but the majority were predicted poorly. So there definitely is room for improvement.
- Also a new evaluation based on Tanimoto coefficients with other fingerprints is advised and a more thorough evaluation of the apparent coupling of the JPLogP and molecular weight descriptors is an interesting possibility.
- The obtained results can be shared in different ways with colleagues to discuss the next actions