MSC

Molecule Set Comparator

An example to use MSC to analyse the predictions from a machine learning model

Jan-Mathis Hein

Westfälische Hochschule

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

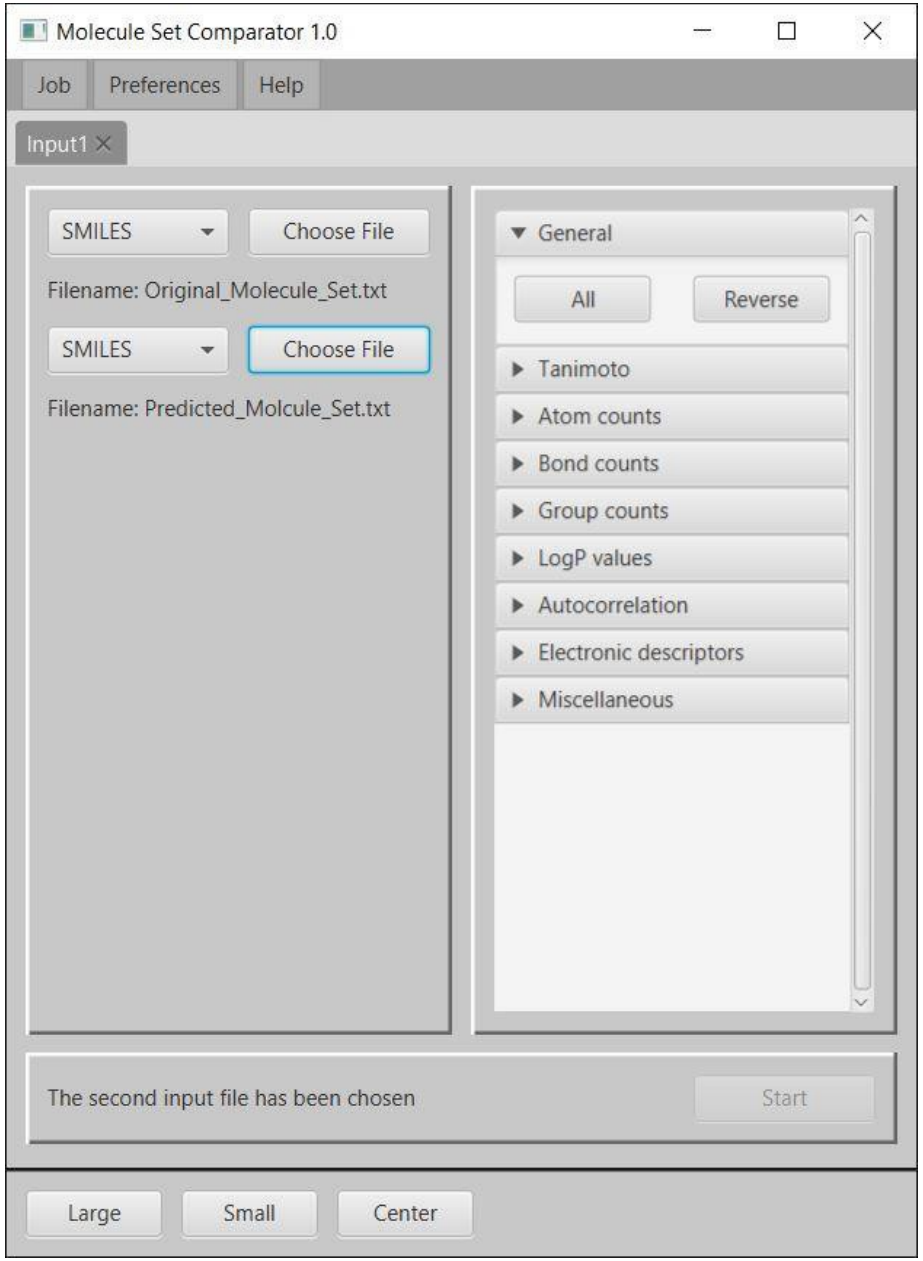
Initializing the job

- In this tutorial we demonstrate how the MSC can be used to evaluate predictions from a machine learning model

- First, the application is started, and a new job is created.

- Then the original and predicted molecule sets are loaded into the MSC.

- Here, both sets are encoded in the SMILES format.



Filename: Original_Molecule_Set.txt

Filename: Predicted_Molcule_Set.txt

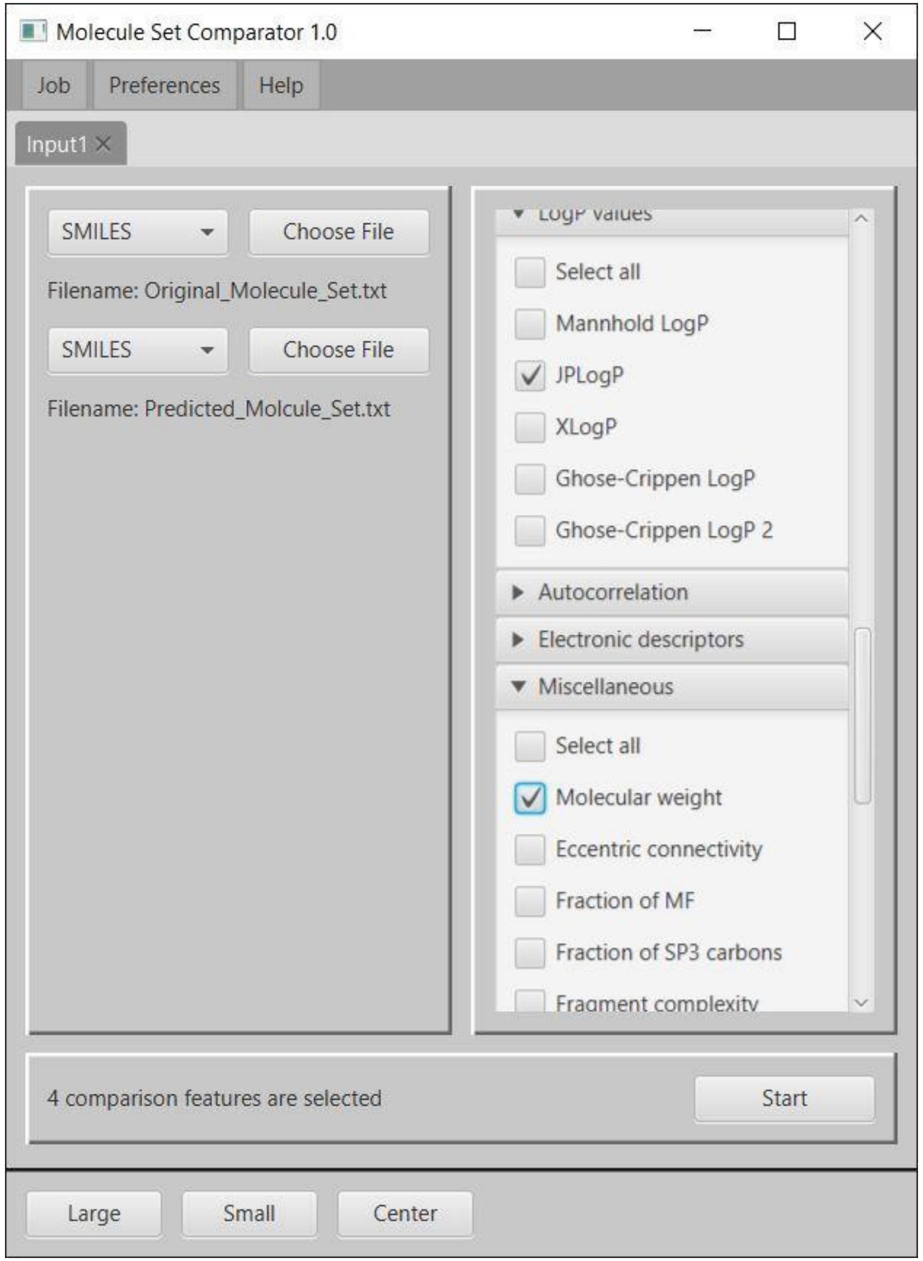
The second input file has been chosen

Initializing the job

- Next, a set of molecular descriptors can be selected to compare the molecule sets based on the user's need.

- Here we have selected the Tanimoto coefficient with the basic CDK fingerprint, the atom count, the JPLogP value, and the molecular weight. For the last three descriptors, each pair is compared based on the difference between the two descriptors.

- Once the job gets started, the molecule sets are compared pairwise (the first original molecule to the first predicted molecule and so on)



SMILES

Filename: Original_Molecule_Set.txt

SMILES

Filename: Predicted_Molcule_Set.txt

▾ LogP values

□ Select all

□ Mannhold LogP

☑ JPLogP

□ XLogP

□ Ghose-Crippen LogP

□ Ghose-Crippen LogP 2

▸ Autocorrelation

▸ Electronic descriptors

▾ Miscellaneous

□ Select all

☑ Molecular weight

□ Eccentric connectivity

□ Fraction of MF

□ Fraction of SP3 carbons

□ Fragment complexity

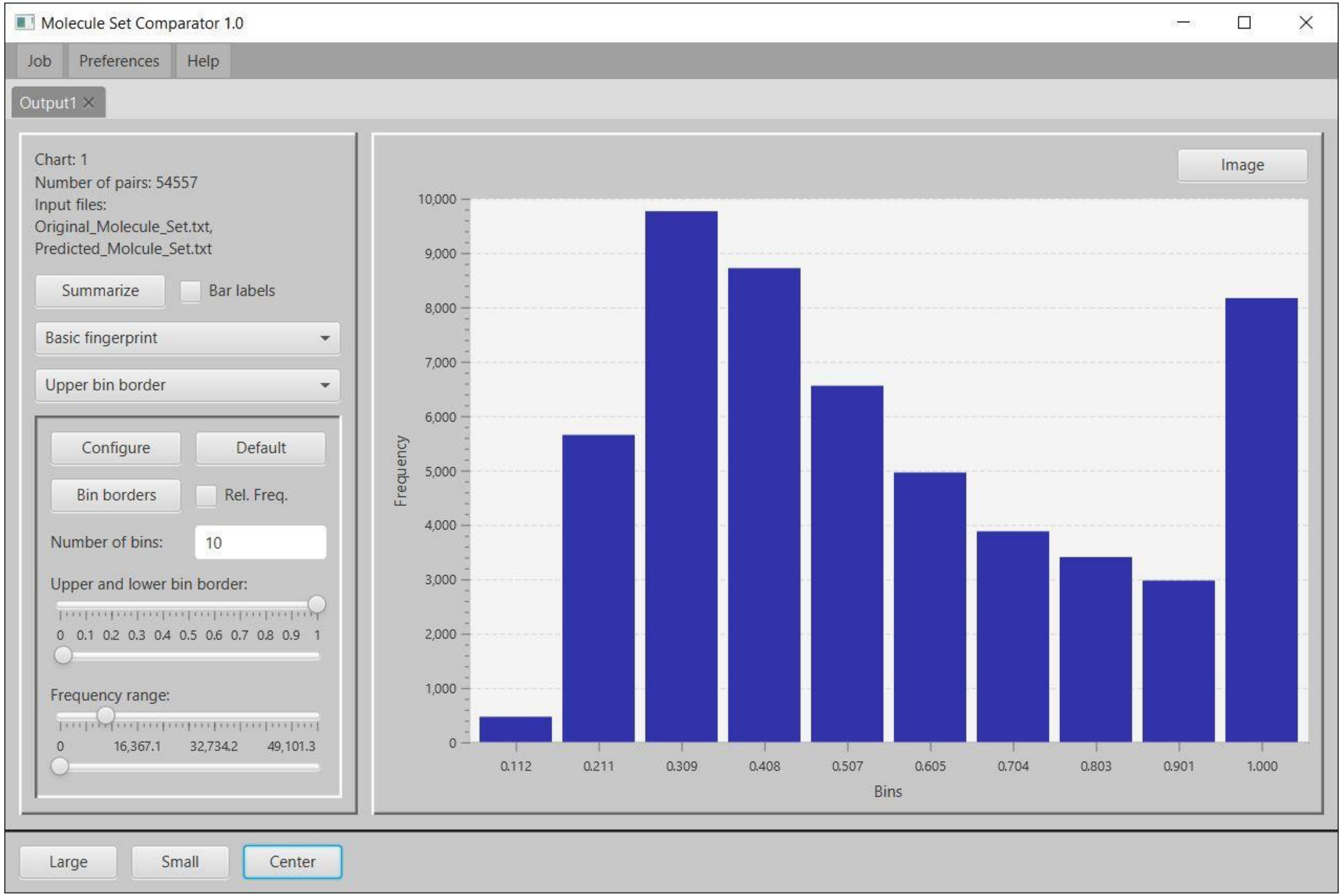
4 comparison features are selected

Analysing the results

- Once the execution completes, a separate tab with the results will pop up.

- The histogram shows the absolute frequencies of molecule pairs for which the Tanimoto coefficient rests in a specific interval. On the x-axis, the intervals are labelled with their upper border (so the last interval goes from 0.901 to 1.0)

- The first step towards analysing these results is to configure the histogram according to your needs.



Molecule Set Comparator 1.0

Chart: 1
Number of pairs: 54557
Input files:
Original_Molecule_Set.txt,
Predicted_Molcule_Set.txt

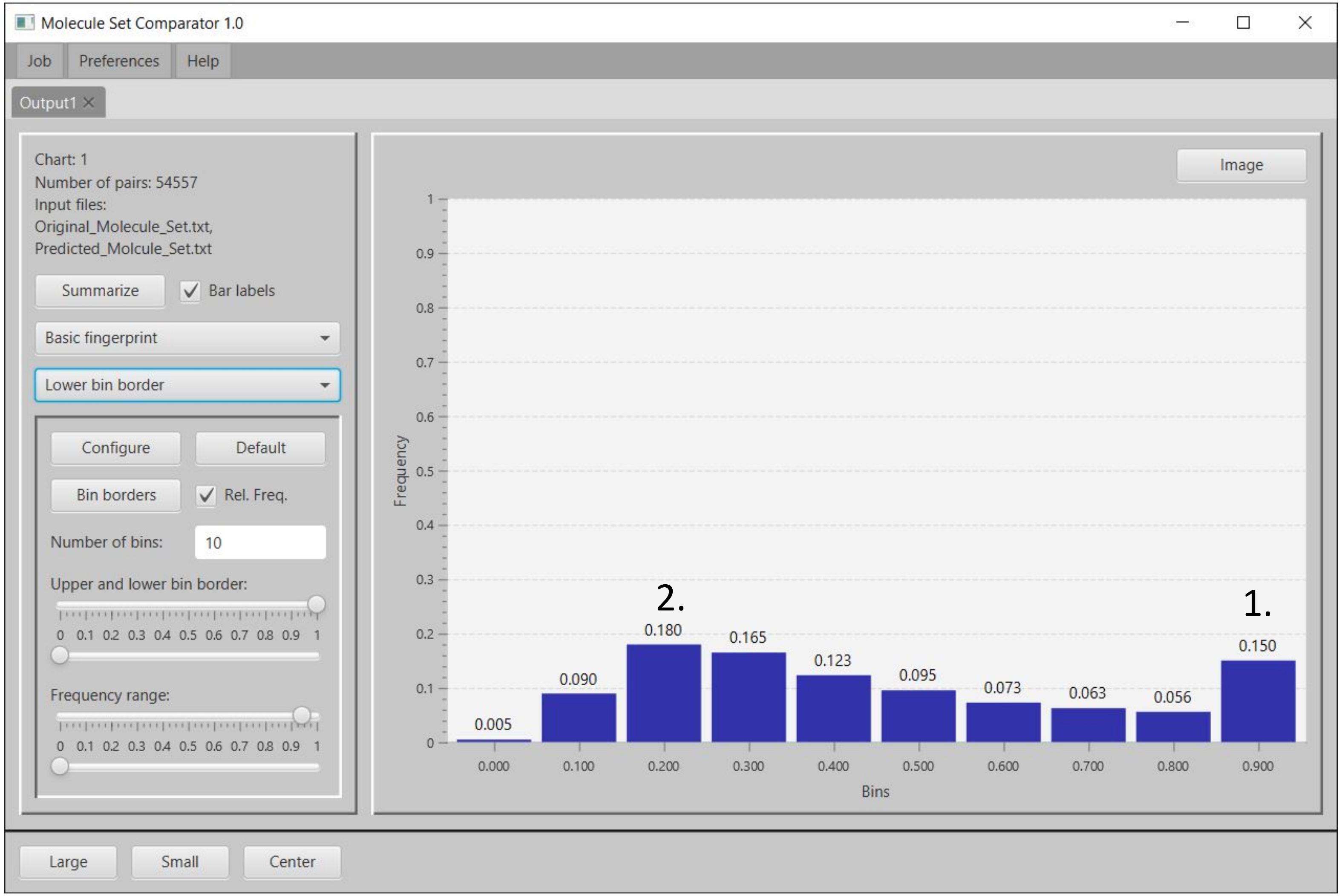
Basic fingerprint

Upper bin border

Number of bins: 10

Analysing the results

- Instead of absolute frequencies, relative ones are now displayed.
- Every bar is labelled with its frequency.
- The y-axis now ranges from 0 to 1
- The x-axis labels now display the lower border of each interval (so the first interval goes from 0.0 to 0.1) and now go from 0.0 to 0.9
- Around 15% of the original set was predicted very well by the algorithm (1.). But many molecules were predicted poorly (2.)



Molecule Set Comparator 1.0

Output1

Chart: 1
Number of pairs: 54557
Input files:
Original_Molecule_Set.txt,
Predicted_Molcule_Set.txt

☑ Bar labels

Basic fingerprint

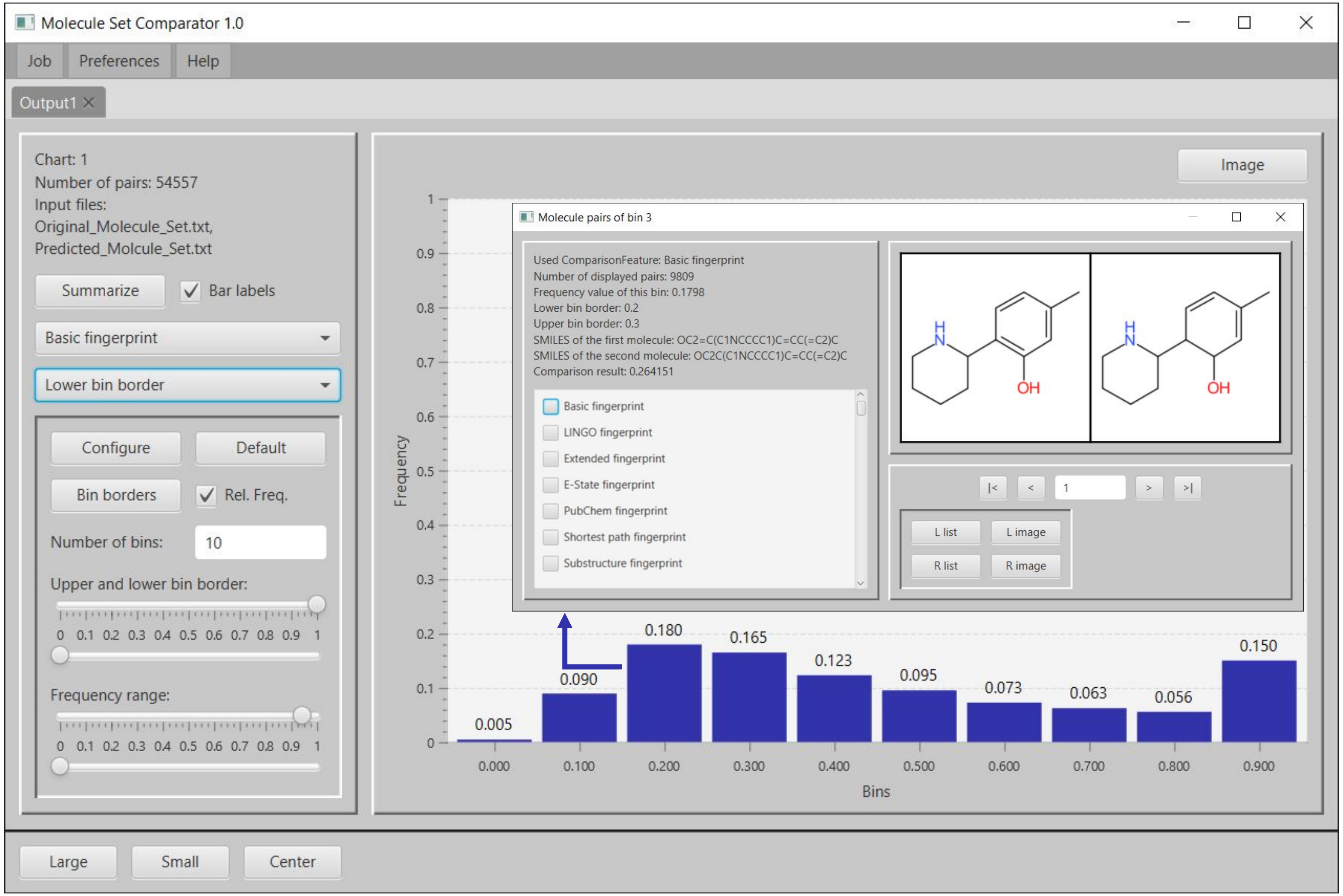
Lower bin border

☑ Rel. Freq.

Number of bins: 10

The detail window

- The poorly predicted molecules can be examined more closely by clicking on one of the bins, for example the third bin (see arrow).
- By clicking on the bin, the detail window for that bin will be opened. Here one can browse through the list of input-output-pairs of the bin.
- It is immediately visible that the first molecule pair doesn't look that different. This intuitive rating contradicts the low Tanimoto coefficient of 0.26



Molecule Set Comparator 1.0

Chart: 1
Number of pairs: 54557
Input files:
Original_Molecule_Set.txt,
Predicted_Molcule_Set.txt

☑ Bar labels

Basic fingerprint

Lower bin border

☑ Rel. Freq.

Number of bins: 10

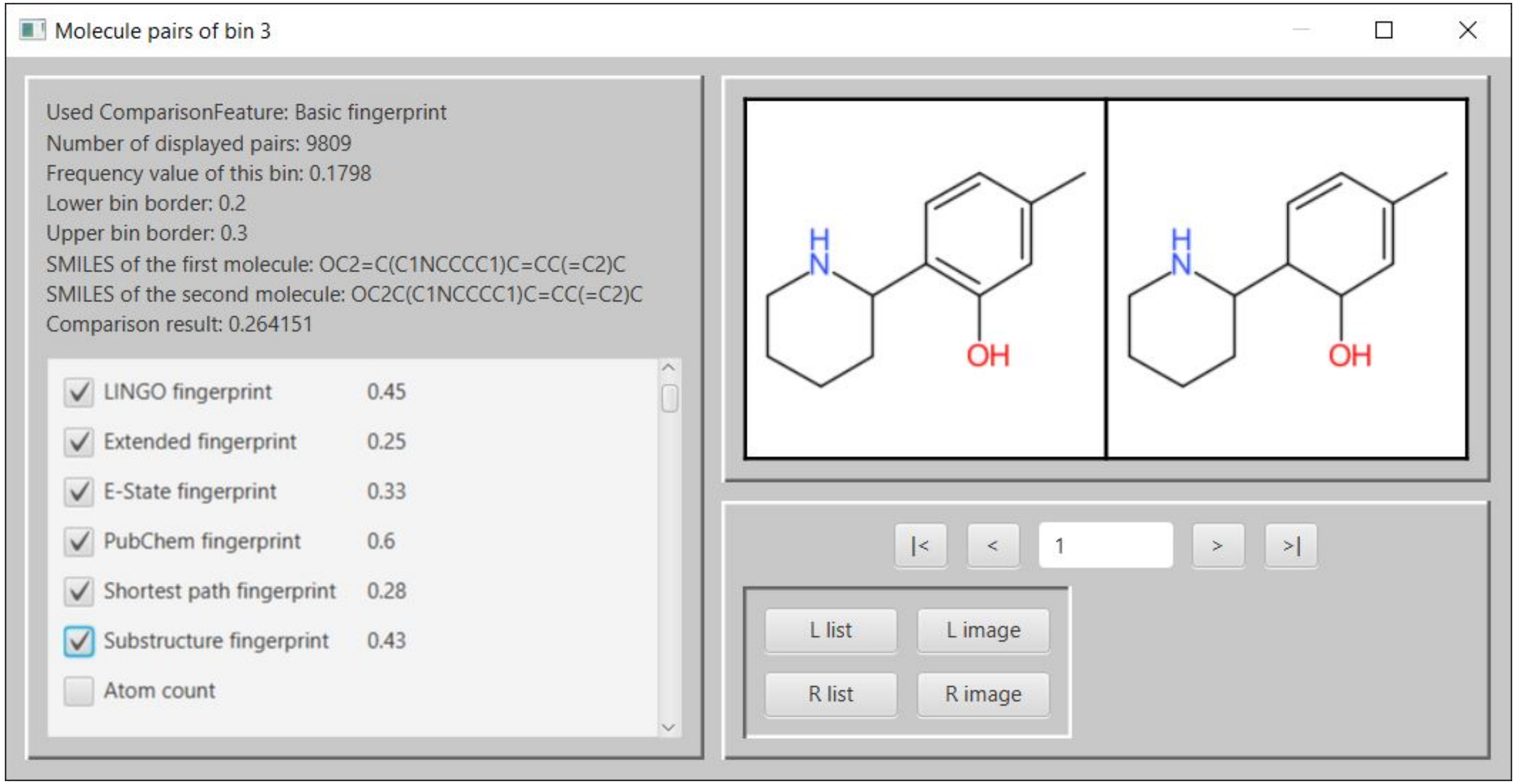
Molecule pairs of bin 3

Used ComparisonFeature: Basic fingerprint
Number of displayed pairs: 9809
Frequency value of this bin: 0.1798
Lower bin border: 0.2
Upper bin border: 0.3
SMILES of the first molecule: OC2=C(C1NCCCC1)C=CC(=C2)C
SMILES of the second molecule: OC2C(C1NCCCC1)C=CC(=C2)C
Comparison result: 0.264151

- □ Basic fingerprint
- □ LINGO fingerprint
- □ Extended fingerprint
- □ E-State fingerprint
- □ PubChem fingerprint
- □ Shortest path fingerprint
- □ Substructure fingerprint

The detail window

- This perceived contradiction can be solved by calculating additional Tanimoto coefficients with different fingerprints.
- For example, the PubChem fingerprint yields a Tanimoto value of 0.6 for the same molecule pair which seems to be a lot more reasonable.
- This is because the Tanimoto coefficient is very dependent on the fingerprint.
- To get a clearer picture we could run a separate job with a different fingerprint.



Used ComparisonFeature: Basic fingerprint
Number of displayed pairs: 9809
Frequency value of this bin: 0.1798
Lower bin border: 0.2
Upper bin border: 0.3
SMILES of the first molecule: OC2=C(C1NCCCC1)C=CC(=C2)C
SMILES of the second molecule: OC2C(C1NCCCC1)C=CC(=C2)C
Comparison result: 0.264151

☑ LINGO fingerprint	0.45
☑ Extended fingerprint	0.25
☑ E-State fingerprint	0.33
☑ PubChem fingerprint	0.6
☑ Shortest path fingerprint	0.28
☑ Substructure fingerprint	0.43
☐ Atom count	

Other descriptors

- In the output tab, we can quickly switch between the different calculated descriptors.
- Here the histogram of the atom counts descriptor is displayed. The intervals are labelled with their lower border (so the first interval goes from 0 to 17, the second from 18 to 34 and so on)
- If you do not configure the display, the histogram can be uninformative.

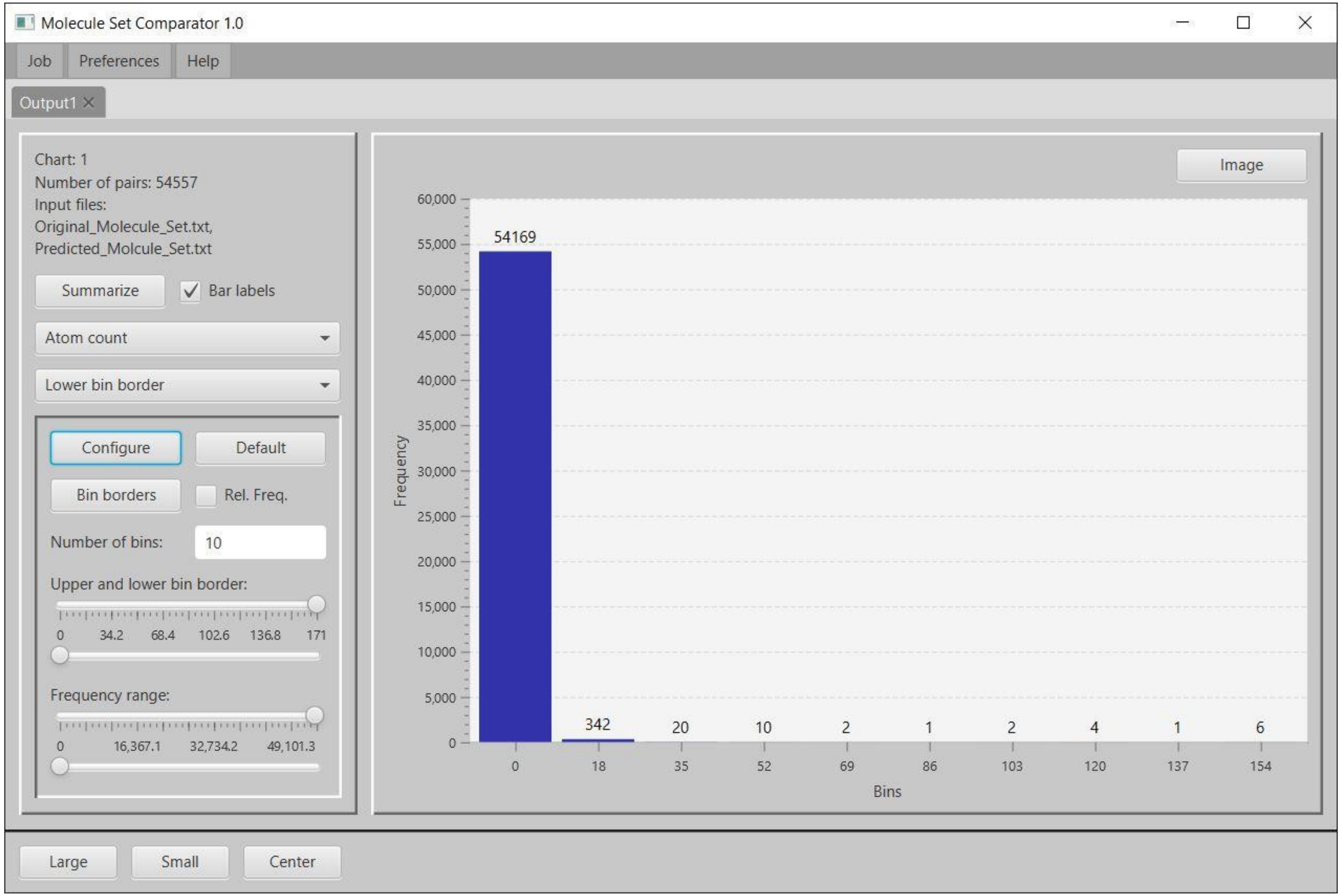


Chart: 1
Number of pairs: 54557
Input files:
Original_Molecule_Set.txt,
Predicted_Molcule_Set.txt

☑ Bar labels

Atom count

Lower bin border

☐ Rel. Freq.

Number of bins: 10

Other descriptors

- After configuring, the histogram conveys a lot more information.
- Around 25% of the molecules that were predicted have the right number of atoms.
- The histogram shows a “nice” distribution except for the 4th bar.
- This anomaly is likely to be caused by the addition of one “C” atom which also leads to two additional "H" atoms.

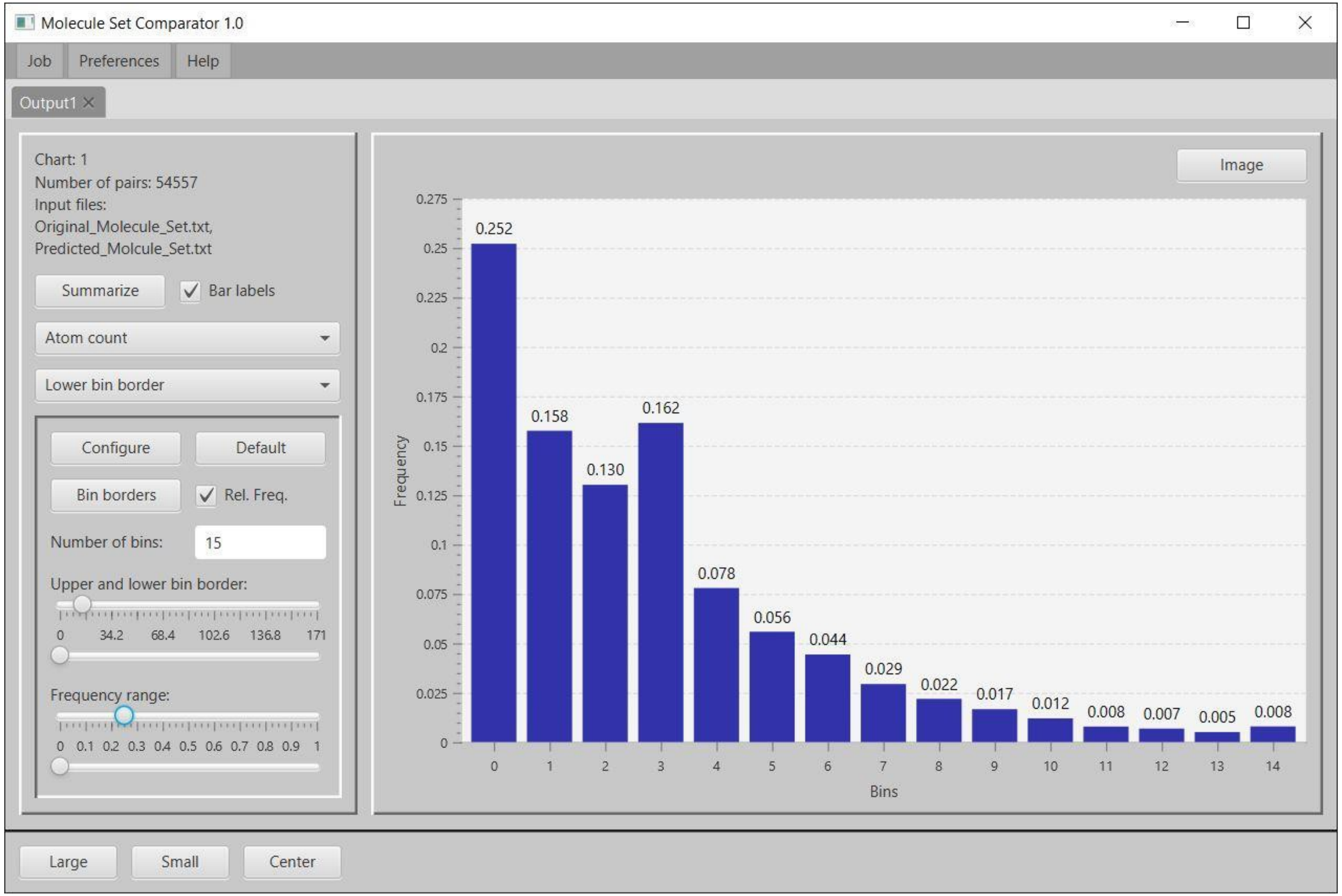


Chart: 1
Number of pairs: 54557
Input files:
Original_Molecule_Set.txt,
Predicted_Molcule_Set.txt

☑ Bar labels

Atom count

Lower bin border

☑ Rel. Freq.

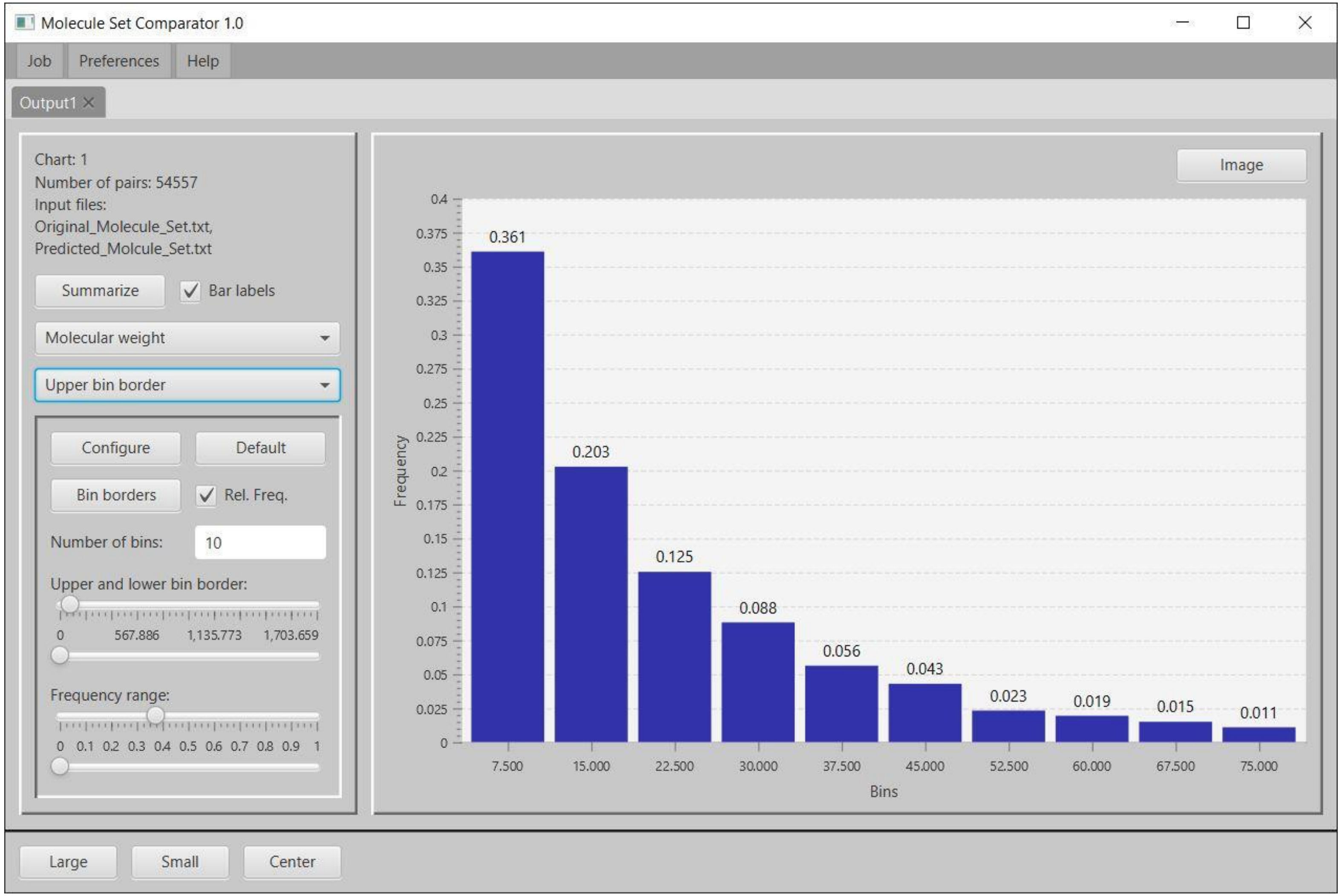
Number of bins: 15

Other descriptors

- The configured histogram of the molecular weight descriptor also exhibits a “nice” distribution without anomalies. The intervals are labelled with their upper border (so the first interval goes from 0 to 7.5)

- Roughly 36% of the predicted molecules have a similar molecular weight that only differs by about 0 to 7.5 daltons (see first bin)

- By clicking on the **Image** button, the histogram can be exported into various file formats for different purposes.



Molecule Set Comparator 1.0

Chart: 1
Number of pairs: 54557
Input files:
Original_Molecule_Set.txt,
Predicted_Molcule_Set.txt

☑ Bar labels

Molecular weight

Upper bin border

☑ Rel. Freq.

Number of bins: 10

Upper and lower bin border:

Frequency range:

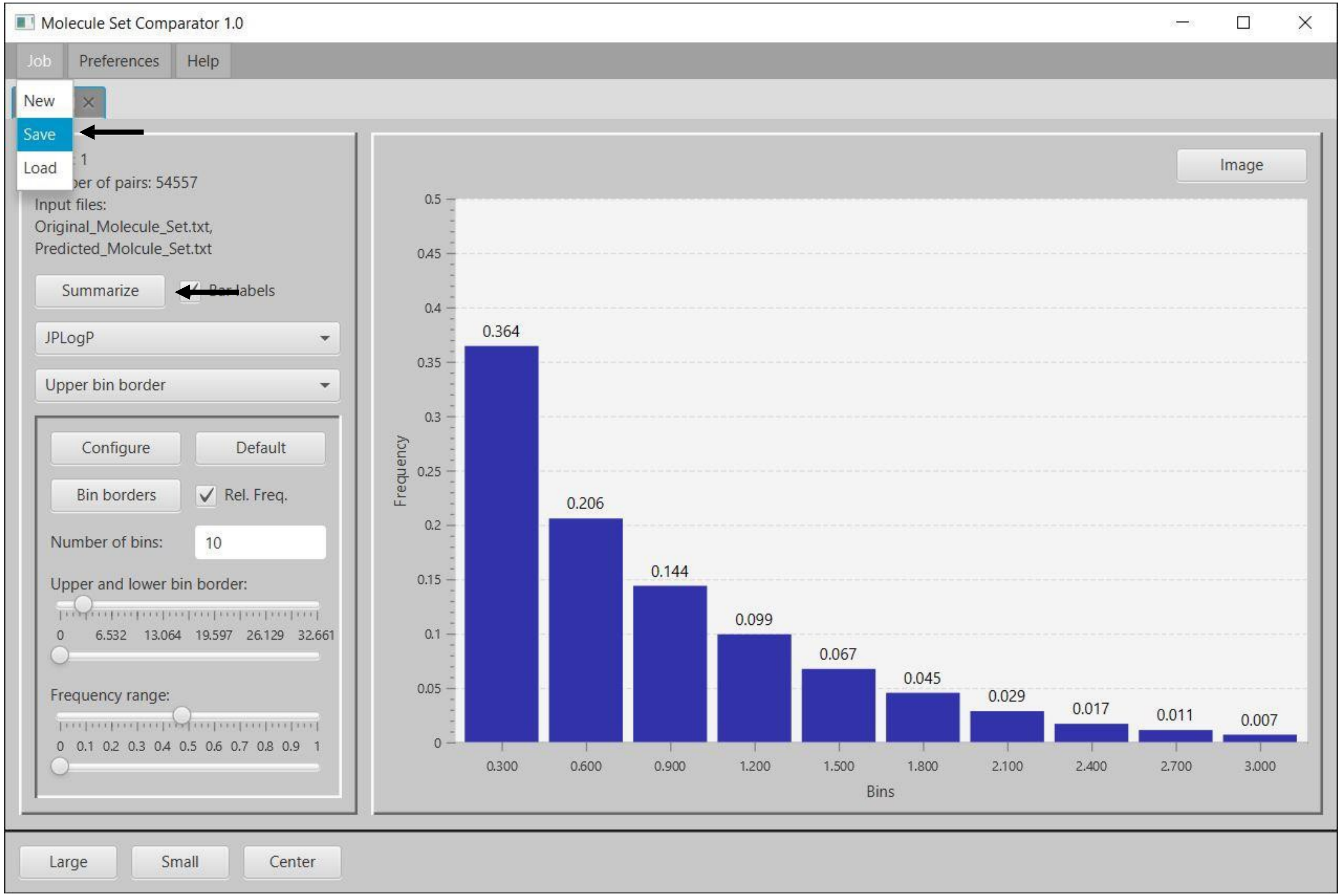
Other descriptors

- The configured histogram of the JPLogP value can be seen on the right. The intervals are labelled with their upper border (so the first interval goes from 0.0 to 0.3)

- The histogram shows a distribution that is very similar to the one of the molecular weights. This connection could be worthy of a more thorough exploration with, e.g. the detail window

- Using the **Summarize** button, a summary document can be generated

- The processed job can be saved to be reloaded later or to be shared with others.



Molecule Set Comparator 1.0

1
ber of pairs: 54557
Input files:
Original_Molecule_Set.txt,
Predicted_Molcule_Set.txt

☑ Bar labels

JPLogP

Upper bin border

☑ Rel. Freq.

Number of bins: 10

Upper and lower bin border:

Frequency range:

Summary

- The goal here was to evaluate the capacity of a machine learning algorithm by comparing the predictions with the original molecule set.

- For the evaluation, four molecular descriptors were used (Tanimoto coefficient, atom count, JPLogP value, and molecular weight).

- Based on the histograms a quick assessment of the predictive power of the algorithm is possible. Here we could see that some molecules were predicted well, but the majority got predicted poorly. So there is room for improvement.

- Also, a new evaluation based on Tanimoto coefficients with other fingerprints is advised for a more thorough evaluation.

- Also, the tool can generate multiple outputs to share the results.