

v0.1 -- still under construction!

- > Start a job
- > Select a job
- > Models
- > MMP Overview
- > Atom Contributions
- > Feature Importance
- > Evaluation
- > WISP

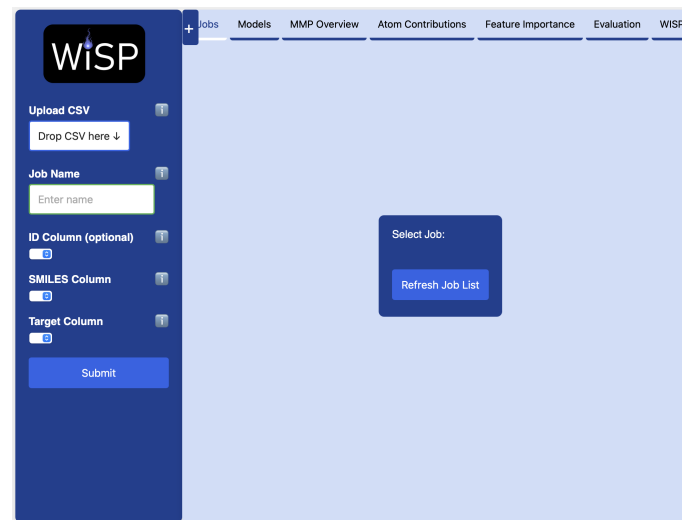
> How-to: Start a job?

visit website (public instance at <https://www.molecularxai-lab.com/wisp>)

Input is a CSV file with molecules in SMILES format and target property as a float number, e.g.:

```
solub,SMILES  
-0.77,OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)C(O)C3O  
-3.3,Cc1occc1C(=O)Nc2ccccc2
```

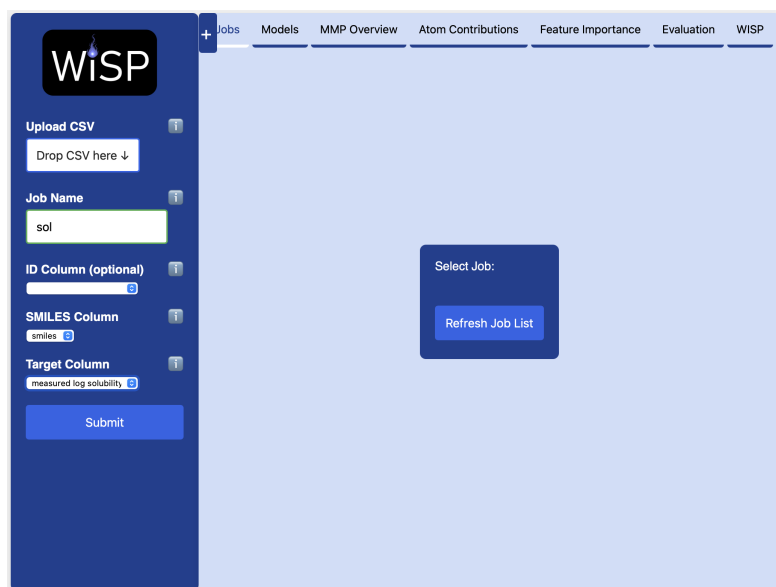
drag and drop the csv file or click on "Drop CSV here" and select the input CSV from the file dialog



drag and drop the csv file or click on "Drop CSV here"
and select the input CSV from the file dialog

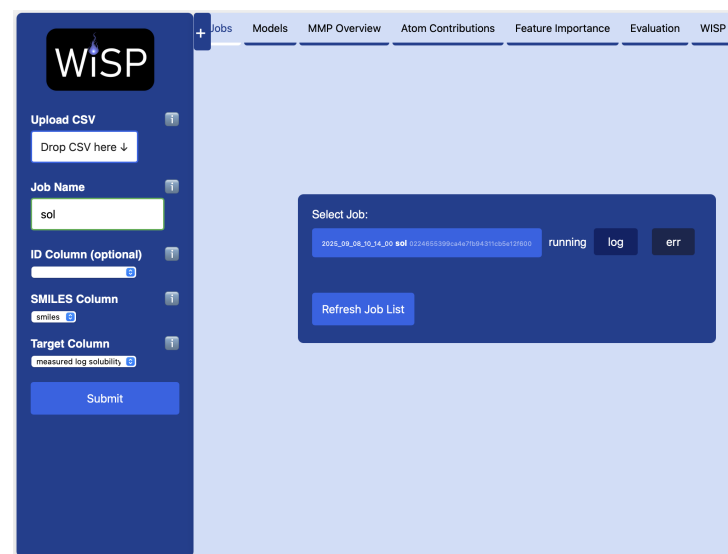
SMILES and target columns are automatically detected,
but you have to check that the right ones were chosen!

For delaney esol dataset, the smiles column was
correctly detected, but I had to correct the target
column from the dropdown list:



The screenshot shows the WiSP web interface. On the left is a configuration sidebar with the following fields: 'Upload CSV' with a 'Drop CSV here' button; 'Job Name' with a text input containing 'sol'; 'ID Column (optional)' with a dropdown menu; 'SMILES Column' with a dropdown menu showing 'smiles'; and 'Target Column' with a dropdown menu showing 'measured log solubility'. A blue 'Submit' button is at the bottom of the sidebar. The main panel on the right has a light blue background and contains a 'Select Job:' box with a 'Refresh Job List' button.

after confirming the right SMILES&target column.
submit the job by clicking the button, and a
corresponding job should appear in the center panel:



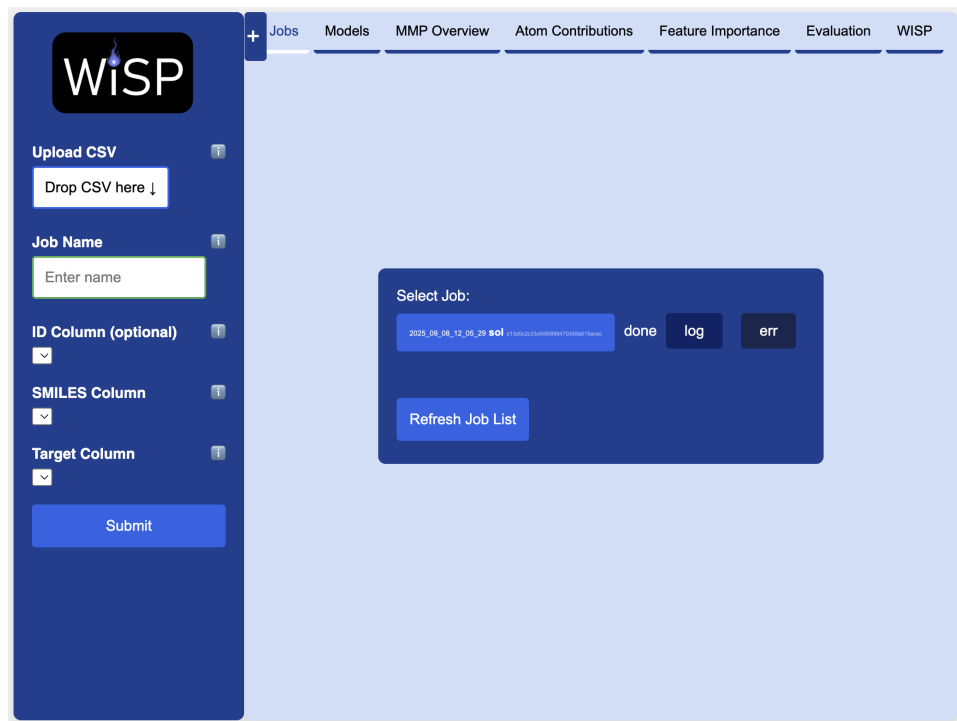
This screenshot is similar to the one on the left, but the 'Target Column' dropdown menu now shows 'measured log solubility' instead of 'measured log solubility'. The 'Submit' button is still present at the bottom of the sidebar. The main panel on the right shows the 'Select Job:' box with a 'Refresh Job List' button.

currently the job is still "running", and it takes typically
20 min. - 1 hour per job to be completed

*you have to click on "Refresh Job List" to update both
the job status and the log/error output information!*

once a job completed successfully, it's status will appear as "done", in case there are any errors, the error log would appear red, and the status would be shown as "failed".

here a screenshot of a successfully completed job:



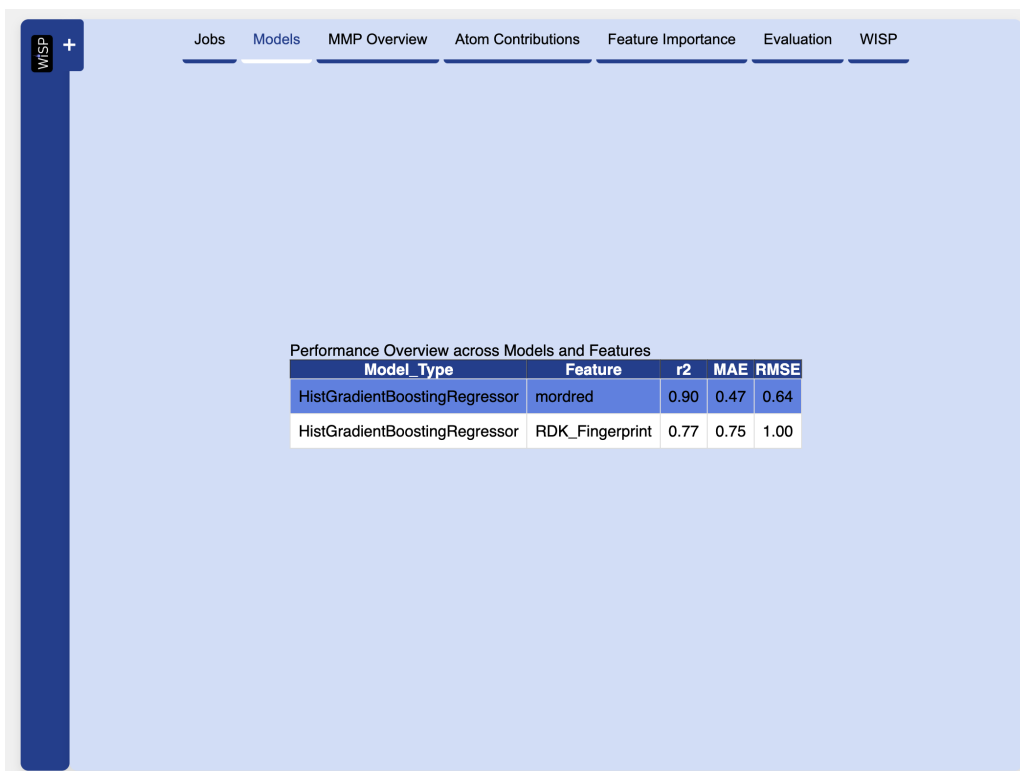
To select a job, go into the central panel and select the desired job by clicking on it. The selected job will be highlighted in grey:



Now with the job selected, click on the tabs on the top right to select the different types of output (Models, MMP Overview, ...)

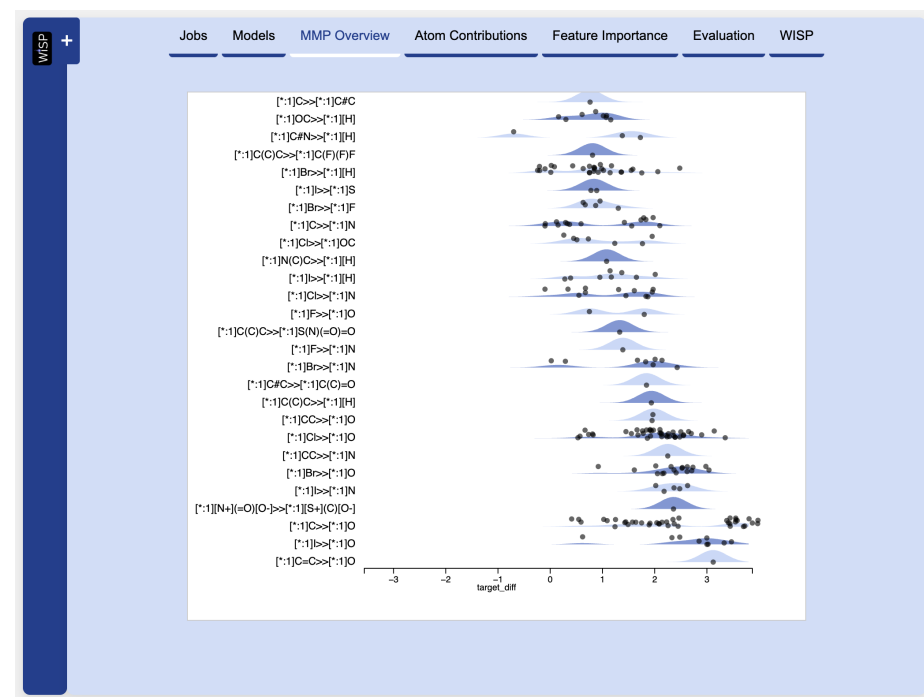
The Models tab gives an overview of the models that were trained together with their performance across a range of selected metrics on a random 80% : 20% split.

For ESOL below, it was found that molecular descriptors (mordred) performed very well with a high pearson correlation (R^2 0.90) and a mean absolute error of about half a log unit. The RDKitFingerprints performed worse. WISP selects the best model automatically for further processing, which is defined as the model with the lowest MAE value.



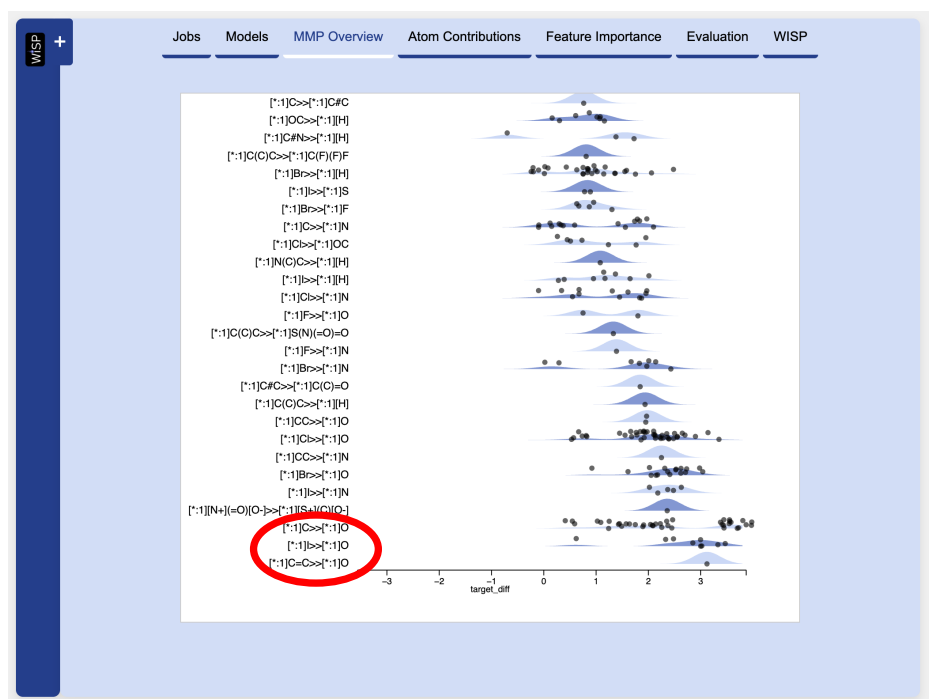
This tab gives an overview of the matched molecular pairs, where:

- in the vertical direction:
each row corresponds to a given MMP rule, e.g. replace hydrogen by chlorine (chlorination)
- in the horizontal direction:
influence on target property (e.g. increase of solubility with positive values on the right, lowering solubility with negative values on the left)
- each datapoint corresponds to a single data instance



Note that the MMP overview panel supports scrolling in the vertical direction.

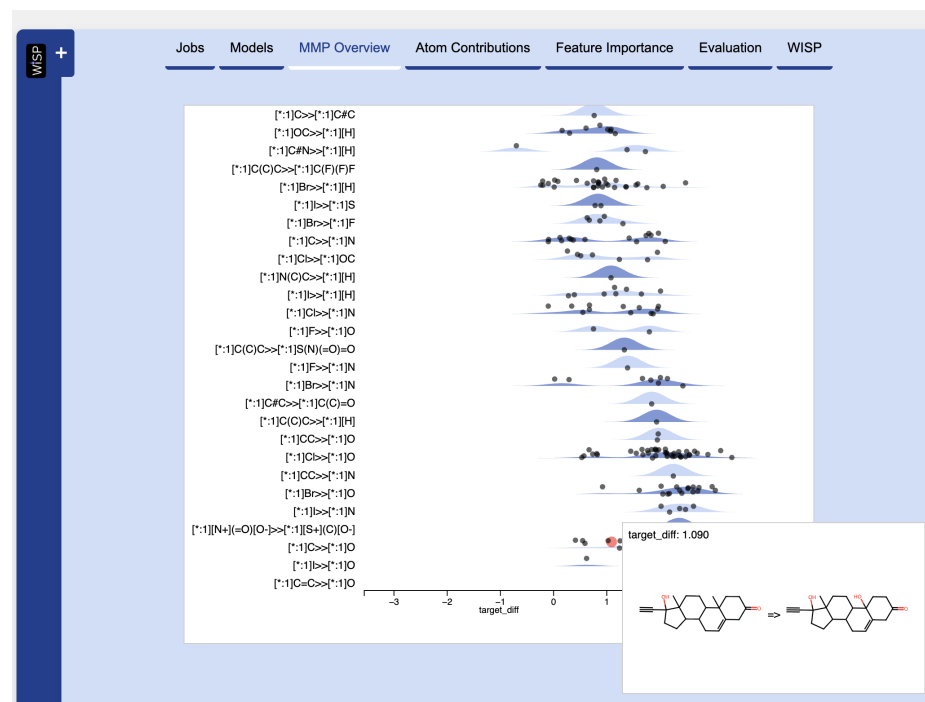
Scrolling all the way down on the delaney / ESOL dataset, we observe that the three rules with the highest positive impact on aqueous solubility are all hydroxylations:



We can further inspect single datapoints corresponding to concrete matched molecular pairs by hovering over them.

On hovering, the selected point will appear in red and a pop-up will show the concrete MMP.

For ESOL, we can see that for the selected methyl-hydroxyl substitution, the induced target diff is likely lower because there is already a hydroxyl group present, arguably lowering the impact of hydroxylation on the overall aqueous solubility.



This tab therefore gives a convenient overview of the MMP patterns for a given dataset, and can already provide valuable insights into the molecular property.

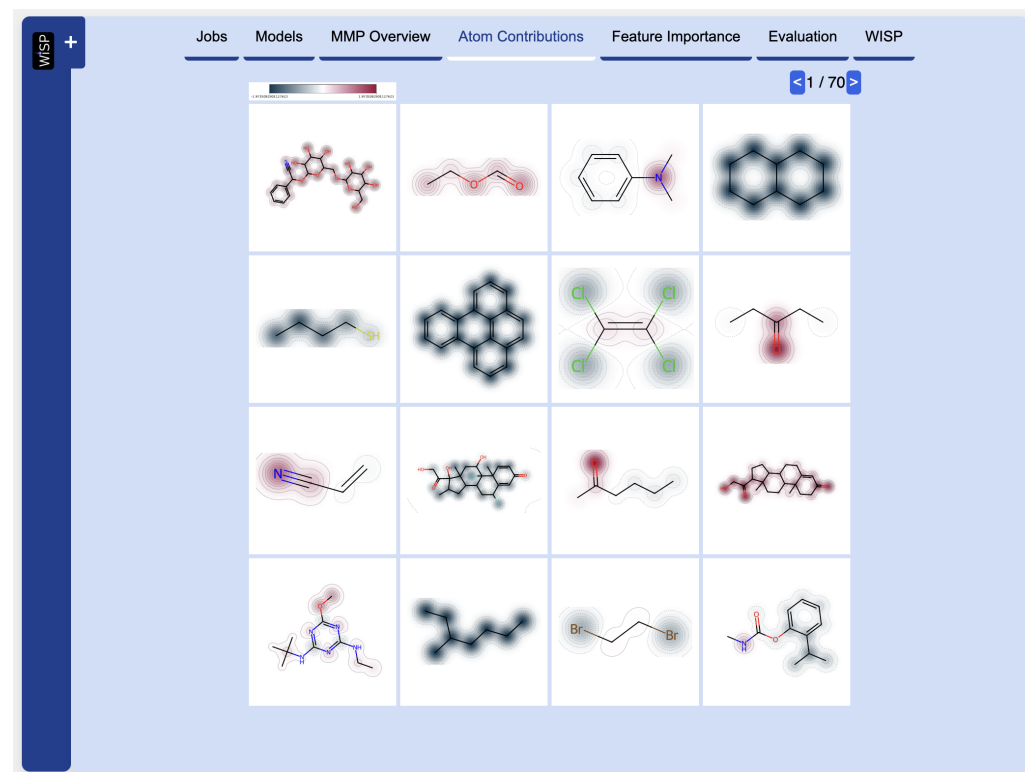
However, it should also be noted that the final tab "WISP" is based solely on the matched molecular pairs shown here. Therefore, if few MMPs are generated, or the MMP rules that are extracted are "trivial" in nature, then this will also affect the final WISP analysis, as it takes the MMPs as its basis.

The MMP Overview should therefore be used to assess whether evaluating the given model on the MMP rules is sensible from the domain perspective.

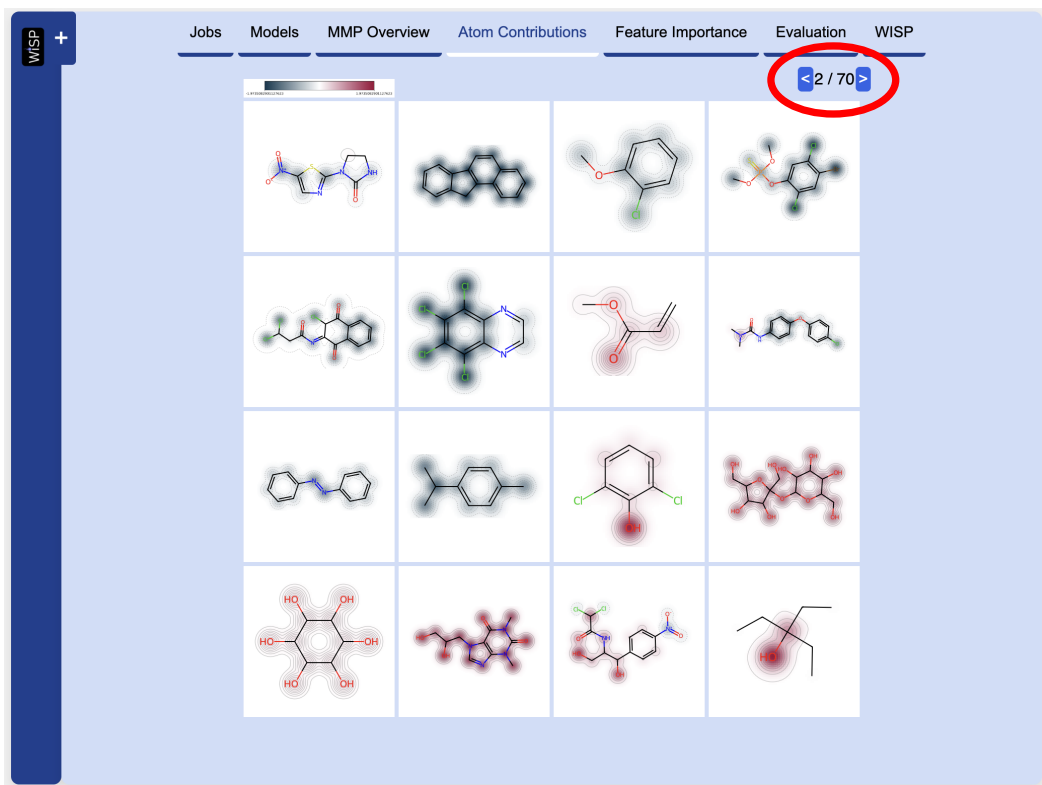
The Atom Contributions tab shows paginated examples of how each atom contributes towards the prediction of the ML model.

Blue areas contribute negatively towards the outcome, while red areas contribute positively towards the outcome.

Here, we can see that the ESOL model learned correctly that e.g. carbonyl groups (hydrogen bond acceptors) positively contribute towards the solubility (increase solubility), while e.g. halogens and isopropyl groups decrease it.



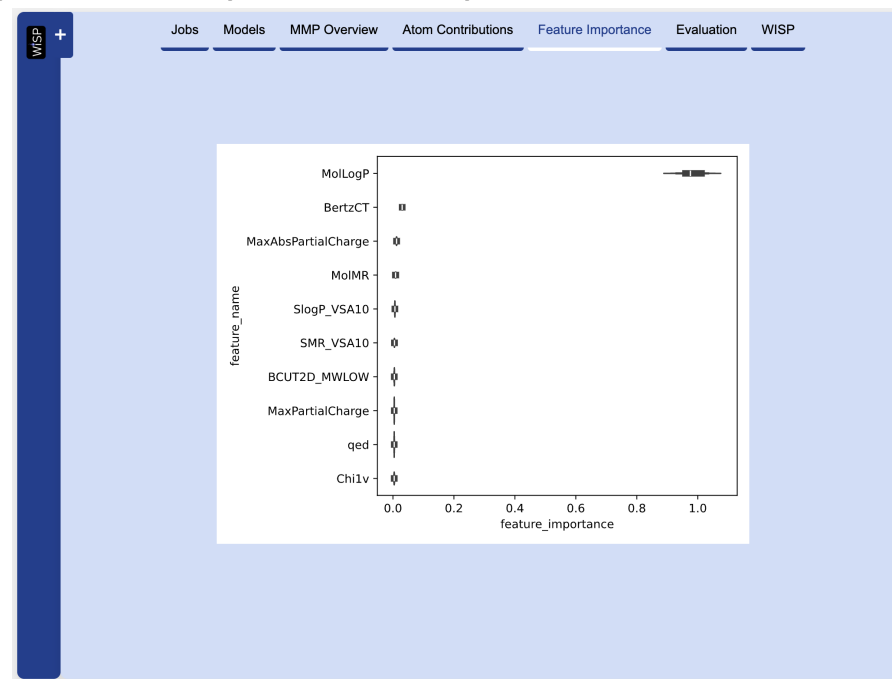
Clicking on the arrows allows to browse through all of the 70 pages for this dataset:



On lower bandwidth connections, loading new pages can take several seconds.

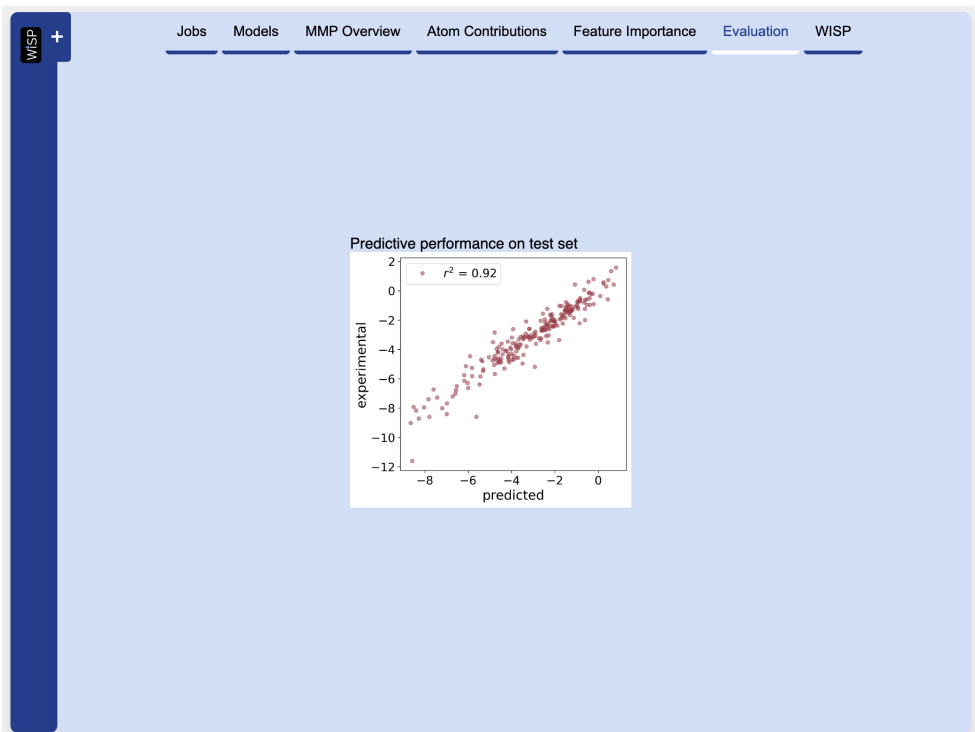
Shows the result of training a tree ensemble model* on all available RDKit descriptors. The goal is to show what classical molecular properties highly correlate with the property of interest.

In the case of ESOL, we see that the log P is the most important feature, as well as several other descriptors (MaxAbsPartialCharge, ... ,) that reflect the polarity of the given compound. Furthermore, several features that reflect the size of the compound (BertzCT, MolMR) are also relevant. This is intuitive, as both types of descriptors are expected to be relevant here.

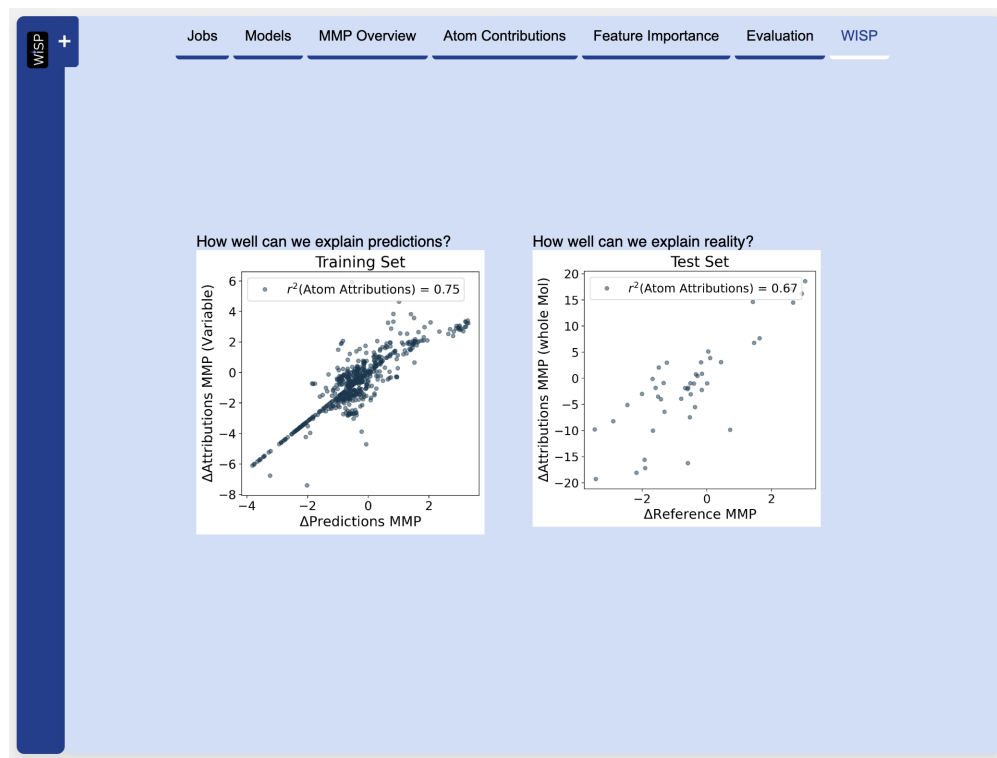


* A HistGradientBoostingRegressor, to be more precise, as described here: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html>

Shows the parity plot on the test set employing a 80% : 20% random train/test split.



Shows the result of the WISP analysis for predictions on the train set (left side, how well can we explain the predictions?), and on the ground truth of the test set (right side, how well can we explain the future/reality?).



For the example dataset (ESOL, delaney)*, we learned:

- model results look promising
(parity plot in "Evaluation" tab,
metrics in "Models" table)
- MMPs show sensible patterns in the data
(e.g. hydroxylation increases aq. solubility),
- the model is capable of describing these MMP changes
well both on past predictions (WISP, left plot) as well
as on future data (WISP, right plot)

However, as we see for ESOL, this still does not mean that the molecule will also generalise towards unseen chemical space, as we only simulated predicting the changes across MMPs drawn from the same data. Thus, there is an implicit assumption of identically distributed data.

We can therefore not conclude anything about the generalizability of models towards new chemistry.

The WISP analysis only works to reject faulty models, when we see that the models already break under the assumptions made.
But the converse is not true:
A positive outcome does not imply that the model works well.

* obtainable from e.g.: <https://moleculenet.org/datasets>

