Troubleshooting

I can't upload my csv file. How is it supposed to be formatted?

ChemX can read csv files in several different ways, but below is an example of an accepted format. If you are having issues it can help to add "smiles" or "SMILES" as header to the relevant column. Note, that if you have very few compounds, ChemX may have a harder time assigning the columns (and if you have fewer than 10 you are unable to train a model).

The csv file must have a column with smiles strings and at least one column with a target parameter - either quantitative or categorical. A name/ID column is not required.

Drug_ID,Drug,Y

22416348,Cc1occc1C(=O)NCc1ccco1,20.17528

26665387,COc1ccc(/C=C2\SC(=S)N(N3CCOCC3)C2=O)c(OC)c1,10.2263

862531,CC(NC(=O)Nc1cccc(C(F)(F)F)c1)C(=O)O,2.0442

I uploaded my csv file but one of the columns were wrongly interpreted as text instead of categorical data. What do I do?

After upload, you can change the parameter type by clicking "Reassign parameter type" in the "Additional actions" panel on the right.

I can't get my target data normally distributed. Can I still train a regression model?

Yes, even though you don't have normally distributed target data, it is still possible to train a regression model using a subset of the machine learning algorithms that ChemX employs under the hood.

All my models have R^2 below 0.5. What am I doing wrong?

If all you models have poor predictive power, then you should go back to check the data. Check if the molecules are missing important information for predicting the target data, e.g. chirality annotation or protonation state. Also check the target data - is it (near)normally distributed? If none of the available transformations improve your dataset, then you should consider changing to a categorical approach instead. This can happen if your data is highly zero-spiked or unbalanced.

The best way forward could be to train a categorical model on the whole dataset that can predict if a molecule is good or poor overall, and then train a regression model on the good subset of the data that can be used to rank the molecules found to be good by the categorical model.

I am trying to model a dataset (% inhibition at 1 uM) that is a perfect normal distribution, but I only create terrible models. What is going on?

This is an excellent example of a major caveat of modelling. On the surface your data looks excellent for training machine learning models; however, due to the way the data was measured, there is likely considerable diversity in both structure and actual inhibitory power between molecules that are grouped together as 100 % inhibtion at i uM. I.e. some of the molecules may continue to invoke complete inhibition at sub-nM concentrations where others only do at uM concentration. This makes it next to impossible to train a good machine learning model.

Instead, we suggest you make a categorical model that splits the best inhibitors from the worst. If you have follow-up data on the good subset at lower inhibitor concentrations, then that can be used to train a regression model that can rank the good subset after the first model has identified them. Be aware that the regression model likely was trained on less diverse chemistry and may thus be less generalizable.

I am getting negative R^2 values for my models. What is happening?

The R^2 metric used in ChemX is calculated as one minus the ratio of the mean squared error (MSE) of the model predictions over the MSE of the trivial model predictions (constant number). The reported R^2 is therefore not the classic R^2 many of us have learned in school when learning to fit regression models.

A value of 1 indicates perfect predictions, while values around 0 indicate that the regression results are no different than the trivial approach of constantly predicting the average value of the outcome. Negative values indicate that the model is worse than the trivial approach.

I have multiple equally good models. How do I know which one is best?

The best regression model may not be the model with the highest R^2. Here are a few things to consider:

1. When models are trained multiple times using different data splits for training and testing sets the calculated performance varies to some extent. Therefore, when comparing model performance, the variation should be considered too to assess if two models actually perform differently. Use the reported 95% confidence interval for each model; if the intervals of the models do not overlap their performances are significantly different. Be aware that the opposite is not necessarily the case!
2. Two models with significantly different R^2 values may not have significantly different MAE and vice versa.
3. Consider how you intend to use the model down the line. Will it be used as a quick filter on large molecule libraries? If so, then it might be best to choose a slightly less predictive model that uses a quicker featurizert and is much faster to screen with.
4. Ensure that the model didn't pick up on random correlations between target parameter and features e.g. when training a model that predicts orbital energies, a simple featurizer like RDKit descriptors does not encode the chemical complexity necessary to predict quantum mechanical properties. For more information on how to assess this, go to the "Probing the model" section of this tutorial.

I used a transformation on my data and now I don't know how to interpret my screening results.

We realize that data transformations are sometimes unkown territory for our users, so we have applied a de-transformation step before you see any of the results - for counterfactuals etc during modelling, and in your screening results later on. The only place you will see data on the transformed scale is for the MAEs reported in the performance heatmap so don't conpare MAE's between transformations, only between featurizations.

I want to use a model on a screening library. How do I know if the model is applicable to the new chemistry?

After you have applied a model to a screening library, you can see a UMAP of both the training and screening data. If there is good overlap between the training and sceening library, then you can assume that the predictions are good (assuming the model has good predictive power). If there are a lot of molecules from the screening library that is far from the training data, then the chemistry in the screening library is unknown to the model and you should proceed with caution and ensure external validation of the predictions.

I know some of my molecules are (de)protonated at the pH I run my experiment, but sanitization removes all changes I make in my csv-file. What do I do?

If you expect the protonation state to be important for predicting your target paramater, then you have to edit your SMILES accordingly and deselect sanitization upon upload to ChemX.

I am trying to export my predictions, but nothing happens

Clicking the "Export data as csv" button from the Workflow panel (only available from within screening libraries, not modelling projects) results in your browser downloading a complete data file to your Downloads folder without further questions. If you have many predictions it may take some time to download, and if you have a blocker added to your browser it may silently block the download and you will have to find out how to allow the download.