


# Frequently Asked Questions

---

# 1. Can I download data?

---

Yes. Any user can download data that

- are uploaded by him/her (including private data);
- are publicly available and freely downloadable (indicated as:  **Public and freely downloadable record**).  
An example of such data are melting point from [Bergstrom et al article](#).
- are shared by providing a public id of a model developed (use **Download descriptors and model statistic link** on the model page)

## 2. The first steps to develop models

---

The suggested way is to use Comprehensive Modelling interface (see Model/Create multiple models menu).

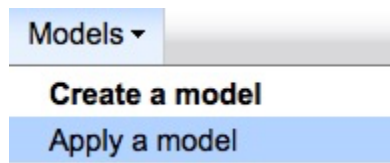
Important steps:

- Specify units with physical meaning (e.g., log(mol/l) and not mol/l or g/l for model development)
- Select only one method with different (all) descriptors
  - one of the longest and CPU consuming part is calculation of descriptors; some descriptors such as PyDescriptor take about one minute per molecule
  - once descriptors are cached the other methods will reuse them from the cache without a need to spend again and again CPU time for their calculation
  - LibSVM requires to select descriptor normalization in advanced options (otherwise results with this method will be really bad)
- Do not use the bagging unless you do the final model development. Bagging will **always** provide better results than cross-validation, but requires much more computational resources.
- Identify the best method and several best descriptors and continue work with them (outliers exclusion)
- For highly imbalanced classification datasets stratified validation can significantly improve the accuracy of models

### 3. How can I predict new data using an OCHEM model?

---

Use the Apply model menu



select models that you would like to apply, click "NEXT" at the bottom of page and proceed to upload page, where you can specify molecules to be predicted.

## 4. Why prediction of even a single molecule could very long?

---

If prediction of a molecule is cached, the result is shown almost instantaneously. The processing of a molecule involves complex steps, such as calculation of descriptors, processing of multiple models, each of which can consist of hundreds of submodels (for example in case of bagging). All these steps can take minutes and sometimes tens of minutes. Sometimes OCHEM is overloaded and the task can wait in a queue before starting calculations. However, if predictions are finished and cached, the result will be instantaneous (e.g., for the next prediction of the same structure). The best way - predict a set of molecules. The calculations for one and few thousands of molecule take approximately the same time.

## **5. I have a set of molecules. Should I predict them (or calculate descriptors for them) one by one?**

---

No. This is a very inefficient way. The result will be much faster (and will require much less computations) if you will send them as a batch (e.g., as an uploaded SDF file with all molecules or using basket).

## **6. Can I calculate and export descriptors?**

---

Yes, unless it was forbidden by the provider of a respective descriptor package. These descriptor packages (e.g., Dragon, Adriana) are not listed in the browser for calculation of descriptors.

## 7. Can I develop models using my own descriptors?

---

Yes. First you need to upload your structures to OCHEM. This will allow you to obtain for each molecule its MOLECULEID in OCHEM (by exporting uploaded structures as a basket). Once it is done, you can map your descriptors to the MOLECULEID and upload descriptors to [the descriptor storage](#) (see further instructions there). The uploaded descriptors will be shown along with other descriptors on the respective descriptor selection page. Naturally, only molecules with uploaded values could be used for model development.



## 8. How can I upload my own data?



---

Yes, of course. You can find instructions how to upload data and interface to do it at [Batch Upload upload](#) page. See also [tutorial](#) which shows the required steps.

Briefly, your data should contain (at least) structural information as SMILES or SDF, molecular name (if available) and the property data. Select the name of the property from the list of [available properties](#). Also, use units as there were originally provided (it will provide you a possibility to easily identify and track errors). OCHEM will provide an automatic conversion to any other units, which could be required for modelling. Importantly, link your data to an article from which they were uploaded. This will again, will help you to be not lost in reviewing them later.

## 9. I have quantitative data. How can I build a classification model?

---

First select records, which you would like to convert to qualitative properties, to a basket (see [this link](#) how to work with data). Click "Edit basket" (e.g., go to [Database/Basket](#) and click basket edit ). After this use  [Discretize the numerical values](#) option to select a threshold (or several of them) to create records with qualitative values. These newly created records can be used for classification models.

## **10. Can I build a classification model for more than two classes?**

---

Yes, there is an experimental support of this option using all models which support MTL (Multi Task Learning) method. However, for this type of analysis some functionality is not fully supported (e.g., applicability domain calculation, etc.) and it is recommended only for exploration purposes.

## 11. My property is missing in the set of OCHEM properties. Can I create a new one?

---

First of all spend few minutes to find this property. Frequently property may exist but with a different name, e.g. logKow or Octanol/water partitioning coefficient is named as logPow in ochem. Just try to search for it using synonymous words. If you found it, there is no need to create it again. If property does not exist, you can easily create it. Go to browser of Properties and click "create new property". After this select its type: Numeric for properties which are numeric, e.g. have float values or Classification for properties like AMES, which could be "active" or "inactive" mutagens. For Numeric property correctly select "System of units". If your data are concentrations, select "Concentration". For properties like logPow, which are just ratios or logarithm of ratios, select default "Dimensionless". This is very important since will allow you to upload data, e.g. in mg/kg but develop models in log(mol/l): OCHEM provides an automatic conversion of units. For classification property add available classes, e.g. "active" or "inactive". Options for classification property should not be float/integer values, e.g. 0/1 will not be allowed by OCHEM. Finally provide some brief description of the property.

Download Ames training or test sets from <http://ochem.eu/article/4027> to see an example how to prepare data for Classification properties.