# Melting Point models

## Dataset profile

The available at OCHEM models predict Melting Point (MP) of organic chemical compounds. The MP is one of the important physico-chemical properties, which is frequently used in drug discovery to estimate aqueous solubility of chemical compounds. The complexity with prediction of this point are connected to purity of compounds, existence of polymorphic forms, degradation of compounds before melting, etc. All these factors influence the quality of models for this point. The data for MP were collected in OCHEM database as well as were provided by ChemExper database (OCHEM dataset) and Enamine Ltd (Enamine dataset) The majority of data were organic chemistry compounds. The models were validated using 277 compounds compiled by [Bergstrom et al Molecular descriptors influencing melting point and their role in classification of solid drugs. J. Chem. Inf. Comput. Sci. 2003; 43 (4) 1177-85] as well as data from Open Notebook.

## Data preprocessing

All chemical structures were processed using OCHEM cleaning and standardization protocols. A specific care was used to eliminate salts and mixtures, and inorganic compound, which could dramatically change MP of molecules. The detection and elimination of outliers was done based on p-value (article in preparation).

## Descriptors

Models were built 11 individual descriptor packages available in OCHEM. A simple average of all 10 models was done to develop consensus model. This model, however, requires rather long calculations, especially if calculations of descriptors have not been previously cached. There is also 2D model, "Melting Point best (Estate)", which was built using Estate descriptors. All other sub-models are not shown, just to avoid confusion with having too many of them in the web browser (they can be accessed using public IDs indicated in the profile of the consensus model as https://ochem.eu/model/MODEID).

## Validation

The model were built using 5-fold cross validation as well as prediction of subsets (e..g, model developed using OCHEM subset was used to predict Enamine, Bergstrom and Bradley sets, etc.)

## Statistical parameters

### Prediction accuracy

The basic prediction accuracy parameters according to the 5-fold cross-validation procedure (N=25547) are:

| **Property** | RMSE | MAE | $R^2$ | $r_2$ (Coefficient of determination) |
|---|---|---|---|---|
| **Consensus model** | 37.1 | 27.6 | 0.78 | 0.78 |

| | | | | |
|---|---|---|---|---|
| **Melting Point best (Estate)** | 39.6 | 29.1 | 0.75 | 0.75 |

The accuracy for drug-like subset (molecules with melting point in [50,250]°C interval is less than 33°C for the consensus model.

## Reference

The full details of the study are published in *How accurately can we predict the melting points of drug-like compounds?* [Tetko IV, Sushko Y, Novotarskyi S, Patiny L, Kondratov I, Petrenko AE, Charochkina L, Asiri AM. *J Chem Inf Model.* 2014 Dec 22;54(12):3320-9. doi: 10.1021/ci5005288.]

# Availability

All data can be publicly downloaded at http://ochem.eu/article/55638.