PREDICTION OF SUPERCRITICAL CARBON DIOXIDE SOLUBILITY BASED ON DECISION TREE FORMALISM

S. Batin[a], P. Gurikov[a], I. Smirnova[b], N. Menshutina[a]

[a]Mendeleev University of Chemical Technology of Russia, Miusskay sq., 9 , Moscow, 125047, Russia

[b]Technical University of Hamburg - Harburg, Chair of Separation Science and Technology - V8 - Eiβendorferstr. 38, 21073 Hamburg, Germany

Nowadays in many areas of chemistry and chemical technology the use of supercritical fluids (SCF) as solvents in various processes is under intensive research. Adsorption, extraction and impregnation with supercritical carbon dioxide are most extensive areas of SCF application because solvent has low critical parameters and solubility can be easily regulated by varying pressure and temperature. In order to collect, store and analyze numerous solubility data published during previous 30 years we have developed unified information system (IC-SCF). IC-SCF includes:

1. Database containing information about molecular structure of substances (more than 10 000 entries), their physicochemical properties (molar mass, melting and boiling points) and solubility values (single-component and binary solubility in supercritical carbon dioxide).
2. Prediction modules are based on QSPR methodology, the theory of molecular similarity and also on data mining methods. Such modules allow to build mathematical models both manual and automatic for the chemical diversity.
3. Module for automatic calculation of descriptors for QSPR analysis. The calculation of descriptors is based on 2D and 3D molecular information (graphs, quantum chemical calculations, string notations).
4. Module for import and export of chemical compounds, that allows to exchange data from one chemical format to another. Module maintains the most extended formats of molecular structure (CML, MDL, ChemDraw CDX, ChemDraw XML, MRV, SVG, etc.), as well as identifiers of chemical substances (SMILES, InChI, CAS).

The proposed information system was used to analyze data on the solubility for a diverse set of chemical compounds (alkanes, esters, aromatics) using the formalism of decision tree. Topological descriptors were used as classifiers. Each node of the tree is associated with their linear regression equation. Residual variance is less than 0.95 for both the test and for the training sample. Shown that the discriminating factor is the Ghose – Grippen molar refraction. The approach has shown its high predictive power and can be recommended for estimating the unknown solubility under different pressure and temperature.