Quantitative Structure-Property Relationship (QSPR) Study of Critical Properties: Database of Predictions and Uncertainty Analysis.

Andrei Kazakov [C, S], Chris D. Muzny, Vladimir V. Diky, Robert D. Chirico and Michael Frenkel
Thermophysical Properties Division, National Institute of Standards and Technology, Boulder, CO, U.S.A.

In our previous work, we presented development of a QSPR model for prediction of critical parameters for pure compounds derived exclusively from available raw experimental data. The approach was shown to be highly competitive with the existing (mainly group contribution-based) methods both in terms of its accuracy and the scope. However, several issues remained unresolved and are addressed in the present work. First, the application of the QSPR-based method to a new compound is rather cumbersome as it necessitates generation of optimized three-dimensional molecular structure followed by calculation of all necessary descriptors, procedures that generally require human expert supervision and specialized software. To facilitate access to the QSPR prediction methods, we initiated development of large-scale database of precomputed model results. A wide range of two-dimensional molecular structure representations was adopted from a number of publicly-accessible depositories as well as generated in-house and indexed according to The IUPAC International Chemical Identifier (InChI) within a relational database system. For each structure, an initial three-dimensional representation was generated using distance geometry and molecular mechanics force field geometry optimization with, where applicable, enforcement of specific stereo assignments. Initial three-dimensional structures were further optimized on a semiempirical PM3 level and subsequently used to calculate descriptors and to produce QSPR predictions that were also stored in the database. As a result, the database can be searched for predicted values by molecular structure of interest. The database is also expected to expand over time to address demands for compounds that are presently not included and will be made available for public access in near future. Another critical issue that needed to be addressed for practical applications is quantification of uncertainties of QSPR predictions. Using the generated database of molecular structures and their descriptors (with diversity much greater than that of the original subset used to produce the QSPR model itself), the prediction confidence intervals were estimated via Monte Carlo sampling procedure.