[L9] De novo Structure Generation Using Inverse QSAR Approach

Kimito Funatsu[1, 2]

[1]*Department of Chemical System Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.*
[2]*Data Science Center, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan.*

Quantitative structure-property relationship (QSPR) or quantitative structure-activity relationship (QSAR) is a way to find a quantitative relation between compounds and their corresponding property or activity in a statistical manner. Property or activity is usually numerical and, therefore, can be represented as an objective variable: y. To treat compounds numerically, they are usually transformed into sets of descriptors (**x**), which are the abstract representation of a molecule. Therefore, a QSPR/QSAR model can be regarded as a regression model (y=f(**x**)) constructed with an experimental dataset. Inverse QSPR/QSAR-based molecular design is to generate chemical structures satisfying a specific y value through backward analysis of a pre-constructed QSPR/QSAR model. Compared with molecular design with QSPR/QSAR, such as virtual screening, inverse QSPR/QSAR is promising since it may generate structures *de novo*. There have not been, however, methodologies for inverse QSPR/QSAR, which can be applied to practical applications, using various descriptors, nonlinear regression methodologies, and considering an applicability domain (AD). ADs limit the chemical space in a way that, only inside AD, predicted values produced by regression models should be trusted. ADs must be considered when applying QSPR/QSAR models to novel chemical structures.

The important contribution of this work is to develop a practical chemical structure generation system based on inverse QSPR/QSAR analysis. In order to make the system feasible, several methodologies have been proposed and implemented by the author, which previous inverse QSPR/QSAR analyses could not consider, namely:

1. To introduce a nonlinear regression methodology for capturing nonlinear relationship between **x** and y.
2. To develop a methodology for considering an AD with a probabilistic density function (PDF) as a Gaussian mixture model (GMM). Based on the premise that inside dense areas in chemical space reliability of predicted value is high, posterior PDFs of **x** given a y value p(**x**|y) possesses both the degree of predictive reliability and closeness to the y.
3. To propose novel algorithms for structure generation. Structures are efficiently constructed by combining ring systems and atom fragments as building blocks by making use of the canonical construction path method proposed by McKay.
4. To introduce and implement monotonous changing descriptors (MCDs) in the developed structure generator and the proposed inverse QSPR/QSAR system. MCDs are descriptors whose values monotonously change by adding a building block to a structure. Since wide range of descriptors can be categorized as a MCD, the construction of regression models with high predictability is expected.