molecular informatics

DOI: 10.1002/minf.200900038

Exhaustive Structure Generation for Inverse-QSPR/QSAR

Tomoyuki Miyao,[a] Masamoto Arakawa,[a] and Kimito Funatsu*[a]

Abstract: Chemical structure generation based on quantitative structure property relationship (QSPR) or quantitative structure activity relationship (QSAR) models is one of the central themes in the field of computer-aided molecular design. The objective of structure generation is to find promising molecules, which according to statistical models, are considered to have desired properties. In this paper, a new method is proposed for the exhaustive generation of chemical structures based on inverse-QSPR/QSAR. In this method, QSPR/QSAR models are constructed by multiple linear regression method, and then the conditional distribution of explanatory variables given the desired properties is estimated by inverse analysis of the models using the framework of a linear Gaussian model. Finally, chemical structures are exhaustively generated by a sophisticated algorithm that is based on a canonical construction path method. The usefulness of the proposed method is demonstrated using a dataset of the boiling points of acyclic hydrocarbons containing up to 12 carbon atoms. The QSPR model was constructed with 600 hydrocarbons and their boiling points. Using the proposed method, chemical structures which had boiling points of 100, 150, or 200 °C were exhaustively generated.

Keywords: Inverse-QSAR · Inverse-QSPR · Molecular design · Structure generation · Chemoinformatics · Drug design

1 Introduction

Quantitative structure property relationships (QSPR) and quantitative structure activity relationships (QSAR) have been widely used in the pharmaceutical industry and other fields to design molecules with a desired combination of properties. The basic purpose for using QSPR and QSAR is to construct statistical models to reveal the relationships between chemical structures and their biological activities or physicochemical properties.[1] QSPR/QSAR models that have been successfully trained and scientifically validated are useful for predicting the properties and activities of chemical structures whose properties are still unknown. In addition, a physicochemical and/or mechanistic interpretation can be expected from the obtained model parameters. The ultimate goal of molecular design, however, is not to construct such models, but to propose promising chemical structures that have desired properties.

Virtual screening[2,3] based on QSPR/QSAR models is a widely used method to effectively find objective molecules. A number of studies involving virtual screening for various targets have been reported, mainly in the field of drug discovery.[4] In virtual screening, chemical descriptors of each molecule included in a chemical library are calculated. Then, the predictive values of their objective properties are estimated by using QSPR/QSAR models.

The success of virtual screening studies depends on the quality of the virtual library and the QSPR/QSAR models used. Although accurate and predictive models are necessary for the success of virtual screening, the quality of the virtual library is a more important factor. Even if QSPR/QSAR models with high predictive power are successfully constructed, we cannot find satisfactory chemical structures for lead candidates from a library which contains only small size and/or the small number of compounds.

One possible way to form a virtual library for practical use is to collect molecules from published chemical libraries such as PubChem,[5] ZINC,[6] and various free or commercially available databases. The advantage of this approach is that one can easily collect many feasible structures whose stability and synthetic feasibility are already assured. However, because such libraries include only known chemicals, one cannot expect to discover a truly new scaffold. Some method of structure generation is therefore needed in order to propose new chemical structures.

Chemical structure generators had been developed in 1970's and 1980's mainly for structure elucidation.[7] These programs first generate all possible chemical structures having the predefined molecular formula. For example, CHEMICS[8] is based on the concept of connectivity stack, which allows an exhaustive and unique enumeration. In the generator proposed by Bangov,[9] atoms and bonding sites of the unknown compound are partitioned into equivalent classes. And then, the fragments are iteratively combined

[a] *T. Miyao, M. Arakawa, K. Funatsu*
Department of Chemical System Engineering,
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113–8656, Japan
phone: (+ 81) 03-5841-7751, fax: (+81) 03–5841-777
**e-mail: funatsu@chemsys.t.u-tokyo.ac.jp*
funatsu@chemsys.t.u-tokyo.ac.jp

each other. For every level an exhaustive combination is performed, and redundant structures are eliminated.

The structure generation system MOLGEN[10] is one of the fastest and most sophisticated programs that applies the concept of homomorphism. MOLGEN can efficiently generate a large number of virtual structures under some primitive constraint such as molecular formula, the number of double bonds, and so on. Rücker et al. have proposed the method MOLGEN-QSPR,[11,12] which aims to deal with the problem of molecular design. Many chemical structures were first systematically generated under certain restrictions by MOLGEN, and then filtered according to their properties which are estimated by a QSPR model.

The GDB proposed by Fink et al.[13] is another notable virtual library; it contains about 26.4 million virtual structures. The GDB was constructed through the exhaustive generation of possible organic structures containing up to 11 atoms of C, N, O, and F. Some additional constraints related to chemical stability and synthetic feasibility were also applied to prevent the undue proliferation of the number of structures generated. The availability of the GDB as a drug screening library was verified by Hert et al. through a comparison of KEGG and DNP databases.[14]

Although MOLGEN and the method proposed by Fink are available for the formation of virtual libraries, it is impossible to cover the entire realm of the chemical universe because of its vastness; it includes more than 10^{60} compounds.[15] For this reason, a generator intended for molecular design has to be able to selectively generate chemical structures that are expected to have desired properties. This problem is known as inverse-QSPR/QSAR.[16] The ultimate goal of inverse-QSPR/QSAR analysis is to enumerate all those chemical structures that indicate desired properties or activities without generating undesired structures.

One possible way to achieve such a selective generation is to use an evolutionary algorithm (EA)[17] or a genetic algorithm (GA).[18] The basic idea of EA and GA is to try to mimic the effects of natural evolution. In EA and GA, chemical structures are evolved so as to have higher predicted activities. The evolution of structures is implemented by mutation and crossover. Mutation is an operator to partially change chemical structures; for example, an atom type and/or bond order are randomly changed to other ones. The crossover operation combines the substructures of two or more structures so as to breed a combined structure.

Brown et al.[19] proposed a novel method for de novo design called median molecule workflow (MMW). MMW consist of three modules: a de novo design module, a molecular scoring module, and a multi-objective ranking module. In the de novo design module, a compound generator (CoG) is used to generate chemical structures. A CoG is GA-based generator that uses graph-based chromosomes to represent chemical structures in a population. The advantage of this kind of generator is that it effectively generates a relatively a large-sized molecules. By using MMW, various properties are taken into account simultaneously.

The obvious drawback of the EA-based method is that the generation of structures is not exhaustive. In other words, the possibility remains that better structures might exist. As mentioned above, most of the research in molecular design based on QSPR/QSAR models has suffered from the limitation that it is quite hard, or almost impossible, to generate exhaustive chemical structures that satisfy desired properties. In addition to addressing this problem, it is also important to consider the applicability domain (AD) [20] of models, because the values of their properties are calculated through QSPR/QSAR models.

In order to resolve these problems, we have developed a structure generation system based on the inverse-QSPR/QSAR method. Inverse-QSPR/QSAR can resolve the problem of explosion of the number of candidate structures. This is because chemical structures which have low activities or undesired properties must not be generated and estimated using QSPR/QSAR models. This useful method can make us generate all of the aiming chemical structures. The proposed system can exhaustively generate chemical structures whose properties are desirable. In the proposed system, chemical structures are generated by adding one atom at a time to a growing structure. The notable idea used in this system is its use of a descriptor, which is monotonically changed by adding an atom. By using monotonically increasing or decreasing descriptors, the structure generation process can be accelerated, as will be discussed later. We call this type of descriptor a monotonically changing descriptor (MCD).

A limited number of studies on structure generation using the inverse-QSPR/QSAR method have been reported. A novel method for generating exhaustive structures with inverse-QSPR was developed by Faulon et al.[21] They used specific descriptors, called signatures, to construct QSPR models. After first constructing models, they inversely analyzed them and generated exhaustive structures by combining the signature descriptors. The difference between Faulon's method and ours is not only the algorithm that is used to generate chemical structures but also the type of descriptors used to construct the QSPR/QSAR models. In our method, any descriptors can be used, as long as their values are monotonically changed by attaching an atom to a growing structure. Therefore, more flexible QSPR/QSAR models can be constructed.

To demonstrate our system, we applied it to a dataset of boiling points of acyclic hydrocarbons. The boiling point is one of the most elementary properties, and thus, has been widely studied for a long time. In the field of chemoinformatics, many studies have been conducted on the QSPR method for predicting boiling points.[22]

This paper is arranged in the following manner. First, an overview of the proposed method for inverse-QSPR/QSAR is described. Second, three modules in the system are explained in detail. Third, the results of the application of the proposed method to a set of acyclic hydrocarbons is provided, along with a discussion of how our system was suc-

cessful. Finally, we briefly summarize our findings and propose future directions for further study.

2 Method

2.1 Overview

The purpose of this study is to propose a method for the exhaustive generation of chemical structures having desired properties using QSPR/QSAR models. An overview of the proposed method is illustrated in Figure 1.

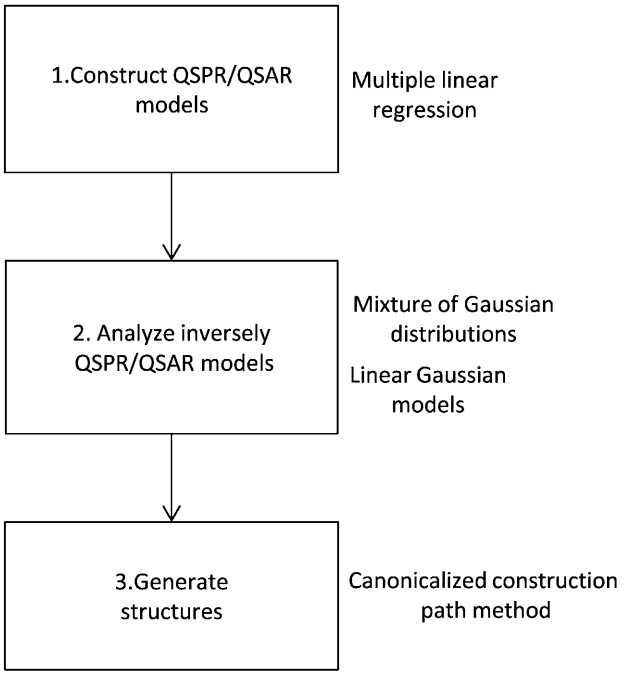


Figure 1. Flowchart of exhaustive generation of chemical structures having desired properties. 1. QSPR/QSAR models are constructed using multiple linear regression with a least square estimator. 2. The constructed QSPR/QSAR models are analyzed inversely using linear Gaussian models. 3. Structures are generated exhaustively through a canonicalized path.

First, QSPR/QSAR models are constructed using the method of multiple linear regression with a least square estimator (MLR/OLS).Second, the constructed MLR/OLS models are analyzed inversely using a linear Gaussian model[23] to obtain a conditional distribution of explanatory variables, given the desired value of a dependent variable. The MLR/OLS method is used in the first step to derive the conditional probability distribution without convergence calculation. In MLR/OLS, given the value of **x**, the corresponding value of the dependent variable has a Gaussian distribution with the mean equal to the predicted value of *y*. Provided that we model the probability distribution of a random variable **x**, given a dataset, by using normal distribution, we can make use of the linear Gaussian models analysis framework. In our method, a mixture of Gaussian distributions[24] is used to characterize the data distribution. Finally, chemical structures having desired properties are generated exhaustively. The generation algorithm is based on the canonical construction path method of McKay.[25] The chemical structures are produced by augmenting, with the acceptance of only structures that are made via a canonical augmentation. In this study, the augmentation is implemented by attaching an atom to a growing structure. As mentioned above, the proposed system adopts MCD. Since the values of MCD monotonically change with the attachment of an atom, effective pruning of the branch of the generating tree is achieved.

2.2 QSPR/QSAR and Inverse-QSPR/QSAR Analysis

In this section, a mathematical formulation of the inverse-QSPR/QSAR analysis of this method is described. The dimension of explanatory variables is *k*, and a dependent variable is scalar. The purpose of the inverse-QSPR/QSAR analysis is to derive the conditional probability distribution for explanatory variables **x**, given the value of dependent property *y*, $p(\mathbf{x}|y)$.

2.2.1 Multiple Linear Regression with a Least Square Estimator (MLR/OLS)

QSPR/QSAR models are constructed using MLR/OLS. It is convenient to define an additional dummy variable so that a bias parameter can be included in the MLR coefficient vector **A**. The conditional probability of the objective variable, given an explanatory variable in MLR/OLS, is

$$p(y|\mathbf{x}) = N(y|\mathbf{A}\mathbf{x},\sigma^2) \tag{1}$$

In Equation 1, the conditional probability distribution of *y* given **x** is a normal distribution with the mean **Ax**, and the variance of σ^2. Consider a training set comprising *n* observations $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^T$ and their corresponding target values $\mathbf{y}=(y_1,\ldots,y_n)^T$. If the data are assumed to be drawn independently from the distribution 1, then the likelihood function is given by the following equation:

$$p(\mathbf{y}|\mathbf{x},\mathbf{A},\sigma^2) = \prod_{i=1}^{n} N(y_i|\mathbf{A}\mathbf{x}_i,\sigma^2) \tag{2}$$

Maximizing the likelihood of Equation 2 with respect to **A** and σ^2, the maximum likelihood solutions $\mathbf{A}_{ML}$ and σ^2_{ML} can be obtained.

$$\begin{aligned}\mathbf{A}_{ML}^T &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\\ \sigma^2_{ML} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{A}_{ML}\mathbf{x}_i)^2\end{aligned} \tag{3}$$

Although the maximization of the likelihood function is achieved without convergence calculation as Equation 3, we need to be careful of the collinearity which is a cause of rank deficient.

An assessment of the statistical significance of constructed MLR/OLS models is estimated by a t-test for the components of regression coefficient, an *F*-test for the regression coefficient in the total, and the value of the determination coefficient for training data R^2 and for test set R_{pred}^2.

2.2.2 Mixture of Gaussian Distributions

In order to derive the conditional probability distribution of explanatory variables **x**, given a dependent property value of *y*, the shape of distribution for an observed dataset in descriptor space is assumed to be a Gaussian distribution. A mixture of Gaussian distributions is employed in order to capture the complex form of the distribution of the dataset. A Gaussian mixture distribution can be written as a linear superposition of *M* Gaussian densities of Equation 4,

$$p(\mathbf{x}) = \sum_{i=1}^{M} \pi_i N(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{4}$$

where *M* is the number of the mixture, π is the mixing coefficient, $\boldsymbol{\mu}$ is the mean vector of each component, and $\boldsymbol{\Sigma}$ is the covariance matrix.

We now introduce the binary latent random variable, z_k. The random variable z_k of 1 represents its belonging to the *k*-th clump. An efficient method for finding maximum likelihood solutions for models with latent variables is the expectation-maximization (EM) algorithm. The following is the likelihood function of a mixture of Gaussian distributions, and is to be maximized using the EM algorithm. The likelihood function to be maximized is in Equation 5

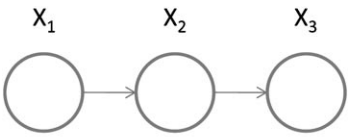
$$p(Data|\theta) = \prod_{j=1}^{n} \sum_{i=1}^{M} \pi_i N(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{5}$$

where *Data* denotes a dataset with n observations. The θ represents a collection of parameters to be estimated. The notation of the EM algorithm is described in the literature of Bishop.[26]

Although a maximum likelihood solution can be effectively found, a problem remains to be addressed, namely how many components of a mixture are needed. In this study, the number of components is determined by the value of the Bayesian information criterion (BIC).[27] The higher the BIC value is, the more appropriate is the probability density function comprised of the mixture of Gaussian distributions. It should be noted that it is possible to determine the number of components automatically by means of a fully Bayesian treatment of a mixture of Gaussian distributions.[28]

2.2.3 Inverse Analysis Using a Linear Gaussian Model

By using the methods described above, both the probability distribution of a dataset $p(\mathbf{x})$ formed by a mixture of Gaussian distributions, and the conditional distribution for the dependent variable given a value of explanatory variables, are obtained. Using these two probability distributions, $p(\mathbf{x})$ and $p(y|\mathbf{x})$, the conditional probability distribution for **x** given *y*, $p(\mathbf{x}|y)$, is calculated by applying the framework of a linear Gaussian model. Several widely used techniques in the field of cheminformatics are regarded as applications of this method, for example, probabilistic principal component analysis (PPCA),[29] factor analysis (FA),[30] and linear dynamical systems.[31]



$$x_i = \sum_{j \in pa_i} w_{ij} x_j + b_i + \sigma_i \varepsilon_i$$

Figure 2. Graphical model for a linear Gaussian model. The upper schematic is a directed graph over three Gaussian variables, with one missing link. The equation is its mathematical formulation.

In Figure 2, the nodes represent random variables in the constructed model. This graphical model implies that the random variable x_3 is conditioned by x_2, and also that the random variable x_2 is conditioned by x_1, whereas x_1 is an independent random variable. The lower part of Figure 2 displays the formulation of the model where σ is a standard deviation, pa_i is a set of nodes of the parent of node *i*, ε is a random variable having a standard normal distribution, and w_{ij} and b_i are parameters related to the mean. By replacing x_i with *y*, it can be applied to QSPR/QSAR models.

As a result of Sections 2.2.1 and 2.2.2, Formulas 1 and 4 are used to derive the conditional distribution of **x** given *y*, $p(\mathbf{x}|y)$. That conditional distribution is obtained without convergence calculation due to Gaussian distributions (see [32]) in Equation 6.

$$p(\mathbf{x}|y) = \sum_{i=1}^{M} \omega_i N(\mathbf{x}|\Delta_i\{\mathbf{A}^T\sigma^{-2}y + \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\}, \Delta_i)$$
$$\Delta_i = \left(\boldsymbol{\Sigma}_i^{-1} + \mathbf{A}^T\sigma^{-2}\mathbf{A}\right)^{-1} \tag{6}$$

where ω represents each of the posterior mixing coefficients. $\boldsymbol{\Delta}$ is the covariance matrix of each component. The conditional probability of **x** given *y* is represented by Equation 7.

$$p(\mathbf{x}|y) = \sum_{k=1}^{M} p(\mathbf{x}|y, z_k)p(z_k|y) \quad (7)$$

The conditional probability distribution **x** given y and z_k is defined as Equation 8.

$$p(\mathbf{x}|y, z_k) = N\left(\mathbf{x}|\Delta_k\{\mathbf{A}^{\mathrm{T}}\sigma^{-2}y + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\}, \Delta_k\right) \quad (8)$$

By comparing Formulation 8 with 7, the following Equation 9 is obtained.

$$\omega_k = p(z_k|y) = \frac{p(y|z_k)p(z_k)}{\sum_{i=1}^{M} p(y|z_i)p(z_i)} \quad (9)$$

where $p(z_k)$ is obtained as the result of parameter estimation with the EM algorithm as Equation 10.

$$p(z_k) = \pi_k = \frac{n_k}{n} \quad (10)$$

n_k is the number of observations assigned to cluster k.

The probability of $p(y|z_k)$ is the conditional distribution of y, given the clump. There is no relationship between probability $p(y|z_k)$ and **x**. According to the formula of marginal Gaussian distribution in the framework of a linear Gaussian model, the following Equation 11 is obtained.[32]

$$p(y|z_k) = N(y|\mathbf{A}\boldsymbol{\mu}_k, \sigma^2 + \mathbf{A}\boldsymbol{\Sigma}_k\mathbf{A}^{\mathrm{T}}) \quad (11)$$

The abovementioned formulation provides a way to derive the conditional distribution of **x** given y in descriptor space.

In this chapter, we illustrate a method with which to obtain the conditional distribution of **x** given y. The important point is to use MLR/OLS models as QSPR/QSAR models. After constructing models, the framework of linear Gaussian models made it possible to obtain the conditional distribution of **x** given y with the help of a mixture of Gaussian distributions.

2.2.4 Target Region

We have to determine the region in descriptor space where chemical structures are exhaustively generated according to the conditional probability distribution of **x** given y. The probability density in the region should be relatively high. In addition, the shape of the region is also important. Considering the generation strategy described in the next section, the shape of the upper-right border of the region must be convex polyhedral. In this study, for the sake of simplicity, we represent the high-density area as hyper-rectangular in high-dimensional space. Using the marginal distribution of **x**, $p(\mathbf{x})$, another region which represents AD is obtained. Thus, we combine these two areas into one targeted region to achieve valid structure generation.

2.3 Structure Generator

2.3.1 Structure Generation Algorithm

For the purpose of the exhaustive generation of chemical structures in a specific region of descriptor space, McKay's canonical construction path method[25] was adopted in our structure generator. It is one of the fastest graph generation methods that does not also generate isomorphic structures.[33] Chemical structures are treated as chemical graphs[34]; vertices represent chemical atoms and edges represent chemical bonds. Vertices are labeled according to the corresponding atoms and edges are labeled with the types of bonds.

Roughly speaking, new chemical graphs are obtained by attaching an atom to a growing structure. A seed structure, which is the first structure entered in the system, is usually a single atom. An atom is then attached in every possible way, but without exceeding the atom valences of an attached atom, and of the atoms in the present structure. This indicates that it is possible to generate structures having plural rings by attaching an atom and connecting it to plural vertices in an attached structure. By repeating this process, a rooted tree is formed whose nodes are the generated chemical structures. In this study, we adopt breadth-first search of a tree-searching algorithm for exhaustive structure generation.

2.3.2 Descriptors

Due to finite time and limited memory capacity, it is important to not generate structures that are located outside of the region. For this reason, we have proposed the use of a specific type of descriptor, MCD. Figure 3 illustrates the concept of how generated structures assume locations in descriptor space. The target area is depicted as a rectangle

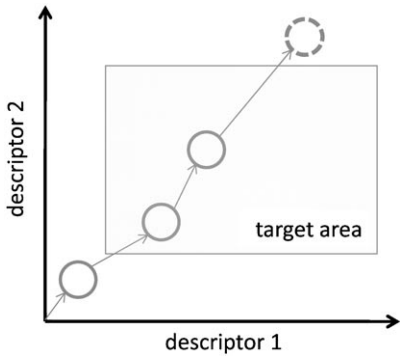


Figure 3. A growing path in descriptor space. Circles represent generating structures and arrows represent the relation between parent and child. In this example, the parent generates only one child, so that the circles and arrows are a path in the generating tree in descriptor space. The target area is the place where exhaustive structure generation is executed. The circle formed by a dotted line represents the structure that is deleted because it is beyond the target area.

in a two-dimensional descriptor space. The arrows and circles represent a path in the generation tree. The circles represent the generated structures, and the arrows represent the relations of two structures, parent and children. The children are born from the parent by the attachment of an atom. By using MCD, the descriptor values naturally move upward and to the right from their present location in the descriptor space, according to the structure modification. Thus, by calculating MCD values during the generation process, and rejecting those structures whose MCD values exceed the target area, efficient pruning is achieved.

Descriptors that are proportional to molecular size are examples of MCDs; the number of carbon atoms, the number of double bonds, and the number of specific fragments are MCDs. Some topological descriptors are also MCDs. Topological descriptors generally have more useful information on chemical structures than do constitutional ones.[35] Thus, by using this type of descriptor, regression models can be constructed that are more appropriate than those constructed with only constitutional descriptors. Some examples of topological MCDs are the Randic index,[36] autocorrelation descriptors,[37] 2-dimensional RDF descriptors,[38] and some kinds of descriptors based on specific bond length (walk count).[39]

2.3.3 Generation Flow

A flowchart for the exhaustive generation of chemical structures in the desired region is provided in Figure 4.

The structures are generated by a breadth-first search with the data structure of the queue, by trying to attach atoms to the head structure of the queue. In the unit where structure is being extended, all kinds of atoms are attached to that structure in every possible way, including the formation of rings. In attaching an atom to the structure, canonicalized path test and descriptor value test are performed. The unit where structure is being extended is the most important unit in this module. Figure 5 illustrates the details of the procedure of this structure extension using as an example the chemical structure of $CH_2NCH_2NCH_2$.

On the left side of Figure 5, the way to extend the structure of $CH_2NCH_2NCH_2$ is illustrated. In our generation module, we treat this chemical structure as an H-depleted chemical structure. First, the atoms in the structure are searched. These atoms must have the capacity for an extra atom. The atom valences of the two nitrogen atoms in the structure are three. An extra atom cannot be attached to the places of these nitrogen atoms. The places where an extra atom can be connected are marked by *.

Second, an oxygen atom is attached to the structure in every possible place and in every possible way, yet without generating isomorphs; this is achieved by considering the automorphism of the structure (see [33]). In this case, the atom type is limited only to oxygen. As a result of the second step, six nonisomorphic candidate structures are obtained.

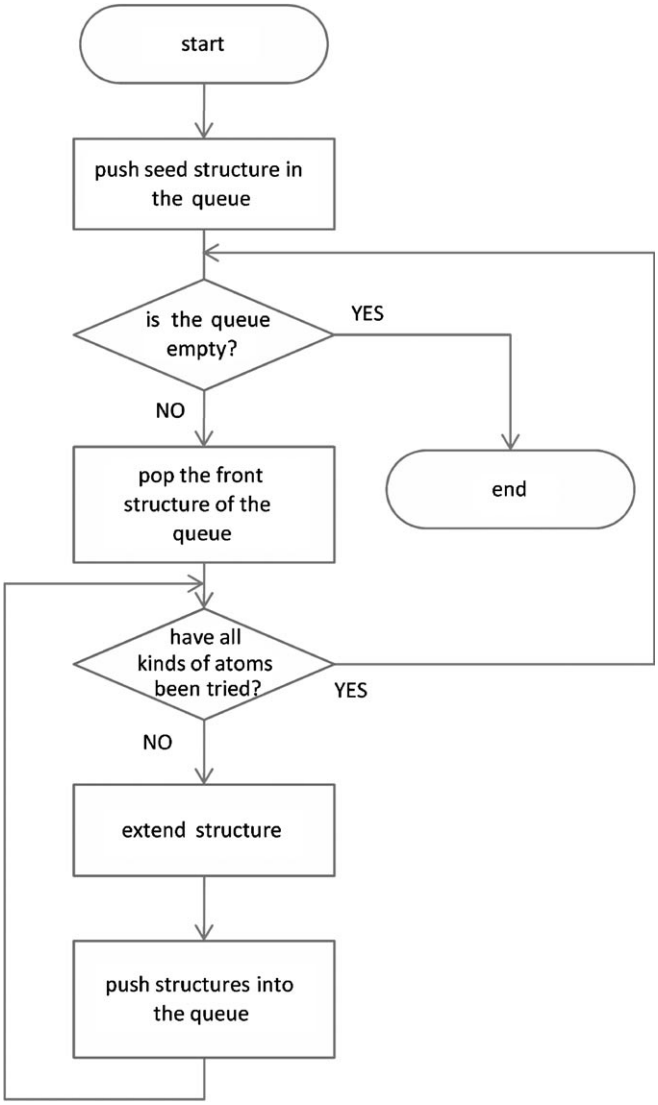


Figure 4. Flowchart of structure generation module. A breadth-first search algorithm is implemented using an FIFO queue. The head structure of the queue becomes the seed structure for the production of the next generation. An atom is attached on the basis of the structure's symmetry.

Third, the canonical construction path is checked by using the canonical choice function. The return value of the function is an orbit of atoms (the orbit is determined according to the concept of automorphism). The canonical path is checked by identifying the newly added atom with one of the atoms in the orbit as a return value of having applied the canonical choice function (see [25]). The input of the function is candidate structures. In this example, we define the canonical choice function as follows:

1: If the lowest degree in the structure is one, return the orbit of atoms having the highest atomic number in the atoms with the lowest degree.

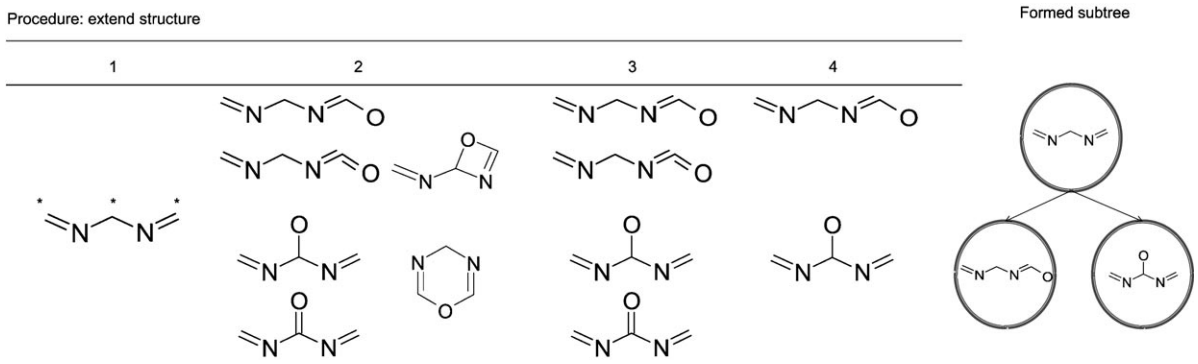


Figure 5. Example of how a structure is extended in the unit of 'extend the structure.' Left: the process of gaining new children. 1: All atoms in the structure are searched for where an extra atom can be connected. The places are marked by *. 2: An oxygen atom is attached to the structure in every possible way. Six structures are obtained as potential children. 3, 4: The result of checking canonical path and descriptor values in the phase 3 or 4, respectively. The surviving structures are illustrated. Right: Formation of subtree in the generation as a result of obtaining descendents.

2: If the lowest degree in the structure is two, return the orbit of atoms having the lowest atomic number in the atoms with the lowest degree.

Applying this canonical choice function, four candidate structures are obtained as a result of the third step.

Finally, we test whether or not the descriptor value of the surviving structures is within the descriptor's range. In this case, the MCD that was used is "the number of double bonds" whose upper bound is 2.

Eventually, the two candidate structures are found to be real children of the parent structure, $CH_2NCH_2NCH_2$. The process of gaining children corresponds to create a subtree in the generation tree. This is illustrated on the right side of Figure 5.

In general, when using the canonical choice function, the information regarding the automorphism (or at least the information regarding the orbit) of the candidates must be known. In this study, we apply the NAUTY algorithm by McKay[40] to recognize it efficiently.

3 Results and Discussion

3.1 Dataset of Boiling Point

In order to demonstrate the superiority of the system of our proposed inverse-QSPR/QSAR method, we have enumerated all possible structures with a desired property. The target property is set to the boiling point (*bp*). We used the following MCDs: *nSK* is the number of main atoms; *nDB* is the number of double bonds; *MWC02* is the molecular walk count with order 2, *MWC03* is the same with order 3, and *MWC04* is with order 4. *SRW04* is a self-returning walk count with order 4. *X1* is the Randic connectivity index. In this work, a set of 882 acyclic hydrocarbons containing from 1 to 12 main atoms was studied. The experimental values of *bp* were measured at near-atmosphere pressure (750–770 mmHg). The data were collected from the Beilstein database.[41] A random selection of 600 samples from the database were used for the training set that was used for constructing models, and the rest of them were used for the test set that was used to validate the constructed models.

Figure 6 shows the histograms of each descriptor value for the training set, and those of the objective property (*bp*) values of both the training and test sets.

The histograms show that the density distributions of the descriptor values were well characterized by continuous probability distributions, except for *nDB*. Because *nDB* is a discrete variable, it is unnatural to approximate the density model of *nDB* by continuous distribution such as a mixture of Gaussian distributions. Because the families of parametric models should be determined according to the type of variables, *nDB* should be described by a categorical distribution in which the random variable can take on one of K-mutually exclusive states. In this study, we characterized the density of *nDB* by a mixture of Gaussian distributions, as well as the rest of the MCDs, in order to make it easy to perform the following analysis.

3.2 Multiple Linear Regression

Six hundred samples were used to construct the MLR/OLS model. The regression analysis of autoscaled data resulted in a determination coefficient, $R^2 = 0.949$, a determination coefficient for the test set, $R_{pred}{}^2 = 0.949$, and *F*-statistics, 1587. We succeeded in constructing an MLR model with adequate generalization capability, since $R_{pred}{}^2$ is greater than 0.9. The constructed model is also significant in terms of *F*-statistics.

The obtained regression equation is given by Equation 12.

$$\begin{aligned} bp\ [^\circ\mathrm{C}] = {} & -217.953 + 37.622\ nSK - 0.943\ nDB \\ & +30.870\ MWC02 + 249.128\ MWC03 \\ & -185.337\ MCW04 - 34.971\ X1 \end{aligned} \quad (12)$$

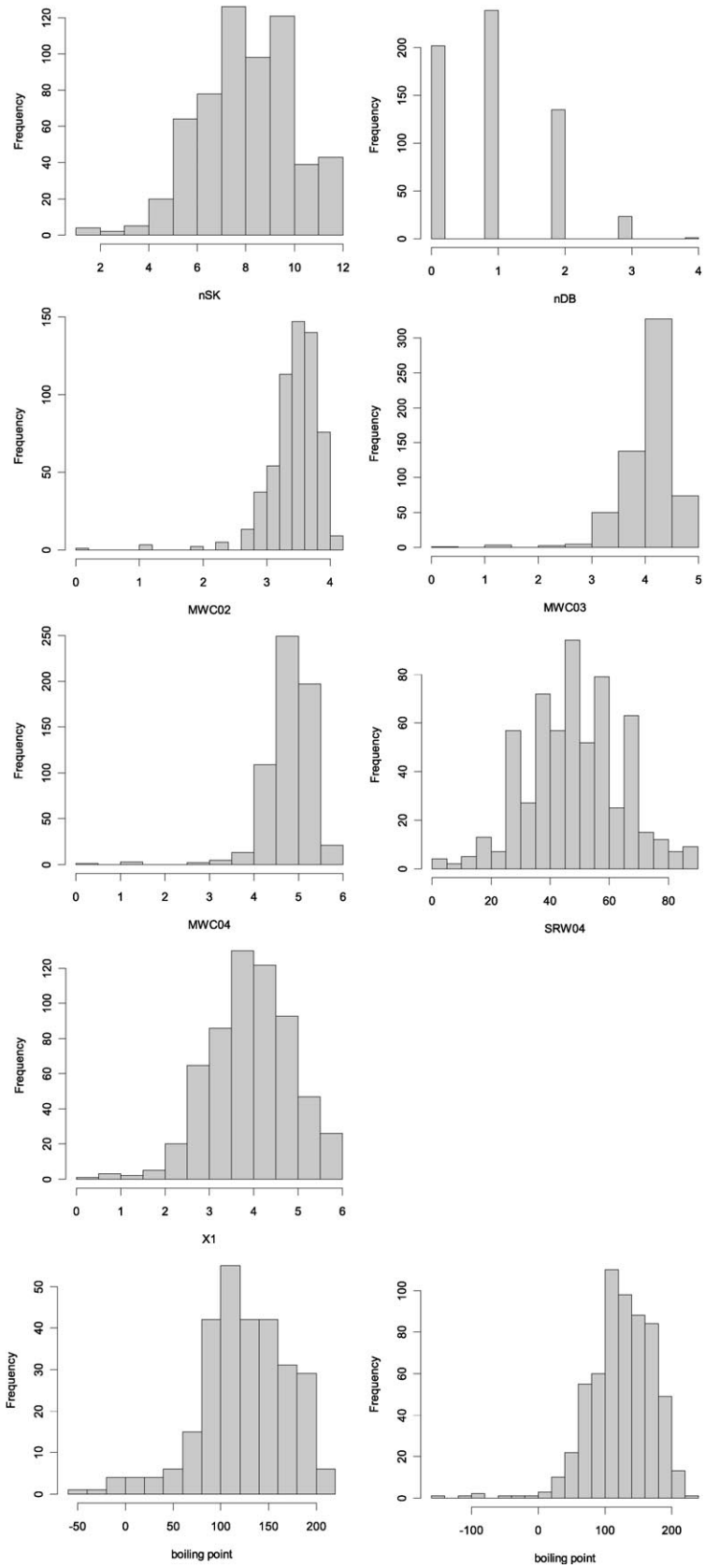


Figure 6. Histograms of descriptor values of training dataset (upper 4 rows). The descriptors are *MCDs*, such as *nSK*, *nDB*, *MWC02*, *MWC03*, *MWC04*, *SRW04*, and *X1*. Only *nDB* is a discrete variable, as shown on the top right. The bottom row shows histograms for the boiling point. The left chart describes the training dataset; the right shows the test dataset used to validate the constructed model.

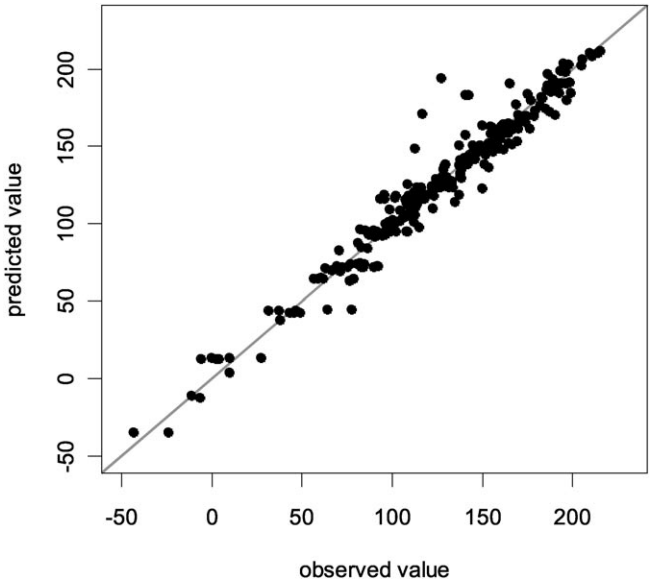


Figure 7. Observed values vs. predicted values for test set.

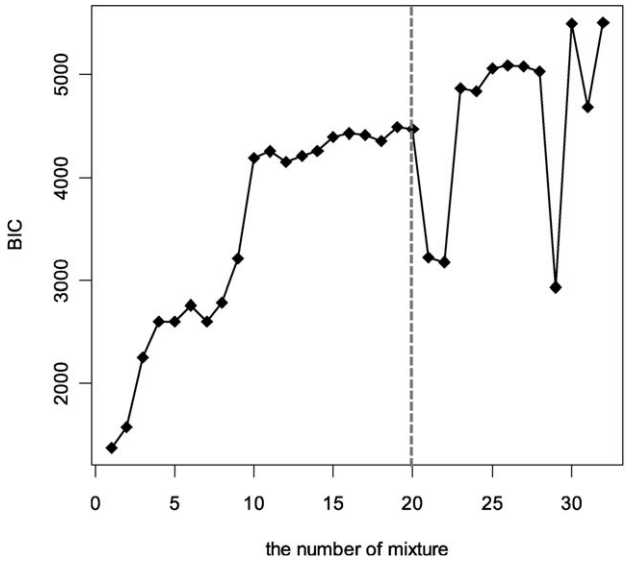


Figure 8. BIC values with different number of mixtures.

The observed values vs. predicted values for the test set are plotted in Figure 7.

As can be seen in Figure 7, the *bp* values of almost all structures in the test set are positioned along the diagonal line. We have therefore confirmed the model's generalization capability. The structure which strayed farthest from the diagonal line was 4,5-diethyl-3,5-octadiene, so we sought the reason for that structure being an outlier. Sabade et al.[42] reported that the boiling point of 4,5-diethyl-3,5-octadiene was 63 °C at 750 mmHg. Other sources, however, reported its boiling point to be around 180 °C. Campbell[43] and Dzhemilev[44] reported its boiling point as being around 90 °C at 7 mmHg or 15 mmHg. Generally speaking, the higher the pressure, the higher the boiling point. This raises the strong possibility that the value of *bp* reported by Sabade was for some reason invalid. It is noted that the existence of an outlier in the validation set does not reflect a deficiency in the constructed MLR/OLS model.

3.3 The Mixture of Gaussian Distributions

For Inverse-QSPR analysis, the density estimation model for training dataset was constructed using Gaussian mixture distributions. In this study, the same covariance matrix with every Gaussian distribution is assumed. The estimating parameters were initialized using the result of hierarchical clustering.[45] In Figure 8, the BIC values for different number of components of a mixture are shown. As the BIC value at 21 components was drastically decreasing, we used the Gaussian mixture distributions models with 20 components.

In order to visualize the mixture of Gaussian distributions in high-dimensional space, we used principal component analysis (PCA) projection. The first principal component describes 76% of total variance (horizontal axis) and the second principal component describes 15% (vertical axis). Figure 9 shows the contour map of probability density on the lower-dimensionality subspace of PCA. Visualized dataset obtained by projecting the training dataset onto the first two principal components is also provided (on the bottom). Figure 9 indicates that the mixture of 20 Gaussian distributions successfully characterizes the density of the training set.

3.4 Inverse-QSPR/QSAR Analysis

The primary purpose of this section is to validate the conditional probability distribution model, $p(\mathbf{x}|y)$, which was derived from both $p(y|\mathbf{x})$ and $p(\mathbf{x})$. The conditional probability distribution of $p(y|\mathbf{x})$ is the result of the QSPR model with MLR/OLS. The marginal probability distribution of $p(\mathbf{x})$ is the result of a density estimation with a mixture of Gaussian distributions. In this study, the desired *bp* values were set at 100, 150, 200, and 300 °C. The conditional distributions of **x** given each of *y* were obtained. Obviously, 300 °C is out of AD, because the maximum value of the boiling point in the training dataset is 224.4 °C. The purpose of an analysis for 300 °C is to reveal how the posterior probability $p(\mathbf{x}|y)$ could be extrapolated.

Figure 10 shows the results of the inverse-QSPR analysis. As can be seen, the high-density region changes in correspondence with the values of the objective variable. According to the loading vectors of the PCA, the first principal component is mainly related to molecular size and the second principal component is related to the number of double bonds (*nDB*). A high second component score means a relatively small value for *nDB*.

Generally speaking, the *bp* of chemical structures usually tends to be higher if the structures are bigger. This is in

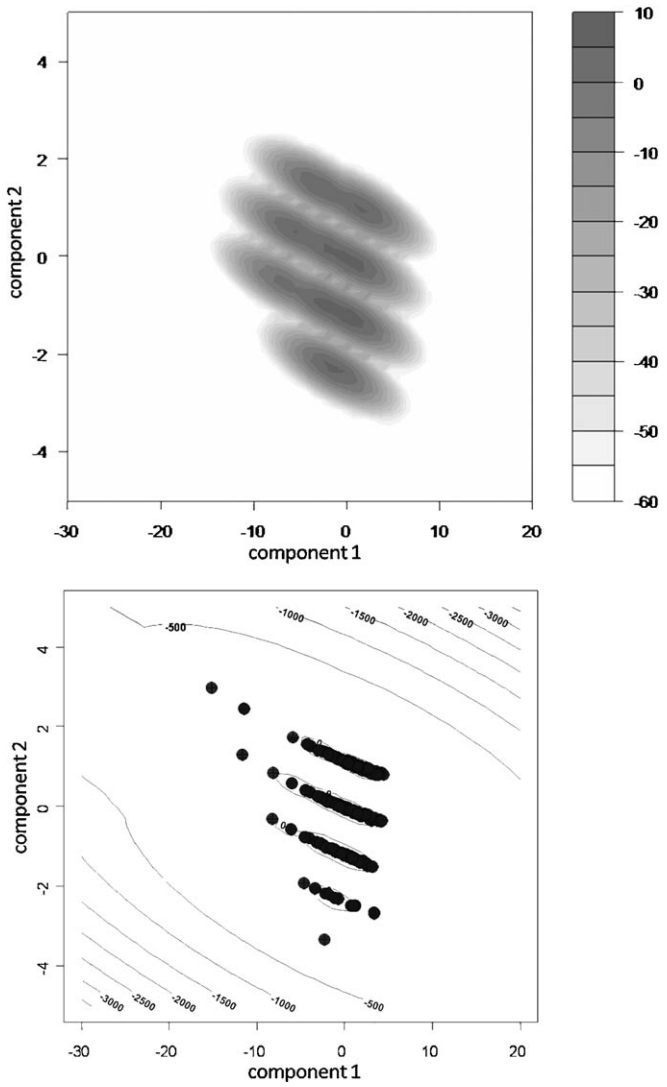


Figure 9. Probability density of the mixture of Gaussian distributions projected onto the first two principal components. Upper: Filled contour map with logarithmic probability density. Lower: Contour of the probability density. Also shown are the training data projected onto the first two principal components.

substantial agreement with the trend in the images in Figure 10. The high-density region on the PCA map moves toward right as the values of the objective variable become higher. This result indicates the validity of our proposed Inverse-QSPR analysis method.

3.5 Decision of Region

The results of our inverse-QSPR/QSAR analysis helped to determine the regions where structures were generated exhaustively. A high-density region was obtained by extracting the area whose probability density was more than a predetermined threshold value. The dense region on the axis of the Randic Index is clearly moving according to the value of objective variable, as illustrated in Figure 11. The shape of the target area is hyper-rectangular, as was mentioned above. Each range of an explanatory variable was determined by its marginal probability distribution. The region in which exhaustive generation is executed was determined independently of the other variables. This assumption might not be applicable to a complicated multivariate probability distribution.

The range for each explanatory variable is listed in Table 1. The result of the case of 300 °C was eliminated because there were no data having the value of *bp* around that temperature.

3.6 Structure Generation

After deciding the target ranges of explanatory variables for each objective variable, the target area was created by combining these ranges into a hyper-rectangular region. The structures were exhaustively generated inside of this region. The results of structure generation were compared with the number of exhaustive structures inside the range of the training set of 153 122 acyclic hydrocarbon structures.

By using these structures, the effectiveness of our proposed methodology was demonstrated. Its degree of effectiveness can be measured by the satisfactory answers it provides to the following two questions.

Table 1. Ranges of high-density region of each descriptor.

bp[a]		*nSK*	*nDB*	*W2*[e]	*W3*[f]	*W4*[g]	*RW*[h]	*X1*
100	LB[b]	6.67	−0.05	3.15	3.76	4.43	33.58	3.09
	UB[c]	8.27	2.09	3.50	4.17	4.93	48.12	3.80
150	LB	8.74	−0.11	3.41	4.03	4.68	45.61	4.10
	UB	10.37	2.06	3.80	4.53	5.10	56.93	4.82
200	LB	10.69	−0.10	3.64	4.27	4.92	56.22	5.07
	UB	12.55	2.03	4.16	4.92	5.80	64.77	5.78
PD[d]	LB	5.48	−0.15	2.92	3.47	4.05	23.39	2.53
	UB	11.60	3.12	3.98	4.74	5.61	74.00	5.43

[a] Objective boiling point value using inverse-QSPR analysis; [b] Lower bound of the determined range; [c] Upper bound of the determined range; [d] Value using prior distribution; [e] *MWC02*; [f] *MWC03;* [g] *MWC04*; [h] *SRW04*; For the prior distribution, the threshold value was set at 0.1. For the others, it was set at 0.5.

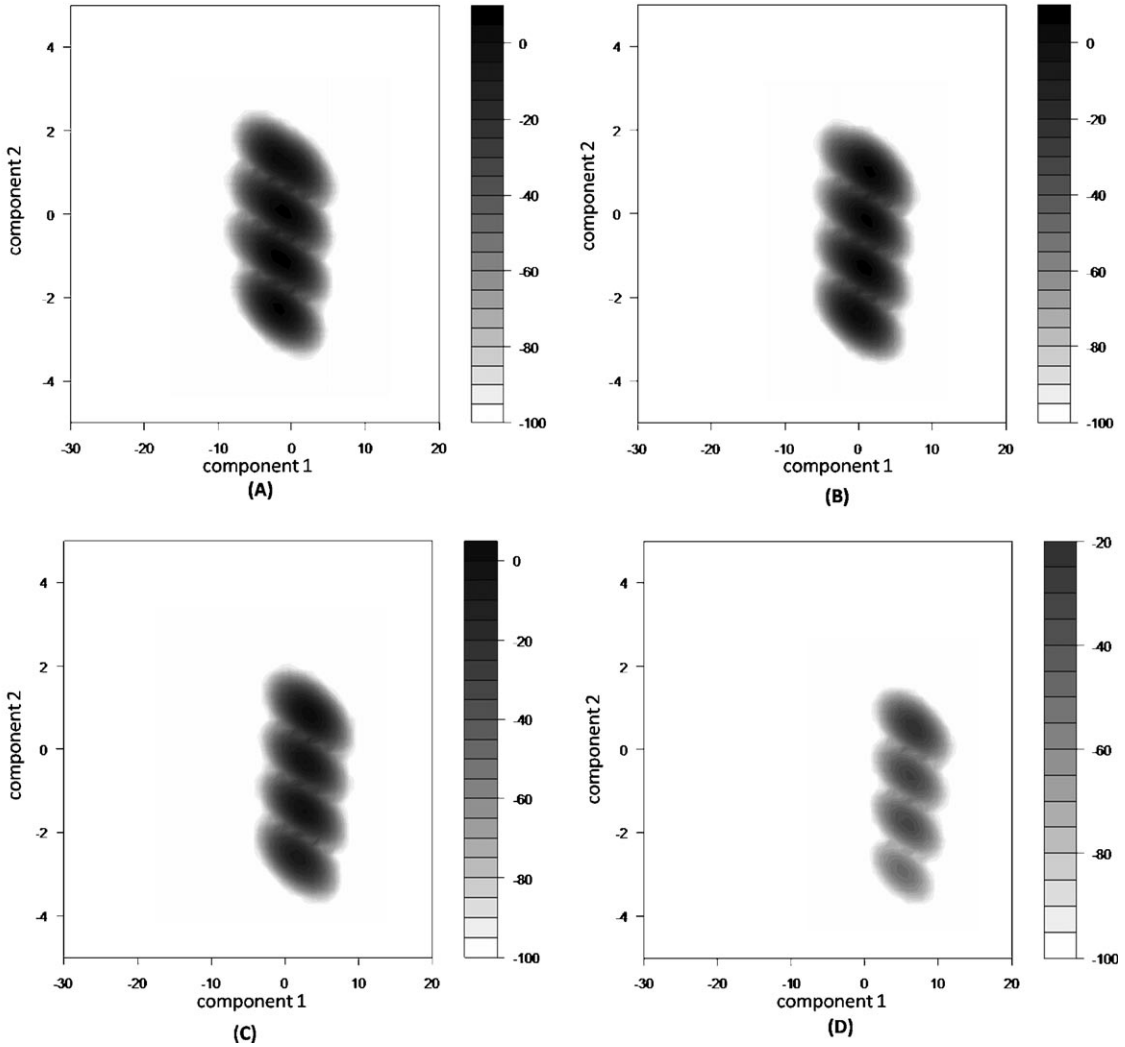


Figure 10. Filled contour maps of the probability density function on a logarithmic scale on the first two PCA axes. (A) $p(x|y=100)$, (B) $p(x|y=150)$, (C) $p(x|y=200)$, and (D) $p(x|y=300)$. The horizontal and vertical axes represent the first and second PCA components, respectively. High density area moves toward right as the value of objective variable increases. There are small regions in (A), (B), and (C), where the probability density is the highest, whereas there are no such regions in (D). These small regions are located at the places where the descriptor values of the training set exist. $y=300$ is obviously outside of AD.

First, how many structures can we generate using this system, as compared to the number of all structures whose *bp* values are the desired values?

Second, using our methodology, how many structures can we avoid generating whose values of *bp* are not the desired values?

The first question is directly related to the detection rate, and the second is related to accuracy. In order to examine these values at one time, receiver operating characteristic (ROC) curves[46] were drawn with the three values of objective variable. In this study, the error range of the value of *bp* was set at 20 °C. Structures having *bp* values within that margin of error were regarded as correct structures.

In Figure 12, the horizontal axis is the false positive rate and the vertical axis is the true positive rate. The points used for drawing the ROC curves were obtained by changing the threshold of the probability density.

Both ROC curves for the values of the objective variable of 100 and 150 °C show a high true positive rate and a low false positive rate. This result indicates that the probability density distributions of the dataset are well separated: one density distribution is for true structures having the target *bp* values and the other is for false structures not having the target *bp* values. By contrast, the ROC curve for 200 °C is below the diagonal line at a low false positive rate. This indicates that the variance of the probability density distribution for the true structures is smaller than that of the false structures. Also, these density distributions overlapped one another to some degree.

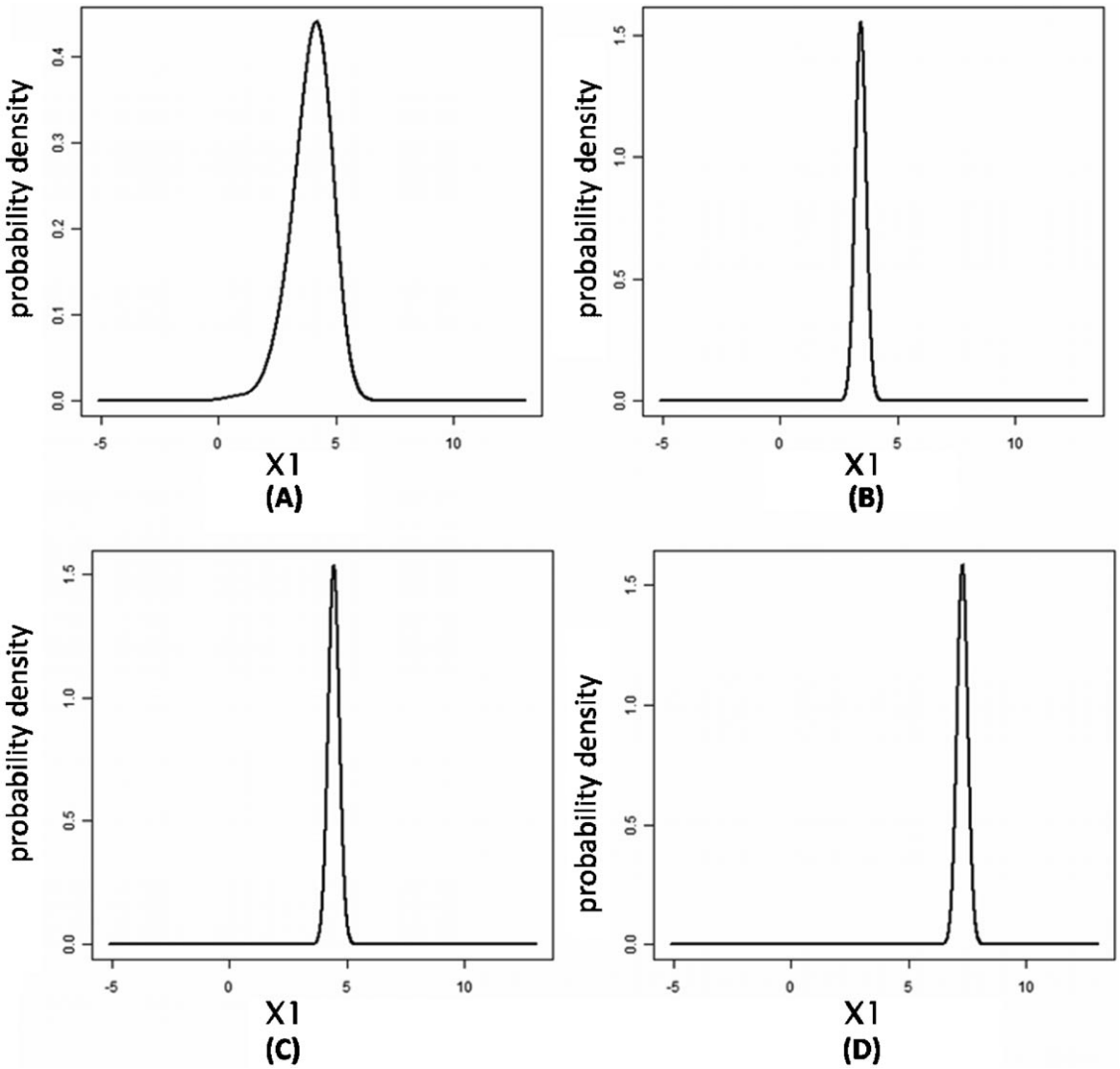


Figure 11. Marginal probability density distributions of Randic connectivity index for various *bp* values: (A) $p(x_1)$, (B) $p(x_1|y=100)$, (C) $p(x_1|y=150)$, and (D) $p(x_1|y=200)$.

Because the purpose of our system is not to generate all the structures whose physicochemical properties are desirable, but to generate all the structures whose properties are desirable inside of the AD (because the validity of the QSPR model was unknown outside of AD), we drew the ROC curves for the dataset by taking the AD into account. In this study, the constructed marginal probability distribution achieved by a mixture of Gaussian distributions was used to define the AD by its probability density. The structures inside of the AD and having desirable values of the objective variable were collected. Those structures whose probability densities values were over 10 were regarded as inside of AD. There were 50 991 structures that fulfilled this criterion.

Figure 13 indicates that our proposed methodology for inverse-QSPR was successful inside of the region obtained by the AD model.

In order to confirm the superiority of structure generation using our proposed methodology, structures inside of the target area were generated using our system. The range of the target area is listed in Table 2. This range was determined with reference to the ROC curve. Because the aim of our study is to exhaustively generate structures having desired property, we determined the narrowest range where the detection rate was 100%. This could not be considered a fair approach to deciding the target region, however, because it is impossible to draw ROC curves for all the structures which we would like to generate. For this reason, drawing ROC curves for the training dataset might be an appropriate approach to deciding the target range.

The results of the structure generation are reported in Table 3. In each case, the chemical structures were pro-

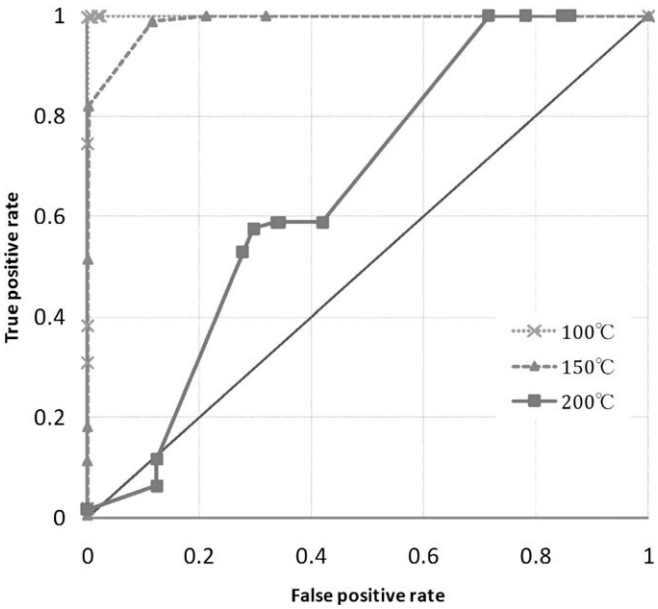


Figure 12. ROC curves for three different *bp* values.

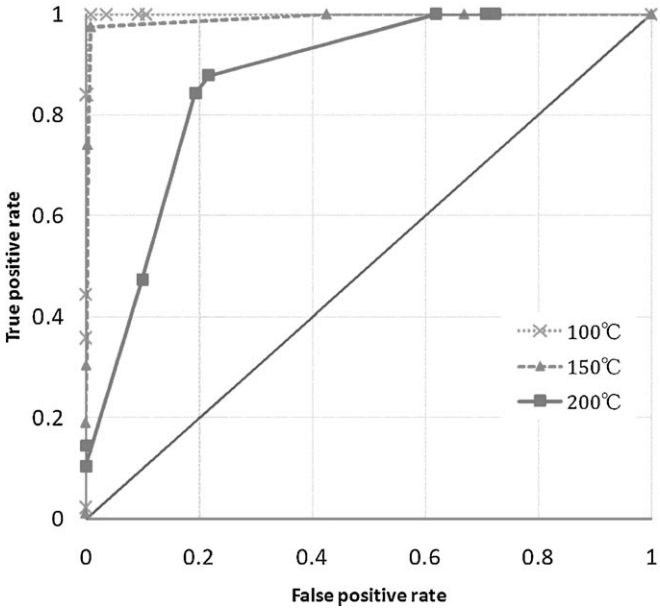


Figure 13. ROC curve for inverse-QSPR method based on three different *bp* values. This figure differs from that in Figure 12 due to its use of data which were obtained by considering AD.

duced on a Windows Vista personal computer with a 3.00 GHz Intel Xeon CPU and 2.00 GB RAM.

The term "ratio" denotes the ratio of the number of generated structures having the correct value of *bp*, divided by the number of total structures whose only constraint is that they have fewer than 12 main atoms. This primitive constraint is appropriate because, when the methodology of virtual screening is used to obtain all the chemical structures with desired physicochemical properties, it is natural to generate structures exhaustively under some primitive constraints. The number of total structures was 181 274 and the generation time was 9.082 s.

The generation result of an objective variable with 100 °C was remarkable. The proposed methodology avoided generating some 99.5% of the number of structures typically generated by traditional virtual screening methodology. The speed of generation was also around 300 times faster than that of the traditional methodology, even as the same generation algorithm was employed. The fact that 83% of the generated structures were desirable is another testament to the superiority of this methodology.

That said, in the case of the value of the objective variable of 150 and 200 °C, the generation time was longer than

Table 2. Ranges for structure generation.

bp[a]		*nSK*	*nDB*	*W2*[d]	*W3*[e]	*W4*[f]	*RW*[g]	*X1*
100	LB[b]	6.22	–0.12	3.01	3.57	4.15	27.73	2.90
	UB[c]	8.78	3.09	3.63	4.31	5.18	57.76	4.00
150	LB	7.85	–0.20	3.24	3.80	4.37	34.90	3.70
	UB	11.49	3.20	4.06	4.84	5.79	87.25	5.23
200	LB	9.76	–0.20	3.48	4.06	4.63	45.96	4.67
	UB	13.72	3.12	4.36	5.16	6.15	103.04	6.19

[a] Objective boiling point value using inverse-QSPR analysis; [b] Lower bound of the determined range; [c] Upper bound of the determined range; [d] *MWC02;* [e] *MWC03;* [f] *MWC04;* [g] *SRW04.*

Table 3. Results of structure generation

bp[a]	*GS*[b]	*CGS*[c]	*Ratio*[d]	*Time [s]*
100	951	788	0.004	0.029
150	23 857	11 564	0.064	14.035
200	113 006	93 837	0.518	20.997

[a] Objective boiling point value; [b] Number of generated structures; [c] Number of correct structures having the targeted boiling point; [d] Ratio of the number of correct structures compared to the number of all acyclic hydrocarbon structures which have fewer than 12 main atoms.

that required by traditional methodology. One reason for this is that we incorporated the function the necessary computing powers of a matrix, such as molecular walk count, into the generation module. The function to calculate these descriptors has yet to be optimized.

Another reason for this is associated with one rather essential factor of inverse-QSPR/QSAR analysis. The physicochemical property of the boiling point seems to be positively correlated with molecular size. In other words, the higher that we set the value of an objective variable, the more structures we have to generate. This causes an "expansion of the number of structures," so to speak, even if we generate these structures effectively. Thus, in the case of 200 °C, the proposed method may not be as effective as one in which structures are generated exhaustively under some primitive constraints.

4 Conclusions

In this paper, we have proposed a method for molecular design based on inverse-QSPR/QSAR models, and demonstrated the usefulness of the proposed method by the exhaustive generation of acyclic hydrocarbon having a specific boiling point. The proposed system consists of three modules: construction of QSPR/QSAR models using MLR/OLS, inverse analysis of the models using the framework of a linear Gaussian model and generation of all chemical structures by the extended canonicalized construction path method. The superiority of the proposed method has been proved by applying it to the analysis of a boiling point dataset.

In order to demonstrate effectiveness of our proposed method, we have exhaustively enumerated the chemical structures having a desired boiling point. Although we did not confirm the experimental boiling points of generated structures, we need not to examine the values this time. This is because whether the boiling points of the generated structures are the desired ones or not depends on the predictive ability of constructed QSPR/QSAR models by MLR/OLS and correctness of the density estimation of a mixture of Gaussian distributions. In this study, we confirmed the correctness of the density estimation with the dataset that have boiling points predicted by MLR/OLS. As long as we use predictive QSPR/QSAR models, the chemical structures are promising. In this study, it is sufficient to show that generation speed is durable and that the number of generated structures is dramatically decreasing compared with those with some primitive constraints.

Although the proposed method has worked well and can be expected to be applied to other molecular design problems, some parts of our system can be improved. The most promising area of improvement regards the approach to deciding the target region from the probability density distribution. Although we used marginal density distributions to decide the shape of the target area, we had to make use of information regarding joint probability distributions. Additional study of this problem should be conducted. In addition, the number of MCDs should be increased in order to construct more predictive models. Since many of MCDs can be defined by extending known chemical descriptors, their properties must be studied. Moreover, the region of AD should be defined more precisely, to improve the predictive power of QSPR/QSAR models and to restrict the number of structures that are generated.

In this paper, only the case study of a boiling point dataset was described. The proposed method, however, can also be applied to any other properties and activities. It is our hope that a number of the challenges of molecular design will be successfully met by using the proposed method.

5 References

[1] *Chemoinformatics: A Textbook* (Ed: J. Gasteiger, T. Engle), Wiley, Chichester **2004**.
[2] A. L. Cheng, K. M. Merz, *J. Med. Chem.* **2003**, *46*, 3572 – 3580.
[3] A. Vernek, D. Fourches, V. P. Solov'ev, V. E. Baulin, A. N. Turanov, V. K. Karandashev, D. Fara, A. R. Katritzky, *J. Chem. Inf. Model.* **2004**, *44*, 1365 – 1382.
[4] W. Walters, M. Stahi, M. Murcko, *Drug Discovery Today* **1998**, *3*, 160 – 178.
[5] D. L. Wheeler, T. Barrett, D. A. Benson, et al., *Nucleic Acids Res.* **2007**, *35*, D5 – D12.
[6] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177 – 182.
[7] M. E. Munk, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 997 – 1009.
[8] K. Funatsu, N. Miyabayaski, S. Sasaki, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18 – 28.
[9] I. P. Bangov, *J. Chem. Inf Comput. Sci.* **1990**, *30*, 277 – 289.
[10] T. Wieland, A. Kerber, R. Laue, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 413 – 419.
[11] C. Rücker, M. Meringer, A. Kerber, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2070 – 2076.
[12] C. Rücker, M. Meringer, A. Kerber, *J. Chem. Inf. Model.* **2005**, *45*, 74 – 80.
[13] T. Fink, J. L. Reymond, *J. Chem. Inf. Model.* **2007**, *47*, 342 – 353.
[14] J. Hert, J. J. Irwin, C. Laggner, M. J. Keiser, B. K Shoichet, *Nat. Chem. Biol.* **2009**, *5*, 479 – 483.
[15] R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3 – 50.
[16] R. Bruggemann, S. P. Udenz, L. Carlsen, P. B. Sorensen, M. Thomsen, R. K. Mishra, *SAR QSAR Environ. Res.* **2001**, *11*, 473 – 487.
[17] R. B. Nachbar, *Genetic Programming and Evolvable Machines* **2000**, *1*, 57 – 94.
[18] R. Van Deursen, J. L. Reymond, *ChemMedChem*, **2007**, *2*, 636 – 640.
[19] N. Brown, B. McKay, J. Gasteiger, *J. Comput. Aided Mol. Des.* **2004**, *18*, 761 – 771.
[20] P. Gramatica, *QSAR Comb. Sci.* **2007**, *26*, 694 – 701.
[21] J. L. Faulon, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707 – 720.
[22] F. A. de Lima Ribeiro, M. M. C. Ferreira, *J. Mol. Struct. (theochem)* **2003**, *663*, 109 – 126.
[23] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Heidelberg **2006**, pp. 370 – 372.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Heidelberg **2006**, pp. 423 – 455.
[25] B. D. McKay, *J. Algorithms* **1998**, *26*, 306 – 324.
[26] R. A. Redner, H. F. Walker, *Siam Rev.* **1984**, *26*, 195 – 237.
[27] A. E. Raftery, *Sociological Methodol.* **1995**, *25*, 111 – 163.
[28] C. M. Bishop, in *Pattern Recognition and Machine Learning*, Springer, Heidelberg **2006**, pp. 474 – 486.
[29] M. E. Tipping, C. M. Bishop, *J. Roy. Stat. Soc.* **1999**, *61*, 611 – 622.
[30] H. Attias, *Neural Comput.* **1999**, *11*, 803 – 851.
[31] S. Roweis, Z. Ghahramani, *Neural Comput.* **1999**, *11*, 305 – 345.
[32] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Heidelberg **2006**, pp. 90 – 94.
[33] J. L. Faulon, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 338 – 348.
[34] J. M. Amigó, J. Gálvez, V. M. Villar, *Naturwissenschaften* **2009**, *96*, 749 – 761.
[35] M. Randic, *J. Math. Chem.* **1991**, *7*, 155 – 168.
[36] M. Randic, *J. Am. Chem. Soc.* **1975**, *97*, 6609 – 6615.
[37] P. Broto, G. Moreau, C. Vandycke, *Eur. J. Med. Chem.* **1984**, *19*, 66 – 70.
[38] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Willey-VCH, Weinheim **2000**, pp. 366 – 367.
[39] G. Rücker, C. Rücker, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683 – 695
[40] B. D. Mckay, *nauty User's Guide*, Version 2.2; http://cs.anu.edu.au/~bdm/nauty.
[41] *Beilstein Database*, http://www.beilstein-inst.de/.
[42] M. B. Sabade, M. F. Farona, *J. Organomet. Chem.* **1986**, *310*, 311 – 316.
[43] J. B. Campbell, H. C. Brown, *J. Org. Chem.* **1980**, *45*, 549 – 550.
[44] U. M. Dzhemilev, A. G. Ibragimov, V. A. D'yakonov, R. A. Zinnurova, *Russ. J. Org. Chem.* **2007**, *43*, 176 – 180.
[45] J. H. Ward, *J. Am. Stat. Assoc.* **1963**, *58*, 236 – 244.
[46] T. Fawcett, *Pattern Recognition Lett.* **2006**, *27*, 861 – 874.

Received: September 19, 2009
Accepted: October 25, 2009
Published online: January 25, 2010