THE UNIVERSITY OF TOKYO

Chemical System Engineering

Chemical Structure Generation Based on Inverse Quantitative Structure-Property Relationship/Quantitative Structure-Activity Relationship

Kimito Funatsu

Chemical System Engineering, The University of Tokyo

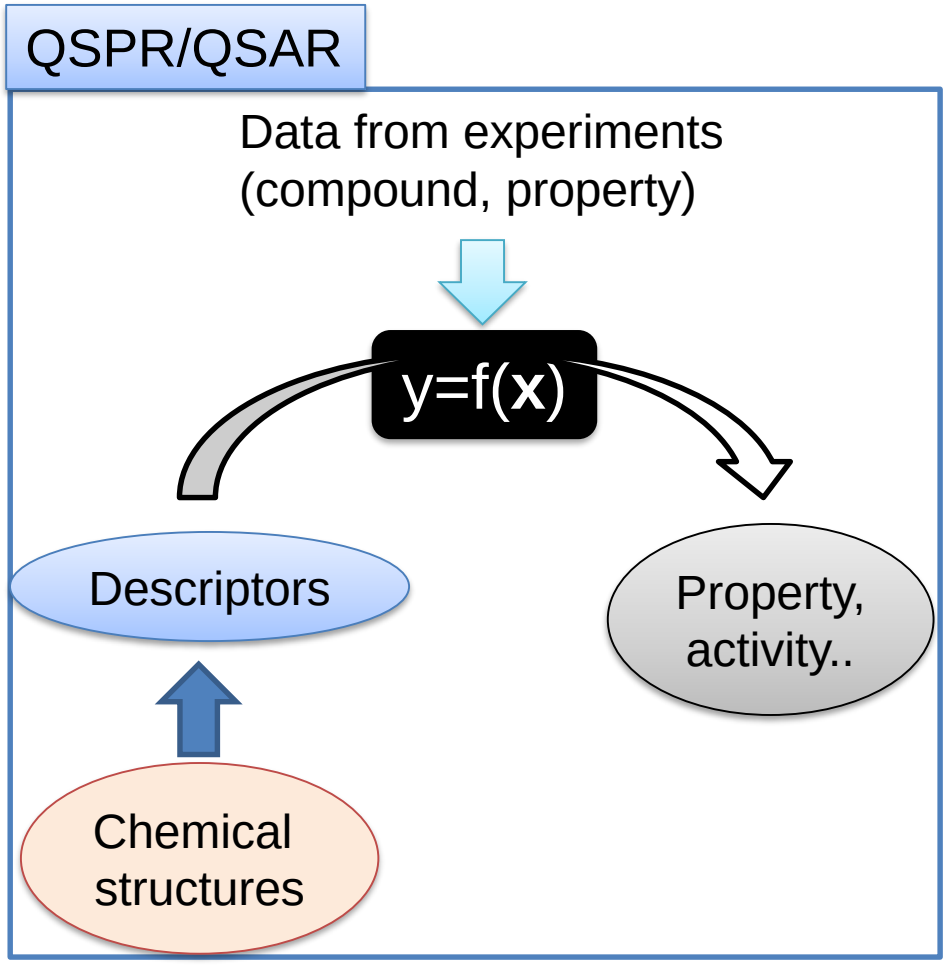
Strasbourg Summer School on Chemoifnormatics
June 27, 2018

Outline

- 1: General introduction
  - Inverse QSPR/QSAR
  - Objective and hypothesis
- 2: Structure generation
- 3: Inverse QSPR/QSAR analysis (from y to **x**)
- 4: Structure generation based on inverse QSPR/QSAR
- 5: Summary

Molecular design with inverse quantitative structure-property/activity relationship (QSPR/QSAR)

Molecular Design with QSPR/QSAR



Quantitative structure-property relationship (QSPR)
Quantitative structure-activity relationship (QSAR)

x

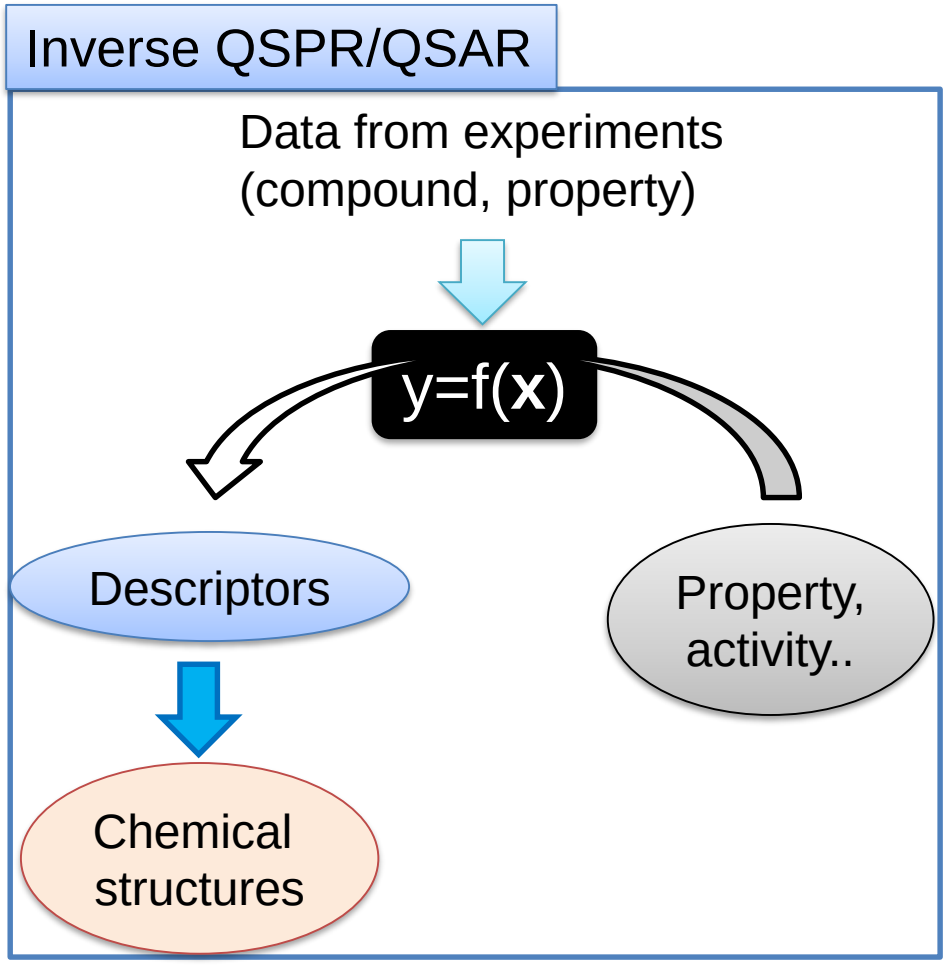
Descriptors	Values
MW [g/mol]	180.04
#HBA	3
#NBD	1
#Aromatic Rings	1
TPSA [Å^2]	63.6

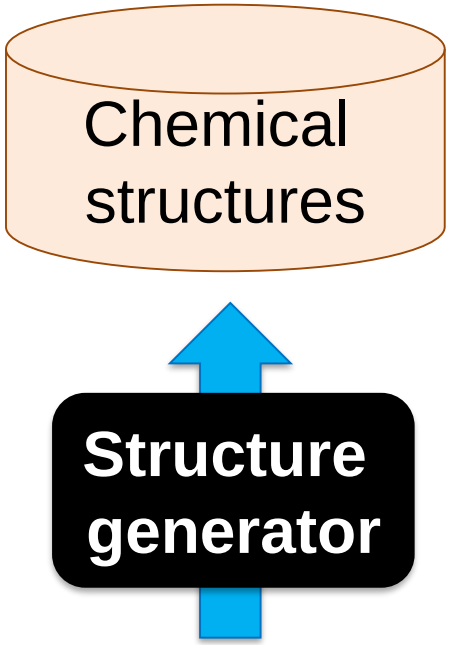
y=f(x)

y

Property	value
MP [degree]	-5
Viscosity [Pa·s]	8.0
LogP	3
LogS	-1

Inverse QSPR/QSAR

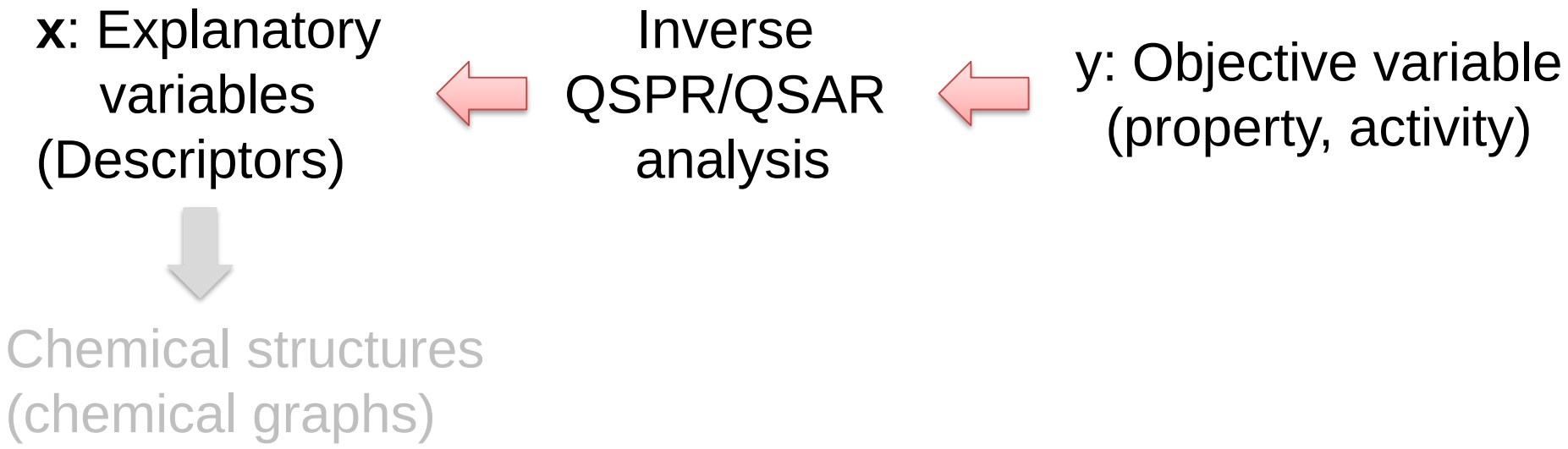




x

Descriptors	Values
MW [g/mol]	180.04
#HBA	3
#NBD	1
#Aromatic Rings	1
TPSA [Å^2]	63.6

Inverse QSPR/QSAR Workflow



Obtaining **x** information from y

Inverse QSPR/QSAR Workflow

x: Explanatory variables (Descriptors)

Inverse QSPR/QSAR analysis

y: Objective variable (property, activity)

Chemical structures (chemical graphs)

Generating structures based on **x** information

Challenges in Inverse QSPR/QSAR

x: Explanatory variables (Descriptors)

Inverse QSPR/QSAR analysis

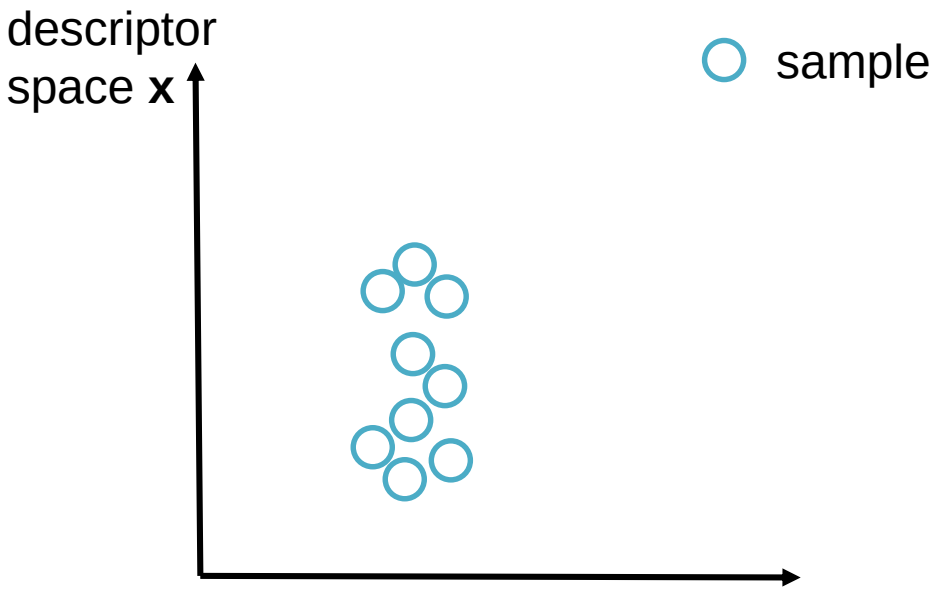
y: Objective variable (property, activity)

Chemical structures (chemical graphs)

Obtaining **x** information from y
Not considering applicability domain (AD)
Poor predictability by multiple linear regression(MLR) model

Applicability Domain (AD)

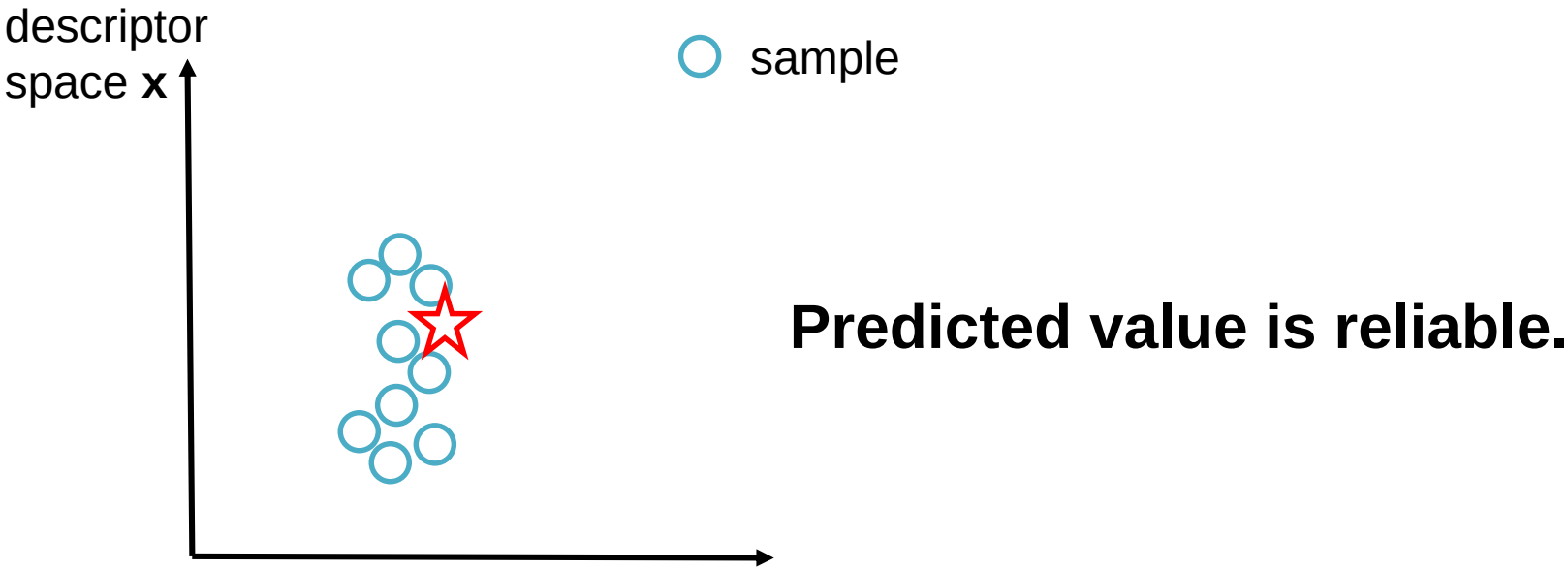
- Only inside AD, predicted values produced by regression models should be reliable.
  - Density-based method
  - Ensemble-based method



Baskin, I. I., Kireeva, N. and Varnek, A. *Mol. Inform*. 29, 581–587, 2010
Kaneko, H. and Funatsu, K. *J. Chem. Inf. Model.* 54, 2469–2482, 2014

Applicability Domain (AD)

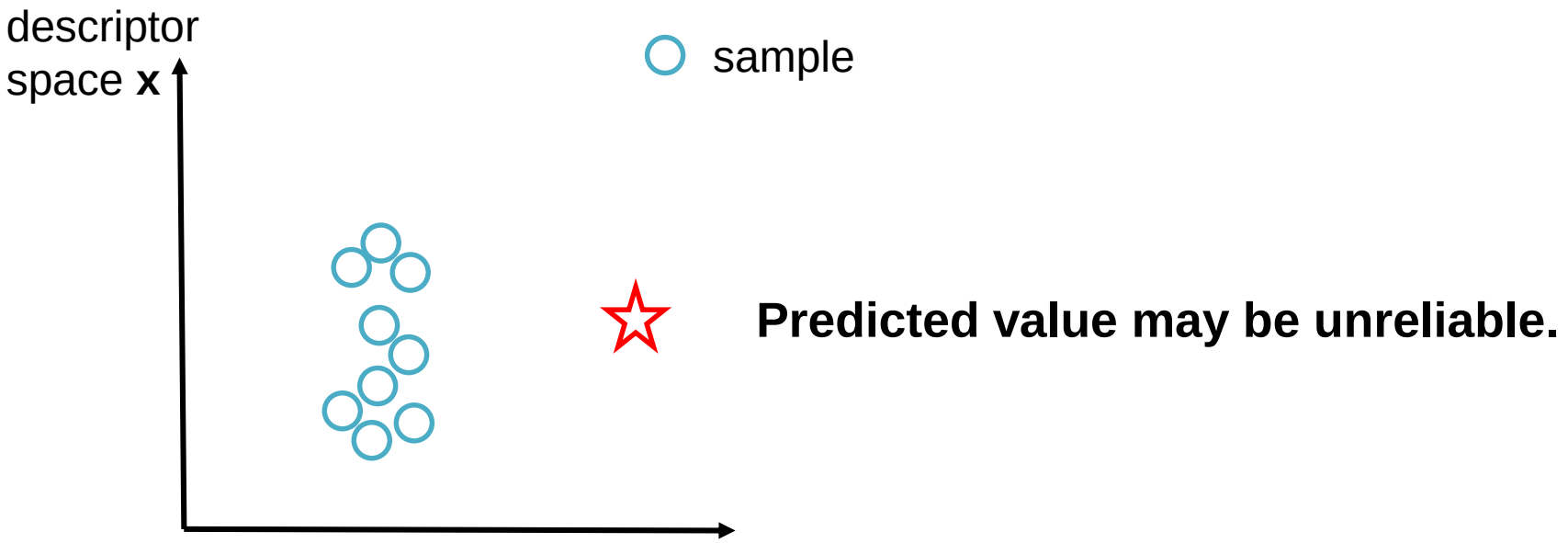
- Only inside AD, predicted values produced by regression models should be reliable.
  - Density-based method
  - Ensemble-based method



Baskin, I. I., Kireeva, N. and Varnek, A. *Mol. Inform*. 29, 581–587, 2010
Kaneko, H. and Funatsu, K. *J. Chem. Inf. Model.* 54, 2469–2482, 2014

Applicability Domain (AD)

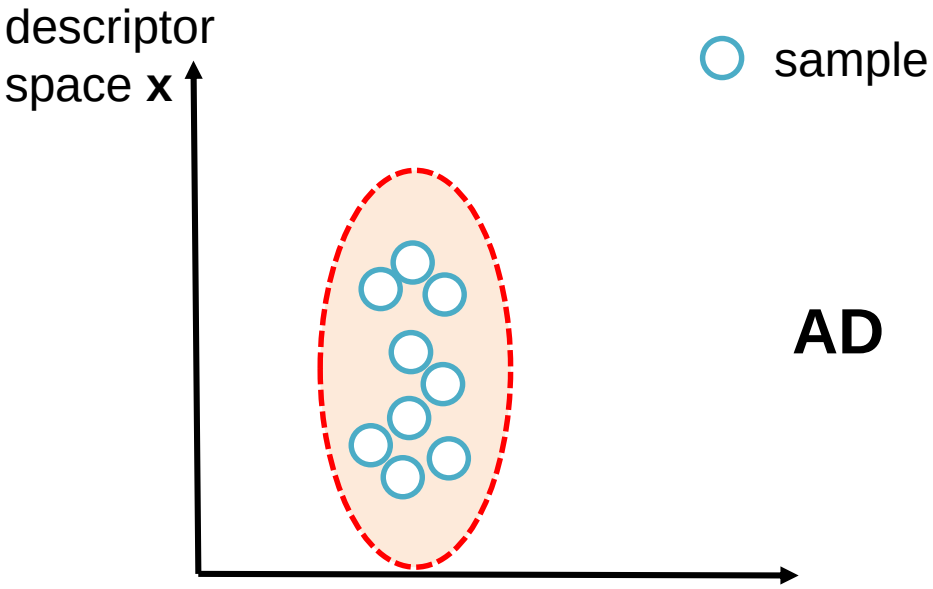
- Only inside AD, predicted values produced by regression models should be reliable.
  - Density-based method
  - Ensemble-based method



Baskin, I. I., Kireeva, N. and Varnek, A. *Mol. Inform*. 29, 581–587, 2010
Kaneko, H. and Funatsu, K. *J. Chem. Inf. Model.* 54, 2469–2482, 2014

Applicability Domain (AD)

- Only inside AD, predicted values produced by regression models should be reliable.
  - Density-based method
  - Ensemble-based method

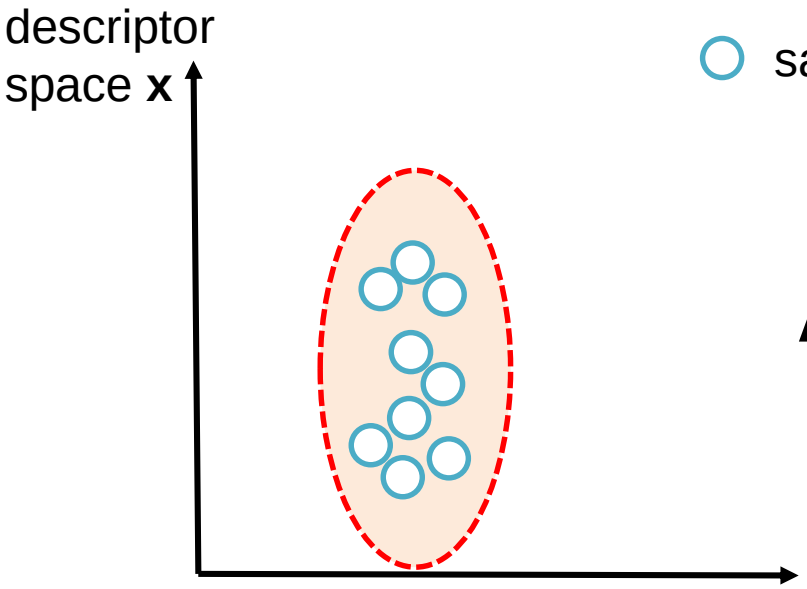


AD In a density-based method, Probability density function p(**x**) (PDF) is a criterion for AD

Baskin, I. I., Kireeva, N. and Varnek, A. *Mol. Inform*. 29, 581–587, 2010
Kaneko, H. and Funatsu, K. *J. Chem. Inf. Model.* 54, 2469–2482, 2014

Applicability Domain (AD)

- In inverse QSPR/QSAR analysis, AD has not been considered.
  - Not considering training data information
    - Extrapolation is allowed without limitation.



AD In a density-based method, Probality density function p(**x**)(PDF) is a criterion for AD

Challenges in Inverse QSPR/QSAR

x: Explanatory variables (Descriptors)

Inverse QSPR/QSAR analysis

y: Objective variable (property, activity)

Chemical structures (chemical graphs)

Chemical structure generation

Treating limited variety (number) of descriptors

Not considering universal AD

Descriptors in Inverse QSPR/QSAR Analysis

- Specific type of descriptors is employed.
  - Kier indices, X indices, Wiener index.
  - Signatures
- Proper descriptor set varies from projects to projects.

M. I. Skvortsova, I. I. Baskin, *et al.*, *J. Chem. Inf. Comput. Sci.*, 33, 4, 630–634, 1993.
C. J. Churchwell, *et al.*, *J. Mol. Graph. Model.*, 22, 263–273, 2004.
Kirkpatrick, P. and Ellis, C. *Nature*, 432, 823–823, 2004.

Universal AD

- Universal AD is an abstract concept, which is irrelevant with models
  - Determined based only on the training data before constructing any QSPR/QSAR models.

Baskin, I. I., Kireeva, N. and Varnek, A. *Mol. Inform*. 29, 581–587, 2010

- Simple example: boiling point model

$$\mathrm{bp}(℃) = -126.19 + 33.42\mathrm{N}_c - 6.286\mathrm{T}_m$$

N_c : Number of carbon atoms
T_m : Number of terminal carbon atoms

n = 39
s = 5.86
r^2 = 0.987

Seybold, P. G., May, M. and Bagal, U. A. *J. Chem. Educ.* 64, 575, 1987

- Simple example: boiling point model

$$\mathrm{bp}(℃) = -126.19 + 33.42\mathrm{N}_c - 6.286\mathrm{T}_m$$

This equation is valid for C2-C8 alkanes.

Structures to be generated
should be restricted to alkanes.

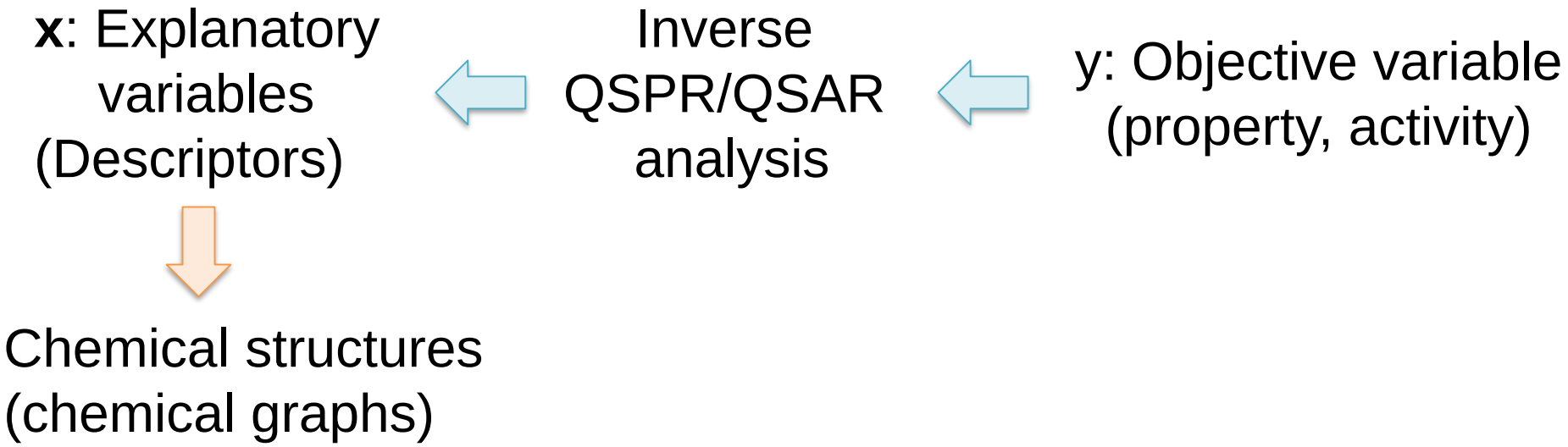
Seybold, P. G., May, M. and Bagal, U. A. *J. Chem. Educ.* 64, 575, 1987

Challenges in inverse QSPR/QSAR analysis

Chemical structure generation	Obtaining **x** information from y
Treating limited variety (number) of descriptors	Not considering AD
Not considering universal AD	Poor predictability by MLR

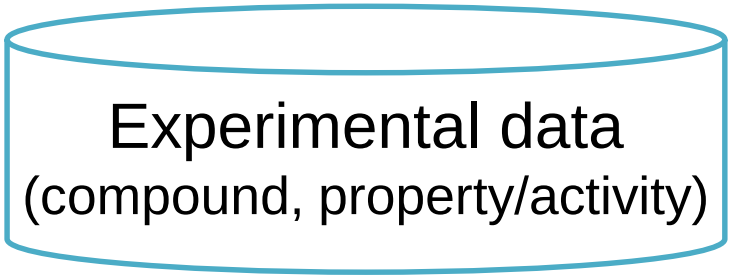
Objective

To develop a practical chemical structure generation system based on inverse QSPR/QSAR by overcoming the challenges.



QSAR and Inverse QSAR Workflow

Workflow of QSPR/QSAR analysis



x: Molecular descriptors

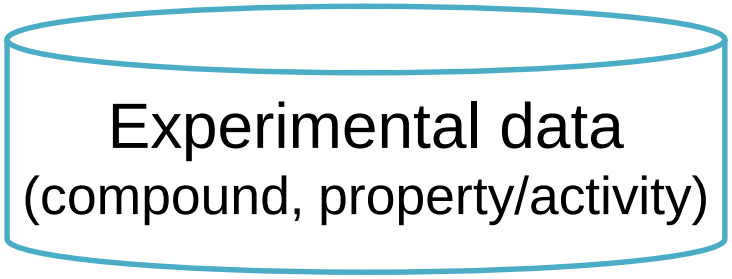
QSPR/QSAR model

y: predicted property/activity

a specific y value

QSPR and Inverse QSPR Workflow

Workflow of QSPR/QSAR analysis



x: Molecular descriptors

QSPR/QSAR model

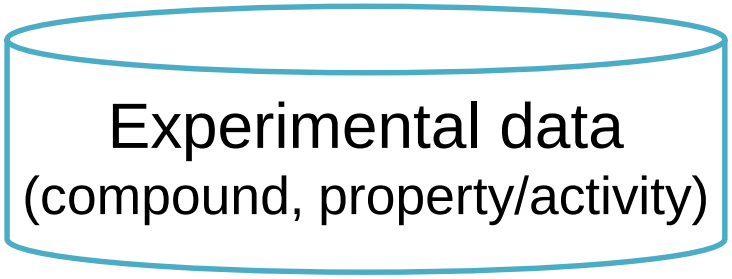
y: predicted property/activity

Inverse QSPR/QSAR analysis

a specific y value

QSPR and Inverse QSPR Workflow

Workflow of QSPR/QSAR analysis



x: Molecular descriptors

QSPR/QSAR model

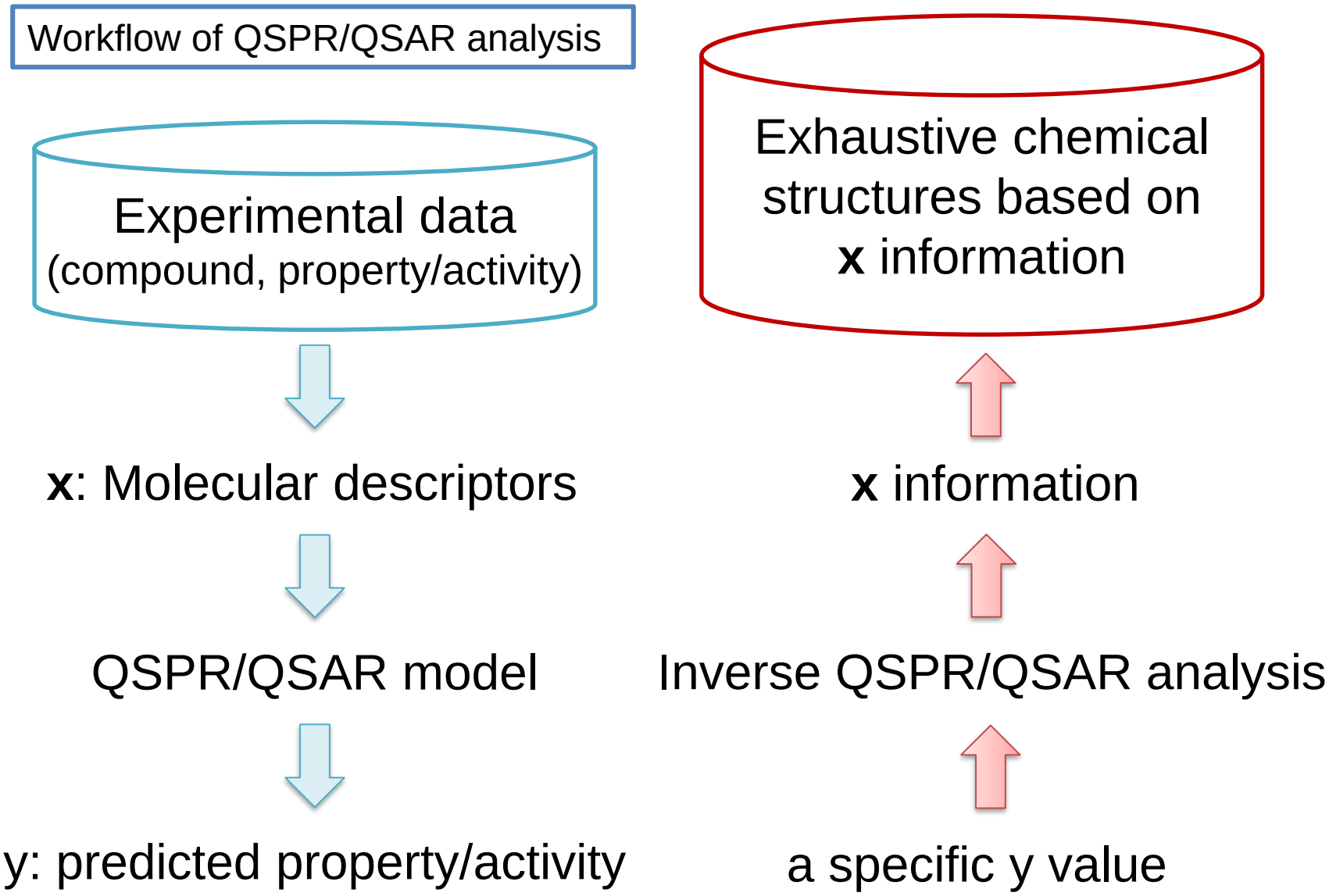
y: predicted property/activity

x information

Inverse QSPR/QSAR analysis

a specific y value

QSPR and Inverse QSPR Workflow



Hypothesis

Chemical structures giving a specific y can be exhaustively generated in inverse QSPR/QSAR by

- ✓ considering local and universal ADs,
- ✓ using efficient structure generator,
- ✓ introducing variety of descriptors.

To develop a practical chemical structure generation system based on inverse QSPR/QSAR

Chemical structure generation	Obtaining **x** information from y
Explaining a methodology for using variety of descriptors	Introducing probability density for treating AD
Describing algorithms for treating chemical graphs	Explaining a non-linear regression methodology

2. Structure Generation

Goal To develop a structure generator that overcomes challenges as follows:

Challenges

Chemical structure generation
Treating limited variety (number) of descriptors
Not considering universal AD

Proposed Methodologies

- Using ring systems and atom fragments as building blocks in structure generation.
- Introducing monotonous changing descriptors (MCDs)

Challenges

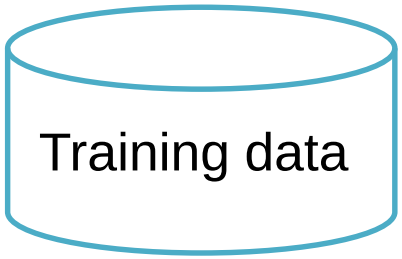
Chemical structure generation

Treating limited variety (number) of descriptors

Not considering universal AD

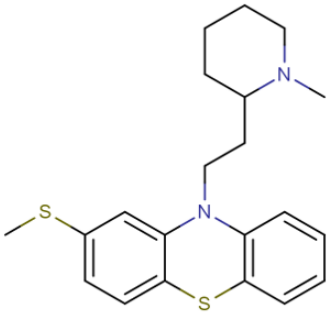
B. D. McKay, *J. Algorithms*, 26, 306-324, 1993
Miyao, T., Arakawa, M. and Funatsu, K. *Mol. Inform.* 29, 111–125, 2010

Generation Strategy for Considering Universal AD

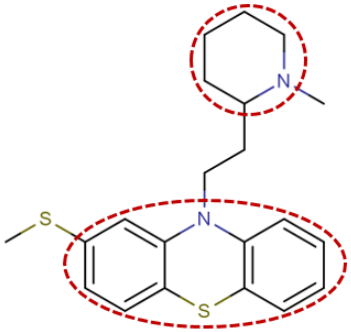


Building blocks (number and kinds)

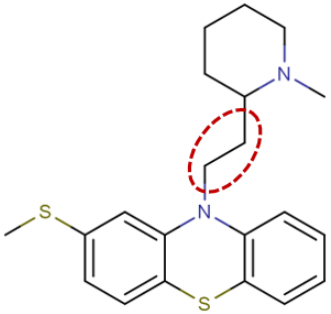
- Ring systems in the training dataset.
- Elements in the training dataset.



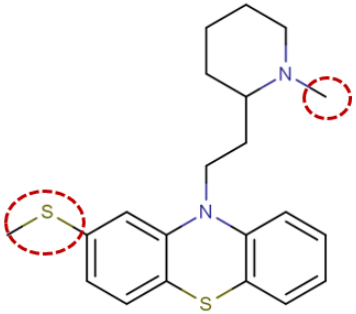
Thioridazine



Ring systems



Linker

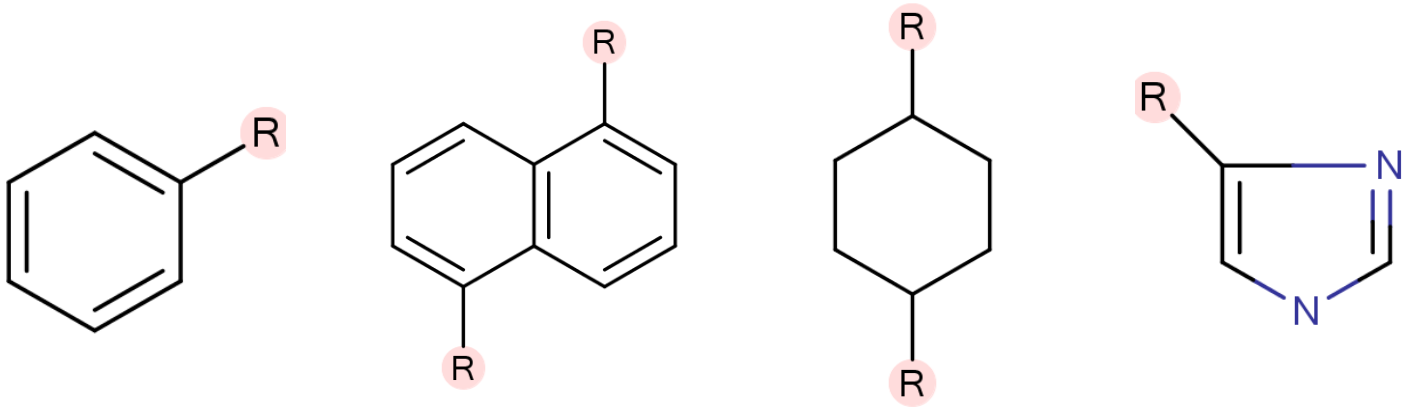


Side chains

Bemis, G. W. and Murcko, M. A. *J. Med. Chem.* 39, 2887–2893, 1996
Taylor, R. D., MacCoss, M. and Lawson, A. D. G. *J. Med. Chem.* 57, 5845–5859, 2014

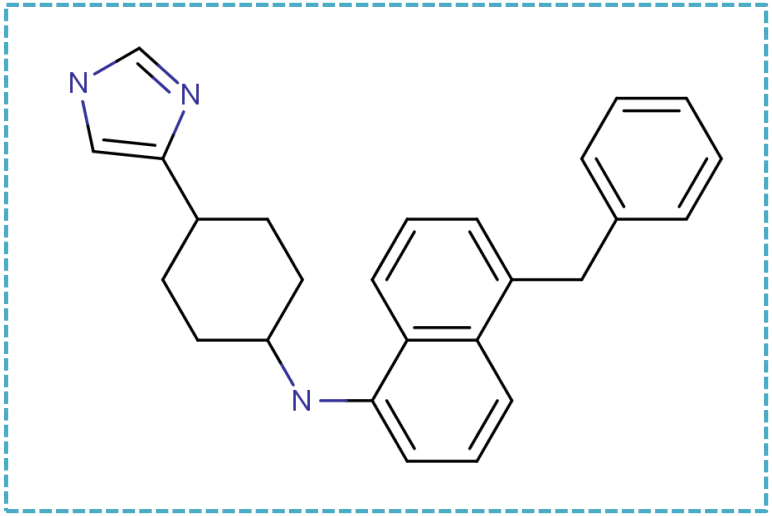
Structure Generation by Combining Building Blocks

Ring systems and atom fragments are combined to form a chemical graph.



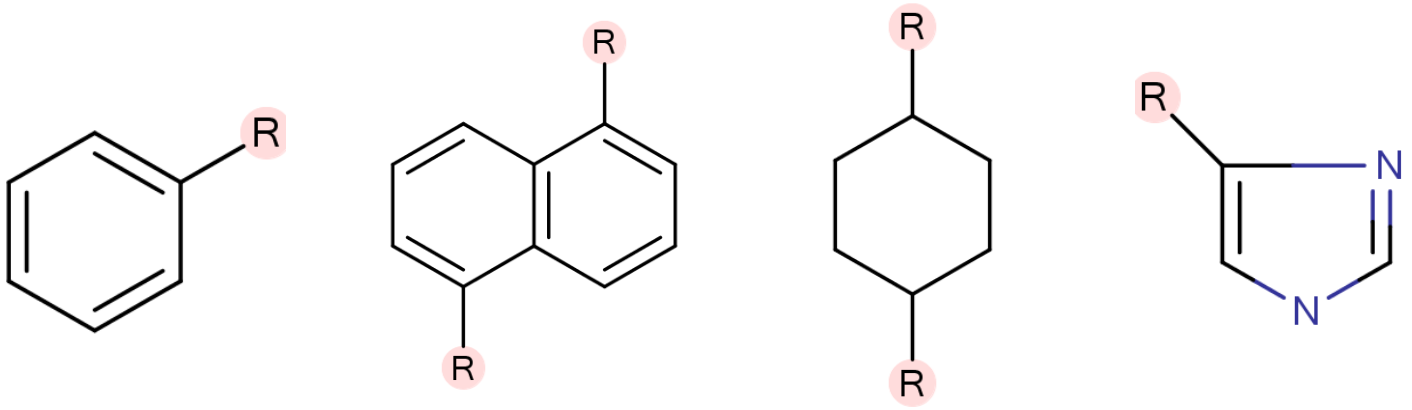
CH_2

NH

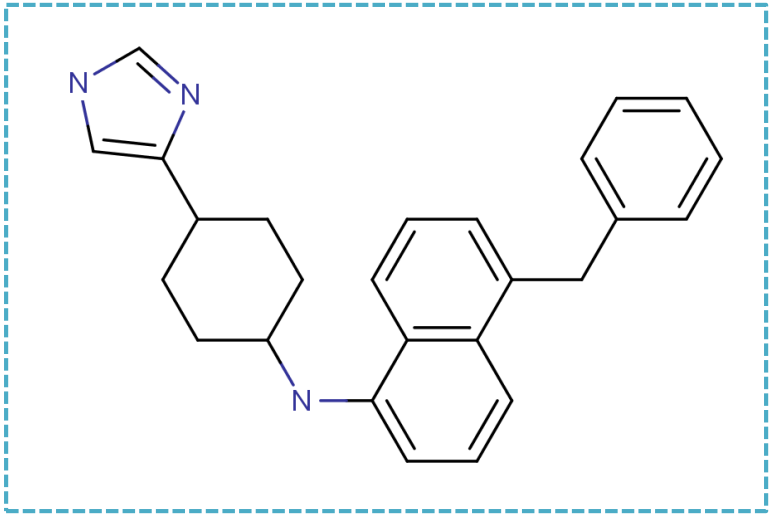


Challenges in Structure Generation

- Generating duplicate structures
  - Combinatorial explosion



CH_2 NH



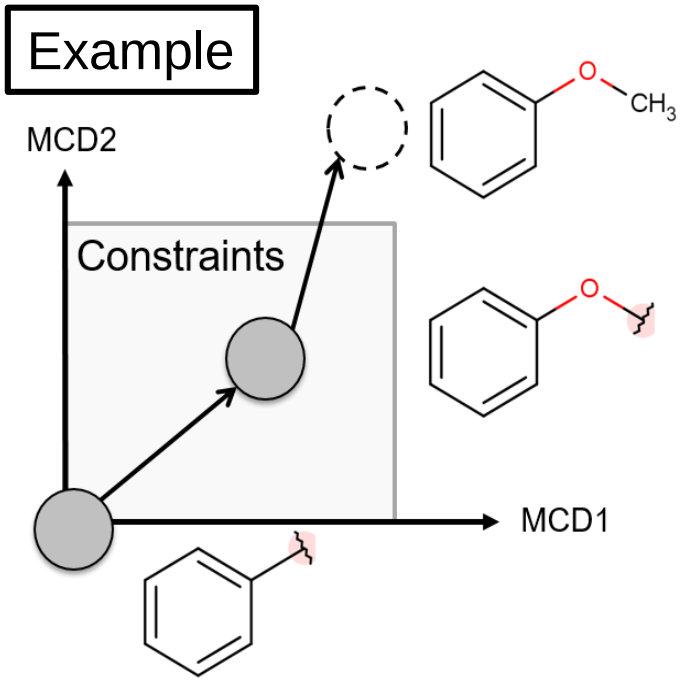
Strategies for Pursuing Computational Efficiency

- Modifying the canonical construction path method to treat building blocks
  - Assure the uniqueness and exhaustiveness of the generated structures
- Using reduced graphs instead of ring systems
  - For speeding up graph operation during structure generation
- Combining building blocks in a tree-like approach.

B. D. McKay, *J. Algorithms*, 26, 306-324, 1998
Miyao, T., Kaneko H., Funatsu, K. *J. Comput. Aided Mol. Des.* 2016, 30, 425-446.

How to Use a Variety Set of Descriptors

- Monotonous changing descriptors (MCDs)
  - MCDs are descriptors whose values change monotonously by adding a building block to a growing structure.
    - molecular weight, topological indices



Miyao, T., Arakawa, M. and Funatsu, K. *Mol. Inform.* 29, 111–125, 2010.

Regression Model with MCDs

- MCDs: 409 extracted from DRAGON 5 (790 descriptors)
- Data set: Ligands for alpha 2A adrenergic receptor (GVK)
  - y: pK_i
  - training data: 500, test data: 143
- Regression method: partial least squares regression

	Opt. Compt. [a]	Q^2_{5fold}	$RMSE_{cv}$	R^2	$RMSE_{pred}$	R^2_{pred}
MCD	10	0.832	0.354	0.891	0.385	0.836
DRAGON	11	0.859	0.324	0.918	0.347	0.867

DRAGON for Windows (Software for Molecular Descriptor Calculation) version 5.4.
GVK data base, *http://www.gvkbio.com*

Partial Conclusions

- Propose efficient structure generation algorithms by combining ring systems and atom fragments

- Using all MCDs in DRAGON has molecular description ability compatible to the comprehensive descriptors
  - High predictability of a PLS regression model for alpha 2A adrenoceptor.

3. Inverse QSPR/QSAR Analysis (from y to **x**)

Goal

To develop a inverse QSPR/QSAR methodology that overcomes challenges as follows:

Challenges
Obtaining **x** information from y
Not considering applicability domain
Poor predictability by MLR

Proposed Methodologies

- Probability density function (PDF)
  - Gaussian mixture models (GMMs)
- Pseudo nonlinear regression methodology
  - GMMs and cluster-wise multiple linear regression (GMMs/cMLR)

Challenges
Obtaining **x** information from y
Not considering applicability domain
Poor predictability by MLR

- GMM: $p(\mathbf{x})$
- GMMs/cMLR: $p(y|\mathbf{x})$

y value ⇨ Posterior density: $p(\mathbf{x}|y)$

In order to consider AD

Miyao, T., Kaneko, H. and Funatsu, K. *J. Chem. Inf. Model.* 56, 286–299, 2016.

Proposed Methodologies

- GMM: p(**x**)
- GMMs/cMLR: p(y|**x**)

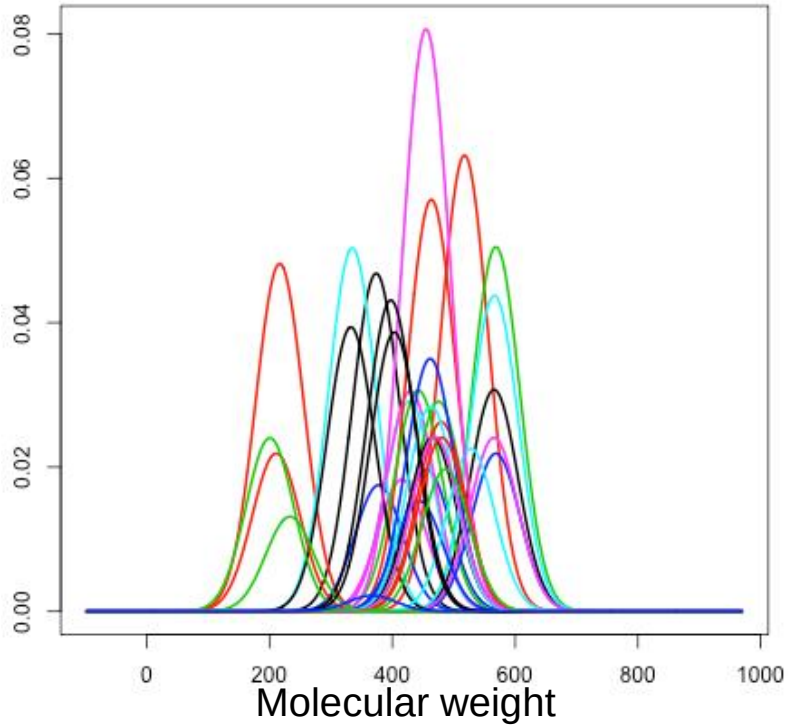
y value ⇨ Posterior density: p(**x**|y)

In order to consider AD

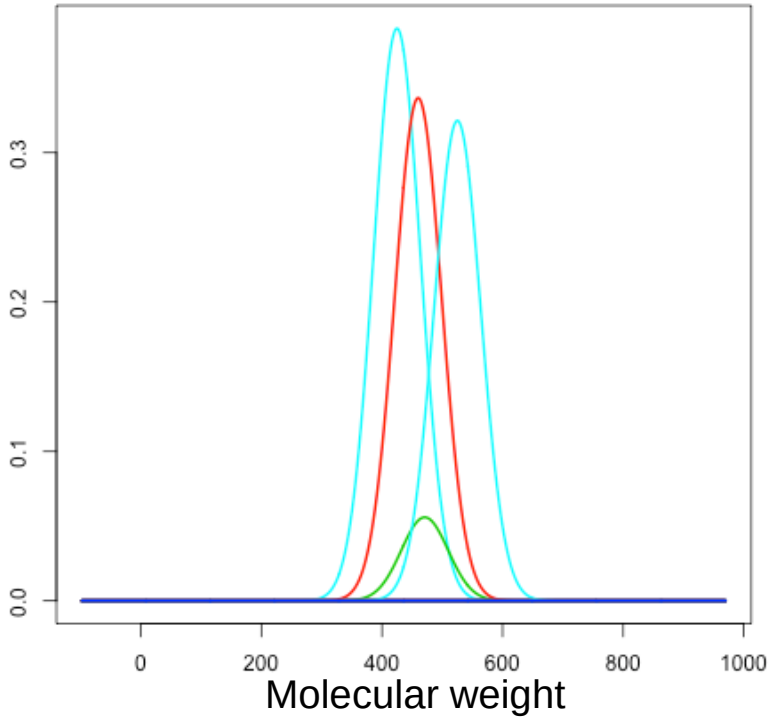
$$p(\mathbf{x}) = \sum_{i=1}^{M} \pi_i N(\mathbf{x} \mid \boldsymbol{\mu}_i, \Sigma_i)$$

$$p(\mathbf{x} \mid y) = \sum_{i=1}^{M} \omega_i N(\mathbf{x} \mid \Delta\{\mathbf{A}^T \sigma^{-2} y + \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i\}, \Delta)$$

Prior distribution (p(**x**))



Posterior distribution (p(**x**|y))



Evaluation of Posterior PDF for a Criterion of AD

- Aqueous solubility dataset
  - training data: 900
  - test data: 254
  - Objective variable: LogS
  - Descriptors (6):
    - Molecular weight (MW)
    - Hydrogen bond donor (HBD)
    - Hydrogen bond acceptor (HBA)
    - Number of rings (CIC)
    - Topological polar surface area (TPSA)
    - Number of rotatable bonds (nBR)

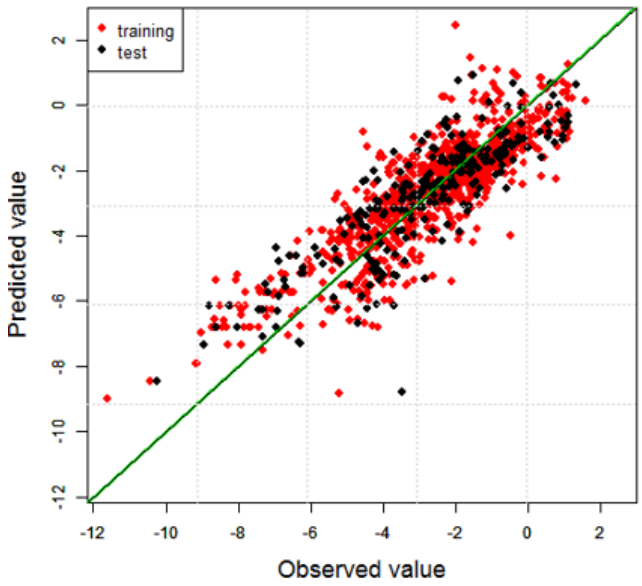
Hou, T J and Xia, K and Zhang, W and Xu, X J, *J. chem. inf. comput. sci.*, 44, 266-275, 2004

QSPR Model Construction

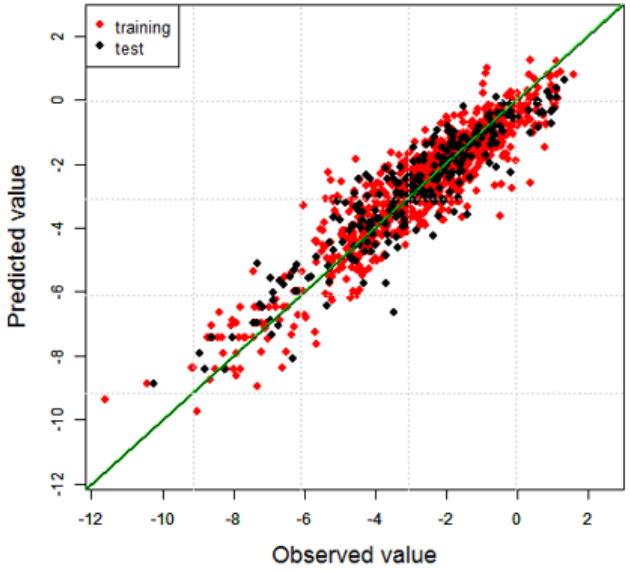
- 7 Gaussians formed a prior PDF: p(**x**)
- With the Gaussians, GMMs/cMLR model was constructed.

	R^2	RMSE	R_{pred}^2	$RMSE_{pred}$
MLR	0.736	1.061	0.722	1.131
GMMs/cMLR	0.853	0.791	0.854	0.820

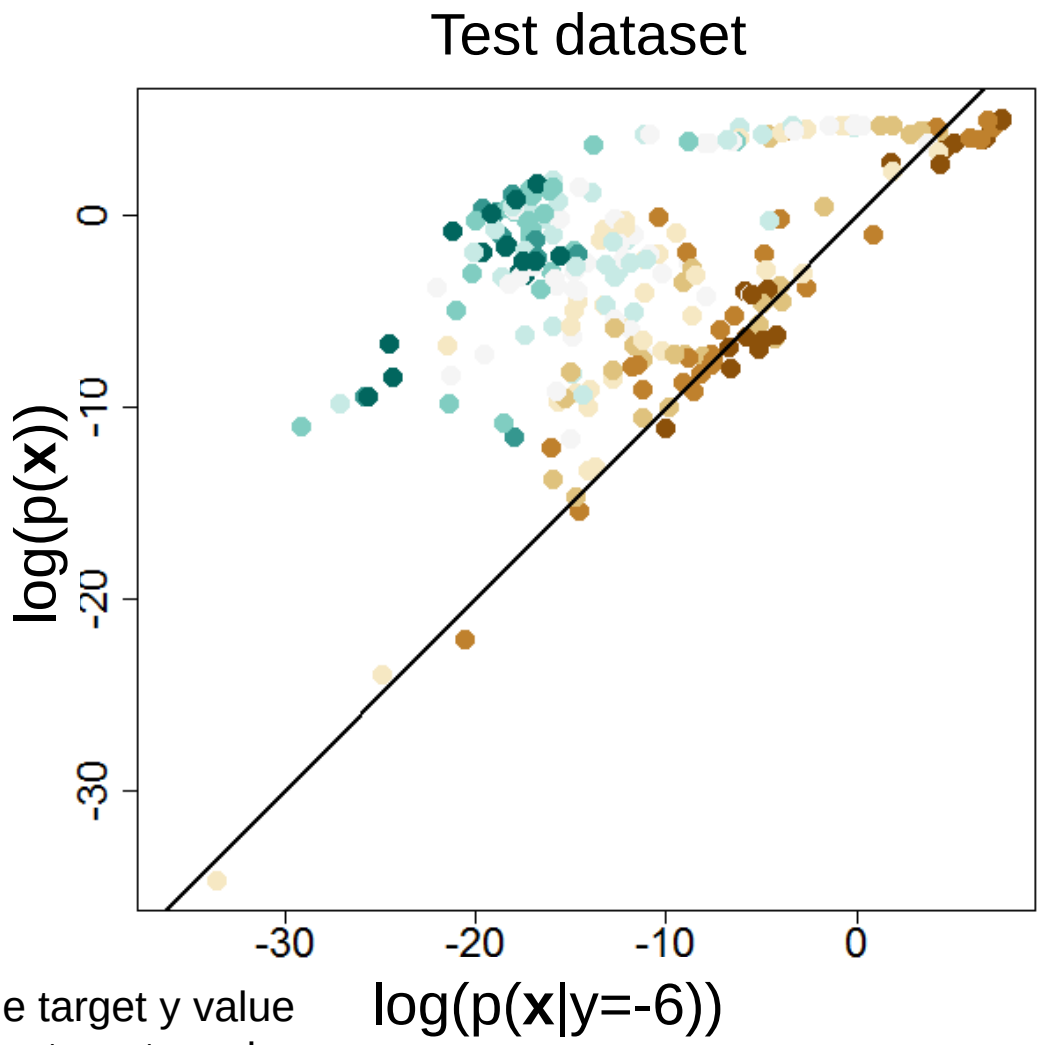
MLR



GMMs/cMLR



Posterior PDF of **x** Given a Specific y Value p(**x**|y)



Green: far from the target y value
Brown: close to the target y value

- Could inherit the prior distribution's feature

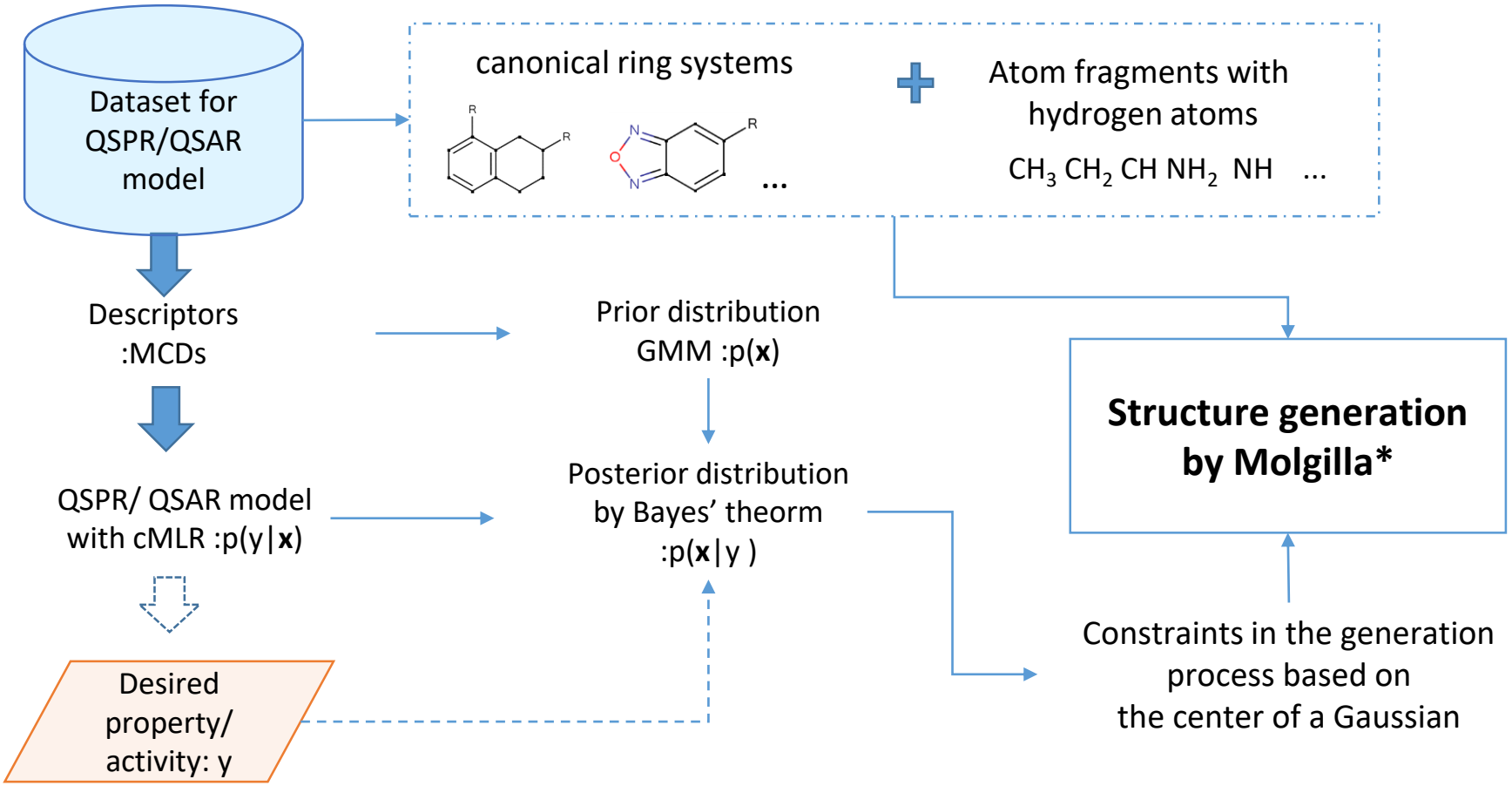
Partial Conclusions

- AD was considered by the posterior PDF with GMMs and cluster-wise multiple linear regression(cMLR)

- Posterior PDF gains information from the degree of closeness to a target y value.

- GMMs/cMLR showed better predictability than MLR did.

4. Proposed Workflow for Inverse QSPR/QSAR



*Molgilla is a structure generator developed by our lab.

Miyao, T., Arakawa, M., Funatsu, K., *Mol. Inform.,* 29, 111–125, 2010
Miyao, T., Kaneko, H., Funatsu, K., *Mol. Inform., 33*, 764-778, 2014

Case Study (Inverse QSAR)

- Target: Thrombin
- Dataset: from ChEBML 20,
  1705 samples annotated with pK_i (inhibition constant)
  - confidence score > 7
  - bioactivity type K_i
  - assay type = B
- Elimination of peptide (more than 10 amide bonds or MW > 1,000)
- Descriptors: 27 MCDs

CIC	R05	aR	ZM1V	nBM	nHAcc Lipin	nCH_2R_2	$nCHR_3$	nCH_3R
nCH_3X	nOH	=O	$nArNR_2$	nArCO	TPSA	LL	LD	LP
AA	AP	AN	DD	RL	RA	RD	RP	RR

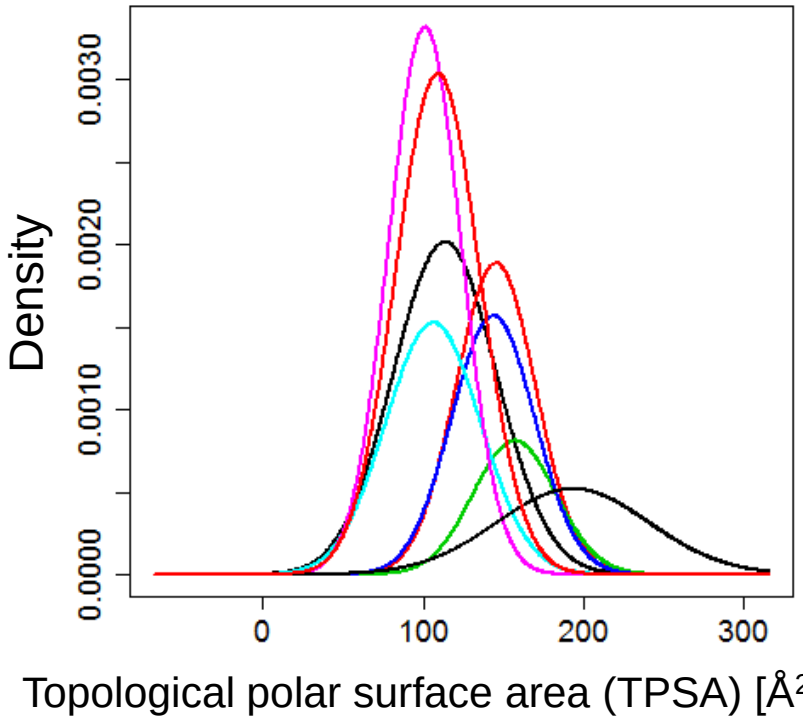
Blue: Sum of topological distance-based descriptors
- inspired by the Chemically Advanced Template Search (CATS)

Reutlinger, M., Koch, C. P., Reker, D., Todoroff, N., Schneider, P., Rodrigues, T., Schneider, G., *Mol. Inform. 32*, 133–138. 2013

p(**x**) and p(y|**x**) with GMMs/cMLR

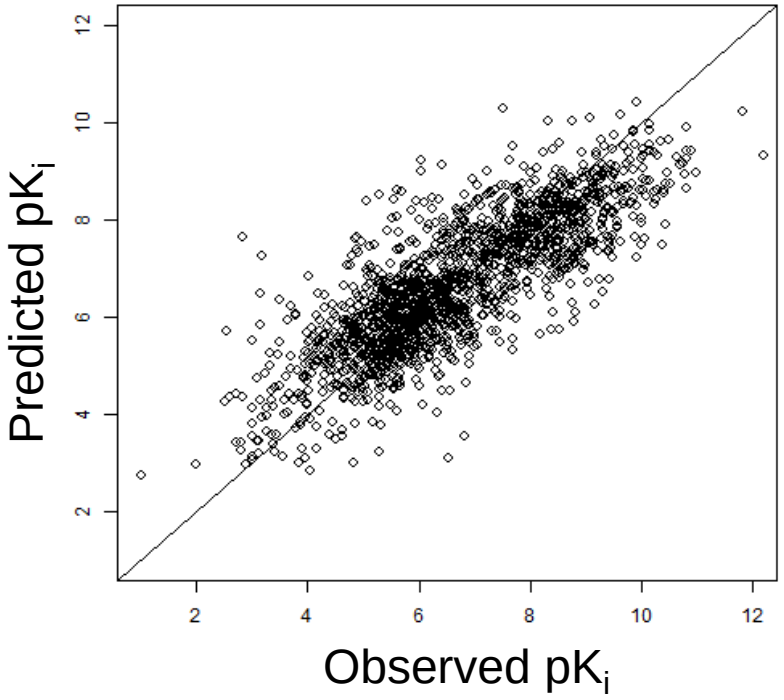
p(**x**)

- 8 Gaussians formed p(**x**)

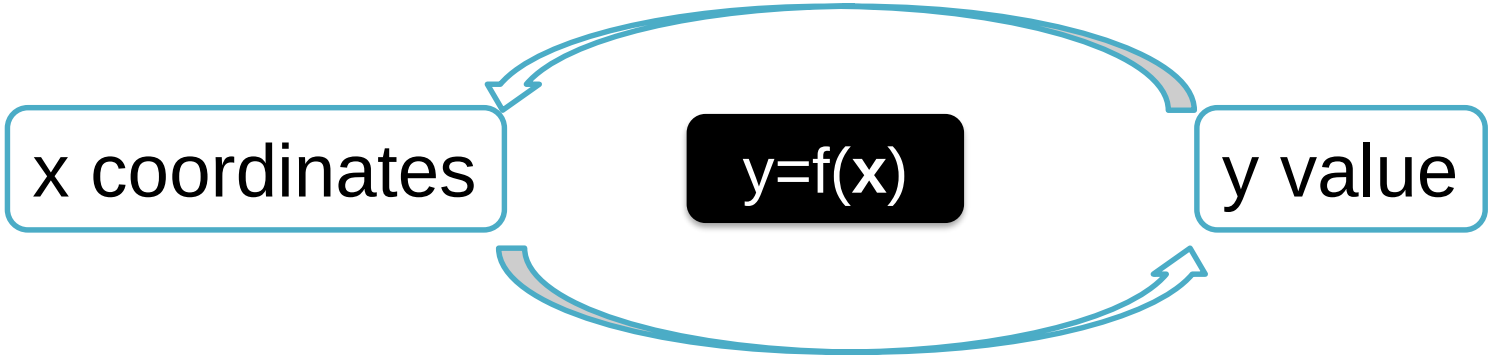


p(y|**x**)

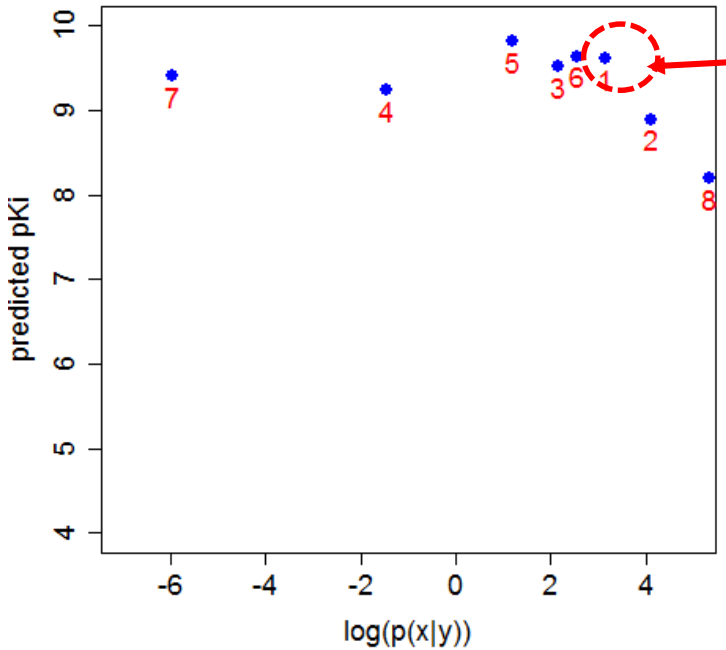
- RMSE: 0.993
- R^2: 0.656



Posterior Density p(**x**|y)



p(**x**|y = 11)



Target Gaussian
for structure generation.

Numbers: the Gaussian centers

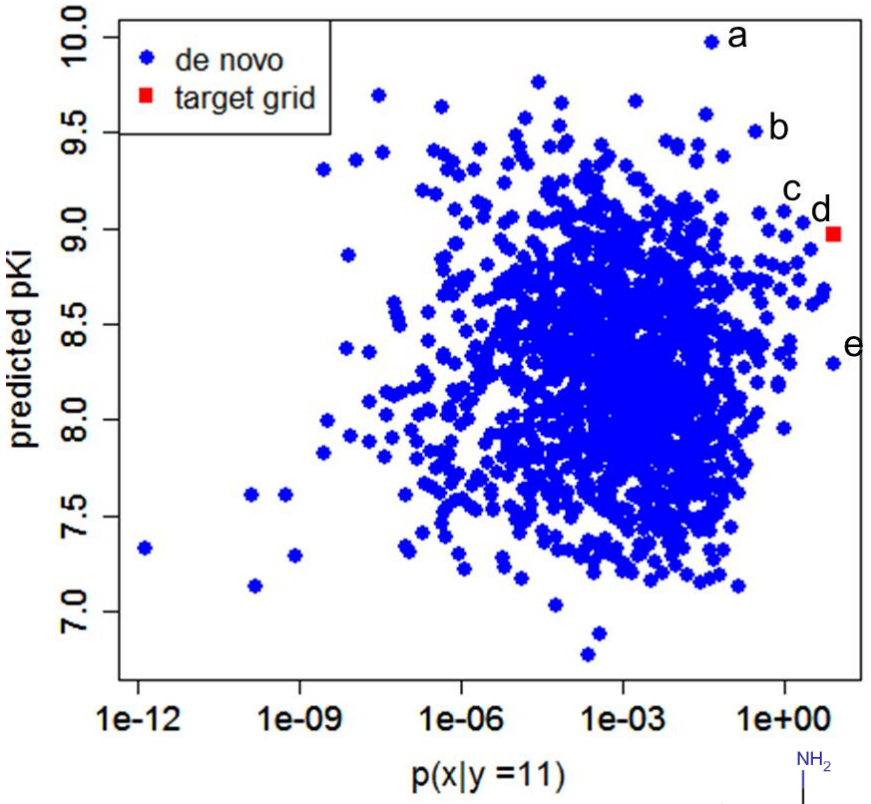
Structure Generation Conditions and Result

- Stochastic generation was conducted in order to generate structures having more building blocks.
- Prohibition rules were applied to avoid generating structures having reactive and unstable substructures
- Ring systems: 289
- Atom fragments: C, N, O, F, Cl, Br, I

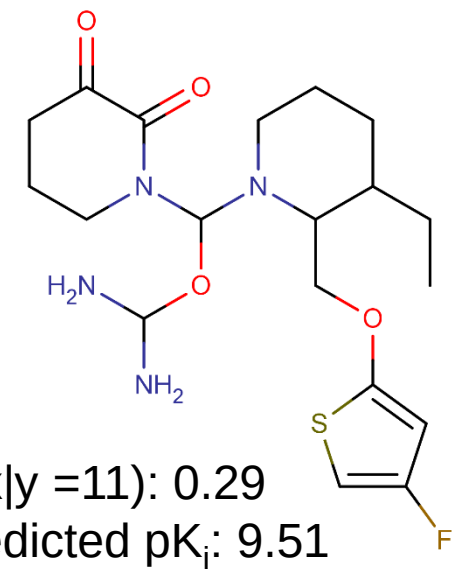
Generation Result

Generated structures	1,739

Generated Structures

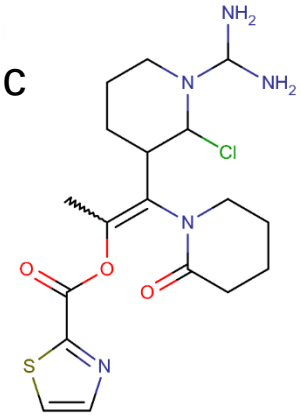


b



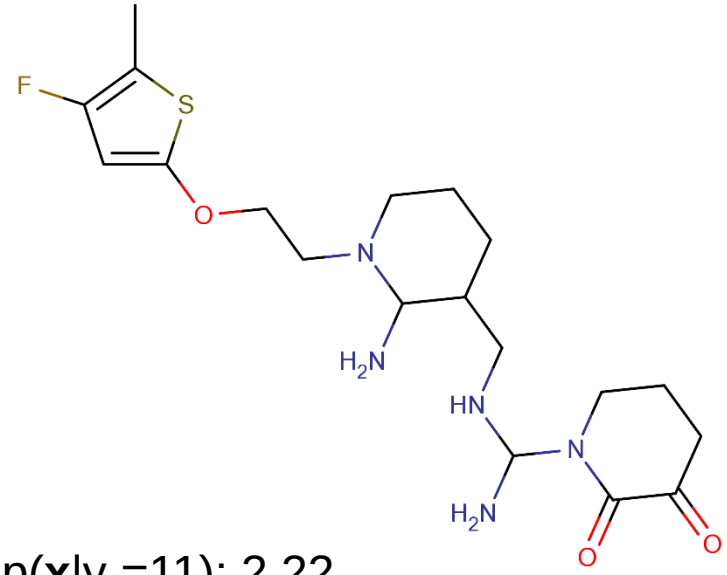
p(**x**|y =11): 0.29
predicted pK_i: 9.51

c



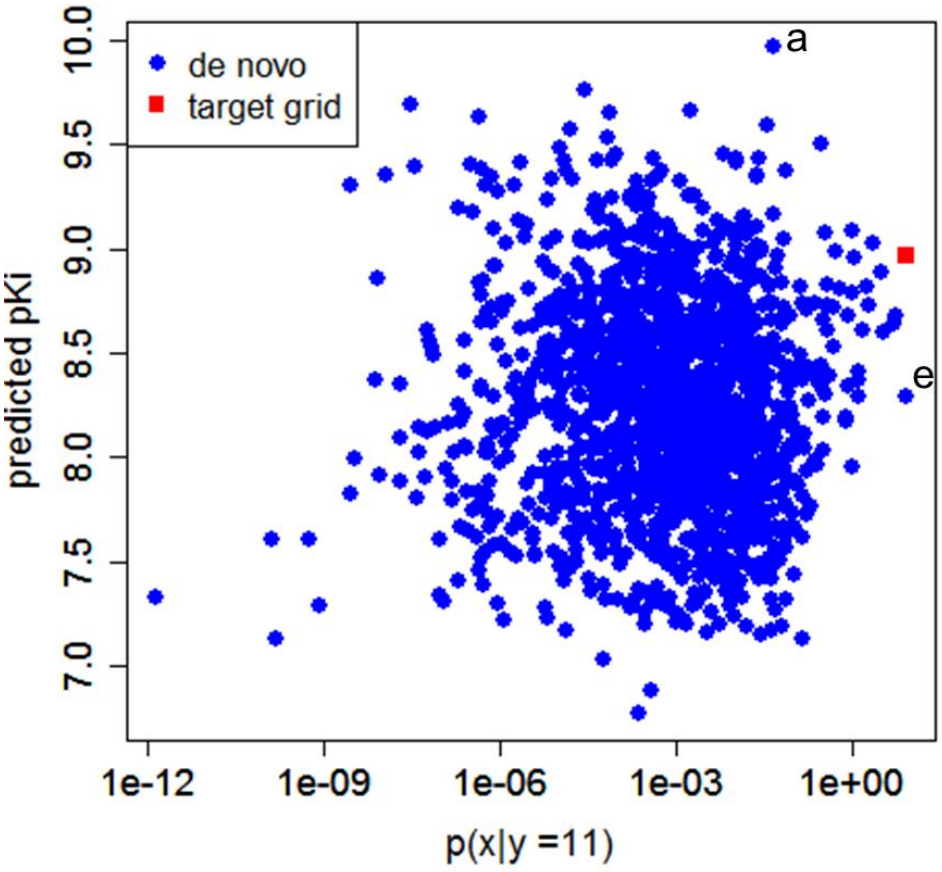
p(**x**|y =11): 1.02
predicted pK_i: 9.09

d

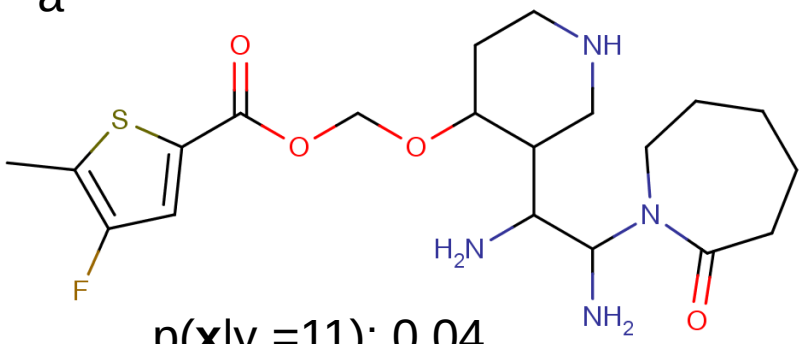


p(**x**|y =11): 2.22
predicted pK_i: 9.03

Generated Structures

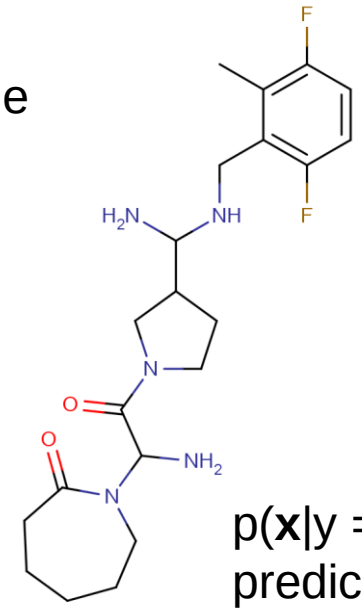


a



p(**x**|y =11): 0.04
predicted pK_i: 9.97

e



p(**x**|y =11): 8.35
predicted pK_i: 8.29

Partial Conclusions

- Structure generation system based on inverse QSPR/QSAR was constructed
  - A Gaussian center of posterior PDF p(**x**|y) is selected for structure generation constraints

- As a practical application, small molecules for thrombin inhibitor candidates were generated.

Summary

- Efficient chemical graph construction algorithms were introduced.
  - ring systems and atom fragments combination
  - constraints by MCDs are considered during generation
- Inverse QSPR/QSAR analysis methodology was proposed.
  - by introducing PDFs with GMMs/cMLR
    - AD consideration
    - higher predictability than with MLR
- A structure generation system by combining these methodologies was proposed in order to generate chemical structures *de novo.*

Acknowledgements

– Dr. Miyao, Tomoyuki

– Dr. Kaneko, Hiromasa

– Dr. Escobar, Matheus

– Dr. Tanaka, Kenichi

– Prof. Dr. Schneider, Gisbert

– Dr. Schneider, Petra