Development of a Structure Generator to Explore Target Areas on Chemical Space

Kimito Funatsu
Department of Chemical System Engineering,
The University of Tokyo

This materials will be published on “Molecular Informatics”

Copyright: *Funatsu Lab.*
The University of Tokyo

Drug Development

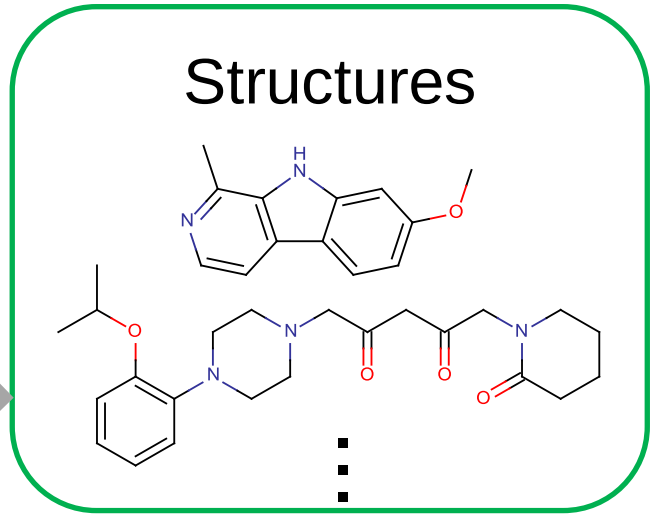
- Conditions lead compounds are required to meet
  - Biological activity for specific target
  - Various properties

    (ADME-Tox, synthetic accessibility, etc…)

➢ On the first stage of drug development, various structures with high activity are required

Copyright: *Funatsu Lab.*
The University of Tokyo

Structure Generators and SAR



Descriptors

Structure-activity relationships (SAR)

Activity

Electronic properties
Molecular weight
Number of rings etc.

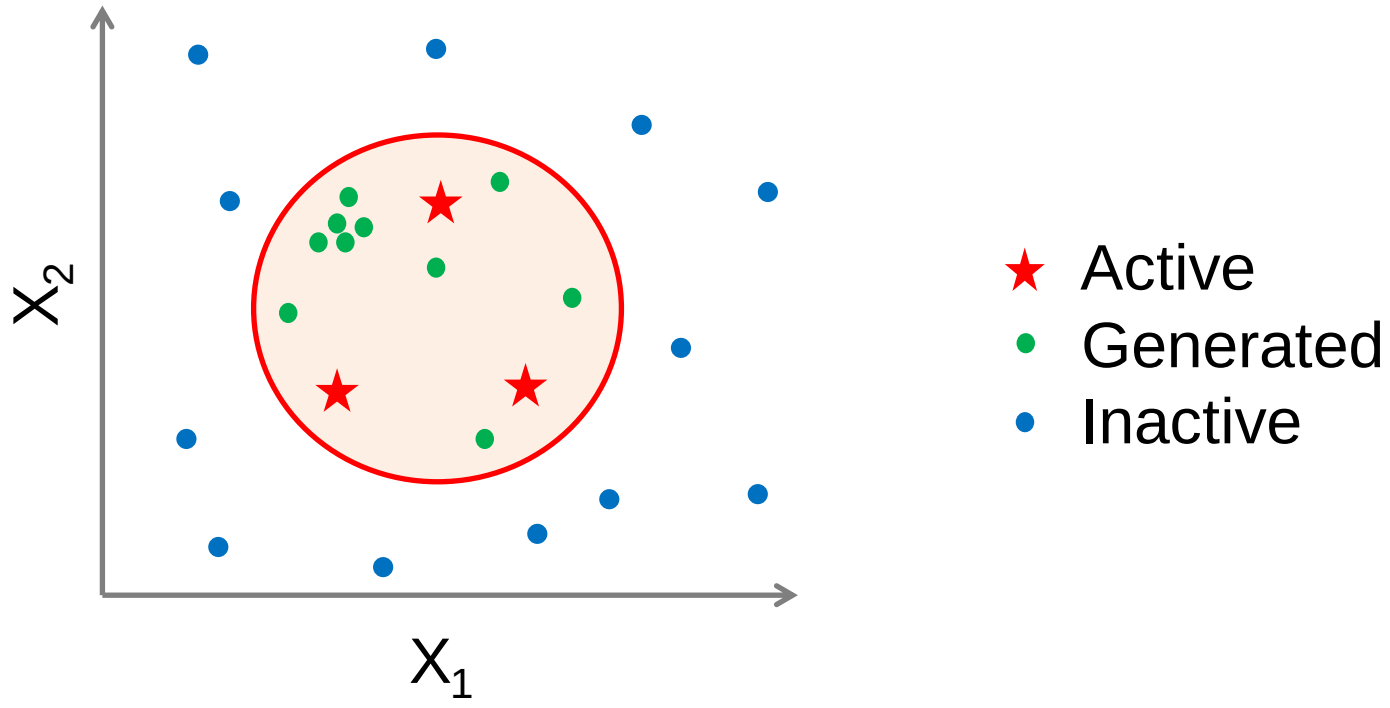
Structure generators that aim to generate highly active structures are proposed.[1][2]

[1] B. Pirard, *Expert Opin. Drug Discov.*, **6**, 225, 2011
[2] H. Mauser, W. Guba, *Curr. Opin. Drug Discov. Devel.*, **11**, 365, 2008

Copyright: *Funatsu Lab.* *The University of Tokyo*

Chemical Space

Chemical space is defined by descriptors.

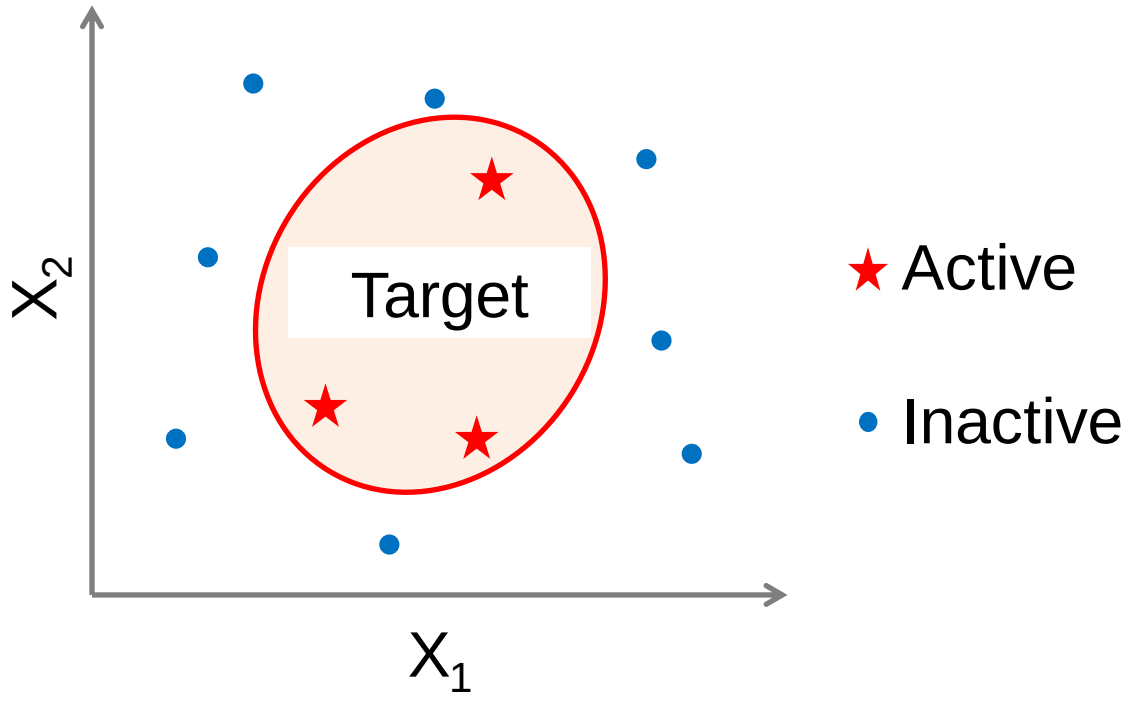


- It is necessary to consider the distribution of ~~the~~ generated structures on the chemical space

Copyright: *Funatsu Lab.*
The University of Tokyo

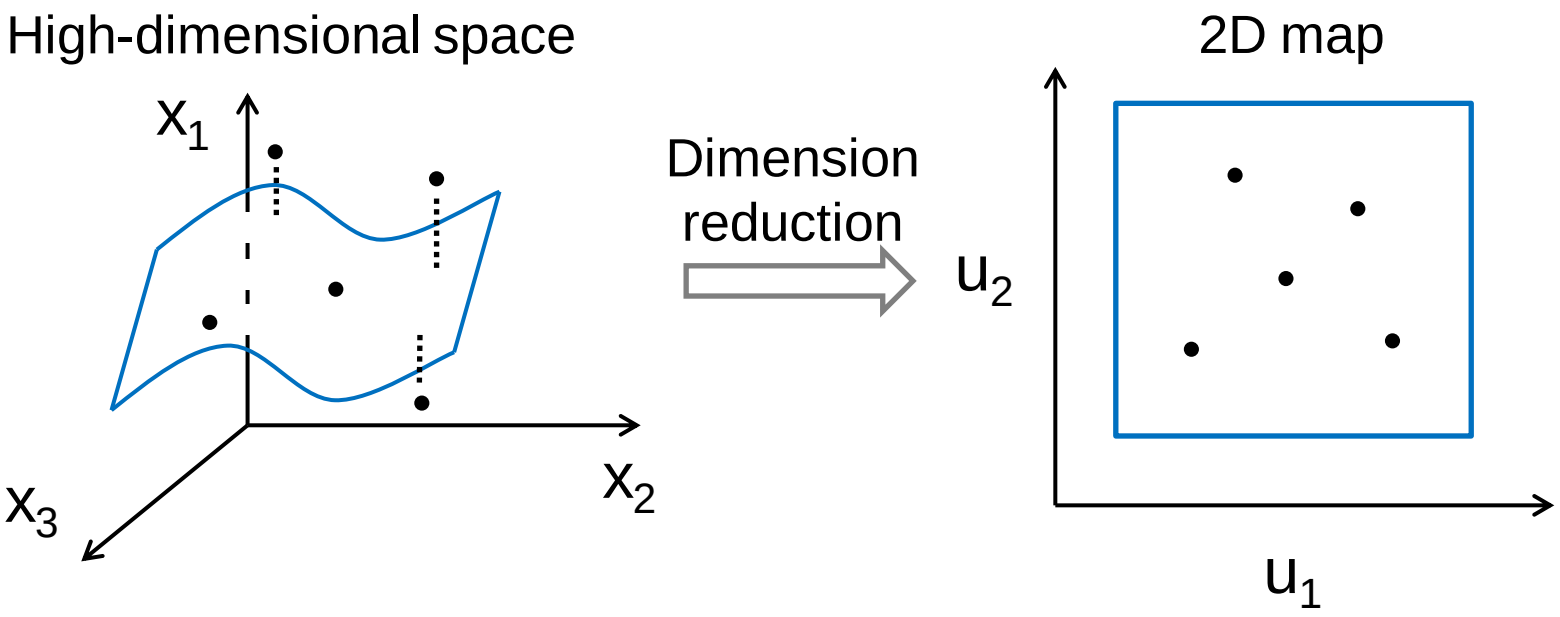
Objective

Development of a structure generator
for searching target areas in chemical space



Copyright: Funatsu Lab.
The University of Tokyo

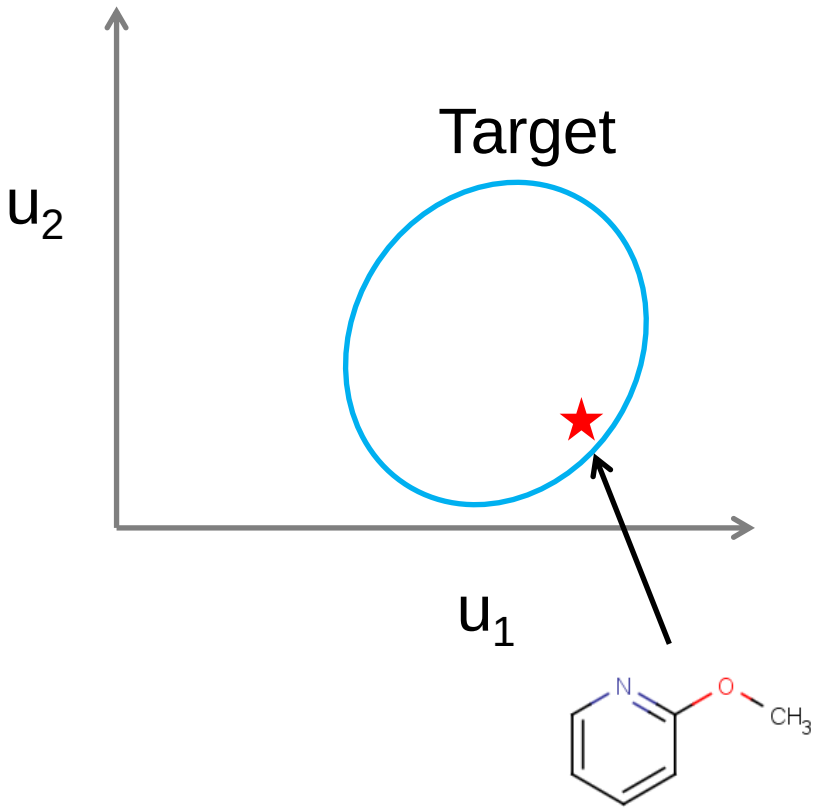
Visualization of chemical space



- A system for structure generation on 2D maps, *de novo* Design Algorithm for Exploring Chemical Space DAECS

Copyright: *Funatsu Lab*
The University of Tokyo

Systems



★ Seed structures

• Pooled structures

①Input of initial seeds

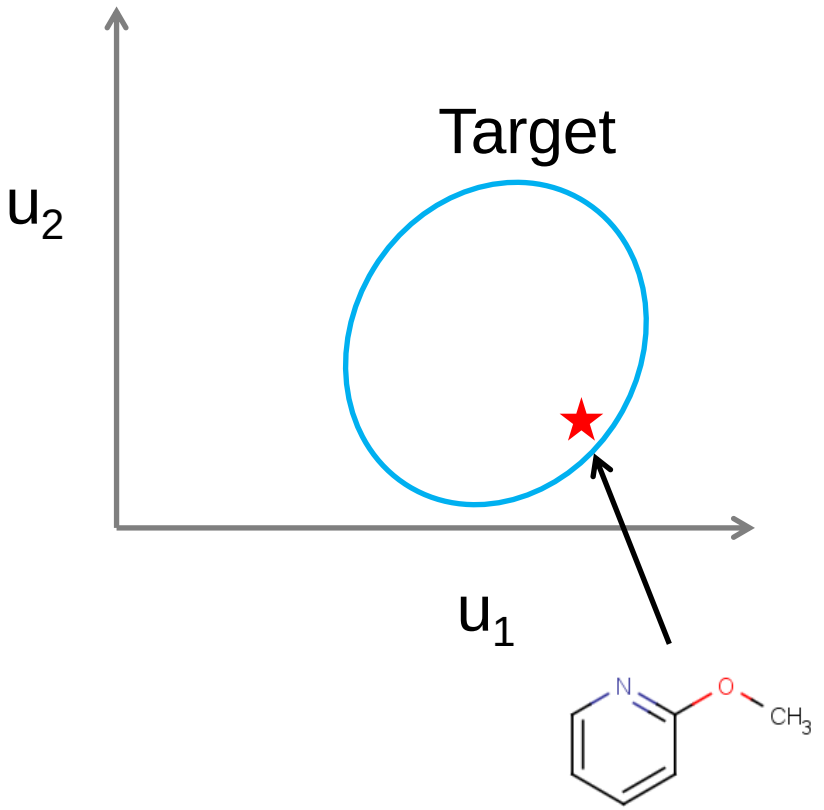
②Generation of new structures

③Filtering and recording

④Probabilistic selection of new seeds

Copyright: *Funatsu Lab.*
The University of Tokyo

Systems



★ Seed structures

• Pooled structures

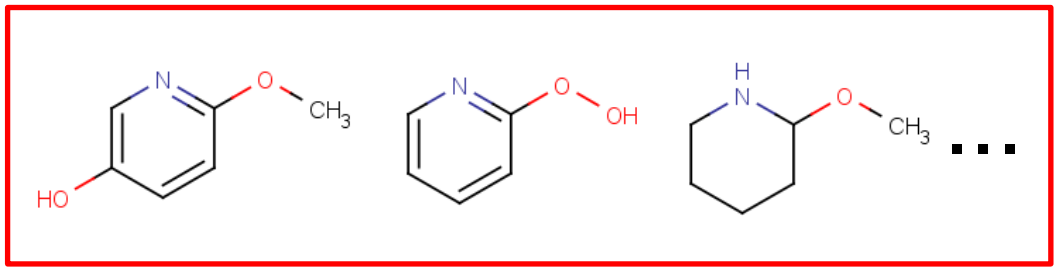
①Input of initial seeds

②Generation of new structures

③Filtering and recording

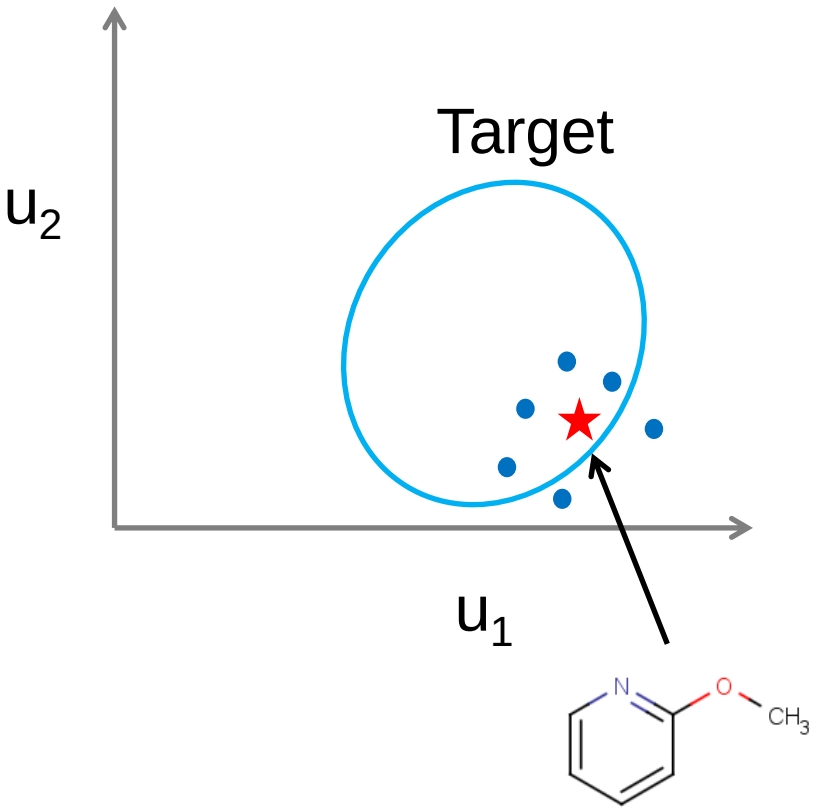
④Probabilistic selection of new seeds

New structures



Copyright: *Funatsu Lab.*
The University of Tokyo

Systems



★ Seed structures

• Pooled structures

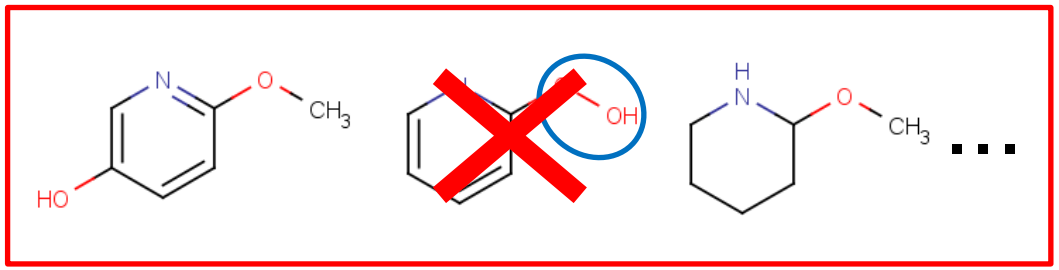
①Input of initial seeds

②Generation of new structures

③Filtering and recording

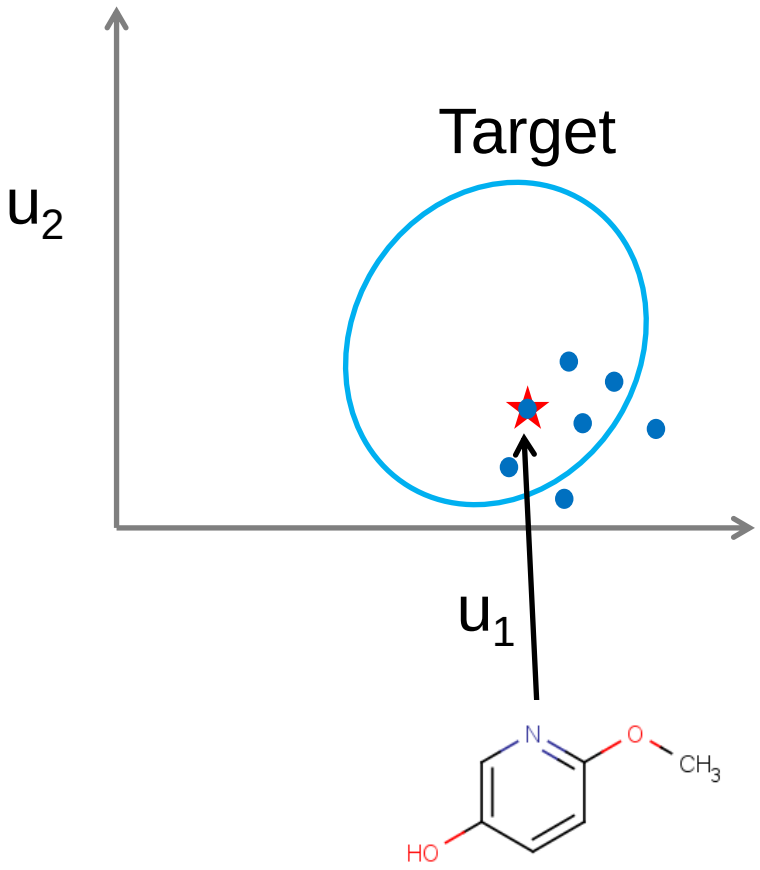
④Probabilistic selection of new seeds

New structures



Copyright: *Funatsu Lab.*
The University of Tokyo

Systems



★ Seed structures

• Pooled structures

①Input of initial seeds

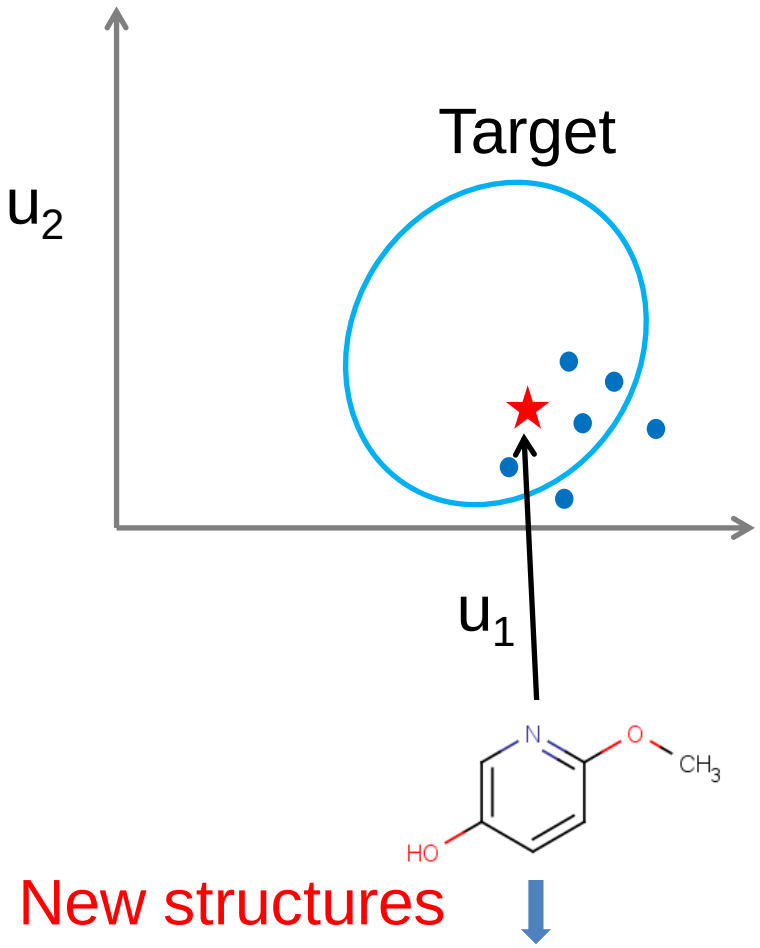
②Generation of new structures

③Filtering and recording

④Probabilistic selection of new seeds

Copyright: *Funatsu Lab*
The University of Tokyo

Systems



★ Seed structures

• Pooled structures

①Input of initial seeds

②Generation of new structures

③Filtering and recording

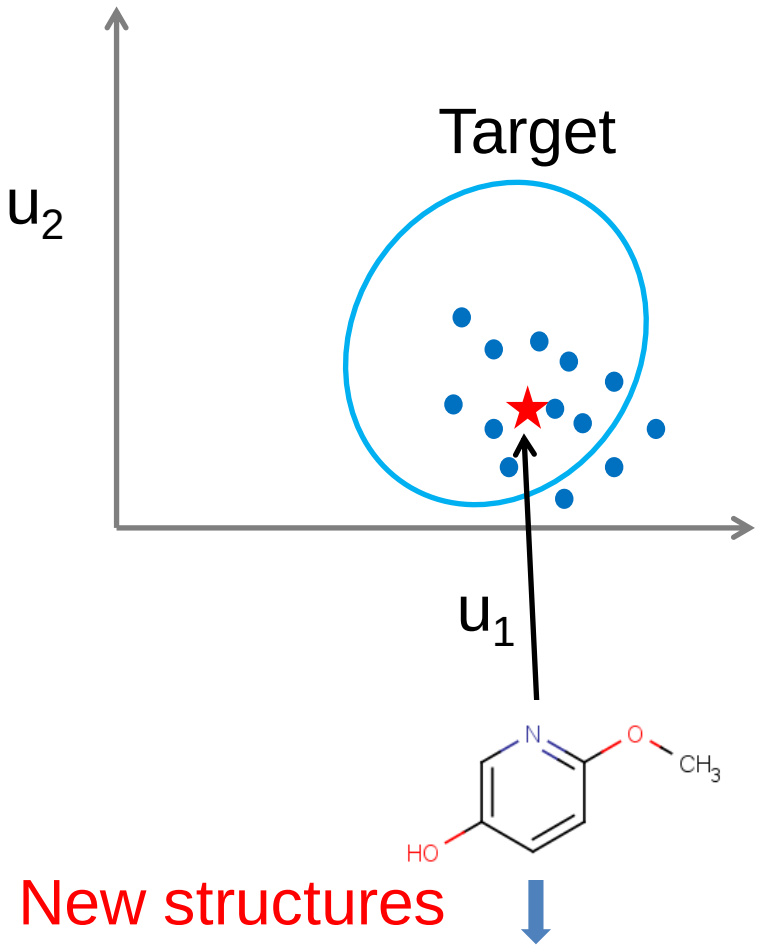
④Probabilistic selection of new seeds

New structures

…

Copyright: *Funatsu Lab.*
The University of Tokyo

Systems



★ Seed structures

• Pooled structures

①Input of initial seeds

②Generation of new structures

③Filtering and recording

④Probabilistic selection of new seeds

New structures

…

Copyright: *Funatsu Lab.*
The University of Tokyo

Case Study

α2A data : GVK[1] database

ligand-binding affinity for α_{2A} adrenergic receptor

training data × 300, test data × 335

Descriptors : Fingerprints from PubChem[2] (460 bit)

Ex.

- >= 4 H
- >= 2 saturated or aromatic nitrogen containing ring size 6
- C(-C)(-H)(=N)

Visualization : Generative Topographic Mapping(GTM)[3]

[1] GVK Bio Databases, http://www.gvkbio.com/informatics.html

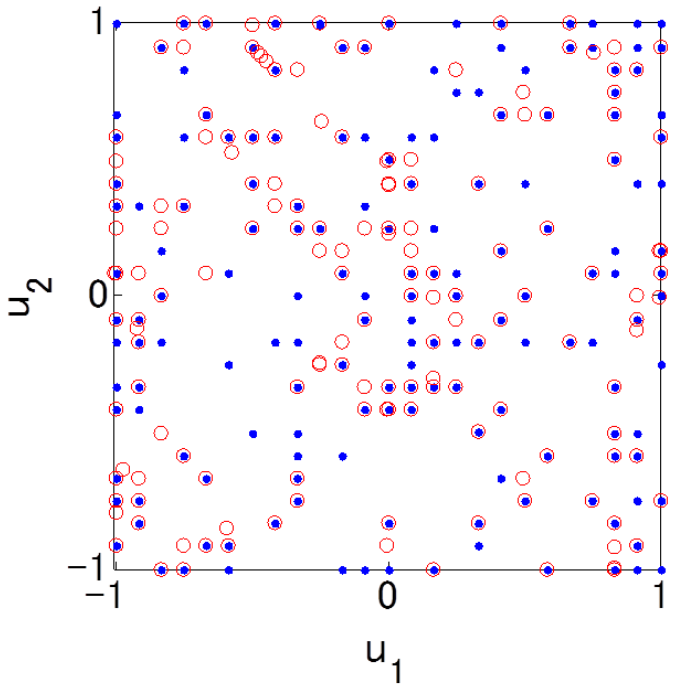
[2] PubChem, http://pubchem.ncbi.nlm.nih.gov/

[3] C. M. Bishop, Markus Svensen, *Neaural Computation*, **10**, 215, 1998.

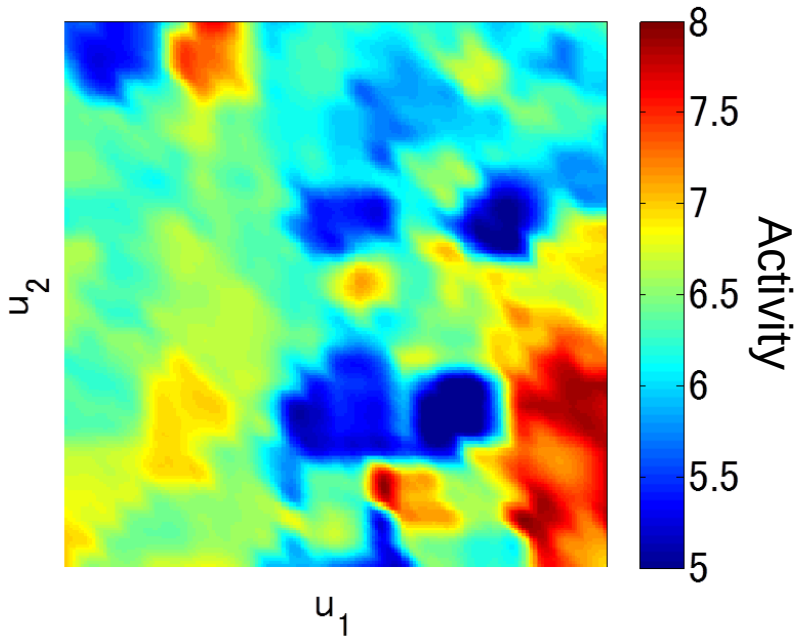
Copyright: *Funatsu Lab.*

The University of Tokyo

Visualization of Chemical Space



- Training data
- Test data



Predicted activity calculated with 3-nearest neighbors method

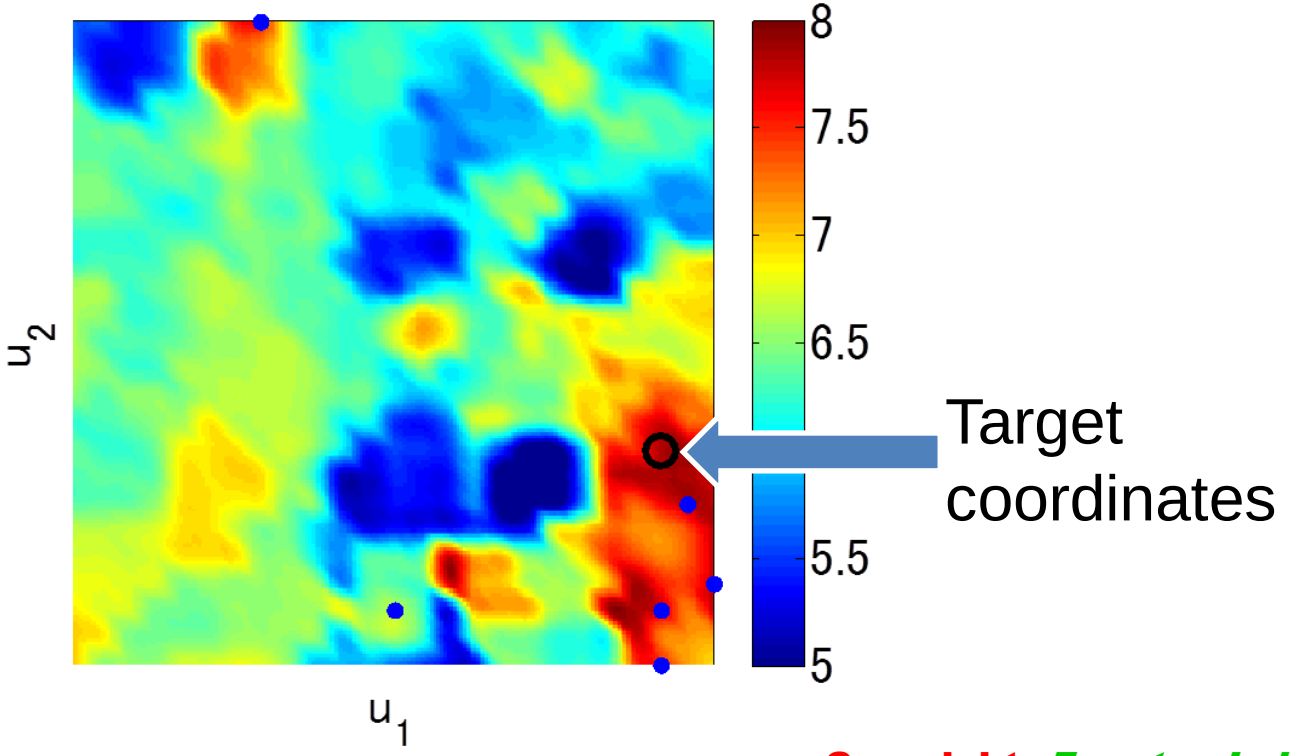
$RMSE_{cv}$ = 0.45
$RMSE_{pred}$ = 0.42

Copyright: *Funatsu Lab.*
The University of Tokyo

Structure Generation

Structure Generation：
1.DAECS
2.Conventional method : selection of new seeds
by predicted activity (SVR[1])

Initial seeds： 7 structures with > 8 activity values
Cycles：100
Trial：10 times



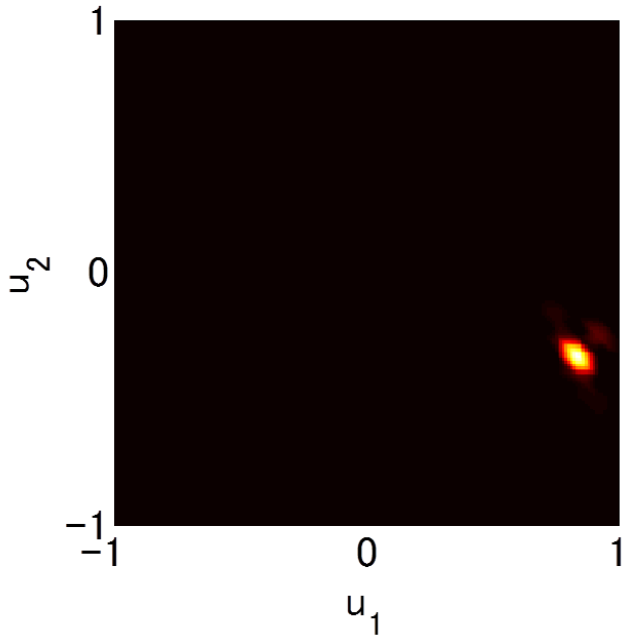
[1] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

Copyright: *Funatsu Lab.*
The University of Tokyo

Structure Generation

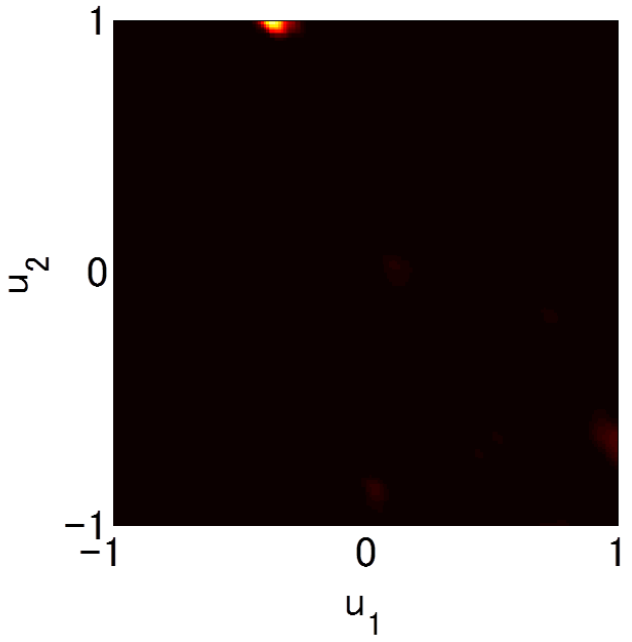
- Distribution of generated structures

① DAECS



258,068 structures

② Conventional method



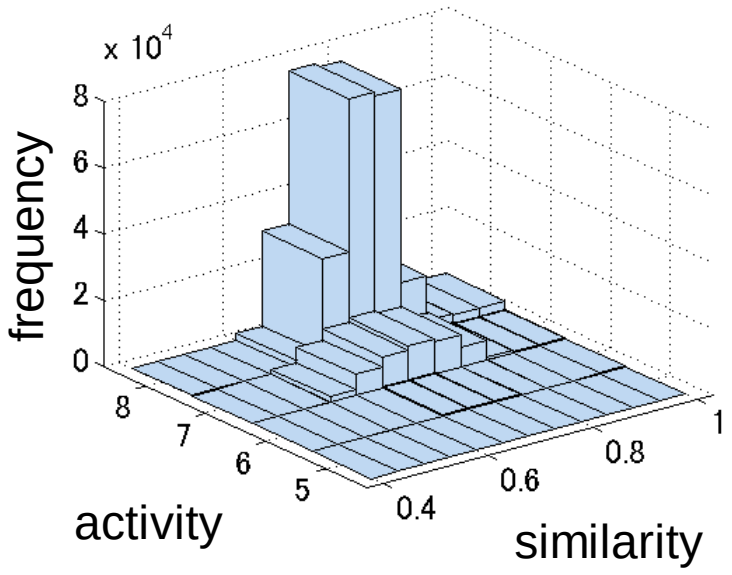
311,975 structures

Copyright: *Funatsu Lab.*
The University of Tokyo

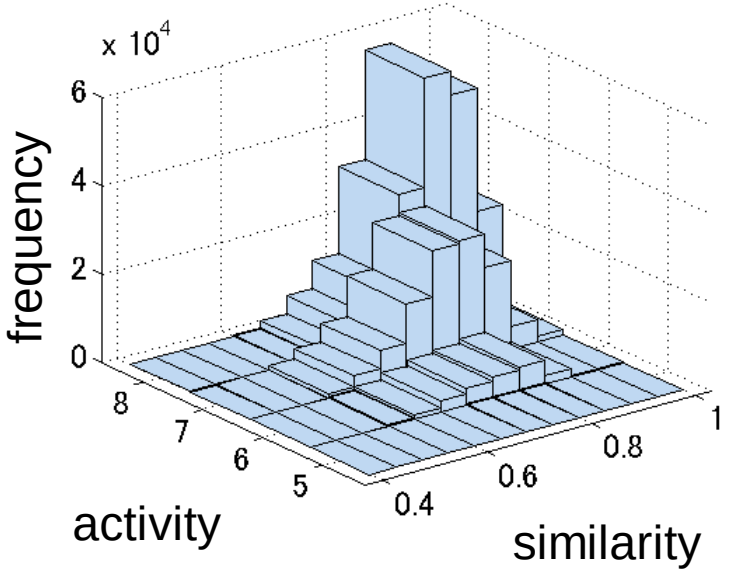
Diversity and Activity

- Distributions of similarity with seed structures and predicted activity
  - Similarity : tanimoto coefficient
  - Activity : 3-NN method

① DAECS

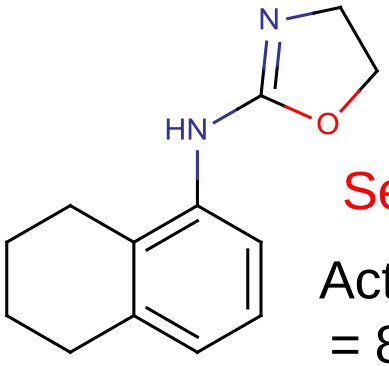


② Conventional method



Copyright: Funatsu Lab.
The University of Tokyo

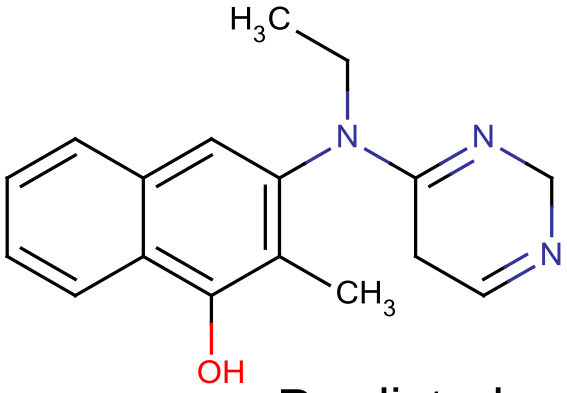
Example of Generated Structures



Seed

Activity
= 8.18

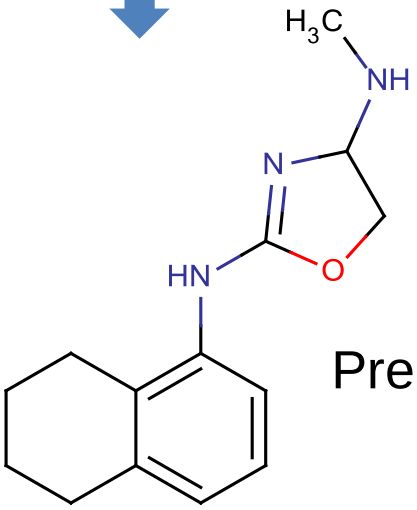
Similarity
= 0.60



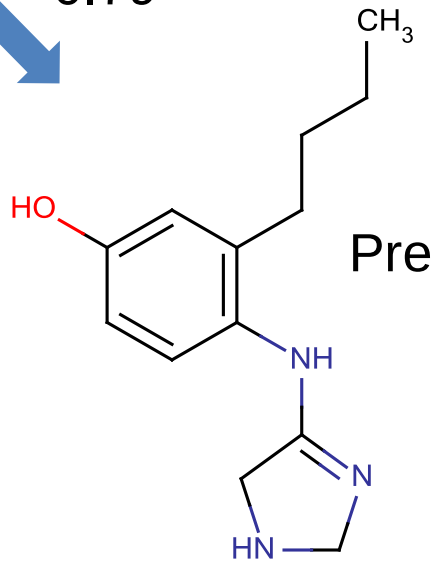
Predicted activity
= 7.48

Similarity
= 0.98

Similarity
= 0.79



Predicted activity
= 7.85



Predicted activity
= 7.84

Copyright: *Funatsu Lab.*
The University of Tokyo

Conclusions

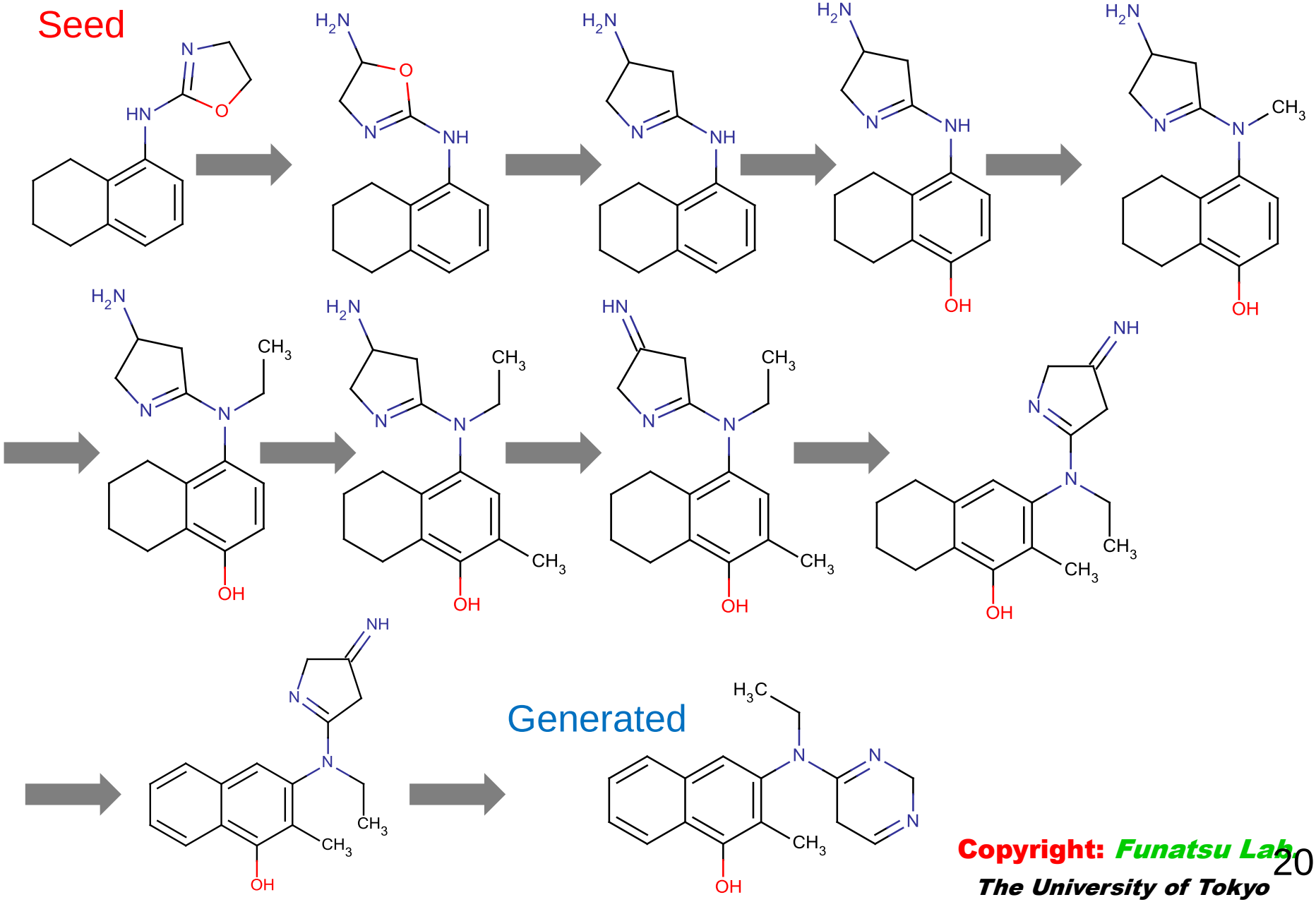
- A structure generator ‘DAECS’ was developed. DAECS aims to generate structures in target areas on visualized chemical space.
- In a case study with GVK data, visualization of the chemical space and structure generation were performed.
- Distribution, diversity and activity of generated structures were verified. It was showed that DAECS can generate diverse structures, which are distributed in the target area on visualized chemical space.

Copyright: *Funatsu Lab.*
The University of Tokyo

APPENDIX

Copyright: *Funatsu Lab.*
The University of Tokyo

Path



Copyright: Funatsu Lab.
The University of Tokyo

Calculation of Similarity

	Fingerprints
Structure A	0 1 0 1 0 0 0 1 0 0 1 1 1 1 0 1 …
Structure B	0 0 1 1 0 0 1 0 1 1 1 1 0 1 0 0 …

$$t = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

Number of bits that satisfy below

M_{01} : A→0, B→1

M_{10} : A→1, B→0

M_{11} : A→1, B→1

Copyright: *Funatsu Lab.*
The University of Tokyo

Selection of Next Seeds

Probabilistic selection (roulette) based on a scoring function

$$SCORE(r,d) = \exp\left(-\frac{r^2}{\sigma_r^2}\right) \times \exp\left(-\frac{d^2}{\sigma_d^2}\right)$$

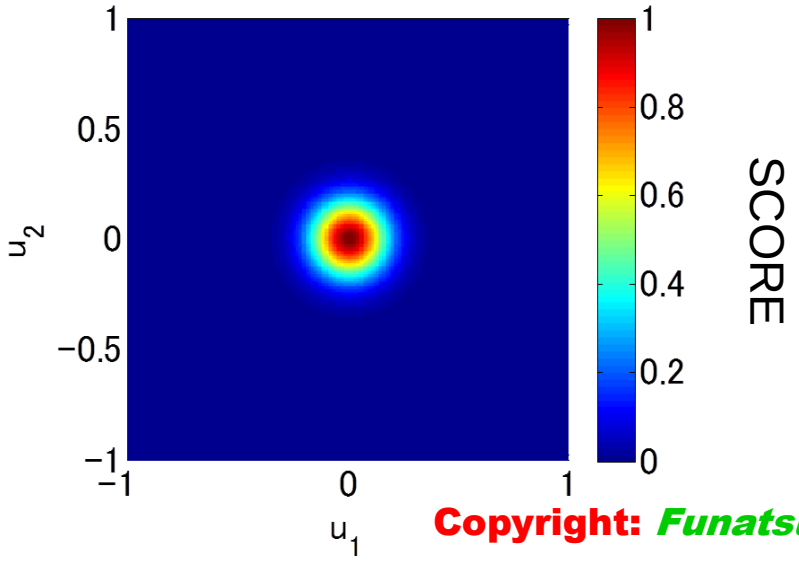
r : Distance from target on the map

d : Distance from the map

σ_r : Hyper-parameter

σ_d : Hyper-parameter

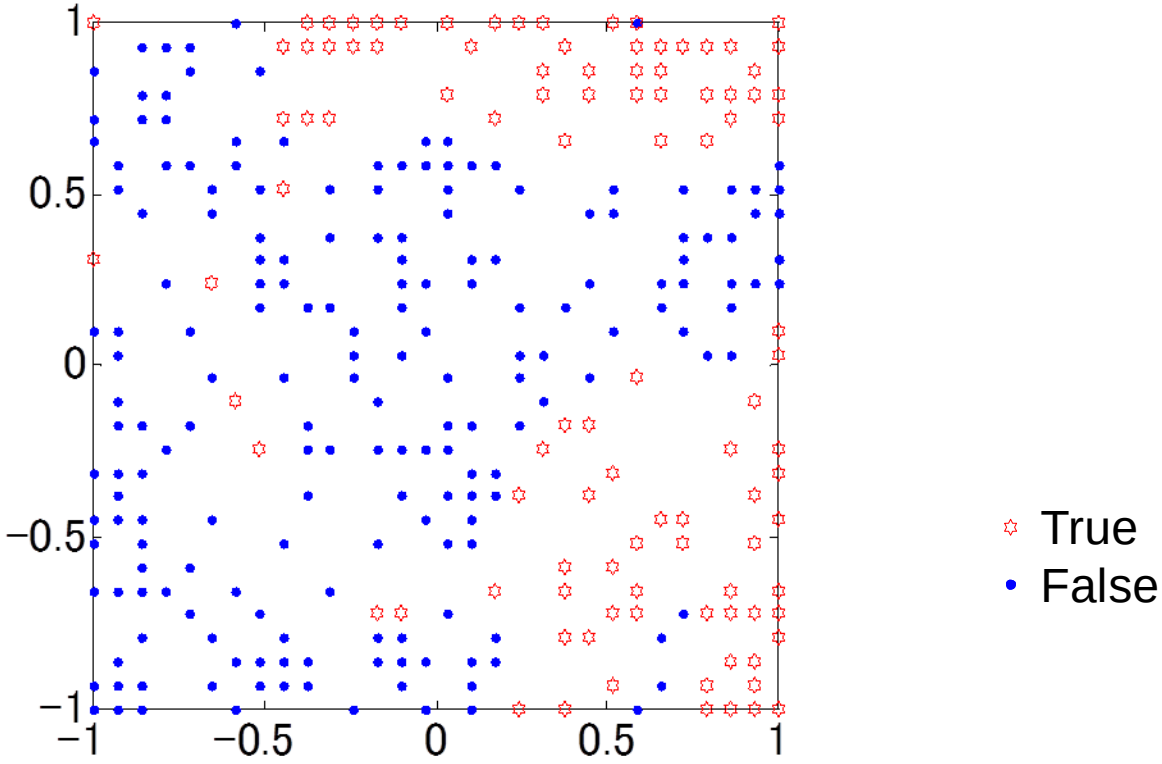
Example of Scoring function



Copyright: Funatsu Lab.
The University of Tokyo

Distribution of a Specified Descriptor

- 145 >= 1 saturated or aromatic nitrogen-containing ring size 5



Copyright: Funatsu Lab.
The University of Tokyo