JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

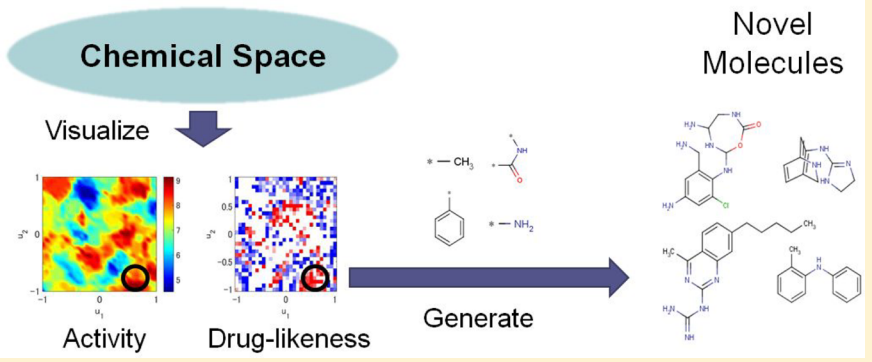
pubs.acs.org/jcim

Chemical-Space-Based de Novo Design Method To Generate Drug-Like Molecules

Shunichi Takeda, Hiromasa Kaneko, and Kimito Funatsu*

Department of Chemical Systems Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Supporting Information



ABSTRACT: To discover drug compounds in chemical space containing an enormous number of compounds, a structure generator is required to produce virtual drug-like chemical structures. The de novo design algorithm for exploring chemical space (DAECS) visualizes the activity distribution on a two-dimensional plane corresponding to chemical space and generates structures in a target area on a plane selected by the user. In this study, we modify the DAECS to enable the user to select a target area to consider properties other than activity and improve the diversity of the generated structures by visualizing the drug-likeness distribution and the activity distribution, generating structures by substructure-based structural changes, including addition, deletion, and substitution of substructures, as well as the slight structural changes used in the DAECS. Through case studies using ligand data for the human adrenergic alpha2A receptor and the human histamine H1 receptor, the modified DAECS can generate high diversity drug-like structures, and the usefulness of the modification of the DAECS is verified.

1. INTRODUCTION

Drug development has dramatically changed with the introduction of computational chemistry. Virtual screening of a chemical structure library is performed to efficiently search drugs. Virtual libraries include structures that are expected to be drug compounds. De novo drug design is a promising way to generate ligand candidates for a target protein. Many de novo design methods have been proposed, and recent studies have been summarized in review papers.[1−3] The compound generator developed by Brown et al.[4] generates structures with high activity and desired properties based on a genetic algorithm (GA). Structures are optimized by exchanging substructures, that is, cross over, changing atoms or bonds (i.e., mutation), and selection of structures. When structures with a specific activity or other property value are desired, the structure generator developed by Miyao et al.[5] is appropriate. Structures are exhaustively generated in a descriptor domain determined based on the inverse quantitative structure property relationship or the inverse quantitative structure activity relationship. Hartenfeller et al. developed the DOGS (design of genuine structures) software,[6] which can generate structures based on known chemical reactions for actual synthesis. A reaction site of a structure is searched for based on a reaction library, and a reaction with a substructure that is searched for in a building block library is applied to the reaction site. The reaction site and the substructure are selected to provide the smallest structural changes. When DOGS is used, it is possible to construct a library of structures similar to an original structure considering synthesis routes.

Drug discovery using molecular fragments that bind to a desired target protein is called fragment-based drug discovery (FBDD).[7] GA-based de novo design of inhibitors (GANDI)[8] is one of the FBDD methods. Structures are generated by connecting fragments with linker fragments, and the connection is based on the simultaneous optimization of the force-field energy and a score representing the two-dimensional (2D) similarity to known inhibitors or the three-dimensional (3D) overlap of known binding modes. It is possible to generate structures by considering docking to a target protein. Mishima et al.[9] developed the de novo design algorithm for exploring chemical space (DAECS), which can generate structures in a target area in chemical space selected by the user. Chemical space is defined by structure descriptors and visualized using

Received: January 28, 2016
Published: September 15, 2016

DOI: 10.1021/acs.jcim.6b00038
J. Chem. Inf. Model. 2016, 56, 1885−1893

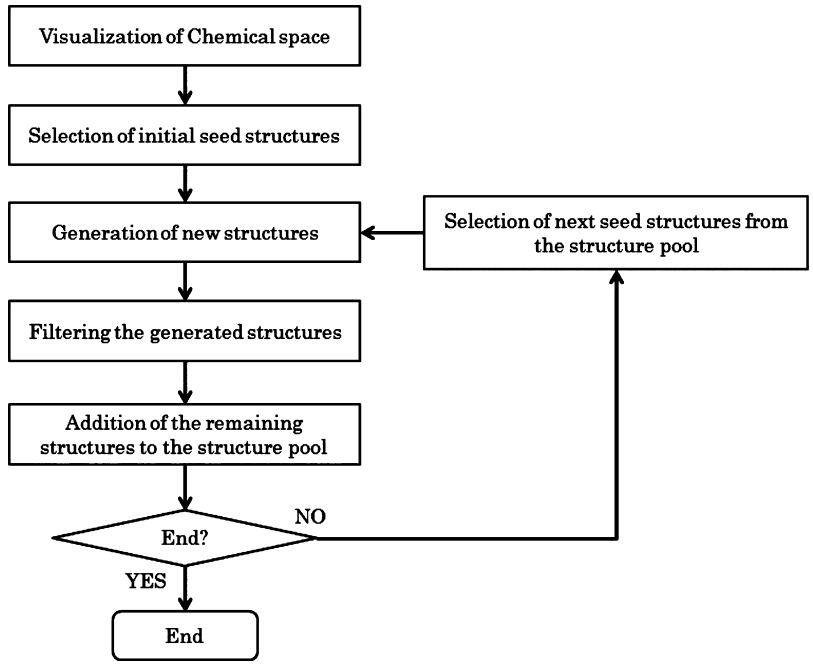


Figure 1. Outline of the DAECS.

training compounds. To aid in the area selection by the user, a 2D plane (map) corresponding to chemical space is constructed using a dimensionality reduction method, such as the self-organizing maps (SOM) method[10] or generative topographic mapping (GTM),[11] and the activity value distribution is displayed on the map by color. Typically, multiple areas of high activity molecules will be found on the map where the activity distribution is shown. One of the regions is selected as a target area to generate chemical structures. A library of generated structures is constructed by repeating the cycle of selecting seed structures and generating new structures, which are transformed by slight structural changes from seed structures, such as adding or deleting an atom. Seed structures for the next cycle are selected from the generated structures that are close to the target region. When an area containing high activity molecules is selected as the target region, many high activity structures could be generated because structures close to each other on the map will have similar activity in the case that optimal structure descriptors are used.[12]

In this study, we modify DAECS to improve two aspects. First, the diversity of generated structures. Although DAECS generates a number of structures in a selected target region, the diversity of the generated structures can be low because structural changes in each structure generation cycle are small. We introduce substructure-based structural changes in addition to slight structural changes to enhance the diversity of the generated structures. New structures are generated by addition, deletion, and substitution of substructures, and slight structural changes. The features of the substructures have a major effect on the generated structures. We use some substructures included in training compounds to visualize chemical space because structures should be generated in the applicability domain (AD)[13] where the activity prediction model and the dimensionality reduction model show their true performance. Second, visualization of the distribution on the map corresponding to chemical space. In DAECS, only the activity value distribution is displayed on the map to help target area selection. Nevertheless, drugs need to satisfy various properties, such as absorption, distribution, metabolism, and excretion (ADME), toxicity, and synthesizability.[14,15] Although DAECS can consider the activity value of molecules, it cannot consider other properties. Therefore, we represent the drug-likeness[16] distribution on the map. Molecules that have the properties required for drugs are called drug-like molecules. The study of drug-likeness starts with Lipinski's Rule of Five[16] and is now actively discussed. Improvement of the utility of quantitative drug-likeness is one of the concepts. Bickerton et al. ranked compounds and extended the utility by applying drug-likeness to the problem of molecular target druggability.[17] Another concept is drug-likeness with machine-learning. Using a classification method is an effective way to determine whether the compound is drug-like or not.[18−20] Arakawa et al.[21] confirmed that it is possible to visualize the region of drug-like compounds in chemical space by displaying the drug-likeness distribution on a map corresponding to chemical space. By setting a target area based on the displayed activity distribution and drug-likeness distribution, the user can generate structures considering both the activity and the drug-likeness.

To verify the usefulness of the modified DAECS, we perform case studies using ligand data for the human adrenergic alpha2A receptor and the human histamine H1 receptor. We show that our modified DAECS can generate more diverse structures than the traditional DAECS considering the drug-likeness of the generated structures.

2. METHODS

2.1. Traditional DAECS. The procedure for DAECS is shown in Figure 1. First, chemical structures in the existing database are represented as structure descriptors and projected onto a 2D map constructed with a dimensionality reduction method, such as SOM or GTM. Second, the activity distribution on the map is predicted using an activity prediction model constructed with a regression method. Third, the user sets a target coordinate on the map and determines initial seed

DOI: 10.1021/acs.jcim.6b00038
J. Chem. Inf. Model. 2016, 56, 1885−1893

structures that are mapped around the target coordinate. New structures are generated by slight structural changes of the seed structures (e.g., adding or deleting an atom). Structures that will be difficult to synthesize are removed, and the rest of the structures are added to the structure pool. Seed structures are then selected from the structure pool based on the following score:

$$S(r, d) = \exp\left(-\frac{r^2}{\sigma_r^2}\right) \times \exp\left(-\frac{d^2}{\sigma_d^2}\right) \quad (1)$$

where r is the distance from the target coordinate on the map, d is the projection error which is calculated as the distance between the descriptor of the structure and the point in descriptor space corresponding to the projected point on the map by mapping function, and σ^2_r and σ^2_d are hyperparameters. The higher the score of a structure is, the higher the probability that the structure is selected. The cycle of selection of seed structures, generation of new structures, filtering, and addition to the structure pool is repeated until the number of generated structures reaches a required number. Please refer to the paper of Mishima et al.[9] for the details of DAECS.

2.2. Modified DAECS. DAECS was modified by introducing substructure-based structural changes and visualization of the drug-likeness distribution. We describe the details of these modifications in this section.

2.2.1. Substructure-Based Structural Changes. The diversity of structures generated by the traditional DAECS can be low because traditional DAECS includes only slight structural changes in the generation step. Therefore, we introduced substructure-based transformation methods.

2.2.1.1. Transformation Methods. The traditional DAECS uses the 10 transformations in Table 1.[9]

Table 1. 10 Transformations Used in DAECS

transformation	description
1. addition of atoms	add atoms to the seed structure
2. insertion of atoms	insert atoms between two bonded atoms
3. removal of atoms	remove atoms on the edge of the molecule
4. cyclization	connect two atoms to form a ring
5. decyclization	remove a bond of a ring
6. changing a ring into a saturated ring	change a ring into a saturated ring
7. changing a ring into an aromatic ring	change a ring into an aromatic ring
8. mutation of an atom	change an atom to another atom
9. mutation of a bond	change the type of a bond
10. transfer of a bond	change the binding destination to a neighbor atom

[a]In the transformations, a hydrogen atom is not included as an atom.

Modified DAECS includes the 10 transformations of the traditional DAECS with the additional five transformations shown in Table 2.

These transformations are exhaustively applied to all possible atoms and bonds.

2.2.1.2. Substructure Selection. Substructures are obtained by decomposing the training structures used for visualization of chemical space to generate structures within AD. One of the main structure decomposing methods is the retrosynthetic combinatorial analysis procedure (RECAP),[22] which obtains substructures by cleaving bonds derived from common chemical reactions. It is possible to consider the synthetic

Table 2. Additional Five Transformations Used in the Modified DAECS

transformation	description
1. addition of a building block	add a building block to the seed structure
2. removal of a building block	remove a building block from the seed structure
3. mutation of a building block	change a building block in the seed structure to another building block
4. insertion of a linker	connect two atoms to form a ring
5. removal of a linker	remove a bond of a ring

[a]In the transformations, a hydrogen atom is not included as an atom.

case of generated structures using substructures obtained with RECAP. However, RECAP has a drawback that new generated structures are often the same as already generated structures. A library constructed with de novo design needs diversity of structures. Therefore, we do not use RECAP. Instead, we cleave bonds between rings and chains to obtain substructures. We extracted rings from structures with the reference of Breadth-First Search.[23]

Although various substructures will be generated by decomposing structures, structures consisting of these substructures are sometimes difficult to synthesize. Structures must be filtered after generation. In this study, structures with hetero−hetero bonds were removed before addition to the structure pool.

When substructures are extracted from many training structures, there are usually too many substructures to handle. Therefore, it is necessary to select substructures used to transform seed structures. Expert knowledge can be applied to structure generation by selecting substructures.

2.2.2. Visualization of the Drug-likeness Distribution. Drugs need to not only have high activity but also suitable properties such as ADME and toxicity. The traditional DAECS only visualizes the activity distribution. To consider properties other than activity, the modified DAECS also visualizes the drug-likeness distribution. Studies on drug-likeness have been performed, and many drug-likeness discrimination models have been proposed.[24,25] Drug-likeness discrimination models that are statistically constructed using drug and nondrug databases can predict the drug-likeness of novel chemical structures. The modified DAECS visualizes the drug-likeness distribution using a drug-likeness discrimination model, which enables the user to set a target coordinate considering both the drug-likeness distribution and the activity distribution.

3. CASE STUDIES

To investigate the performance of the modified DAECS, we analyzed two data sets of ligands.

3.1. Data Sets. One of the data sets is a set of 635 ligands for the human adrenergic alpha2A receptor obtained from the GVK database (data set 1), and the other is a set of 522 ligands for the human histamine H1 receptor obtained from the ChEMBL database (data set 2). Compounds including P, Si, charges, and irregular bonds that do not obey the valency rule were excluded from both the data sets as is the case with the traditional DAECS.[9] These data sets contain the activity values, which are the inverse logarithm of the inhibition constant (K_i). Data set 1 was randomly divided into a training data set (300 compounds) and a test data set (335 compounds) to validate the prediction ability of the modified DAECS. For data set 2, all of the data were used to construct an activity prediction model

DOI: 10.1021/acs.jcim.6b00038
J. Chem. Inf. Model. 2016, 56, 1885−1893

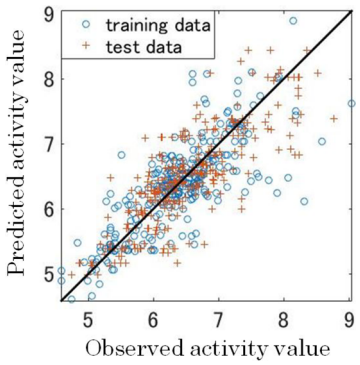


Figure 2. Observed and predicted activity values for data set 1. The activity values for the training data were predicted with 5-fold cross-validation.

and the activities of the generated structures were checked by docking simulations. The 3D conformations of these structures were calculated by Marvin Sketch[26] using geometry optimization based on the Dreiding force field.[27]

3.2. Visualization of Chemical Space. We used data sets 1 and 2 to visualize the chemical space and to construct activity prediction models using PubChem's fingerprints (Section 1-5, 460 bits)[28] as molecular descriptors. Descriptors that have the same value for all structures and descriptors with the same pattern as another descriptor were removed for each data set. We constructed a 2D map corresponding to the descriptors using the GTM proposed by Bishop[11] (see Appendix). Netlab toolbox[29] (a MATLAB[30] toolbox) was used for visualization with GTM.

To predict the activity distribution on a map, a prediction model was constructed using support vector regression,[31] which is a nonlinear regression method. The data point corresponding to a point on the map by weighted RBFs (see Formula A-1 in APPENDIX) is used as input of the prediction model for calculation of the activity of structures corresponding to the point, and activity distribution on the map is represented by calculation of the activity of structures corresponding to each grid on the map. The Gaussian kernel was used as the kernel function.

In DAECS, a 2D map is required to accurately represent the descriptor space around the training compounds and the activity of structures corresponding to each grid on the map. The values of the four hyperparameters of GTM should be selected to provide high accuracy mapping and high prediction ability of an activity prediction model. To select the optimal values of the hyperparameters, we used 5-fold cross-validation (AC_{cv}) to evaluate the accuracy of the mapping

$$AC_{cv} = \frac{\text{Number of coinciding bits between } \mathbf{X} \text{ and } \mathbf{X}'}{\text{Number of bits in } \mathbf{X}} \quad (2)$$

where $\mathbf{X}$ is the original data points. Here, $\mathbf{X}$ is projected to coordinates on a map and compared with the data points $\mathbf{X}'$ corresponding to the projected coordinates on the map by weighted RBFs. The calculated r^2 value with 5-fold cross-validation (r^2_{cv}) used to evaluate the prediction ability is

$$r^2_{cv} = 1 - \frac{\sum_{i=1}^{n} (y_{obs,i} - y_{pred,i})^2}{\sum_{i=1}^{n} (y_{obs,i} - y_{mean})^2} \quad (3)$$

where y_{obs} is the observed y value, y_{pred} is the predicted y value, y_{mean} is the mean of y_{obs}, and n is the number of training data. In this study, we used the number of radial base functions (RBFs) M = 16, 36, 64, 100, 144, 196, 256, 324 and 400, grid resolution K = 20 × 20, 25 × 25, and 30 × 30, RBF width σ = 0.0625, 0.125, 0.250, and 0.500, and weight regularization coefficient λ = 0.0010, 0.10, and 10.0. A reasonable compromise between AC_{cv} and r^2_{cv} was M = 256, K = 20 × 20, σ = 0.0625, and λ = 0.0010 for data set 1 and M = 64, K = 30 × 30, σ = 0.0648, and λ = 10.0 for data set 2. These parameters gave AC_{cv} = 0.954 and r^2_{cv} = 0.651 for data set 1 and AC_{cv} = 0.956 and r^2_{cv} = 0.337 for data set 2. Mapping and prediction on the map of data set 1 were validated using test data. For the test data, the accuracy of the mapping test data was AC_{test} = 0.954, and the activity prediction ability for the test data was r^2_{test} = 0.684. The relationship between the observed and predicted activity values for data set 1 is shown in Figure 2. The mapping accuracy and the activity prediction ability for the test data are similar to those by cross-validation, and the mapping and prediction are applicable to structures other than the training data.

We visualized the drug-likeness distribution on the maps. To predict drug-likeness, we used a data set of structures extracted from the comprehensive medicinal chemistry (CMC)[32] and available chemicals directory (ACD).[33] CMC is a database composed of compounds launched as drugs, and ACD is a database including commercially available compounds. To construct a drug-likeness discrimination model, the CMC database was used as a drug data set and the ACD database was used as a nondrug data set. However, ACD is not a nondrug database because it contains some drug compounds. We removed compounds in CMC from the ACD database to reduce the number of drug compounds in ACD and removed compounds with atoms other than C, H, N, O, S, P, and halogen atoms, compounds with special atoms, and compounds duplicated in each database from the CMC and ACD databases. The training data set for the drug-likeness discrimination model consisted of 1000 compounds from CMC and 1000 compounds from ACD, and the test data set consisted of another 1000 compounds from CMC and 1000 compounds from ACD. To consider the AD,[12] these 4000 compounds were selected to be similar to the structures in data sets 1 and 2. The structures were described by PubChem's fingerprints (Section 1-5, 460 bits).[28] The drug-likeness discrimination model was constructed using the support vector machine (SVM),[34] which is a nonlinear classification method. In the SVM, the Gaussian kernel was used as the kernel function. The results of discrimination for drug-likeness of the test data are shown in Table 3. The accuracy of discrimination for data set 1 is higher than that for data set 2. That is, molecules that are similar to

Table 3. Results of Discrimination for Drug-likeness of Test Data

	TP[a]	TN[b]	FP[c]	FN[d]	accuracy	precision	sensitivity
data set 1	891	883	117	109	0.8870	0.8839	0.8910
data set 2	842	797	203	158	0.8105	0.8057	0.8420

[a]True positive. [b]True negative. [c]False positive. [d]False negative.

DOI: 10.1021/acs.jcim.6b00038
J. Chem. Inf. Model. 2016, 56, 1885−1893

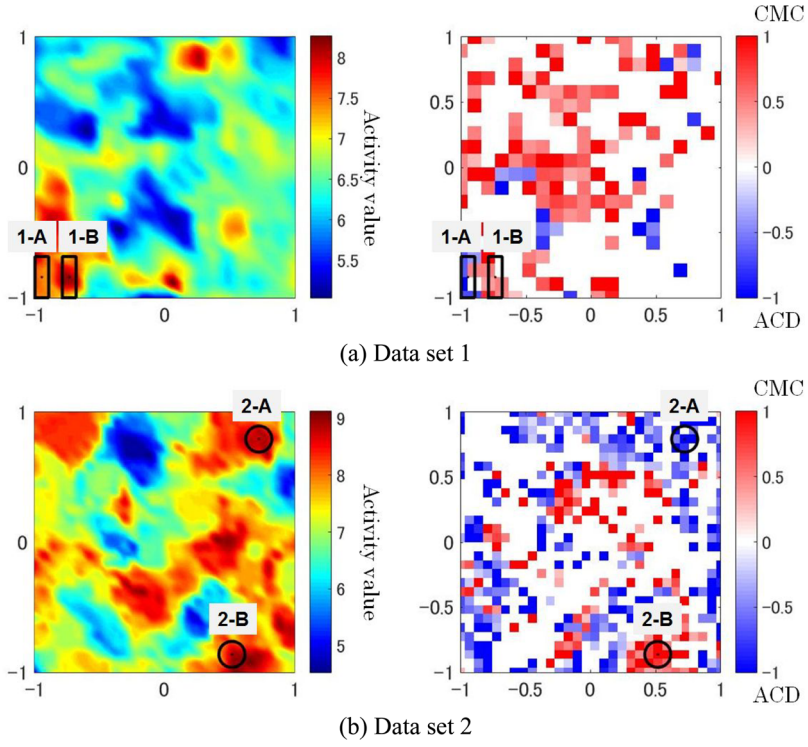


Figure 3. Activity distribution (left), drug-likeness distribution (right), and selected target areas on the constructed maps. 1-A and 2-A are the areas of nondrug-like compounds, and 1-B and 2-B are the areas of drug-like compounds.

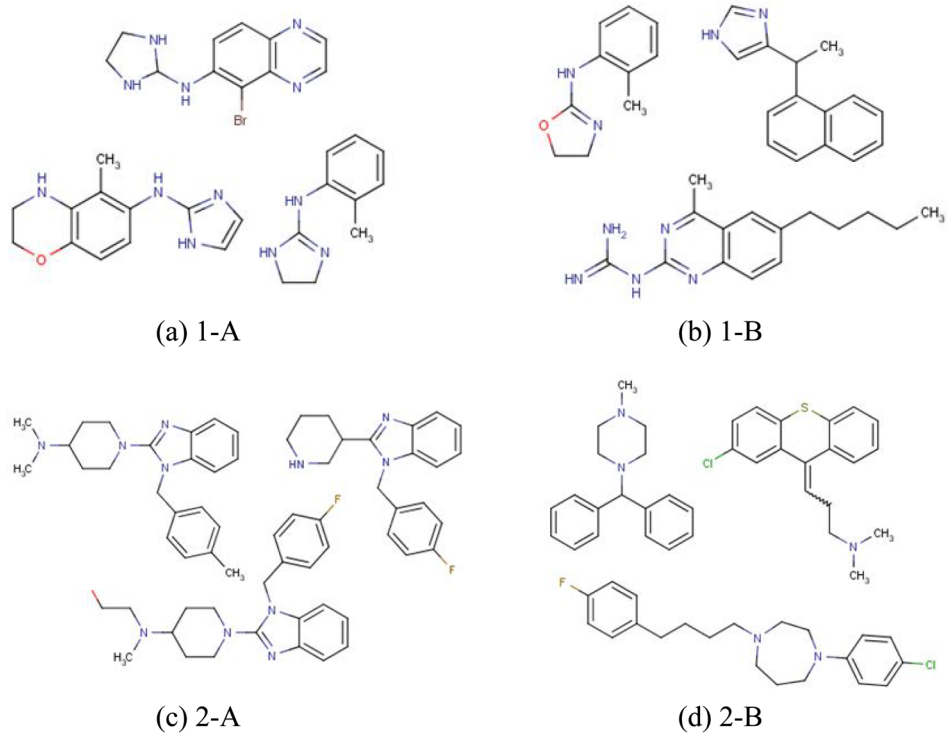
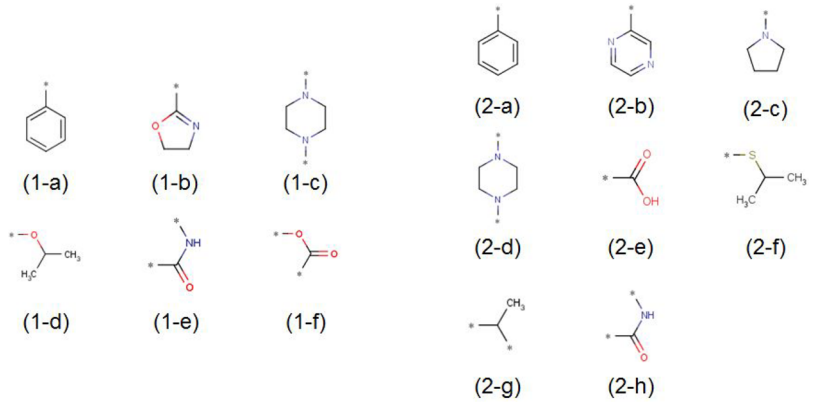


Figure 4. Initial seed structures for each target area.

molecules in data set 1 are easier to discriminate than molecules similar to those in data set 2.

3.3. Selection of Target Area. To compare the structure generation performance of the traditional DAECS with that of the modified DAECS, we selected four target areas on the constructed maps. The activity distributions, drug-likeness distributions, and selected target areas on the constructed maps are shown in Figure 3. In the drug-likeness distributions, red represents compounds projected onto the grid that are drug-like, and blue represents compounds projected onto the grid that are nondrug-like. The depth of the color on each grid represents the reliability of the discrimination, and it

DOI: 10.1021/acs.jcim.6b00038
J. Chem. Inf. Model. 2016, 56, 1885−1893



(a) Data set 1 (b) Data set 2

Figure 5. Selected substructures for (a) data set 1 and (b) data set 2.

Table 4. Results of Structure Generation

target area	use of substructures	average similarity	drug-like ratio	molecular weight	
				max	min
1-A	no	0.8308	0.2509	310.2	172.2
	yes	0.7830	0.3687	370.3	159.2
1-B	no	0.7410	0.9778	310.4	146.2
	yes	0.7114	0.9854	360.5	107.2
2-A	no	0.8666	0.1458	556.4	294.4
	yes	0.8441	0.1078	557.6	294.4
2-B	no	0.8954	0.8620	436.8	279.4
	yes	0.8296	0.9107	456.1	246.4

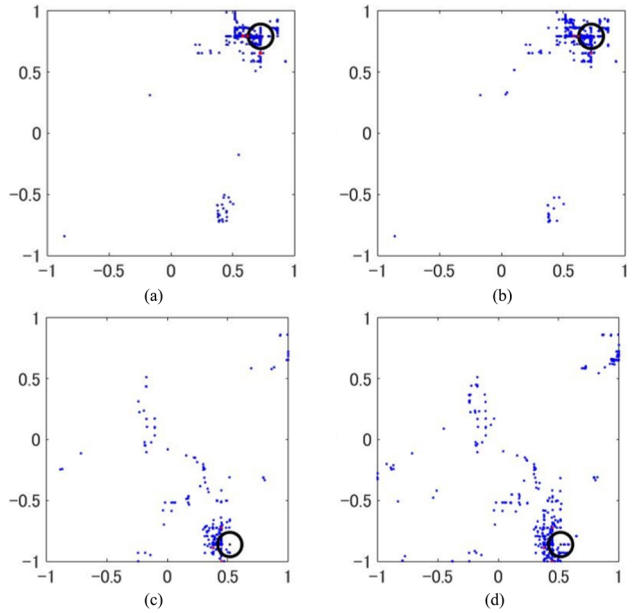


Figure 6. Coordinates of generated structures plotted on the map. Red points on each map are coordinates of initial seed structures. (a) Target area is set to 2A and substructures are not used. (b) Target area is set to 2A and substructures are used. (c) Target area is set to 2B and substructures are not used. (d) Target area is set to 2B and substructures are used.

corresponds to the ratio of the number of structures whose drug-likeness is correctly predicted to the number of structures projected onto the grid. White regions are regions where the drug-likeness prediction is unreliable. We confirmed that high

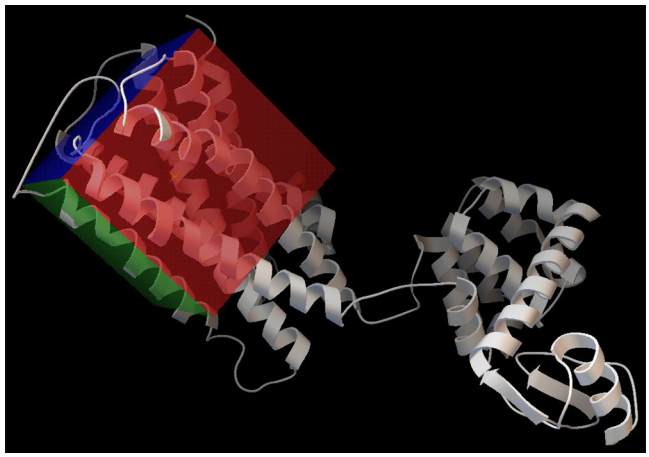


Figure 7. Structure of the histamine H1 receptor. The cube with red, green, and blue sides represents the search area.

activity compounds and drug-like compounds are projected on specific regions. In Figure 3, selected area 1-A is an area of nondrug-like compounds on the map of data set 1; 2-A is an area of nondrug-like compounds on the map of data set 2; 1-B is an area of drug-like compounds on the map of data set 1; 2-B an area of drug-like compounds on the map of data set 2. All of the target areas were areas of high activity compounds. For each area, the target coordinate was set on the center of the area, and three initial seed structures were selected from each data set. The selected initial seed structures are shown in Figure 4. These structures are projected around the target area.

3.4. Substructures. The substructures for structure generation were extracted from data sets 1 and 2. The selected substructures are shown in Figure 5. These substructures were selected because they are frequently found in each data set, and the structures that contain these substructures often have high activity. Substructures 1-a, 1-c, and 1-e are the same as substructures 2-a, 2-d, and 2-h, respectively. These substructures are frequently observed in both data sets 1 and 2. Substructure 1-b is a specific structure that is not frequently observed in data set 2.

3.5. Structure Generation. The number of seed structures in each cycle was set to two, and the cycles of generation were terminated when the number of generated structures in the target areas reached 10,000. The results are shown on Table 4.

DOI: 10.1021/acs.jcim.6b00038
J. Chem. Inf. Model. 2016, 56, 1885−1893

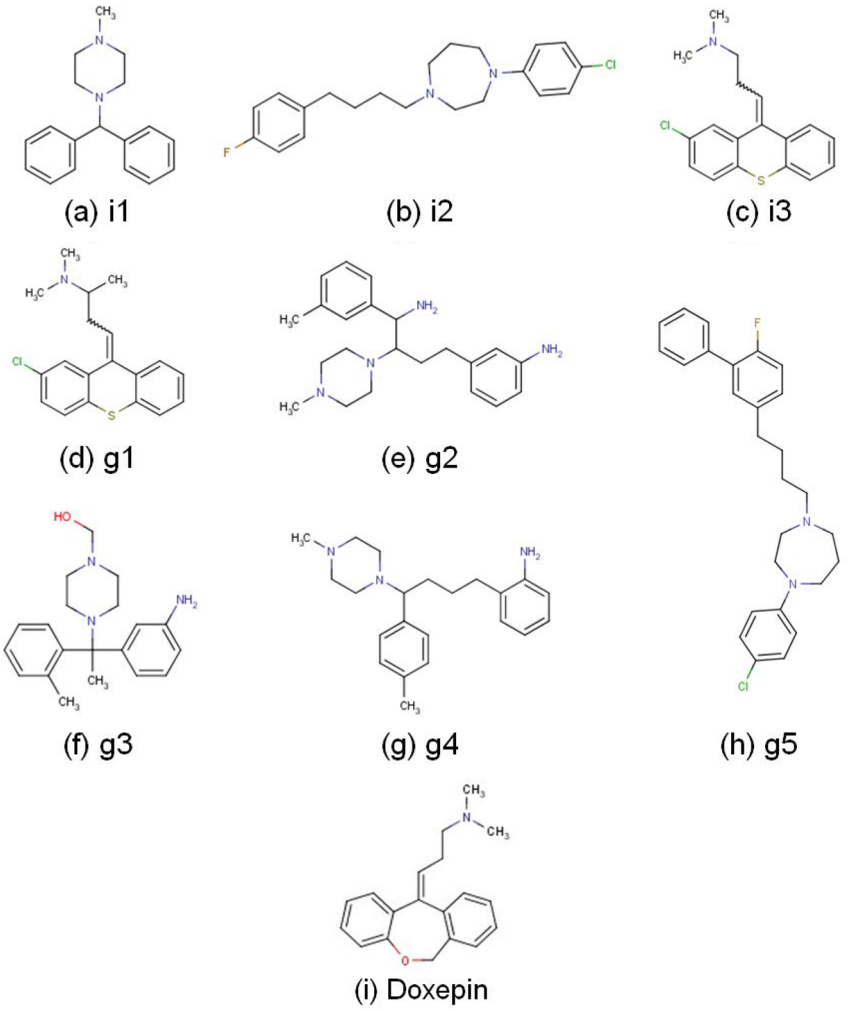


Figure 8. (a)−(c) Initial seed structures and (d)−(h) some of the generated structures. (i) Doxepin, a reference ligand structure.

The similarity between two structures is represented with the Tanimoto coefficient (TC). The TC is calculated with fingerprints between molecule a and molecule b as follows:

$$\mathrm{TC} = \frac{N_{ab}}{N_a + N_b - N_{ab}} \tag{4}$$

where N_a is the number of bits set to one in the fingerprint of a, N_b is the number of bits set to one in the fingerprint of b, and N_{ab} is the number of bits set to one in both fingerprints. TC has a value from 0 to 1, and structures are similar when the value is close to 1. In Table 4, the average similarity is the average of the similarity values between all possible combinations of generated structures in the target area. A lower similarity value mean higher diversity of generated structures. The drug-like ratio is the ratio of structures estimated to be drug-like to all of generated structures projected on the target area. Drug-likeness was estimated by constructed SVMs (Section 3.2).

When substructures were used for structure generation, the average similarity of the generated structures was lower and the range of molecular weight was wider than those without substructure transformation, which means that using substructures improves the diversity of the generated structures. When the target area was set as the area of drug-like compounds (target areas 1-B and 2-B), the ratio of drug-like structures was higher than that in the other case (target areas 1-A and 2-A). Therefore, selection of a target area considering the drug-likeness distribution affects the drug-likeness of the generated structures.

Generated structures in the case that the target area is set to 2-A or 2-B are plotted on the map as shown in Figure 6. Seed structures are selected from structures which are plotted in or near the target area at each cycle.

3.6. Docking Simulations. To evaluate the activity of the generated structures, we simulated docking of the generated structures to a protein. Docking simulations are an effective way to theoretically predict the activity. Structures with low calculated binding energies are expected to have high activity.

The free software AutoDock Vina,[35] which was designed by Oleg Trott, was used for the docking simulations, and the crystal structure of the histamine H1 receptor (PDB ID: 3RZE) downloaded from the RCSB PDB[36] was used as the protein. This contains doxepin data as a ligand. The search area of the docking simulations was set around the conformation of doxepin. The structure of the histamine H1 receptor and the search area are shown in Figure 7. However, it needs to be note that the accuracy of docking simulation with AutoDock Vina is low. We performed the simulation as a reference.

We verified the activity of the generated structures when the target area was set on 2-B, and substructures were used for structure transformation. Docking simulations were performed for the three initial seed structures and some of the generated

DOI: 10.1021/acs.jcim.6b00038
J. Chem. Inf. Model. 2016, 56, 1885−1893

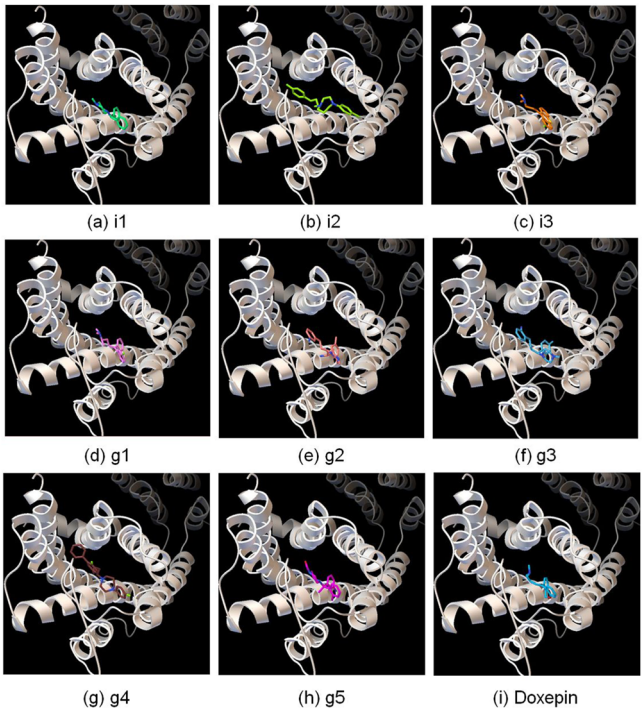


Figure 9. Simulated conformations of molecules docking with the histamine H1 receptor.

Table 5. Binding Energies from Docking Simulations

	structure	binding energy (kcal/mol)
initial seed structures	i1	−8.4
	i2	−8.0
	i3	−7.3
some generated structure	g1	−8.7
	g2	−8.2
	g3	−9.5
	g4	−8.4
	g5	−9.2
reference structure	doxepin	−9.4

structures, which are shown in Figure 8. Structures i1−i3 are the initial seed structures, and g1−g5 are the generated structures. The docking conformations are shown in Figure 9, and the calculated binding free energies are given in Table 5. There are generated structures with lower binding energies than the initial structures. Structure g3 is expected to have high activity and predicted to be drug-like. Thus, diverse structures that are expected to have high activity and be drug-like can be generated using the proposed method.

4. CONCLUSIONS

In this study, we modified DAECS by visualization of the drug-likeness distribution and structure transformation using substructures. Two data sets were used to validate the modified DAECS. A number of drug-like structures were generated when the target area was set on a high activity drug-like area or a high activity nondrug-like area, and the diversity of the generated structures was higher when substructures were used for structure transformation. However, improvement of 2D similarity is slight (0.07 at max and is 0.03 at minimum), and further modifications are desired, for example, seed selections considering diversity in each cycle.

Visualization of the distribution of a property required for drugs enables DAECS to generate structures considering this property. In this study, although the distribution of the activity and the drug-likeness is visualized, visualization of the distributions of other properties required for drugs enables DAECS to generate structures considering these properties.

When activity data to a target protein exists, the proposed method can be used to explore more desirable structures. For example, when a high-activity nondrug-like structure is found, the modified DAECS can generate many structures that are expected to have similar activity to the found structure and be drug-like.

The substructures for structure transformations need to be carefully selected. When substructures that are often included in drug compounds are selected, structures with drug-like features are likely to be generated. Conversely, when substructures that are often included in nondrug compounds are selected, structures with nondrug-like features are likely to be generated. Although substructures need be selected by the user in our proposed method, it is difficult to select optimal substructures. Therefore, a method to automatically select optimal substructures is required. Nevertheless, the modified DAECS is a promising approach to construct virtual libraries and discover new drugs.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00038.

Information as mentioned in the text. (PDF)
Structures. (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: funatsu@chemsys.t.u-tokyo.ac.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank CTC Laboratory Systems, Japan, and GVK Biosciences, India, for allowing us to use the data set. The authors acknowledge support from the Core Research for Evolutionary Science and Technology (CREST) project "Development of a knowledge-generating platform driven by big data in drug discovery through production processes" of the Japan Science and Technology Agency (JST).

■ REFERENCES

(1) Honma, T. Recent Advances in de Novo Design Strategy for Practical Lead Identification. *Med. Res. Rev.* **2003**, *23*, 606−632.
(2) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.
(3) Pirard, B. The Quest for Novel Chemical Matter and the Contribution of Computer-aided de Novo Design. *Expert Opin. Drug Discovery* **2011**, *6*, 225−231.
(4) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A Graph-based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1079−1087.
(5) Miyao, T.; Arakawa, M.; Funatsu, K. Exhaustive Structure Generation for Inverse-QSPR/QSAR. *Mol. Inf.* **2010**, *29*, 111−125.
(6) Hartenfellar, M.; Proschak, E.; Schüller, A.; Schneider, G. DOGS: Reaction-driven de Novo Design of Bioactive Compounds. *PLoS Comput. Bio.* **2012**, *8*, 1−12.

DOI: 10.1021/acs.jcim.6b00038
J. Chem. Inf. Model. 2016, 56, 1885−1893

(7) Bembenek, S. D.; Tounge, B. A.; Reynolds, C. H. Ligand Efficiency and Fragment-based Drug Discovery. *Drug Discovery Today* **2009**, *14*, 278−283.
(8) Dey, F.; Caflisch, A. Fragment-based de Novo Ligand Design by Multiobjective Evolutionary Optimization. *J. Chem. Inf. Model.* **2008**, *48*, 679−690.
(9) Mishima, K.; Kaneko, H.; Funatsu, K. Development of a New de Novo Design Algorithm for Exploring Chemical Space. *Mol. Inf.* **2014**, *33*, 779−789.
(10) Kohonen, T. The Self-organizing Map. *Proc. IEEE* **1990**, *78*, 1464−1480.
(11) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215−234.
(12) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inf.* **2015**, *34*, 348−356.
(13) Kaneko, H.; Funatsu, K. Applicability Domain Based on Ensemble Learning in Classification and Regression Analyses. *J. Chem. Inf. Model.* **2014**, *54*, 2469−2482.
(14) Van de Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.
(15) Li, A. P. Screening for Human ADME/Tox Drug Properties in Drug Discovery. *Drug Discovery Today* **2001**, *6*, 357−366.
(16) Lipinski, C. A. Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.
(17) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90−98.
(18) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 2048−2056.
(19) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1882−1889.
(20) Korkmaz, S.; Zararsiz, G.; Goksuluk, D. Drug/nondrug classification using support vector machines with various feature selection strategies. *Comp. Meth. Prog. Biomed.* **2014**, *117*, 51−60.
(21) Arakawa, M.; Miyao, T.; Funatsu, K. Development of Drug-likeness Model and Its Visualization. *J. Comput. Aided Chem.* **2008**, *9*, 70−80.
(22) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comp. Sci.* **1998**, *38*, 511−522.
(23) Figueras, J. Ring perception using breadth-first search. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 986−991.
(24) Li, Q.; Bender, A.; Pei, J.; Lai, L. A Large Descriptor Set and a Probabilistic Kernel-based Classifier Significantly Improve Druglikeness Classification. *J. Chem. Inf. Model.* **2007**, *47*, 1776−1786.
(25) Wagener, M.; van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comp. Sci.* **2000**, *40*, 280−292.
(26) ChemAxon Marvin 5.1.3_2 program, November 13, 2008. www.chemaxon.com/products.html (accessed September 2016).
(27) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: a Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897−8909.
(28) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a Public Information System for Analyzing Bioactivities of Small Molecules. *Nucleic Acids Res.* **2009**, *37*, W623−W633.
(29) Nabney, I. *NETLAB: Algorithms for Pattern Recognition*; Springer Science & Business Media: London, 2002.
(30) MATLAB R2015a, February 12, 2015. www.mathworks.com/ (accessed September 2016).
(31) Li, H.; Liang, Y.; Xu, Q. Support Vector Machines and Its Applications in Chemistry. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 188−198.
(32) *Comprehensive Medicinal Chemistry* is available from Dassault Systemes Biovia K.K., ThinkPark Tower 21F, 2-1-1 Osaki, Shinagawa-ku, Tokyo 141-6020.
(33) *Available Chemicals Directory* is available from Dassault Systemes Biovia K.K., ThinkPark Tower 21F, 2-1-1 Osaki, Shinagawa-ku, Tokyo 141-6020.
(34) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273−297.
(35) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.
(36) RCSB Protein Data Bank. http://www.rcsb.org/pdb/home/home.do (accessed September 2016).

DOI: 10.1021/acs.jcim.6b00038
J. Chem. Inf. Model. 2016, 56, 1885−1893