ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

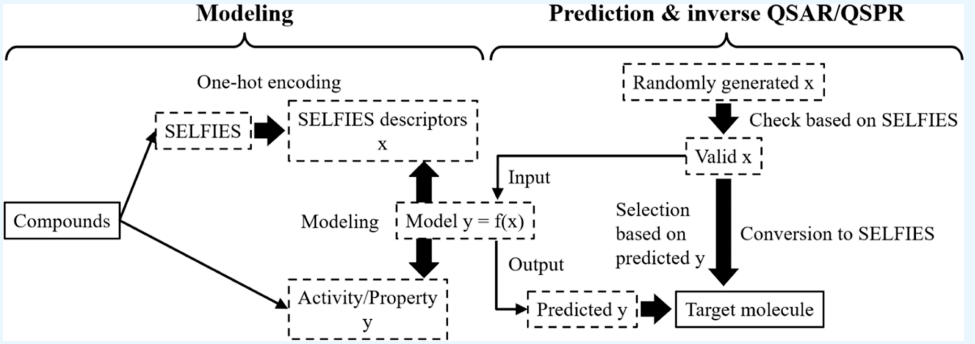
Molecular Descriptors, Structure Generation, and Inverse QSAR/QSPR Based on SELFIES

Hiromasa Kaneko*

Cite This: *ACS Omega* 2023, 8, 21781−21786

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: For inverse QSAR/QSPR in conventional molecular design, several chemical structures must be generated and their molecular descriptors must be calculated. However, there is no one-to-one correspondence between the generated chemical structures and molecular descriptors. In this paper, molecular descriptors, structure generation, and inverse QSAR/QSPR based on self-referencing embedded strings (SELFIES), a 100% robust molecular string representation, are proposed. A one-hot vector is converted from SELFIES to SELFIES descriptors x, and an inverse analysis of the QSAR/QSPR model $y = f(x)$ with the objective variable y and molecular descriptor x is conducted. Thus, x values that achieve a target y value are obtained. Based on these values, SELFIES strings or molecules are generated, meaning that inverse QSAR/QSPR is performed successfully. The SELFIES descriptors and SELFIES-based structure generation are verified using datasets of actual compounds. The successful construction of SELFIES-descriptor-based QSAR/QSPR models with predictive abilities comparable to those of models based on other fingerprints is confirmed. A large number of molecules with one-to-one relationships with the values of the SELFIES descriptors are generated. Furthermore, as a case study of inverse QSAR/QSPR, molecules with target y values are generated successfully. The Python code for the proposed method is available at https://github.com/hkaneko1985/dcekit.

1. INTRODUCTION

To develop compounds with specific activities and properties, artificial intelligence and machine learning are commonly applied to design molecules or chemical structures. A mathematical model $y = f(x)$ with the objective variable y and molecular descriptor x was constructed using a dataset of compounds with measured activities and properties. Here, the objective variable y represents activities and properties, and the molecular descriptor x represents the structural features of molecules. The model can predict the y values for new molecules based on an input of the x values of the chemical structures of the molecules. Furthermore, the x values at which y values have the desired values can be predicted by conducting an inverse analysis of the model. By generating numerous chemical structures or molecules, predicting y from x, and selecting molecules based on the predicted y values, molecules with the desired y values can be designed.

In molecular design, constructing a model with high predictive ability, where the x values appropriately represent the characteristics of the chemical structures, is crucial. To this end, RDKit,[1] Mordreds,[2] MOE,[3] and AlvaDesc[4] are used to calculate x. In addition, fingerprints such as extended-connectivity fingerprints (ECFPs),[5] functional connectivity fingerprints (FCFPs),[6] and MACCS keys[7] are used to quantify chemical structures and are considered as x.

Various structure generators have been developed to generate chemical structures as inputs for a model that is constructed between x and y. Further information on the structure of the generators can be found in refs 8−91011. In addition, inverse analysis of such a model and inverse QSAR/QSPR are effective

Received: February 27, 2023
Accepted: May 29, 2023
Published: June 5, 2023

https://doi.org/10.1021/acsomega.3c01332
ACS Omega 2023, 8, 21781−21786

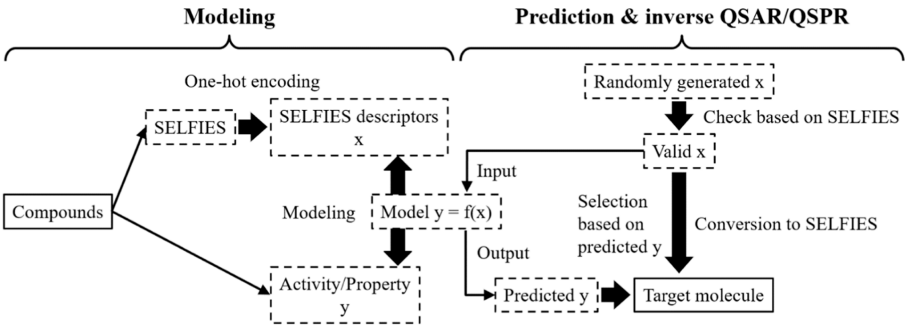


Figure 1. Schematic of the proposed method.

for molecular design with target y values.[12−14] To design molecules with target y values, it is necessary not only to find the x values corresponding to each y value but also to ensure a one-to-one relationship between the x values and the chemical structures. Even if a y value predicted from an x value is promising, it is futile if the chemical structure corresponding to the x value cannot be generated.

In this study, based on self-referencing embedded strings (SELFIES),[15] a 100% robust molecular string representation, SELFIES-based molecular descriptors, structure generation, and inverse QSAR/QSPR are proposed. The SELFIES can be converted into a one-hot vector, which is a data representation method that converts a categorical variable into a numeric vector. Specifically, for each element in a category, a vector is created, where the length of the vector represents the number of elements in that category; the value of the corresponding element position is set to one, and the values of all other positions are set to zero. The one-hot vector converted from the SELFIES is proposed as a SELFIES descriptor and is used as x. By appropriately selecting the vectors from randomly generated one-hot vectors, a one-to-one relationship between the SELFIES, and the one-hot vector can be established. Furthermore, owing to the characteristics of the SELFIES, chemical structures can be generated from all the SELFIES, and the relationship between the chemical structures and the SELFIES is one-to-one. Therefore, the SELFIES, the one-hot vector, and the chemical structure exhibit a one-to-one relationship. This implies that the x values determined by the inverse analysis of the model correspond to chemical structures.

Datasets comprising the boiling point (BP), aqueous solubility, and toxicity of compounds were used to generate chemical structures from the one-hot vector of the SELFIES and to test the predictive performance of the model by setting the SELFIES descriptors as x. In addition, a case study of inverse QSAR/QSPR was performed to generate molecules such that the target y values were attained.

2. METHOD

2.1. SELFIES. SELFIES is a string-based molecular graph representation that is 100% robust, even when randomly generated. Each set of SELFIES corresponds to a valid molecule, and it is possible to convert all molecules to SELFIES. Information on the length and size of the ring is stored with the corresponding identifiers, namely "Branch" and "Ring." The symbols after Branch and Ring represent numbers that are interpreted as lengths, preventing the possibility of strings with an invalid syntax. In addition, each molecule is represented by one-hot encoding in SELFIES.

2.2. Molecular Descriptors, Structure Generation, and Inverse QSAR/QSPR Based on SELFIES. A one-hot vector encoding SELFIES was used as the SELFIES descriptor. The one-hot vector contains information on the presence or absence and position of SELFIES symbols, and the SELFIES descriptors represent the features of the chemical structures of the molecules. Furthermore, because the one-hot vector can be directly converted into SELFIES, that is, chemical structures or molecules, molecules can be generated directly from the values of the SELFIES descriptors, resulting from the inverse analysis of the model $y = f(x)$ using the SELFIES descriptors x.

Molecular generation can be achieved by randomly generating zero or unit vectors as SELFIES descriptors (one-hot vectors) and converting them into SELFIES as follows:

- A: A randomly generated set of SELFIES descriptors.
- B: Sets of SELFIES descriptors converted from the SELFIES from A.

However, A and B do not always match.

If A and B do not match, generating molecules from the x values resulting from the inverse analysis of the model $y = f(x)$ using the SELFIES descriptors as x does not make sense. Therefore, the identities of A and B are checked, and if A and B do not match, the generated set of SELFIES descriptors is deleted.

Figure 1 shows the inverse QSAR/QSPR based on the SELFIES descriptors. First, the chemical structures of the compounds are represented by SELFIES, and the SELFIES descriptors are calculated by one-hot encoding and denoted as x. Activities and properties are denoted as y, and the model $y = f(x)$ is constructed between x and y based on a dataset of compounds.

Each 0 or 1 vector randomly generated as the SELFIES descriptor is converted to SELFIES; the SELFIES are then converted to SELFIES descriptors, their identity is checked, and if they are different, the vector is deleted. Thus, only valid sets of the SELFIES descriptor values are obtained. These sets are then input into the $y = f(x)$ model to predict the y values. Only sets of SELFIES descriptor values, that is, molecules that have promising y values and predicted y values close to the target y values, are selected. This allows the design of molecules that achieve the target y values. Additionally, the predicted y values and their variance can be considered when selecting the sets of SELFIES descriptor values or molecules. Alternatively, the selection can be based on the acquisition function in Bayesian optimization.[16]

The greater the number of 0 or 1 vectors generated in the inverse QSAR/QSPR process, the greater the amount of time and memory required. However, instead of generating a large number of vectors simultaneously, it is possible to iteratively conduct the process of generating vectors and performing

https://doi.org/10.1021/acsomega.3c01332
ACS Omega 2023, 8, 21781−21786

predictions and inverse QSAR/QSPR, which requires only a small amount of memory.

Python codes for the proposed method are available in ref 17.

3. RESULTS AND DISCUSSION

Datasets of the BP,[18] solubility in water (log *S*, *S* = solubility at 20−25 °C in moles per liter),[19] and environmental toxicity (Tox)[20] were used to verify the effectiveness of the proposed method. The Tox dataset was sourced from an online challenge inviting researchers to estimate the toxicity of molecules against *Tetrahymena pyriformis* and contained entries corresponding to the logarithm of the 50% growth inhibitory concentration.

First, chemical structures or molecules were generated based on the SELFIES data obtained for each dataset. The number of one-hot vectors or sets of SELFIES descriptor values to be generated was set to 10,000. Molecules were generated with the number of SELFIES symbols as the maximum number of SELFIES symbols in each dataset, 250, 500, 750, 1000, 2500, 5000, 7500, and 10,000, and the duplicated molecules were removed. The number of molecules generated is listed in Table 1. For each dataset, approximately half of the generated

Table 1. Number of Valid and Unique Generated Molecules Based on SELFIES for Each Number of SELFIES Symbols Used for Each Dataset

	BP	log *S*	Tox
max	3687	4685	4279
250	4199	4831	4380
500	4364	4948	4598
750	4418	4872	4543
1000	4439	4894	4674
2500	4505	5038	4641
5000	4564	4896	4543
7500	4535	4929	4696
10,000	4552	4937	4696

SELFIES were not identical to the original one-hot vector when re-converted to the one-hot vector; however, approximately 40−50% of the SELFIES or chemical structures generated successfully were valid and unique. Hence, the results confirm that the proposed structure-generation method based on SELFIES can be used to effectively generate molecules.

Table 1 shows that increasing the number of SELFIES symbols does not significantly change the number of molecules generated. Considering *x* and the inverse QSAR/QSPR process, a large number of SELFIES symbols indicates a large number of descriptors. This reduces the prediction performance of the QSAR/QSPR models. Therefore, it is preferable not to increase the number of SELFIES symbols and use the maximum number of actual SELFIES symbols in the training data.

Second, to test the performance of the SELFIES descriptors, the predictive ability of the regression model based on the SELFIES descriptors *x* was compared with that of regression models based on *x* values calculated with the RDKit, ECFP, FCFP, and MACCS keys. For each dataset, 70% of the compounds were randomly selected as training data, and the remaining 30% were used as test data. Subsequently, regression models were constructed with the training data to predict the test data. The following regression methods were used:

- Ordinary least squares regression (OLS).
- Partial least squares regression (PLS).
- Ridge regression (RR).
- Least absolute shrinkage and selection operator (LASSO).
- Elastic net (EN).
- Support vector regression with a linear kernel (SVRL).
- Support vector regression with a Gaussian kernel (SVRG).
- Decision tree (DT).
- Random forests (RF).
- Gaussian process regression (GPR).
- Gradient boosting decision tree (GBDT).
- XGBoost (XGB).
- LightGBM (LGB).
- Gaussian mixture regression (GMR).
- Variational Bayesian Gaussian mixture regression (VBGMR).
- Deep neural networks (DNN).

For each set of *x* values, the regression method that could construct the model with the highest predictive ability for the test data was used.

The R^2 and root-mean-squared error (RMSE) of the test data for each set of *x* values are listed in Tables 2, 3, and 4 for BP, log *S*, and Tox, respectively. The plots of the actual and estimated *y* values for the test data are shown in Figures 2, 3, and 4 for BP, log *S*, and Tox, respectively. The prediction results show that the RDKit descriptors, which are continuous, have a higher R^2 and lower RMSE than the other fingerprint-type descriptor sets. Furthermore, the RDKit descriptors imparted a high predictive ability to the regression models on all datasets. Because *y* has continuous values, continuous RDKit descriptors provided the best prediction results in the regression analyses.

Table 2. R^2 and RMSE for Each Descriptor Set in the BP Test Data

	R^2	RMSE
RDKit (GPR)	0.98	10
ECFP (XGB)	0.74	37
FCFP (GBDT)	0.63	45
MACCS keys (EN)	0.81	32
SELFIES (GPR)	0.76	36

Table 3. R^2 and RMSE for Each Descriptor Set in the Log *S* Test Data

	R^2	RMSE
RDKit (GPR)	0.94	0.51
ECFP (LGB)	0.76	1.04
FCFP (GBDT)	0.76	1.04
MACCS keys (GPR)	0.82	0.90
SELFIES (GPR)	0.78	0.99

Table 4. R^2 and RMSE for Each Descriptor Set in the Tox Test Data

	R^2	RMSE
RDKit (GPR)	0.85	0.39
ECFP (GBDT)	0.65	0.60
FCFP (RF)	0.65	0.59
MACCS keys (XGB)	0.72	0.54
SELFIES (GPR)	0.62	0.62

A comparison between the fingerprints demonstrates that the SELFIES descriptors exhibit almost the same prediction

https://doi.org/10.1021/acsomega.3c01332
ACS Omega 2023, 8, 21781−21786

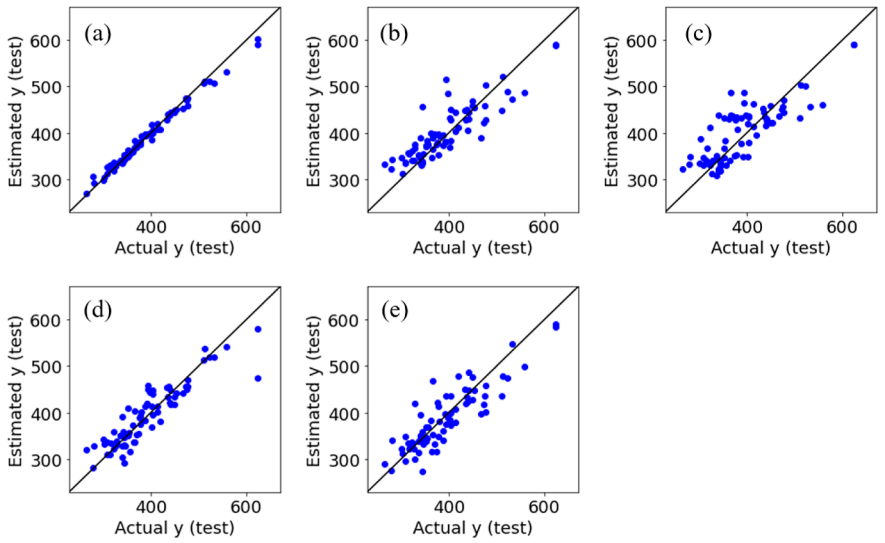


Figure 2. Actual *y* vs estimated *y* for the test BP data. (a) RDKit (GPR), (b) ECFP (XGB), (c) FCFP (GBDT), (d) MACCS keys (EN), and (e) SELFIES (GPR).

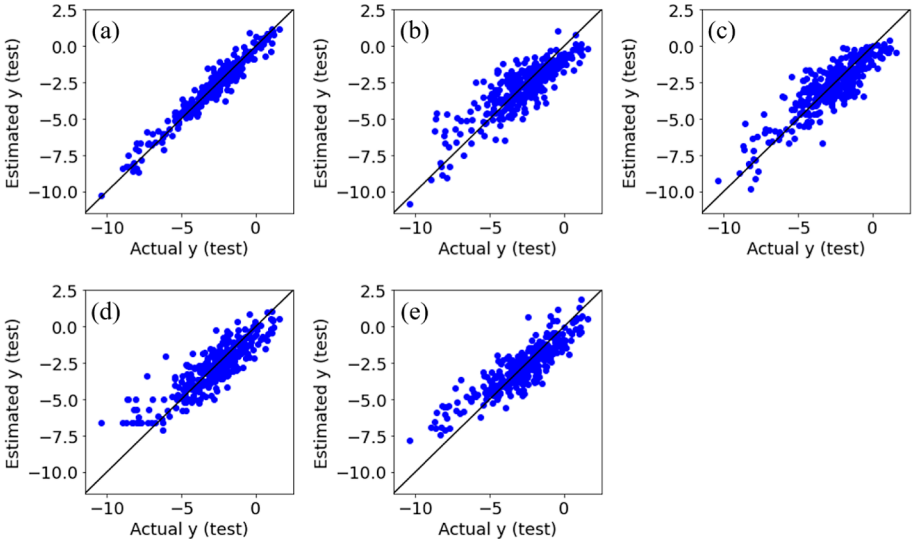


Figure 3. Actual y vs estimated *y* for the test log *S* data. (a) RDKit (GPR), (b) ECFP (LGB), (c) FCFP (GBDT), (d) MACCS keys (GPR), and (e) SELFIES (GPR).

performance as the other fingerprints. However, unlike the other fingerprints, the SELFIES descriptors are interpretable because the values 0 and 1 are directly related to the chemical structures. Moreover, they can be directly transformed into chemical structures, making them superior to the other fingerprints. For example, the values of the SELFIES descriptors with promising y values can be used to generate the corresponding molecules, as predicted by the regression model.

Finally, a regression model was constructed between the SELFIES descriptors and *y*, and inverse QSAR/QSPR was performed based on the constructed model. Using the BP dataset, the SELFIES descriptors were calculated as *x*, and the molecular weight (MW), quantitative estimate of drug-likeness (QED), and log *P* calculated by RDKit[1] from SMILES were used as *y*. QED is a quantitative representation of whether a structure is drug-like. QED values close to 0 indicate that the chemical structures are not drug-like, whereas QED values close to 1 indicate drug-like chemical structures. Log *P* is the logarithmically transformed octanol−water partition coefficient. It is related to oral absorption and is an important index for drug development.

GPR[21] was used to calculate not only the predicted *y* values but also their variance and prediction reliability. A total of 70% of the compounds were randomly selected as training data, and the remaining 30% were used as test data. The regression models were constructed using the training data to predict the test data. Plots of the actual and estimated *y* values for the test data are shown in Figure 5. The plots show that each molecule is close to the diagonal line, and the *y* values are accurately predicted from the SELFIES descriptors MW, QED, and log *P*.

Subsequently, 100,000 one-hot vectors or sets of SELFIES descriptors were generated as inputs for the constructed GPR model. SELFIES were generated based on these vectors, and 26019 valid and unique molecules were generated. These molecules were input into the GPR model to predict the *y* values. Molecules with low variance in the predictions were selected, and the actual *y* was calculated. The plots of the predicted *y* values and target *y* values vs the actual *y* values are shown in

https://doi.org/10.1021/acsomega.3c01332
ACS Omega 2023, 8, 21781−21786

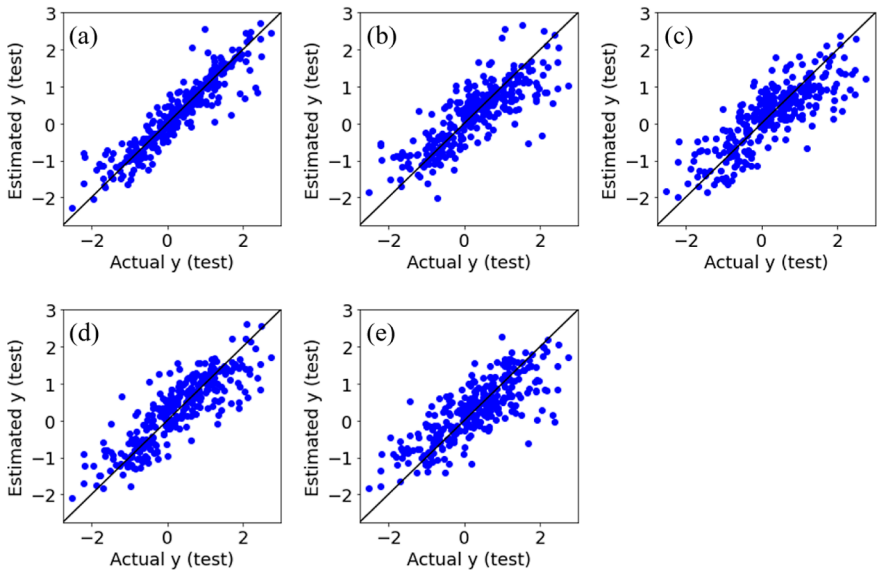


Figure 4. Actual y vs estimated y for the test Tox data. (a) RDKit (GPR), (b) ECFP (GBDT), (c) FCFP (RF), (d) MACCS keys (XGB), and (e) SELFIES (GPR).

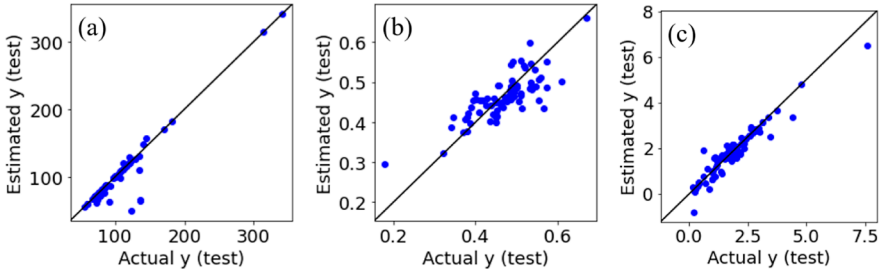


Figure 5. Actual y vs estimated y for the test data for MW, QED, and log P. (a) MW (GPR), (b) QED (GPR), and (c) log P (GPR).

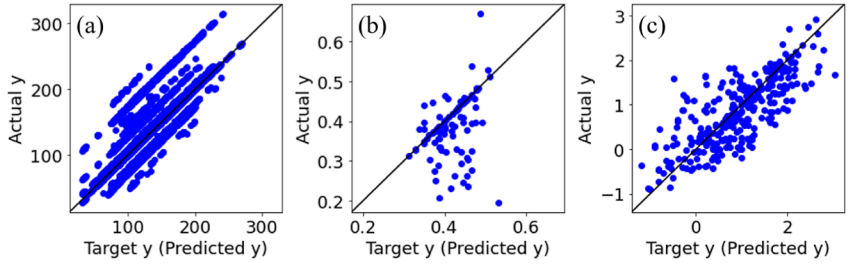


Figure 6. Target y (predicted y) vs actual y for the inverse analysis with SELFIES. (a) MW, (b) QED, and (c) log P.

Figure 6. The molecules can be generated such that the actual y falls within the range of the target y for MW, QED, and log P. In contrast with the cases of the MW and log P, the molecules are far from the diagonal line in the case of QED because of the lower predictive ability of the QED prediction model compared with the MW and log P prediction models, as shown in Figure 5. However, it was confirmed that even if y is obtained with QED, generating molecules close to the diagonal is possible. Furthermore, inverse QSAR/QSPR can be appropriately performed using the proposed method.

4. CONCLUSIONS

In this paper, SELFIES-based molecular descriptors, chemical structure generators, and inverse QSAR/QSPR methods are proposed. The SELFIES descriptor x, which is the one-hot vector of SELFIES, does not have one-to-one correspondence with chemical structures. Thus, chemical structures or molecules can be generated from the x values obtained by inverse analysis of the regression model $y = f(x)$ constructed between x and objective variables y, such as activities and properties. Hence, chemical structure generation, where y has the target values (i.e., inverse QSAR/QSPR), is possible. The proposed method was validated using compound datasets of BP, log S, and toxicity. It was confirmed that important chemical structures could be generated with the SELFIES descriptors or the one-hot vectors of SELFIES. By setting the SELFIES descriptors as x, regression models with the same prediction performance as that of models based on other fingerprints were constructed. Furthermore, a case study of inverse QSAR/QSPR was conducted. The results demonstrate that the proposed method can be used to generate molecules with MW, QED, and log P target values. It is expected

https://doi.org/10.1021/acsomega.3c01332
ACS Omega 2023, 8, 21781−21786

that inverse QSAR/QSPR based on the proposed SELFIES descriptors will improve the efficiency of molecular design.

Python codes for the proposed method are available at https://github.com/hkaneko1985/dcekit.

■ ASSOCIATED CONTENT

Data Availability Statement

Data supporting the findings of this study are available in ref 17.

■ AUTHOR INFORMATION

Corresponding Author

Hiromasa Kaneko − *Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan;* orcid.org/0000-0001-8367-6476; Phone: +81-44-934-7197; Email: hkaneko@meiji.ac.jp

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c01332

Notes

The author declares no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by a Grant-in-Aid for Scientific Research (KAKENHI) (grant number 19K15352) from the Japan Society for the Promotion of Science.

■ REFERENCES

(1) RDKit: Open-Source Cheminformatics Software. https://rdkit.org (accessed Nov 24, 2022).
(2) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018**, *10*, 4.
(3) *MOE (Molecular Operating Environment), 2008.10*; Chemical Computing Group, Inc.: Montreal, Canada, 2008.
(4) Alvascience. alvaDesc. https://www.alvascience.com/alvadesc/ (accessed Nov 24, 2022).
(5) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(6) Finn, E.; Shen, X.; Scheinost, D.; Rosenberg, M. D.; Huang, J.; Chun, M. M.; Papademetris, X.; Constable, R. T. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **2015**, *18*, 1664−1671.
(7) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(8) Schneider, G. *De novo Molecular Design*; Wiley, 2013.
(9) Lin, X.; Li, X.; Lin, X. A Review on Applications of computational methods in drug screening and design. *Molecules* **2020**, *25*, 1375.
(10) Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expet Opin. Drug Discov.* **2021**, *16*, 949−959.
(11) Wang, M.; Wang, Z.; Sun, H.; Wang, J.; Shen, C.; Weng, G.; Chai, X.; Li, H.; Cao, D.; Hou, T. Deep learning approaches for de novo drug design: An overview. *Curr. Opin. Struct. Biol.* **2022**, *72*, 135−144.
(12) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(13) Gantzer, P.; Creton, B.; Nieto-Draghi, C. Inverse-QSPR for de novo design: A review. *Mol. Inf.* **2020**, *39*, 1900087.
(14) Gebauer, N. W. A.; Gastegger, M.; Hessmann, S. S. P.; et al. Inverse design of 3d molecular structures with conditional generative neural networks. *Nat. Commun.* **2022**, *13*, 973.
(15) Krenn, M.; Häse, F.; Nigam, A. K.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.
(16) Ando, T.; Shimizu, N.; Yamamoto, N.; Matsuzawa, N.; Maeshima, H.; Kaneko, H. Design of molecules with low hole and electron reorganization energy using DFT calculations and Bayesian optimization. *J. Phys. Chem. A* **2022**, *126*, 6336−6347.
(17) DCEKit. hkaneko1985/dcekit. https://github.com/hkaneko1985/dcekit (accessed Nov 24, 2022).
(18) Hall, L. H.; Story, C. T. Boiling point and critical temperature of a heterogeneous data set: qsar with atom type electrotopological state indices using artificial neural networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004−1014.
(19) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.
(20) ICANN'09. Environmental Toxicity Prediction Challenge. http://www.cadaster.eu/node/65.html (accessed Nov 24, 2022).
(21) scikit-learn. sklearn.gaussian_process.GaussianProcessRegressor. https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html (accessed Nov 24, 2022).

https://doi.org/10.1021/acsomega.3c01332
ACS Omega 2023, 8, 21781−21786