ACS OMEGA

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

Article

Molecular Design Learned from the Natural Product Porphyra-334: Molecular Generation via Chemical Variational Autoencoder versus Database Mining via Similarity Search, A Comparative Study

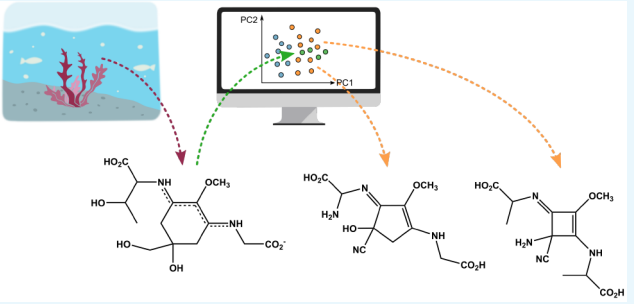
Yuki Harada, Makoto Hatakeyama, Shuichi Maeda, Qi Gao, Kenichi Koizumi, Yuki Sakamoto, Yuuki Ono, and Shinichiro Nakamura*

Cite This: *ACS Omega* 2022, 7, 8581−8590

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A comparative study is presented. The method via chemical variational autoencoder (VAE) and the method via similarity search are compared, focusing on their generation ability for new functional molecular design. Focusing on the natural porphyra-334 as a model molecule, we generated three groups: molecules of mycosporine-like amino acids (MAAs) as seeds ($\mathbf{G}_{\mathbf{SEEDS}}$), molecules generated via chemical VAE ($\mathbf{G}_{\mathbf{VAE}}$) and molecules gathered via similarity search ($\mathbf{G}_{\mathbf{SIM}}$). The number of molecules that satisfy the condition for the light absorption ability of porphyra-334 in $\mathbf{G}_{\mathbf{SEEDS}}$, $\mathbf{G}_{\mathbf{VAE}}$, and $\mathbf{G}_{\mathbf{SIM}}$ are 52, 138, and 6, respectively. The method via chemical VAE shows a promising potential for future molecular design. By using quantum chemistry wave function properties for chemical VAE, we find new molecules that are comparable to porphyra-334, including some with unexpected geometries. At the end, we show a group of molecules found with this method.



1. INTRODUCTION

UV radiation (UVR) has become one of the subjects of environmental and green chemistry because of the decrease of the thickness of the ozone layer, which hinder the transmission of UVR from the sun to the Earth's surface. Sunlight is the primary energy source of living organisms; however, UVR damages human skin. It may act as the origin of skin cancers. Therefore, the development of efficient sunscreens without side effects is necessary. Porphyra-334 is a UV-resistive molecule in nature. Mycosporine-like amino acids (MAAs), including porphyra-334, are chemicals that prevent UVR-induced damage. They have attracted attention due to having a strong anti-UV effect.[1−4]

We reported previously a study on the molecular-level mechanism in energy transformations from sunlight to heat in porphyra-334 using first-principles molecular dynamics simulations and by quantum chemistry.[5,6] It revealed that the UV-excited porphyra-334 releases its kinetic energy via vibrational modes to surrounding water molecules. The structure of porphyra-334, which contains many hydrophilic functional groups, favors effective hydrogen bond formation with surrounding water molecules. Thus, the vibrational modes of water molecules absorb the energy from the excited molecule. This study provided an interpretation of *excellence* in a natural molecule, namely porphyra-334. An ambitious extension in molecular science is the design of such molecules. Therefore, we explore a design principle in an attempt to advance toward the natural products.

The design and selection of environmentally friendly and harmless materials and molecules are critical to establishing a sustainable society. They are mandatory for the development of functional molecules, drugs, and a wide range of materials. To achieve the sustainable conditions, many expensive experiments are in fact necessary. However, considering the time and cost of the society, we must provide, in parallel, computational support for the design and selection of these molecules and materials. Historically, the methodology so far has been based on the analogy of geometrical appearances (shapes) in molecules and materials starting from a lead molecule that is found more or less by chance. If such a methodology was sufficient, we would not be suffering from the current environmental problems.

Chemical space consists of the union of compounds. While the number of all feasible compounds is extremely high, estimated to be 10^{60} possible structures, only a small fraction

Received: November 16, 2021
Accepted: February 18, 2022
Published: March 2, 2022

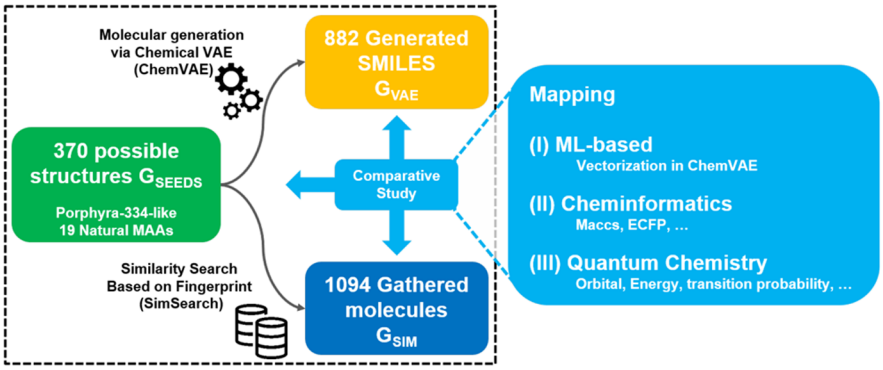


Figure 1. Scheme of the comparative study.

can be processed and analyzed at the same time.[7] Exploring the new horizon of chemical space is a challenge for cheminformatics and computational molecular design. An alternative approach that does not depend on the appearance or similarity of molecular shapes is necessary. We conducted a comparative study to find search criterions other than shapes and appearances, and the results are reported here.

One of the hopeful design approaches is learning from nature-made molecules such as porphyra-334. Porphyra-334, a molecule that survived in the long process of evolution, is considered to be the goal of UV-resistive natural products. We compare the approaches, one via the shapes and appearances method and the other that uses something different, as a clue to reach this goal. In fact, we are comparing the different processes of *lead-optimization*. We have carried out a comparison of the molecules generated via chemical variational autoencoder (ChemVAE)[8] versus the molecules gathered via Similarity Search (SimSearch).[9] Chemical VAE is a promising approach proposed recently that is based on machine learning. This provides great opportunities to generate a new molecule and to explore the search method in chemical space. In contrast, similarity search is a powerful conventionally applied method. Notice that Winter et al. proposed the application of chemical VAE in drug discovery,[10] and Gao et al. reported the availability of chemical VAE in application for the generation of novel alternative drug candidates for eight existing market drugs.[11] We compare lead-optimization processes starting from the natural product porphyra-334.

The group obtained via SimSearch is based on fingerprints from a chemical database. This cheminformatics method is a conventional search that is based on an existing chemical space. The molecular generation via ChemVAE is based on machine learning structural recognition; it transforms the input data from SMILES into the vector representation. There is no need to manually specify the mutation rules. As a result, unexpected jumps (to desired properties) in chemical space are possible. In the future, gradient-based optimization will be performed in combination with Bayesian statistics.[8]

In Figure 1, the scheme of current study is presented. The design approaches begin from the seeds, which are derivatives of the molecule porphyra-334; hereafter, we will refer to them as $\mathbf{G}_{SEEDS}$ (in green in Figure 1). The first molecular group was gathered via SimSearch, and the second was generated via ChemVAE; hereafter, we will call them $\mathbf{G}_{SIM}$ (in blue) and $\mathbf{G}_{VAE}$ (in orange), respectively.

For each group of molecules, SMILES data; 3D MOL data, that is, (x, y, z) coordinates; and properties by quantum chemical calculations were obtained. Data for each molecule are represented by vector elements. Then, the following three data mapping methods for $\mathbf{G}_{SEEDS}$, $\mathbf{G}_{VAE}$, and $\mathbf{G}_{SIM}$ were compared: (I) a machine learning (ML)-based comparison, (II) the cheminformatic comparison from 3D MOL, and (III) the quantum chemistry properties comparison from DFT calculations (see right in Figure 1, light blue). In this paper, we will show a demonstrative result that the new lead-optimization process produces promising results via Chem-VAE, especially in connection with quantum chemical calculations. We believe that the current study provides an example of machine learning applications in the search for desired molecule from the vast chemical space.

2. MATERIALS AND METHODS

2.1. Preparation of Three Molecular Groups. *The Seeds Structures (370) from MAAs Molecules (19) ($\mathbf{G}_{SEEDS}$).* A variety of UV-absorbing molecules, termed mycosporine-like amino acids (MAAs), have been reviewed by several researchers.[1−3,12−14] The MAAs from a marine organism are imine derivatives of mycosporines, as shown in Figure 2a. The MAA motif contain an amino-cyclohexen imine ring linked to an amino acid, an amino alcohol, or an amino group, which absorbs UV light from 320 to 362 nm[12] and shows photoprotective and antioxidant functions.

As an extension of a previous study on porphyra-334,[6] we study here the same family of molecules with a stable structure. Taking the ubiquitous photosensitive component of marine algae in a liquid water environment into account, we systematically and exhaustively obtained all possible structural isomers and tautomers that existed in the aqueous phase. Thus, derived from the 19 molecules shown in Figure 2a as porphyra-334 derivatives, 370 seeds structures ($\mathbf{G}_{SEEDS}$) were generated on account of the equilibrium in water. From the thus-prepared $\mathbf{G}_{SEEDS}$, two groups of molecules, namely $\mathbf{G}_{VAE}$ and $\mathbf{G}_{SIM}$, were obtained via the ChemVAE method and the SimSearch method, respectively.

Given the excellent properties of porphyra-334 in UV energy absorption and its dispersion mechanism,[1,6,14−18] we must include the protonated MAA motifs. The typical examples of protonated MAA motifs are shown in Figure 2b (see the SI for others). Thus, we added structures reflecting protonated and zwitterionic molecules (the 99 structures, which are included in the total 370 of $\mathbf{G}_{SEEDS}$; see the SI).

2.2. Molecular Generation via ChemVAE ($\mathbf{G}_{VAE}$). Gómez et al. reported a deep neural network model consisting of three coupled functions: an encoder, a decoder, and a

https://doi.org/10.1021/acsomega.1c06453
ACS Omega 2022, 7, 8581−8590

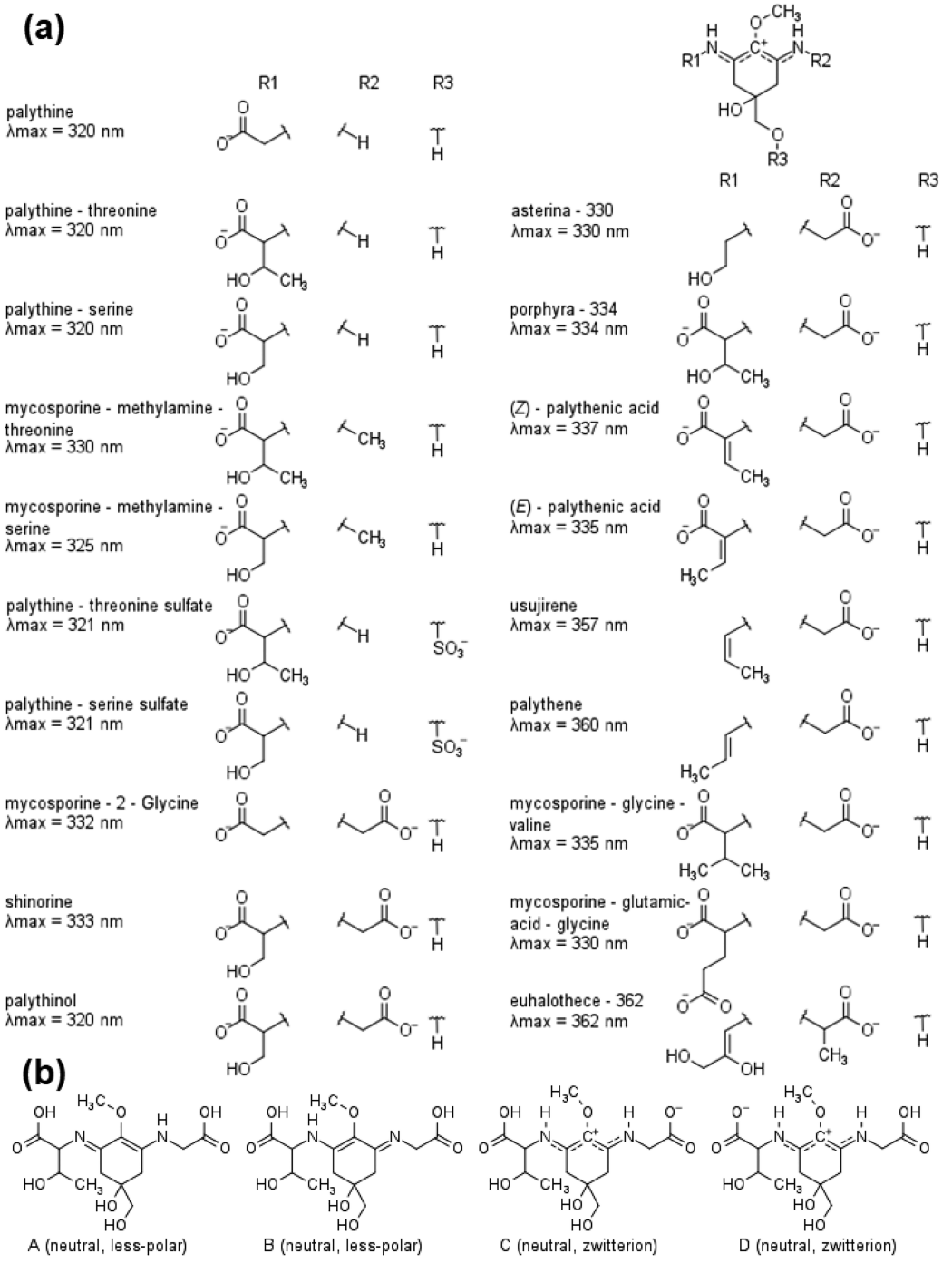


Figure 2. (a) Structures of 19 natural MAAs molecules. (b) Examples of protonated MAA motifs of porphyra-334 (see the SI for others).

predictor. It provides a machine learning-based *de novo* molecular design method.[8] The code and full training data sets are disclosed at their GitHub page.[19] This model was trained on hundreds of thousands of existing chemical structures, which allowed us to automatically generate novel chemical structures. Owing to this system, we could carry out the current study, that is, the group of molecules $\mathbf{G}_{\mathbf{VAE}}$ generated via ChemVAE.

Their autoencoder architecture is illustrated in Figure 3. Notations follow those from the paper by Gómez et al.[8] This trained autoencoder system has three latent representations: an embedding vector (X_1), a latent vector (z_1), and embedding vector (X_r); hereafter, we will call them X_1, z_1, and X_r, respectively. During the training, the canonical SMILES strings were assigned as an input to avoid confusion among chemically equivalent string representations. The encoder and the decoder shown in Figure 3 are recurrent neural networks (RNNs).

The encoder RNN that processes from a given SMILES string and the decorder RNN that processes from a given X_r are stochastic operations. As a result, the same input (smi) may be decoded into different outputs (smi_r), reflecting the different intermediates (X_1, z_1, or X_r). There is a possibility that the decoder RNN (from X_r to SMILES

https://doi.org/10.1021/acsomega.1c06453
ACS Omega 2022, 7, 8581−8590

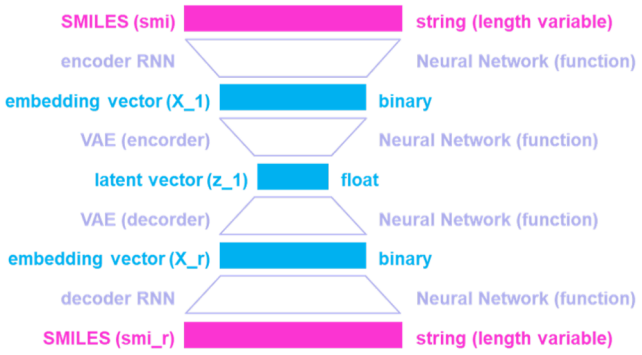


Figure 3. Scheme of the current Chem VAE.

(smi)) might result in chemically invalid strings. We collected the generated molecules, 2000 per one SMILES decoding attempt, iteratively for the ChemVAE method. After removing duplicated strings, we obtained 550784 strings for which we employed RDkit[20] to validate the chemical structures of the output molecules and discard invalid ones. Thus, we finally obtained 2454 SMILES strings. Meaningless structures were ruled out for the following reasons: having less than four heavy atoms, failing generate a 3D structure for quantum chemistry calculations, having unrealistic termination during quantum chemistry calculations, or having an unstable radical species. In total, 1572 molecules were excluded. Finally, 882 molecules ($\mathbf{G_{VAE}}$) were generated via ChemVAE (in orange, left in Figure 1).

2.3. Similarity Search by Fingerprint ($\mathbf{G_{SIM}}$). The SimSearch procedure in chemical databases is a well-known and widely used process.[9,21,22] We downloaded the "Annotated" subset of 1 458 577 582 molecules from ZINC15 (as shown in Figure S23).[23,24] It includes compounds that are in catalogs (but not for sale). We did not apply any other specific standardization to the molecular database. We gathered SMILES strings in accordance with Tanimoto similarity by utilizing MACCS, ECFP, and FCFP fingerprints (see the SI for details). ZINC15 is a research tool for investigators to search chemical and biological targets. Notice that fingerprints can be used for applications such as the current SimSearch as well as for molecular characterization, molecular diversity, and chemical database clustering. The MACCS keys have 166 bit structural key descriptors (vector with 166 elements) in which each bit is associated with a SMARTS pattern.[25,26] Extended-connectivity fingerprints (ECFPs) are circular topological fingerprints designed for various wide molecular studies and structure−activity modeling.[27,28] The ECFP encodes substructure patterns from molecules to a bit string length of 1024 (the length can be varied). The FCFP is a variant of this ECFP that is intended to capture precise atom environment substructural features. The FCFPs are intended to capture more abstract role-based substructural features.

These keys were implemented in the open-source cheminformatics software package RDkit. We gathered 1125 compounds from a database derived from $\mathbf{G_{SEEDS}}$ (in green in Figure 1). We removed some chemicals because of their failure to prepare 3D structures for quantum chemistry calculations. At the final stage, we obtained 1094 chemicals ($\mathbf{G_{SIM}}$) to be considered in the chemical space exploration (in blue, left in Figure 1).

2.4. Quantum Chemistry Properties. To prepare geometric data for quantum chemistry calculations, the MMFF94 force field implemented with RDkit was applied to construct 3D structures for $\mathbf{G_{SEEDS}}$, $\mathbf{G_{VAE}}$, and $\mathbf{G_{SIM}}$. We then performed the calculations for the ground and excited states using density functional theory (DFT). We used the B3LYP hybrid functional and the 6-31G(d) basis sets. The solvent effect of water was taken account by the integral equation formalism of the polarization continuum model (IEFPCM). We used the Gaussian 16 program package.[29] We first carried out the geometry optimizations of the ground states, starting from the structure generated by RDkit. We then performed the single-point calculation of the excited states using time-dependent density functional theory (TD-DFT).

As shown in Table 1, we extracted 23 properties from the calculated results, such as total energies, the HOMO (highest occupied molecular orbital)−LUMO (lowest unoccupied molecular orbital) gap energies, three orbital energies around the HOMO and the LUMO, viral coefficients, dipole moments, quadrupole moments, the degrees of freedom in the structures, the trace of the quadrupole moment, and the coordinate invariants of the quadrupole moment (Table 1). Ground-state properties are selected for versatility. In total, there are 84 elements for each vector. Then, we carried out the PCA analyses, to be mentioned later.

Table 1. Quantum Chemistry Properties Obtained from DFT Calculations and Some Physical Chemical Properties

detail	number of elements
estimated molecular volume[30]	1
difference of the orbital energies (eigen values) of the HOMO and LUMO	1
quadrupole moment	3
total dipole moment	1
total energy and the viral coefficient	1
electronic spatial extent	1
absorption wavelength (nm) of the *n*th excited state	20
absorption energy (eV) of the *n*th excited state	20
oscillation strength of the *n*th excited state	20
number of electrons	1
orbital energy (eigen value) of first through third highest occupied molecular orbitals	3
orbital energy (eigen value) of first through third lowest unoccupied molecular orbitals	3
rotational constants	3
degree of freedom	1
number of (H, C, N, O, and S) atoms	5

2.5. Mapping. Representations of various vectors in chemical space[7,31] were applied for the comparison or exploration of the internal relations. It is necessary to map higher-ordered complex information onto a low-dimensional space. One typical mapping method is principal component analysis (PCA),[32] which is used for exploratory data analysis and to make predictive models. It is commonly used for dimensional reduction by projecting each data point onto only the few principal components to obtain lower-dimensional data. We show the first two principal components, and the cumulative contribution rate data are shown in the SI.

3. RESULTS AND DISCUSSIONS

3.1. Representation for Three Groups: $\mathbf{G_{SEEDS}}$, $\mathbf{G_{VAE}}$, and $\mathbf{G_{SIM}}$. We present here the results obtained via ChemVAE generation and SimSearch mining. The comparison of the

https://doi.org/10.1021/acsomega.1c06453
ACS Omega 2022, 7, 8581−8590

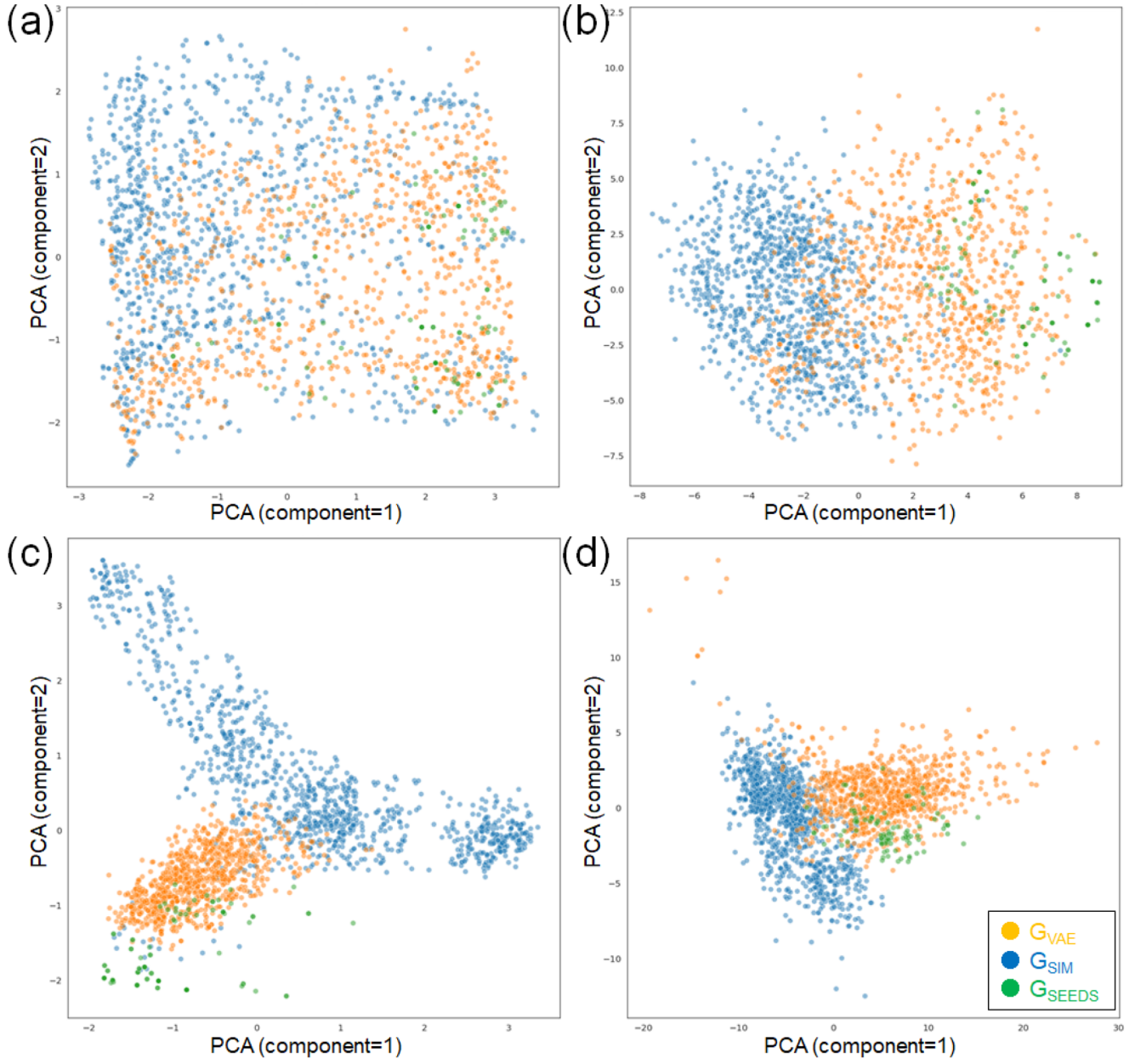


Figure 4. Mapping of principal components analyses for three groups, namely $\mathbf{G}_{\mathbf{SEEDS}}$ (green), $\mathbf{G}_{\mathbf{VAE}}$ (orange, via ChemVAE) and $\mathbf{G}_{\mathbf{SIM}}$ (blue, via SimSearch), using (a) ML-based vector X_1, (b) ML-based vector z_1, (c) cheminformatics (ECFP), and (d) (III) quantum chemistry.

three groups ($\mathbf{G}_{\mathbf{SEEDS}}$, $\mathbf{G}_{\mathbf{VAE}}$, and $\mathbf{G}_{\mathbf{SIM}}$) was carried out by mapping three different viewpoints: (I) ML-based, (II) cheminformatics, and (III) quantum chemistry (right in Figure 1). It is noteworthy that we used the ChemVAE method again in the mapping process. That is, in the process of (I) the ML-based process (Figure 1), we use SMILES strings for $\mathbf{G}_{\mathbf{VAE}}$ and $\mathbf{G}_{\mathbf{SIM}}$ as the input (the second time) for the ChemVAE procedure, then we obtained output vectors of X_1, z_1, and X_r with which we carried out the PCA mapping . The results of X_1 and z_1 from the ChemVAE vectorization are shown below. For the X_r results, see the SI.

3.2. Mapping (I): ML-Based Comparison. Two chemical space representations were mapped by PCA via ChemVAE vectorization as shown in Figure 4 a and b (see the SI for the X_r results). At first, the mapping results for the vectors (X_1) are shown in Figure 4 a, where $\mathbf{G}_{\mathbf{VAE}}$ is distributed slightly closer to $\mathbf{G}_{\mathbf{SEEDS}}$ than $\mathbf{G}_{\mathbf{SIM}}$. For the second mapping, the vector (z_1) is shown in Figure 4 b. Now, we observe that $\mathbf{G}_{\mathbf{VAE}}$ is distributed distinctly closer to $\mathbf{G}_{\mathbf{SEEDS}}$ than $\mathbf{G}_{\mathbf{SIM}}$. The PCA mapping is one of the various methods used. We stay with the method due to its well-known versatility.[33,34] We also show the results from t-SNE in Figures S15−21 in the SI. The main arguments are the same.

3.3. Mapping (II): Cheminformatics Comparison. Chemical space is usually described by molecular descriptors, so-called descriptor space. We adopted the ECFP fingerprint for these three groups, namely $\mathbf{G}_{\mathbf{SEEDS}}$, $\mathbf{G}_{\mathbf{VAE}}$, and $\mathbf{G}_{\mathbf{SIM}}$. The PCA mapping results are shown in Figure 4 c (see the SI for results by MACCS and FCFP). The results show that $\mathbf{G}_{\mathbf{VAE}}$ is closer to $\mathbf{G}_{\mathbf{SEEDS}}$ than $\mathbf{G}_{\mathbf{SIM}}$. Interestingly, the groups $\mathbf{G}_{\mathbf{VAE}}$ and $\mathbf{G}_{\mathbf{SIM}}$ are located in different areas of the chemical space. This result shows that the two methods, ChemVAE and SimSearch,

https://doi.org/10.1021/acsomega.1c06453
ACS Omega 2022, 7, 8581−8590

provide two distinct groups of molecules, suggesting the high potential of ChemVAE as a method for searching through criteria different from similarity toward new areas in chemical space.

3.4. Mapping (III): Quantum Chemistry Properties. The chemical space spanned by vectors consisting of quantum chemistry properties is expressed by PCA and shown in Figure 4 d. It can be seen from this result that the distribution of $\mathbf{G}_{\mathbf{VAE}}$ is located closer to $\mathbf{G}_{\mathbf{SEEDS}}$ than $\mathbf{G}_{\mathbf{SIM}}$. Contrasting with the other mappings shown above, as shown in Figure 4d, the distribution of the two groups $\mathbf{G}_{\mathbf{VAE}}$ and $\mathbf{G}_{\mathbf{SIM}}$ scarcely overlap. Therefore, we can infer the fact that the molecules in $\mathbf{G}_{\mathbf{VAE}}$ are differentiated well from those in $\mathbf{G}_{\mathbf{SIM}}$ when these vector elements consist of quantum chemistry properties.

The results shown in Figure 4a−d indicate why it is so critical that we adopt a relevant vector for each molecule. As shown in Figure 4d, we have arrived at a mapping that enables us to distinguish among three groups of our samples. By adopting a vector whose elements consist of quantum chemical properties, reflecting the wave function of each molecule, we can differentiate the groups well. The results suggest that we can obtain molecules (in orange) that might be comparable to porphyra-334. These differentiated molecules may potentially be new molecules.

Here, we had better mention that there may be another possibility for the vector selection. The relevant vectors led us to the best mapping in the molecular space to find molecules comparable to porphyra-334. What is a rational procedure to find such an optimal vector? To the best of our knowledge, there is no established methodology. This is a very important issue in future. Recently, some ML-based fingerprints have been published. Among them is the promising fingerprint Mol2vec,[35] which has been applied for drug discovery,[36,37] solvation free energy prediction,[38] the prediction of pK_a values of CH acids,[39] and other material designs. Examples include other ML-based fingerprints such as one that uses graph-convolution models[40] and another proceeds by the evolution of the embedding step[41] (including an application for SAR/SPR). Obtaining a rational procedure for creating a linkage between classical fingerprints and ML-based fingerprints will be a future subject.

3.5. Differences among the Three Groups from a Quantum Chemistry Point of View. The purpose of the current study is to find excellent molecules. Therefore, we examine the obtained molecules in three groups from a physical chemistry point of view. The MAAs are known to possess high stabilities even under relatively strong UV irradiation.[42] The absorbed energy is expected to be dissipated very efficiently to the surrounding water environment.[5,7,29,31,42−45] It is the typical mechanism for porphyra-334 and its charasteristics of UV-resistance and the non-destructive release of energy properties.

Among many properties of porphyra-334, we must consider the critical ones, that is, its hydrophilic property (log P), absorption wavelength (λ_{max}), and oscillator strength (f). Although log P is widely used, we focus here on quantum chemistry properties and did not include log P. The results with log P included did not change our conclusion described below. The details of the results and arguments for log P are explained in the SI. Since the excitation wavelength (λ_{max}) in UV−visible range and the oscillator strength (f) are the indispensable properties for the optical property in porphyra-334, we employed the TD-DFT method to calculate the excitation energies and oscillator strengths of the three groups $\mathbf{G}_{\mathbf{SEEDS}}$, $\mathbf{G}_{\mathbf{VAE}}$, and $\mathbf{G}_{\mathbf{SIM}}$.

Among the various UV regions, namely UVB (280−315 nm), UVA1 (315−340 nm), and UVA2 (340−400 nm), we filtered molecules whose calculated spectral characteristics were in the 300−350 nm range, reflecting the absorbing range of porphyra-334. We paid special attention on the zwitterionic isomers, since the protonated MAA motifs for photoprotective and antioxidant functions are critical isomers, as was reported in our previous study.[6] We extracted charge-neutral and zwitterionic forms of $\mathbf{G}_{\mathbf{SIM}}$ via SimSearch and $\mathbf{G}_{\mathbf{VAE}}$ via ChemVAE. The histogram of the calculated oscillator strengths is shown in Figure 5. Thus, the number of molecules that

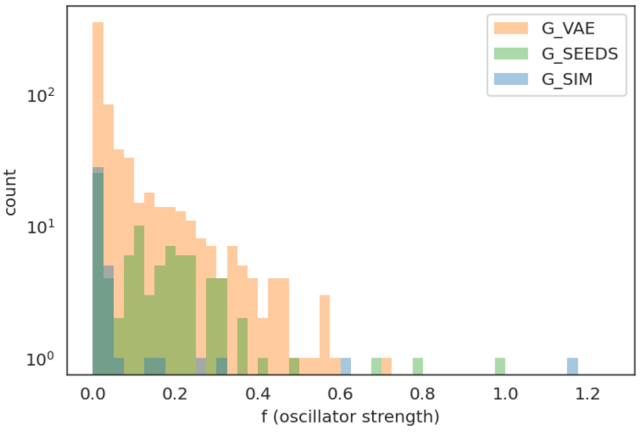


Figure 5. Histogram of calculated oscillator strengths in the 300 < λ < 350 nm range for the three groups, namely $\mathbf{G}_{\mathbf{SEEDS}}$ (green), $\mathbf{G}_{\mathbf{VAE}}$ (orange, via ChemVAE), and $\mathbf{G}_{\mathbf{SIM}}$ (blue, via SimSearch).

satisfied the threshold of spectral properties f > 0.1 and 300 < λ < 350 for $\mathbf{G}_{\mathbf{SEEDS}}$, $\mathbf{G}_{\mathbf{VAE}}$, and $\mathbf{G}_{\mathbf{SIM}}$, are 52, 138, and 6, respectively. These molecules were finally filtered and scrutinized described below.

3.6. Mapping of the Final Selected Molecules. The results shown in Figure 4 for ML-based, cheminformatics-based, and quantum chemistry-based mappings were filtered by the criteria f > 0.1 and 300 < λ < 350, and results are shown in Figure 5. We then focused on the selected molecules and examined the features of these molecules. The results are shown in Figure 6.

All the plots in Figure 6 satisfy the conditions f > 0.1 and 300 < λ < 350. As shown in Figure 6, the data points (each plot corresponds to each molecule expressed by one vector from X_1 or z_1 of the ChemVAE vectorization) cannot be clearly divided into clusters. This is quite natural in the sense that the results at the X_1 or z_1 level still correspond to these bu way of machine learning.

By contrast, the data shown in Figure 6 c show relatively separated features in two clusters. One is the $\mathbf{G}_{\mathbf{VAE}}$ group (orange) and the other is the $\mathbf{G}_{\mathbf{SEEDS}}$ (green) and $\mathbf{G}_{\mathbf{SIM}}$ (blue) groups. In the latter, the two groups ($\mathbf{G}_{\mathbf{SEEDS}}$ and $\mathbf{G}_{\mathbf{SIM}}$) are mostly overlapped. These results suggest the possibility that we can somehow explore new chemical space using vectors generated via ChemVAE, even though at this stage the elements consist only of structural information and do not yet include quantum chemistry information.

At the final stage, as shown in Figure 6d, the plots show a promising feature. These data were generated via the vectors

https://doi.org/10.1021/acsomega.1c06453
ACS Omega 2022, 7, 8581−8590

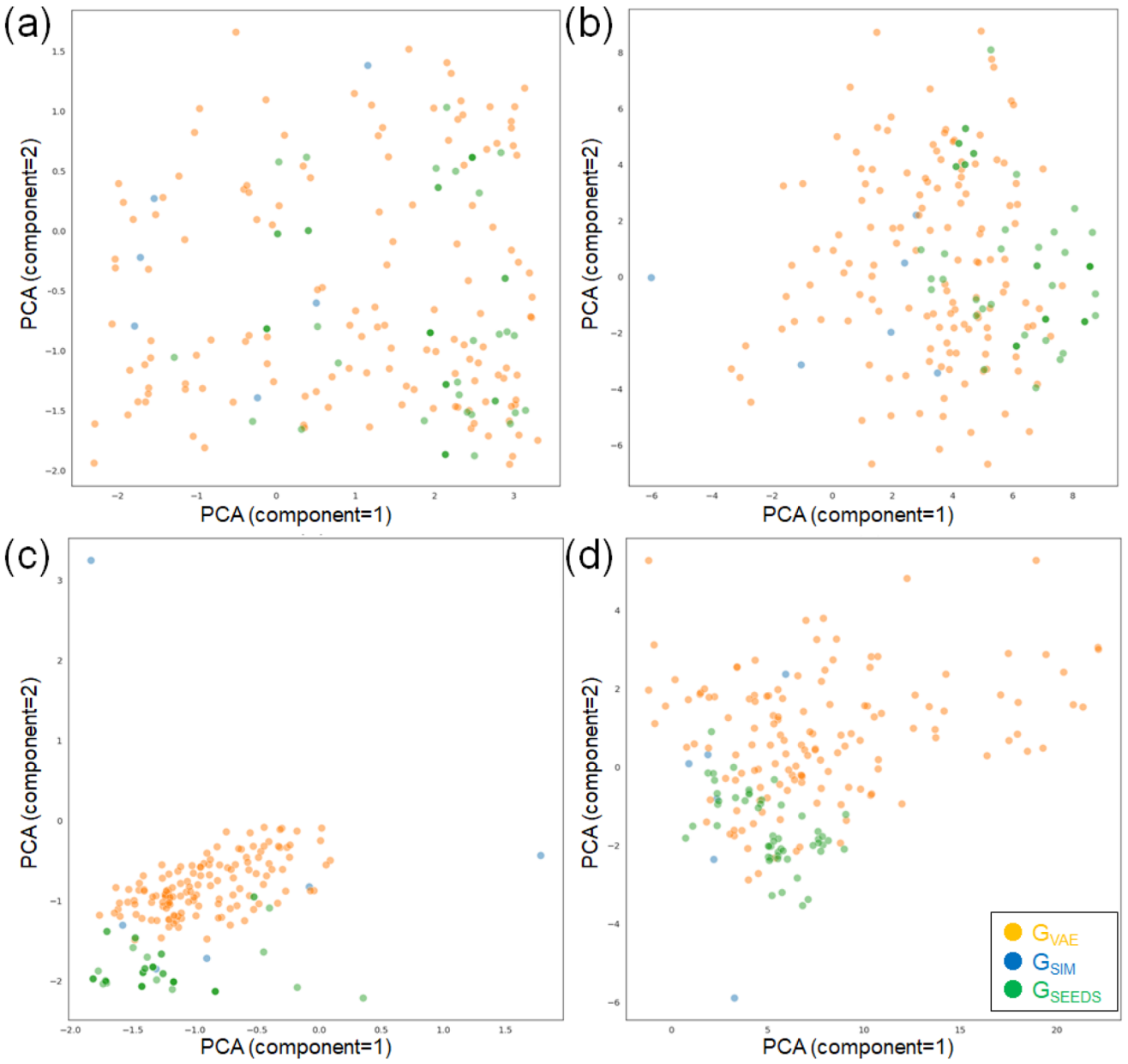


Figure 6. Filtered molecules ($f > 0.1$ and $300 < \lambda < 350$ nm) from those shown in Figure 4 for (a) X_1, (b) z_1, (c) ECFP (fingerprint), and (d) quantum chemistry.

whose elements consisted of quantum chemical properties. The $\mathbf{G_{VAE}}$ (orange) data show a distribution with a large diversity, whereas the other two, $\mathbf{G_{SIM}}$ (blue) and $\mathbf{G_{SEEDS}}$ (green), are covered by the $\mathbf{G_{VAE}}$ (orange) zone; they stay in one section and do not spread, suggesting their properties have less diversity. From the aspects shown in Figure 4d and Figure 6d, as a matter of fact, many molecules belonging to $\mathbf{G_{SIM}}$ were rejected by the filtration criteria (f and λ). When we take the quantum chemical properties into account, we can explore the chemical space more widely via ChemVAE than via SimSearch.

It may be relevant to cite here the arguments given by Gómez et al.[8] and various researchers[46−49] as well as the reported studies in which quantum chemical properties were predicted by machine learning.[50,51] Moreover, some studies using transfer learning have been published.[47,52] A future subject remains, specifically how to find new strings of molecular representation beyond SMILES. Currently we are using only SMILES strings, therefore the performance of machine learning for chemical information is still limited. It is noteworthy that recently some research examples beyond SMILES have appeared, such as those from graph theory[51] and those from linear string.[46]

The current mapping in Figure 6d shows that quantum chemical properties *do* extend a new horizon of the search area. Methodologies based on molecular machine learning (ChemVAE) are thus promising when we add quantum chemical properties.

The excellence of porphyra-334 may not be limited only to its intramolecular properties. The excellence may exist further in its ability to form intermolecular interactions such as subtle hydrogen bond networks. If we can include molecular information derived from other dimensions such as wave

https://doi.org/10.1021/acsomega.1c06453
ACS Omega 2022, 7, 8581−8590

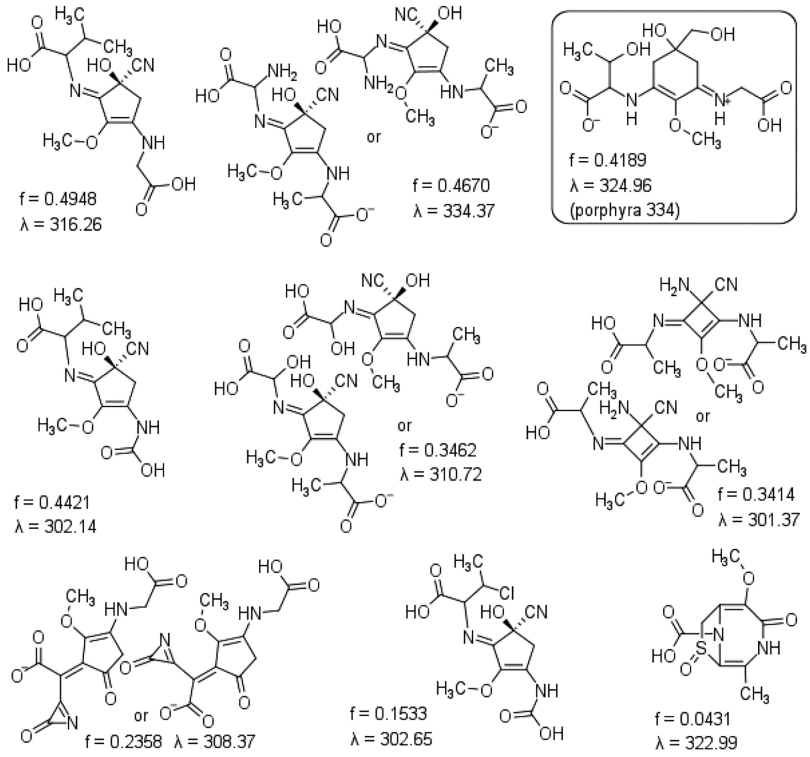


Figure 7. Selected molecular structures from G_{VAE}.

functions and responsive properties to the environment instead of solely structures, the potential of machine learning will be further realized. The inclusion of such properties will be a future subject.

3.7. *De Novo* Molecules Generated via ChemVAE. According to calculated spectral properties and the mappings after filtration, we have now demonstrated a promising performance of the method via ChemVAE. We show representative examples of the filtered and selected final structures from $\mathbf{G}_{VAE}$ in Figure 7.

To show the currently obtained promising feature of ChemVAE molecular generation together with quantum chemistry properties, we display eight representative molecules in Figure 7. Among the filtered (selected) molecules shown in Figure 6d, these eight representative molecules are located in the vicinity of $\mathbf{G}_{SEEDS}$ plots. The other $\mathbf{G}_{VAE}$ molecules are also shown in the SI. By contrast, only six molecules from the $\mathbf{G}_{SIM}$ group satisfied the calculated spectral requirements (see the SI).

As shown in Figure 7, the presence of molecules with a five-membered ring is noteworthy. In their molecular molecular paper, Losantos et al.[17,18] reported the protonated MAA motifs and also proposed protonated five-membered-ring motifs. Since natural bioactive MAAs have six-membered-ring motifs, their rational design shows the significance. Indeed, the thus-proposed five-membered-ring photoactive molecules were not registered in the database of ZINC15 until now. Even among the molecules in the $\mathbf{G}_{SIM}$ group obtained via SimSearch, we could not find the molecules that they designed. By contrast, we generated the molecules with five-membered rings, as shown in Figure 7, in the $\mathbf{G}_{VAE}$ group via ChemVAE.

4. CONCLUSIONS

This study reports the results of a comparative study between the ChemVAE method and the SimSearch method, which was focused on their generation ability for new functional molecular designs. Defining the natural porphyra-334 as a model molecule, we generated three groups: molecules of MAAs as seeds, molecules generated via ChemVAE, and molecules gathered via SimSearch ($\mathbf{G}_{SEEDS}$, $\mathbf{G}_{VAE}$, and $\mathbf{G}_{SIM}$, respectively). There were 52, 138, and 6 molecules that satisfied the condition of the light absorption ability of porphyra-334 at $f > 0.1$ and $300 < \lambda < 350$ in $\mathbf{G}_{SEEDS}$, $\mathbf{G}_{VAE}$, and $\mathbf{G}_{SIM}$, respectively. The ChemVAE method shows promising potential for future molecular design capability. When we use quantum chemistry properties for the ChemVAE method, we can obtain molecules significantly comparable to porphyra-334, including unexpected ones (five-membered ring).

4.1. Data and Software Availability. We used the Gaussian 16 program package[29] for the quantum chemistry calculations. We used RDkit[20] for the 3D structure construction (MMFF94 force field), the fingerprints (MACCS, ECFP, and FCFP), and the Tanimoto similarity of the fingerprints. We used the OpenBabel toolkit[53] for the data I/O. The multivariate analysis and mapping are proprietary but not restricted to our program.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c06453.

Gaussian input files for the three groups ($\mathbf{G}_{SEEDS}$, $\mathbf{G}_{VAE}$, and $\mathbf{G}_{SIM}$) (ZIP)

Additional experimental details and methods (PDF)

https://doi.org/10.1021/acsomega.1c06453
ACS Omega 2022, 7, 8581−8590

Data for mappings and structures for the three groups (G_{SEEDS}, G_{VAE}, and G_{SIM}) (ZIP)

AUTHOR INFORMATION

Corresponding Author

Shinichiro Nakamura − *Cluster for Science, Technology, and Innovation Hub, Nakamura Laboratory, RIKEN, Wako, Saitama 351-0198, Japan;* orcid.org/0000-0002-6437-6993; Email: snakamura@riken.jp

Authors

Yuki Harada − *Cluster for Science, Technology, and Innovation Hub, Nakamura Laboratory, RIKEN, Wako, Saitama 351-0198, Japan*

Makoto Hatakeyama − *Cluster for Science, Technology, and Innovation Hub, Nakamura Laboratory, RIKEN, Wako, Saitama 351-0198, Japan; Sanyo-Onoda City University, Sanyo-Onoda, Yamaguchi 756-0884, Japan*

Shuichi Maeda − *Cluster for Science, Technology, and Innovation Hub, Nakamura Laboratory, RIKEN, Wako, Saitama 351-0198, Japan*

Qi Gao − *Mitsubishi Chemical Corporation Science & Innovation Center, Yokohama, Kanagawa 227-8502, Japan*

Kenichi Koizumi − *Cluster for Science, Technology, and Innovation Hub, Nakamura Laboratory, RIKEN, Wako, Saitama 351-0198, Japan*

Yuki Sakamoto − *Cluster for Science, Technology, and Innovation Hub, Nakamura Laboratory, RIKEN, Wako, Saitama 351-0198, Japan;* orcid.org/0000-0003-2249-0779

Yuuki Ono − *Mitsubishi Chemical Corporation Science & Innovation Center, Yokohama, Kanagawa 227-8502, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c06453

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors are grateful to Professor Kunio Tanabe for helpful discussions. We thank the HOKUSAI system of RIKEN for computational resources. We also thank the computer facilities of RCCS at Okazaki, Japan.

REFERENCES

(1) Abiola, T. T.; Whittock, A. L.; Stavros, V. G. Unravelling the Photoprotective Mechanisms of Nature-Inspired Ultraviolet Filters Using Ultrafast Spectroscopy. *Molecules* **2020**, *25*, 3945.

(2) Geraldes, V.; Pinto, E. Mycosporine-Like Amino Acids (MAAs): Biology, Chemistry and Identification Features. *Pharmaceuticals* **2021**, *14*, 63.

(3) Bedoux, G.; Pliego-Cortés, H.; Dufau, C.; Hardouin, K.; Boulho, R.; Freile-Pelegrín, Y.; Robledo, D.; Bourgougnon, N. Chapter Seven - Production and properties of mycosporine-like amino acids isolated from seaweeds. In *Advances in Botanical Research*, Vol. *95*; Bourgougnon, N., Ed.; Elsevier, 2020; pp 213−245.

(4) Lawrence, K. P.; Long, P. F.; Young, A. R. Mycosporine-like amino acids for skin photoprotection. *Curr. Med. Chem.* **2019**, *25*, 5512−5527.

(5) Boggio-Pasqua, M.; Robb, M. A.; Bearpark, M. J. Photostability via a sloped conical intersection: A CASSCF and RASSCF study of pyracylene. *J. Phys. Chem. A* **2005**, *109*, 8849−8856.

(6) Hatakeyama, M.; Koizumi, K.; Boero, M.; Nobusada, K.; Hori, H.; Misonou, T.; Kobayashi, T.; Nakamura, S. Unique structural relaxations and molecular conformations of porphyra-334 at the excited state. *J. Phys. Chem. B* **2019**, *123*, 7649−7656.

(7) Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S. Progress in visual representations of chemical space. *Expert Opin. Drug Discovery* **2015**, *10*, 959−973.

(8) 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.

(9) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Model.* **1998**, *38*, 983−996.

(10) Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D.-A. Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **2019**, *10*, 8016−8024.

(11) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative network complex for the automated generation of drug-like molecules. *J. Chem. Inf. Model.* **2020**, *60*, 5682−5698.

(12) Carreto, J. I.; Carignan, M. O. Mycosporine-like amino acids: relevant secondary metabolites. Chemical and ecological aspects. *Mar. Drugs* **2011**, *9*, 387−446.

(13) Whittock, A. L.; Turner, M. A.; Coxon, D. J.; Woolley, J. M.; Horbury, M. D.; Stavros, V. G. Reinvestigating the Photoprotection Properties of a Mycosporine Amino Acid Motif. *Front. Chem.* **2020**, *8*, 2296−2646.

(14) Woolley, J. M.; Stavros, V. G. Unravelling photoprotection in microbial natural products. *Sci. Prog.* **2019**, *102*, 287−303.

(15) Koizumi, K.; Hatakeyama, M.; Boero, M.; Nobusada, K.; Hori, H.; Misonou, T.; Nakamura, S. How seaweeds release the excess energy from sunlight to surrounding sea water. *Phys. Chem. Chem. Phys.* **2017**, *19*, 15745−15753.

(16) Woolley, J. M.; Staniforth, M.; Horbury, M. D.; Richings, G. W.; Wills, M.; Stavros, V. G. Unravelling the photoprotection properties of mycosporine amino acid motifs. *J. Phys. Chem. Lett.* **2018**, *9*, 3043−3048.

(17) Losantos, R.; Funes-Ardoiz, I.; Aguilera, J.; Herrera-Ceballos, E.; García-Iriepa, C.; Campos, P. J.; Sampedro, D. Rational design and synthesis of efficient sunscreens to boost the solar protection factor. *Angew. Chem.* **2017**, *129*, 2676−2679.

(18) Losantos, R.; Lamas, I.; Montero, R.; Longarte, A.; Sampedro, D. Photophysical characterization of new and efficient synthetic sunscreens. *Phys. Chem. Chem. Phys.* **2019**, *21*, 11376−11384.

(19) Wei, J.; Sanchez-Lengeling, B.; Sheberal, D.; Gomez-Bomberelli, R.; Aspuru-Guzik, A.. *chemical VAE*; GitHub, 2019. https://github.com/aspuru-guzik-group/chemical_vae.

(20) *RDKit: Open-Source Cheminformatics Software*; RDKit. https://www.rdkit.org.

(21) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry: miniperspective. *J. Med. Chem.* **2014**, *57*, 3186−3204.

(22) Kunkel, C.; Schober, C.; Oberhofer, H.; Reuter, K. Knowledge discovery through chemical space networks: The case of organic electronics. *J. Mol. Model.* **2019**, *25*, 87.

(23) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

(24) Sterling, T.; Irwin, J. J. ZINC 15−ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.

(25) Nguyen, D. D.; Wei, G.-W. DG-GL: Differential geometry-based geometric learning of molecular datasets. *Int. J. Number. Meth. Biomed. Engng.* **2019**, *35*, No. e3179.

(26) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Model.* **2002**, *42*, 1273−1280.

(27) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208−220.

(28) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(29) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H. et al. *Gaussian 09*, rev. D.01.; Gaussian, Inc.: Wallingford, CT, 2009 (b) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H. et al. *Gaussian 16*, rev. A; Gaussian, Inc.: Wallingford, CT, 2016, *3*.
(30) Yanagisawa, K. *python_tools/calc_mol_volume.py*; GitHub, 2016. https://github.com/keisuke-yanagisawa/python_tools/blob/master/calc_mol_volume.py.
(31) Vogt, M. How do we optimize chemical space navigation? *Expert Opin. Drug Discovery* **2020**, *15*, 523−525.
(32) Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer-Verlag: Berlin, Germany, 2006.
(33) Moreno-Torres, J. G.; Raeder, T.; Alaiz-Rodríguez, R.; Chawla, N. V.; Herrera, F. A unifying view on dataset shift in classification. *Pattern recognition* **2012**, *45*, 521−530.
(34) Quiñonero-Candela, J.; Sugiyama, M.; Lawrence, N. D.; Schwaighofer, A. *Dataset shift in machine learning*; MIT Press: Cambridge, MA, 2009.
(35) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.
(36) Cai, C.; Guo, P.; Zhou, Y.; Zhou, J.; Wang, Q.; Zhang, F.; Fang, J.; Cheng, F. Deep learning-based prediction of drug-induced cardiotoxicity. *J. Chem. Inf. Model.* **2019**, *59*, 1073−1084.
(37) Shan, X.; Wang, X.; Li, C.-d.; Chu, Y.; Zhang, Y.; Xiong, Y.; Wei, D.-Q. Prediction of CYP450 Enzyme− Substrate Selectivity Based on the Network-Based Label Space Division Method. *J. Chem. Inf. Model.* **2019**, *59*, 4577−4586. PMID: 31603319.
(38) Lim, H.; Jung, Y. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chem. Sci.* **2019**, *10*, 8306−8315.
(39) Roszak, R.; Beker, W.; Molga, K.; Grzybowski, B. A. Rapid and Accurate Prediction of p K a Values of C−H Acids Using Graph Convolutional Neural Networks. *J. Am. Chem. Soc.* **2019**, *141*, 17142−17149.
(40) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019**, *10*, 1692−1701.
(41) Zheng, S.; Yan, X.; Yang, Y.; Xu, J. Identifying Structure−Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism. *J. Chem. Inf. Model.* **2019**, *59*, 914−923. PMID: 30669836.
(42) Klisch, M.; Richter, P.; Puchta, R.; Häder, D.-P.; Bauer, W. The stereostructure of porphyra-334: an experimental and calculational NMR investigation. Evidence for an efficient 'proton sponge. *Helv. Chim. Acta* **2007**, *90*, 488−511.
(43) Karlov, D. S.; Sosnin, S.; Tetko, I. V.; Fedorov, M. V. Chemical space exploration guided by deep neural networks. *RSC Adv.* **2019**, *9*, 5151−5157.
(44) Frauenheim, T.; Seifert, G.; Elsterner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. A self-consistent charge density-functional based tight-binding method for predictive materials simulations in physics, chemistry and biology. *Phys. Status Solidi B* **2000**, *217*, 41−62.
(45) Martinez, T. J. Seaming is believing. *Nature* **2010**, *467*, 412−413.
(46) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **2020**, *1*, 045024.
(47) Karpov, P.; Godin, G.; Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminf.* **2020**, *12*, 17.
(48) Hinton, G. E.; Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science* **2006**, *313*, 504−507.
(49) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI-the worldwide chemical structure identifier standard. *J. Cheminf.* **2013**, *5*, 7.
(50) Peng, S.-P.; Zhao, Y. Convolutional neural networks for the design and analysis of non-fullerene acceptors. *J. Chem. Inf. Model.* **2019**, *59*, 4993−5001.
(51) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584−4601.
(52) Li, X.; Fourches, D. Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT. *J. Cheminf.* **2020**, *12*, 27.
(53) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2008**, *2*, 5.

https://doi.org/10.1021/acsomega.1c06453
ACS Omega 2022, 7, 8581−8590