ACS OMEGA

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

Article

MOLGENGO: Finding Novel Molecules with Desired Electronic Properties by Capitalizing on Their Global Optimization

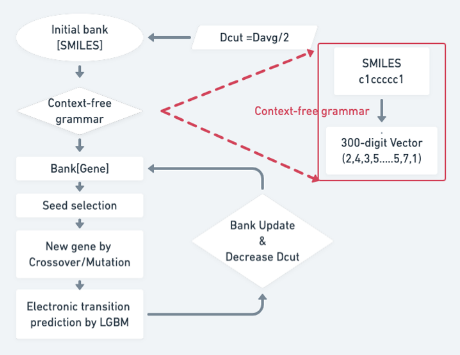
Beomchang Kang, Chaok Seok,* and Juyong Lee*

Cite This: *ACS Omega* 2021, 6, 27454−27465

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: The discovery of novel and favorable fluorophores is critical for understanding many chemical and biological studies. High-resolution biological imaging necessitates fluorophores with diverse colors and high quantum yields. The maximum oscillator strength and its corresponding absorption wavelength of a molecule are closely related to the quantum yields and the emission spectrum of fluorophores, respectively. Thus, the core step to design favorable fluorophore molecules is to optimize the desired electronic transition properties of molecules. Here, we present MOLGENGO, a new molecular property optimization algorithm, to discover novel and favorable fluorophores with machine learning and global optimization. This study reports novel molecules from MOLGENGO with high oscillator strength and absorption wavelength close to 200, 400, and 600 nm. The results of MOLGENGO simulations have the potential to be candidates for new fluorophore frameworks.



1. INTRODUCTION

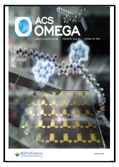
Fluorophores play crucial roles in various disciplines such as medicine, biochemistry, spectroscopy, and analytical chemistry.[1−4] They are widely used to screen toxic compounds at the molecular level and observe protein−protein interactions.[5,6] The discovery of novel fluorophores will open new possibilities in biology and biochemistry because only a small number of fluorophores are commonly used at present.[7] It is essential to optimize their maximum oscillator strength ($f_{\max}$) and the corresponding absorption wavelength ($\lambda_{\max}$) to the desired values to design bright and fluorophores with various colors.[8]

Conventionally, the discovery of most of the new fluorophores has been accomplished by the established rules, guidelines, and strategies.[9−11] Finding a novel scaffold by conventional experimental molecular discovery approach without the established rules, guidelines, and strategies demands astronomical amounts of resources and time to synthesize and experimentally verify the properties of candidate molecules.[7] Despite decades of endeavors by chemists, only a few fluorophores are commonly used, such as fluorescein,[12] bodify,[13] cyanine,[14] bisbenzimide,[15] coumarin,[16] rhodamine,[17] and others.[18] In this study, we aimed to develop a computational approach to find novel scaffolds, which are distinct from known ones without using explicit rules, guidelines, and strategies.

Various computational approaches have been suggested for efficient optimization of the desired electronic transition properties.[19−21] Sumita and co-workers developed ChemTS, which utilized Monte Carlo tree search with a recurrent neural network as a molecular generator and density functional theory (DFT) calculation as the evaluator of the desired electronic transition.[19] They synthesized and validated five fluorophores. Henault and co-workers applied a graph-based genetic algorithm (GB-GA) and the tight-binding-based simplified Tamm−Dancoff approximation (sTDA-xTB) as a molecular generator and a desired electronic transition evaluator.[20] They reported nine molecules, which are expected to have favorable electronic properties. Leguy and co-workers combined a graph-based evolutionary algorithm (EvoMol) to generate new molecules with density functional theory calculations.[21] They reported 15 molecules with low E_{LUMO} and 15 molecules with high E_{HOMO}.

Favorable fluorophores must have high $f_{\max}$ and $\lambda_{\max}$, which is close to the target $\lambda_{\max}$. In a practical sense, toxicity, biocompatibility, chemical stability, and photostability are also important properties for fluorophores. However, considering all properties simultaneously is a highly challenging task and out of the scope of this study. This study focused on optimizing $f_{\max}$ and $\lambda_{\max}$ as the first step to discovering new fluorophores. However, many previous approaches did not optimize $f_{\max}$. Sumita and co-workers only optimized $\lambda_{\max}$.[19] Leguy and co-workers optimized only E_{HOMO} and E_{LUMO}.[21] They did not optimize fluorescent strength.

Received: August 13, 2021
Accepted: September 17, 2021
Published: October 5, 2021



https://doi.org/10.1021/acsomega.1c04347

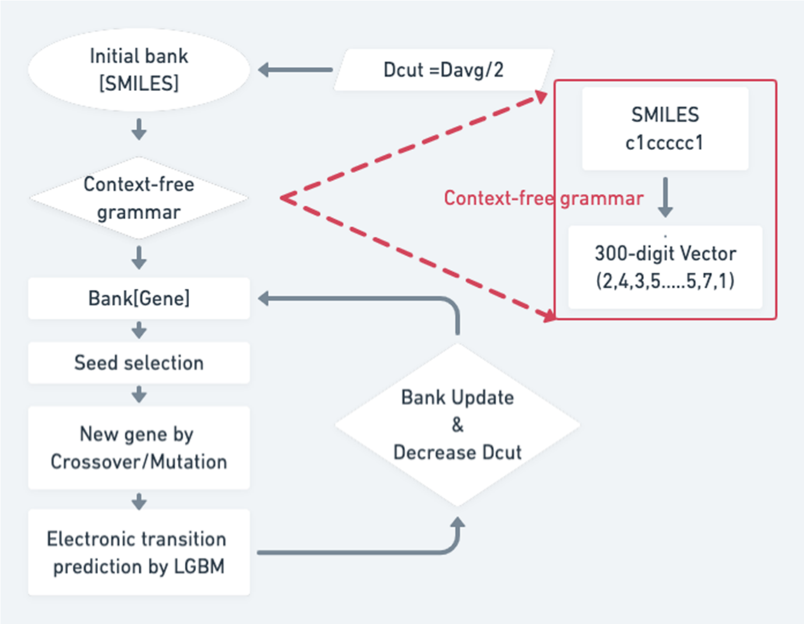


Figure 1. Flow chart of MOLGENGO in discovering novel fluorophores.

An extensive search of the chemical space is essential to discover various favorable fluorophores from unexplored areas of chemical space. For extensive searches, molecular evaluators should have fast speed, and populations should hold diversity and not be trapped in local minima at the early-stage optimization. All current methods used DFT calculations that require extensive computation resources and time as evaluators to the best of our knowledge.[19−23] Also, they did not consider the diversity of generated molecules during optimization.[19−21]

Here, we present the MOLecular generator using Light Gradient boosting machine, Grammatical Evolution aNd Global Optimization (MOLGENGO) approach to find novel molecules that have a targeted absorption wavelength, λ_{max}, and high oscillator strength, f_{max}. Our method optimizes both f_{max} and λ_{max} simultaneously, unlike previous approaches (Figure 1).[19,21] The light gradient boosting machine (LGBM)[24] method, one of the tree-based machine learning (ML) methods, was used to predict f_{max} and λ_{max}. Our ML-based predictions require much fewer computation resources than quantum mechanical (QM) calculations without sacrificing accurate characterization of electronic excitation properties.[22,24] Furthermore, we implemented the conformational space annealing (CSA) algorithm as a global optimization method that searches global minimum solutions while considering the diversity of molecules.[25−27] As a result, we observed that the faster evaluation of f_{max} and λ_{max} and consideration of diversity enabled an extensive search of fluorophores with broad chemical space coverage.

This paper is organized as follows. First, the details of the molecular descriptors and LGBM models to predict f_{max} and λ_{max} are described. Second, the components of the CSA algorithm, genetic operators for the molecular generator, and diversity control scheme are described. Third, the detailed molecular optimization results of MOLGEGO are discussed. We show that our method successfully optimized f_{max} and λ_{max} and maintained the diversity of the pool of generated molecules better than the existing genetic algorithm through the benchmark test. Finally, this article suggests novel molecules with optimized f_{max} and λ_{max} verified by time-dependent (TD)-DFT.

2. RESULTS AND DISCUSSION

2.1. Prediction of Maximum Oscillator Strength and the Corresponding Excitation Energy. LGBM models predicted the desired electronic transition properties with similar accuracy and correlation but faster training speed than the previous random forest (RF)-based models (Table 1). The mean prediction times of LGBM models were less than 1/10 of the RF method. These results demonstrate that our LGBM models facilitate global optimization efficiency due to their faster prediction time. Furthermore, the LGBM models showed higher Pearson correlation coefficients on f_{max} and maximum corresponding excitation energy (E_{max}) predictions. The root mean square error (RMSE) of E_{max} from LGBM was 0.02 lower than that of RF and the RMSE of f_{max} from LGBM was identical to that of RF.

Table 1. Comparison of Accuracy and Efficiency of LGBM and RF Models

model	quantum property	RMSE[a]	Pearson R[a]	mean prediction time (s)[b]
LGBM	excitation energy (eV)	0.43	0.89	0.006
LGBM	f_{max}	0.08	0.85	0.004
RF[7]	excitation energy (eV)	0.45	0.88	0.065
RF[7]	f_{max}	0.08	0.83	0.068

[a] The test set includes 50 000 molecules. The number of heavy atoms is up to 38. [b] Intel Xeon CPU E5-2620 v4 2.10 GHz 1 core, 1 processor, 128 GB memory.

2.2. Finding Novel Fluorescent Molecules via Global Optimization. *2.2.1. Benchmarking Optimization Performance.* MOLGENGO successfully generated molecules with high $f_{max,pred}$ and desired $\lambda_{max,pred}$ (Figure 2). MOLGENGO was executed from the identical first bank to optimize $f_{max,pred}$ and $\lambda_{max,pred}$ with three $\lambda_{max,target}$ values: 200, 400, and 600 nm.

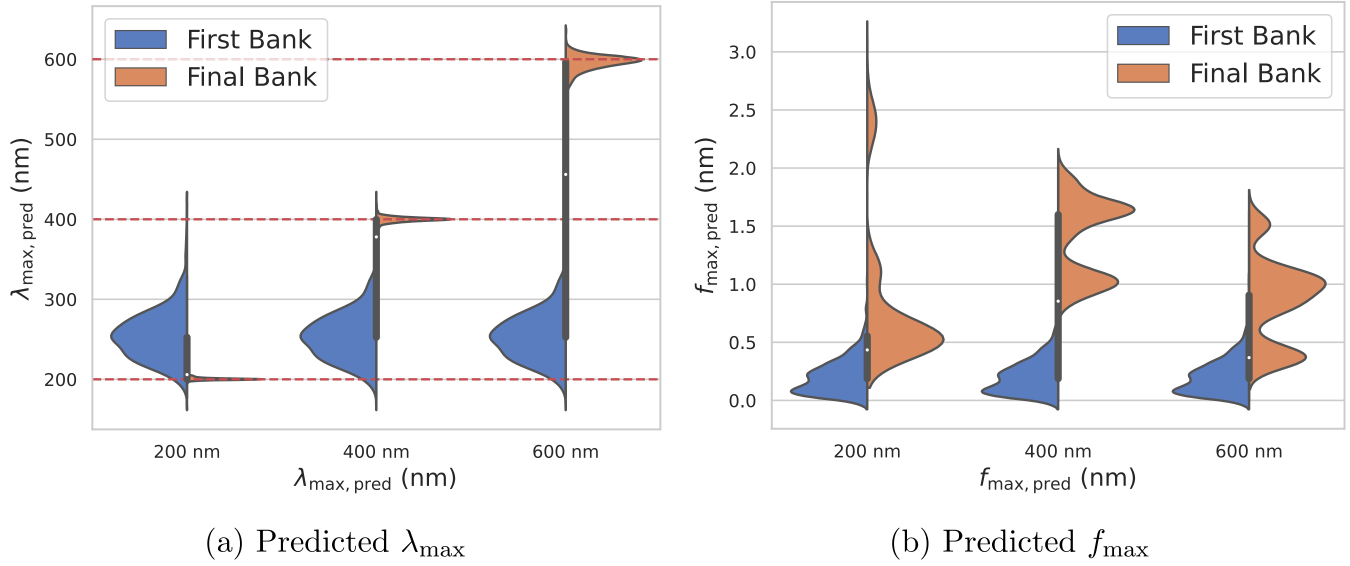


Figure 2. Distribution change of predicted (a) λ_{max} and (b) f_{max} using the vilin plot. Cyan and orange correspond to the first bank and final bank, respectively. The first plot, second plot, and third plot correspond to 200, 400, and 600 nm target λ_{max}, respectively.

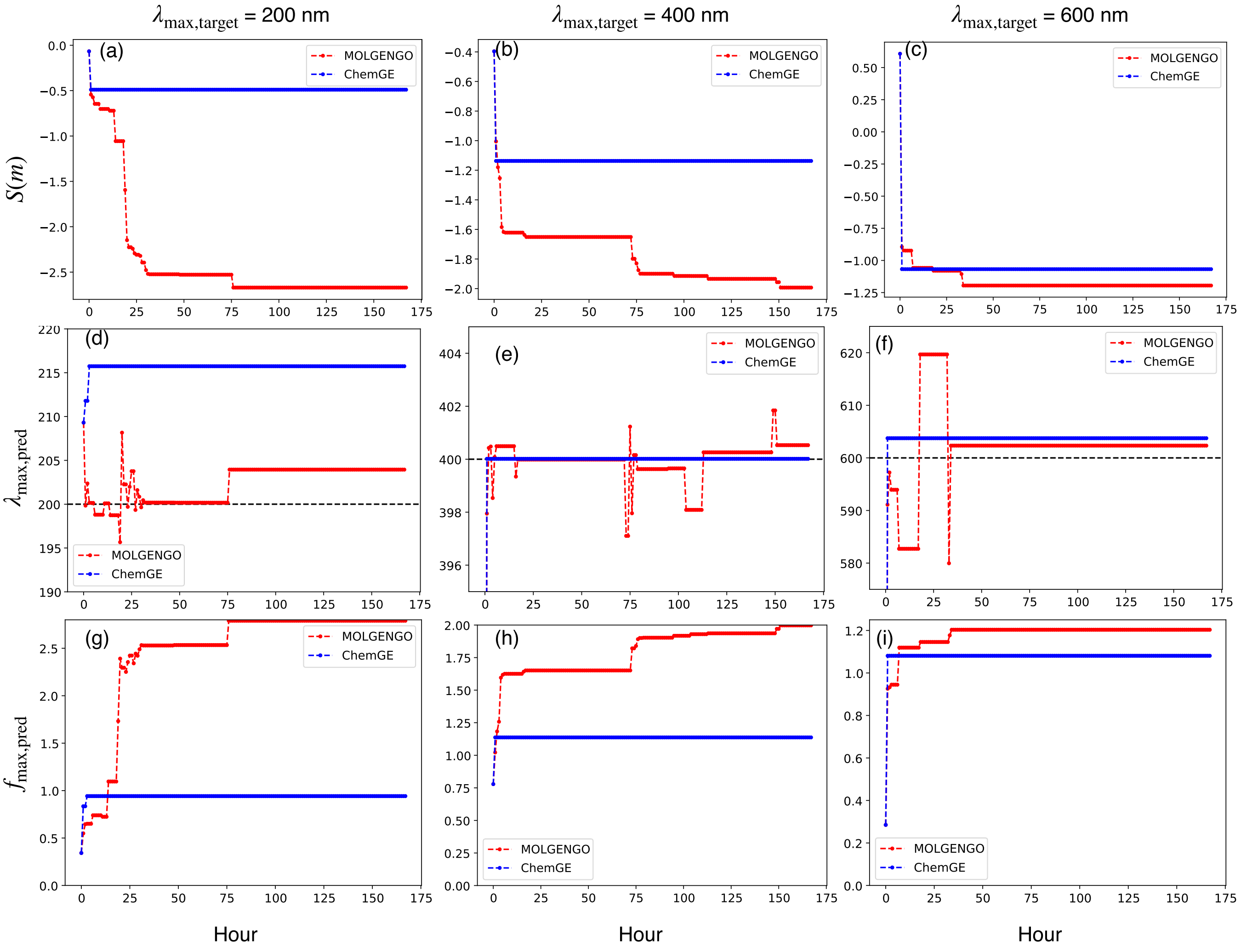
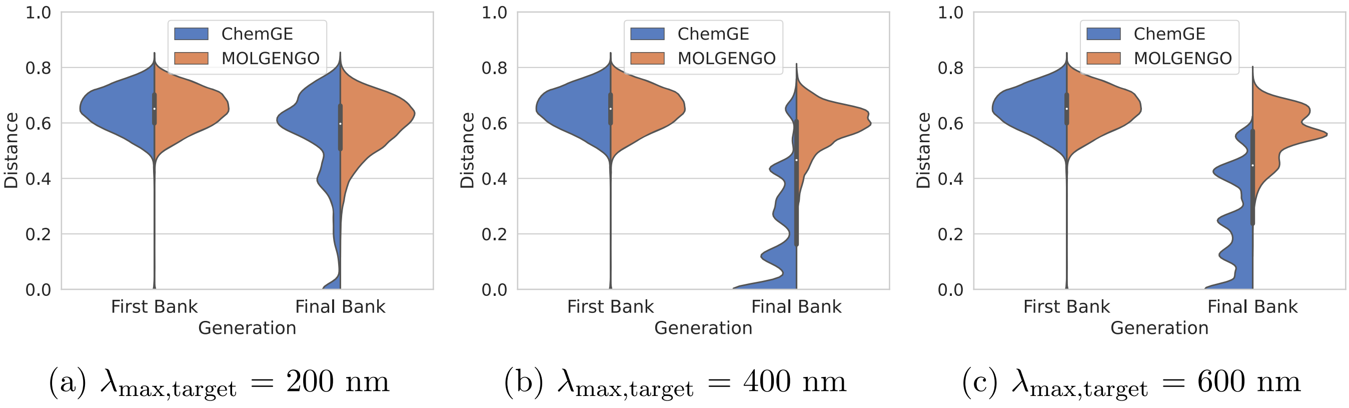


Figure 3. Average of the best $S(m)$, $\lambda_{max,pred}$ and $f_{max,pred}$ from 10 simulations by ChemGE and CSA. The change of $S(m)$ with target λ_{max} (a) 200 nm, (b) 400 nm, and (c) 600 nm. $\lambda_{max,pred}$ with target λ_{max} (d) 200 nm, (e) 400 nm, and (f) 600 nm. $f_{max,pred}$ with target λ_{max} (g) 200 nm, (h) 400 nm, and (i) 600 nm during search. The X-axis represents running time (hour). Red and blue correspond to the CSA and ChemGE, respectively.

https://doi.org/10.1021/acsomega.1c04347



(a) $\lambda_{\text{max,target}} = 200$ nm (b) $\lambda_{\text{max,target}} = 400$ nm (c) $\lambda_{\text{max,target}} = 600$ nm

Figure 4. Pairwise distance $(1 - J_c)$ distribution of MOLGENGO and ChemGE. Cyan and orange colors correspond to the ChemGE and MOLGENGO results, respectively.

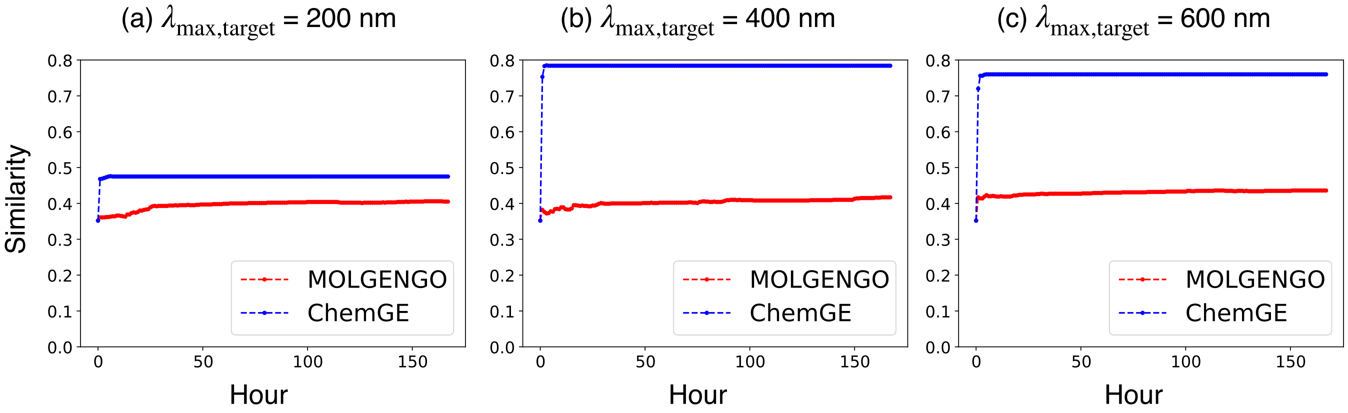


Figure 5. Change of average similarities of generated molecules measured by the Jaccard coefficient. Red and blue lines correspond to the MOLGENGO and ChemGE, respectively.

The population of the predicted λ_{max} of the first bank had its peak near 250 nm. After optimization, all predicted λ_{max} values in the final banks deviated from $\lambda_{\text{max,target}}$ less than 50 nm for all simulations (Figure 2a). The $\lambda_{\text{max,pred}}$ distributions became narrow, changing from the first banks to the final banks for all $\lambda_{\text{max,target}}$, which indicated that our method successfully generated molecules with desired properties. The width of $\lambda_{\text{max,pred}}$ distribution for $\lambda_{\text{max,target}}$ = 600 nm was broader than that of $\lambda_{\text{max,target}}$ = 200 nm. This may be due to a sparse population of molecules whose λ_{max} were close to 600 nm in PubChemQC. In PubChemQC, the number of molecules whose λ_{max} were in the range of 600 ± 10 nm is about 10^2, 1/100 of the molecules with λ_{max} = 400 ± 10 nm, and 1/10 000 of the molecules with λ_{max} = 200 ± 10 nm[7,28] (Figure S1).

For all $\lambda_{\text{max,target}}$ values, many molecules whose predicted f_{max} surpassed 1.5 were found (Figure 2b). All predicted f_{max} values of the first banks were less than 1.0 for all $\lambda_{\text{max,target}}$. When $\lambda_{\text{max,target}}$ = 200 nm, we even discovered molecules with f_{max} over 3.0. When $\lambda_{\text{max,target}}$ = 400 and 600 nm, two peaks that exceeded 1.0 were found.

For a fair comparison between the two methods tested here, we performed global optimization using both methods for 7 days with a single CPU. MOLGENGO requires a longer computation time to be converged but finds more optimized molecules than ChemGE[29] (Figure 3). ChemGE simulations converged within 3 h for all $\lambda_{\text{max,target}}$. However, MOLGENGO kept finding more optimized molecules until 168 h for all $\lambda_{\text{max,target}}$.

The objective function of our global optimization was designed to maximize the highest oscillator strength (f_{max}) and make the corresponding absorption wavelength (λ_{max}) close to the target wavelength ($\lambda_{\text{max,target}}$). The objective function of a molecule m used for in this study is defined as follows

$$S(m) = f_{\text{max,pred}}(m) - w|\lambda_{\text{max,pred}}(m)^{-1} - \lambda_{\text{max,target}}(m)^{-1}| \tag{1}$$

where $f_{\text{max,pred}}(m)$ and $\lambda_{\text{max,pred}}(m)$ are the predicted values obtained from the LGBM regressors[24] and w is the weight of the λ_{max} deviation term.

In terms of finding molecules whose $\lambda_{\text{max,pred}}$ are close to $\lambda_{\text{max,target}}$, MOLGENGO outperformed the ChemGE method significantly except the case of $\lambda_{\text{max,target}}$ = 400 nm. When $\lambda_{\text{max,target}}$ = 200 nm, the best result of MOLGENGO approached 200 nm but that of ChemGE departed from 200 nm as a simulation proceeded (Figure 3d). When $\lambda_{\text{max,target}}$ = 600 nm, MOLGENGO results converged to λ_{max} = 602 nm and ChemGE converged at 604 nm (Figure 3f). When $\lambda_{\text{max,target}}$ = 400 nm, the difference between the results of both methods was not significant, under 1.0 nm (Figure 3e).

In terms of finding molecules with high f_{max}, MOLGENGO outperformed the ChemGE method significantly for all $\lambda_{\text{max,target}}$. When $\lambda_{\text{max,target}}$ = 200 nm, $f_{\text{max,pred}}$ of the best molecule obtained with MOLGENGO was 2.8, almost 3 times that of the ChemGE result, 1.0 (Figure 3g). When $\lambda_{\text{max,target}}$ = 400 nm, the highest $f_{\text{max,pred}}$ obtained with MOLGENGO was 2.0 while that of the ChemGE result was 1.1 (Figure 3h).

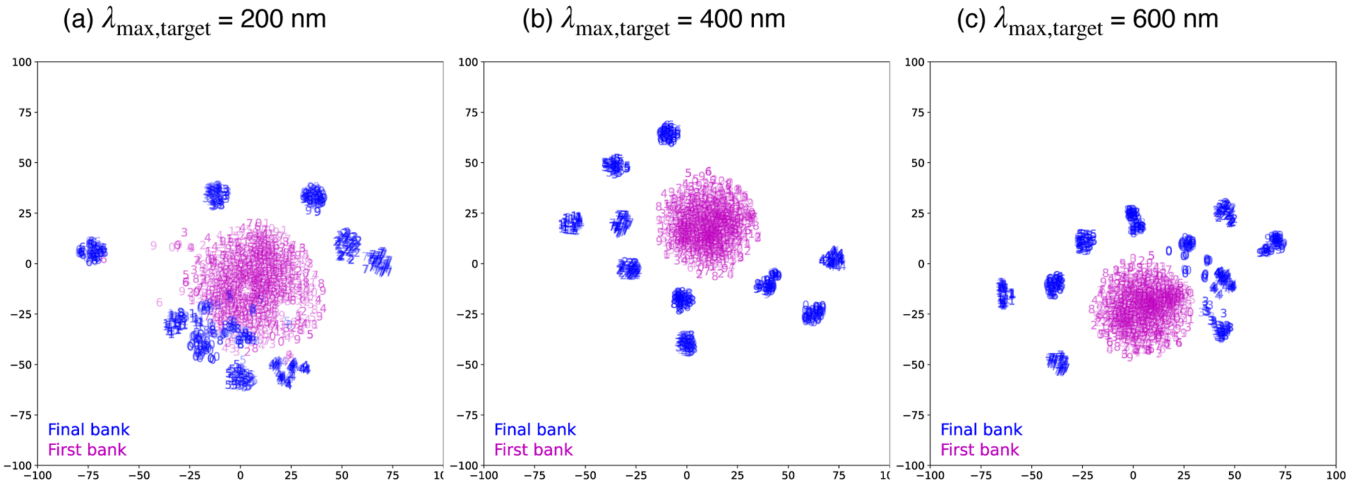


Figure 6. t-SNE visualization of molecules in first and final banks with target λ_{max} (a) 200 nm, (b) 400 nm, and (c) 600 nm. Magenta and blue digits represent molecules included in the first and final banks, respectively. Numbers mean indices of runs. A 4096-dimensional ECFP4 vector was projected onto a two-dimensional space.

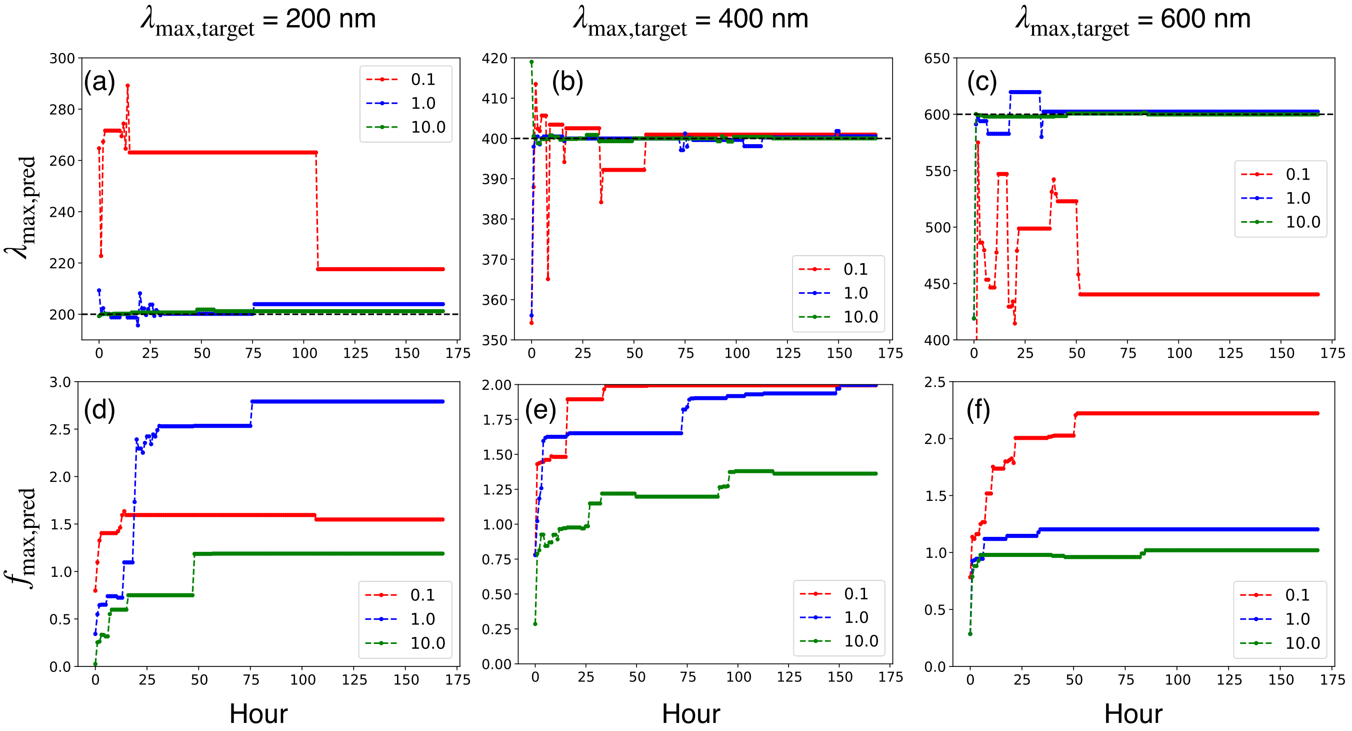


Figure 7. Electronic property optimization results with different weight coefficients. Red, blue, and green represent the results obtained with w = 0.1, 1.0, and 10.0, respectively. $\lambda_{max,pred}$ for target λ_{max} (a) 200 nm, (b) 400 nm, and (c) 600 nm during searches. $f_{max,pred}$ for target λ_{max} (d) 200 nm, (e) 400 nm, and (f) 600 nm during searches.

When $\lambda_{max,target}$ = 600 nm, the best $f_{max,pred}$ of the MOLGENGO simulation was 1.2, slightly higher than that of the ChemGE result, 1.1 (Figure 3i).

With a weight factor of 1.0, MOLGENGO showed slightly worse results in terms of λ_{max} than ChemGE (Figure 3e). However, MOLGENGO found molecules with much higher f_{max} values than ChemGE, which compensates slightly worse results of λ_{max} (Figure 3h). Because we optimized the objective function defined as the linear combination of λ_{max} and f_{max} terms, each individual component may not show consistent improvement over ChemGE. However, the overall objective values obtained with MOLGENGO are better than those of ChemGE consistently. These results show that our approach using CSA performs more extensive exploration of chemical space than ChemGE based on the conventional genetic algorithm, resulting in better molecules.

2.2.2. *Diversity of Generated Molecules.*

MOLGENGO sampled more diverse molecules than ChemGE[29] for all target λ_{max} (Figures 4 and 5). The pairwise distance distribution of MOLGENGO shows that the diversities of the pools of molecules were well-maintained in the final banks, which are the results of global optimization (Figure 4). However, the optimization by ChemGE led to lower distances between optimized molecules, indicating that they are highly similar to each other. For all $\lambda_{max,target}$, the pairwise distance distributions of the final banks obtained with ChemGE had their peaks near

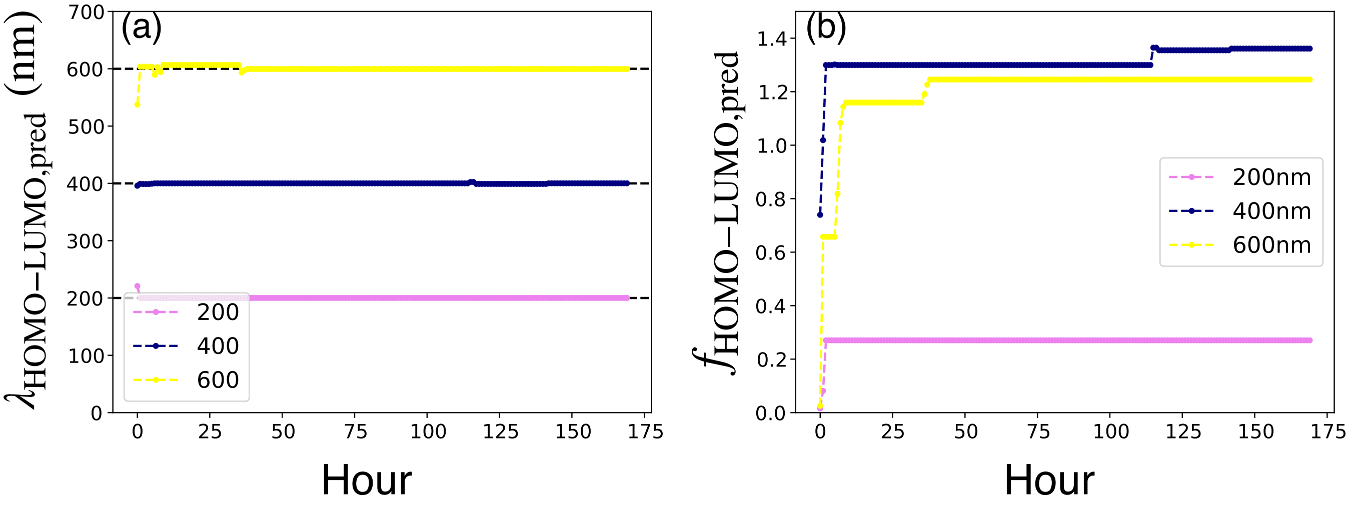
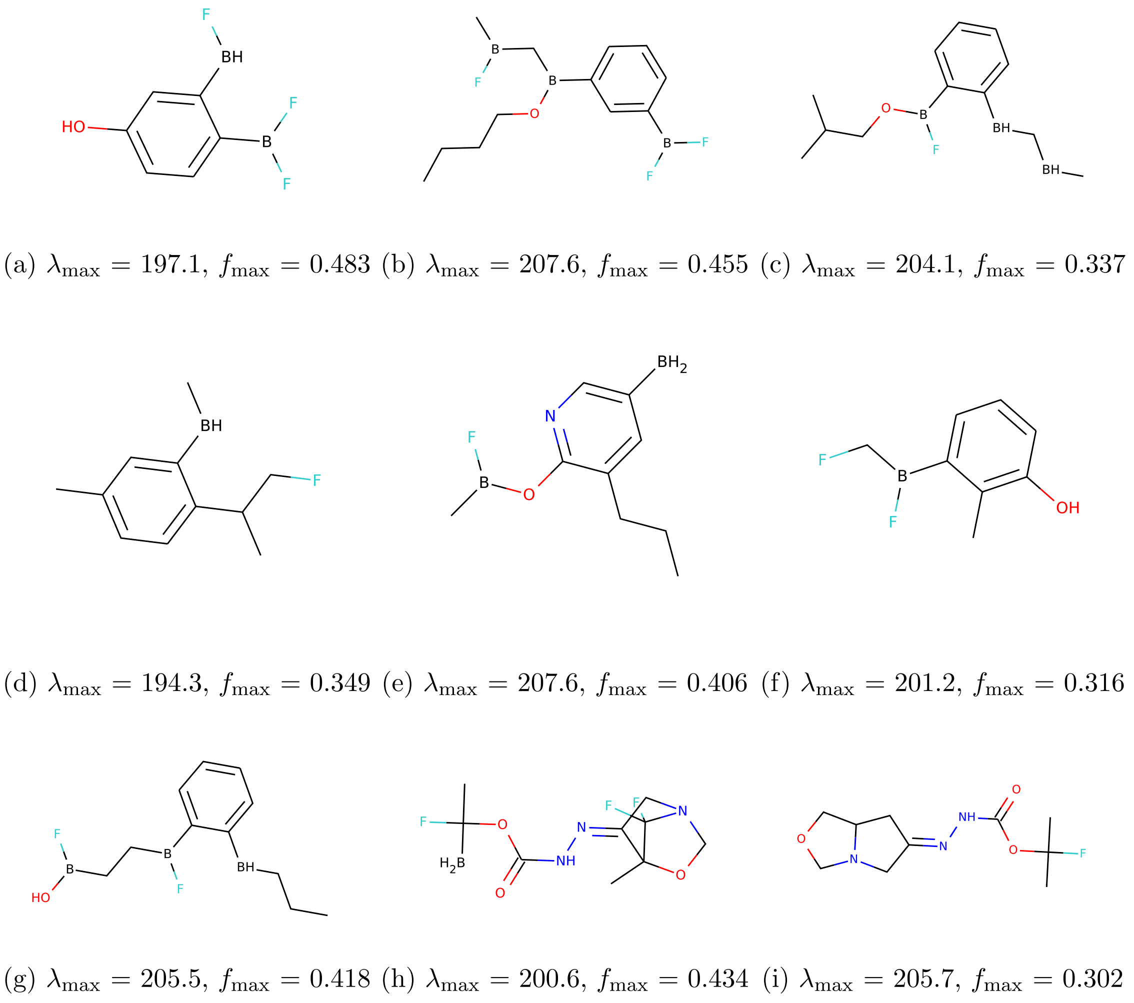


Figure 8. Optimization of (a) $\lambda_{HOMO-LUMO}$ and (b) $f_{HOMO-LUMO}$ for λ_{target} = 200 (magenta), 400 (navy), and 600 (yellow) nm.



(a) $\lambda_{max} = 197.1$, $f_{max} = 0.483$ (b) $\lambda_{max} = 207.6$, $f_{max} = 0.455$ (c) $\lambda_{max} = 204.1$, $f_{max} = 0.337$

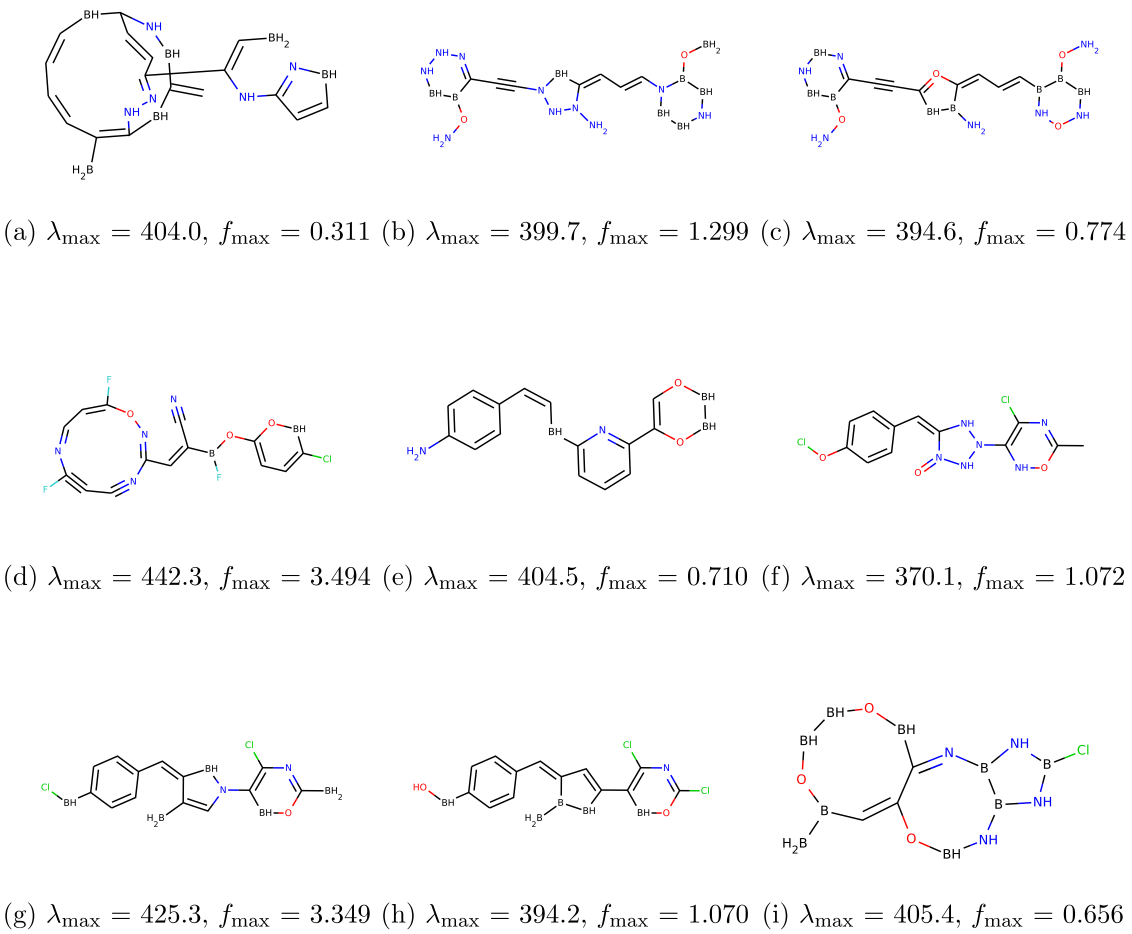
(d) $\lambda_{max} = 194.3$, $f_{max} = 0.349$ (e) $\lambda_{max} = 207.6$, $f_{max} = 0.406$ (f) $\lambda_{max} = 201.2$, $f_{max} = 0.316$

(g) $\lambda_{max} = 205.5$, $f_{max} = 0.418$ (h) $\lambda_{max} = 200.6$, $f_{max} = 0.434$ (i) $\lambda_{max} = 205.7$, $f_{max} = 0.302$

Figure 9. Novel molecules found by MOLGENGO with $\lambda_{max,target}$ = 200 nm. TD-DFT results of λ_{max} and f_{max} are represented below the molecular structures.

0.0. Furthermore, the highest peaks of pairwise distance distribution of the final banks from ChemGE were located near 0.0 when $\lambda_{max,target}$ = 400 and 600 nm. However, the MOLGENGO results formed their peaks around 0.6 for all $\lambda_{max,target}$ values.

The average similarities of banks increased and converged at an early stage, within 7 h, with ChemGE (Figure 5). In contrast, that of MOLGENGO results rose gradually until the end of simulations and saturated near 0.4 (Figure 5). When $\lambda_{max,target}$ = 400 and 600 nm, the average similarities of the final bank of ChemGE simulations were almost twice that of



(a) $\lambda_{\max} = 404.0$, $f_{\max} = 0.311$ (b) $\lambda_{\max} = 399.7$, $f_{\max} = 1.299$ (c) $\lambda_{\max} = 394.6$, $f_{\max} = 0.774$

(d) $\lambda_{\max} = 442.3$, $f_{\max} = 3.494$ (e) $\lambda_{\max} = 404.5$, $f_{\max} = 0.710$ (f) $\lambda_{\max} = 370.1$, $f_{\max} = 1.072$

(g) $\lambda_{\max} = 425.3$, $f_{\max} = 3.349$ (h) $\lambda_{\max} = 394.2$, $f_{\max} = 1.070$ (i) $\lambda_{\max} = 405.4$, $f_{\max} = 0.656$

Figure 10. Novel molecules found by MOLGENGO with $\lambda_{\max,\text{target}}$ = 400 nm. TD-DFT results of $\lambda_{\max}$ and $f_{\max}$ are represented below the molecular structures.

MOLGENGO. Preservation of diversity with MOLGENGO explains its slower convergence and broader search in the chemical space. In addition, because MOLGENGO covered broader space than ChemGE, it may have had more chances to discover compounds whose $\lambda_{\max,\text{pred}}$ was closer to $\lambda_{\max,\text{target}}$ and $f_{\max}$ higher.

2.3. Chemical Space Coverage. t-Distributed stochastic neighbor embedding (t-SNE) visualization was utilized to show how MOLGENGO searches chemical space widely (Figure 6). t-SNE is a statistical model for visualizing high-dimensional data by giving each datapoint a coordinate in a two or three-dimensional map.[30] A 4096-dimensional ECFP4 vector was projected onto two-dimensional space to deal with structural diversity. The final bank of each run formed clusters, which extended from the first banks (Figure 6). Most final banks' molecules resided outside of ZINC-250k. It shows that MOLGENGO could discover novel molecules that are not present in the initial DB.[31]

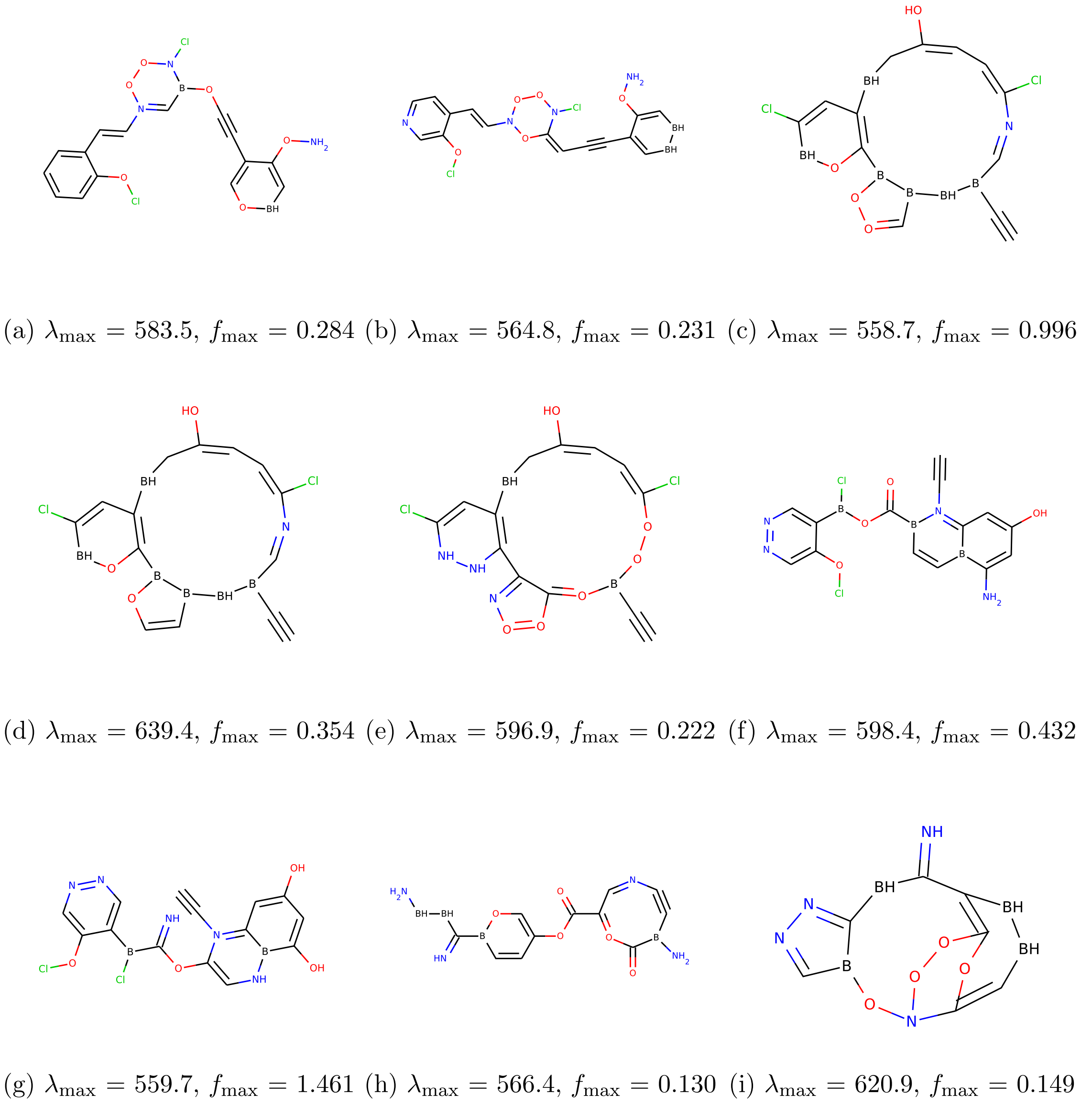
2.4. Optimization of Objective Function. As the first trial for the weight parameter of the objective function, we tried three values: 0.1, 1.0, and 10.0.

Desirable optimization results must satisfy two conditions: high $f_{\max,\text{pred}}$ and convergence to target $\lambda_{\max}$. Excitation energy optimization results with w = 0.1 did not converge to $\lambda_{\max,\text{target}}$ = 200 nm (Figure 7a) and 600 nm (Figure 7c). Both optimization results with w = 1.0 and 10.0 converged close to target $\lambda_{\max}$ for all target $\lambda_{\max}$ (Figure 7a–c). However, $f_{\max,\text{pred}}$ optimization results with w = 1.0 were higher than those of w = 10.0 for all target $\lambda_{\max}$. When $\lambda_{\max,\text{target}}$ = 200 nm, $f_{\max,\text{pred}}$ with w = 1.0 was twice of $f_{\max,\text{pred}}$ with w = 10.0 (Figure 7d). Similarly, when $\lambda_{\max,\text{target}}$ = 400 nm, $f_{\max,\text{pred}}$ with w = 1.0 was 0.7 higher than that with w = 10.0 (Figure 7e). When $\lambda_{\max,\text{target}}$ = 600 nm, $f_{\max,\text{pred}}$ with w = 1.0 was also higher than that with w = 10.0 (Figure 7f). Thus, we determined the best weight of eq 1 as 1.0 among the three tested values. This hyperparameter optimization is not extensive and a more rigorous and systematic parameter tuning is required for more accurate results.

2.5. Optimization of Highest Occupied Molecular Orbital (HOMO)–Lowest Unoccupied Molecular Orbital (LUMO) Gap and Its Oscillator Strength. In solution, HOMO–LUMO gap's wavelength ($\lambda_{\text{HOMO-LUMO}}$), and oscillator strength ($f_{\text{HOMO-LUMO}}$) play important roles in fluorescence.[32] We optimized $\lambda_{\text{HOMO-LUMO}}$ and $f_{\text{HOMO-LUMO}}$ with the same manner of $\lambda_{\max}$ and $f_{\max}$ (Figure 8). For all λ_{target}, $\lambda_{\text{HOMO-LUMO}}$ entered the stationary stage after 50 h. The convergence values were the same for all calculations with λ_{target} = 200, 400, and 600 nm. Except for λ_{target} = 200 nm, $f_{\text{HOMO-LUMO}}$ values exceeded 1.0. This also supports the possibility of discovering favorable fluorescence using MOLEGNGO.

Validation of optimized molecules is described in the Supporting Information.

2.6. Validation of Optimization Results Using TD-DFT Calculations. We executed TD-DFT calculations of molecules generated by MOLGENGO simulations and obtained



(a) $\lambda_{\max} = 583.5$, $f_{\max} = 0.284$ (b) $\lambda_{\max} = 564.8$, $f_{\max} = 0.231$ (c) $\lambda_{\max} = 558.7$, $f_{\max} = 0.996$

(d) $\lambda_{\max} = 639.4$, $f_{\max} = 0.354$ (e) $\lambda_{\max} = 596.9$, $f_{\max} = 0.222$ (f) $\lambda_{\max} = 598.4$, $f_{\max} = 0.432$

(g) $\lambda_{\max} = 559.7$, $f_{\max} = 1.461$ (h) $\lambda_{\max} = 566.4$, $f_{\max} = 0.130$ (i) $\lambda_{\max} = 620.9$, $f_{\max} = 0.149$

Figure 11. Novel molecules found by MOLGENGO with $\lambda_{\max,\text{target}}$ = 600 nm. TD-DFT results of $\lambda_{\max}$ and $f_{\max}$ are represented below the molecular structures.

their maximum oscillator strength ($f_{\max,\text{TD-DFT}}$) and its corresponding wavelength ($\lambda_{\max,\text{TD-DFT}}$) to verify the properties of novel fluorophores discovered by MOLGENGO. Quantum calculation results from PubChemQC are based on the optimized ground state geometry. Molecules in the final banks whose $f_{\max,\text{TD-DFT}}$ values exceeded 0.1 were divided into nine clusters using k-means clustering algorithm[33] and the ECFP4 of the molecules folded into 4096 bits. From each cluster, the molecule with the lowest $|\lambda_{\max,\text{TD-DFT}} - \lambda_{\max,\text{pred}}|$ was selected. In summary, 27 molecules, 9 molecules for $\lambda_{\max,\text{target}}$ = 200, 400, and 600 nm, are displayed (Figures 9−11). All 27 molecules are novel molecules, which are not in PubChem.

When $\lambda_{\max,\text{target}}$ = 200 nm, all absolute deviation between $\lambda_{\max,\text{TD-DFT}}$ and $\lambda_{\max,\text{target}}$, $|\lambda_{\max,\text{TD-DFT}} - \lambda_{\max,\text{target}}|$, were less than 8 nm and all $f_{\max}$ exceeded 0.3 (Figure 9). $|\lambda_{\max,\text{TD-DFT}} - \lambda_{\max,\text{target}}|$ of molecules for $\lambda_{\max,\text{target}}$ = 200 nm were less than those of compounds for $\lambda_{\max,\text{target}}$ = 400 and 600 nm. The number of molecules that satisfy $|\lambda_{\max,\text{TD-DFT}} - \lambda_{\max,\text{target}}| < 10$ nm was 9, 6, and 2 for 200, 400, and 600 nm, respectively. The $\lambda_{\max}$ values of molecules in PubChemQC are densely populated around 200 nm[28] (Figure S1). This may have led to the better prediction accuracy of $\lambda_{\max}$. This also may be related to better TD-DFT results of $\lambda_{\max}$ compared to $\lambda_{\max,\text{target}}$ = 400 and 600 nm.

The structures of novel molecules found with $\lambda_{\max,\text{target}}$ = 200 nm (Figure 9) are simpler and smaller than compounds with $\lambda_{\max,\text{target}}$ = 400 and 600 nm (Figures 10 and 11). Molecules with $\lambda_{\max}$ near 200 nm do not have extensive π-conjugations.[34]

When $\lambda_{\max,\text{target}}$ = 400 nm, all $|\lambda_{\max} - \lambda_{\max,\text{target}}|$ values were under 50 nm and all $f_{\max}$ values exceeded 0.3 (Figure 10). We found five molecules whose $f_{\max}$ were over 1.0 and two molecules with $f_{\max}$ exceeding 3.0 (Figure 10d,g). The molecule in Figure 10g may have high quantum yield because it may have a high $f_{\max}$ over 3.0 and a rigid structure that prevents excited-state molecular twisting.[35,36]

When $\lambda_{\max,\text{target}}$ = 600 nm, all $|\lambda_{\max} - \lambda_{\max,\text{target}}|$ values were under 50 nm and all $f_{\max}$ values were over 0.1 (Figure 11). We found one molecule whose $f_{\max}$ was over 1.0. The insufficient number of molecules whose $\lambda_{\max}$ are in a range of 600 ± 10

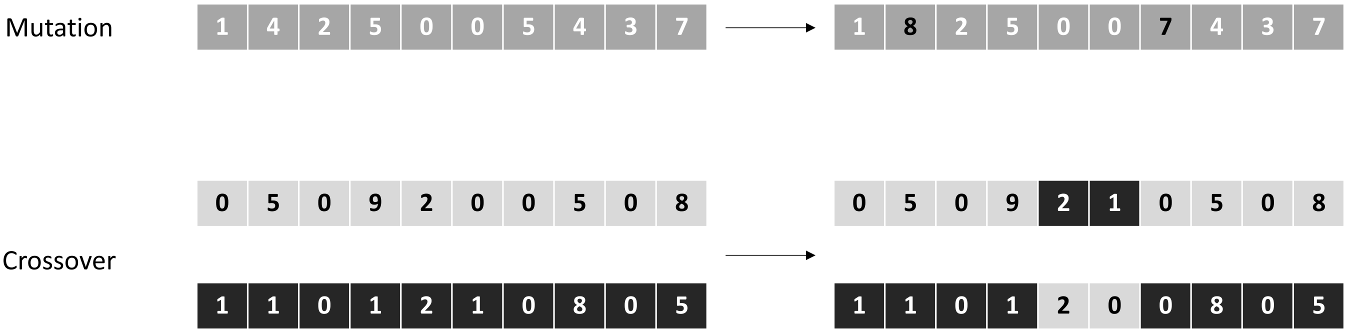


Figure 12. Genetic operators used in MOLGENGO: mutation and crossover.

nm[7,28] (Figure S1) appears to be a reason for relatively worse results to discover novel molecules with $\lambda_{max} \approx 600$ nm.

2.7. Limitation of the Current Study. We identified that some of the generated molecules appear to be hard to synthesize. To overcome this problem, two approaches will be tried in future studies. First, scores to measure the synthesizability of molecules, such as SA-score[37] or RA-score,[38] can be directly incorporated into an objective function. Second, after generating many novel molecules, they can be screened based on synthesizability and hand-crafted rules based on synthetic chemists' expertise.

Second, the accuracy of MOLGENGO is tightly coupled with the PubChemQC database. Therefore, MOLGENGO may not explore the near-infrared (NIR)-I/II region accurately, which has been drawing much attention recently.[39] The ratio of molecules whose excitation energy and oscillator strength lying in the NIR-I/II region is less than 0.1% of our training set. This severely prevents efficient and accurate prediction of the corresponding region of chemical space. Once more molecules in the NIR-I/II region are accumulated, we will be able to explore the NIR-I/II region accurately.

3. CONCLUSIONS

In this work, we developed a new molecular discovery approach by combining the global optimization method and LGBM predictors to find molecules with high oscillator strength and targeted excitation energy. Unlike previous approaches,[19−21] which used quantum calculation, we used machine learning to characterize the desired electronic transition properties. We also performed global optimization of properties on chemical space to consider the diversity. LBGM models predicted f_{max} and λ_{max} efficiently without deteriorating the accuracy of predictions. MOLGENGO successfully found novel molecules with $\lambda_{max,pred}$ = 200, 400, and 600 nm and $f_{max,pred}$ over 3.0. We identified that MOLGENGO covers a wide range of chemical space outside the existing databases' coverage. Many novel molecules with high f_{max} and desired λ_{max} were found and they were verified via TD-DFT calculations. We expect that MOLGENGO is an efficient tool for discovering novel molecules, which can be candidates for favorable fluorophores.

From the experimental results, the limitations of the current version of MOLGENGO are identified. First, the results of MOLGENGO targeting the longer absorption wavelength, 600 nm, were worse than those targeting the shorter absorption wavelengths, 200 or 400 nm. We believe that this is due to the bias of the training set, PubChemQC. The majority of molecules in PubChemQC have their absorption wavelength in a relatively short-wavelength region, i.e., shorter than 600 nm. This bias of input data appears to be the reason for relatively worse results for molecules targeting absorption wavelength in the IR region. To overcome this limitation, more information on molecules with longer absorption wavelengths is necessary. Second, many newly discovered molecules appear to be hard to synthesize. MOLGENGO performs global optimization of molecular properties in a combinatorial fashion. Thus, unlike generative models, it does not require any training or assumption on chemical structure. Instead, a global optimization approach heavily relies on an objective function and assumes that the objective function quantifies the quality of a molecule accurately. Currently, synthetic accessibility is not considered in the objective functions used in this study. Therefore, incorporating synthetic accessibility scores[37,38] will help MOLGENGO generate more synthesizable molecules.

4. METHOD

4.1. Overview of Workflow. In this study, we applied the CSA global optimization algorithm to discover novel fluorescent molecules. CSA is a powerful global optimization approach and includes components of GA.[25,40−42] The flow chart of the MOLGENGO is illustrated in Figure 1.

The inputs of the algorithm were the simplified molecular input line entry system (SMILES) format.[43] SMILES represents molecules as strings.[43] However, the string type is not efficient to be handled with genetic operators because the grammatical rules of SMILES are so complex that SMILES-based genetic operators easily generate many invalid molecules.[44] Thus, we converted SMILES strings to integer arrays using simple grammatical rules used in the context-free grammar method[29,45] (Figure 1). The example of converting from SMILES string to integer array using context-free grammar is introduced in the Supporting Information. Integer-based genetic operations allow larger changes of molecules compared to string-based genetic operations.[29]

New gene populations, molecules, were generated at each generation from the initial bank to the final bank using genetic operations. A bank includes a fixed number of genes. Two types of genetic operators were used to create diverse genes. One is mutation type and the other is crossover type (Figure 12).

One of the crucial features of MOLGENGO is that it keeps the diversity of its population by (1) defining a distance measure and setting the criterion of the similarity between two genes as D_{cut} and (2) using D_{cut} to control the diversity of the bank while D_{cut} is slowly decreased from the first value to the final value.[25] Details of the CSA method in MOLGENGO will be described in the CSA Algorithm section.

Our method only covered molecules that only include H, B, C, N, O, F, and Cl atoms with no net charge.[7]

4.2. Data Set. The PubChemQC database was used to train LGBM regressors.[28] We randomly sampled 0.5 million molecules from PubChemQC. The data set was split into a 9:1 ratio to generate the training and test sets.[7] The first banks were selected from the ZINC-250k set, which was first compiled by Kusner and co-workers and consists of 0.25 million random molecules ZINC.[31,46] ZINC-250k was used to sample starting molecules in other studies: ChemGE and GB-GA.[20,29,47]

4.3. Descriptors and Prediction Models for Objective Function. We applied the LGBM[24] algorithm to predict $f_{\max}$ and its corresponding $\lambda_{\max}$. Our previous research developed random forest (RF) machines to predict a given molecule's maximum oscillator strength and the corresponding excitation energy.[7] However, a more efficient prediction method with comparable accuracy was necessary to perform an extensive search with CSA. To satisfy this requirement, we trained the prediction models using LGBM.[24]

To convert a molecular feature vector, we utilized three descriptors, extended connectivity fingerprint with a diameter of 4 (ECFP4),[48] MACCS keys,[49] and RDKit molecular properties.[50] ECFP is a circular fingerprint for molecular characterization that accounts for the relationships between the molecular substructure efficiently. The MACCS keys were one-hot encoded fingerprints that describe the 166 crucial molecular substructures. Implemented RDKit molecular descriptors contain the real values of molecular features such as molecular weight, charge, and many more. The list of used RDKit molecular descriptors is described in the Supporting Information (RDkit Molecular Descriptors used for LGBM training section in the Supporting Information). In summary, 4301-dimensional vectors were used as input features. The vector contained 4096 bits of ECFP4, 166 MACCS keys, and 39 RDKit molecular descriptors. The number of the estimator was 1000, the number of data points per leaf node was 50, and the feature fraction was 1/3 in our LGBM model.

4.4. CSA Algorithm. The bank size of CSA was set to 100. SMILES representations were converted to 300-dimensional integer vectors using grammatical evolution (GE).[29] Thus, each bank contains 100 integer vectors. The number of vectors in a bank was maintained identical throughout the sampling.

Here, we aimed to optimize $f_{\max}$ and its corresponding $\lambda_{\max}$. Ten MOLGENGO runs were performed for three target $\lambda_{\max}$ values: 200, 400, and 600 nm. Simulations were performed for 7 days to guarantee their convergence to target wavelength using a machine with Intel Xeon CPU E5-2650 v4 (2.20 GHz 1 core, 1 processor, 128 GB memory). We also performed 10 ChemGE[29] simulations to compare performance with MOLGENGO.

4.4.1. New Gene Generation. To generate new genes, we randomly selected 50 seed genes from a current bank. Then, we implemented three operators, one mutation and two crossovers, to create new genes.[25] As a result, 30 chemically valid children's genes were generated from each seed gene, 10 by mutation, 10 by crossover1, and 10 by crossover2 (Figure 12).

The mutation operator mutated up to three randomly selected variables of each seed. The crossover1 operator performed a crossover between a seed gene and a randomly selected gene from the current bank. The size of the crossover did not exceed half of the total number of variables. The crossover2 operator performed a crossover between a seed and a randomly selected gene from the first bank. The size of the crossover did not exceed 20% of the total number of variables.

4.4.2. Bank Update and D_{cut} Control. All children genes were used to update the bank one at a time as follows. For each child gene G_{child}, its distances to all of the bank solutions were measured to identify its closest neighbor gene G_{closest}. If the distance between them, $d(G_{\text{child}}, G_{\text{closest}})$, was less than or equal to D_{cut} and the objective of G_{child}, $S(G_{\text{child}})$, was more optimized than $S(G_{\text{closest}})$, G_{child} replaced G_{closest}.

A distance between two genes was defined by $1 - J_c$. J_c is the Jaccard coefficient, also known as Tanimoto similarity, between two genes. Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the sample sets. If $d(G_{\text{child}}, G_{\text{cloeset}}) > D_{\text{cut}}$ and $S(G_{\text{child}})$ was more optimized than the worst optimized gene (G_{worst}) in the current bank, G_{child} replaced G_{worst}. Otherwise, G_{child} was abandoned.

At each generation, D_{cut} was reduced by R_{D}. At the first bank, D_{cut} started as $D_{\text{cut,init}} = D_{\text{mean,init}}/2$ where $D_{\text{mean,init}}$ is the mean distance among the first bank genes. After each CSA iteration, D_{cut} was decreased by multiplying it with R_{D} = 0.999995945357139. D_{cut} became $D_{\text{mean,init}}/3$ after 100 000 generations.

In summary, eq 2 represents the D_{cut} of the nth generation

$$D_{\text{cut}}(n) = R_{\text{D}}^{n-1} D_{\text{avg,init}} \tag{2}$$

The value of R_{D} controls the annealing schedule of CSA. D_{cut} played the role of the temperature in conventional simulated annealing.

4.5. TD-DFT Calculations of the Generated Molecules. Quantum mechanical (QM) calculations were executed to verify the desired electronic transition properties of the designed molecules. We utilized the TeraChem program, whose computational efficiency is accelerated by GPU.[51,52] Because our LGBM models were trained with PubChemQC, QM calculations followed the procedure of the PubChemQC paper.[28] Density functional theory (DFT) calculation with the B3LYP functional and the 6-31G* basis set was operated to optimize the geometries of molecules in the ground state.[53] Next, we applied time-dependent-density functional theory (TD-DFT) calculations with B3LYP/6-31+* to predict up to 10 excitation levels and identified $f_{\max}$ and corresponding $\lambda_{\max}$.[54] We used the VWN5 correlation for B3LYP to be compatible with GAMESS, which was used in PubChemQC.[28,51,55]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c04347.

Example of converting from SMILES string to integer array using context-free grammar, RDkit molecular descriptors used for LGBM training, excitation energy distribution of PubChemQC, and validation of HOMO−LUMO gap optimization results using TD-DFT calculations (PDF)

AUTHOR INFORMATION

Corresponding Authors

Chaok Seok − *Department of Chemistry, Seoul National University, 08826 Seoul, Republic of Korea;* orcid.org/0000-0002-1419-9888; Email: chaok@snu.ac.kr

Juyong Lee − *Department of Chemistry, Division of Chemistry and Biochemistry, Kangwon National University, 24341 Chuncheon, Republic of Korea;* orcid.org/0000-0003-1174-4358; Email: juyong.lee@kangwon.ac.kr

Author

Beomchang Kang − *Department of Chemistry, Seoul National University, 08826 Seoul, Republic of Korea;* orcid.org/0000-0003-3665-1086

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c04347

Notes

The authors declare no competing financial interest.
The MOLGENGO program and the zinc-250k data file are available free of charge at our homepage (http://galaxy.seoklab.org/suppl/molgengo.html).

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2016M3C4A7952630, NRF-2019M3E5D4066898 and NRF-2020M3A9G7103933). B.K. was supported by the BK21Plus Program funded by the Ministry of Education, Republic of Korea (21A20131312240). This work was supported by the Korea Environment Industry & Technology Institute (KEITI) through the Technology Development Project for Safety Management of Household Chemical Products, funded by the Korea Ministry of Environment (MOE) (KEITI:2020002960002 and NTIS:1485017120).

REFERENCES

(1) Murphy, K. R.; Stedmon, C. A.; Wenig, P.; Bro, R. OpenFluor−an online spectral library of auto-fluorescence by organic compounds in the environment. *Anal. Methods* **2014**, *6*, 658−661.

(2) Cartlidge, E. The light fantastic. *Science* **2018**, *359*, 382−385.

(3) Zinchuk, V.; Grossenbacher-Zinchuk, O. Recent advances in quantitative colocalization analysis: Focus on neuroscience. *Prog. Histochem. Cytochem.* **2009**, *44*, 125−172.

(4) Evanko, D. A. 'flaky' but useful fluorophore. *Nat. Methods* **2005**, *2*, 160−161.

(5) Martin, S. F.; Tatham, M. H.; Hay, R. T.; Samuel, I. D. Quantitative analysis of multi-protein interactions using FRET: Application to the SUMO pathway. *Protein Sci.* **2008**, *17*, 777−784.

(6) Moczko, E.; Mirkes, E. M.; Cáceres, C.; Gorban, A. N.; Piletsky, S. Fluorescence-based assay as a new screening tool for toxic chemicals. *Sci. Rep.* **2016**, *6*, No. 33922.

(7) Kang, B.; Seok, C.; Lee, J. Prediction of Molecular Electronic Transitions Using Random Forests. *J. Chem. Inf. Model.* **2020**, *60*, 5984−5994.

(8) Kim, E.; Lee, Y.; Lee, S.; Park, S. B. Discovery, Understanding, and Bioapplication of Organic Fluorophore: A Case Study with an Indolizine-Based Novel Fluorophore, Seoul-Fluor. *Acc. Chem. Res.* **2015**, *48*, 538−547.

(9) Fahrni, C. J. Biological applications of X-ray fluorescence microscopy: exploring the subcellular topography and speciation of transition metals. *Curr. Opin. Chem. Biol.* **2007**, *11*, 121−127.

(10) Fahrni, C. J. Fluorescent Probes and Labels for Cellular Imaging. *CHIMIA Int. J. Chem.* **2009**, *63*, 714−720.

(11) Morgan, M. T.; McCallum, A. M.; Fahrni, C. J. Rational design of a water-soluble, lipid-compatible fluorescent probe for Cu(i) with sub-part-per-trillion sensitivity. *Chem. Sci.* **2016**, *7*, 1468−1473.

(12) Kobayashi, H.; Ogawa, M.; Alford, R.; Choyke, P. L.; Urano, Y. New Strategies for Fluorescent Probe Design in Medical Diagnostic Imaging. *Chem. Rev.* **2010**, *110*, 2620−2640.

(13) Loudet, A.; Burgess, K. BODIPY Dyes and Their Derivatives: Syntheses and Spectroscopic Properties. *Chem. Rev.* **2007**, *107*, 4891−4932.

(14) Mishra, A.; Behera, R. K.; Behera, P. K.; Mishra, B. K.; Behera, G. B. Cyanines during the 1990s: A Review. *Chem. Rev.* **2000**, *100*, 1973−2012.

(15) Swanson, L.; Kuypers, H. A direct projection from the ventromedial nucleus and retrochiasmatic area of the hypothalamus to the medulla and spinal cord of the rat. *Neurosci. Lett.* **1980**, *17*, 307−312.

(16) Stefanachi, A.; Leonetti, F.; Pisani, L.; Catto, M.; Carotti, A. Coumarin: A Natural, Privileged and Versatile Scaffold for Bioactive Compounds. *Molecules* **2018**, *23*, No. 250.

(17) Kubin, R.; Fletcher, A. Fluorescence quantum yields of some rhodamine dyes. *J. Lumin.* **1982**, *27*, 455−462.

(18) Song, H.-O.; Lee, B.; Bhusal, R. P.; Park, B.; Yu, K.; Chong, C.-K.; Cho, P.; Kim, S. Y.; Kim, H. S.; Park, H. Development of a Novel Fluorophore for Real-Time Biomonitoring System. *PLoS One* **2012**, *7*, No. e48459.

(19) Sumita, M.; Yang, X.; Ishihara, S.; Tamura, R.; Tsuda, K. Hunting for Organic Molecules with Artificial Intelligence: Molecules Optimized for Desired Excitation Energies. *ACS Cent. Sci.* **2018**, *4*, 1126−1133.

(20) Henault, E. S.; Rasmussen, M. H.; Jensen, J. H. Chemical space exploration: how genetic algorithms find the needle in the Haystack. *PeerJ Phys. Chem.* **2020**, *2*, No. e11.

(21) Leguy, J.; Cauchy, T.; Glavatskikh, M.; Duval, B.; Mota, B. D. EvoMol: a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation. *J. Cheminf.* **2020**, *12*, No. 55.

(22) Dral, P. O.; Barbatti, M. Molecular excited states through a machine learning lens. *Nat. Rev. Chem.* **2021**, 388.

(23) Grimme, S.; Bannwarth, C. Ultra-fast computation of electronic spectra for large systems by tight-binding based simplified Tamm-Dancoff approximation (sTDA-xTB). *J. Chem. Phys.* **2016**, *145*, No. 054103.

(24) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. In *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, Advances in Neural Information Processing Systems, 2017.

(25) Joung, I.; Kim, J. Y.; Gross, S. P.; Joo, K.; Lee, J. Conformational Space Annealing explained: A general optimization algorithm, with diverse applications. *Comput. Phys. Commun.* **2018**, *223*, 28−33.

(26) Shin, W.-H.; Kim, J.-K.; Kim, D.-S.; Seok, C. GalaxyDock2: Protein-ligand docking using beta-complex and global optimization. *J. Comput. Chem.* **2013**, *34*, 2647−2656.

(27) Floudas, C. A.; Gounaris, C. E. A review of recent advances in global optimization. *J. Global Optim.* **2009**, *45*, 3−38.

(28) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **2017**, *57*, 1300−1308.

(29) Yoshikawa, N.; Terayama, K.; Sumita, M.; Homma, T.; Oono, K.; Tsuda, K. Population-based De Novo Molecule Generation, Using Grammatical Evolution. *Chem. Lett.* **2018**, *47*, 1431−1434.

(30) Pezzotti, N.; Thijssen, J.; Mordvintsev, A.; Hollt, T.; Lew, B. V.; Lelieveldt, B. P.; Eisemann, E.; Vilanova, A. GPGPU Linear Complexity t-SNE Optimization. *IEEE Trans. Visualization Comput. Graphics* **2020**, *26*, 1172−1181.

(31) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. In *Grammar Variational Autoencoder*, Proceedings of the 34th International Conference on Machine Learning, 2017; pp 1945−1954.

(32) Valeur, B.; Berberan-Santos, M. *Molecular Fluorescence: Principles and Applications*; Wiley, 2012.

(33) Pelleg, D.; Moore, A. In *Accelerating Exact k-Means Algorithms with Geometric Reasoning*, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '99, 1999.

(34) Yin, N.; Wang, L.; Lin, Y.; Yi, J.; Yan, L.; Dou, J.; Yang, H.-B.; Zhao, X.; Ma, C.-Q. Effect of the π-conjugation length on the properties and photovoltaic performance of A−π−D−π−A type oligothiophenes with a 4, 8-bis (thienyl) benzo [1, 2-b: 4, 5-b] dithiophene core. *Beilstein J. Org. Chem.* **2016**, *12*, 1788−1797.

(35) Sun, F.; Jin, R. DFT and TD-DFT study on the optical and electronic properties of derivatives of 1,4-bis(2-substituted-1,3,4-oxadiazole)benzene. *Arabian J. Chem.* **2017**, *10*, S2988−S2993.

(36) Suhina, T.; Amirjalayer, S.; Mennucci, B.; Woutersen, S.; Hilbers, M.; Bonn, D.; Brouwer, A. M. Excited-State Decay Pathways of Molecular Rotors: Twisted Intermediate or Conical Intersection? *J. Phys. Chem. Lett.* **2016**, *7*, 4285−4290.

(37) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, No. 8.

(38) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic accessibility score (RAscore) − rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **2021**, *12*, 3339−3349.

(39) Schnermann, M. J. Organic dyes for deep bioimaging. *Nature* **2017**, *551*, 176−177.

(40) Lee, J.; Scheraga, H. A.; Rackovsky, S. New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *J. Comput. Chem.* **1997**, *18*, 1222−1232.

(41) Lee, J.; Lee, I.-H.; Joung, I.; Lee, J.; Brooks, B. R. Finding multiple reaction pathways via global optimization of action. *Nat. Commun.* **2017**, *8*, No. 15443.

(42) Lee, J.; Zhang, Z.-Y.; Lee, J.; Brooks, B. R.; Ahn, Y.-Y. Inverse Resolution Limit of Partition Density and Detecting Overlapping Communities by Link-Surprise. *Sci. Rep.* **2017**, *7*, No. 12399.

(43) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(44) Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096−1108.

(45) Dempsey, I.; O'Neill, M.; Brabazon, A. *Foundations in Grammatical Evolution for Dynamic Environments*; Springer, 2009; Vol. *194*.

(46) Maziarka, Ł.; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Danel, T.; Warchoł, M. Mol-CycleGAN: a generative model for molecular optimization. *J. Cheminf.* **2020**, *12*, No. 2.

(47) Sterling, T.; Irwin, J. J. ZINC 15 − Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.

(48) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(49) 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.

(50) Landrum, G. RDKit: Open-Source Cheminformatics Software, 2016.

(51) Ufimtsev, I. S.; Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5*, 2619−2628.

(52) Titov, A. V.; Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. Generating Efficient Quantum Chemistry Codes for Novel Architectures. *J. Chem. Theory Comput.* **2013**, *9*, 213−221.

(53) Kästner, J.; Carr, J. M.; Keal, T. W.; Thiel, W.; Wander, A.; Sherwood, P. DL-FIND: An Open-Source Geometry Optimizer for Atomistic Simulations†. *J. Phys. Chem. A* **2009**, *113*, 11856−11865.

(54) Isborn, C. M.; Luehr, N.; Ufimtsev, I. S.; Martínez, T. J. Excited-State Electronic Structure with Configuration Interaction Singles and Tamm−Dancoff Time-Dependent Density Functional Theory on Graphical Processing Units. *J. Chem. Theory Comput.* **2011**, *7*, 1814−1823.

(55) Barca, G. M. J.; Bertoni, C.; Carrington, L.; Datta, D.; De Silva, N.; Deustua, J. E.; Fedorov, D. G.; Gour, J. R.; Gunina, A. O.; Guidez, E.; Harville, T.; Irle, S.; Ivanic, J.; Kowalski, K.; Leang, S. S.; Li, H.; Li, W.; Lutz, J. J.; Magoulas, I.; Mato, J.; Mironov, V.; Nakata, H.; Pham, B. Q.; Piecuch, P.; Poole, D.; Pruitt, S. R.; Rendell, A. P.; Roskop, L. B.; Ruedenberg, K.; Sattasathuchana, T.; Schmidt, M. W.; Shen, J.; Slipchenko, L.; Sosonkina, M.; Sundriyal, V.; Tiwari, A.; Galvez Vallejo, J. L.; Westheimer, B.; Wloch, M.; Xu, P.; Zahariev, F.; Gordon, M. S. Recent developments in the general atomic and molecular electronic structure system. *J. Chem. Phys.* **2020**, *152*, No. 154102.