ACS OMEGA

Open Access

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

Article

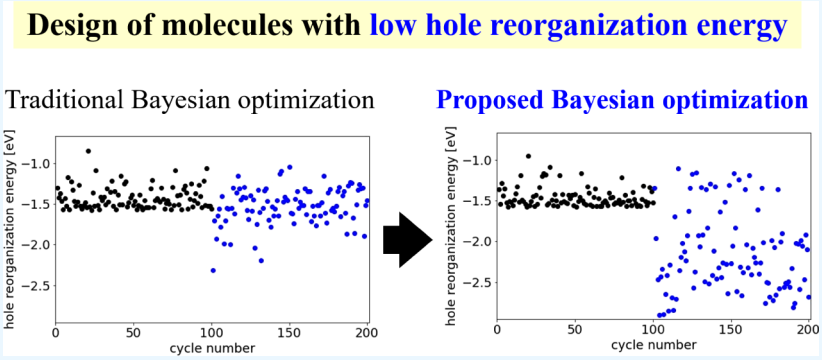
Exploring Molecular Descriptors and Acquisition Functions in Bayesian Optimization for Designing Molecules with Low Hole Reorganization Energy

Rinta Kawagoe, Tatsuhito Ando,*,§ Nobuyuki N. Matsuzawa, Hiroyuki Maeshima, and Hiromasa Kaneko*,§

Cite This: *ACS Omega* 2024, 9, 48844−48854

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: Organic semiconductors have been widely studied owing to their potential applications in various devices, such as field-effect transistors, light-emitting diodes, solar cells, and image sensors. However, they have a limitation of significantly lower carrier mobility compared to silicon, which is a widely used inorganic semiconductor. Therefore, to address such limitations, these molecules should be further explored. Hole reorganization energy has been known to influence carrier mobility; that is, lower energy results in higher mobility. This study uses Bayesian optimization (BO) to identify molecules with low hole reorganization energies. While several acquisition functions (AFs), including probability of improvement, expected improvement, and mutual information, have been proposed for use in BO, it is well established that the performance of AFs can vary depending on the data set. We evaluate the performance of AFs applied to a data set of organic semiconductor molecules and propose a novel approach that alternates the use of AFs in the BO process. Our findings conclude that alternating AFs in BO enhance the stability of the search for molecules with low reorganization energy.

INTRODUCTION

Organic semiconductors have been intensively investigated because of their potential applications in various devices, such as field-effect transistors, light-emitting diodes, solar cells, and image sensors.[1−5] Using organic materials for semiconductors instead of traditional inorganic materials offers various advantages. These include simple processing, low-cost and large-area manufacturing, lightweight structures, mechanical flexibility, and versatile molecular design, owing to their vast chemical design space. Carrier mobility is an important characteristic of semiconductors, and significant efforts have been made to obtain organic materials with enhanced mobilities. For this purpose, various classes of molecules, such as oligoacenes,[6] polythiophenes,[7] heteroarenes,[8] and fullerenes,[9] have been investigated. To this date, the highest electron and hole mobilities reported are 11 $cm^2/(V\ s)$ for C_{60}[10] and 40 $cm^2/(V\ s)$ for rubrene,[11] respectively. However, these mobility values are still significantly lower than those of silicon, which has mobilities of 1200 $cm^2/(V\ s)$ for electrons and 500 $cm^2(V\ s)$ for holes.[12] Therefore, there is a need to explore molecules with even higher carrier mobilities.

The carrier transport phenomena of molecules can be described based on the Marcus theory,[13−16] where the charge-hopping rate of the local dimer of molecules, k, is formulated by eq 1.

Received: October 7, 2024
Revised: November 19, 2024
Accepted: November 22, 2024
Published: November 27, 2024

$$k = \frac{2\pi}{\hbar}\left(\frac{H_{ab}^2}{\sqrt{4\pi\lambda k_B T}}\right)\exp\left(-\frac{(\Delta G + \lambda)^2}{4\lambda k_B T}\right) \quad (1)$$

where $\hbar$ is Planck's constant divided by 2π, k_B is Boltzmann's constant, T is the temperature, ΔG is the free energy difference for charge transfer, λ is reorganization energy, and H_{ab} is the intermolecular electronic coupling. The four parameters H_{ab}, ΔG, λ, and T in this equation determine the charge transfer rate. Among them, λ has a considerable effect on charge transfer as it decreases the transfer rate exponentially.

Multiple studies have explored molecules with low reorganization energies to identify molecules with enhanced carrier mobilities. For example, Schober et al.[17] calculated the reorganization energy and intermolecular electronic coupling of 95,445 molecular crystal structures obtained from the Cambridge Structural Database.[18] They successfully extracted four promising molecules with reorganization energies ranging from 0.121 to 0.169 eV.[17] A comprehensive theoretical screening of hole-conducting heteroacene molecules has also been conducted, where the hole reorganization energies of a quarter million heteroacenes were calculated using a cloud-computing environment.[19] This screening led to the identification of various promising structures.

Using the aforementioned quarter-million values of hole reorganization energy, we evaluate the effect of acquisition functions (AFs) in Bayesian optimization (BO)[20,21] on exploring molecules with enhanced reorganization energy. A Gaussian process regression (GPR)[22] model was constructed using an initial data set of molecules. The candidate molecules were searched based on the predictions of the objective variable y and its variance using the constructed GPR model. While using a basic regression model $y = f(x)$ to identify candidates with enhanced predicted y values may lead to locally optimal solutions, BO mitigates this risk by exploring the extrapolation and interpolation of x.

In BO, an AF is calculated using the predicted y values and their variance, and the next candidates are selected to maximize the value of the function. The probabilities of improvement (PI),[23] expected improvement (EI),[23] and mutual information (MI)[24] are commonly used AFs. However, it is important to note that no single AF is universally effective across different optimization tasks. Additionally, the calculation of x requires the application of molecular descriptors, such as RDKit,[25] Mordred,[26] and Morgan fingerprints.[27] Similarly, there is no universal set of molecular descriptors that works for every scenario.

In this study, we examine the effect of AFs and molecular descriptors on molecular design aimed at minimizing the hole reorganization energy using BO. Our study uses a data set containing a quarter million values of hole reorganization energy obtained from density functional theory (DFT) calculations.[28] We applied BO to identify molecules with low hole reorganization energy in the data set. The process began with a limited data set, where some molecules exhibiting high hole reorganization energy in the quarter-million data set were selected. Based on our findings, we propose a method that combines multiple AFs in BO to efficiently explore a diverse design space for enhanced molecular designs.

METHOD

Data Set. We utilized the hole reorganization energies of a quarter-million heteroacenes obtained from DFT calculations.[28] The acene structures consist of 2−8 fused rings, incorporating hydrogen, carbon, oxygen, sulfur, and selenium elements in various positions according to combinatorial rules. The building blocks of the fused ring structures were benzene, acenaphthylene, pyracyclene, pyrene, perylene, cyclopentadiene, furan, thiophene, and selenophene.

The hole reorganization energies in the data set were in the range of 0.0548−0.490 eV, and their distribution on a logarithmic scale (base: 10) is shown in Figure 1, where the energy values span from −1.26 eV to −0.31 eV.

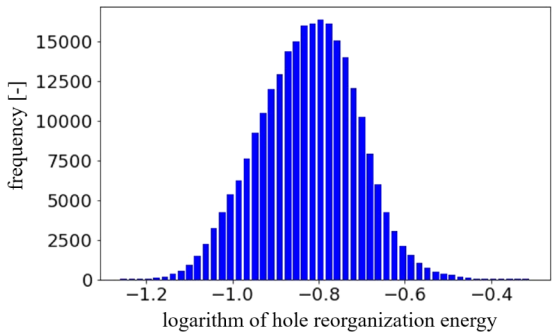


Figure 1. Distribution of hole reorganization energy within the quarter-million data set.

Molecular Descriptors. We used RDKit,[25] Mordred,[26] and Morgan fingerprints[27] to calculate the molecular descriptors. RDKit calculated 200 descriptors, including molecular weight and number of atoms, based on the two-dimensional (2D) structure of molecules. Mordred calculates 1600 descriptors from both the 2D and three-dimensional (3D) structures of molecules, with the Boruta method[29] used for variable selection to reduce the number of descriptors. Morgan fingerprints are calculated based on structural information on atoms that are located within a certain distance of a given atom; in our study, we set the parameters to radius = 2 and nBits = 2048.

GPR. GPR[22] is a linear regression method in which input variables $\boldsymbol{x}$ are transformed to high-dimensional space using basis functions φ. Linear regression is then performed on the vector $\varphi(\boldsymbol{x})$ as follows:

$$y(x) = \mathbf{w}^{\mathrm{T}}\varphi(\mathbf{x}) \quad (2)$$

where $\boldsymbol{w}$ is the weight vector. Further, as a prior distribution of $\boldsymbol{w}$, an isotropic Gaussian distribution with mean and variance equal of 0 and $\alpha^{-1}\mathbf{I}$ is assumed as follows:

$$p(\mathbf{w}) = \mathrm{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) \quad (3)$$

where, α^{-1} and $\boldsymbol{I}$ denote the hyperparameters corresponding to the inverse and identity matrix of variance σ^2, respectively. Additionally, N denotes the probability density function of the Gaussian distribution, which for $\boldsymbol{x}$ with mean μ and variance σ^2 is expressed as follows:

$$N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \quad (4)$$

$y(\boldsymbol{x})$ can be obtained for $\boldsymbol{x}$ by determining the probability distribution of $\boldsymbol{w}$ from eq 2 and deriving the probability distribution of $y(\boldsymbol{x})$ from eq 3. For example, if there are N samples used to construct model, the output vector y with

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854

elements $y_i = y(x_i)$ can be expressed using the matrix Φ, where the elements are $\Phi_i = \Phi(x_i)$, as follows:

$$\mathbf{y} = \mathbf{\Phi} w \tag{5}$$

Because y is represented by a linear combination of w following a Gaussian distribution, $\boldsymbol{y}$ also follows a Gaussian distribution. Therefore, the probability distribution of $y(\boldsymbol{x})$ can be obtained by determining the mean and covariance of $\boldsymbol{y}$. The mean and covariance of y are denoted as $E[\boldsymbol{y}]$ and $cov[\boldsymbol{y}]$, respectively. From eq 4, we have the following equation:

$$E[\mathbf{y}] = \mathbf{\Phi} E[\mathbf{w}] = 0 \tag{6}$$

$$cov[\mathbf{y}] = E[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = \mathbf{\Phi} E[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\mathbf{\Phi}^{\mathrm{T}} = \frac{1}{\alpha}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}} = \mathbf{K} \tag{7}$$

where $\boldsymbol{K}$ can be expressed as follows:

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\alpha}\varphi(\mathbf{x}_i)^{\mathrm{T}}\varphi(\mathbf{x}_j) \tag{8}$$

The matrix $\boldsymbol{K}$ is a Gram matrix with elements $k(x_i, x_j)$, where $k(x_i, x_j)$ is the kernel function. From eqs 6 and 7, we can obtain K_{ij} and $cov[\boldsymbol{y}]$ by calculating the dot product of $\Phi(x_i)$ and $\Phi(x_j)$. This dot product is known as the kernel function. We used the following 11 types of kernel functions:

ConstantKernel() * DotProduct() + WhiteKernel()
ConstantKernel() * RBF() + WhiteKernel()
ConstantKernel() * RBF() + WhiteKernel() + ConstantKernel() * DotProduct()
ConstantKernel() * RBF(np.ones(n_features)) + WhiteKernel()
ConstantKernel() * RBF(np.ones(n_features)) + WhiteKernel() + ConstantKernel() * DotProduct()
ConstantKernel() * Matern(nu = 1.5) + WhiteKernel()
ConstantKernel() * Matern(nu = 1.5) + WhiteKernel() + ConstantKernel() * DotProduct()
ConstantKernel() * Matern(nu = 0.5) + WhiteKernel()
ConstantKernel() * Matern(nu = 0.5) + WhiteKernel() + ConstantKernel() * DotProduct()
ConstantKernel() * Matern(nu = 2.5) + WhiteKernel()
ConstantKernel() * Matern(nu = 2.5) + WhiteKernel() + ConstantKernel() * DotProduct()

n_features: number of explanatory variable $\boldsymbol{X}$

In this study, the sklearn.gaussian_process.GaussianProcessRegressor module from scikit-learn[30] was used to construct the GPR model. Five functions, including DotProduct, WhiteKernel, RBF, ConstantKernel, and Matern, were used to calculate the kernel functions. The kernel parameters were optimized with the L-BFGS[30] for each kernel function. The best kernel was selected from 11 different kernels using cross-validation (CV).[31] This process involved dividing the data set into several groups, building models for all but one group, and estimating the excluded samples. The CV process was repeated for each group, and the estimates were computed. The actual and estimated values were then evaluated using performance indicators to select appropriate parameters for model building. In this study, the scikit-learn module "sklearn.model_selection.cross_val_predict"[32] was used for CV with a holdout number of 5. The sklearn.metrics.r2_score[33] module was used to calculate the performance measures, and the kernel function with the largest determinant coefficient R^2 after cross validation was selected.

We also note that in our study, y represents the logarithm of hole reorganization energy multiplied by −1. The negative sign was used to minimize the hole reorganization energy, whereas the GPR is designed to maximize the target y.

BO. In this study, we applied BO,[20,21] which utilizes the variance of the estimates from GPR to transform them into an AF. We used three AFs, including PI,[23] EI,[23] and MI.[24] The BO method identifies the next candidate experiment as the one with a higher AF value of the AF compared to the calculated value. New molecules that are not included in the data set can be proposed, as well as extrapolation in x can be performed, by applying BO. The source code of BO is available at DCEKit.[34]

PI. The PI[23] is calculated as the probability of exceeding the objective variable in the existing sample, using the following function:

$$PI(x_{new}) = \int_{Y_{max}+\varepsilon}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2(x_{new})}} \exp\left\{-\frac{1}{2\sigma^2(x_{new})} (x - \mu(x_{new}))^2\right\} dx \tag{9}$$

where $u(x_{new})$ represents the estimated y value (i.e., -log(reorganization energy)) obtained GPR, and $\sigma^2(x_{new})$ is its variance. Additionally, the term ε is defined as the product of relaxation and the standard deviation of the reorganization energy of the model data, with relaxation set to 0.01.

EI. The EI[23] represents the expected increase in the maximum y value within the existing sample. Assuming that the EI value for a new sample x_{new} is denoted as $EI(x_{new})$, it can be expressed by the following equation:

$$EI(x_{new}) = (\mu(x_{new}) - Y_{max} - \epsilon)PI(x_{new}) + \sigma^2 \frac{1}{\sqrt{2\pi\sigma^2(x_{new})}} \exp\left\{-\frac{1}{2\sigma^2(x_{new})} (Y_{\max} + \varepsilon - \mu(x_{new}))^2\right\} \tag{10}$$

It is noteworthy that ε in EI is the same as that in PI.

MI. The MI[24] method selects the maximum sum of prediction and variance and updates this variance with each new experiment. Assuming that the value of MI in a new sample x_{new} is denoted as $MI(x_{new})$, it can be expressed by the following equation:

$$MI(x_{new}) = \mu(x_{new}^{\ i}) + \varphi_i(x_{new}^{\ i}) \tag{11}$$

$$\varphi_i(x_{new}^{\ i}) = \sqrt{\alpha}\left(\sqrt{\sigma^2(x_{new}^{\ i}) + \gamma_{i-1}} - \sqrt{\gamma_{i-1}}\right) \tag{12}$$

$$\gamma_i = \gamma_{i-1} + \sigma^2(x_{new}^{(i)}) \tag{13}$$

where $\gamma_0 = 0$, $\alpha = \log(2/\delta)$, $\delta = 10^{-6}$, and $\sigma^2(x^{(i)}_{new}) = 0$. It is noteworthy that γ_i is always set to 0.

Combination of AF. In this study, BO was applied by calculating various combinations of the AFs, as follows.

Using only PI to search for next molecules.
Using only EI to search for next molecules.
Using only MI to search for next molecules.
Alternating between EI and MI in the molecular search process. Specifically, after searching for candidate

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854

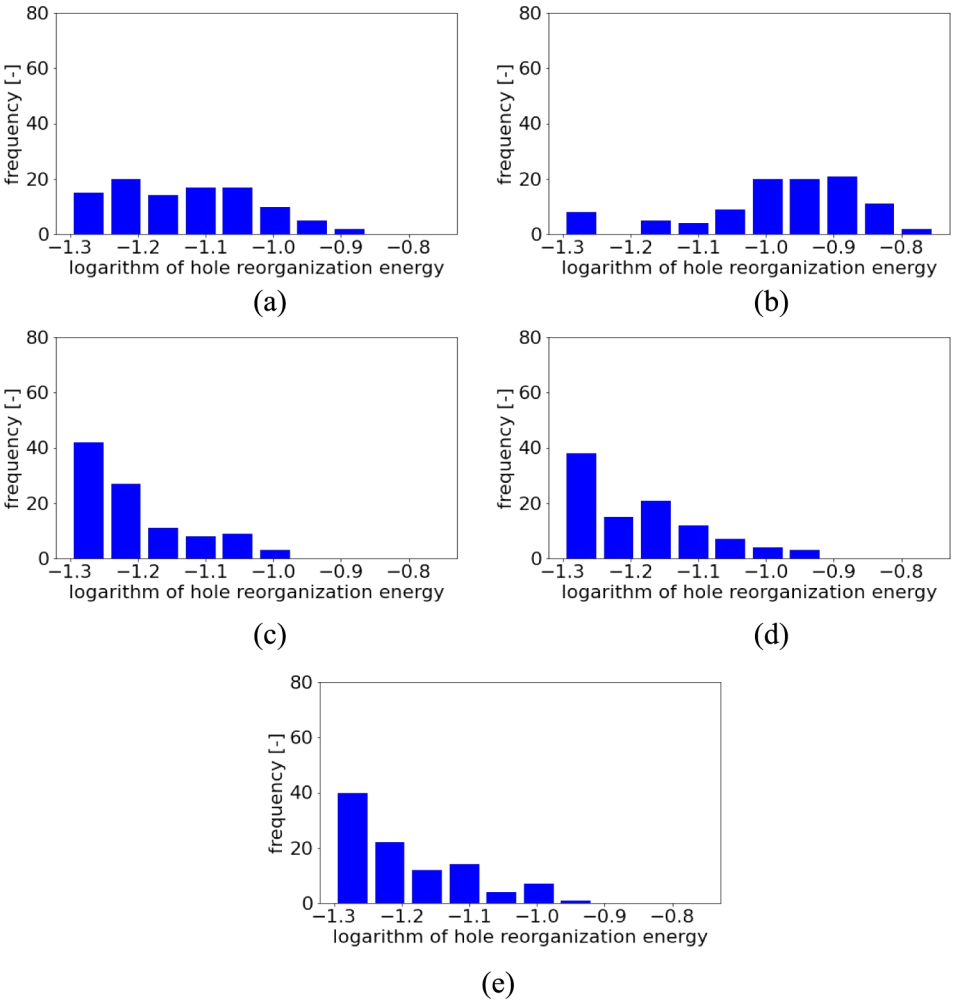


Figure 2. Histograms of the minimum values of the logarithm of hole reorganization energy (in eV) for 100 BO trials with different initial samples (RDKit), obtained using different AFs and their combinations: (a) PI, (b) EI, (c) MI, (d) EI-MI, and (e) EI-MI-PI. The base of the logarithm is 10.

molecules using EI, MI was used, and this process was repeated using EI. (EI-MI)

Alternating among PI, EI, and MI in the molecular search process. Initially, PI was used to search for candidate molecules, followed by EI and then MI. This cycle was repeated multiple times. (PI-EI-MI)

Protocol of Optimization. From the quarter-million data set, we first selected molecules exhibiting high hole reorganization energies. We then applied BO to search for molecules with lower reorganization energies. The maximum number of BO cycles required to search for molecules with lower reorganization energy was fixed at 100, and we examined whether such molecules could be explored within the given number of cycles. The details of the BO protocol are summarized as follows:

The quarter-million values of y (i.e., -log(reorganization energy)) were sorted in descending order. Subsequently, 100 molecules were randomly selected from the bottom 10% of these values to build the initial model.

A GPR model using these 100 selected molecules was built.

Variances were calculated by applying the GPR model to molecules other than those used to build the model.

AFs were calculated with predictions and their variances.

A molecule with the largest AF was added to the molecules for model building.

Steps 2−5 were repeated 100 times, and a molecule with the maximum y value was selected.

Step 6 was repeated 100 times, and a histogram of the maximum values was created.

RESULTS AND DISCUSSION

One hundred cycles of 100 BO calculations were performed for each AF using the RDKit descriptor. Figure 2 shows a histogram of the minimum values of the hole reorganization energy for 100 trials of BO with different initial samples. Figure 3 shows examples of the hole reorganization energy per cycle number in BO. The black points represent the initial data set before the BO calculations, while the blue points represent the molecules selected for the BO calculations. When PI was used, the molecule with the lowest value (−2.90 = log (0.0548 eV)) of reorganization energy was identified. However, as shown in Figures 2 and 3a,b, the search for molecules with low reorganization energies was highly dependent on the selection of the initial samples; hence, a stable search for molecules with low reorganization energies was not possible. When EI was used, the molecule with the lowest reorganization energy was identified; however, the histogram in Figure 2b and the scatter

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854

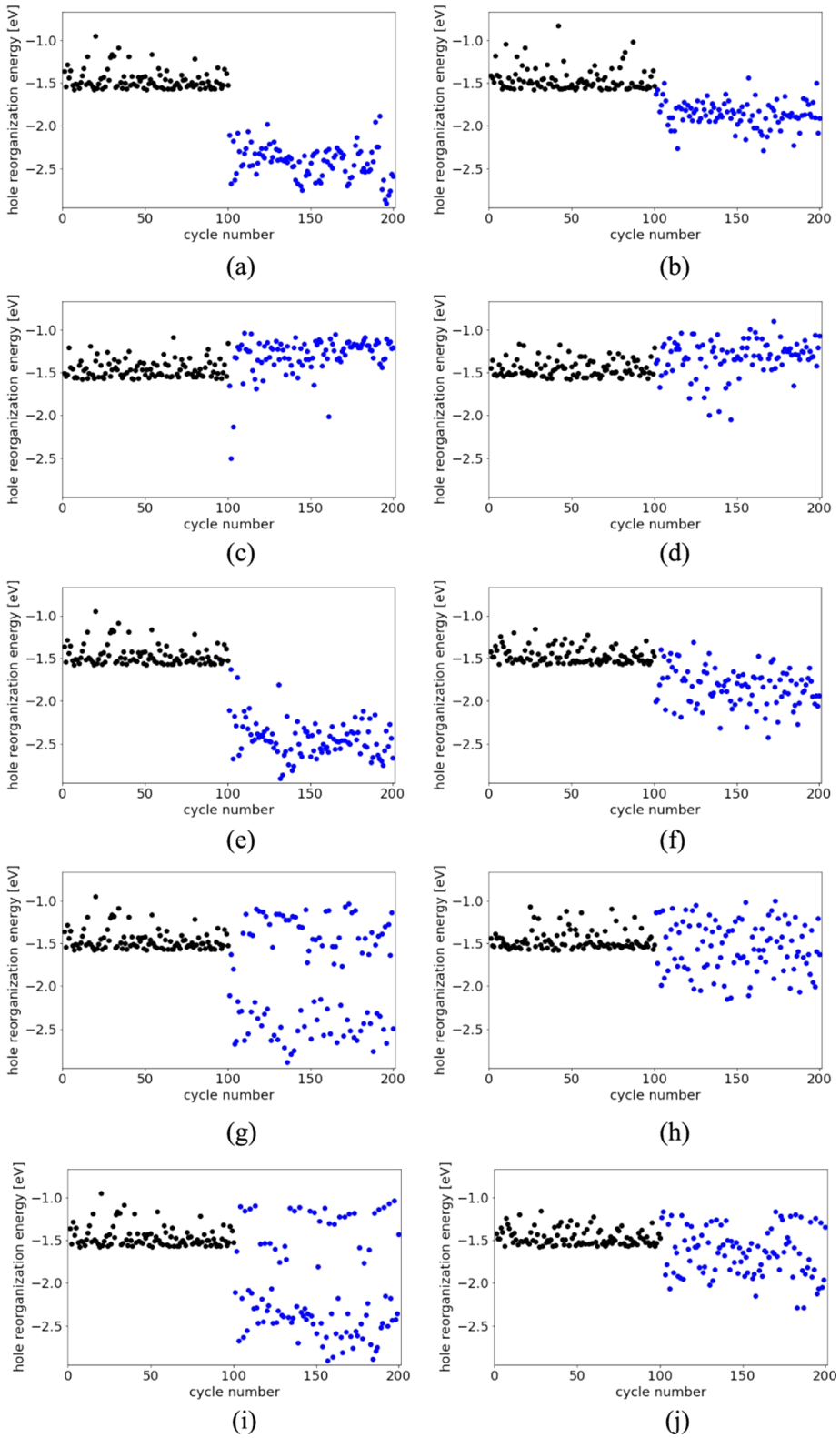


Figure 3. Examples of hole reorganization energy per cycle number in BO (RDKit) performed using different AFs and their combinations: (a,b) PI, (c,d) EI, (e,f) MI, (g,h) EI-MI, and (i,j) EI-MI-PI. The black points indicate the initial data set before the BO calculations, while the blue points indicate the molecules selected in BO.

plots in Figure 3c,d indicate that only a limited number of molecules with low reorganization energies were identified. When MI was used, the molecules with the lowest reorganization energy were identified. Figures 2c and 3e,f show that while some samples yielded low values, some initial samples were not explored at all. This suggests that the effectiveness of the search for molecules with low reorganization energies is highly influenced by the selection of the initial samples. Therefore, a stable search for molecules with low reorganization energies cannot be conducted using MI. When

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854

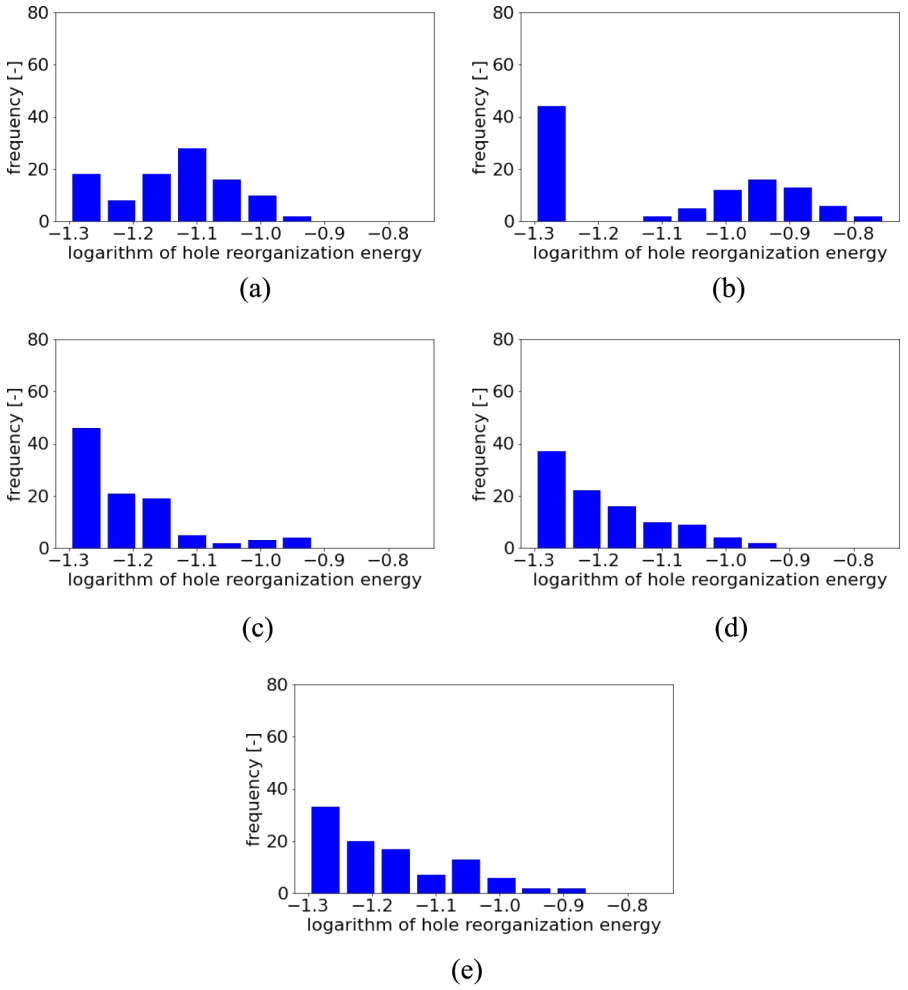


Figure 4. Histograms of the minimum values of hole reorganization energy for 100 BO trials with different initial samples (Mordred), obtained using different AFs and their combinations: (a) PI, (b) EI, (c) MI, (d) EI-MI, and (e) EI-MI-PI. The base of the logarithm is 10.

EI and MI were used alternately to select the next molecules in the BO, the combination successfully identified the molecule with the lowest reorganization energy. The histogram in Figure 2d shows that the frequency increases as the reorganization energy decreases. Although BO did not consistently select molecules with low reorganization energies, 100 BO trials resulted in the selection of molecules with reorganization energies close to the lowest value at least twice. The above results indicate that when EI and MI are used alternately, molecules with low reorganization energies can be stably selected. When BO was performed by alternating between EI, MI, and PI, the combination succeeded in identifying the molecule with the lowest reorganization energy. Figure 3i shows that the selection of molecules with high reorganization energies was minimized, whereas a higher number of molecules with low reorganization energies were selected compared to the combination of EI and MI, as shown in Figure 3g,h. In addition, even in instances where molecules with lower reorientation energies were not immediately identified, as shown in Figure 3j, the reorientation energy values tended to decrease as the search progressed. From the above results, it can be concluded that EI, MI, and PI should be used alternately for the stable selection of molecules with low reorganization energies.

In addition, 100 cycles of 100 BO calculations were performed for each AF using Mordred descriptors. Figure 4 shows a histogram of the minimum values of the hole reorganization energy for 100 BO trials with different initial samples. Figure 5 shows examples of the hole reorganization energy per cycle number in BO. The black points represent the initial data set before the BO calculations, while the blue points represent the molecules selected for the BO calculations. Using PI, the molecule with the lowest organization energy was identified; however, the histograms in Figures 4a and 5a,b show that the search for molecules with low reorganization energies was highly dependent on the selection of the initial samples. Consequently, performing a stable search for molecules with low reorganization energies using this AF was not feasible. The molecule with the lowest reorganization energy was also identified using EI. However, results in Figure 5c,d show that the AF tends to select molecules with high values of reorganization energy, and molecules with low reorganization energies could only be selected for a limited number of trials. Thus, EI may not be suitable for performing stable searches of molecules with low reorganization energies because it mostly identifies molecules with high reorganization energies. Further, MI also successfully identified molecules with the lowest reorganization energy. The histogram in Figure 4c shows an increasing frequency trend toward lower reorganization energies. Figure 5e,f shows that some initial samples could not be searched for molecules with low reorganization energies; however, most of the initial samples

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854

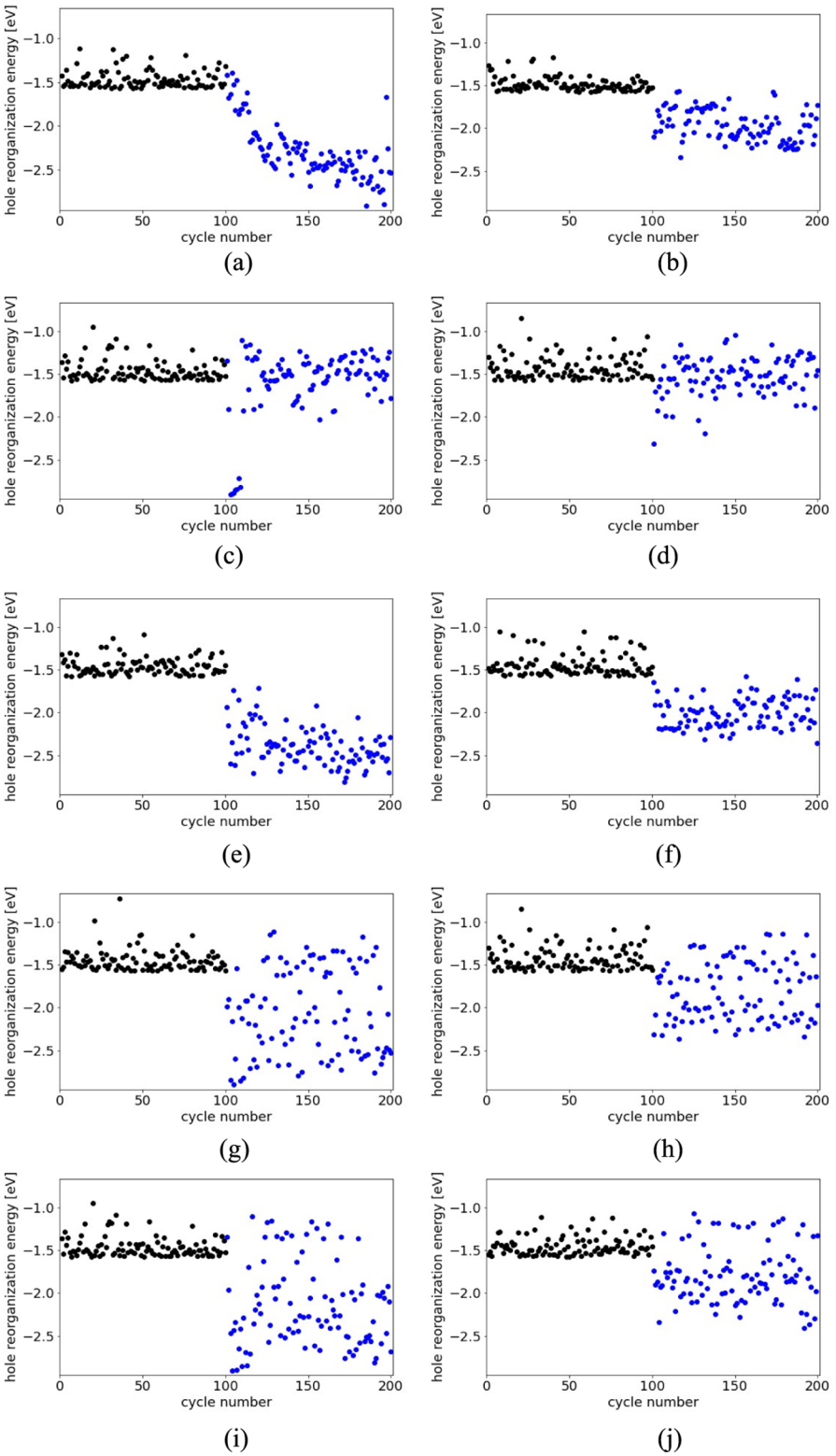


Figure 5. Examples of hole reorganization energy per cycle number in BO (Mordred) performed using different AFs and their combinations: (a,b) PI, (c,d) EI, (e,f) MI, (g,h) EI-MI, and (i,j) EI-MI-P. The black points represent the initial data set before the BO calculations, while the blue points represent the molecules selected in BO.

succeeded in continuously searching for molecules with low reorganization energies. These observations indicate that MI can stably search for molecules with low reorganization energies. When EI and MI were alternately used for selection, molecules with the lowest reorganization energies were identified. The histogram in Figure 4d shows that the frequency increased with decreasing reorganization energy, although the frequency at the lowest reorganization energy was lower than that in MI. Figure 5g,h shows that molecules with higher reorganization energies tended to be selected over those

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854

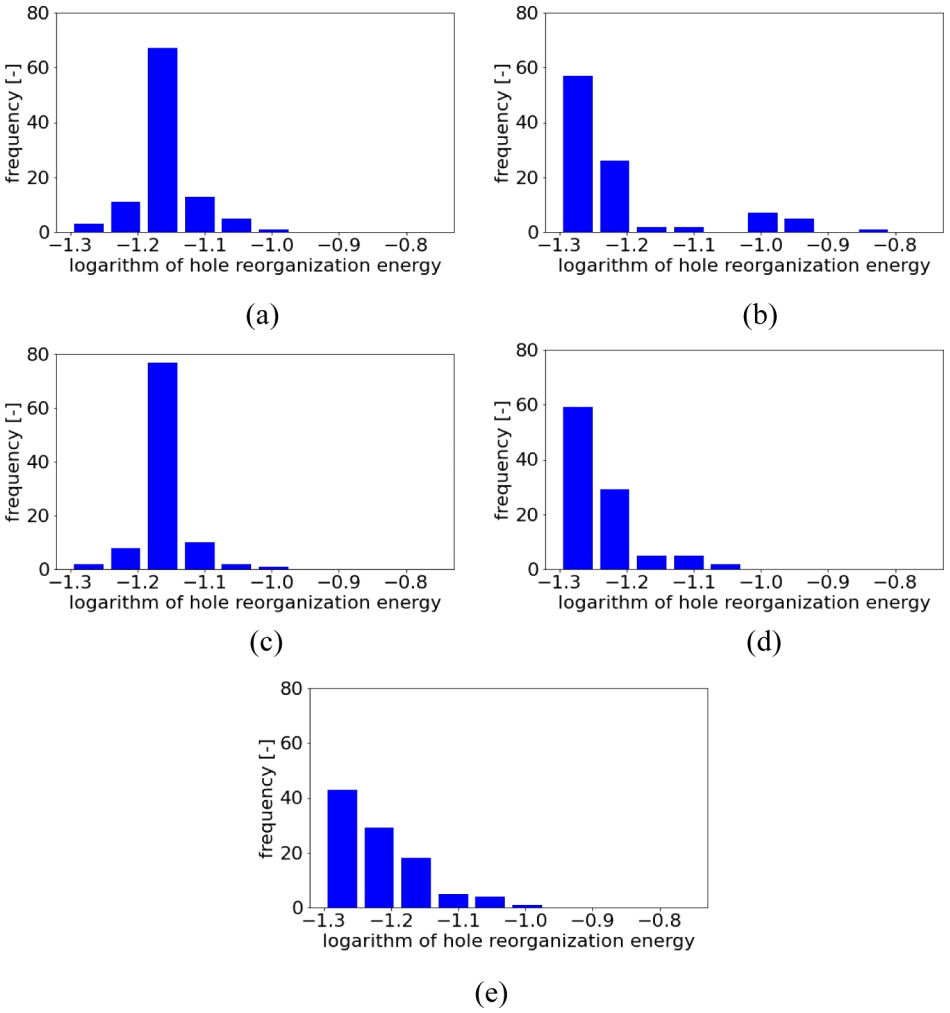


Figure 6. Histograms of the minimum values of hole reorganization energy for 100 BO trials with different initial samples (Morgan fingerprint), obtained using different AFs and their combinations: (a) PI, (b) EI, (c) MI, (d) EI-MI, and (e) EI-MI-PI. The base of the logarithm is 10.

with MI; however, molecules with lower reorganization energies were always searched after 100 BO cycles. Thus, when used alternately, EI was able to stably select molecules with low reorganization energies. The selection of the BO using EI, MI, and PI allowed us to identify the molecule with the lowest reorganization energy. Figure 4e shows that the frequency increased with decreasing reorganization energy, although the frequency for the lowest reorganization energy range was lower than that when MI was applied. Figure 5i,j shows that molecules with lower reorganization energies can be selected, suppressing the selection of molecules with higher reorganization energies compared to EI and MI when used alternately. Thus, EI, MI, and PI can be alternately used to select molecules with low reorganization energies.

A total of 100 cycles of 100 BO calculations were performed for each AF using the Morgan fingerprint descriptors. Figure 6 shows a histogram of the minimum values of the hole reorganization energy for 100 BO trials with different initial samples. Additionally, Figure 7 shows examples of the hole reorganization energy per cycle number in BO. The black points represent the initial data set before the BO calculations, while the blue points represent the molecules selected in BO. Using PI, the molecule with the lowest reorganization energy could not be identified. Figure 6a shows that the selection was not affected by the choice of initial sample; however, the search for molecules with low reorganization energies was unsuccessful. Figure 7a,b shows that most of the initial samples stably selected molecules with low reorganization energies; however, there were some initial samples in which the selection of BO did not proceed. These observations indicate that PI stably selected molecules with low reorganization energies, although it failed to identify the molecule with the lowest reorganization energy. Figure 6b shows that using EI, molecules with low reorganization energies in almost all initial samples were identified. Figure 7c,d shows that molecules with the lowest reorganization energy were identified; however, the search was not as stable, as it frequently selected molecules with reorganization energy values similar to those of the initial samples. Thus, EI can identify molecules with the lowest reorganization energy but not in a stable manner. When MI was used, molecules with the lowest reorganization energy were obtained. Figure 6c shows that the selection progressed in a stable manner in all initial samples; however, molecules with reorganization energies close to the lowest were not effectively selected. Figure 7(e) and (f) show that most of the initial samples consisted of stably selected molecules with low reorganization energies; however, there were some initial samples in which the BO selection did not progress effectively. Thus, while MI could reliably search for molecules with low reorganization energies, its effectiveness is dependent on the

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854

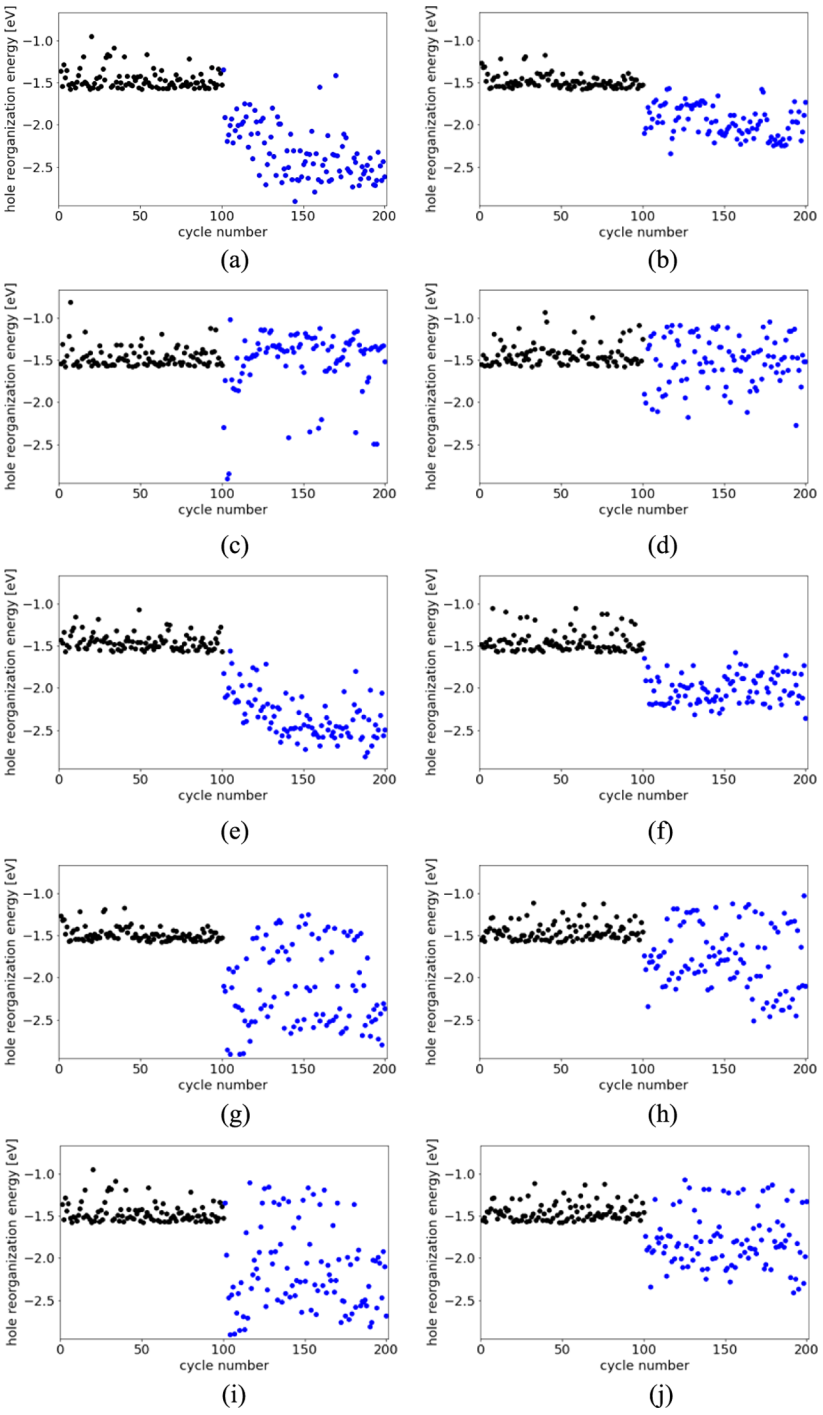


Figure 7. Examples of hole reorganization energy per cycle number in BO (Morgan fingerprint) performed using different AFs and their combinations: (a,b) PI, (c,d) EI, (e,f) MI, (g,h) EI-MI, and (i,j) EI-MI-P. The black points represent the initial data set before the BO calculations, while the blue points represent the molecules selected in BO.

choice of initial samples. When alternating between EI and MI for selections, molecules with the lowest reorganization energy were obtained. Figure 6d shows that, compared to MI, the combination of EI and MI can identify molecules with the lowest reorganization energy more efficiently, and the frequency increases as the reorganization energy decreases. Figure 7g,h shows that compared to MI, the combination of EI and MI selected molecules with high reorganization energies; however, molecules with low reorganization energies were always explored after 100 BO cycles. Thus, alternating between EI and MI effectively enabled stable exploration of molecules with low reorganization energies. When EI, MI, and PI were alternately used for selection, molecules with the lowest reorganization energy were selected. In the case of alternating EI, MI, and PI, the frequency increased with decreasing reorganization energy, as shown in Figure 6e. However, compared to EI-MI, the combination of EI, MI, and PI did not identify molecules with the lowest reorganization energy as effectively. Figure 7i,j illustrates that the combinations of EI,

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854

MI, and PI identified molecules with low reorganization energies.

When using the RDKit descriptor, AFs differed significantly in their ability to search for molecules with lower reorganization energies. Without using the combinations of the AFs, BO did not consistently locate molecules with low reorganization energies. However, when the combinations of the AFs were used, the search for molecules with low reorganization energies progressed. When the Mordred descriptor was used, BO without AF combination was effective in consistently searching for molecules with low reorganization energies or those with the lowest reorganization energy. Combining AFs also allowed for the stable identification of molecules with low reorganization energies. When using the Morgan fingerprint, BO without AF combinations successfully identified molecules with low reorganization energies or those with the lowest reorganization energies. When the AF combinations were used, the search for molecules with low reorganization energies remained stable.

When using PI, it was challenging to identify molecules with reorganization energy logarithms in the range of −1.3 to −1.2 eV. Nevertheless, it was still possible to identify molecules with the lowest reorganization energy. Using EI, stable identification of molecules with low reorganization energies was not achieved; however, it was possible to search for molecules with the lowest reorganization energies. Using MI, molecules with the lowest reorganization energy were successfully identified, and depending on the descriptor selected, the search for lower reorganization energies could be conducted. When alternating between EI and MI, molecules with reorganization energies close to the minimum were identified in a stable manner. In contrast, when EI, MI, and PI were used together, while it was less effective in identifying molecules with reorganization energies close to the minimum compared to using EI and MI alone, it still allowed for a more stable selection of molecules with lower reorganization energies.

As described above, the data set varied significantly with the use of the three descriptors; however, in all cases, combining the AFs led to a more stable identification of molecules with lower reorganization energies. These results suggest that the combined use of AF is effective for search tasks.

CONCLUSION

This study uses BO to identify molecules with a low hole reorganization energy, which dominates the carrier mobility of organic semiconductors. BO was performed using RDKit, Mordred, and Morgan fingerprints to quantify the molecules and combine PI, EI, and MI as AFs. One hundred different sets of initial data sets were prepared, and searches to minimize the reorganization energy were conducted for each initial data set and descriptor set, comparing individual AFs and their combinations for BO. The use of AF alone did not enable stable searches for molecules with low reorganization energy for all descriptors. However, combining acquisition functions during Bayesian optimization resulted in more stable searches for molecules with lower reorganization energies compared to the initial data set and increased the likelihood of identifying molecules with the lowest reorganization energy. These results suggest that the AFs should be combined when performing BO for the reorientation energy. The combinations of different AFs enhance the diversity of the selected molecules by incorporating diverse selection criteria. Although PI, EI, and MI were used in this study, other AFs such as upper and lower confidence bounds can be also combined with them to further improve the performance of BO, which should be tested. The proposed BO method is expected to accelerate the development of not only organic semiconductors but also other materials.

ASSOCIATED CONTENT

Data Availability Statement

The data that support the findings of this study are available in ref. 28. The source code of BO is available at DCEKit.[34] The code is for Bayesian optimization, and it is possible to select PI, EI, and MI. In the proposed method, PI, EI, and MI are performed in order. Therefore, it is possible to reproduce the proposed method by selecting PI, EI, and MI in order.

AUTHOR INFORMATION

Corresponding Authors

Tatsuhito Ando − *Engineering Division, Panasonic Industry Co., Ltd., Kadoma, Osaka 571-8506, Japan;* orcid.org/0000-0001-9755-1009; Email: ando.tatsuhito@jp.panasonic.com

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

Authors

Rinta Kawagoe − *Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan*

Nobuyuki N. Matsuzawa − *Engineering Division, Panasonic Industry Co., Ltd., Kadoma, Osaka 571-8506, Japan;* orcid.org/0000-0002-2388-7797

Hiroyuki Maeshima − *Engineering Division, Panasonic Industry Co., Ltd., Kadoma, Osaka 571-8506, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c09124

Author Contributions

[§]T.A. and H.K. contributed equally. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research (KAKENHI) (Grant Numbers 24K08152 and 24K01234) from the Japan Society for the Promotion of Science.

ABBREVIATIONS

AF, acquisition functions; BO, Bayesian optimization; EI, expected improvement; GPR, Gaussian process regression; MI, mutual information; PI, probabilities of improvement

REFERENCES

(1) Heo, J. S.; Eom, J.; Kim, Y.-H.; Park, S. K. Recent Progress of Textile-Based Wearable Electronics: A Comprehensive Review of Materials, Devices, and Applications. *Small* **2018**, *14* (3), 1703034.

(2) Lee, E. K.; Lee, M. Y.; Park, C. H.; Lee, H. R.; Oh, J. H. Toward Environmentally Robust Organic Electronics: Approaches and Applications. *Adv. Mater.* **2017**, *29* (44), 1703638.

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854

(3) Liu, Y.; Pharr, M.; Salvatore, G. A. Lab-On-Skin: A Review of Flexible and Stretchable Electronics for Wearable Health Monitoring. *ACS Nano* **2017**, *11* (10), 9614−9635.
(4) Root, S. E.; Savagatrup, S.; Printz, A. D.; Rodriquez, D.; Lipomi, D. J. Mechanical Properties of Organic Semiconductors for Stretchable, Highly Flexible, and Mechanically Robust Electronics. *Chem. Rev.* **2017**, *117* (9), 6467−6499.
(5) Isono, S.; Satake, T.; Hyakushima, T.; Taki, K.; Sakaida, R.; Kishimura, S.; Hirao, S.; Nomura, K.; Torazawa, N.; Tsutsue, M.et al.A 0.9 μm Pixel Size Image Sensor Realized by Introducing Organic Photoconductive Film into the BEOL Process*Proceedings of the 2013 IEEE International Interconnect Technology Conference*IEEE20136615587
(6) Anthony, J. E. Functionalized Acenes and Heteroacenes for Organic Electronics. *Chem. Rev.* **2006**, *106* (12), 5028−5048.
(7) Perepichka, I. F.; Perepichka, D. F.; Meng, H.; Wudl, F. Light-Emitting Polythiophenes. *Adv. Mater.* **2005**, *17* (19), 2281−2305.
(8) Kang, M. J.; Doi, I.; Mori, H.; Miyazaki, E.; Takimiya, K.; Ikeda, M.; Kuwabara, H. Alkylated Dinaphtho[2,3-b: 2′, 3′-f]Thieno[3,2-b]Thiophenes (Cn-DNTTs): Organic Semiconductors for High-Performance Thin-Film Transistors. *Adv. Mater.* **2011**, *23* (10), 1222−1225.
(9) Sieval, A. B.; Hummelen, J. C. Fullerene-Based Acceptor Materials. In *Organic Photovoltaics*; Wiley, 2014. .
(10) Li, H.; Tee, B. C.-K.; Cha, J. J.; Cui, Y.; Chung, J. W.; Lee, S. Y.; Bao, Z. High-Mobility Field-Effect Transistors from Large-Area Solution-Grown Aligned C_{60} Single Crystals. *J. Am. Chem. Soc.* **2012**, *134* (5), 2760−2765.
(11) Takeya, J.; Yamagishi, M.; Tominari, Y.; Hirahara, R.; Nakazawa, Y.; Nishikawa, T.; Kawase, T.; Shimoda, T.; Ogawa, S. Very High-Mobility Organic Single-Crystal Transistors with in-Crystal Conduction Channels. *Appl. Phys. Lett.* **2007**, *90* (10), 102120.
(12) Prince, M. B. Drift Mobilities in Semiconductors. II. Silicon. *Phys. Rev.* **1954**, *93* (6), 1204−1206.
(13) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. I. *J. Chem. Phys.* **1956**, *24* (5), 966−978.
(14) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. II. Applications to Data on the Rates of Isotopic Exchange Reactions. *J. Chem. Phys.* **1957**, *26* (4), 867−871.
(15) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. III. Applications to Data on the Rates of Organic Redox Reactions. *J. Chem. Phys.* **1957**, *26* (4), 872−877.
(16) Marcus, R. A. On the Theory of Electron-Transfer Reactions. VI. Unified Treatment for Homogeneous and Electrode Reactions. *J. Chem. Phys.* **1965**, *43* (2), 679−701.
(17) Schober, C.; Reuter, K.; Oberhofer, H. Virtual Screening for High Carrier Mobility in Organic Semiconductors. *J. Phys. Chem. Lett.* **2016**, *7* (19), 3973−3977.
(18) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58* (3), 380−388.
(19) Matsuzawa, N. N.; Arai, H.; Sasago, M.; Fujii, E.; Goldberg, A.; Mustard, T. J.; Kwak, H. S.; Giesen, D. J.; Ranalli, F.; Halls, M. D. Massive Theoretical Screen of Hole Conducting Organic Materials in the Heteroacene Family by Using a Cloud-Computing Environment. *J. Phys. Chem. A* **2020**, *124* (10), 1981−1992.
(20) Shahriari, B.; Swersky, K.; Wang, Z. Y.; Adams, R. P.; de Freitas, N. Taking the Human out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104* (1), 148−175.
(21) Gramacy, R. B. *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*; CRC Press, 2020.
(22) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press, 2006.
(23) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems 25*, NIPS, 2012, pp 2951−2959.
(24) Contal, E.; Perchet, V.; Vayatis, N.Gaussian Process Optimization with Mutual Information*Proceedings of the 31st International Conference on Machine Learning*PMLR2014253
(25) https://rdkit.org (accessed August 23 2024).
(26) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminf.* **2018**, *10* (1), 4.
(27) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(28) Staker, J.; Marshall, K.; Leswing, K.; Robertson, T.; Halls, M. D.; Goldberg, A.; Morisato, T.; Maeshima, H.; Ando, T.; Arai, A.; Sasago, M.; Fujii, E.; Matsuzawa, N. N. De Novo Design of Molecules with Low Hole Reorganization Energy Based on a Quarter-Million Molecule DFT Screen: Part 2. *J. Phys. Chem. A* **2022**, *126* (34), 5837−5852.
(29) Kursa, M. B.; Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Software* **2010**, *36* (11), 1−13.
(30) https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html (accessed 8 November 2024).
(31) Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., pp 1193−1199.
(32) https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_predict.html (accessed November 8 2024).
(33) https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (accessed 8 November 2024).
(34) https://github.com/hkaneko1985/dcekit (accessed 8 November 2024).

https://doi.org/10.1021/acsomega.4c09124
ACS Omega 2024, 9, 48844−48854