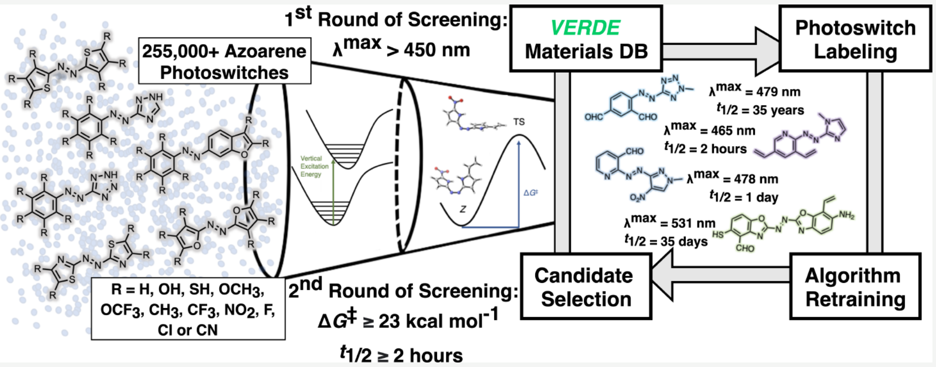
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Efficient Discovery of Visible Light-Activated Azoarene Photoswitches with Long Half-Lives Using Active Search

Fatemah Mukadum, Quan Nguyen, Daniel M. Adrion, Gabriel Appleby, Rui Chen, Haley Dang, Remco Chang, Roman Garnett, and Steven A. Lopez*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 5524−5534 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

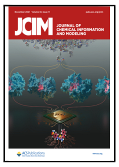


ABSTRACT: Photoswitches are molecules that undergo a reversible, structural isomerization after exposure to certain wavelengths of light. The dynamic control offered by molecular photoswitches is favorable for materials chemistry, photopharmacology, and catalysis applications. Ideal photoswitches absorb visible light and have long-lived metastable isomers. We used high-throughput virtual screening to predict the absorption maxima ($\lambda_{\max}$) of the *E*-isomer and half-life ($t_{1/2}$) of the *Z*-isomer. However, computing the photophysical and kinetic stabilities with density functional theory of each entry of a virtual molecular library containing thousands or millions of molecules is prohibitively time-consuming. We applied active search, a machine-learning technique, to intelligently search a chemical search space of 255 991 photoswitches based on 29 known azoarenes and their derivatives. We iteratively trained the active search algorithm on whether a candidate absorbed visible light ($\lambda_{\max}$ > 450 nm). Active search was found to triple the discovery rate compared to random search. Further, we projected 1962 photoswitches to 2D using the Uniform Manifold Approximation and Projection algorithm and found that $\lambda_{\max}$ depends on the core, which is tunable by substituents. We then incorporated a second stage of screening to predict the stabilities of the *Z*-isomers for the top candidates of each core. We identified four ideal photoswitches that concurrently satisfy the following criteria: $\lambda_{\max}$ > 450 nm and $t_{1/2}$ > 2 h.These candidates had $\lambda_{\max}$ and $t_{1/2}$ range from 465 to 531 nm and hours to days, respectively.

■ INTRODUCTION

Light is an ideal external stimulus to promote organic reactions. Photoswitches are a class of molecules that absorb light and reversibly interconvert between their thermodynamically stable and metastable forms to create photostationary states. Azobenzenes are a class of well-studied photoswitches that undergo efficient isomerization from their thermodynamically stable form (i.e., *E*) to their metastable form (i.e., *Z*) using ultraviolet light (314 nm).[1] The *Z* → *E* isomerization is promoted with 365 nm light.[1] This relatively high-energy light (e.g., ultraviolet) may promote undesired side reactions that compete with the isomerization pathway (e.g., electrocyclic ring-closing reactions). UV light can also promote [2 + 2]-dimerizations that alter the structure and function of nucleotides and has a limited (epidermal depth, 0.1 mm) tissue penetration depth,[2] thus limiting the therapeutic potential of photoswitches in photopharmacology.[3] The *Z*-isomer of azobenzene has a thermal half-life ($t_{1/2}$) of 2 days.[4] Photoswitches ideally suited for photopharmacology applications feature long absorption wavelengths and long $t_{1/2}$; unfortunately, the simultaneous optimization of these parameters is challenging and has been empirically observed to compete. Functionalizing the phenyl rings has been shown to shift the $\lambda_{\max}$ of azobenzene-based photoswitches into the visible range. Konrad et al.[5] recently demonstrated that functionalizing the phenyl rings with halogens at the ortho positions led to a substantial red shift to

Received: August 6, 2021
Published: November 9, 2021



https://doi.org/10.1021/acs.jcim.1c00954

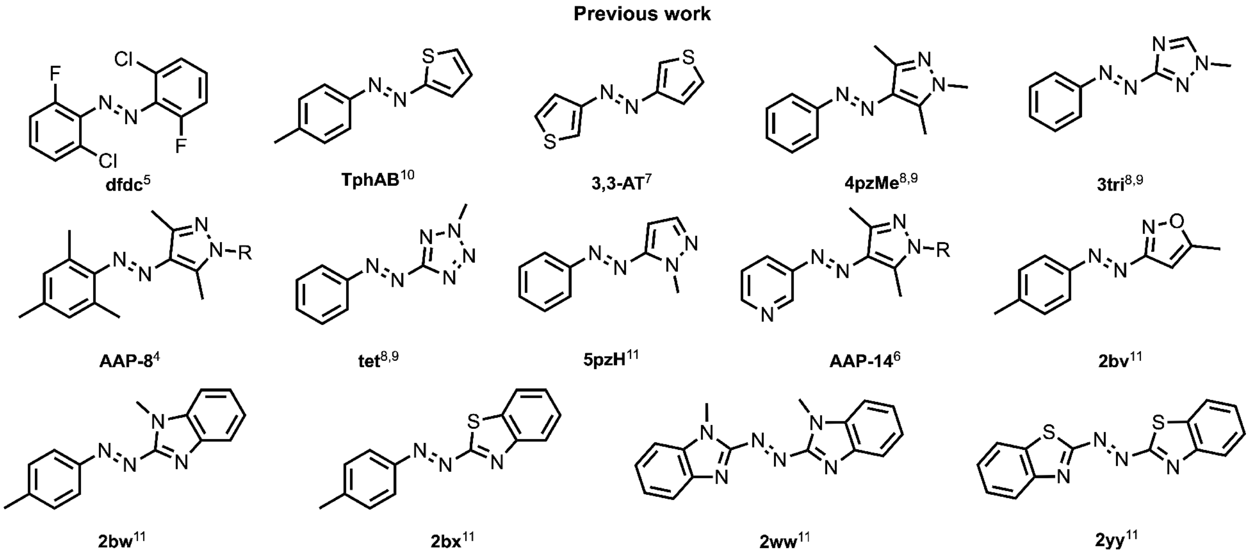
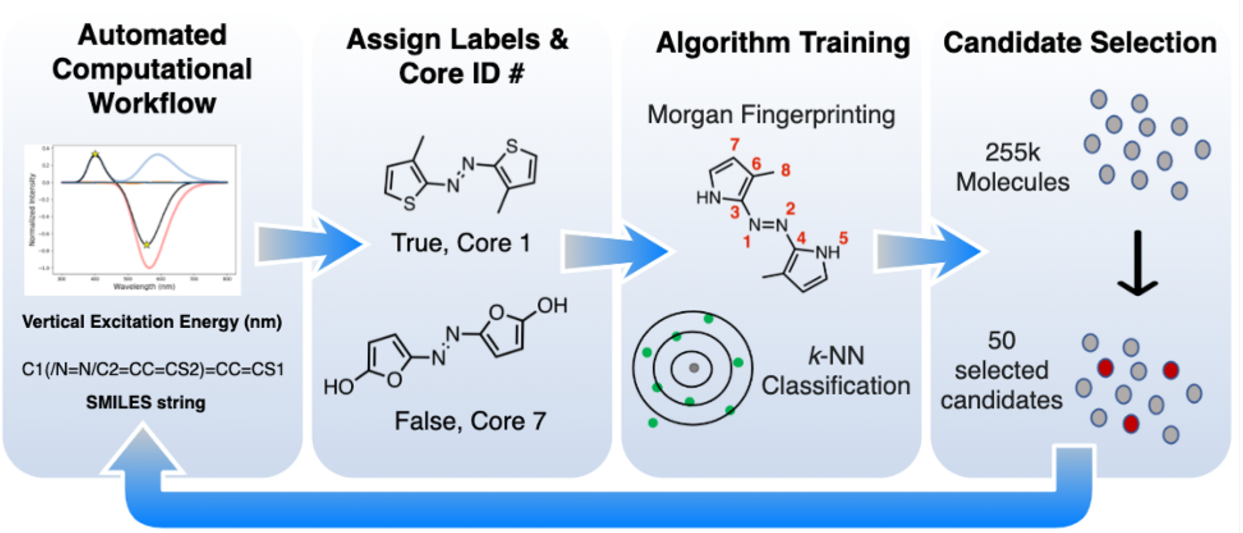


Figure 1. Fourteen azoarene photoswitches used to generate cores for a new molecular library.

Scheme 1. Multipronged Iterative Procedure Used to Update the Active Search Algorithm with DFT Results



410 nm. This functionalization strategy also increased the $t_{1/2}$ to 16 h. Another strategy involves replacing one or both phenyl rings with heteroaryl ring(s), thus creating a more general class of photoswitches, azoarenes. Azoarenes are substantially more diverse than azobenzenes, and multiple examples show λ_{max} values in the visible range and $t_{1/2}$ exceeding 1.5 h. Figure 1 highlights some of the most promising synthesized azoarenes with respect to λ_{max} and $t_{1/2}$.[5−11]

While this relatively new class of azoarene photoswitches is attractive, the complete enumeration of the chemical space approaches 10^6. Density functional theory (DFT) calculations are used to predict structures and photophysical properties at a relatively low computational cost.[12,13] Thus, DFT has been previously used in high-throughput virtual screening (HTVS)[14−17] for virtual libraries containing 500−500 000 molecules. The vastness of the chemical space cannot be understated; conservative estimates suggest that at least 10^{23} organic molecules are theoretically possible.[18] This figure can be narrowed to roughly 10^6 for azoarenes by focusing on those already experimentally realized. Abreha et al.[19] recently published a suite of HTVS tools and the Virtual Excited State Reference for the Discovery of Electronic Materials Database (the VERDE materials DB). The VERDE materials DB is unique because it was the first open-access database to include excited state structures (S_0, S_1, and T_1), photophysical, and redox properties. Further, Adrion et al.[20] published the *EZ*-TS code, which predicts thermal *Z* → *E* activation barriers efficiently and accurately.

Even with high-performance computing and efficient quantum chemistry codes, computing the photophysical properties and stabilities of 10^6 photoswitches is a substantial undertaking. Previous machine-learning (ML) techniques have involved photoswitch property prediction with experimental data. Ju et al.[21] developed a ML algorithm to predict emission wavelengths and quantum yield without quantum chemical data. They used 4300 experimental samples to train and found comparable results between their ML algorithm and TD-DFT calculations. Thawani et al.[22] curated a data set of photoswitches with azobenzene and azoarene derivatives. They also included

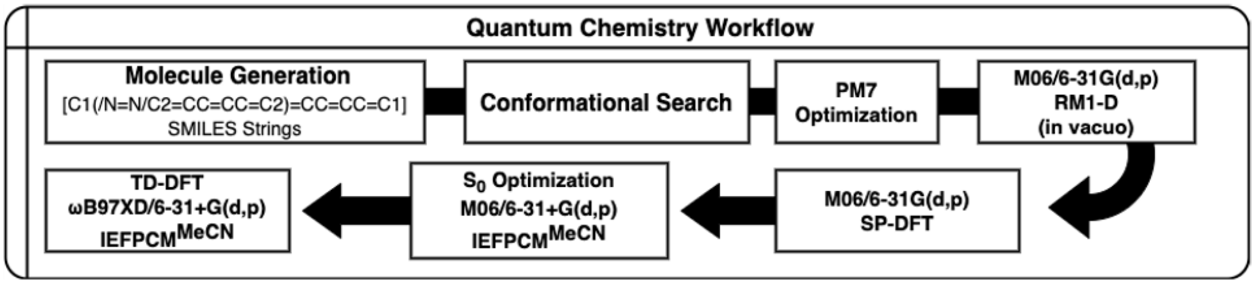


Figure 2. Quantum chemical workflow for computing the λ_{max} for all molecules selected by AS in this study.

experimental data and trained a ML algorithm to predict the transition wavelengths of *E*- and *Z*-isomers.

We have employed the machine-learning algorithm "active search"[23] to intelligently search the vast chemical space (255 991 candidates) of azoarene photoswitches. Active search (AS) was created to discover as many target molecules as possible while balancing computational resources. AS uses the data observed at any given point throughout a search and adaptively makes decisions based on the latest observations.

We now combine these existing tools (the VERDE materials DB,[19] *EZ*-TS,[20] and active search[23]) to automatically identify top photoswitch candidates featuring visible-light λ_{max} and long $t_{1/2}$. The prediction accuracy of our predictive model improves as we frequently query from quantum chemical calculations. After completing the active search iterations, we visualize our results with Uniform Manifold Approximation and Projection (UMAP). The combination of DFT, AS, and UMAP has never been employed to search the chemical space of photoswitches. Scheme 1 shows an illustration of the iterative processes used to identify ideal photoswitches

Phase 1: An initial screen of 50−100 molecules is processed through an automated computational workflow developed by Abreha et al.[19] RDKit[24] is used to generate 3-D coordinates from a simplified molecular-input line-entry system (SMILES)[25] string, followed by a low-mode conformational search where each conformer (four in total) is minimized with the Universal Force Field.[26] The lowest energy conformer is determined through semiempirical optimizations and a single-point energy calculation. The lowest energy structure is optimized with M06/6-31+G(d,p)[27−,29] and IEFPCM^{MeCN},[30] and a vibrational analysis confirms the stationary point as the true minimum if it has only positive frequencies. The λ_{max} is calculated with a single-point energy calculation using ωB97XD/6-31+G(d,p)//M06/6-31+G(d,p).[27,31] Figure 2 shows the automated workflow of quantum chemical calculations used to compute the excitation energies and corresponding λ_{max} for selected molecules from our virtual library.

Phase 2: An in-house Python script assigns a "core ID" (**1**−**29**) to each computed structure. Cores are determined using a substructure analysis included in RDKit. `True` or `False` labels are assigned to each smiles string based on the predetermined threshold, $\lambda_{max} > 450$ nm.

Phase 3: A machine-learning model is trained on the set of labeled molecules to guide the search algorithm. First, we generate the Morgan fingerprint[32] of each molecule and compute the Tanimoto similarity coefficient[33] between each pair of molecules. We then build a *k*-nearest neighbors (*k*-NN)[34] predictive model that computes the probability of a given unlabeled molecule having a positive label, given the data we have observed thus far. This *k*-NN model is then utilized by the search algorithm. The Morgan fingerprints and Tanimoto similarity coefficients only need to be computed once, while the *k*-NN is updated with newly labeled data at each iteration.

Phase 4: The active search algorithm builds the set of 50 recommendations, selecting among all unlabeled molecules (i.e., the chemical space). These recommendations are then sent to Phase 1 to be computed and labeled. This procedure repeats for a total of 40 iterations, sampling 1962 molecules from the space. We include a more detailed description of our methods in the following section.

■ METHODS

We adapted the active search method, which has shown impressive performance in molecular discovery in previous studies.[35−39] The method was first introduced by Garnett et al.[23] and extended to the batch setting by Jiang et al.[39] Formally, suppose we have a large set of elements $\mathcal{X} = \{x_i\}$, among which there is a small subset $\mathcal{R} \subset \mathcal{X}$ of valuable elements that we wish to search for (i.e., molecules exhibiting a desired property). We do not know which members of $\mathcal{X}$ belong to $\mathcal{R}$ *a priori*, but whether a specific element *x* belongs to $\mathcal{R}$ can be determined by querying an oracle, requesting for the binary label $y = 1\{x \in \mathcal{R}\}$, where $1\{\cdot\}$ is the indicator function. In this work, the binary label denotes whether a molecule exceeds the λ_{max} threshold of 450 nm. Further, we assume that at each iteration of the search, *b* elements are inspected simultaneously, requiring that queries to the oracle be made in batches of size *b*. This models experimental settings in which multiple experiments may be run in parallel to maximize throughput, contrasting with the fully sequential setting where queries are made one after another; here, $b = 50$. The goal is to design a sequence of queries limited by a predetermined budget, such that the number of target elements uncovered by querying the oracle is maximized. As such, we naturally define the utility of a given set of observations $\mathcal{D} = \{(x_i, y_i)\}$ to be the total number of targets found:

$$u(\mathcal{D}) = \sum_{y_i \in \mathcal{D}} y_i$$

We aim to determine the sequence of queries that maximizes our definition of utility in the expected case using Bayesian decision theory. This framework first requires a classification model that computes the posterior probability that an unlabeled point *x* belongs to $\mathcal{R}$, given the elements we have inspected thus far in $\mathcal{D}$, $\Pr(y = 1 \mid x, \mathcal{D})$. The active search method is model-agnostic and does not make any further assumptions about this predictive model. In the next section, we describe the *k*-nearest neighbors model we use for this classification task.

We denote $T = tb$ to be the total number of queries allowed to be made given our budget, where *t* is the number of search

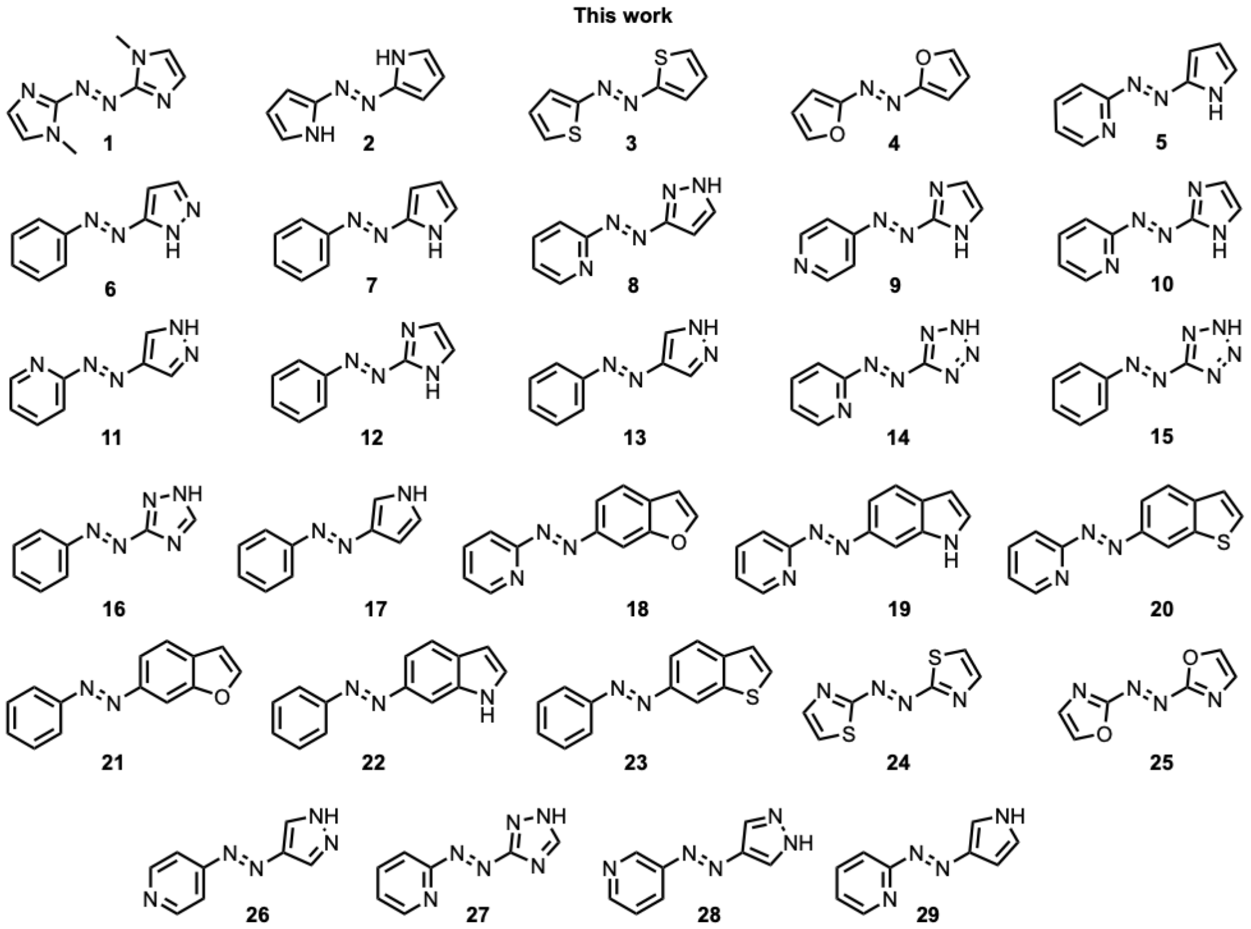


Figure 3. Twenty-nine cores explored in this study.

iterations. We further denote by $\mathcal{D}_i$ the observations collected at the end of iteration i. At iteration $i + 1 \leq t$, the best batch of queries (of size b) we can make, denoted as X_{i+1}, maximizes the expected value of the utility of the data set at termination $\mathcal{D}_t$:

$$X_{i+1} = \arg\max_{X} \mathbb{E}[u(\mathcal{D}_t)|X, \mathcal{D}_i]$$

Jiang et al.[39] analyzed and provided further interpretation for this expected utility. Specifically, it decomposes into the sum of the expected number of positives in the current batch and the expected number of positives found in the future, assuming optimal behavior.[39] The second term in this sum is large relative to the first when the remaining budget is large; as such, the objective naturally balances between exploration and exploitation. This quantity only coincides with the expected number of molecules in the current batch at the very last iteration where the greedy batch is optimal (the second term is zero). Eq 3 in Jiang et al.[39] has more discussion on this objective function.

Although this objective can be derived using the standard procedure of backward induction,[40] it involves $t - i$ nested steps of sampling over unknown labels of candidate queries and maximizing the future expected utility. This computation is prohibitively expensive for horizons $t - i \geq 3$, rendering the optimal query infeasible to calculate in practice.

We adopt the *sequential simulation* strategy proposed by Jiang et al.[39] as an efficient approximation to the optimal batch of queries. First, the strategy builds on the efficient nonmyopic search algorithm ENS[38] in the sequential setting where only one query is made at each iteration. ENS itself approximates the optimal sequential strategy by assuming that all future queries after the current iteration are made at the same time. Jiang et al.[38] demonstrated that ENS actively explores the search space when the remaining budget is large, recommends increasingly promising molecules as the search progresses, and achieves significant improvements in performance over greedy strategies. Our sequential simulation active search algorithm under the batch setting builds its recommendations by iteratively adding elements to an initially empty set using the ENS algorithm until the desired size ($b = 50$) is reached. As a new element is added, we assume that this element will return a negative label (i.e., the element is assumed to lack the desired property). Jiang et al.[39] demonstrated that by taking on this pessimistic view, the algorithm encourages the elements within the same batch to be diverse, which helps explore the search space more effectively. During batch construction, as each point is assumed to be negative, the probabilities of the neighbor points are lowered, effectively causing future points in the same batch to be "pushed away" from the current one. Please see Section 5.2 in Jiang et al.[39] for further discussion and theoretical motivation for this interpretation (this policy also greedily maximizes the probability that *at least one* batch member is positive). The authors further showed that the algorithm significantly outperformed popular baselines in the machine -learning literature such as the greedy and the upper confidence bound (UCB) policy.

Finally, we aim to distribute our queries equally across the 29 cores. Our sequential simulation strategy may be naturally modified in service of this goal as follows. As a new element is

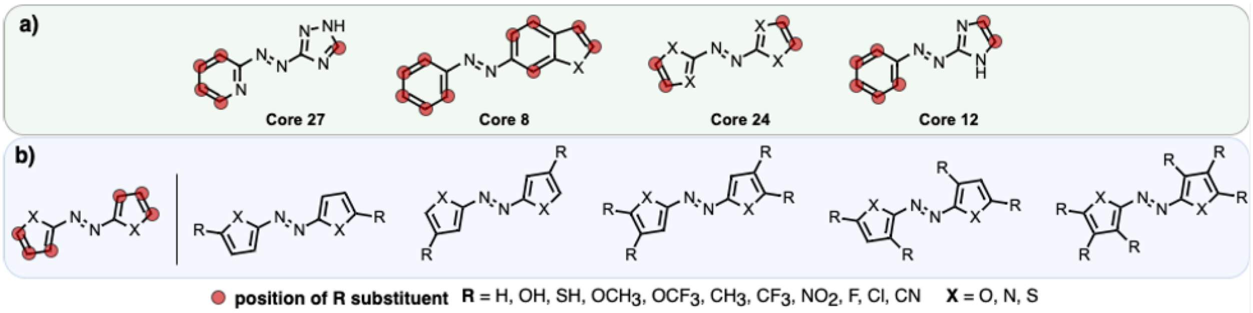

a)
Core 27
Core 8
Core 24
Core 12
b)
● position of R substituent R = H, OH, SH, OCH3, OCF3, CH3, CF3, NO2, F, Cl, CN X = O, N, S

Figure 4. Schematic representation of the substitution patterns of azoheteroarene cores. (a) Subset of 4 cores from the 29. (b) Red circles indicate positions substituted asymmetrically with terminal groups from Figure 2. H, OH, SH, OCH_3, OCF_3, CH_3, CF_3, NO_2, F, Cl, or CN, and X represent endocyclic heteroatoms (oxygen, nitrogen, or sulfur). The 11 substituents are functional groups that range from electron-withdrawing (e.g., NO_2) to electron-donating (e.g., OH).

added to the running batch in the iterative procedure described above, we temporarily remove other candidates having the same core ID as the newest batch member from the search space. When no candidate remains, we add all removed molecules back to our search space. This simple procedure effectively forces each batch of queries to be constructed to span the available cores equally.

As previously described, our active search algorithm requires a probabilistic model that computes the probability that an unlabeled element has a positive label (i.e., exhibiting the desired property), given the current set of observations we have made so far. We first generate the Morgan fingerprint[32] of each molecule in our search space and compute the Tanimoto similarity coefficient[33] between each pair of elements x and x', denoted as $t(x, x')$. We then implement a k-nearest neighbor (k-NN)[34] predictive model, which computes the probability of an uninspected molecule being an active compound as

$$\Pr(y = 1|x, \mathcal{D}) = \frac{\gamma + \sum_{x' \in \mathrm{NN}(x)} t(x, x')y'}{1 + \sum_{x' \in \mathrm{NN}(x)} t(x, x')}$$

where NN(x) is the labeled subset of the k nearest neighbors of x in X. γ is a parameter of the model that acts as a "pseudocount" to define the prior probabilities for molecules that do not have any labeled neighbor; we set $\gamma = 0.1$. This k-NN performs well in previous work,[23,35,36,38,39] as well as in our experiment; please refer to the accompanying Supporting Information for more details regarding the performance of the model. It can further be rapidly updated in light of new observations, allowing for efficient lookahead computations that are central in active search.

Data Set Generation. We aimed to curate a chemical space of novel azoarene photoswitches that could provide insight into molecular design. We searched the literature and designed 29 bisdiazoarene cores (Figure 3) to apply the trained algorithm. The substituents were added in a combinatorial fashion, at positions that would be easily synthesized. This was done with an in-house molecule generation script that takes in a SMILES string and uses RDKit to substitute denoted positions. Each of these has at least one functionalization site substituted with functional groups (i.e., **terminals**).

The **cores** were selected based on a literature search of previously synthesized azoarenes; **1**−**29** range from symmetric bisazoarenes to azoheteroarenes, and known functionalization strategies inspire the substitution sites. Figure 4 describes these positions for a smaller subset of cores.

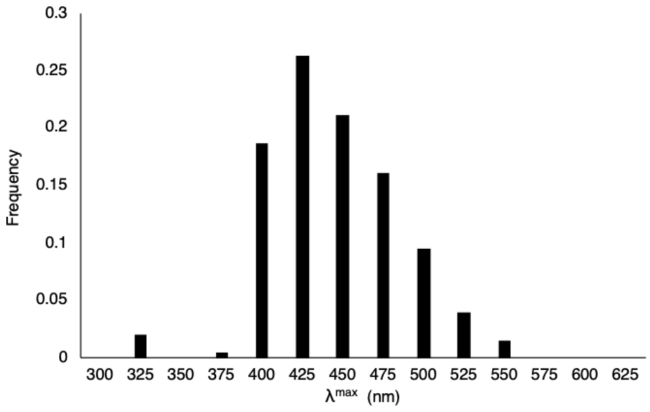


Figure 5. Distribution of the $\lambda_{\max}$ values of the photoswitch training set.

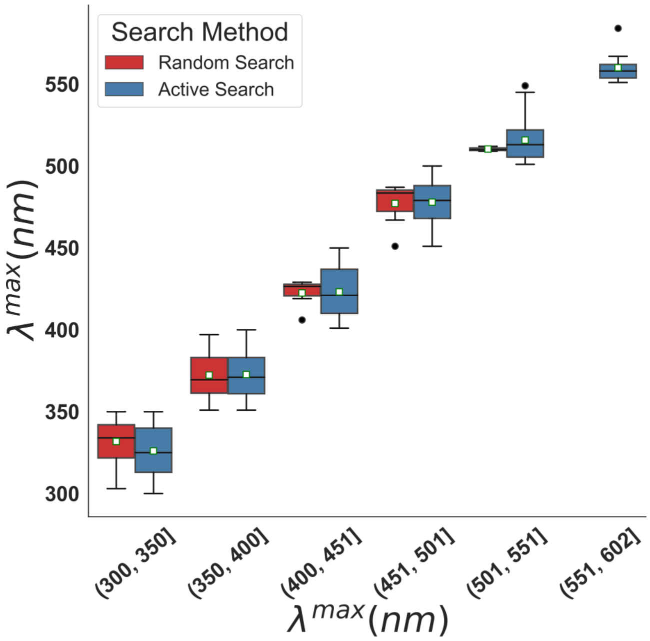


Figure 6. Box plot of the random search compared to the active search. For the random search, molecules are sampled for each core, resulting in a total of 87 molecules. Active search calculations entail 1962 computed azoarenes. The bin size is 50 nm. The median is denoted by the horizontal line within each box, and the average for each bin is denoted by the white squares. Outliers are shown as black circles.

https://doi.org/10.1021/acs.jcim.1c00954
J. Chem. Inf. Model. 2021, 61, 5524−5534

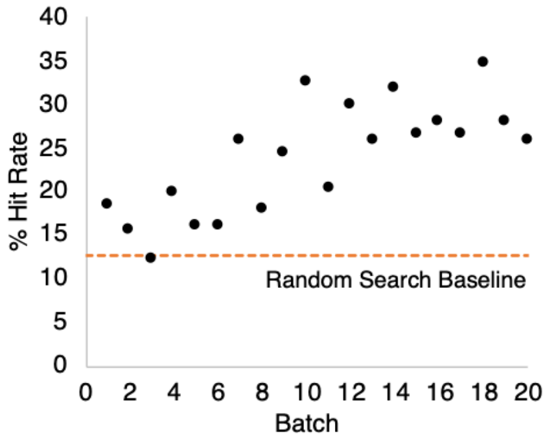


Figure 7. Hit rate of the first 20 iterations of the search with the reset policy. The orange dotted line indicates the hit rate for the random search of 87 molecules which was 13%. A linear regression gave the following equation describing the correlation between the hit rate and batch number: %HR = 0.82(batch) + 15.26; R^2 = 0.57.

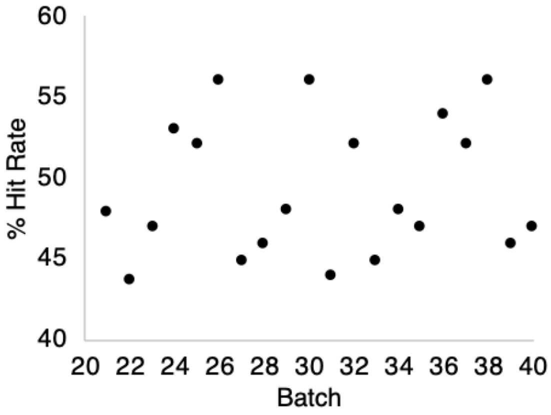


Figure 9. Hit rate of the second 20 iterations of the search policy with 15 cores.

The cores were substituted asymmetrically to systematically enumerate the chemical space. After applying this approach to each of the 29 cores, we created a virtual molecular library of 255 991 azoheteroarene photoswitches.

■ RESULTS AND DISCUSSION

We used a data set of 1436 azoarenes that were previously computed (λ_{max}) using the method described in Figure 2. These azoarenes were generated with a different substitution policy than previously described, and a detailed description can be found in the Supporting Information. Out of 1436, 981 azoarenes had vertical excitation energies that corresponded to $\lambda_{max} > 450$ nm. We initially ran two iterations of the active search with 100 molecules each. These two AS iterations selected molecules from the chemical space of 255 991 molecules. However, we realized that the algorithm would exploit a single-core structure, and the time required to perform 100 calculations for each iteration was expensive. We then decided to create a core restriction policy where the algorithm would equally sample all cores. We retrained the algorithm with the 198 molecules previously selected. Of the 198, 20 molecules absorbed >450 nm. We also decreased the batch size to 50 molecules per batch to decrease the turnaround time for the quantum chemical calculations. A histogram of the λ_{max} of these 198 azoarenes is shown in Figure 5.

Figure 5 shows that the λ_{max} ranges from 301 to 541 nm for the selected 198 azoarenes. To train the AS algorithm, we assigned each candidate a label of `True` or `False`, depending on whether the following expression is satisfied, $\lambda_{max} > 450$ nm. Sixty-two of the 198 azoarenes were assigned `True`, and 136 were assigned `False`.

We then iteratively applied the algorithm 40 times on our new molecular data set. Each molecular batch featured 50 AS-suggested candidates that would enter our computational workflow. The first 20 iterations used an "equidistributed" policy, which equally sampled molecules belonging to each core family of the 29. Since the AS selected 50 molecules for each iteration, we sampled the 29 cores by constraining the algorithm to select at least one molecule per core. The remaining 21 slots for each batch were selected in a similar fashion where no more than two molecules were selected for each core. The remaining iterations (21−40) used a "targeted" policy that only selected molecules from a subset of 15 cores that had previously selected derivatives where $\lambda_{max} > 450$ nm. Cores that did not show derivatives that fit the criteria were excluded from the subset.

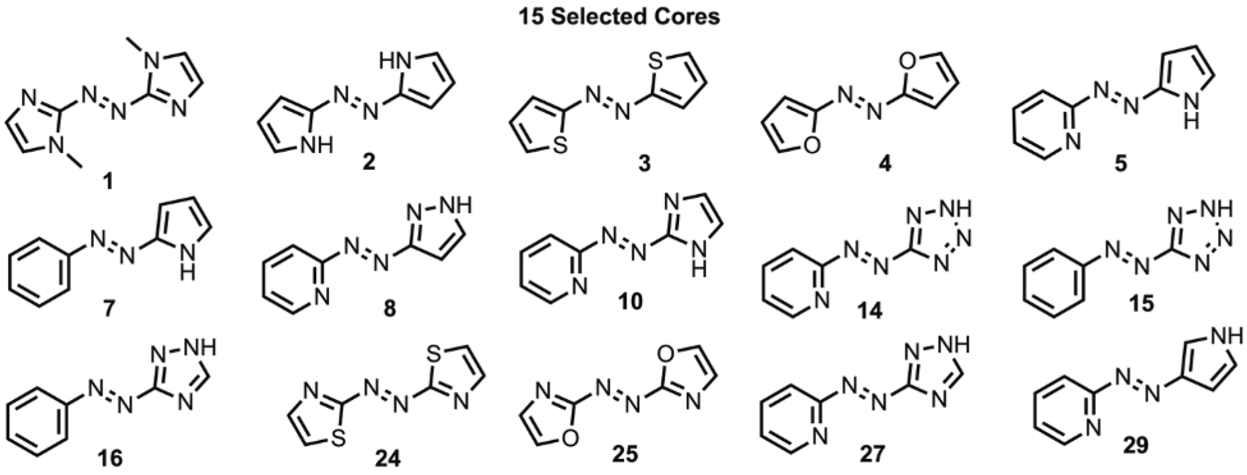


Figure 8. Subset of cores searched for the second half of iterations from 21 to 40. Cores represented yielded at least one substituted molecule that had a λ_{max} exceeding 450 nm.

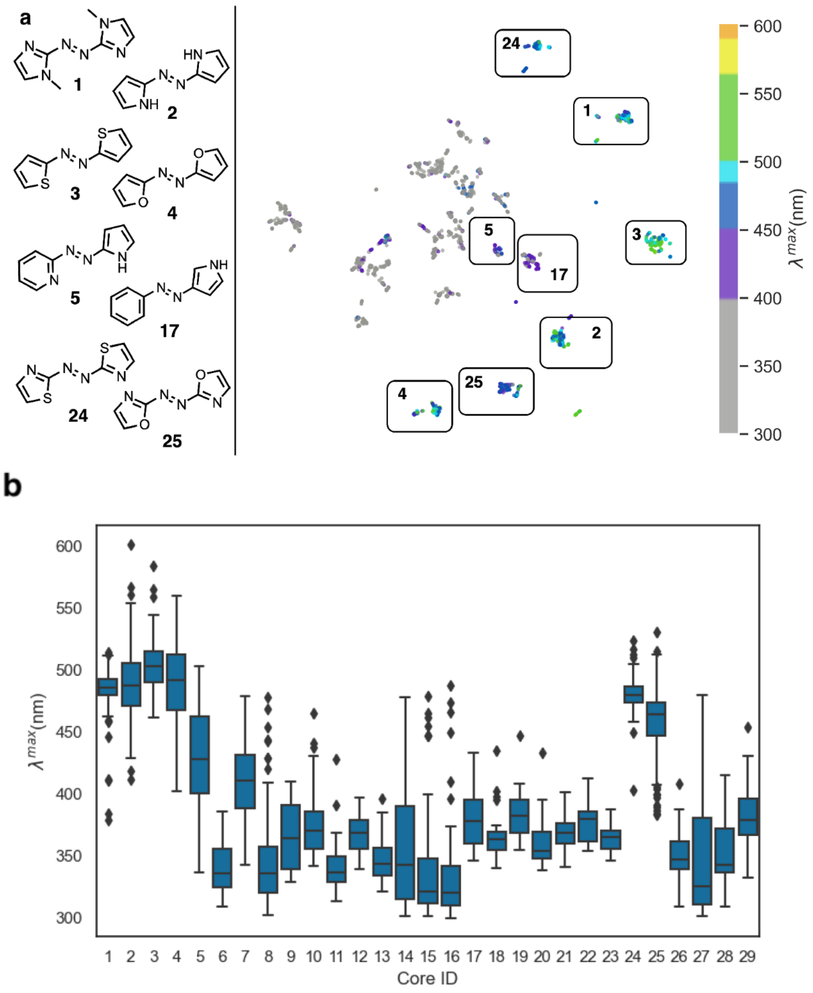


Figure 10. (a) Projection of 1962 azoarene photoswitches suggested by active search using UMAP, computed with a 2048-bit Morgan fingerprint (radius 2), 10 nearest neighbors, a minimum distance of 0.1, and the Tanimoto similarity. (b) Range of λ_{max} of 1962 azoarene photoswitches by core ID. Lines within each box represent the median, while the box represents the interquartile range that includes 50% of values near the median. Tails of each box show the high and low excitation energies of each core ID. Black diamonds represent outliers.

After each iteration, we added a binary label to each molecule based on whether λ_{max} > 450 nm. Figure 3 summarizes this iterative procedure. We compared the AS strategy to the performance of a random search strategy by sampling three molecules selected at random from each of the 29 cores. Figure 6 shows the distribution of the λ_{max} values from AS and the random search.

Figure 6 describes the effect of applying AS. The random search showed that 11 out of the 87 molecules (13%) had λ_{max} > 450 nm. The active search increases the number of molecules that are selected with λ_{max} > 450 nm. The overall increase in average can be seen for bins (501, 551] and (551, 602] where the random search was unable to select any molecules in the latter bin. Figure 7 shows how the proportion of hits change with respect to the first 20 iterations using the equidistributed policy. We define the hit rate as the percentage of molecules with a λ_{max} > 450 nm from the current batch.

The dotted orange line indicates a random search hit rate of 13%. The black data points indicate the hit rate as the active search is iteratively applied. The equidistributed search shows a range of hit rates from 12% to 35% (batch 3 and 18, respectively). The slope is +0.82; the hit rate is improved relative to the random search in nearly all iterations. We then turned our attention to the targeted AS policy to maximize the number of hits corresponding to the subset of cores with molecules that had a λ_{max} > 450 nm, shown in Figure 8.

For iterations 21−40, the AS algorithm selected three derivatives corresponding to each of the 15 cores for a total of 45 selected molecules. To keep the batch size consistent at 50, AS chooses 5 more from the top-ranked derivatives of the 15 core subset. Figure 9 shows the hit rate for iterations 21−40 with the targeted policy.

In the targeted policy, the hit rate varied from 44% to 56%; the average hit rate was 49%. Unlike the equidistributed policy, Figure 9 does not show an increase in the hit rate as a function of the batch number. The relatively high hit rate led to the rapid discovery of 485 candidates with λ_{max} > 450 nm in batches 21−40.

Overall, we identified a total of 717 photoswitches with λ_{max} > 450 nm after the 40 batches (1962 molecules) of AS-assisted virtual screening. The resulting hit rate is 37%, corresponding to a tripling of the 13% hit rate from the random search. A two-sample z-test rejects the null hypothesis that the two strategies result in equal hit rates with overwhelming confidence, yielding a p-value of 5×10^{-6}.

We represented the complex molecular data with a Uniform Manifold Approximation (UMAP)[41] to visualize the molecular motifs responsible for candidates with λ_{max} > 450 nm. Thawani

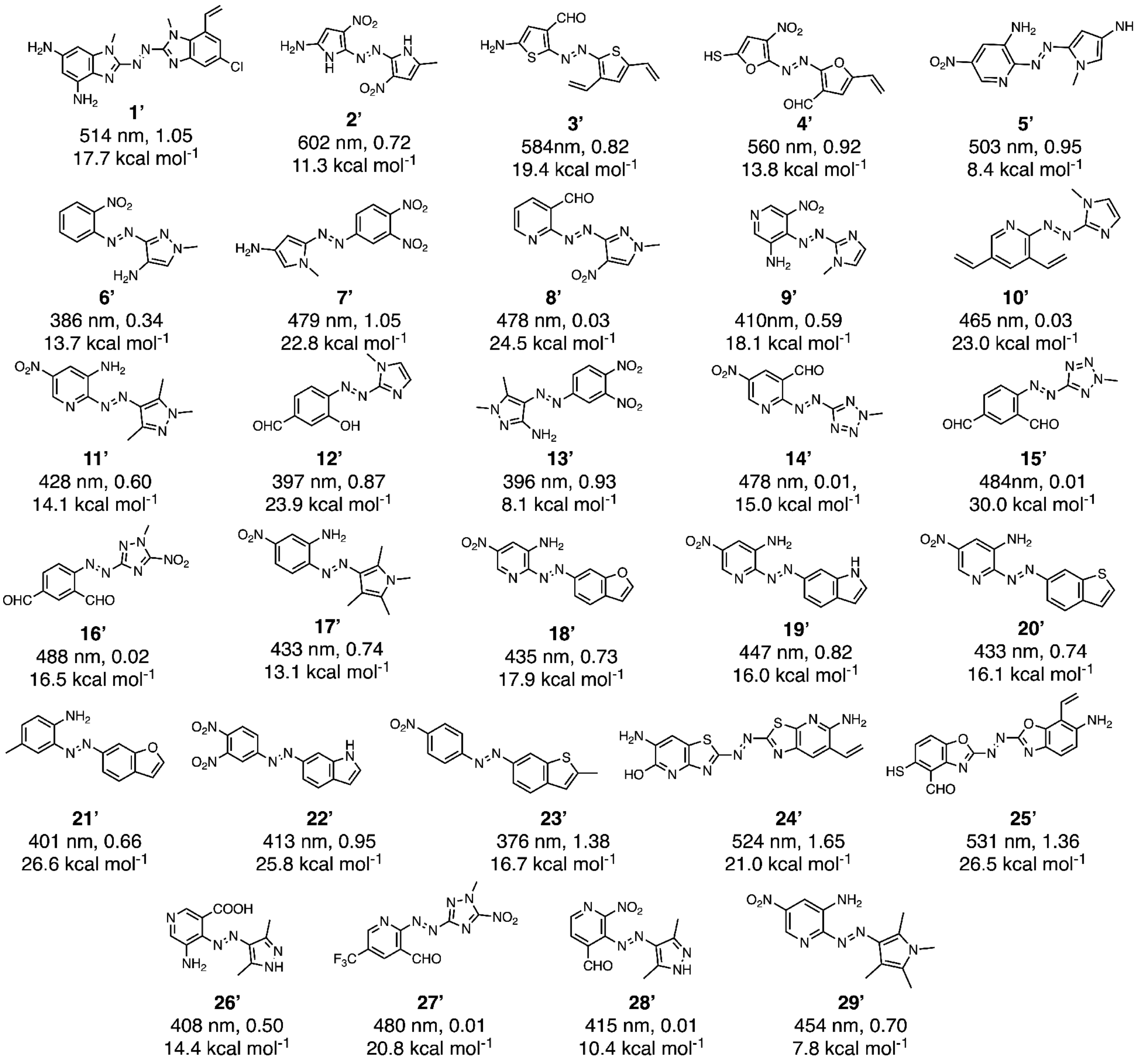


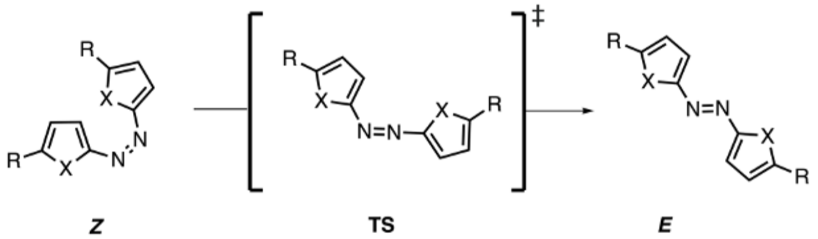
Figure 11. Structures of the 29 highest absorbing azoarene photoswitches for each core. Molecules are labeled by their core ID (in bold), their λ_{max} in nanometers, and the corresponding oscillator strength and activation barrier in kcal mol^{-1}.

et al.[22] previously applied the UMAP algorithm to understand the implications in their choice of molecular representation. They explored the different clusters formed by a Morgan fingerprint-based representation and an RDKit fragment-based representation. They found that the Morgan fingerprint-based representation created more meaningful clusters when compared to the fragment-based representation, which also confirms our use of Morgan fingerprints. We aimed to apply UMAP to explore the relationship between the core structure and the λ_{max} absorbance. Each of the 1962 structures was plotted based on the Tanimoto similarity[33] in Figure 10. The clusters are grouped based on structural similarity and color-coded based on computed λ_{max} results.

Figure 10a shows the UMAP results with each azoarene candidate overlaid with the color corresponding to the λ_{max}. The data points shown in gray correspond to the ultraviolet range of the electromagnetic spectrum (λ_{max} < 400 nm). Cores **1−5**, **17**, **24**, and **25** formed distinct clusters, indicated by the solid lines in the UMAP projection. These cores also had considerably more derivatives with a λ_{max} in the visible range, suggesting that these cores have especially tunable λ_{max} values and should be explored experimentally in the future.

We examined the influence of substituents on each core by plotting the distribution of λ_{max}. Figure 10b shows the range of λ_{max} for 1 962 azoarenes. Spacings within each box represent the degree of dispersion and skewness within the data. Cores with larger boxes indicate a higher variation in absorbance due to the substitution pattern. We compared unsubstituted cores **1−5**, **17**, **24**, and **25** to the derivative with the highest λ_{max}. These values are summarized in Table S2 of the Supporting Information. **1** showed the highest λ_{max} at 514 nm with a range of 139 nm. **2** had the largest λ_{max} value of 602 nm and featured an impressive range of 213 nm within the corresponding derivatives. This suggests that the family of derivatives corresponding to **2** has the most tunable λ_{max}. **3**, **4**, and **5** had their highest absorbing derivatives at 584, 560, and 503 nm, with similar ranges at 193, 186, and 166 nm, respectively. **24** and **25** had their largest λ_{max} values at 524 and 531 nm, respectively. Their derivatives had ranges of 121 and 148 nm, respectively.

Scheme 2. Illustration of the $Z \rightarrow E$ Thermal Isomerization Transition Structure



The ideal $t_{1/2}$ of photoswitches depends on the desired application. The $t_{1/2}$ and λ_{max} are typically in competition because the π-delocalization effects that generally red-shift the λ_{max} also decrease the $t_{1/2}$ by lowering the transition state energies. However, longer $t_{1/2}$ values are generally desirable; we chose those candidates with $t_{1/2} > 2$ h as "hits". Determining $t_{1/2}$ values requires the computation of $Z \rightarrow E$ thermal isomerization transition structures, which reveal the activation free energies. Adrion et al.[20] recently benchmarked 140 model chemistries to predict azoarene isomerization barriers and published the open-access code, *EZ*-TS. We thus applied *EZ*-TS to compute the $t_{1/2}$ of the *Z*-isomers of core derivatives with the longest λ_{max}, identified with the active search. Figure 11 illustrates the candidate from each family of cores subjected to transition state calculations with PBE0-D3/6-31+G(d,p) to optimize the transition states.[42] This was reported to give activation free energies that approach chemical accuracy. Scheme 2 shows the $Z \rightarrow E$ isomerization transition state.

The λ_{max} for these top 29 candidates ranges from 382 to 602 nm. The range of activation free energies is 8.1−30.0 kcal mol^{-1}. We plotted the activation free energies ($\Delta G^{\ddagger}$) against the λ_{max} for these 29 candidates to determine if there was a relationship between these values (Figure 12).

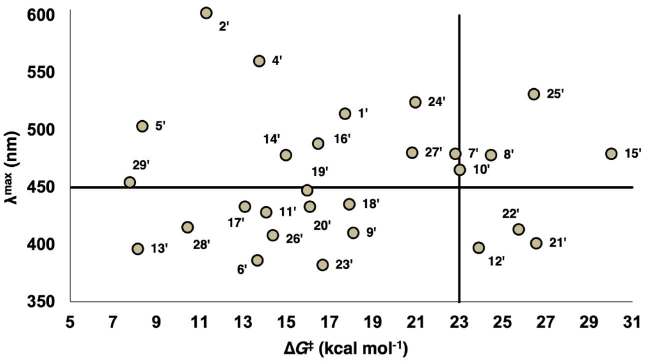


Figure 12. Activation free energy against the λ_{max} of 29 azoarene photoswitches selected by the active search. Their core ID indexes the data points. Quadrant B is where both criteria for an ideal photoswitch (λ_{max} > 450 nm and $\Delta G^{\ddagger}$ > 23 kcal mol^{-1}) have been satisfied. Quadrants A and D are where one criterion has been satisfied, and Quadrant C is where none of the criteria have been satisfied.

Figure 12 shows no linear relationship between the λ_{max} and activation free energy (R^2 of 0.0002). However, we divided the plot into four quadrants to highlight those candidates that meet both, one, or none of the λ_{max} and $t_{1/2}$ optimization criteria. Quadrants A and B contain molecules that have λ_{max} > 450 nm or 2.6 eV. Quadrants A and C are populated with molecules with an activation free energy less than 23.0 kcal mol^{-1}. Quadrants C and D contain molecules that absorb UV light or have λ_{max} greater than 450 nm. Quadrants B and D have molecules with an activation free energy greater than 23.0 kcal mol^{-1}. The ideal candidates fall in Quadrant B that satisfy both criteria; Quadrants A and D are partially optimized; Quadrant C has candidates that do not meet any of the requirements. Molecules **8′**, **10′**, **15′**, and **25′** have a high λ_{max} value of 478, 465, 479, and 531 nm, respectively. They also have high activation free energies of 24.5, 23.0, 30.0, and 26.5 kcal mol^{-1}, respectively. Figure 12 shows the λ_{max} and activation free energies of the highest absorbing molecules for each core. The activation free energies represent the energy required for the *Z*-isomer to revert back to the *E*-isomer. The larger the activation free energy, the more stable the *Z*-isomer.[43] Depending on application, tuning the half-life of the *Z*-isomer is highly desirable. Core structures presented in Figure 12 can be used to influence molecular design for applications that may require long-lived *Z*-isomers in order to activate therapeutic properties, for example, targeted protein degradation.[44] By providing a set of starting structures with long-lived *Z* half-lives, we can increase the number of therapeutic modalities that incorporate photoswitches for site specific treatment.

■ CONCLUSION AND FUTURE WORK

We created a molecular data set of 255 991 azoarenes to find photoswitches with high λ_{max} values and high activation energies for therapeutic applications. We leveraged quantum mechanical calculations to sample just 1% of the search space and computed 2117 DFT calculations in total over 40 iterations. The iterative process of applying AS to photoswitch screening was highly effective and tripled the discovery rate of novel photoswitches compared to a random search. The AS algorithm identified 717 photoswitches with high λ_{max} values ranging from 451 to 602 nm. We also conducted a second layer of screening to identify photoswitches with long. We computed the activation free energies of 29 photoswitches for the molecules with the largest λ_{max} by core identified by the active search. As with previous studies, we found no general correlation between the activation free energies and λ_{max} values. However, we identified four azoarene photoswitches with high λ_{max} values and high activation free energies. We are currently applying the AS technique to identify long-lived *Z*-isomers and will report these findings in due course. There are also other avenues that could be explored related to representing molecular data in machine learning applications. Specifically, RDKit fragments could replace Morgan fingerprints when training. New 3D fingerprints would also be useful to implement when predicting the half-lives in order to better understand structure−property relationships.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00954.

A description of the data and code we release with the submission and detailed search results in each iteration of our procedure (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Steven A. Lopez − *Department of Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States;* orcid.org/0000-0002-8418-3638; Email: s.lopez@northeastern.edu

Authors

Fatemah Mukadum − *Department of Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States*

Quan Nguyen − *Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, United States*

Daniel M. Adrion − *Department of Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States*

Gabriel Appleby − *Department of Computer Science, Tufts University, Medford, Massachusetts 02155, United States*

Rui Chen − *Department of Computer Science, Tufts University, Medford, Massachusetts 02155, United States*

Haley Dang − *Department of Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States*

Remco Chang − *Department of Computer Science, Tufts University, Medford, Massachusetts 02155, United States*

Roman Garnett − *Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00954

Notes

The authors declare no competing financial interest.

We release all data and software used in this work, as detailed in our Supporting Information. Specifically, our data are hosted at https://figshare.com/articles/figure/Efficient_discovery_of_visible_light-activated_azoarene_photoswitches_with_long_half-lives_using_active_search_and_quantum_chemical_calculations/15093933, while code implementations of the scripts and algorithms described in our work are hosted at https://github.com/KrisNguyen135/photoswitch.

■ ACKNOWLEDGMENTS

All authors acknowledge the National Science Foundation (NSF-OAC-1940307) for funding this research. Q.N. and R.G. additionally acknowledge support under award number NSF IIS-1845434. F.M. and S.L. appreciate the assistance from the Northeastern Research Computing Team and access to the computing resources of the Discovery cluster. F.M. and S. L. also acknowledge the Massachusetts Life Science Center (G00006360) for providing computational resources to make this research possible.

■ REFERENCES

(1) Griffiths, J. Photochemistry of Azobenzene and its Derivatives. *Chem. Soc. Rev.* **1972**, *1*, 481−493.

(2) Lawrence, K. P.; Douki, T.; Sarkany, R. P.; Acker, S.; Herzog, B.; Young, A. R. The UV/Visible Radiation Boundary Region (385—405 nm) Damages Skin Cells and Induces "Dark" Cyclobutane Pyrimidine Dimers in Human Skin *in vivo*. *Sci. Rep.* **2018**, *8*, 1−12.

(3) Jia, S.; Sletten, E. M. Spatiotemporal Control of Biology: Synthetic Photochemistry Toolbox with Far-Red and Near-Infrared Light. *ACS Chem. Biol.* **2021**,.

(4) Broichhagen, J.; Frank, J. A.; Trauner, D. A Roadmap to Success in Photopharmacology. *Acc. Chem. Res.* **2015**, *48*, 1947−1960. PMID: 26103428.

(5) Konrad, D. B.; Savasci, G.; Allmendinger, L.; Trauner, D.; Ochsenfeld, C.; Ali, A. M. Computational Design and Synthesis of a Deeply Red-Shifted and Bistable Azobenzene. *J. Am. Chem. Soc.* **2020**, *142*, 6538−6547.

(6) Stricker, L.; Böckmann, M.; Kirse, T. M.; Doltsinis, N. L.; Ravoo, B. J. Arylazopyrazole Photoswitches in Aqueous Solution: Substituent Effects, Photophysical Properties, and Host−Guest Chemistry. *Chem. - Eur. J.* **2018**, *24*, 8639−8647.

(7) Huddleston, P. R.; Volkov, V. V.; Perry, C. C. The Structural and Electronic Properties of 3, 3′-Azothiophene Photo-switching Systems. *Phys. Chem. Chem. Phys.* **2019**, *21*, 1344−1353.

(8) Weston, C. E.; Richardson, R. D.; Haycock, P. R.; White, A. J.; Fuchter, M. J. Arylazopyrazoles: Azoheteroarene Photoswitches Offering Quantitative Isomerization and Long Thermal Half-Lives. *J. Am. Chem. Soc.* **2014**, *136*, 11878−11881.

(9) Calbo, J.; Weston, C. E.; White, A. J.; Rzepa, H. S.; Contreras-García, J.; Fuchter, M. J. Tuning Azoheteroarene Photoswitch Performance Through Heteroaryl Design. *J. Am. Chem. Soc.* **2017**, *139*, 1261−1274.

(10) Slavov, C.; Yang, C.; Heindl, A. H.; Wegner, H. A.; Dreuw, A.; Wachtveitl, J. Thiophenylazobenzene: An Alternative Photoisomerization Controlled by Lone-Pair ···π Interaction. *Angew. Chem., Int. Ed.* **2020**, *59*, 380−387.

(11) Okumura, S.; Lin, C.-H.; Takeda, Y.; Minakata, S. Oxidative Dimerization of (Hetero)aromatic Amines Utilizing t-BuOI Leading to (Hetero)aromatic Azo Compounds: Scope and Mechanistic Studies. *J. Org. Chem.* **2013**, *78*, 12090−12105.

(12) Abburu, S.; Venkatraman, V.; Alsberg, B. K. TD-DFT Based Fine-Tuning of Molecular Excitation Energies Using Evolutionary Algorithms. *RSC Adv.* **2016**, *6*, 3661−3670.

(13) Luo, Y.-W.; Chou, C.-H.; Lin, P.-C.; Chiang, C.-M. Photochemical Synthesis of Azoarenes From Aryl Azides on Cu(100): A Mechanism Unraveled. *J. Phys. Chem. C* **2019**, *123*, 12195−12202.

(14) Chansen, W.; Yu, J.-S. K.; Kungwan, N. A TD-DFT Molecular Screening for Fluorescence Probe Based on Excited-State Intramolecular Proton Transfer of 2-Hydroxychalcone Derivatives. *J. Photochem. Photobiol., A* **2021**, *410*, 113165.

(15) Lopez, S. A.; Sanchez-Lengeling, B.; de Goes Soares, J.; Aspuru-Guzik, A. Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics. *Joule* **2017**, *1*, 857−870.

(16) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120−1127.

(17) Kim, S.; Noh, J.; Gu, G. H.; Aspuru-Guzik, A.; Jung, Y. Generative Adversarial Networks for Crystal Structure Prediction. *ACS Cent. Sci.* **2020**, *6*, 1412−1420.

(18) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.

(19) Abreha, B. G.; Agarwal, S.; Foster, I.; Blaiszik, B.; Lopez, S. A. Virtual Excited State Reference for the Discovery of Electronic Materials Database: An Open-Access Resource for Ground and Excited State Properties of Organic Molecules. *J. Phys. Chem. Lett.* **2019**, *10*, 6835−6841.

(20) Adrion, D. M.; Kaliakin, D. S.; Neal, P.; Lopez, S. A. Benchmarking of Density Functionals for Z-Azoarene Half-Lives via

Automated Transition State Search. *J. Phys. Chem. A* **2021**, *125*, 6474−6485. PMID: 34260236.

(21) Ju, C.-W.; Bai, H.; Li, B.; Liu, R. Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields. *J. Chem. Inf. Model.* **2021**, *61*, 1053−1065.

(22) Thawani, A. R.; Griffiths, R.-R.; Jamasb, A.; Bourached, A.; Jones, P.; McCorkindale, W.; Aldrick, A. A.; Lee, A. A. *The Photoswitch Dataset: A Molecular Machine Learning Benchmark for the Advancement of Synthetic Chemistry*; arXiv preprint arXiv:2008.03226 [physics.chem-ph]; 2020.

(23) Garnett, R.; Krishnamurthy, Y.; Xiong, X.; Schneider, J.; Mann, R. Bayesian Optimal Active Search and Surveying. *Proceedings of the 29th International Conference on Machine Learning*, 2012.

(24) Landrum, G. *RDKit: Open-Source Cheminformatics Software*; Open-source software, 2016.

(25) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(26) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard, W. A., III; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(27) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−241.

(28) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XXIII. A Polarization-Type Basis Set for Second-Row Elements. *J. Chem. Phys.* **1982**, *77*, 3654−3665.

(29) Ditchfield, R.; Hehre, W.; Pople, J. Self-Consistent Molecular Orbital Methods. VI. Energy Optimized Gaussian Atomic Orbitals. *J. Chem. Phys.* **1970**, *52*, 5001−5007.

(30) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999−3094.

(31) Chai, J.-D.; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom−Atom Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615−6620.

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

(33) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(34) Fix, E.; Hodges, J. L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review/Revue Internationale de Statistique* **1989**, *57*, 238−247.

(35) Garnett, R.; Gärtner, T.; Vogt, M.; Bajorath, J. Introducing the 'Active Search' Method for Iterative Virtual Screening. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 305−314.

(36) Jiang, S.; Malkomes, G.; Moseley, B.; Garnett, R. Efficient Nonmyopic Active Search with Applications in Drug and Materials Discovery. *Machine Learning for Molecules and Materials Workshop at NeurIPS*; 2018.

(37) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating High-Throughput Virtual Screening Through Molecular Pool-Based Active Learning. *Chem. Sci.* **2021**, *12*, 7866.

(38) Jiang, S.; Malkomes, G.; Converse, G.; Shofner, A.; Moseley, B.; Garnett, R. Efficient Nonmyopic Active Search. *Proceedings of the 34th International Conference on Machine Learning*; 2017; pp 1714−1723.

(39) Jiang, S.; Malkomes, G.; Abbott, M.; Moseley, B.; Garnett, R. Efficient Nonmyopic Batch Active Search. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1099−1109.

(40) Bertsekas, D. P. *Dynamic Programming and Optimal Control*; Athena Scientific: Belmont, MA, 1995; Vol. *1*.

(41) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861.

(42) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.

(43) Knie, C.; Utecht, M.; Zhao, F.; Kulla, H.; Kovalenko, S.; Brouwer, A. M.; Saalfrank, P.; Hecht, S.; Bléger, D. ortho-Fluoroazobenzenes: Visible Light Switches with Very Long-Lived Z Isomers. *Chem. - Eur. J.* **2014**, *20*, 16492−16501.

(44) Reynders, M.; Trauner, D. In *Targeted Protein Degradation: Methods and Protocols*; Cacace, A. M., Hickey, C. M., Békés, M., Eds.; Springer US: New York, 2021; pp 315−329.