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 Lopez[*,†]

†*Department of Chemistry and Chemical Biology, Northeastern University*

‡*Department of Computer Science and Engineering, Washington University in St. Louis*

¶*Department of Computer Science, Tufts University*

E-mail: s.lopez@northeastern.edu

Abstract

Photoswitches are molecules that undergo a reversible, structural isomerization after exposure to different wavelengths of light. The dynamic control offered by molecular photoswitches is favorable for applications in materials chemistry, photopharmacology, and catalysis. 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-lives ($t_{1/2}$) of the Z-isomer. However, computing the photophysical and kinetic properties of each entry of a virtual molecular library containing 10^3–10^6 entries with density functional theory 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 based 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 1 962 photoswitches to 2D using the Uniform Manifold Approximation and Projection (UMAP) algorithm and found that $\lambda_{\max}$ depends on the core, which is tunable with substituents. We then incorporated a second stage of screening with to predict the stabilities of the Z-isomers for the top 1% of candidates. We identified four ideal photoswitches that concurrently satisfy $\lambda_{\max} >$ 450 nm and $t_{1/2} > 2$ hours; the range of $\lambda_{\max}$ 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 meta-stable 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 $Z \rightarrow E$ 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)[2] tissue penetration depth, thus limiting the therapeutic potential of photoswitches in photopharmacology. The Z-isomer of azobenzene has a thermal half-life ($t_{1/2}$) of 4.7 hours, which prevents the establishment of photostationary states. Ideal photoswitches 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.[3] recently demonstrated that functionalizing the phenyl rings with halogens at the ortho positions led to a substantial red shift to 410 nm. This functionalization strategy also increased the ($t_{1/2}$) to 16 hours. 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 $\lambda_{\max}$ in the visible range and $t_{1/2}$ exceeding 1.5 hours. Figure 1 highlights some of the most promising synthesized azoarenes with respect to $\lambda_{\max}$ and $t_{1/2}$.[3–9]

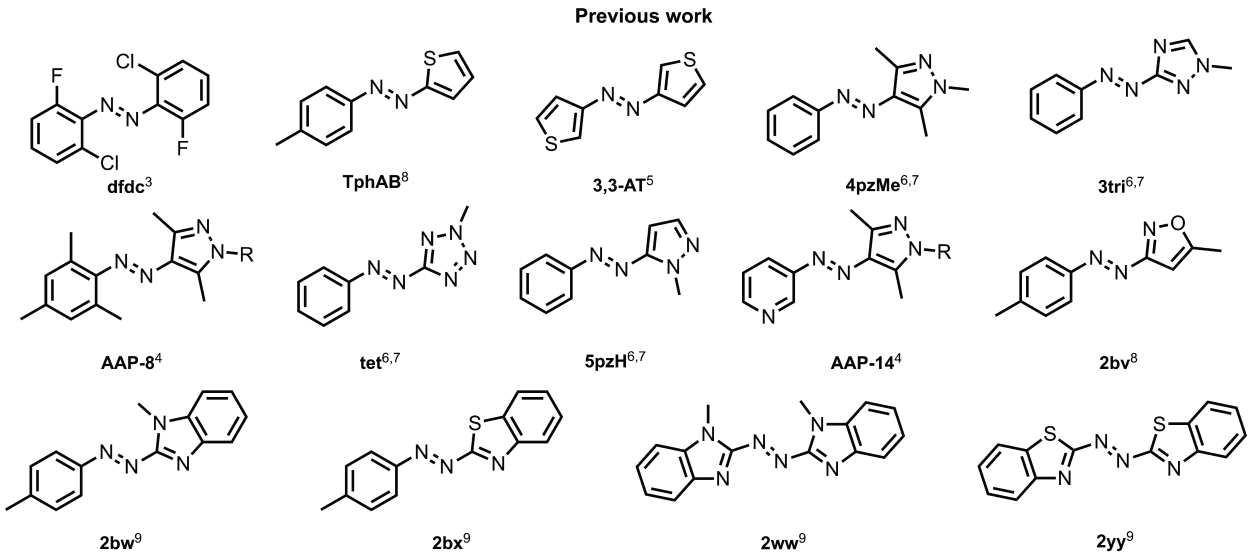


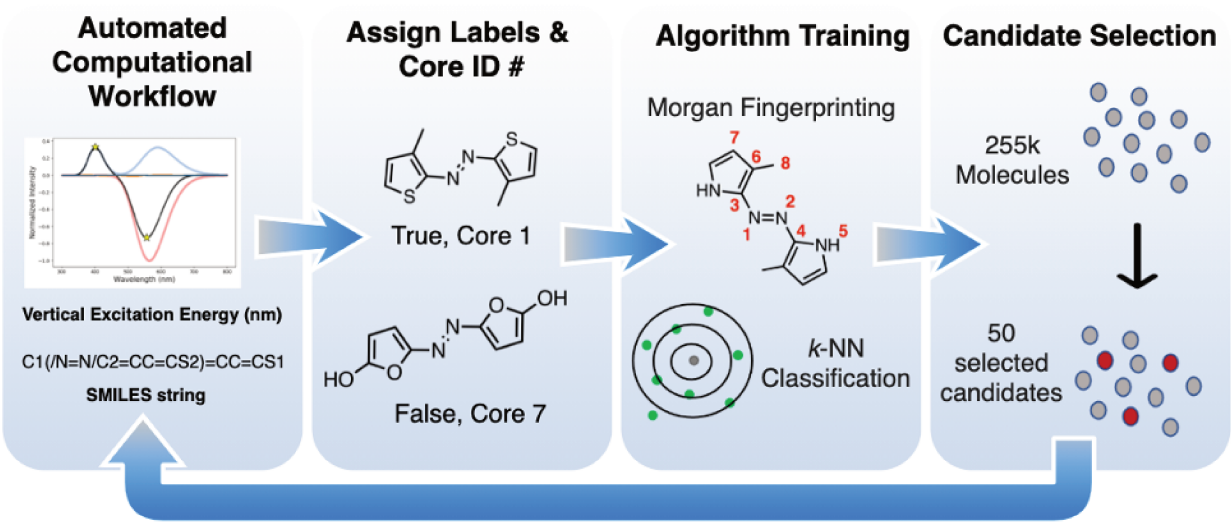
Figure 1: 14 azoarene photoswitches used to generate a new molecular library.

While this relatively new class of azoarene photoswitches is attractive, the full 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.[10,11] Thus, DFT has been previously used in high throughput virtual screening (HTVS)[12–15] for virtual libraries containing 500–500 000 molecules. The vastness of the chemical space cannot be understated; conservative estimates suggest that 10^{23} organic molecules are theoretically possible.[16] This figure can be narrowed to roughly 10^6 for azoarenes by focusing on those already experimentally realized. Abreha et al.[17] 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 (S0, S1, and T1), photophysical, and redox properties. Further, Adrion et al.[18] published the *EZ*–TS code,

which predicts thermal $Z \rightarrow E$ activation barriers efficiently and accurately.

Even with high-performance computing and efficient quantum chemistry codes, computing the photophysical properties and stabilities of 10^5 photoswitches is a substantial undertaking. We have employed the machine learning algorithm 'active search'[19] 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 informed by the latest observations. The prediction accuracy of our predictive model improves as we frequently query from quantum chemical calculations.

We now combine these existing tools (the VERDE materials DB,[17] EZ-TS,[18] and active search[19]) to automatically identify top photoswitch candidates featuring visible-light $\lambda_{\max}$ and long $t_{1/2}$. Scheme 1 shows an illustration of the iterative processes used to identify ideal photoswitches



Scheme 1: The multipronged iterative procedure used to update the active search algorithm with DFT results.

Phase 1: An initial screen of 50–100 molecules is processed through an automated computational workflow developed by Abreha et al.[17]. RDKit[20] is used to generate 3-D coordinates from a simplified molecular-input line-entry system (SMILES)[21] string, followed by a low-

mode conformational search where each conformer (4 total) is minimized with the Universal Force Field.[22] The lowest energy conformer is determined through semi-empirical optimizations and a single-point energy calculation. The lowest energy structure is optimized with M06[23]/6-31+G(d,p)[24,25] and IEFPCM^{MeCN},[26] and a vibrational analysis confirms the stationary point as the true minimum if it has only positive frequencies. The $\lambda_{\max}$ is calculated with a single point energy calculation using ωB97XD[27]/6-31+G(d,p)//M06[23]/6-31+G(d,p). Figure 2 shows the automated workflow of quantum chemical calculations used to compute the excitation energies and corresponding $\lambda_{\max}$ for selected molecules from our virtual library.

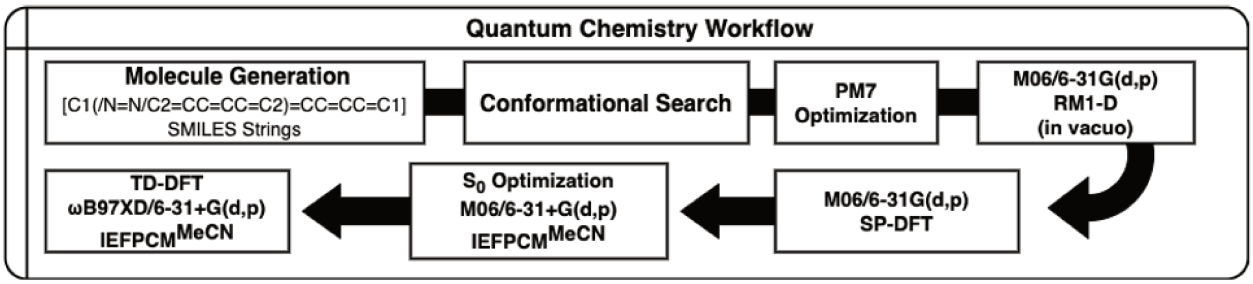


Figure 2: Quantum chemical workflow for computing the $\lambda_{\max}$ for all molecules considered in this study.

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 pre-determined threshold, $\lambda_{\max} > 450$nm.

Phase 3: A machine learning model is trained on the set of molecules that are labeled to guide the search algorithm. First, we generate the Morgan fingerprint[28] of each molecule and compute the Tanimoto similarity coefficient[29] between each pair of molecules. We then build a k-nearest neighbors (k-NN)[30] 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. Note that 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. These recommendations are then sent to Phase 1 to be computed and labeled. This procedure repeats for a total of 40 iterations, sampling 1 962 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.[31–35] The method was first introduced by Garnett et al.[19] and extended to the batch setting by Jiang et al.[35]. 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 $y = \mathbb{1}\{x \in \mathcal{R}\}$, where $\mathbb{1}\{\cdot\}$ is the indicator function. In this work, the binary label denotes whether a molecule exceeds the $\lambda_{\max}$ threshold of 450nm. 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 $T = t\,b$ to be the total number of queries allowed to be made given our budget, where t is the number of search 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 dataset at termination $\mathcal{D}_t$:

$$X_{i+1} = \arg\max_{X} \mathbb{E}\Big[u\left(\mathcal{D}_T\right) \mid X, \mathcal{D}_i\Big].$$

Although this expected utility can be derived using the standard procedure of backward induction,[36] 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.[35] as an efficient approximation to the optimal batch of queries. First, the strategy builds on the efficient nonmyopic search algorithm ENS[34] 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.[34] 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.[35]

showed 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.

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 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[28] of each molecule in our search space and compute the Tanimoto similarity coefficient[29] between each pair of elements x and x', denoted as $t(x, x')$. We then implement a k-nearest neighbor (k-NN)[30] predictive model, which computes the probability of an uninspected molecule being an active compound as:

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

where $\mathrm{NN}(x)$ is the labeled subset of the k nearest neighbors of x in X. γ is a parameter of the model that acts as a "pseudo count" 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,[19,31,32,34,35] as well as in our experiment. It can further be rapidly updated in light of new observations, allowing for efficient lookahead computations that are central in active search.

Results and discussion

We generated a relatively small virtual molecular library of 1 636 azobenzene, bisazopyrrole, bisazothiophene, and bisazofuran photoswitches (Figure 3). The substituent sites (red circles) were replaced with the disubstituted alkenyl, alkynyl, or aryl (spacer) groups. The unfunctionalized end of the growing molecule (–R) was substituted with functional (terminal) groups. Figure 3 shows the sites where a set of 4 azoarene cores were substituted with spacer and terminal groups to generate the initial training set.

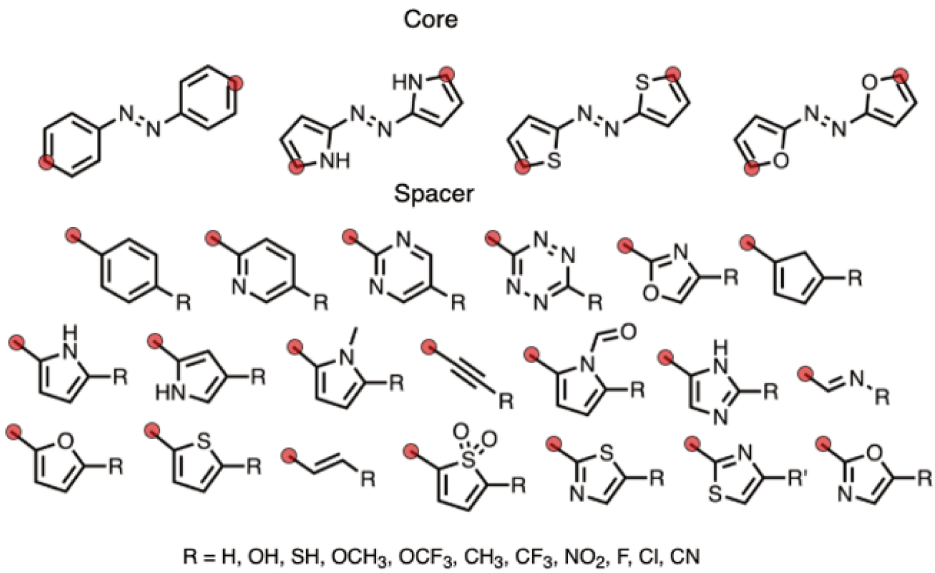


Figure 3: The combinatorial method was used to generate an initial library of 1 636 photoswitches with four azobenzene and azoarene core structures. An in-house Python algorithm symmetrically substitutes red circles with spacer groups and R with H, OH, SH, OCH_3, OCF_3, CH_3, CF_3, NO_2, F, Cl, or CN.

From the 1 636 initial azoarenes, 198 were selected to train the active search algorithm. A histogram of the $\lambda_{\max}$ of these 198 azoarenes is shown in Figure 4.

Figure 4 shows that the $\lambda_{\max}$ 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. 62 of the 198 azoarenes were assigned `True` and 136 were assigned `False`. We designed a virtual molecular library

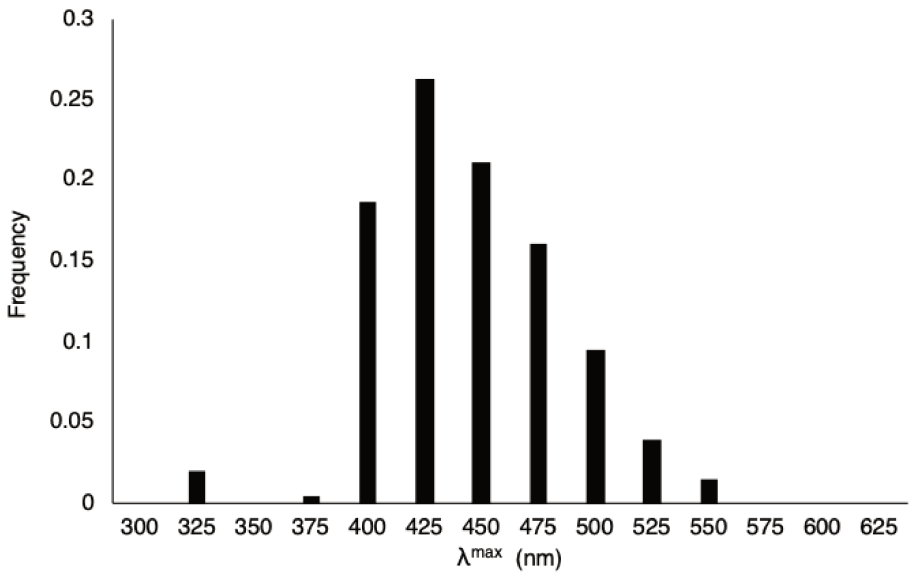


Figure 4: Distribution of the $\lambda_{\max}$ values of the photoswitch training set.

with 29 bisdiazoarene cores (Figure 5) to apply the trained algorithm. 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 6 describes these positions for a smaller subset of cores.

We then iteratively applied the algorithm 40 times on our new molecular dataset. 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 derivatives where the $\lambda_{\max} > 450$ nm. Cores that did not show derivatives that fit the criteria were excluded from the subset. After each iteration, we added a binary label to each

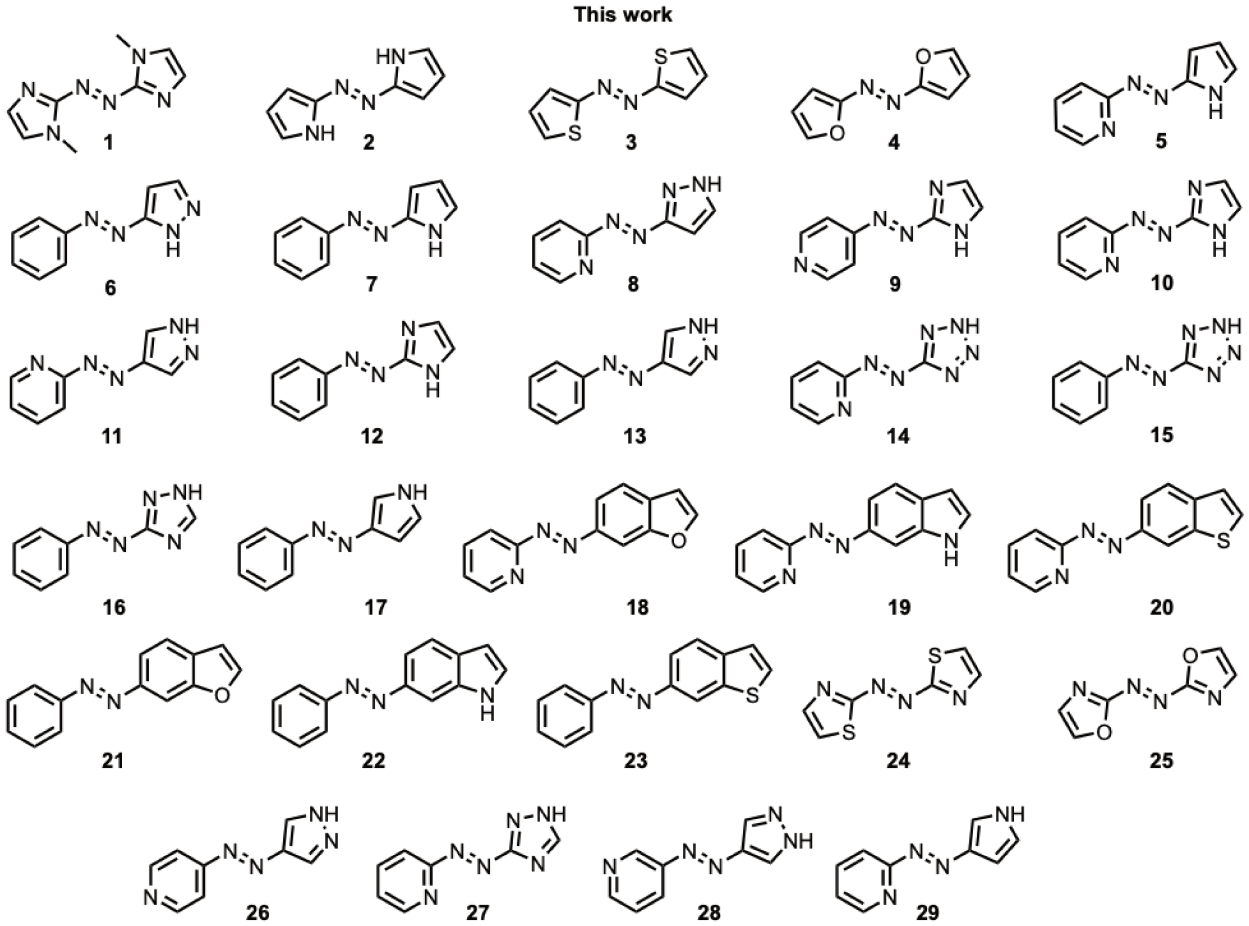


Figure 5: 29 cores explored in this study.

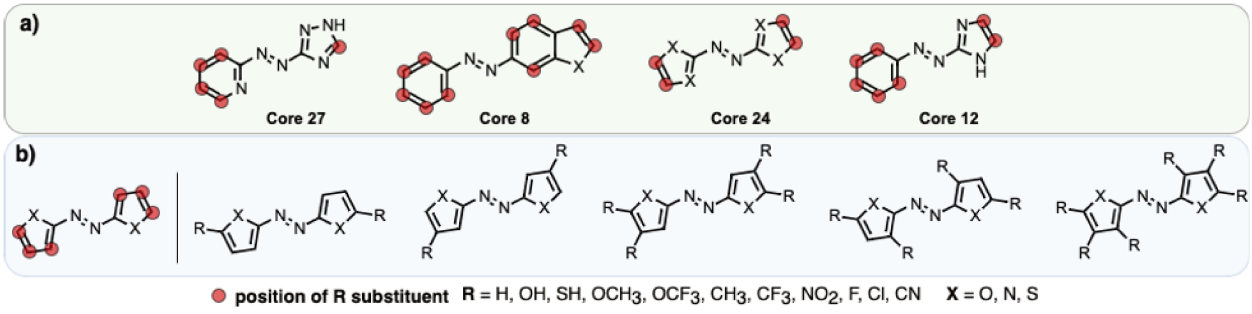


Figure 6: A schematic representation of the substitution patterns of azoheteroarene cores. a) a subset of 4 cores from the 29. b) The cores were substituted asymmetrically to enumerate the chemical space systematically. Red circles indicate positions substituted asymmetrically with terminal groups from Figure 2, H, OH, SH, OCH_3, OCF_3, CH_3, CF_3, NO_2, F, Cl, or CN, and X represents 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).

molecule based on whether $\lambda_{\max} > 450$ 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 7 shows the distribution of the $\lambda_{\max}$ values from AS and the random search.

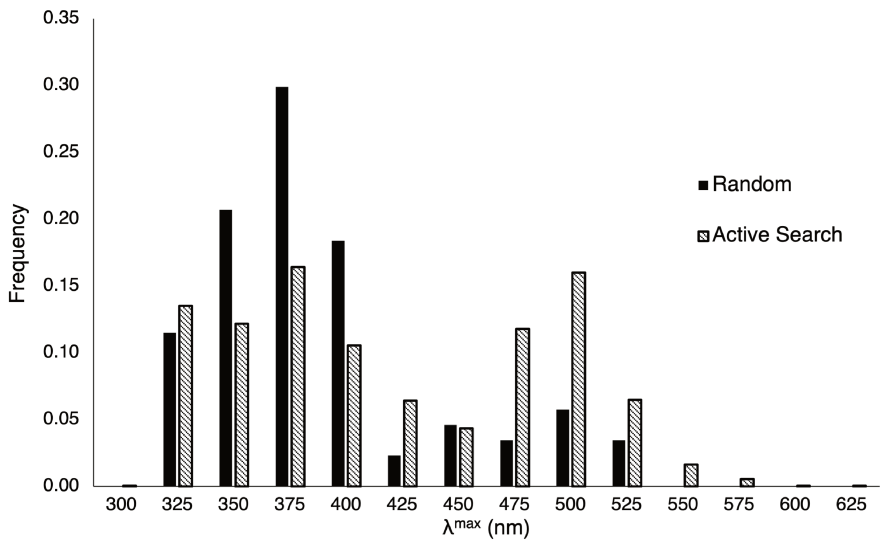


Figure 7: Distribution of the random search compared to active search. Three molecules are sampled for each core, resulting in a total of 87 randomly selected molecules. Active search calculations entail 1 962 computed azoarenes. The values are normalized, and the bin size is 25nm.

The random search showed that 11 out of the 87 molecules (13%) had $\lambda_{\max} > 450$ nm. Figure 8 shows how the proportion of hits changes with respect to the first 20 iterations using the equidistributed policy. We define the hit rate as the percentage of molecules with a $\lambda_{\max} > 450$ 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 $\lambda_{\max} > 450$ nm, shown in Figure 9.

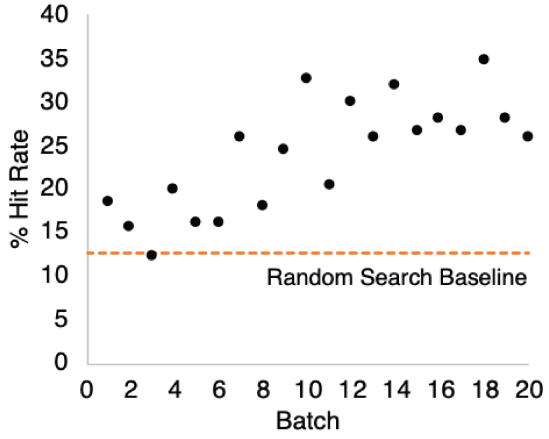


Figure 8: The hit rate of the first 20 iterations of the search with the reset policy. The orange dotted line indicates the hit rate for the random search of 87 molecules which was 13%. A linear regression gave the following equation describing the correlation between the hit rate and batch number, [%HR=0.82(batch) + 15.26] with an R^2 of 0.57.

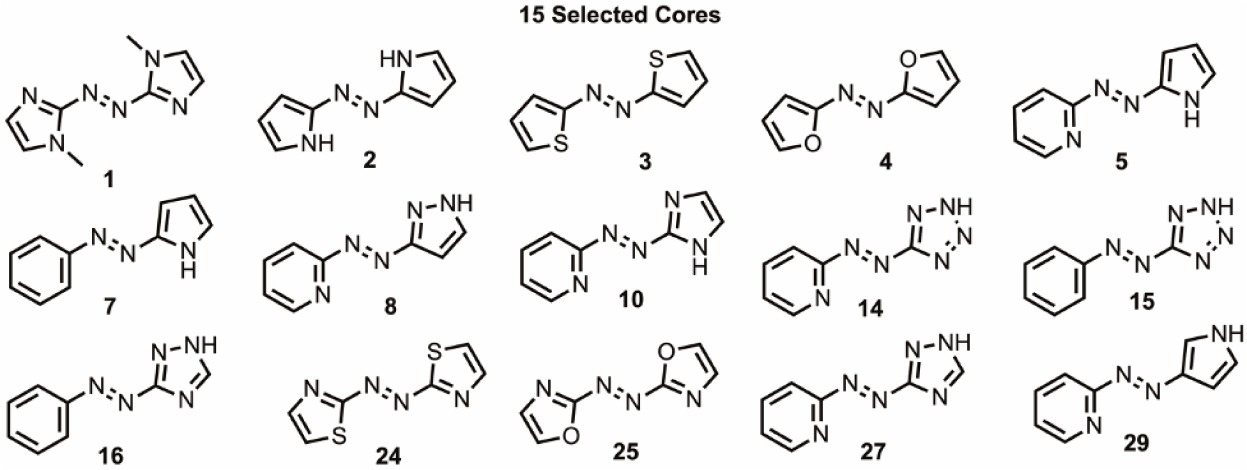


Figure 9: A subset of cores searched for the second half of iterations from 21–40. Cores represented yielded at least one substituted molecule that had a $\lambda_{\max}$ exceeding 450 nm.

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 to 50, AS chooses five more from the top-ranked derivatives of the 15 core subset. Figure 10 shows the hit rate for iterations 21–40 with the targeted policy.

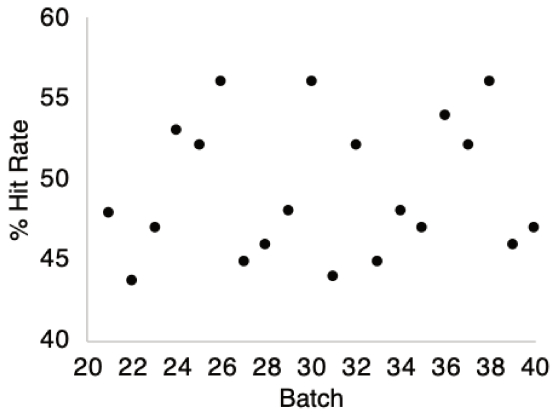


Figure 10: The hit rate of the second 20 iterations of the search policy with 15 cores.

In the targeted policy, the hit rate varied from 44% to 56%; the average hit rate was 49%. Unlike the equidistributed policy, Figure 10 does not show an increase in hit rate as a function of the batch number. The relatively high hit rate led to the rapid discovery of 485 candidates with $\lambda_{\max} > 450$ nm in batches 21–40.

Overall, we identified a total of 717 photoswitches with $\lambda_{\max} > 450$ nm after the 40 batches (1 962 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)[37] to visualize the molecular motifs responsible for candidates with $\lambda_{\max} > 450$ nm. Each of the 1 962 structures was plotted based on the Tanimoto similarity[29] in Figure 11. The clusters are grouped based on structural similarity and color-coded based on computed

λ_{max} results.

Figure 11a shows the UMAP results with each azoarene candidate overlaid with the color corresponding to the λ_{max}. The data points shown in grey correspond to the ultraviolet range of the electromagnetic spectrum ($\lambda_{\text{max}} < 400$ nm). Cores **1**–**5**, **17**, **24**, and **25** formed distinct clusters, indicated by the dotted lines in the UMAP plot. 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 11b 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 602nm 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.

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 $t_{1/2} > 2$ hours as 'hits.' Determining $t_{1/2}$ values requires the computation of $Z \rightarrow E$ thermal isomerization transition structures, which inform the activation free energies. Adrion et al.[18] 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

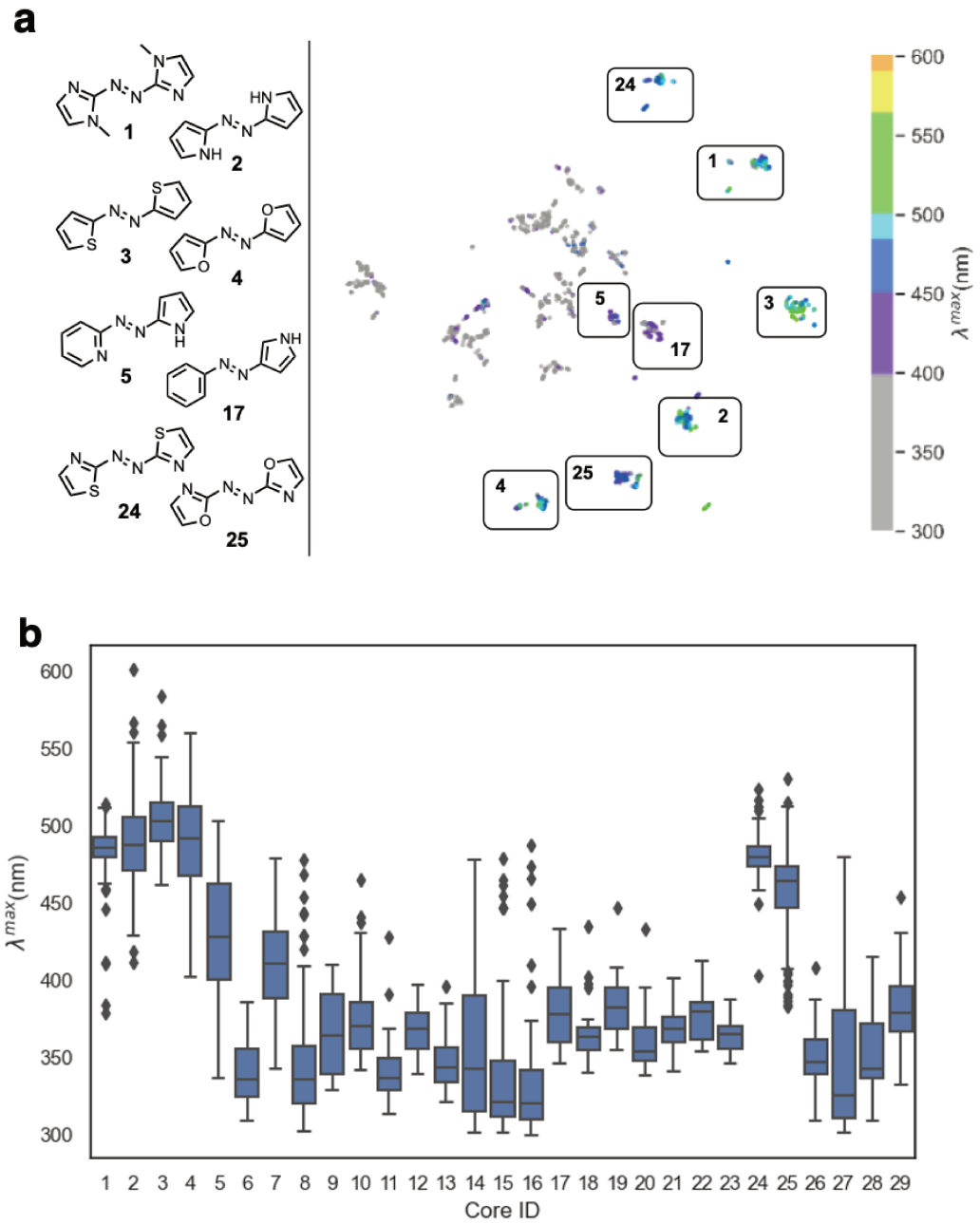


Figure 11: a) Projection of 1 962 azoarene photoswitches suggested by active search using UMAP, computed with a 2048-bit Morgan fingerprint (radius 2), ten nearest neighbors, a minimum distance of 0.1, and the Tanimoto similarity. b) Range of $\lambda_{\max}$ of 1 962 azoarene photoswitches by core ID. 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 circles represent outliers.

with the longest $\lambda_{\max}$, identified with active search. Figure 12 illustrates the candidate from each family of cores subjected to transition state calculations with PBE036-D3/6-31+G(d,p) to optimize the transition states. This was reported to give activation free energies that approach chemical accuracy. Scheme 2 shows the $Z \rightarrow E$ isomerization transition state.

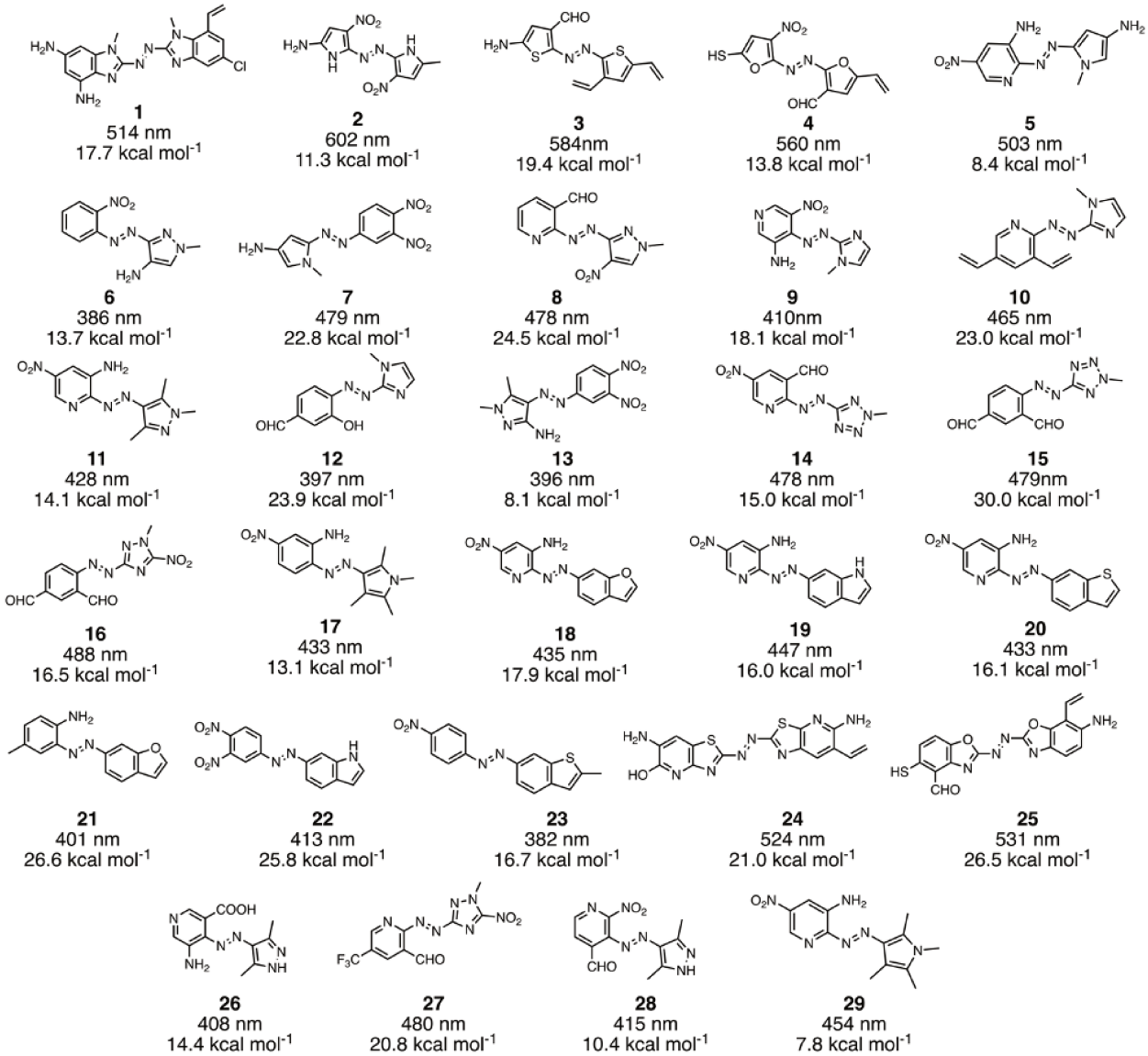
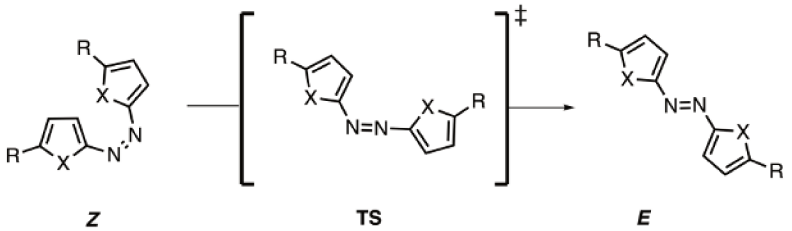


Figure 12: Structures of the 29 highest absorbing azoarene photoswitches for each core. Molecules are labeled by their core ID (in bold), their $\lambda_{\max}$ in nanometers, and activation barrier in kcal mol^{-1}.

The $\lambda_{\max}$ for these top 29 candidates ranges from 382 to 602 nm. The range of activation free energies is 8.1 to 30.0 kcal mol^{-1}. We plotted the activation free energies ($\Delta G^{\ddagger}$) against



Scheme 2: Illustration of the $Z \rightarrow E$ thermal isomerization transition structure.

the λ_{max} for these 29 candidates to determine if there was a relationship between these values (Figure 13).

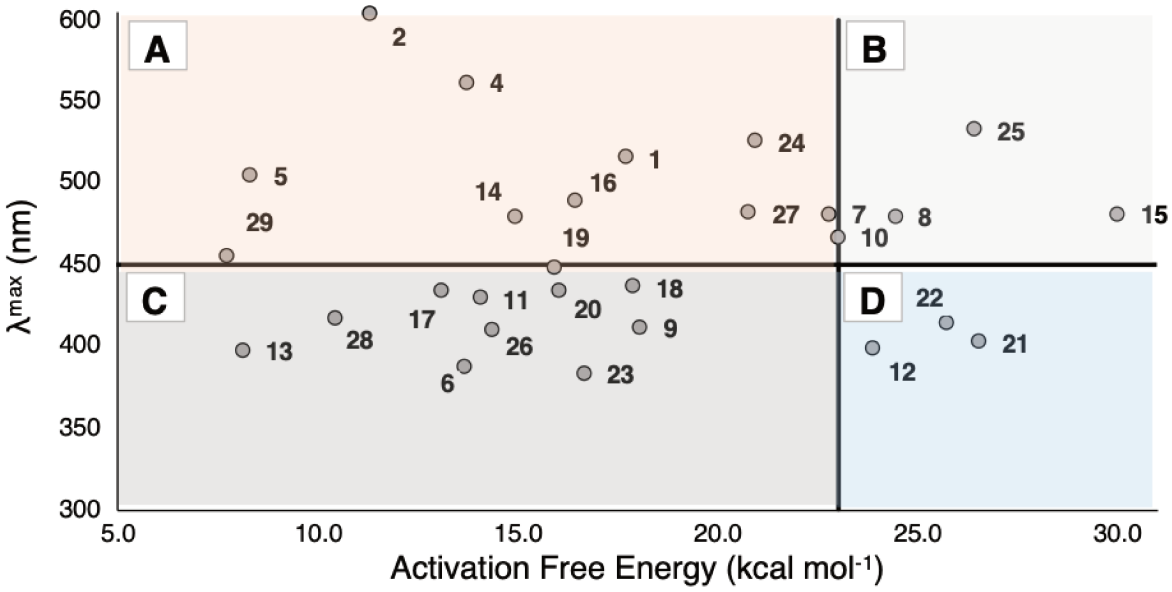


Figure 13: The activation free energy against the λ_{max} of 29 azoarene photoswitches selected by the active search. Their core ID indexes the data points. Quadrant B is where both criterion for an ideal photoswitch ($\lambda_{\text{max}} > 450$ nm and $\Delta G^{\ddagger} > 23$ kcal mol^{-1}) have 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. A linear regression gave the following equation describing the correlation between the activation barrier and λ_{max}, [$\lambda_{\text{max}} = 0.1189\ \Delta G^{\ddagger} + 456$] with an R^2 of 0.0002.

Figure 13 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 (red tint) and B (green tint) contain molecules that have $\lambda_{\text{max}} > 450$ nm or 2.6 eV. Quadrants

A and C (purple tint) are populated with molecules with an activation free energy less than 23.0 kcal mol^{-1}. Quadrants C and D (blue tint) 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, denoted by two checks that satisfy both criterion; Quadrant 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.

Acknowledgement

All authors acknowledge the National Science Foundation (NSF-OAC-1940307) for funding this research. FM and SAL appreciate the assistance from the Northeastern Research Computing Team and access to the computing resources of the Discovery cluster.

Supporting Information Available

The file `supporting_information.pdf` contains supporting information for this manuscript, including a description of the code that we release with the submission and detailed search results in each iteration of our procedure.

References

(1) Griffiths, J. Photochemistry of Azobenzene and its Derivatives. *Chemical Society Reviews* **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 In-

duces "dark" Cyclobutane Pyrimidine Dimers in Human Skin *in vivo*. *Scientific Reports* **2018**, *8*, 1–12.

(3) 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. *Journal of the American Chemical Society* **2020**, *142*, 6538–6547.

(4) 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. *Chemistry a European Journal* **2018**, *24*, 8639–8647.

(5) Huddleston, P. R.; Volkov, V. V.; Perry, C. C. The structural and electronic properties of 3, 3'-azothiophene photo-switching systems. *Physical Chemistry Chemical Physics* **2019**, *21*, 1344–1353.

(6) 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. *Journal of the American Chemical Society* **2014**, *136*, 11878–11881.

(7) Calbo, J.; Weston, C. E.; White, A. J.; Rzepa, H. S.; Contreras-García, J.; Fuchter, M. J. Tuning Azoheteroarene Photoswitch Performance through Heteroaryl Design. *Journal of the American Chemical Society* **2017**, *139*, 1261–1274.

(8) Slavov, C.; Yang, C.; Heindl, A. H.; Wegner, H. A.; Dreuw, A.; Wachtveitl, J. Thiophenylazobenzene: An Alternative Photoisomerization Controlled by Lone-Pair $\cdots\pi$ Interaction. *Angewandte Chemie International Edition* **2020**, *59*, 380–387.

(9) 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. *The Journal of Organic Chemistry* **2013**, *78*, 12090–12105.

(10) Abburu, S.; Venkatraman, V.; Alsberg, B. K. TD-DFT based fine-tuning of molecular excitation energies using evolutionary algorithms. *RSC Advances* **2016**, *6*, 3661–3670.

(11) 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. *The Journal of Physical Chemistry C* **2019**, *123*, 12195–12202.

(12) Chansen, W.; Jen-Shiang, K. Y.; Kungwan, N. A TD-DFT molecular screening for fluorescence probe based on excited-state intramolecular proton transfer of 2'-hydroxychalcone derivatives. *Journal of Photochemistry and Photobiology A: Chemistry* **2021**, *410*, 113165.

(13) Lopez, S. A.; Sanchez-Lengeling, B.; de Goes Soares, J.; Aspuru-Guzik, A. Design Principles and Top Non-FullereneAcceptor Candidates for OrganicPhotovoltaics. *Joule* **2017**, *1*, 857–870.

(14) 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., et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **2016**, *15*, 1120–1127.

(15) Kim, S.; Noh, J.; Gu, G. H.; Aspuru-Guzik, A.; Jung, Y. Generative Adversarial Networks for Crystal Structure Prediction. *ACS Central Science* **2020**, *6*, 1412–1420.

(16) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *Journal of the American Chemical Society* **2009**, *131*, 8732–8733.

(17) 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. *The Journal of Physical Chemistry Letters* **2019**, *10*, 6835–6841.

(18) Adrion, D.; Kaliakin, D.; Neal, P.; Lopez, S. Benchmarking of Density Functionals for Z-Azoarene Half-Lives via Automated Transition State Search. **2021**, ChemRxiv preprint.

(19) 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.

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

(21) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(22) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

(23) 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. *Theoretical Chemistry Accounts* **2008**, *120*, 215–241.

(24) 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. *The Journal of Chemical Physics* **1982**, *77*, 3654–3665.

(25) Ditchfield, R.; Hehre, W.; Pople, J. Self-Consistent Molecular Orbital Methods. VI. Energy Optimized Gaussian Atomic Orbitals. *The Journal of Chemical Physics* **1970**, *52*, 5001–5007.

(26) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chemical Reviews* **2005**, *105*, 2999–3094.

(27) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics* **2008**, *10*, 6615–6620.

(28) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

(29) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 983–996.

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

(31) Garnett, R.; Gärtner, T.; Vogt, M.; Bajorath, J. Introducing the 'active search' method for iterative virtual screening. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 305–314.

(32) 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**,

(33) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical Science* **2021**,

(34) 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.

(35) Jiang, S.; Malkomes, G.; Abbott, M.; Moseley, B.; Garnett, R. Efficient nonmyopic batch active search. Advances in Neural Information Processing Systems 31. 2018; pp 1099–1109.

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

(37) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **2018**, *3*.

Graphical TOC Entry