

Supporting Information

Reaction Array Fingerprinting

Babak Mahjour¹, Clinton Regan¹, Daniel Schorin², Tim Cernak^{*1,3}

¹Department of Medicinal Chemistry, University of Michigan

²Department of Information, University of Michigan

³Department of Chemistry, University of Michigan

*Email: tcernak@umich.edu

Table of Contents

1. General Information
2. Python Pseudocode for Reaction Array Fingerprint Generation
3. Comparison of SOMs, TSNEs, PCAs, and UMAPs for Suzuki Reaction Dataset Reaction Array Fingerprints
4. Hyperparameters and weights used to generate figure 5 visualizations directly from phactor™ output data
5. Cluster to Reaction Array Matching
6. Identifying reagents with generality
7. Unsupervised vs supervised UMAP
8. Weighted Reaction Fingerprints vs Concatenated Fingerprints vs Difference Fingerprints
9. Chi-squared contingency analysis of the reactivity cliff identified in figure 4.
10. Webapp Instructions and screenshots
11. Diffusion Map Experiments
12. CC Coupling Experimental

- General Information.** Code to generate and visualize reaction array fingerprints was written in Python (version 3.9.10). Scikit-learn (version 1.0.1) was used to run TSNE and PCA dimensionality reduction algorithms. RDKit (version 2021.09.4) was used to convert reaction SMILES into fingerprints, umap-learn (version 0.5.3) was used to generate UMAPs, and sklearn-som (version 1.1.0) was used to run the SOM algorithm. Pandas (version 1.4.1) or SQLAlchemy (version 1.4.44) was used to load reaction data. Code for the webapp was written in Python (version 3.9.10) and ReactJS (version 18.2.0) with minimal dependencies. Python dependencies were limited to Flask (version 2.0.2), Numpy (version 1.22.2), Pandas (1.4.1), Matplotlib (version 3.5.1), and RDKit (version 2021.09.4), all installed via pip (version 22.0.3). JavaScript dependencies were limited to ReactJS (version 18.2.0) for the underlying user interface infrastructure and react-csv-reader (version 3.3.0). API endpoints were written in Flask and exposed via HTTPS.

All datafiles used to make the figures in this manuscript alongside the entirety of the code needed to generate the figures are provided in a GitHub repository.

2. Pseudocode for Reaction Array Fingerprint Generation

```
reagentTypes = ["electrophile", "nucleophile", "catalyst_smiles", "base_smiles", "solvent"]
weights = {"electrophile":1, "nucleophile":3, "catalyst_smiles":1, "base_smiles":1, "solvent":1}
for i,k in data.iterrows():
    this_fp = np.zeros(2048)
    for rt in reagentTypes:
        mol = Chem.MolFromSmiles(k[rt])
        # generic fingerprint function, returned weighted fingerprint
        fp = getFP(mol, weights[rt])
        this_fp = this_fp + fp
    rfps.append(this_fp)

# use any embedding algorithm
X_TSNE_RFP = TSNE(n_components=2, n_jobs=-1, perplexity=15).fit_transform(np.array(rfps))
```

Figure S1. Basic template code in python to create the weighted reaction fingerprint.

3. SOM vs TSNE vs PCA for Figure 1 Data

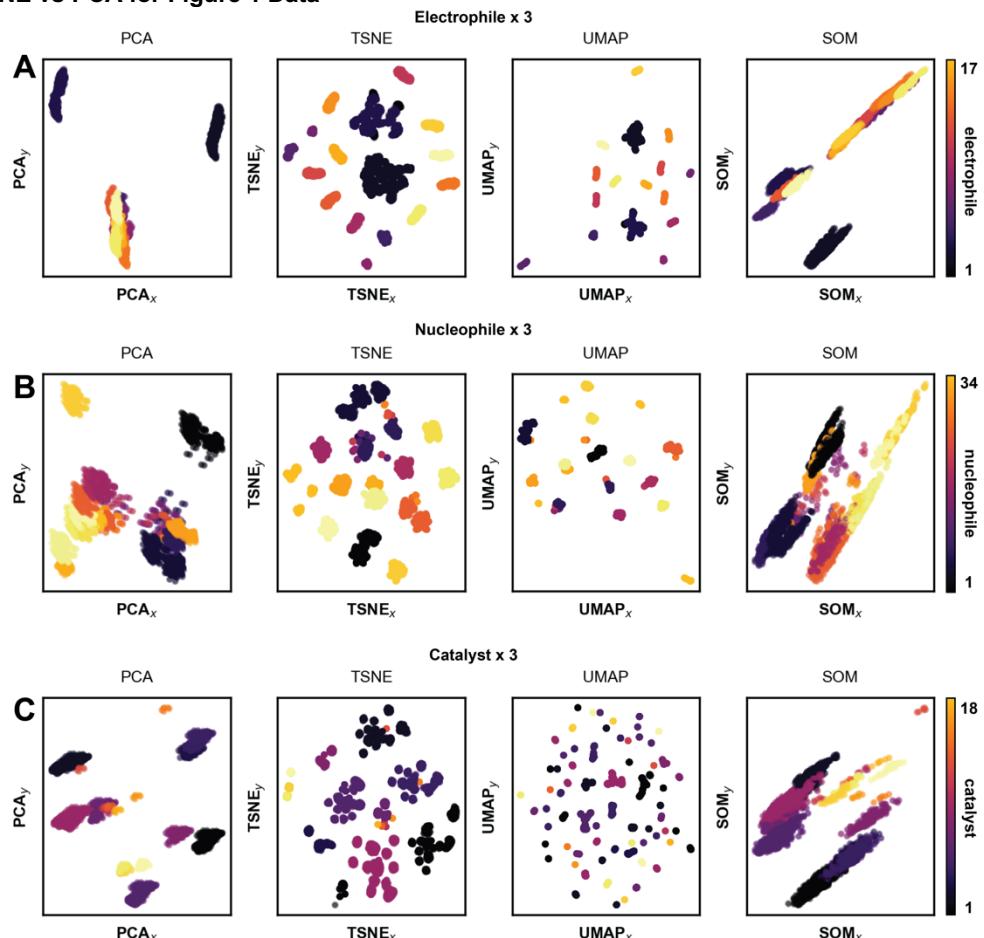


Figure S2. Different dimensionality reduction algorithms performed on the Suzuki dataset weighted reaction fingerprints. **A)** Electrophile fingerprints were multiplied by three. T-SNE and UMAP embeddings formulate distinct clusters. **B)** Nucleophile fingerprints were multiplied by three. **C)** Catalyst fingerprints were multiplied by three.

4. Figure 5 Hyperparameters

Note that the t-SNE algorithm may produce variable results with the same data if a random state seed is not used. Clusters may be relocated or changed entirely.

- 5A) ReductantOxidant Weight: 3, Perplexity: 25
- 5B) Electrophile Weight: 3, Perplexity: 20
- 5C) Ligand1 Weight: 3, Perplexity: 20
- 5D) Nucleophile Weight: 5, Perplexity: 20
- 5E) Nucleophile Weight: 5, Perplexity: 20
- 5F) Nucleophile Weight: 5, Perplexity: 20

5. Cluster to Reaction Array Matching

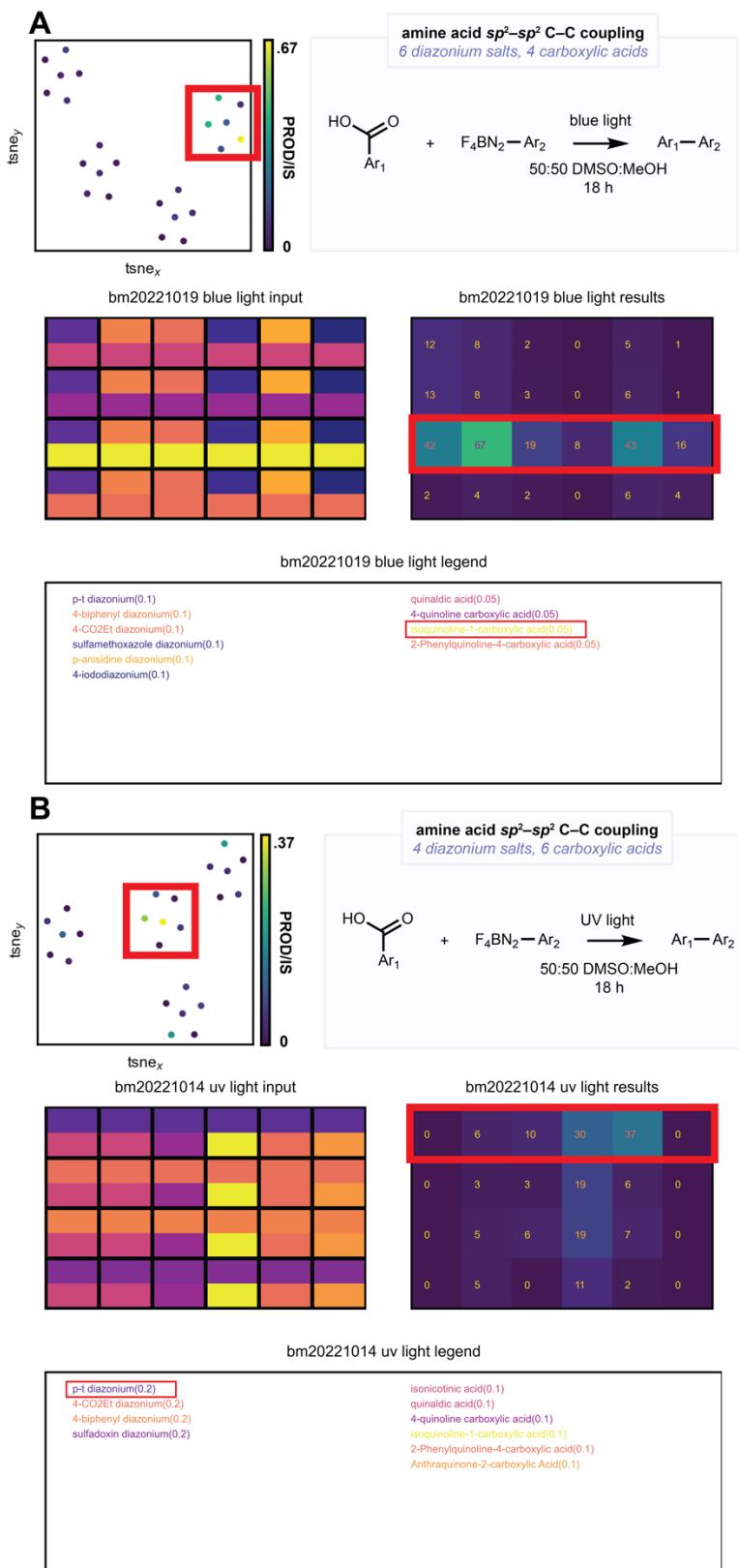


Figure S3. Reaction array outputs paired with weighted reaction fingerprint manifolds. Two reaction arrays ran in the discovery of a catalyst-free sp^2-sp^2 deaminative-decarboxylative C–C coupling have distinct rows and columns that cluster well in the t-SNE. Carboxylic acid and diazonium salt are irradiated by blue or UV light in 50:50 Methanol:DMSO for 18 hours. **A)** Row C is identified as a cluster in the t-SNE. This row corresponds to carboxylic acid 1-isoquinoline carboxylic acid. **B)** Row A corresponds to the boxed cluster in the t-SNE and the diazonium of p-toluidine. This acid/amine pair is one of the best performing substrate pairs with this reactivity.

6. Identifying reagents with generality

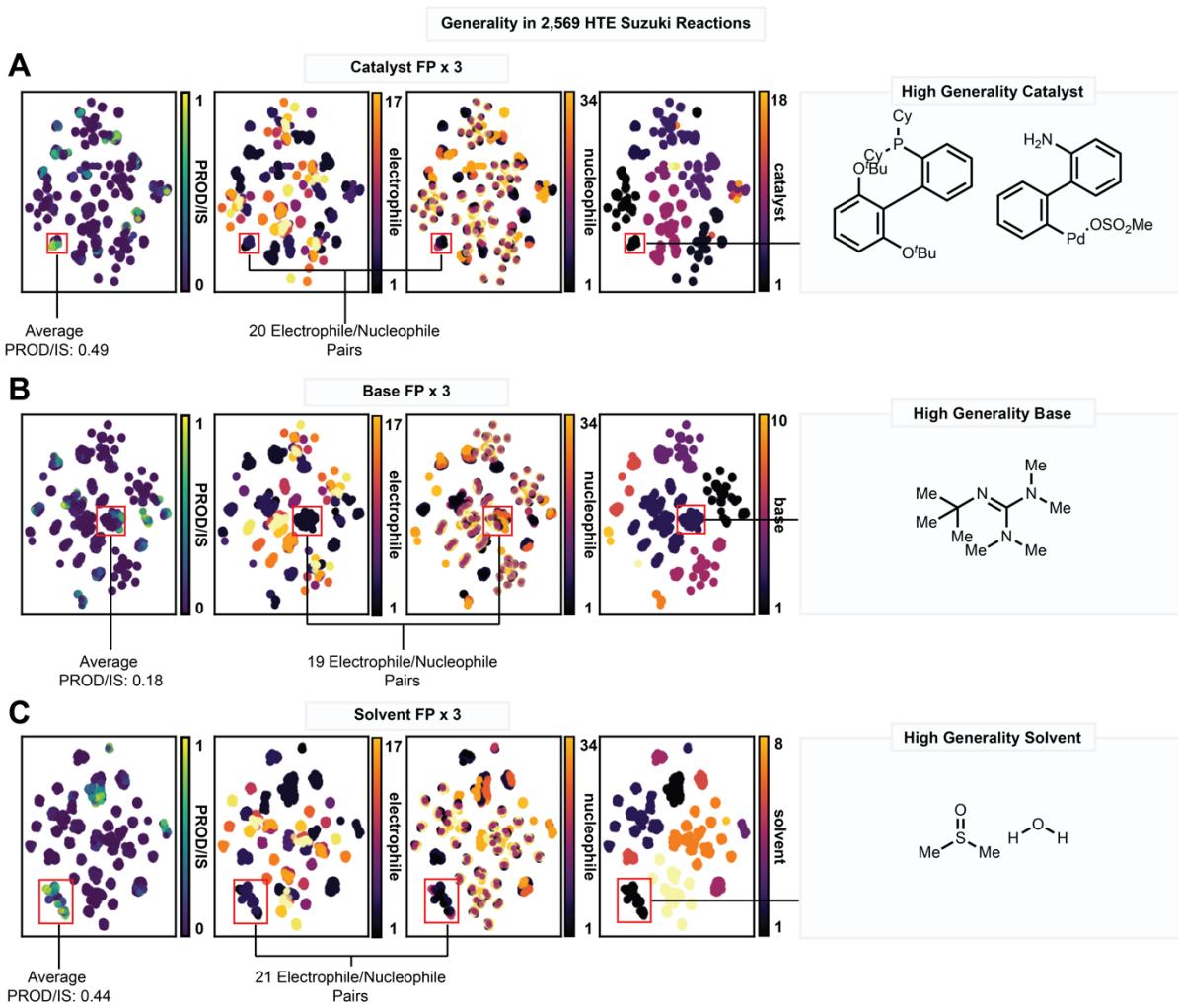


Figure S4. Condition generality demonstrated through the manipulation of weighted reaction fingerprint manifolds of the Suzuki dataset. **A)** When multiplying catalyst fingerprints by three, clusters containing many nucleophile and electrophile substrate pairs that work well with a specific catalyst can be identified. In this case RuPhos Pd G3 was found to produce an average of 49% product/internal standard integration for 20 substrate pairs. **B)** When weighing base fingerprints by three, the high generality base BTMG was found to produce an average of 18% product/internal standard in 19 substrate pairs. **C)** A mixture of DMSO and water generated an average 44% product/internal standard integration in 21 substrate pairs.

7. Unsupervised vs supervised UMAP

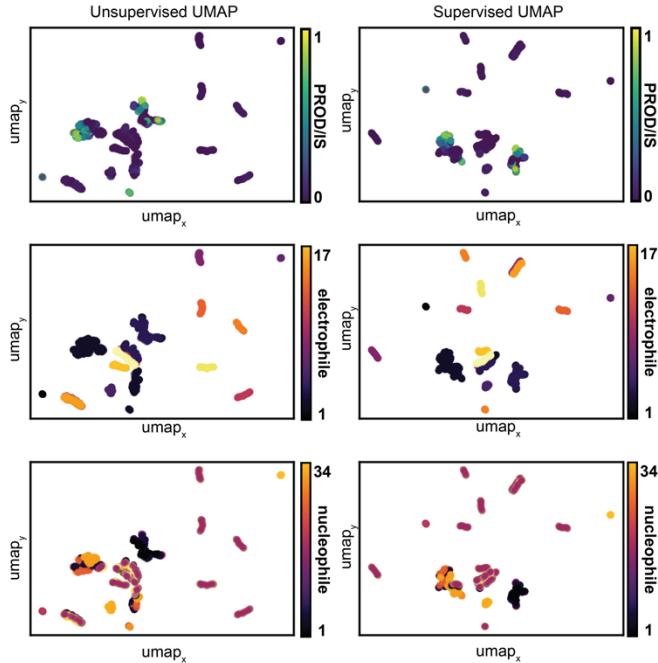


Figure S5. Unsupervised and supervised UMAP dimensionality reduction algorithms are compared. Little difference is indicated between the manifolds.

8. Weighted Reaction Fingerprints vs Concatenated Fingerprints vs Difference Fingerprints

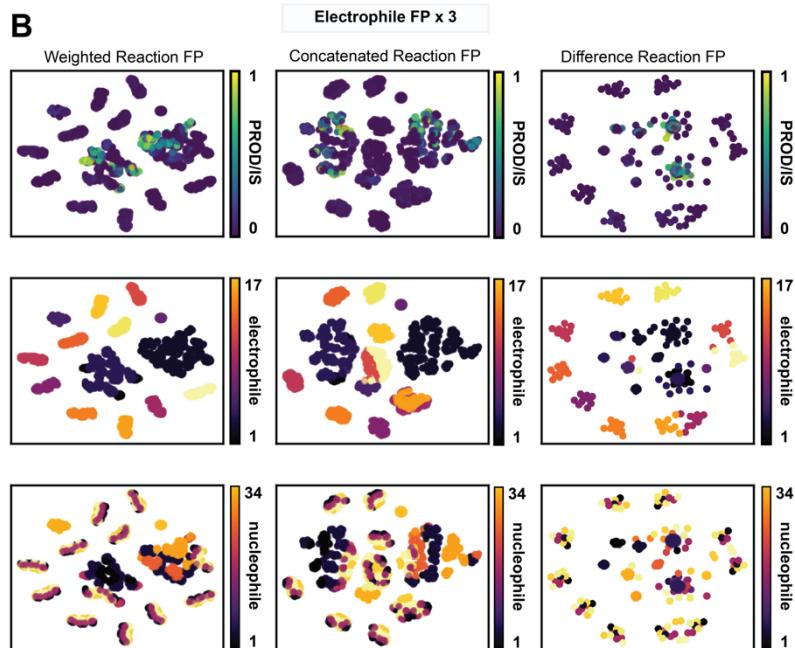
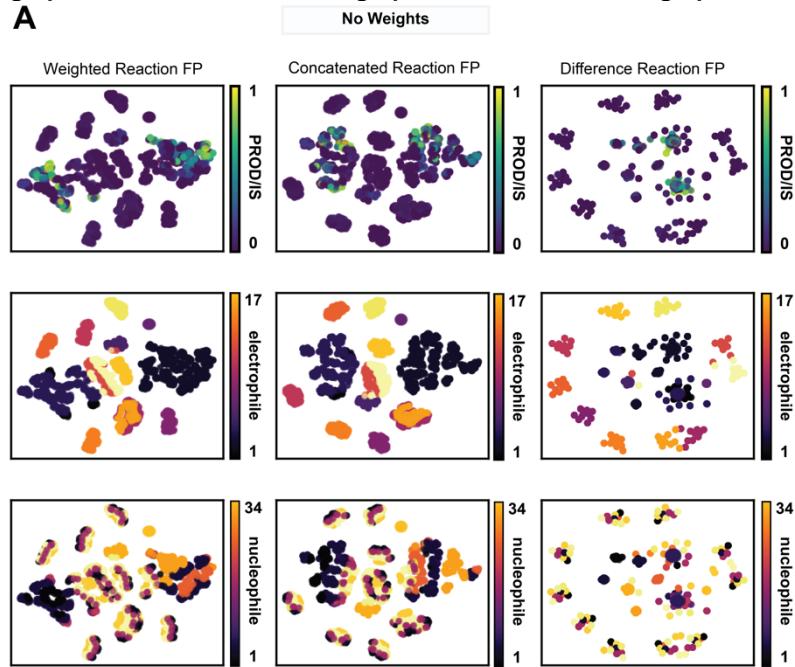


Figure S6. The introduced weighted reaction fingerprint is compared with the commonly used concatenated reaction fingerprints and difference reaction fingerprints found in the literature. **A)** The three fingerprint type manifolds without any feature weights compared side-by-side and colored by product/internal standard, electrophile, and nucleophile. **B)** The three fingerprint type manifolds where the electrophile fingerprints were given a weight of three. The change in the reaction embedding is only noticed in the new weighted reaction fingerprint method (c.f. S6A)

9. Chi-squared validation of the reactivity cliff identified in figure 4.

high	med	low	zero	label
129	151	553	194	[w/ water]
0	8	827	707	[w/o water]

Figure S7. The contingency table of reactions in the Suzuki dataset split by those containing water as a co-solvent and those that do not. A chi-squared statistic is calculated to test the hypothesis of independence of the observed multivariate

frequencies of the table. There are three degrees of freedom, and the chi-squared value is 522 with a p-value of 8.8e-113. This indicates that it is statistically likely that the difference between the observed distributions is not due to chance.

10. Webapp Instructions

To generate weighted reaction fingerprint manifolds using the web app provided on <https://fingerprints.cernaklab.com>, a reaction array output dataset is required (see ref. 8 for assistance in running reaction arrays). At minimum, each row in the CSV file must contain an output_value, in addition to reagents and their SMILES. Nine example datasets are provided as a dropdown on the webpage. Simply load a reaction dataset, enter the weighting scheme you wish to utilize (only integers are accepted but multiple weights for different components can be used.), and hit “generate” to create the manifold. Wait for the algorithm to run. Once the manifold is displayed, the points can be clicked to highlight the corresponding reaction in the reaction table. Points can be colored by output value or by a particular reagent class.

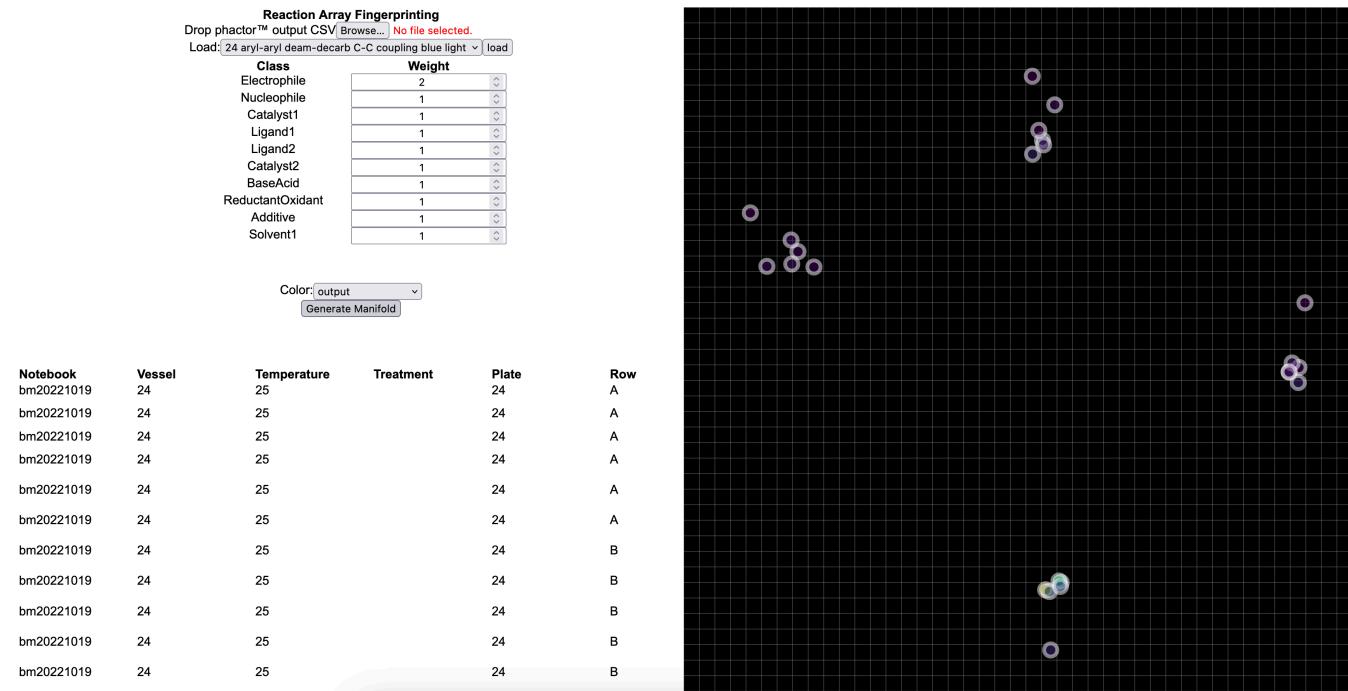


Figure S8. A 24-well manifold preloaded into the webapp. This calculation is performed nearly instantly.

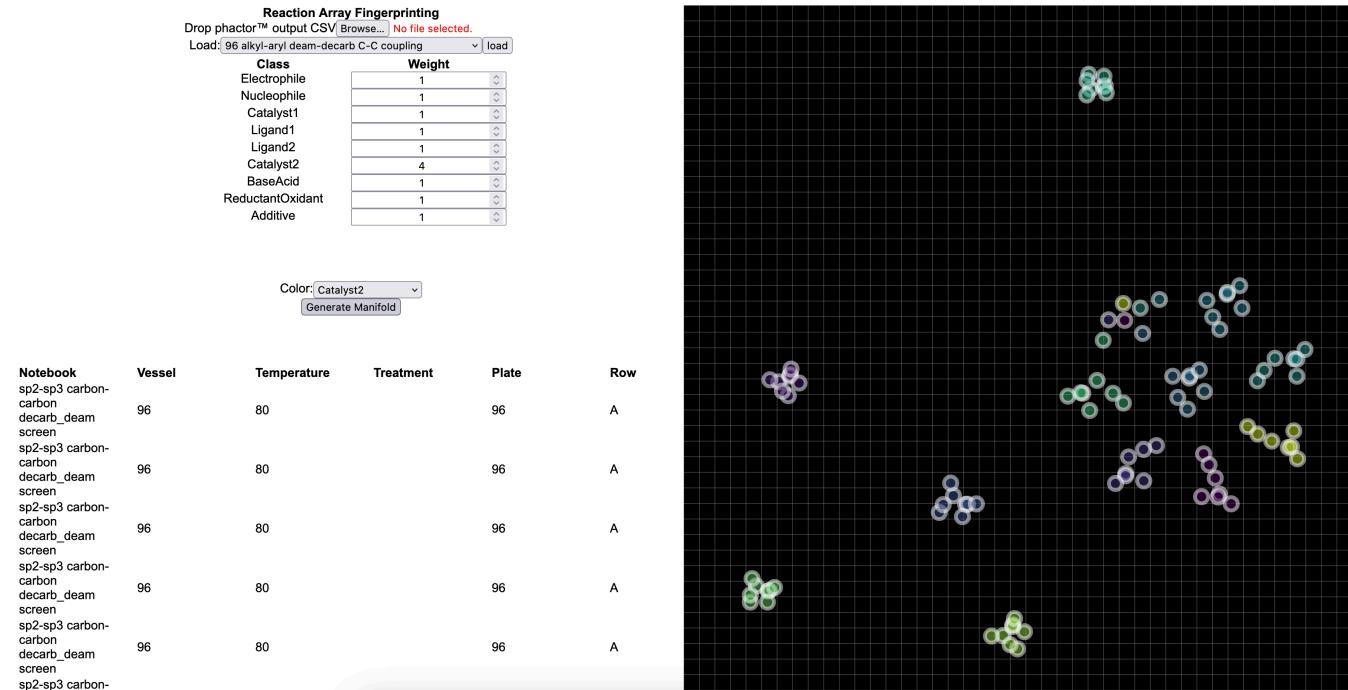


Figure S9. A 96-well manifold preloaded into the webapp. This calculation is performed nearly instantly.

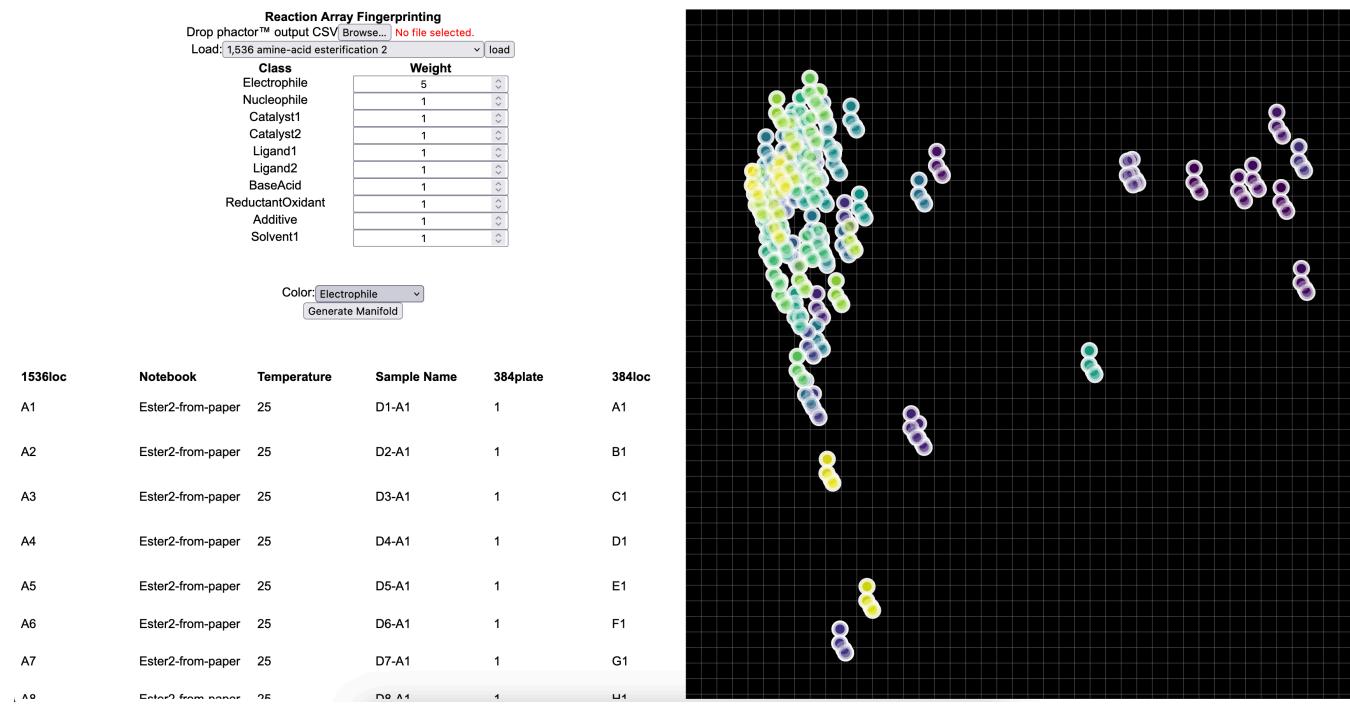


Figure S10. A 1,536-well manifold preloaded into the webapp. This calculation is performed within 10 seconds.

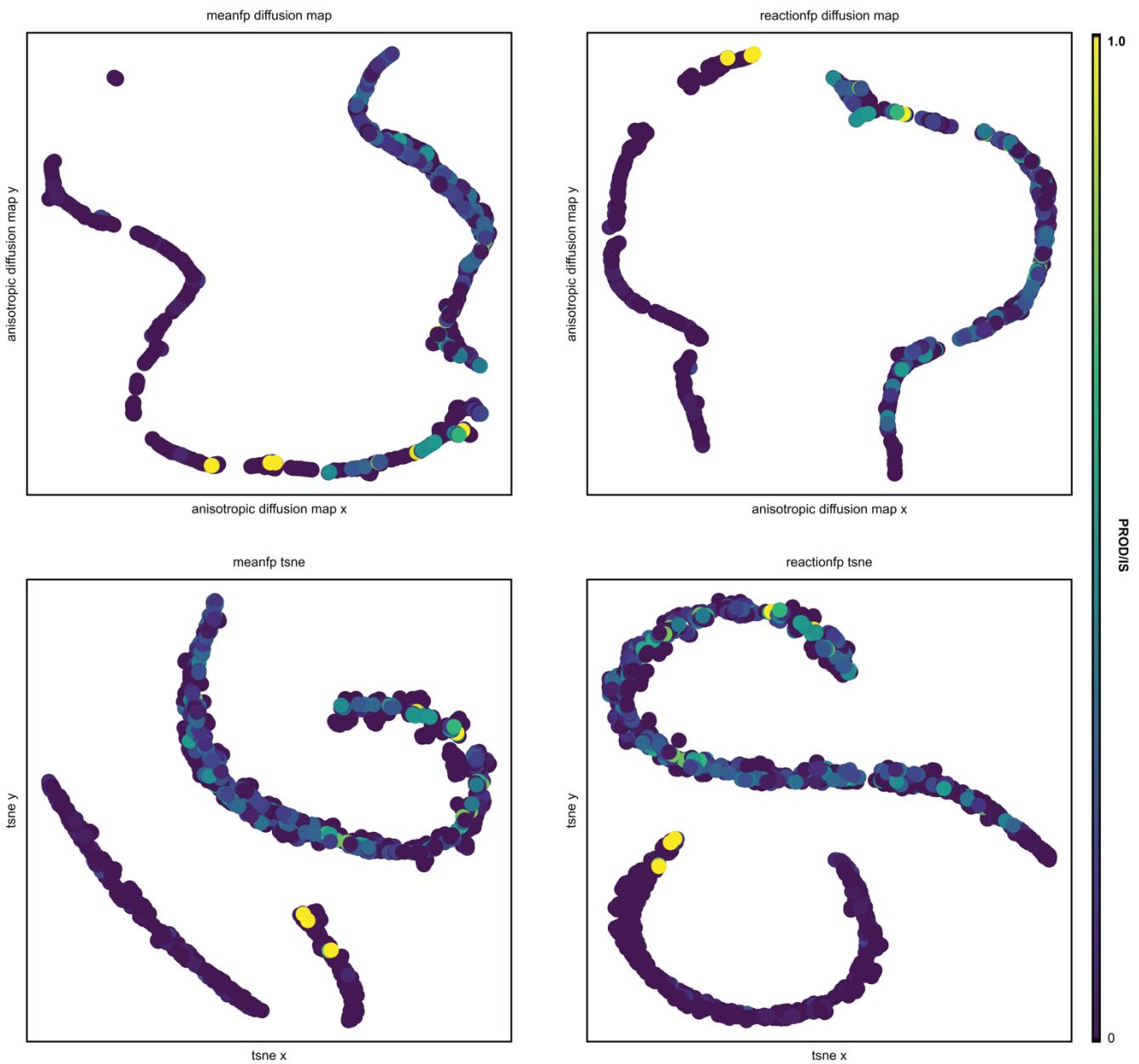
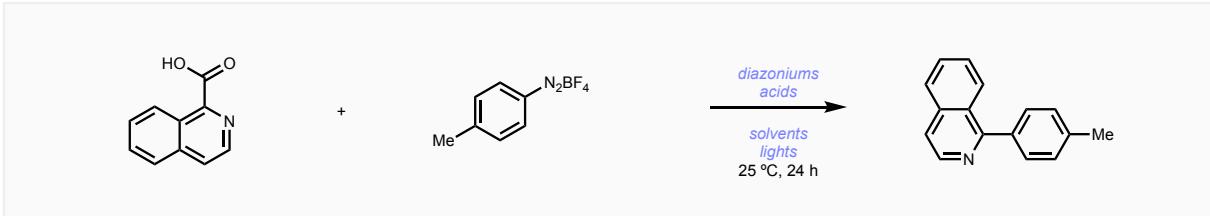


Figure S11. Comparison of mean fingerprints against summed fingerprints when including molecular weight and log_p in tSNE and UMAP manifolds.

a. Decarboxylative-deaminative sp^2 - sp^2 carbon–carbon coupling



- i. **General Screening Procedure.** Acids and diazonium salts were prepared as stock solutions as calculated by phactor™ in DMSO, methanol, or 1:1 DMSO:methanol solution. Diazonium stock solutions were kept in a freezer until dosage into the reaction plate was required. 50 microliters of each acid stock solution were added to corresponding wells, then diazonium solutions were sequentially removed from the freezer and 50 microliters of each was dosed into corresponding wells. The reactor block was then transferred to the LED reactor and allowed to stir for 10 minutes without irradiation. After 10 minutes, irradiation was turned on and the reaction was run for 18 hours. Reactions were quenched with 900 microliters of acetonitrile, and 100 microliter aliquots were transferred to an analytical plate containing equimolar caffeine and 900 microliters of acetonitrile in each well for UPLC-MS analysis. Product/Internal Standard values were calculated by taking the normalized ratio between the integrals of the total wavelength chromatogram peak corresponding to the product's ionized mass ($M+1$) and the peak corresponding to the caffeine internal standard.

Both acids and diazoniums are far more soluble in DMSO than methanol, and slurry loading or vigorous continued stirring may be required to dose suspensions when compound are poorly soluble.

Each of the following plates were run in tandem under two different light systems. For instance, bm20222101blue is the bm20221001 plate run under blue light, and bm20221001uv is the same plate instead run under UV light.

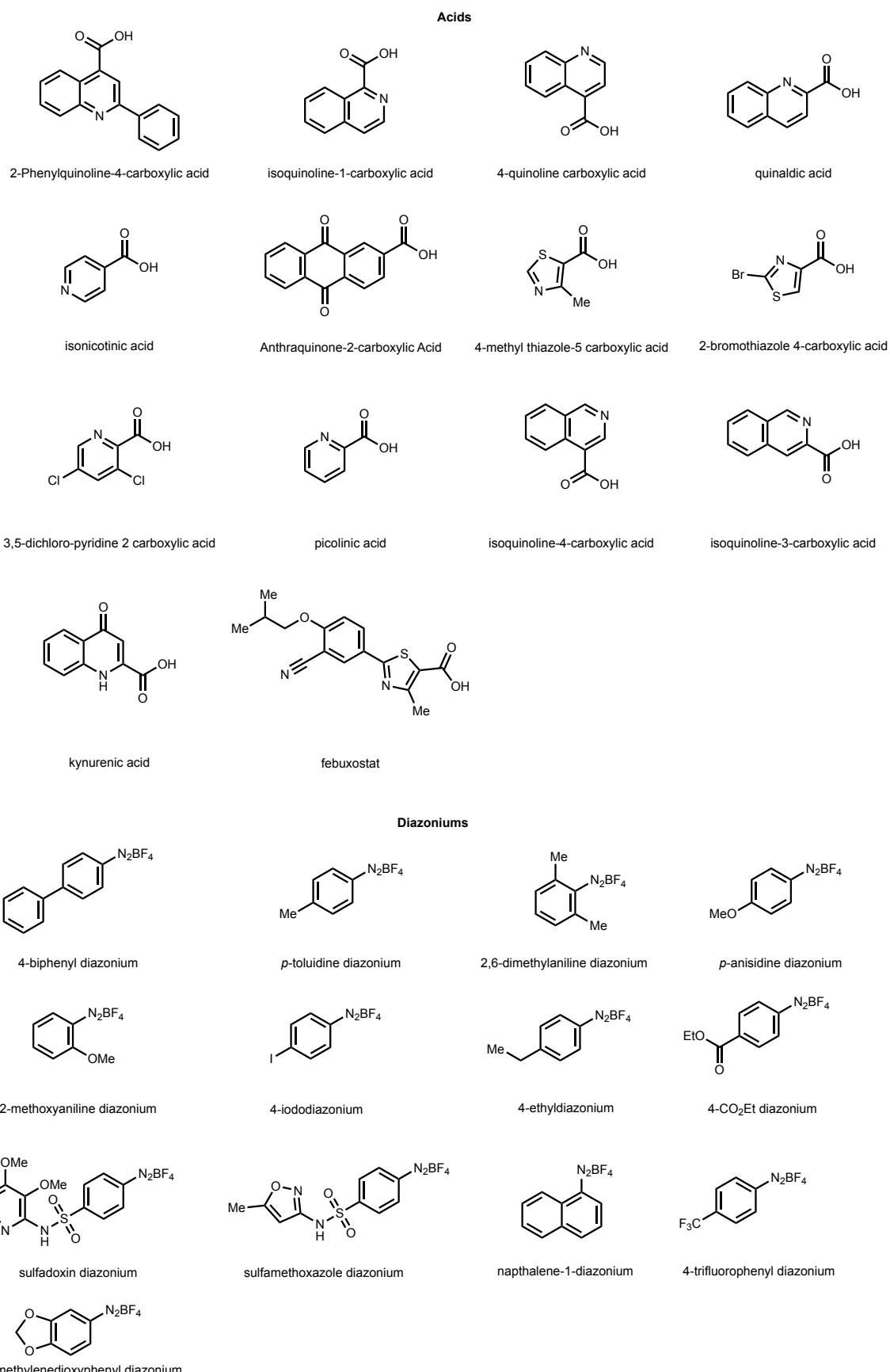


Figure S12. Acids and diazoniums used in catalyst-free sp^2 - sp^2 deaminative decarboxylative C–C coupling reaction arrays

- ii. **bm20222101 – 12 diazonium salts, 8 acids, 2 lights.** The general screening procedure was followed. Two reaction plates were dosed with 12 diazonium salts and eight heterocyclic sp^2 carboxylic acids. The solvent was pure methanol and the limiting reagent was the acid at a concentration of 0.1M. One plate was irradiated by blue light (top) while the other was irradiated by UV light (bottom). Two equivalents of diazonium were added.

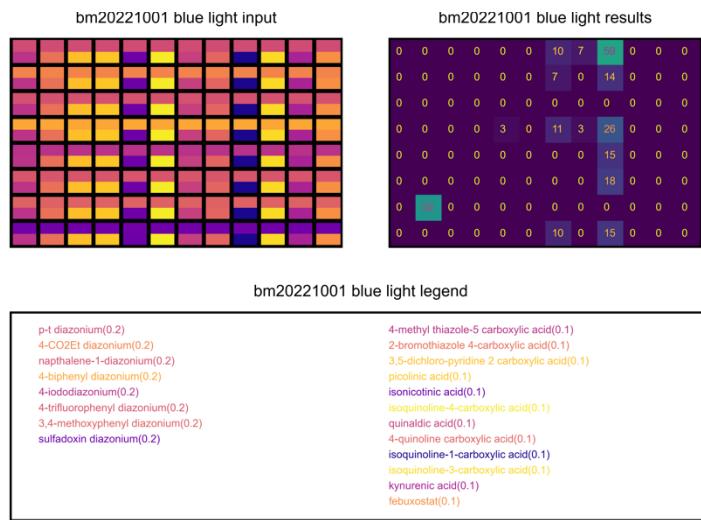


Figure S13. Input and outputs of 96-well array bm20221001blue.

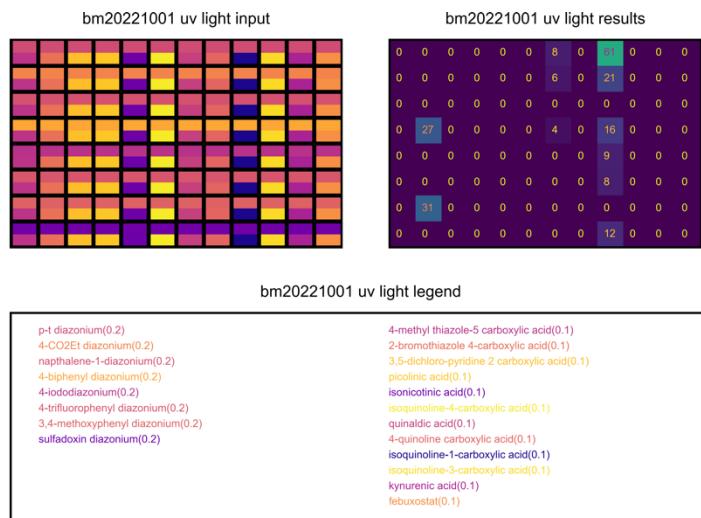


Figure S14. Input and outputs of 96-well array bm20221001uv.

- iii. **bm20221014 – 4 diazonium salts, 6 acids, 2 lights.** The general screening procedure was followed. Two reaction plates were dosed with four diazonium salts and six acids. The solvent was 50:50 DMSO:methanol and the limiting reagent was the acid at a concentration of 0.1M. One plate was irradiated by blue light (top) while the other was irradiated by UV light (bottom). Two equivalents of diazonium were added.

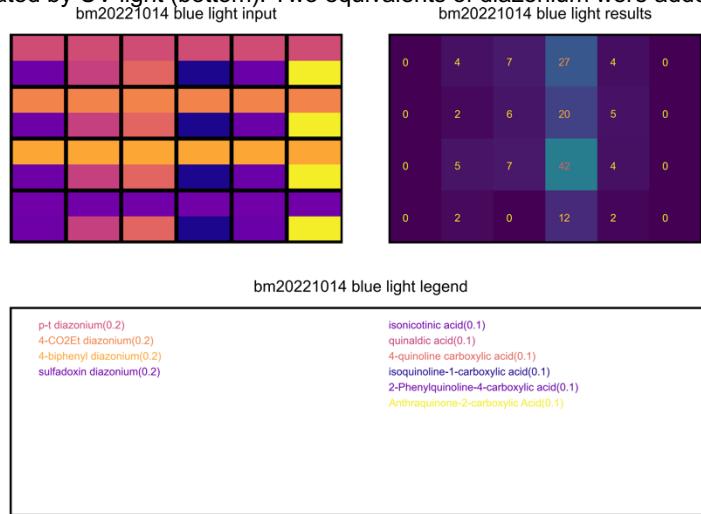


Figure S15. Input and outputs of 24-well array bm20221014blue.

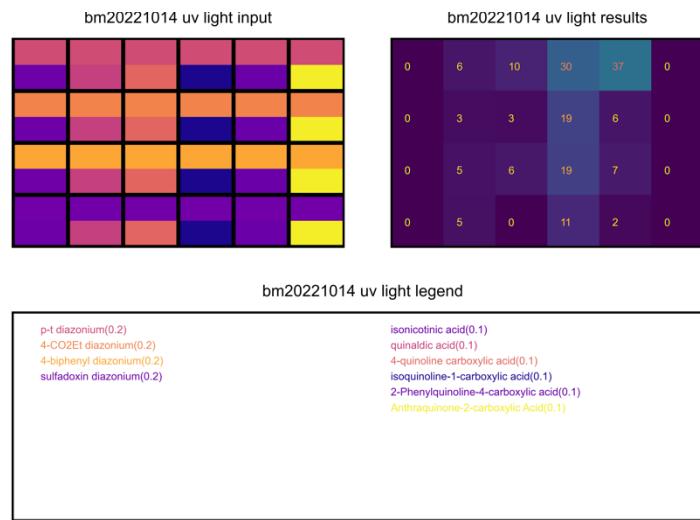


Figure S16. Input and outputs of 24-well array bm20221014uv. (Figure 4C)

- iv. **bm20221019 – 6 diazonium salts, 4 acids, 2 lights.** The general screening procedure was followed. Two reaction plates were dosed with six diazonium salts and four acids. The solvent was 50:50 DMSO:methanol and the limiting reagent was the acid at a concentration of 0.1M. One plate was irradiated by blue light (top) while the other was irradiated by UV light (bottom). Two equivalents of diazonium were added.

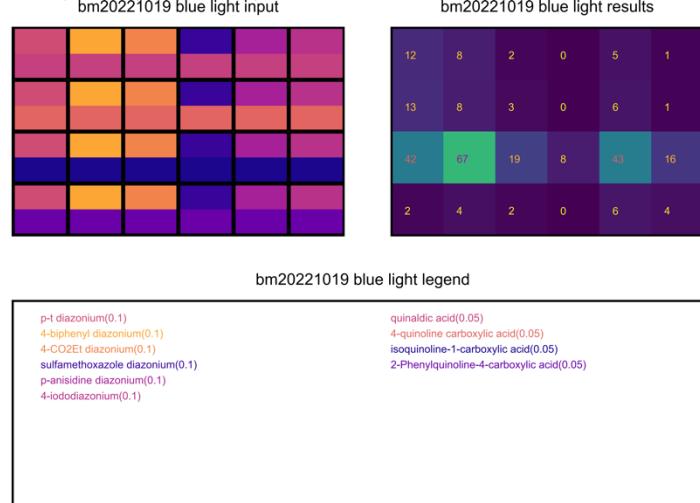


Figure S17. Input and outputs of 24-well array bm20221019blue. (Figure 4B)

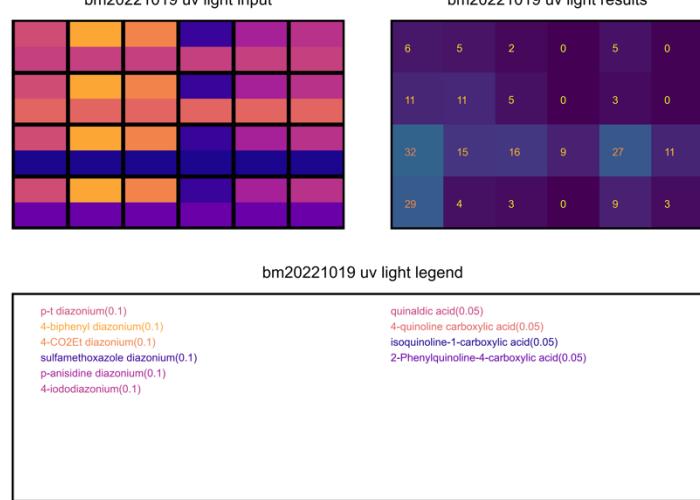


Figure S18. Input and outputs of 24-well array bm20221019uv.

- v. **bm20221021 – 5 diazonium salts, 1 acids, 4 acid concentrations, 2 lights.** The general screening procedure was followed. Two reaction plates were dosed with five diazonium salts and 2-Phenylquinoline-4-carboxylic acid (**4**) at four different concentrations (0.05 M, 0.75 M, 0.1 M, 0.125 M). The solvent was pure DMSO. One plate was

irradiated by blue light (top) while the other was irradiated by UV light (bottom). One equivalent of diazonium was added.

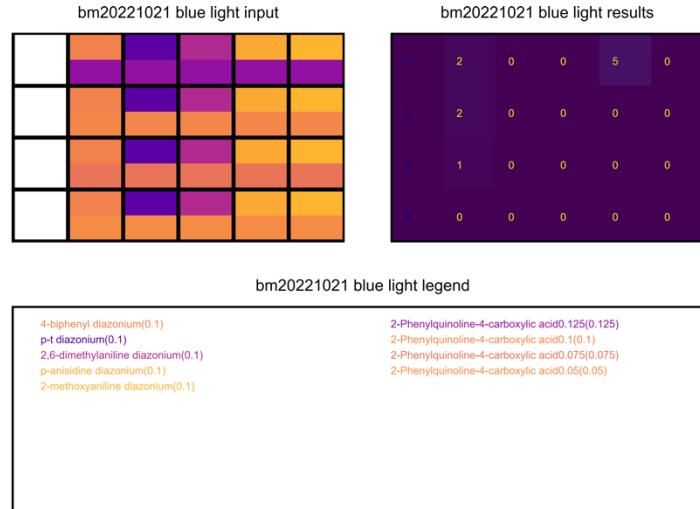


Figure S19. Input and outputs of 24-well array bm20221021blue.

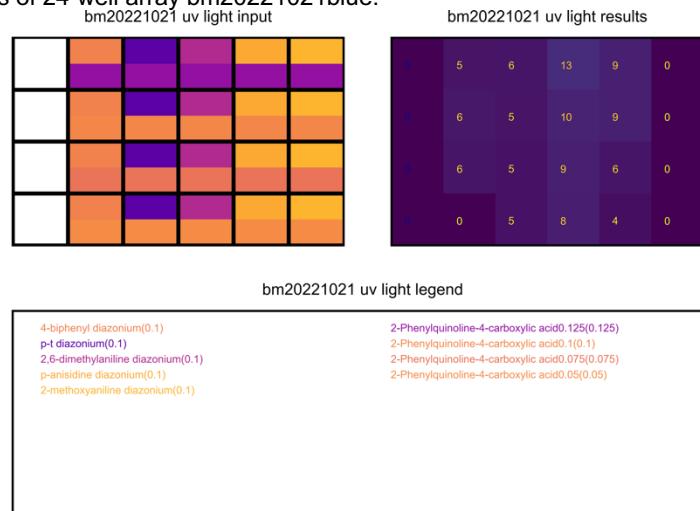


Figure S20. Input and outputs of 24-well array bm20221021uv.

- vi. **bm20221024 – 6 diazonium salts, 4 acids, 2 lights.** The general screening procedure was followed. Two reaction plates were dosed with six diazonium salts and four acids. The solvent was pure DMSO in rows B and D and pure methanol in rows A and C. The limiting reagent was the acid at a concentration of 0.05M. One plate was irradiated by blue light (top) while the other was irradiated by UV light (bottom). One equivalent of diazonium was added.

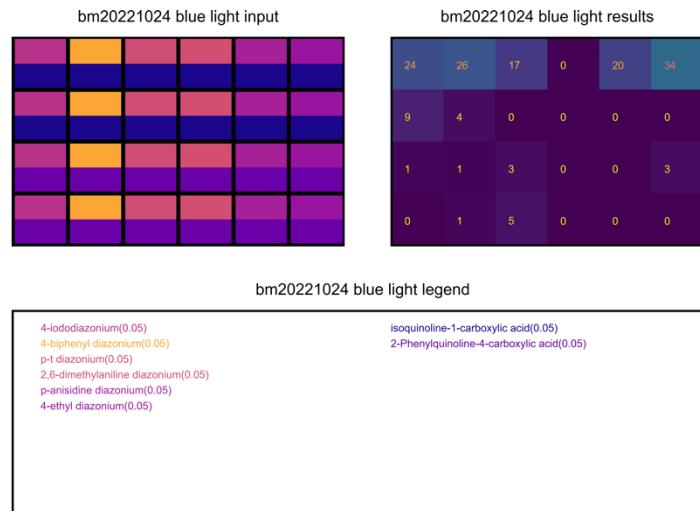


Figure S21. Input and outputs of 24-well array bm20221024blue.

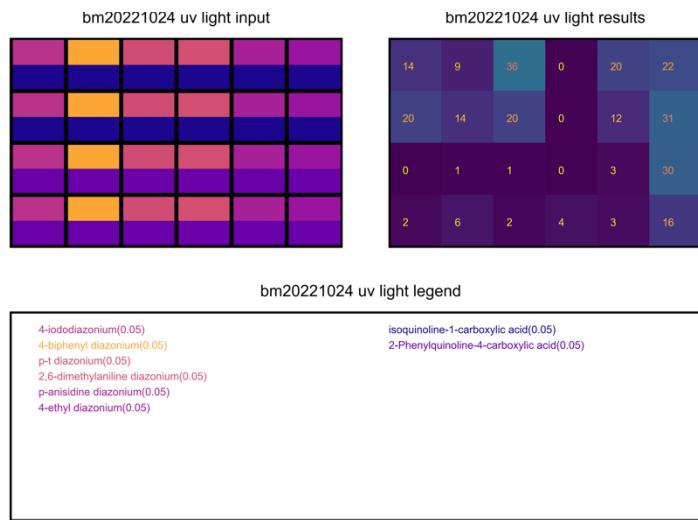


Figure S22. Input and outputs of 24-well array bm20221024uv.

vii. **bm20221026 – 6 diazonium salts, 4 acids, 2 lights.** The general screening procedure was followed. Two reaction plates were dosed with six diazonium salts and four acids. The solvent was pure methanol and the limiting reagent was the acid at a concentration of 0.05M. One plate was irradiated by white light (bottom). While the other was irradiated by UV light (top). Two equivalents of diazonium were added.

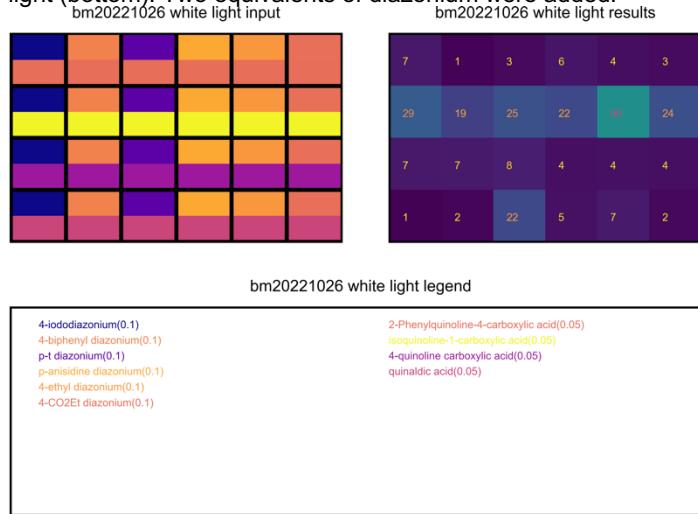


Figure S23. Input and outputs of 24-well array bm20221026white.

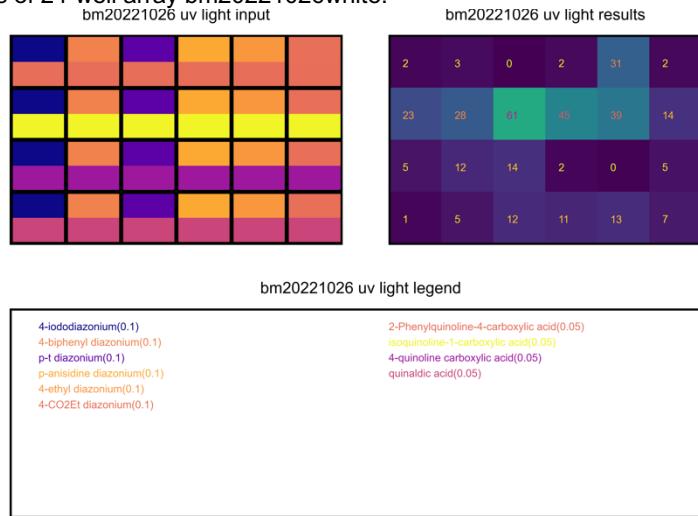


Figure S24. Input and outputs of 24-well array bm20221026uv.

viii. Scale-up Procedure

.1 mmol carboxylic acid and four equivalents of diazonium salt was added to a 2-dram vial with a stirbar. 1 mL of methanol was added, and the reaction was stirred without irradiation for 10 minutes. After 10 minutes, blue light was turned on and the reaction was run for 24 hours. Reactions were run at 0.1 M and at room

temperature. Reactions were quenched with bicarbonate and brine and the aqueous layer is washed with ethyl acetate twice. The organic layer is dried and solvent is removed. Crude mixture was then redissolved in dichloromethane and the product is isolated via flash chromatography.

ix. Scale up of 1-(*p*-tolyl)isoquinoline

The scale-up procedure was followed. 22 mg (97%) was isolated via column chromatograph using 20% EtOAc:Hexanes as eluent. Proton NMR for this compound is displayed in NMR section of the Supporting Information.

^1H NMR (400 MHz, CDCl_3) δ 8.61 – 8.63 (d, J = 5.8 Hz, 1H), 8.17 – 8.19 (dq, J = 8.6, 1.0 Hz, 1H), 7.91 – 7.93 (m, 1H), 7.73 – 7.76 (ddd, J = 8.2, 6.9, 1.2, Hz, 1H), 7.70 – 7.71 (d, J = 5.8 Hz, 1H), 7.62 – 7.65 (d, J = 8.1 Hz, 2H), 7.57 – 7.60 (ddd, J = 8.3, 6.9, 1.3 Hz, 1H), 7.36 – 7.38 (d, J = 7.5 Hz, 2H), 2.47 (s, 3H)

The characterization data matched spectral values from literature.¹

x. Scale up of 1-(4-ethylphenyl)isoquinoline

The scale-up procedure was followed. 21 mg (95%) was isolated via column chromatograph using 20% EtOAc:Hexanes as eluent. Proton NMR for this compound is displayed in NMR section of the Supporting Information.

^1H NMR (400 MHz, CDCl_3) δ 8.60 – 8.61 (d, J = 5.7 Hz, 1H), 8.16 – 8.18 (d, J = 8.5 Hz, 1H), 7.89 – 7.90 (d, J = 8.2 Hz, 1H), 7.69 – 7.72 (t, J = 7.5 Hz, 1H), 7.63 – 7.66 (m, 3H), 7.54 – 7.57 (t, J = 7.1 Hz, 1H), 7.37 – 7.38 (d, J = 8.0 Hz, 2H), 2.74 – 2.79 (q, J = 7.6 Hz, 2H), 1.30 – 1.33 (t, J = 7.6 Hz, 3H).

The characterization data matched spectral values from literature.²

xi. Scale up of 1-[1,1'-biphenyl]-4-yl]isoquinoline

The scale-up procedure was followed. 29 mg (72%) was isolated via column chromatograph using 20% EtOAc:Hexanes as eluent. Proton NMR for this compound is displayed in NMR section of the Supporting Information.

^1H NMR (400 MHz, CDCl_3) δ 8.66 (s, 1H), 8.23 – 8.25 (d, J = 8.6 Hz, 1H), 7.94 – 7.95 (d, J = 8.2 Hz, 1H), 7.82 – 7.84 (d, J = 8.2 Hz, 3H), 7.79 – 7.81 (d, J = 8.2 Hz, 3H), 7.69 – 7.71 (dd, J = 8.3, 1.3 Hz, 2H), 7.61 – 7.64 (t, J = 7.7 Hz, 1H), 7.48 – 7.51 (t, J = 7.7 Hz, 2H), 7.38 – 7.42 (m, 1H)

The characterization data matched spectral values from literature.³

1. 1-(*p*-tolyl)isoquinoline (**3**)

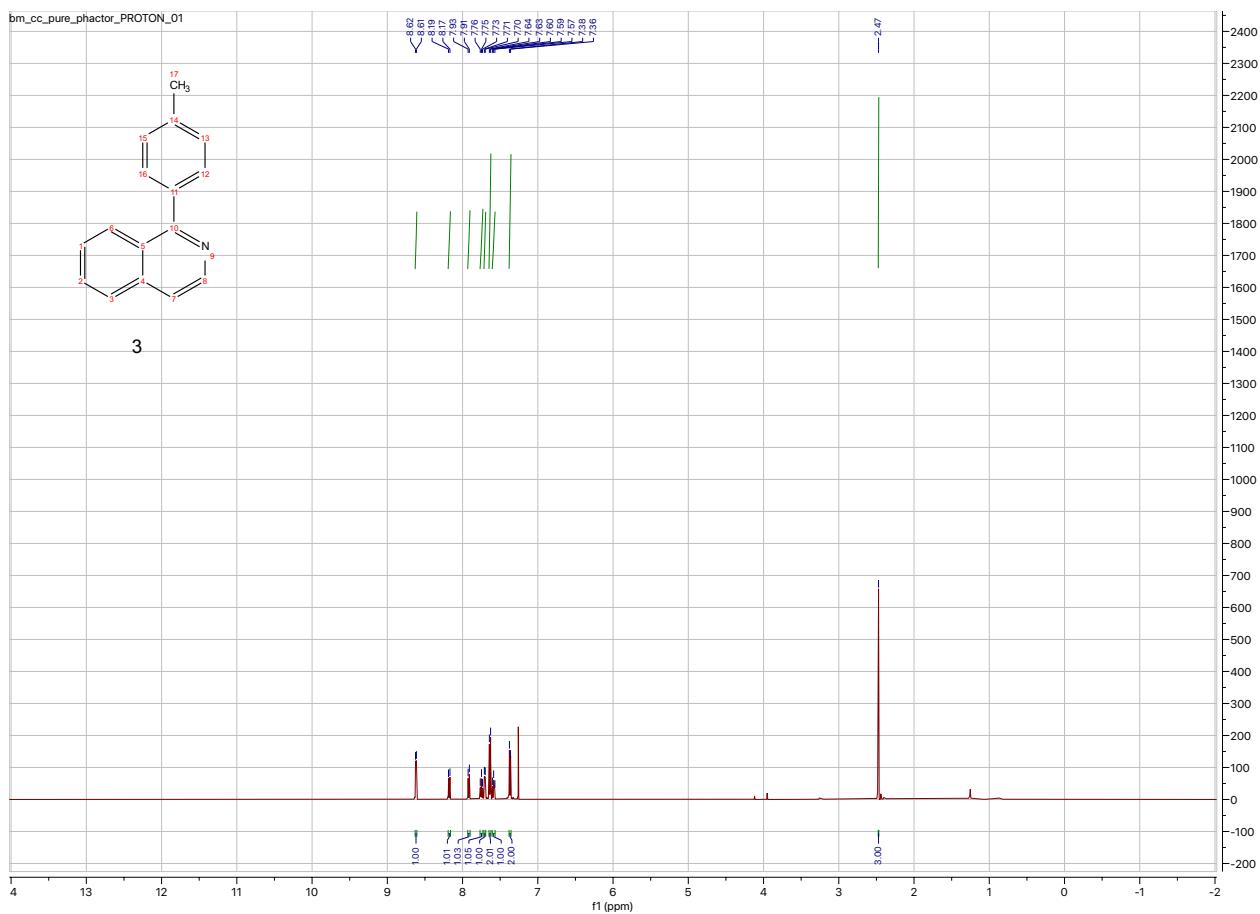


Figure S25. Proton NMR of 1-(*p*-tolyl)isoquinoline.

2. 1-(4-ethylphenyl)isoquinoline

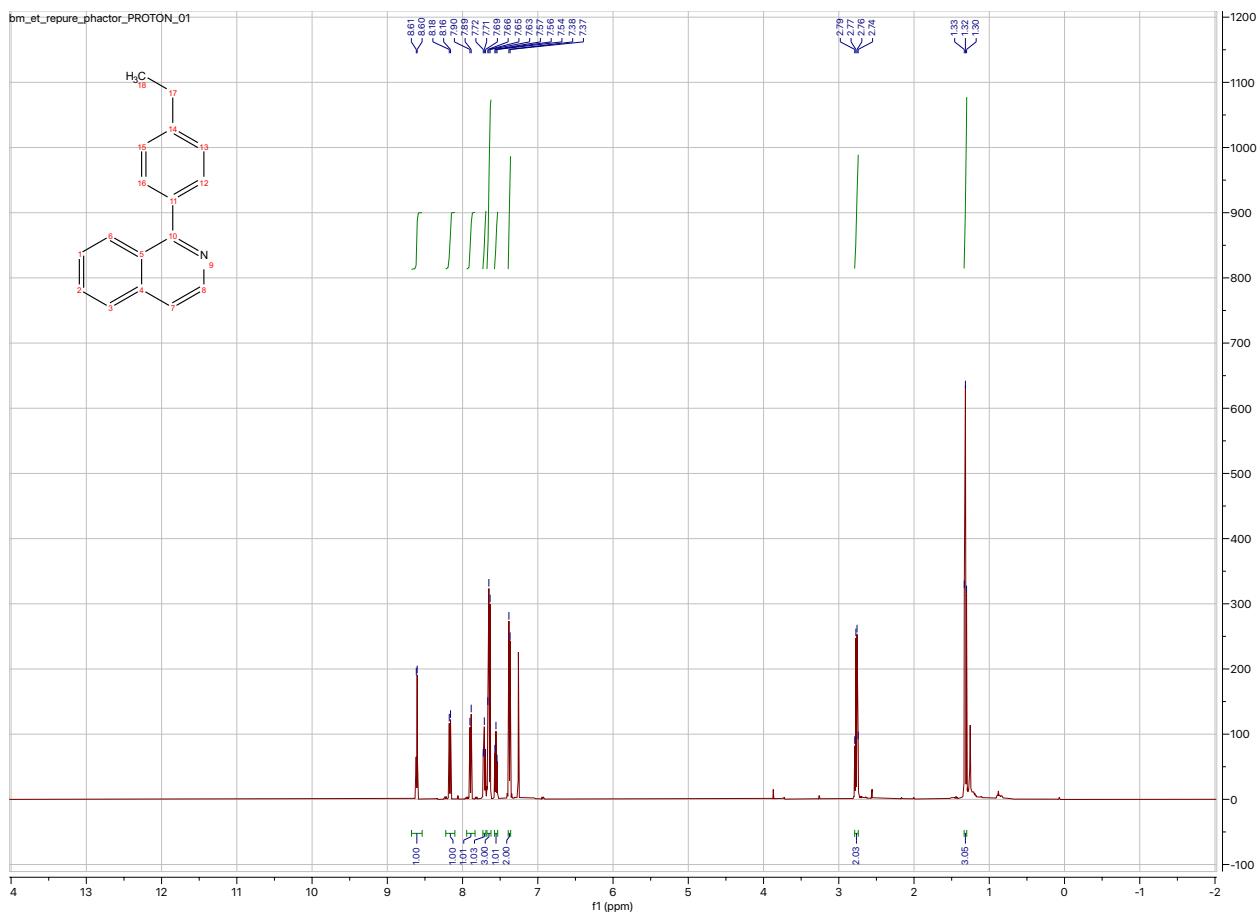


Figure S26. Proton NMR of 1-(4-ethylphenyl)isoquinoline.

3. 1-([1,1'-biphenyl]-4-yl)isoquinoline

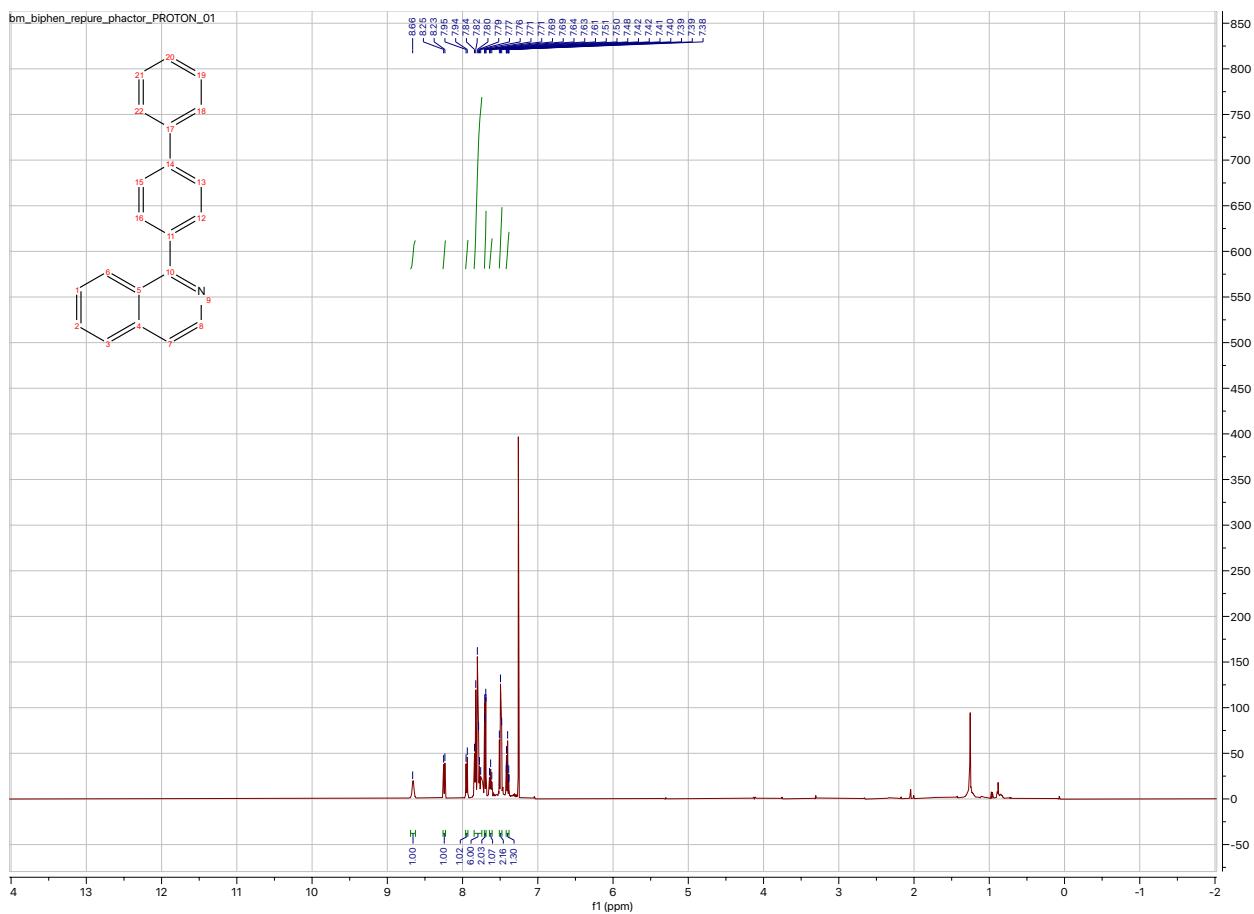


Figure S27. Proton NMR of 1-([1,1'-biphenyl]-4-yl)isoquinoline.