Supporting Information

Reaction Array Fingerprinting

Babak Mahjour[1], Clinton Regan[1], Daniel Schorin[2], Tim Cernak*[1,3]

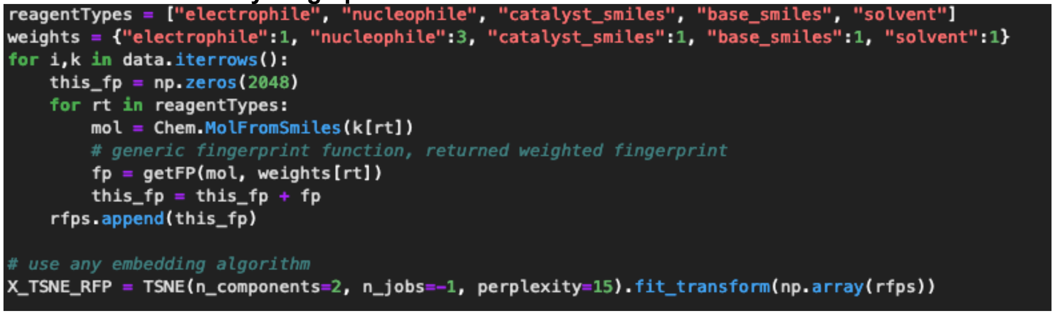
[1]Department of Medicinal Chemistry, University of Michigan
[2]Department of Information, University of Michigan
[3]Department of Chemistry, University of Michigan
*Email: tcernak@umich.edu

Table of Contents

1. **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

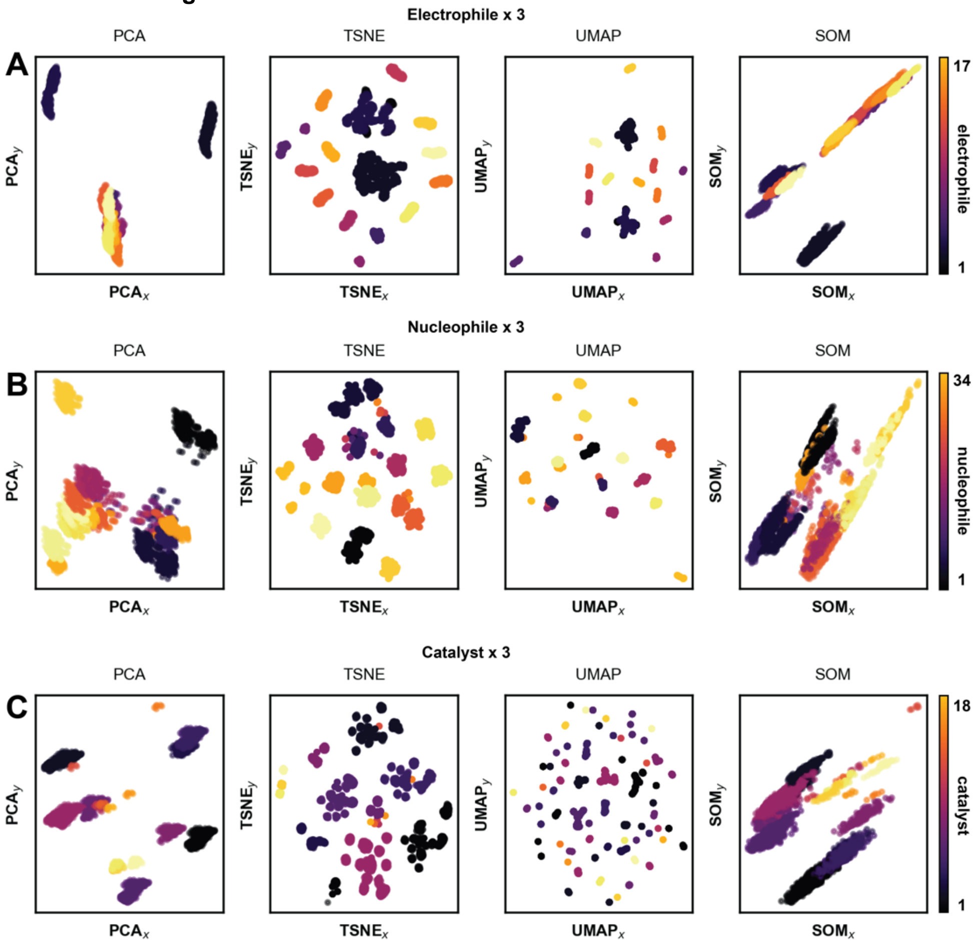


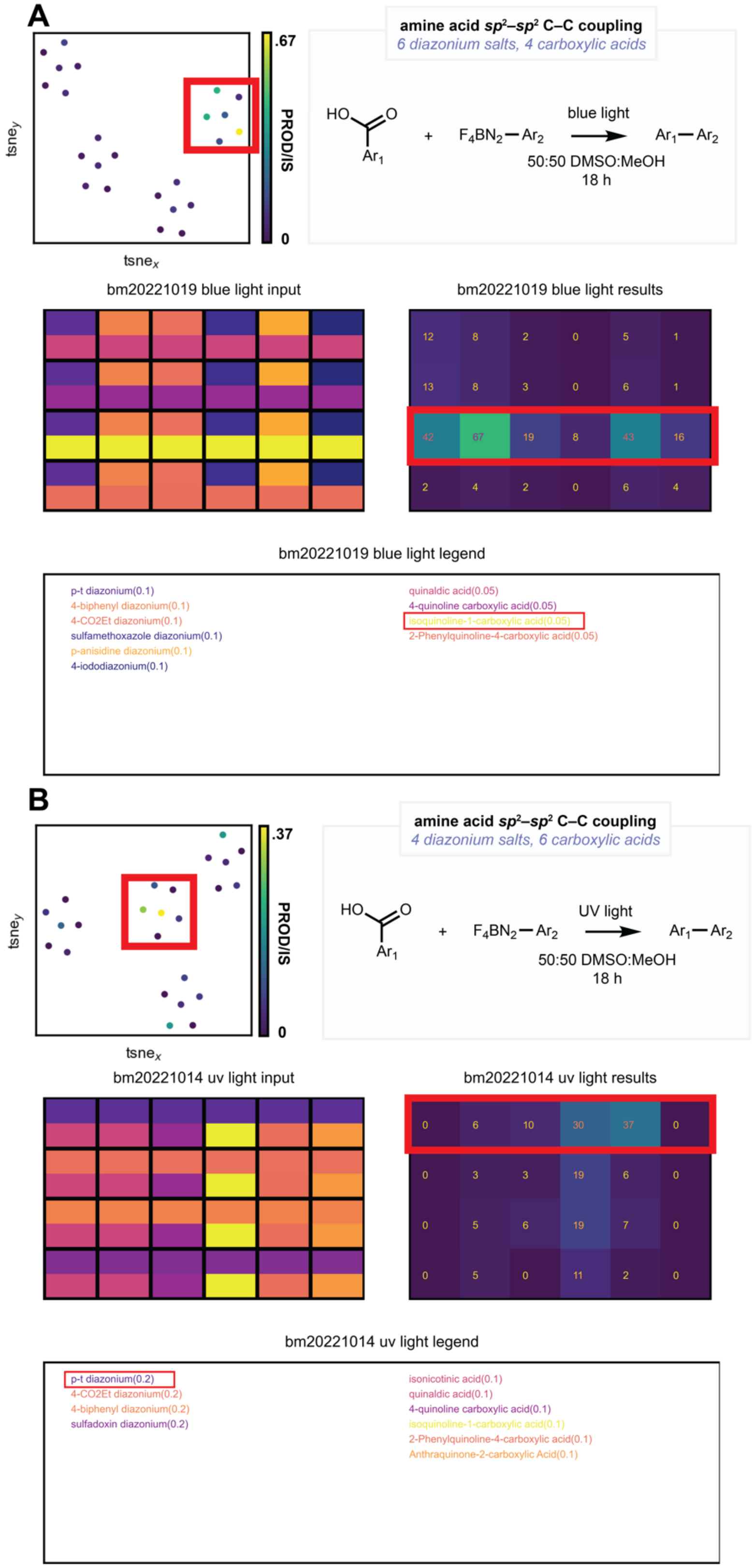
Figure S2. Different dimensionality reduction algorithms performed on the Suzuki dataset weighted reaction fingerprints. **A)** Electrophile fingerprints were multiplied by three. T-SNE and UMAP embeddings formulate distinct clusters. **B)** Nucleophile fingerprints were multiplied by three. **C)** Catalyst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



bm20221019 blue light results

12	8	2	0	5	1
13	8	3	0	6	1
42	67	19	8	43	16
2	4	2	0	6	4

bm20221019 blue light legend

p-t diazonium(0.1)
4-biphenyl diazonium(0.1)
4-CO2Et diazonium(0.1)
sulfamethoxazole diazonium(0.1)
p-anisidine diazonium(0.1)
4-iododiazonium(0.1)

quinaldic acid(0.05)
4-quinoline carboxylic acid(0.05)
isoquinoline-1-carboxylic acid(0.05)
2-Phenylquinoline-4-carboxylic acid(0.05)

bm20221014 uv light results

0	6	10	30	37	0
0	3	3	19	6	0
0	5	6	19	7	0
0	5	0	11	2	0

bm20221014 uv light legend

p-t diazonium(0.2)
4-CO2Et diazonium(0.2)
4-biphenyl diazonium(0.2)
sulfadoxin diazonium(0.2)

isonicotinic acid(0.1)
quinaldic acid(0.1)
4-quinoline carboxylic acid(0.1)
isoquinoline-1-carboxylic acid(0.1)
2-Phenylquinoline-4-carboxylic acid(0.1)
Anthraquinone-2-carboxylic Acid(0.1)

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

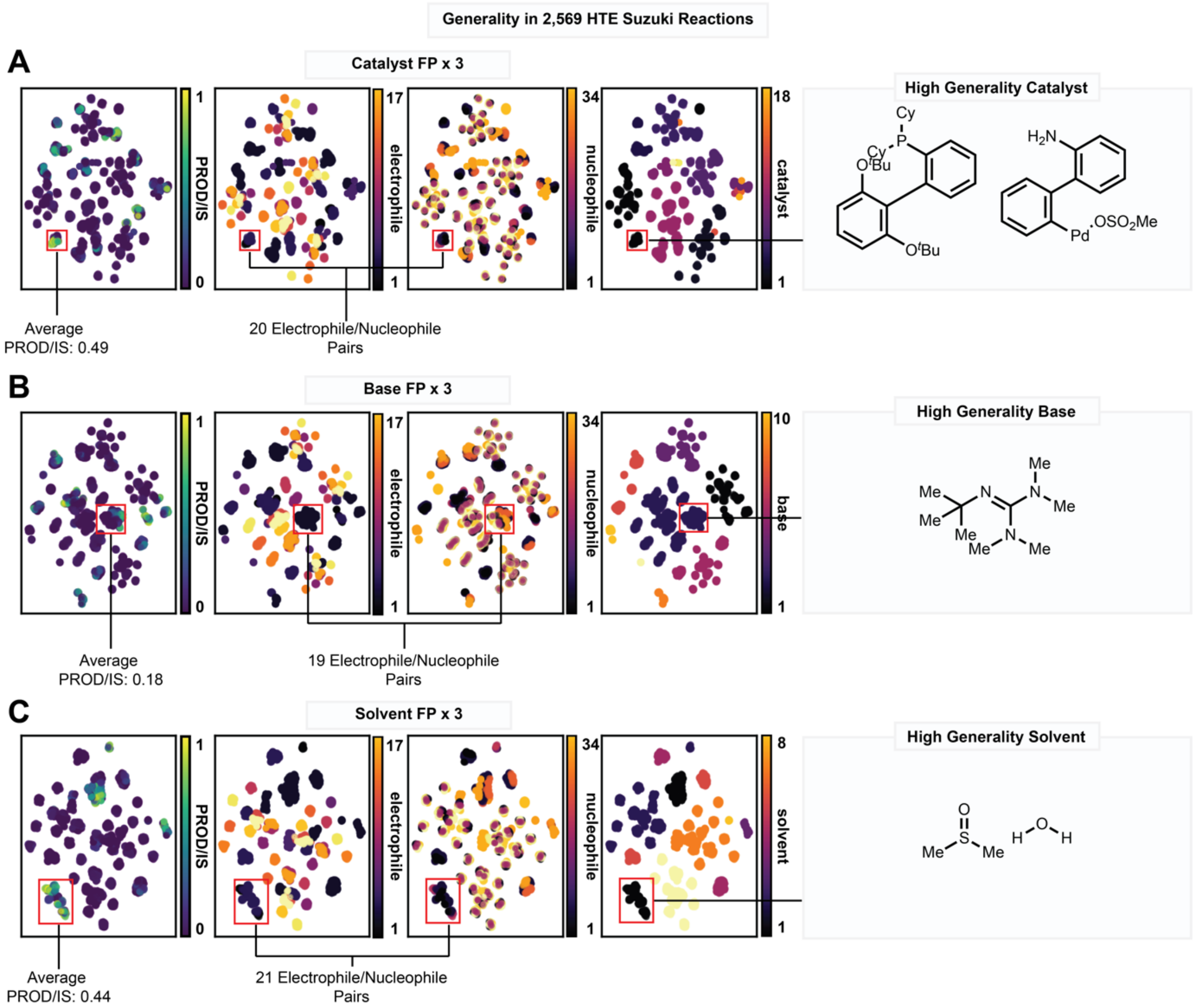


Figure S4. Condition generality demonstrated through the manipulation of weighted reaction fingerprint manifolds of the Suzuki dataset. **A)** When multiplying catalyst fingerprints by three, clusters containing many nucleophile and electrophile substrate pairs that work well with a specific catalyst 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 product 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

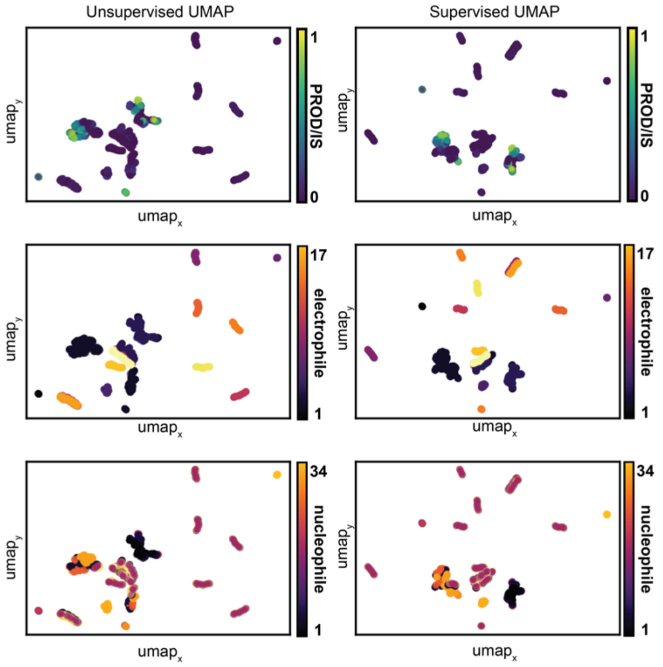


Figure S5. Unsupervised and supervised UMAP dimensionality reduction algorithms are compared. Little difference is indicated between the manifolds.

8. Weighted Reaction Fingerprints vs Concatenated Fingerprints vs Difference Fingerprints

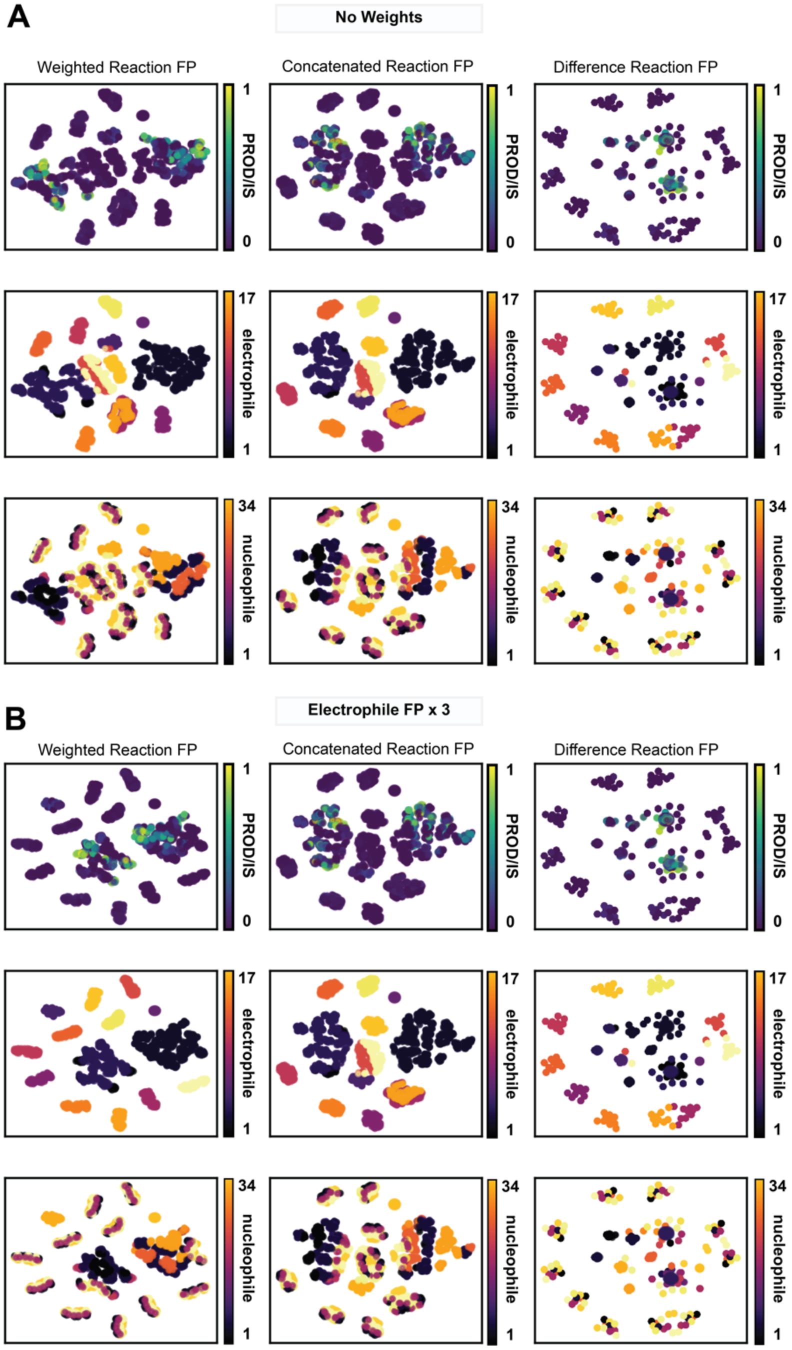
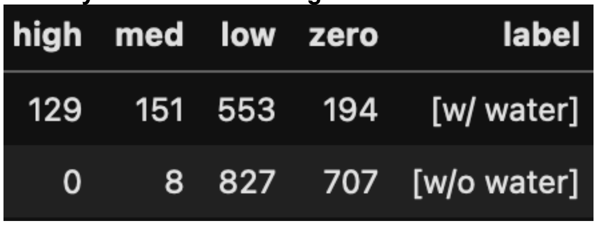


Figure S6. The introduced weighted reaction fingerprint is compared with the commonly used concatenated reaction fingerprints and difference reaction fingerprints found in the literature. **A)** The three fingerprint type manifolds without any feature weights compared side-by-side and colored by product/internal 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.



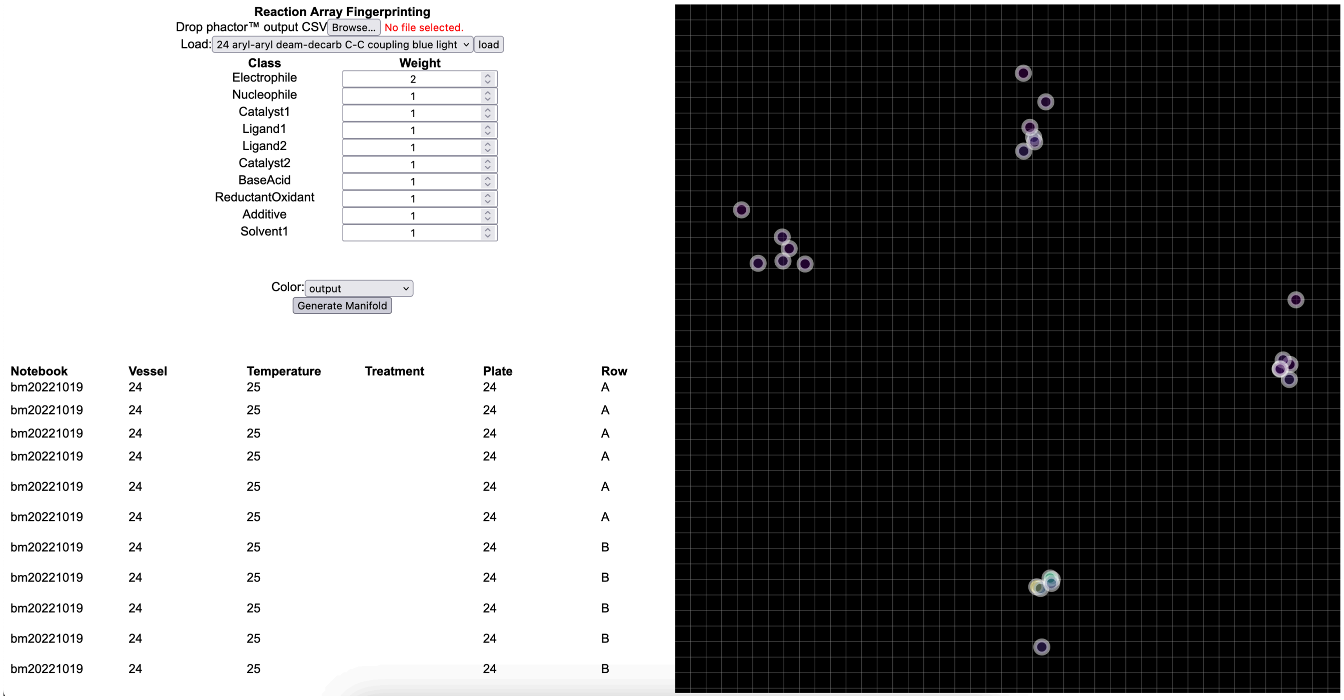
high	med	low	zero	label
129	151	553	194	[w/ water]
0	8	827	707	[w/o water]

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.



Reaction Array Fingerprinting

Drop phactor™ output CSV Browse... No file selected.

Load: 24 aryl-aryl deam-decarb C-C coupling blue light load

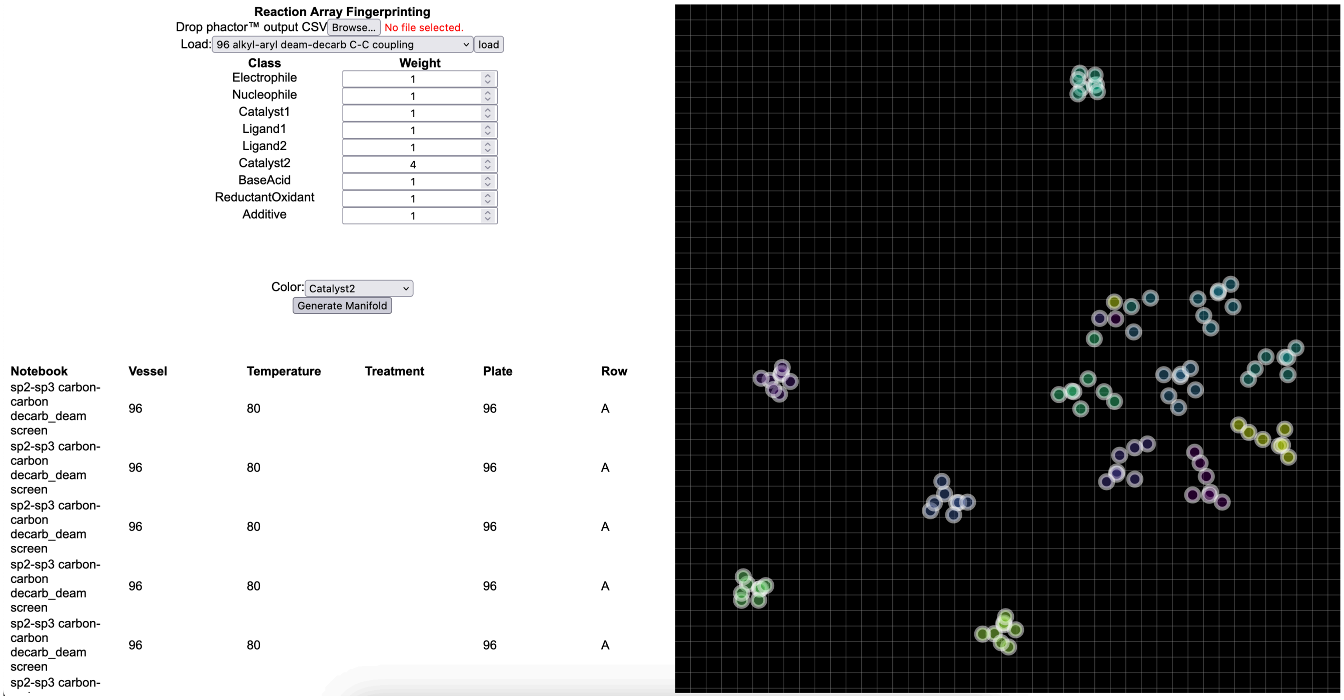
Class	Weight
Electrophile	2
Nucleophile	1
Catalyst1	1
Ligand1	1
Ligand2	1
Catalyst2	1
BaseAcid	1
ReductantOxidant	1
Additive	1
Solvent1	1

Color: output

Generate Manifold

Notebook	Vessel	Temperature	Treatment	Plate	Row
bm20221019	24	25		24	A
bm20221019	24	25		24	A
bm20221019	24	25		24	A
bm20221019	24	25		24	A
bm20221019	24	25		24	A
bm20221019	24	25		24	A
bm20221019	24	25		24	B
bm20221019	24	25		24	B
bm20221019	24	25		24	B
bm20221019	24	25		24	B
bm20221019	24	25		24	B

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



Reaction Array Fingerprinting

Drop phactor™ output CSV Browse... No file selected.

Load: 96 alkyl-aryl deam-decarb C-C coupling load

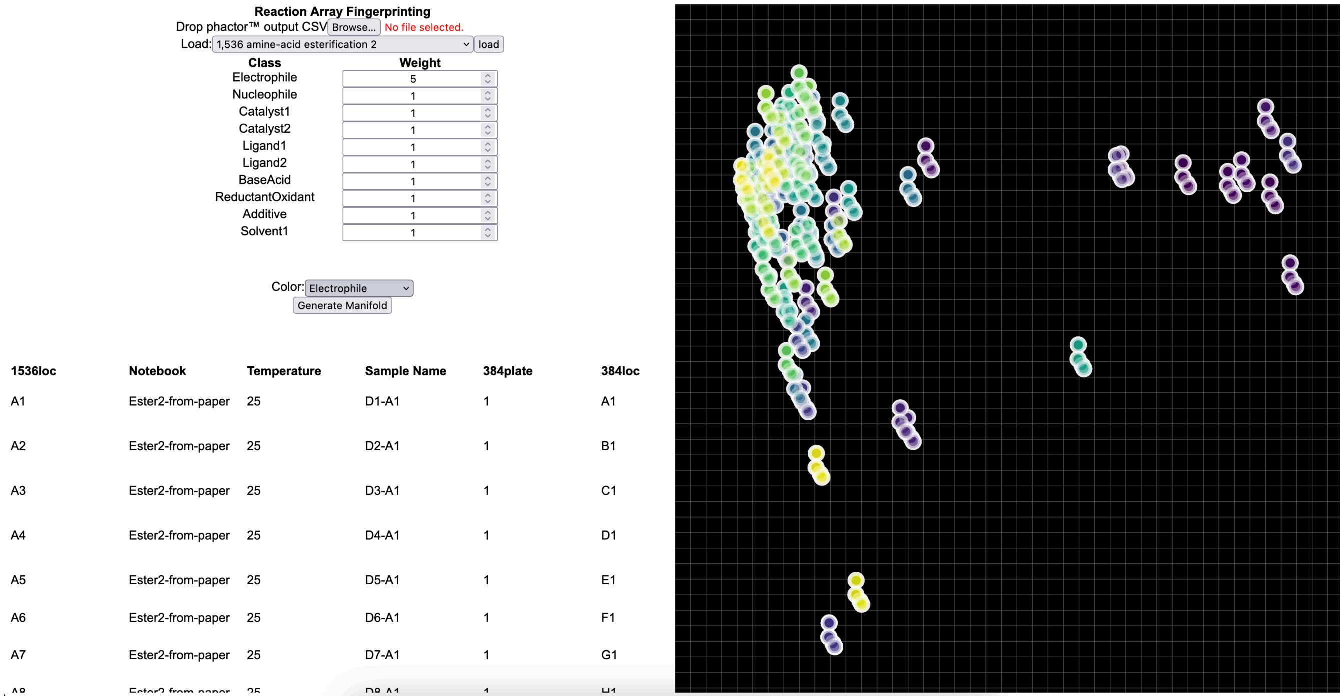
Class	Weight
Electrophile	1
Nucleophile	1
Catalyst1	1
Ligand1	1
Ligand2	1
Catalyst2	4
BaseAcid	1
ReductantOxidant	1
Additive	1

Color: Catalyst2

Generate Manifold

Notebook	Vessel	Temperature	Treatment	Plate	Row
sp2-sp3 carbon-carbon decarb_deam screen	96	80		96	A
sp2-sp3 carbon-carbon decarb_deam screen	96	80		96	A
sp2-sp3 carbon-carbon decarb_deam screen	96	80		96	A
sp2-sp3 carbon-carbon decarb_deam screen	96	80		96	A
sp2-sp3 carbon-carbon decarb_deam screen	96	80		96	A
sp2-sp3 carbon-					

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



Reaction Array Fingerprinting

Drop phactor™ output CSV Browse... No file selected.

Load: 1,536 amine-acid esterification 2 load

Class	Weight
Electrophile	5
Nucleophile	1
Catalyst1	1
Catalyst2	1
Ligand1	1
Ligand2	1
BaseAcid	1
ReductantOxidant	1
Additive	1
Solvent1	1

Color: Electrophile

Generate Manifold

1536loc	Notebook	Temperature	Sample Name	384plate	384loc
A1	Ester2-from-paper	25	D1-A1	1	A1
A2	Ester2-from-paper	25	D2-A1	1	B1
A3	Ester2-from-paper	25	D3-A1	1	C1
A4	Ester2-from-paper	25	D4-A1	1	D1
A5	Ester2-from-paper	25	D5-A1	1	E1
A6	Ester2-from-paper	25	D6-A1	1	F1
A7	Ester2-from-paper	25	D7-A1	1	G1
A8	Ester2-from-paper	25	D8-A1	1	H1

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

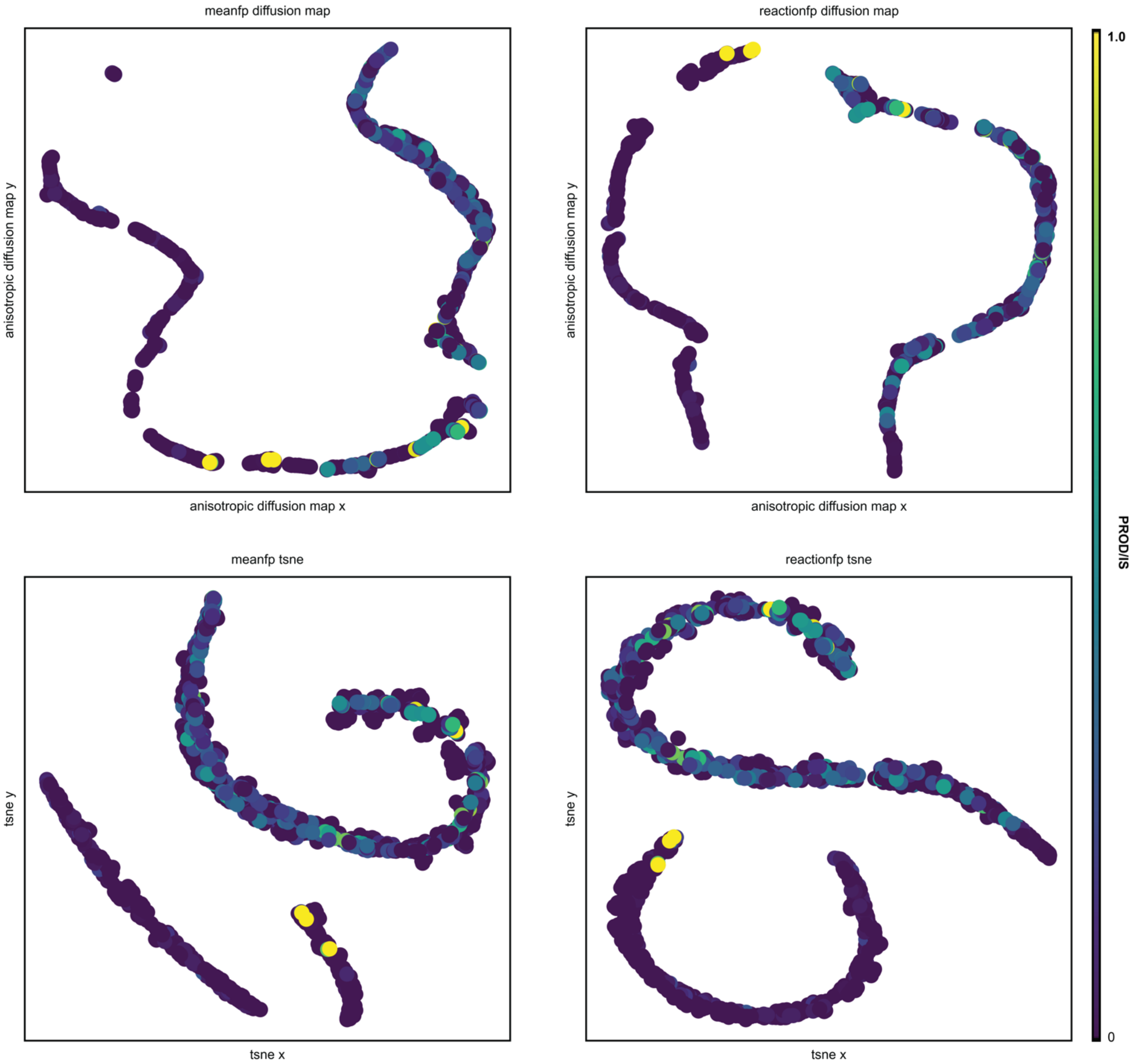


Figure S11. Comparison of mean fingerprints against summed fingerprints when including molecular weight and logp in tSNE and UMAP manifolds.