Supporting Information

Reaction Array Fingerprinting

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

[1]Department of Medicinal Chemistry, University of Michigan, Ann Arbor, Michigan 48109
[2]School of Information, University of Michigan, Ann Arbor, Michigan 48109
[3]Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109
*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. **Python 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. Comparison of SOMs, TSNEs, PCAs, and UMAPs for Suzuki Reaction Dataset Reaction Array Fingerprints

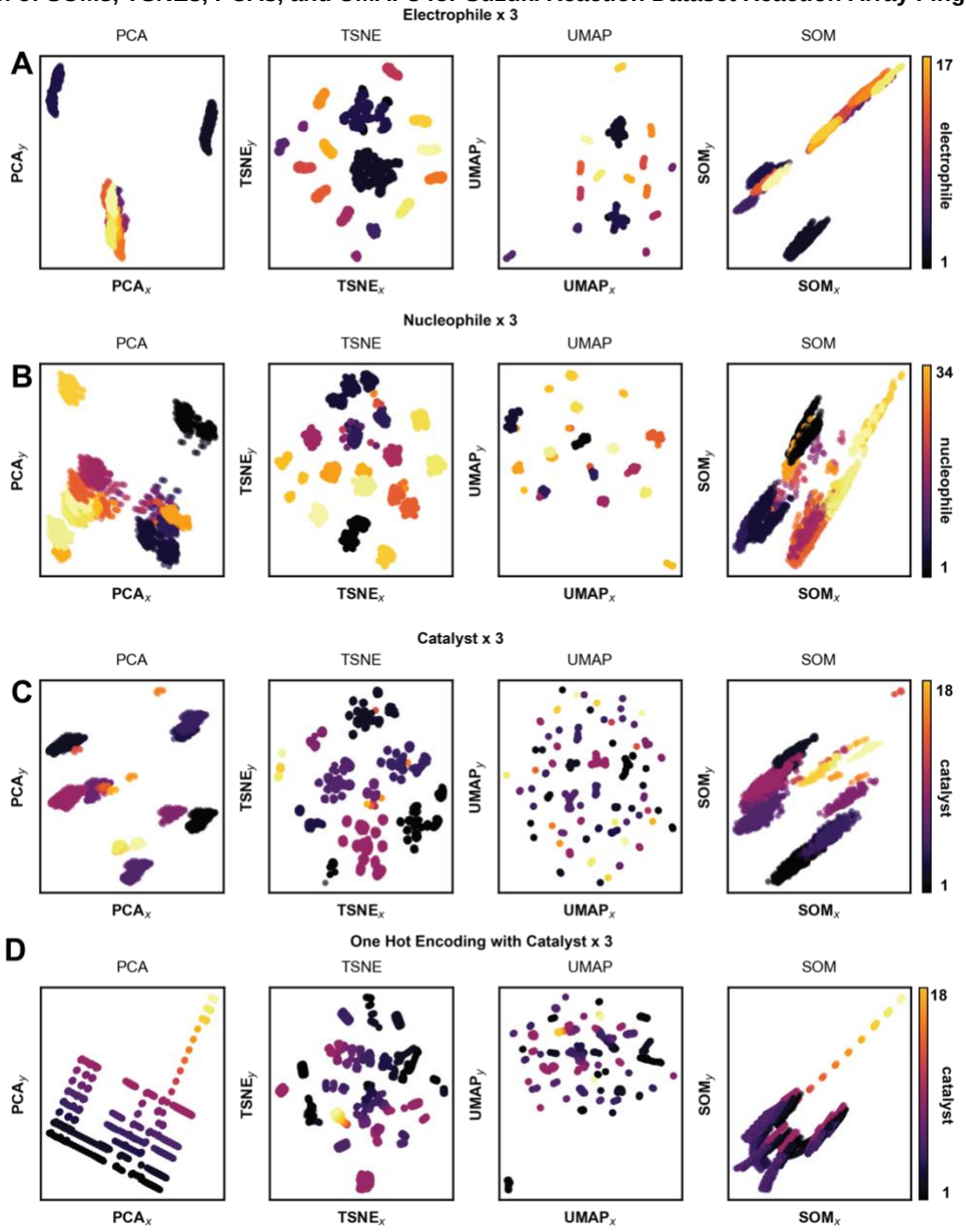

Figure S2. Different dimensionality reduction algorithms performed on the Suzuki dataset weighted reaction fingerprints. **A)** Electrophile fingerprints were multiplied by three. T-SNE and UMAP embeddings formulate distinct clusters. **B)** Nucleophile fingerprints were multiplied by three. **C)** Catalyst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

4. Perplexity Sweep Experiment

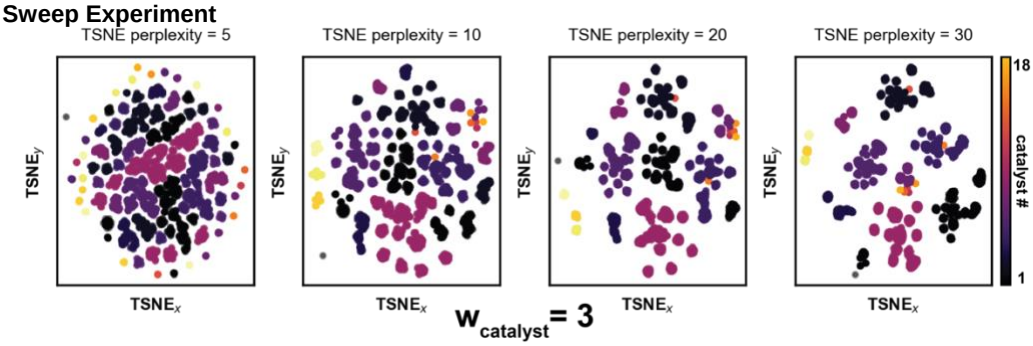

Figure S3. The perplexity value was changed for the Suzuki dataset where the weight of the catalyst was elevated to 3. The standard range for tSNE perplexity is 5-30.

5. Perplexity Sweep Experiment with MACCS fingerprints

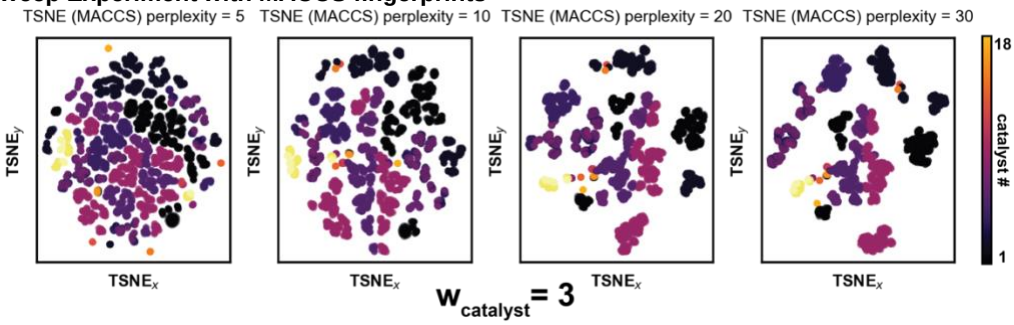

Figure S4. The perplexity value was changed for the Suzuki dataset where the weight of the catalyst was elevated to 3 when using MACCS as a fingerprint.

6. Unsupervised vs supervised UMAP Experiment

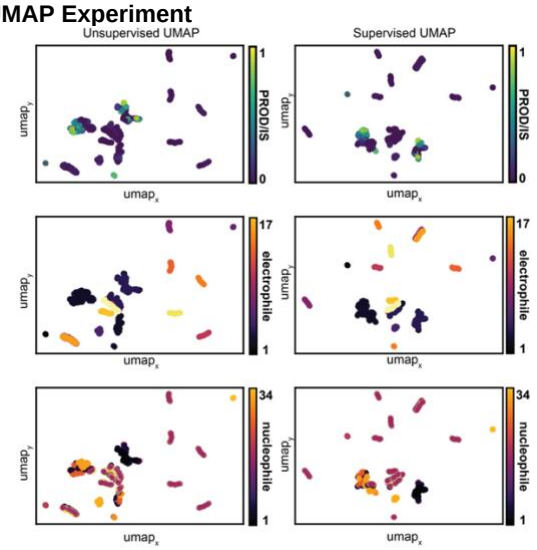

Figure S5. Unsupervised and supervised UMAP dimensionality reduction algorithms are compared. Little difference is indicated between the manifolds.

7. Weighted Reaction Fingerprints vs Concatenated Fingerprints vs Difference Fingerprints

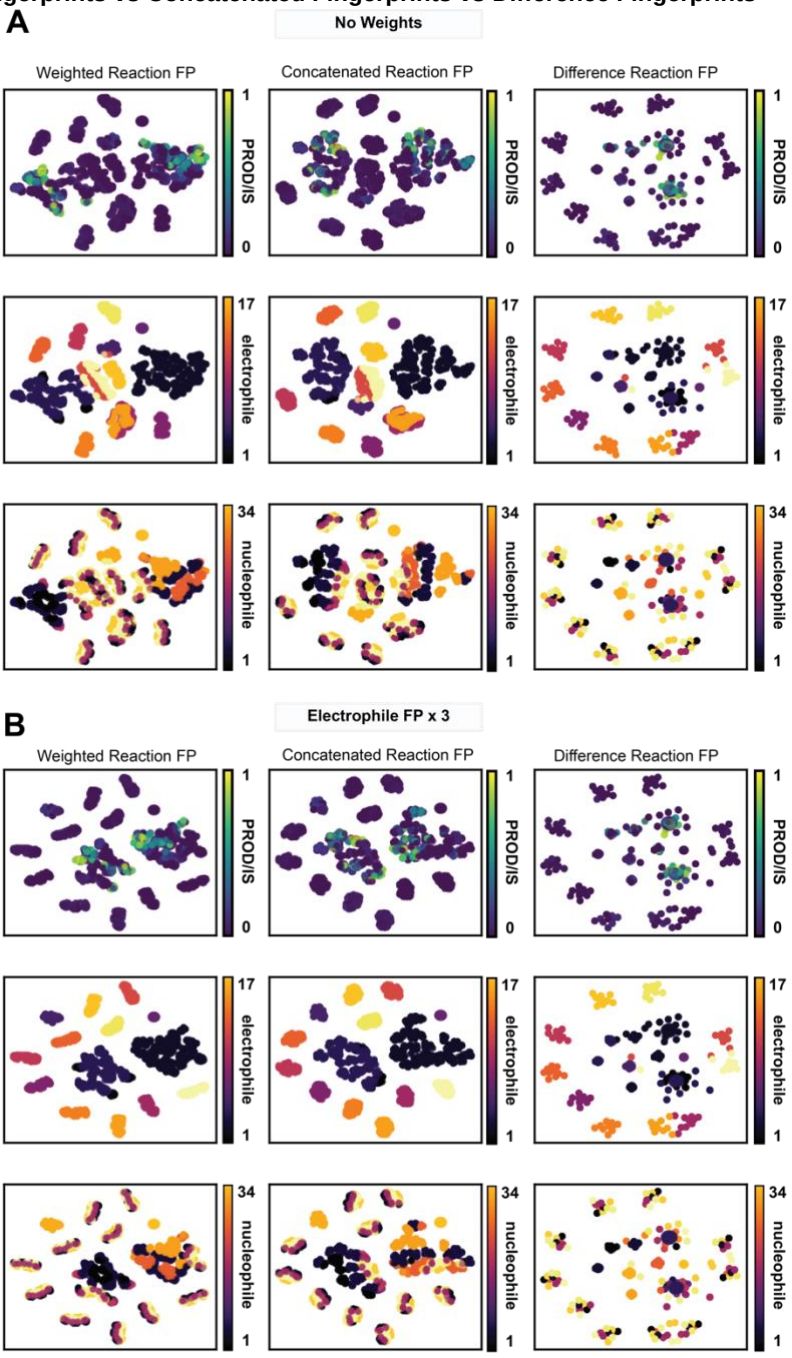

Figure S6. The introduced weighted reaction fingerprint is compared with the commonly used concatenated reaction fingerprints and difference reaction fingerprints found in the literature. **A)** The three fingerprint type manifolds without any feature weights compared side-by-side and colored by product/internal 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 noticed in the new weighted reaction fingerprint method and the concatenated reaction fingerprint (c.f. S6A).

8. Identifying reagents with generality

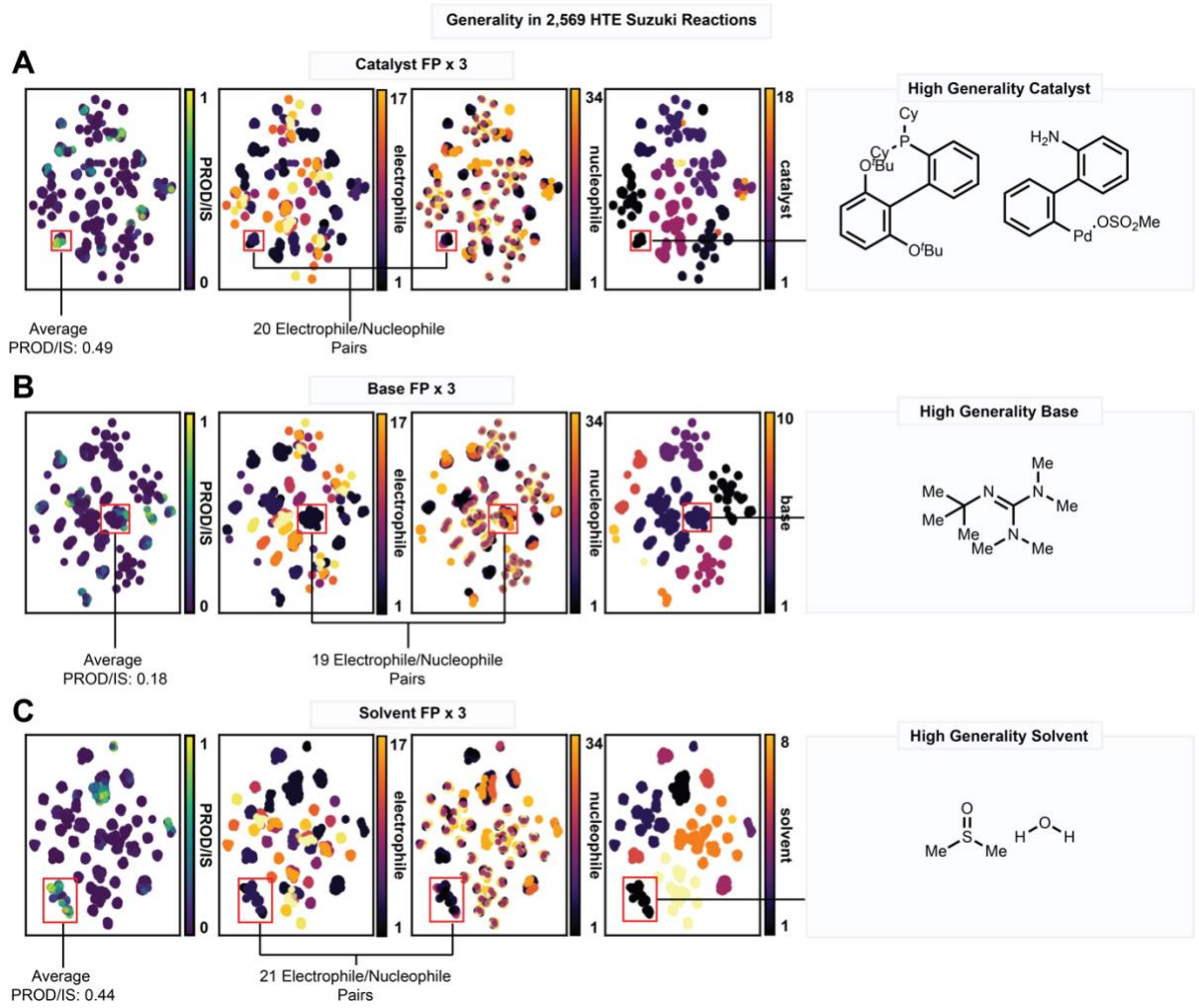

Figure S7. Condition generality demonstrated through the manipulation of weighted reaction fingerprint manifolds of the Suzuki dataset. **A)** When multiplying catalyst fingerprints by three, clusters containing many nucleophile and electrophile substrate pairs that work well with a specific catalyst 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.

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 S8. 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. Box plots of solvent systems of data for figure 4

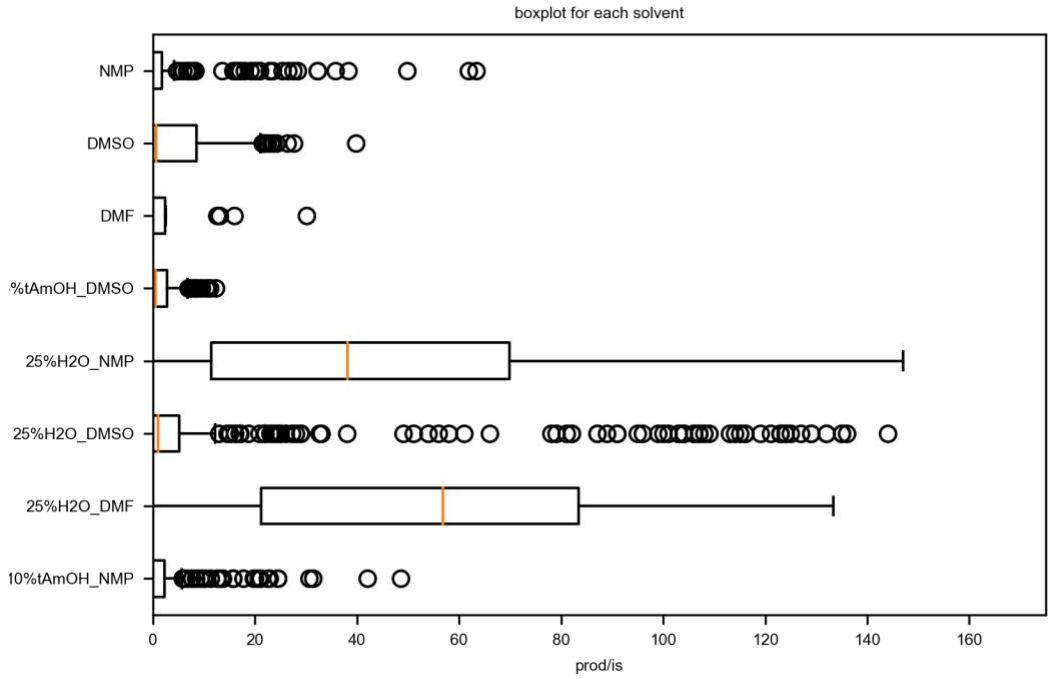

Figure S9. Boxplots grouped by solvent of reactions in the Suzuki dataset. While the best solvents are evident by average, the relation between solvent structure is lost.

11. Hyperparameters and weights used to generate figure 5 visualizations directly from phactor™ output data

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. In this case the random state 1 was used in all experiments.

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

12. Pivot Table Heatmaps for D2B chemistry data of Figure 5E

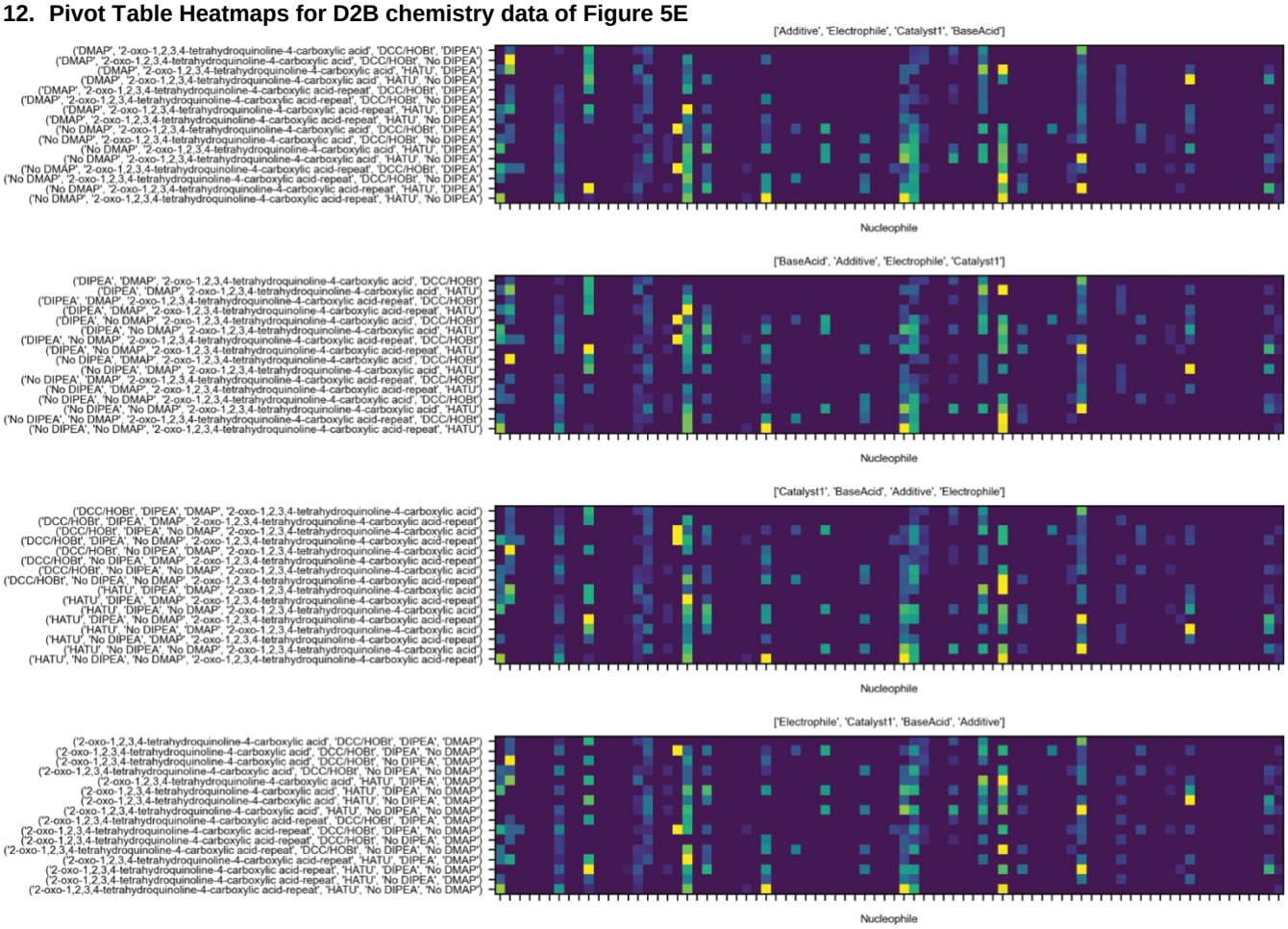

Figure S10. Pivot table heatmaps for the data of Figure 5E. Changing the hierarchy of row indices facilitates comparisons between datapoints.