Supporting Information

Chespa: streamlining expansive chemical space evaluation of molecular sets

Jamie R. Nuñez[1,2], Monee Mcgrady[1], Yasemin Yesiltepe[1,2], Ryan S. Renslow[1,2,*], Thomas O. Metz[1,†]

[1] Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA 99352

[2] The Gene and Linda Voiland School of Chemical Engineering and Bioengineering, Washington State University, Pullman, WA, USA 99164

* ryan.renslow@pnnl.gov
† thomas.metz@pnnl.gov

This document contains additional information to support the manuscript cited above. Included here are additional method details, captions for the data provided in SupportingData.xlsx, and additional figures that support claims made in the manuscript. Please see manuscript for context.

Results and Discussion

RD1. Grouping and Clustering.

Property chemical space clusters: On average, the size of these clusters is 543.3 (σ: 442.3) molecules, with the largest cluster (ChemSpace Cluster 6) covering 1,340 compounds and the smallest cluster (ChemSpace Cluster 5) covering 12 compounds.

DarkChem clusters: On average, the size of these clusters is 408.0 (σ: 246.1), with the largest cluster (DarkChem Cluster 6) covering 774 compounds and the smallest cluster (DarkChem Cluster 5) covering 49 compounds. A total of 3,264 compounds are included in these groups, as some molecules fell outside of the restrictions placed on DarkChem space (such as only including compounds that are composed of sulfur, phosphorous, oxygen, nitrogen, carbon, and/or hydrogen).

ClassyFire clusters: On average, the size of these groups is 543.3 (σ: 401.6), with the largest group (ClassyFire Superclass Group 1) covering 1,396 compounds and the smallest group (ClassyFire Superclass Group 7) covering 106 compounds.

MACCS clusters: On average, the size of these clusters is 543.0 (σ: 171.9), with the largest cluster (MACCS Cluster 4) covering 844 compounds and the smallest cluster (MACCS Cluster 5) covering 256 compounds. A total of 4,344 compounds are included in these groups since 4 compounds failed to return results when using Open Babel's tool.

SPECTRe clusters: On average, the size of these clusters is 148.9 (σ: 70.6), with the largest cluster (SPECTRe Cluster 3) covering 272 compounds and the smallest cluster (SPECTRe Cluster 8) covering 45 compounds. A total of 1,191 compounds (all from the list of spiked in compounds) were included in these groups, due to time and memory requirements.

RD2. Analysis of Trends in Observability.

Molecules within MACCS Cluster 8 also appear to not have been easily observable (5 observed members and 71 not observed members, which is 14.2 times more). Significant trends that stand out here, however, are not the same as ChemSpace Cluster 1. Molecules in this cluster appear to contain many more oxygens, as indicated by its O/C and O/H ratios (0.39 and 0.35, respectively) compared to all spiked in compounds (0.21, p-value: 5.9E-23, and 0.20, p-value: 6.4E-15, respectively) while also having very few nitrogens as indicated by its N/C and N/H ratios (0.01 and 0.01, respectively) compared to all spiked in compounds (0.13, p-value: 3.4E-13, and 0.11, p-value: 7.8E-14, respectively). The average mass of this cluster (221.1) is still significantly lower than all spiked in compounds (p-value: 4.2E-5), though not by as much as ChemSpace Cluster 1.

Clustering with SPECTRe substructures also provides a couple interesting clusters that do not appear to be easily observable, SPECTRe Cluster 1 observable (8 observed members and 56 not observed members, which is 7 times more) and 8 (3 observed members and 39 not observed members, which is 16 times more). These clusters follow the same trend as that seen in MACCS Cluster 8, with the exception that SPECTRe Cluster 8 actually has significantly *lower* oxygens compared to all spiked in compounds as indicated by its O/C and O/H ratios (0.15, p-value: 3.7E-3 and 0.11, p-value: 7.7E-5, respectively).

Some clusters contained many of the compounds that were observed (ChemSpace Cluster 3 and 5 and DarkChem Cluster 5), but the number of members observed or not observed is quite low (n=25, 4, and 11, respectively). Additionally, in general, ClassyFire Superclass does not appear to trend well with observability for this study.

When considering all compounds observed (n=545) vs. not observed (n=853), properties that are significantly different are mass (p-value = 3E-58), N/C (p-value = 1E-9), N/H (p-value = 9E-11), O/C (2E-3), O/H (p-value = 8E-4), most acidic pKa (p-value = 1.3E-5), Balaban index (p-value = 5E-12), and Harary index (p-value = 9E-55) (Figure S34).

SupportingData.xlsx Captions

SD Tab 1. Suspect library.

Library derived from the original ToxCast library sent by the EPA using the method described in Nunez et al.[1]

SD Tab 2. ClassyFire.

ClassyFire Superclass results and label assigned to each compound. For Superclass to group label mappings, see Counts tab.

SD Tab 3. ChemSpace.

Chemical Space calculated properties, their transform in PC space (PC1-10), and cluster label assigned to each compound.

SD Tab 4. DarkChem.

DarkChem latent space vectors and cluster label assigned to each compound.

SD Tab 5. MACCS_Substructures.

MACCS substructures (fingerprints) present in each compound (1 = Yes/Present, 0 = No/Not Present) and cluster label assigned to each compound. For the 166 substructures pre-defined by Openbabel, see Maccs_SMARTS.txt file provided with this paper.

SD Tab 6. SPECTRe_Substructures.

SPECTRe substructures present in each compound (1 = Yes/Present, 0 = No/Not Present) and cluster label assigned to each compound.

SD Tab 7. SPECTRe_Substructures_6.

Substructures found in each compound (1 = Yes/Present, 0 = No/Not Present) for the top 6 most common and largest substructures found in the spiked in compounds. These substructures are labeled as B1-6. For substructure to group label mappings, see Counts tab.

SD Tab 8. Counts.

Number of compounds in each group/cluster and name->label mappings.

SD Tab 9. SpikedIn.

Compounds spiked into each sample for the ENTACT challenge.

SD Tab 10. Observed.

Compounds observed by ESI positive and/or negative mode (meaning they received a score of 11.23 or greater) in each sample from the ENTACT challenge.

SD Tab 11. NotObs.

Compounds not observed (meaning they received a score of 0) in each sample from the ENTACT challenge.

Tables

Tool	Tool Run Time	Clustering Overhead	Clustering (Input Size)	Plotting Overhead	Plotting
ClassyFire	<1 s	NA	NA	4 min	30 sec
ChemSpace (cxcalc)	5 – 7 s*	35 s	2 s	4 min	30 sec
DarkChem	<1 s	38 s	3 s	4 min	30 sec
OpenBabel (MACCS)	<1 s	45 s	47 min	4 min	30 sec
SPECTRe	<1 s – 48 min	65 s	6 hr	4 min	30 sec

Table S1. Timing when using Chespa. “Clustering Overhead” and “Plotting Overhead” accounts for time to read in associated data used in this paper. “Clustering” represents the time to perform KMeans or binary clustering on the full dataset and “Plotting” the time to generate all images seen in the manuscript and this Supporting Information document. “Tool Run Time” is the time for the tool itself to run on a single molecule from the library, considering a single process (though the calculation of these properties is trivially parallelizable). *only the 10 properties used in this paper were calculated.

Figures

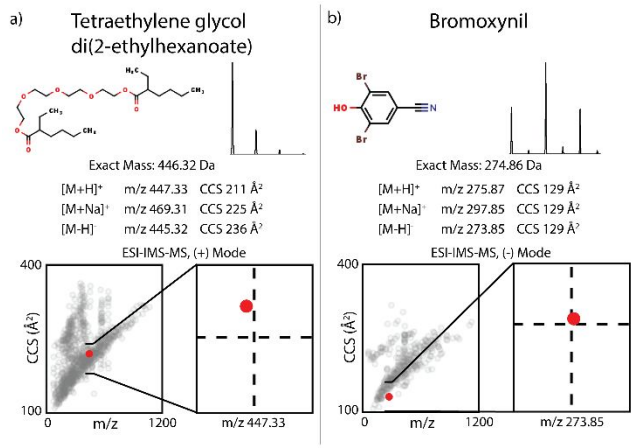

Figure S1. Example observed (spiked in and labeled as "present") compounds and data from the ENTACT challenge, where all evidence of presence came from only (a) (+)-ESI or (b) (-)-ESI. Dashed lines represent where the feature was expected to be (based on adduct *m/z* and predicted CCS), and red dots represent where the experimental feature was found.

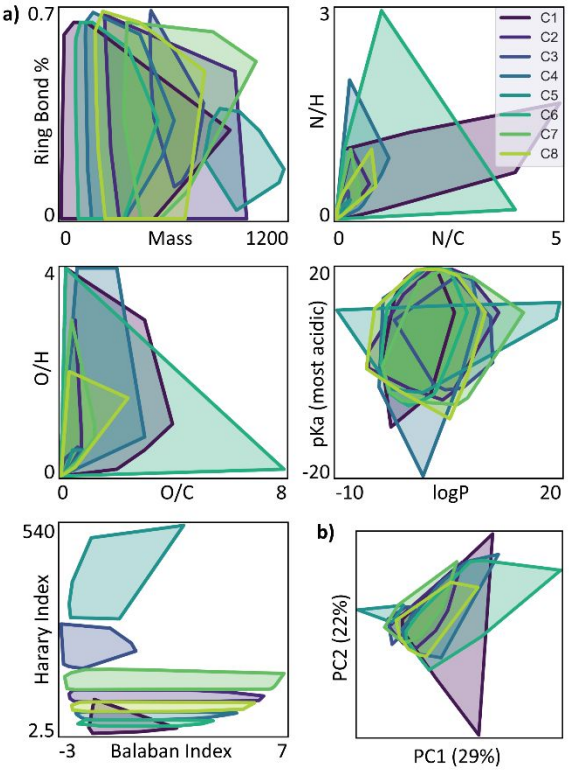

Figure S2. Chemical space covered by each KMeans cluster: (a) Distribution of predicted properties. (b) PC1 and 2 from the principal component analysis performed on the properties plotted in (a).

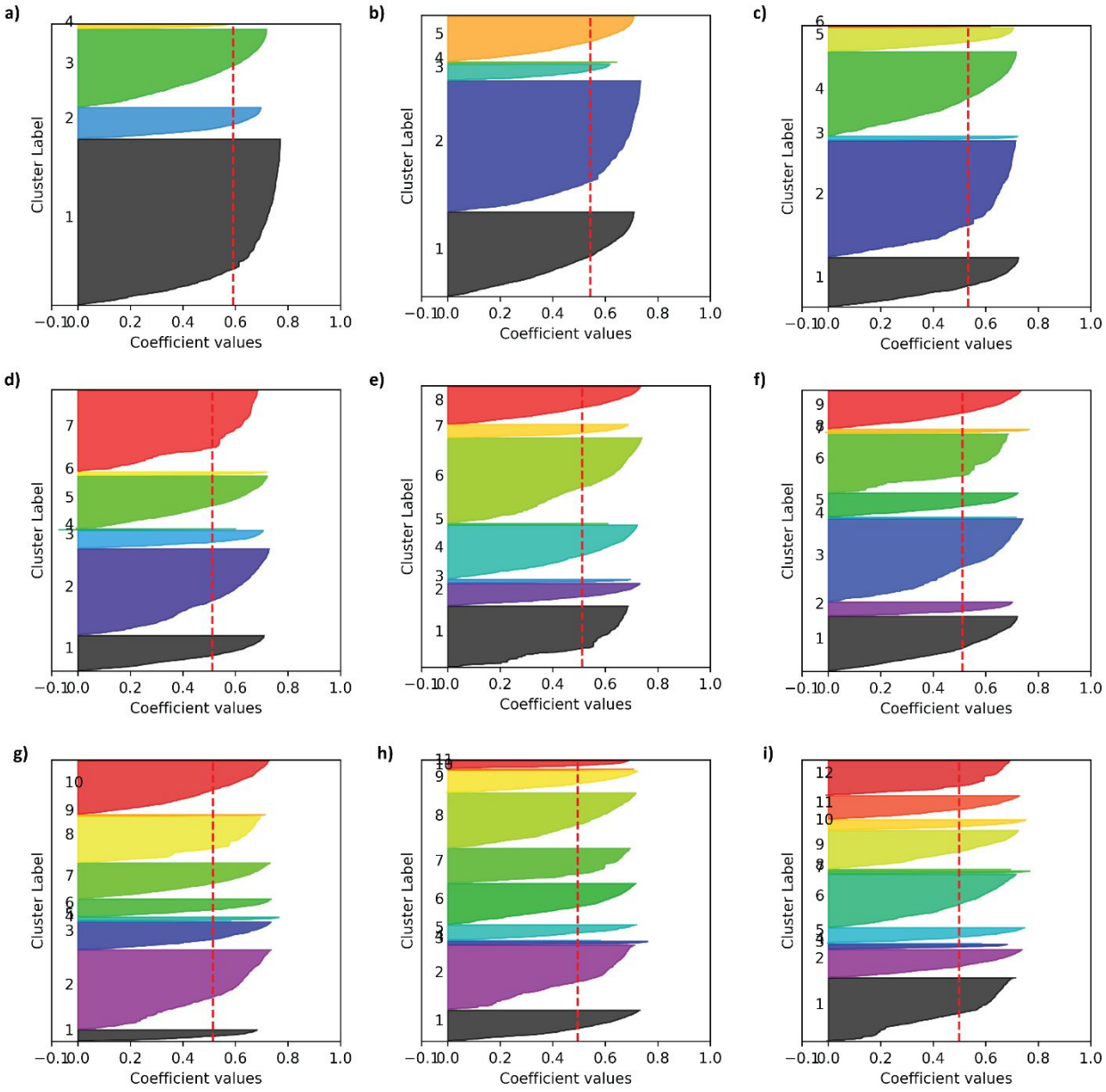

Figure S3. Result of silhouette analysis performed on Chemical Space data, using 4-12 clusters (plots a-i, respectively).

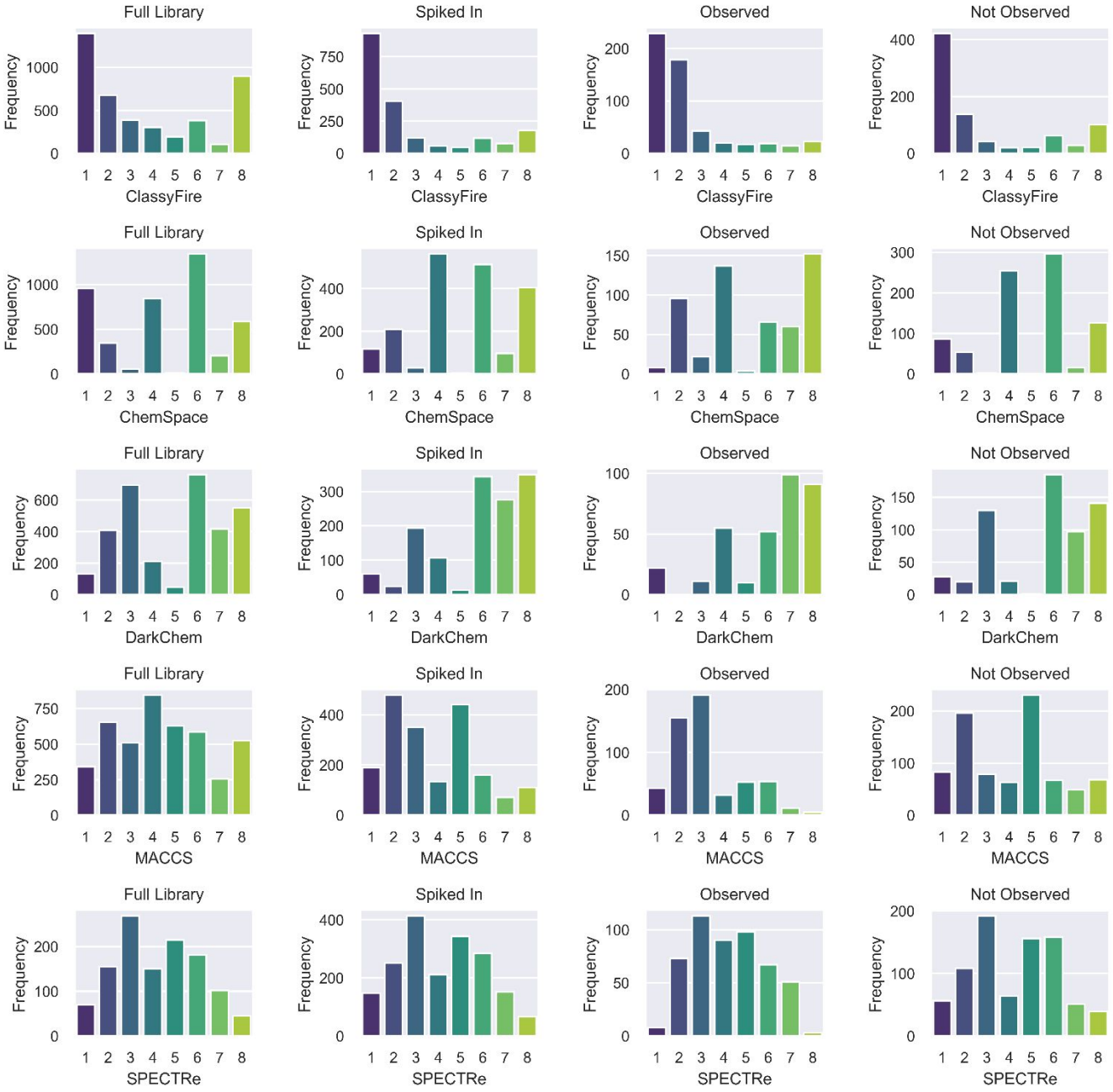

Figure S4. Number of compounds in each group/cluster, given the full suspect library or only spiked in, observed, or not observed compounds.

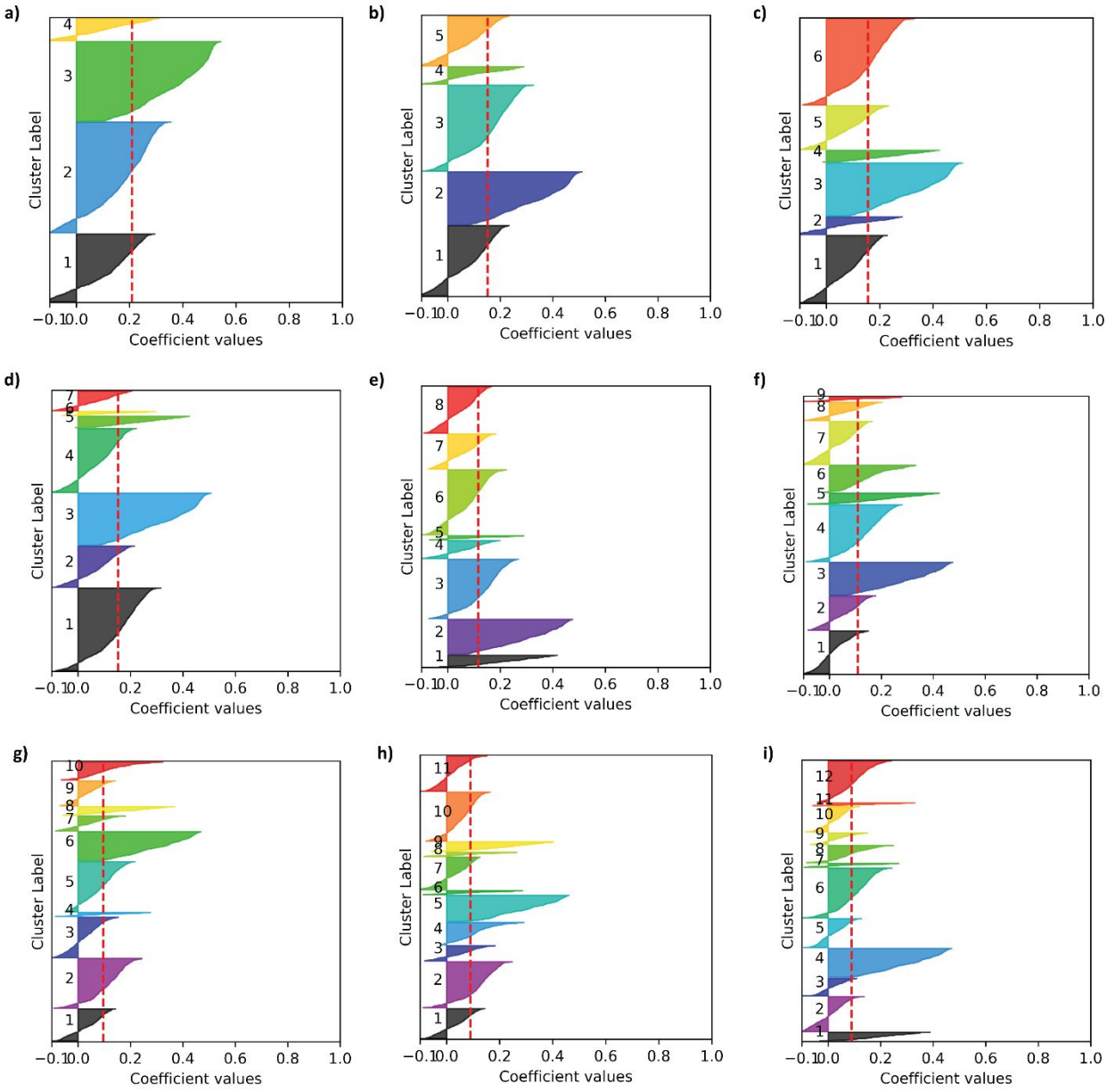

Figure S5. Result of silhouette analysis performed on DarkChem data, using 4-12 clusters (plots a-i, respectively).

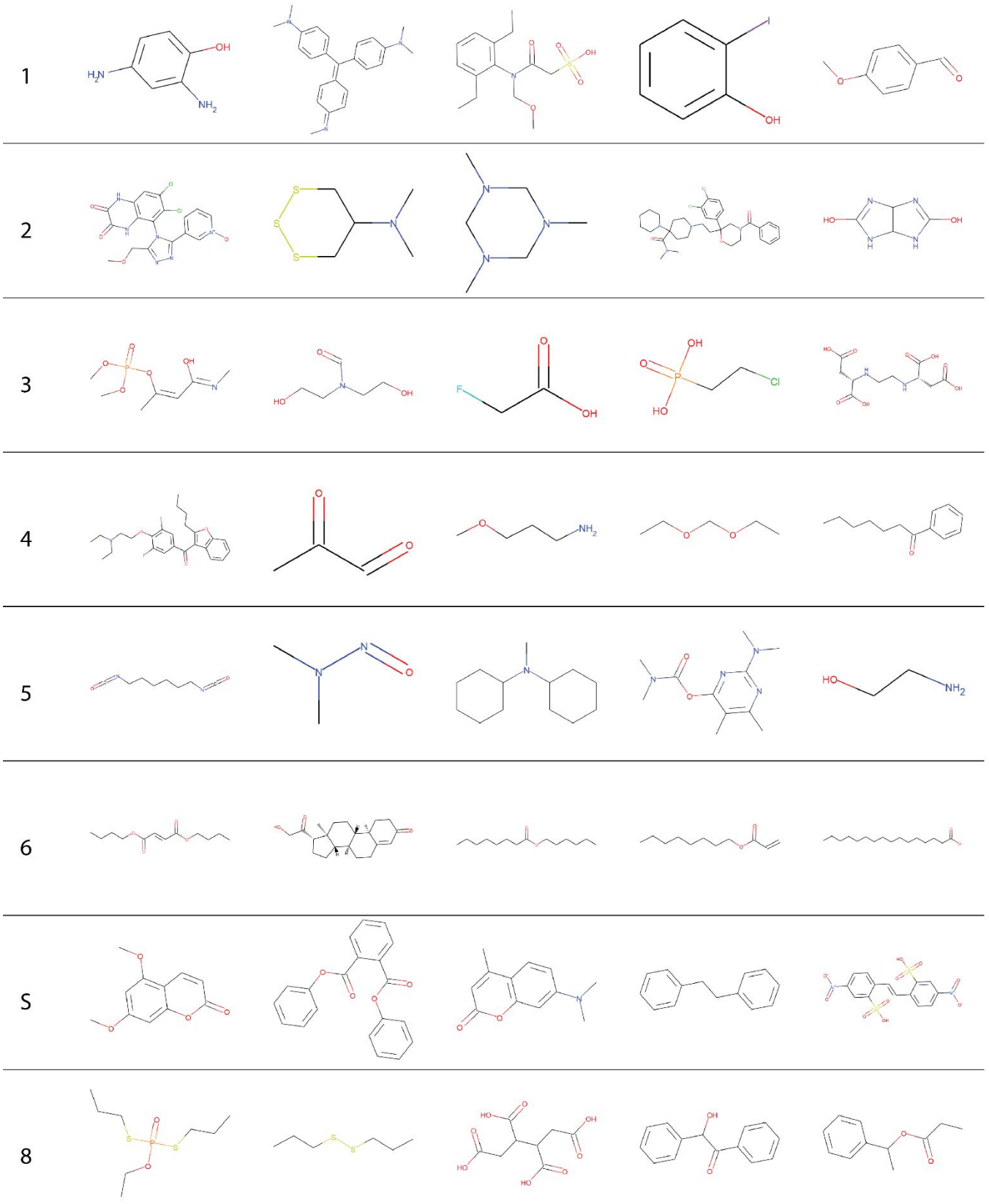

Figure S6. Five randomly chosen compounds in each of the 8 ClassyFire Superclass groups.

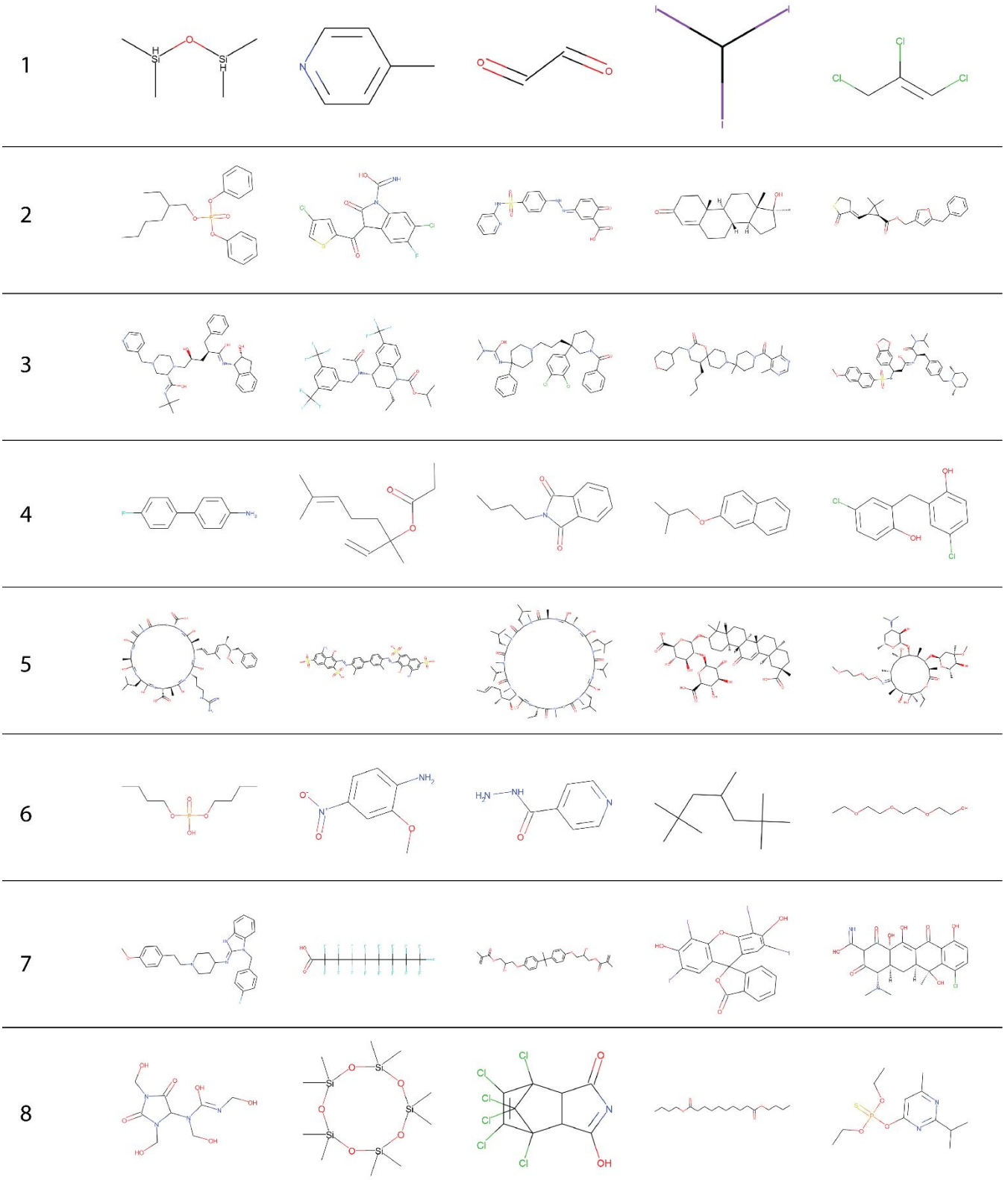

Figure S7. Five randomly chosen compounds in each of the 8 Chemical Space clusters.

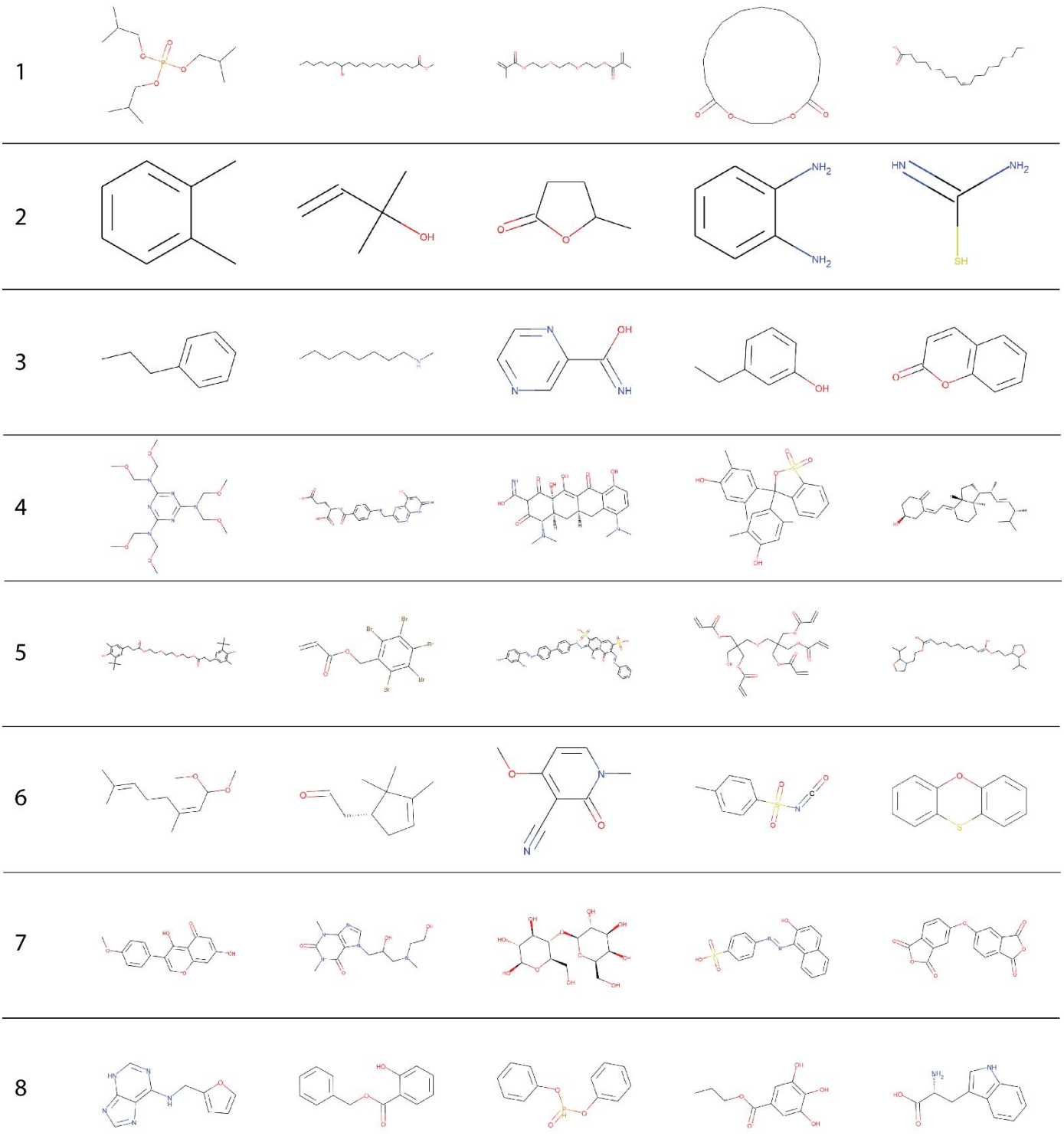

Figure S8. Five randomly chosen compounds in each of the 8 DarkChem clusters.

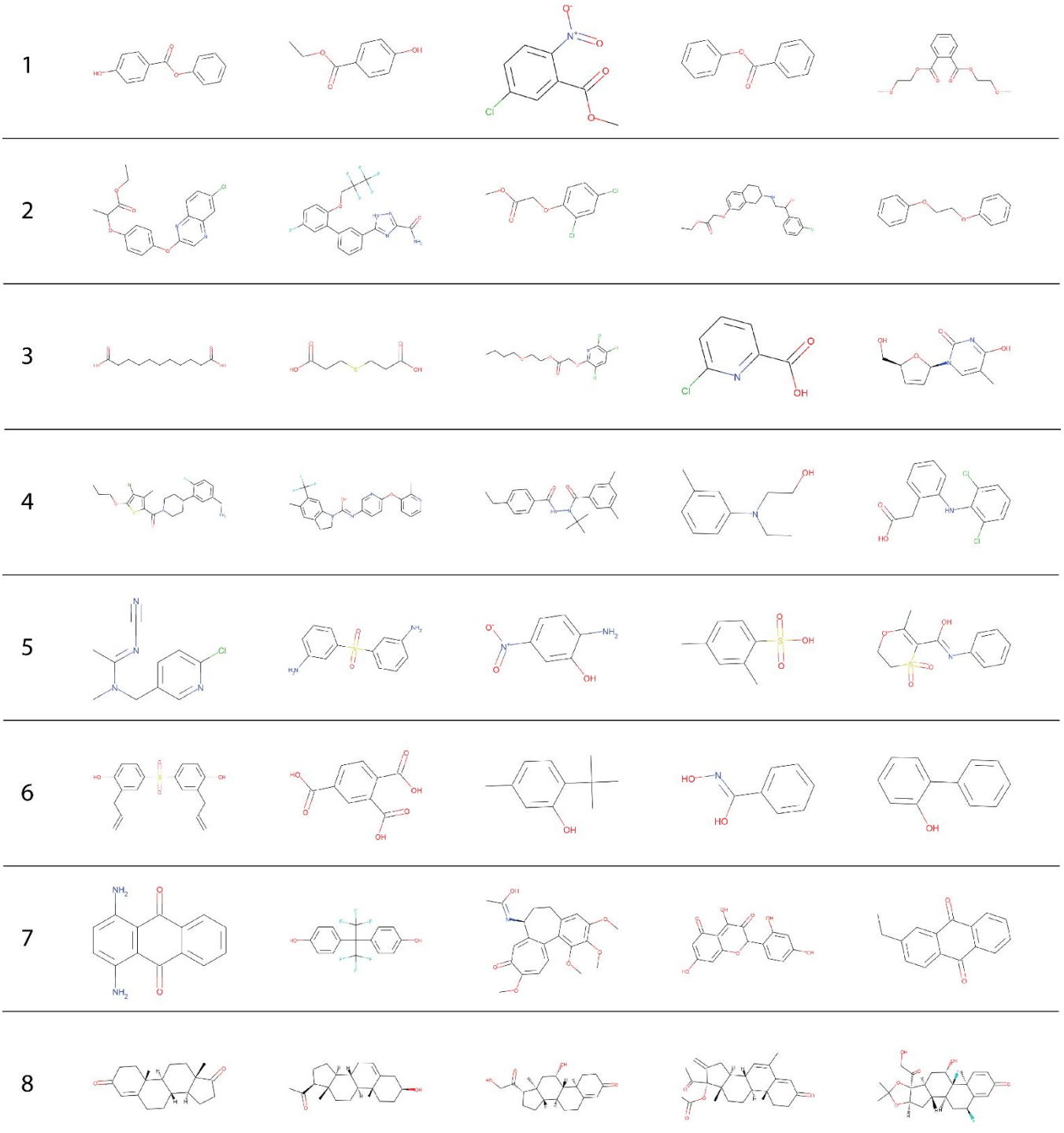

Figure S9. Five randomly chosen compounds in each of the 8 MACCS substructure clusters.

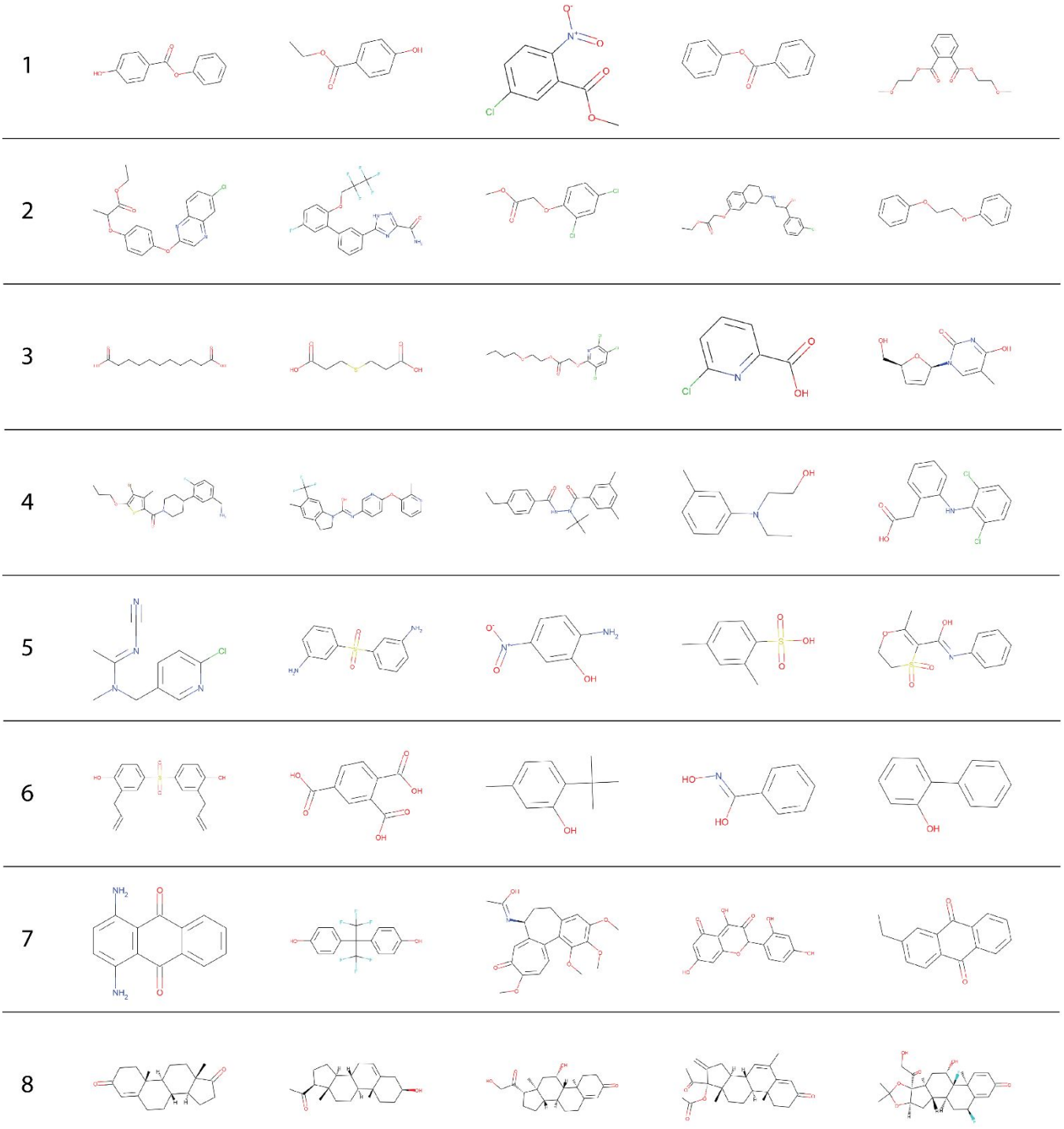

Figure S10. Five randomly chosen compounds in each of the 8 SPECTRe substructure clusters.

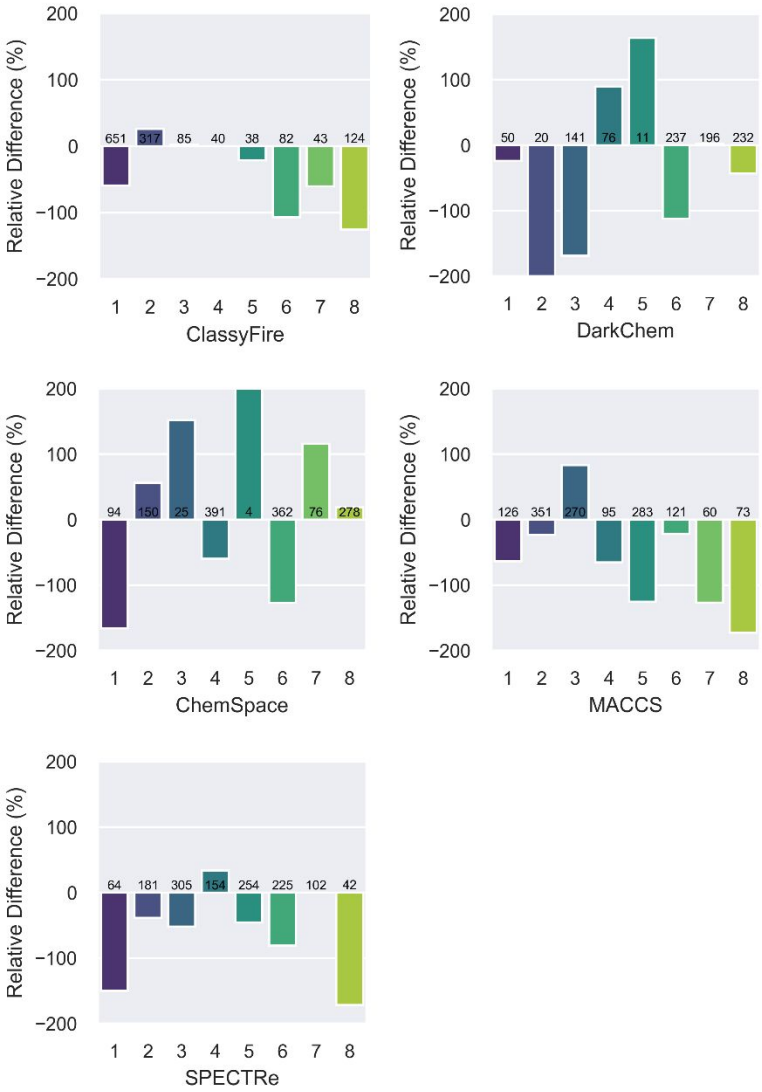

Figure S11. Relative difference of observed and non-observed compounds in each group/cluster, calculated as $\frac{\text{observed} - \text{not observed}}{\text{observed} + \text{not observed}} * 200$. The number of members in each cluster is shown for each bar.

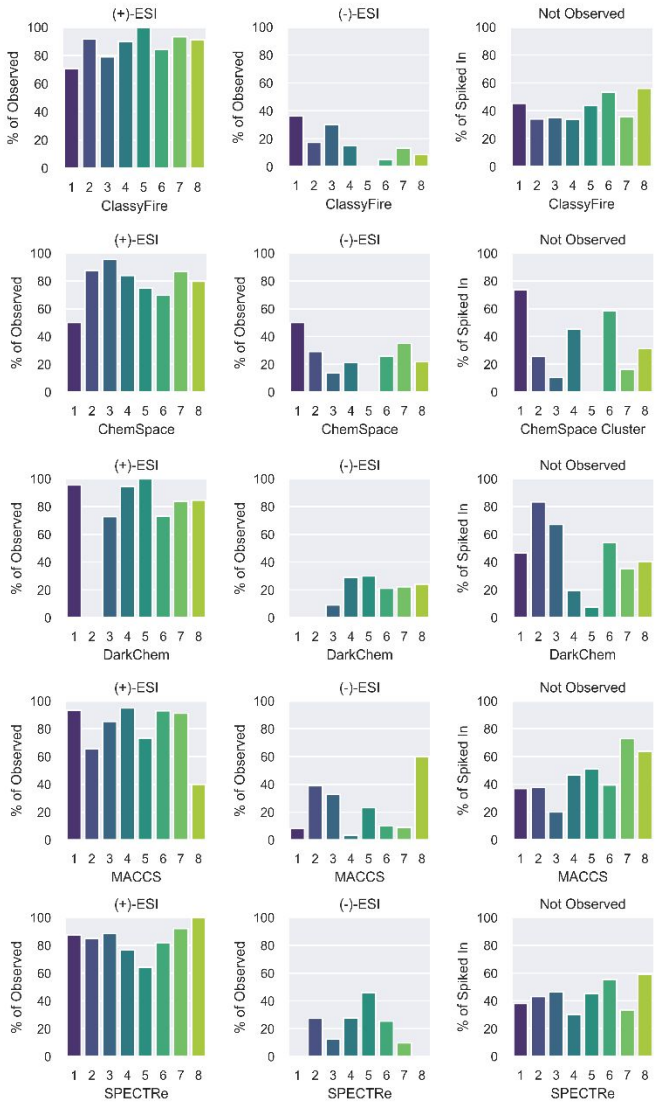

Figure S12. Distribution of compounds observed using positive vs. negative mode ESI, divided by the number of total compounds observed, and the distribution of those that were not observed in any sample.

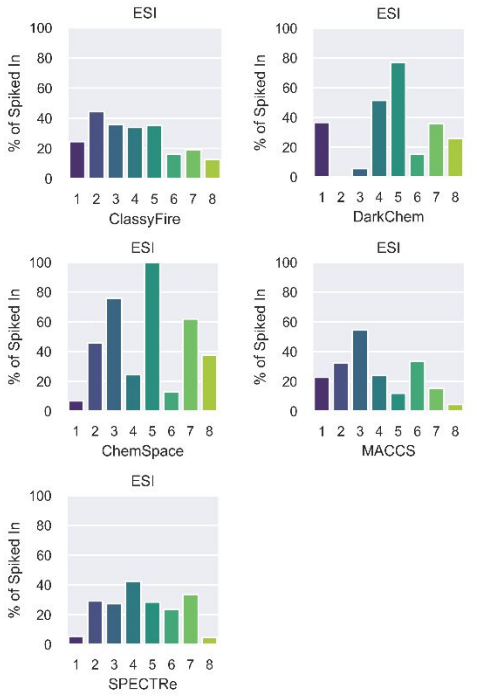

Figure S13. Distribution of compounds observed using ESI (positive or negative mode).

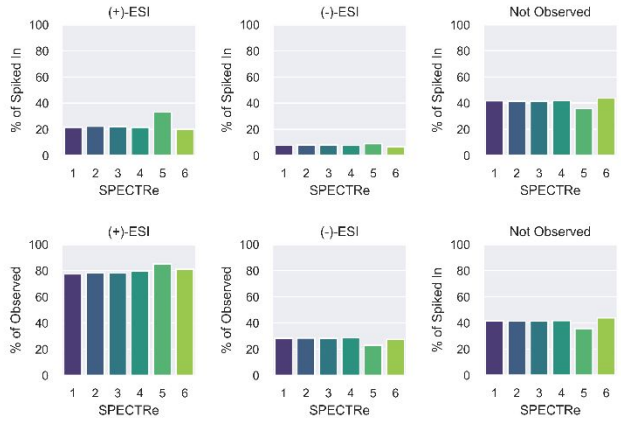

Figure S14. Distribution of spiked in compounds observed using positive vs. negative mode ESI, and the distribution of those that were not observed in any sample for the 6 maximum common substructures (i.e. compounds containing 1-6 of the most common substructures that do not contain one another) found using SPECTRe. These substructures (Cc1ccccc1, CC=CC=CC, CCC=CC=C, CC=C(C)C, CCN, CCO) were found using SPECTRe, and compounds that had these substructures were placed into six groups. On average, the size of these groups is 700.5 (σ: 57.9), with the largest group (SPECTRe Substructures Group 6) covering 828 compounds and the smallest cluster (SPECTRe Substructures Group 1) covering 657 compounds. A total of 1,188 compounds (all from the list of spiked in compounds) were considered for these groups, due to time and memory requirements. Of these, 1,149 contained at least 1 of the 6 maximum common substructures. Compounds do have the opportunity to land in more than one group for this method, leading to a total of 4,206 compound to substructure pairings, where 292, 326, 48, 220, 2090, and 1230 compounds were seen 1-6 times, respectively.

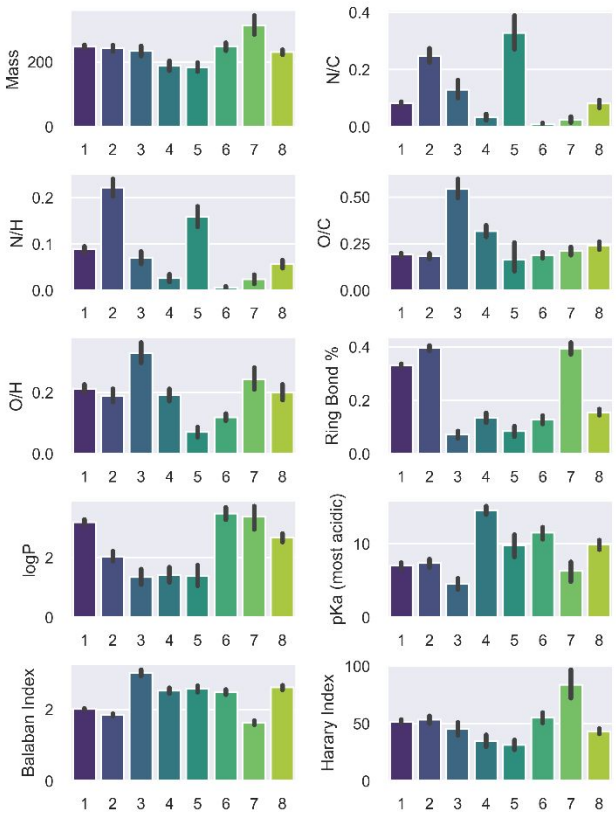

Figure S15. Average properties of compounds in ClassyFire Superclass groups, given the full suspect library. Error bars represent the standard deviation.

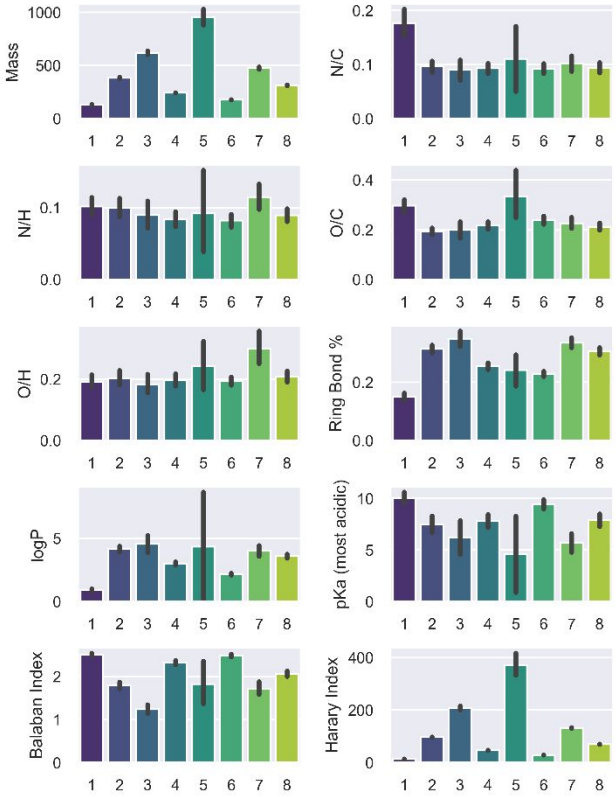

Figure S16. Average properties of compounds in Chemical Space clusters, given the full suspect library. Error bars represent the standard deviation.

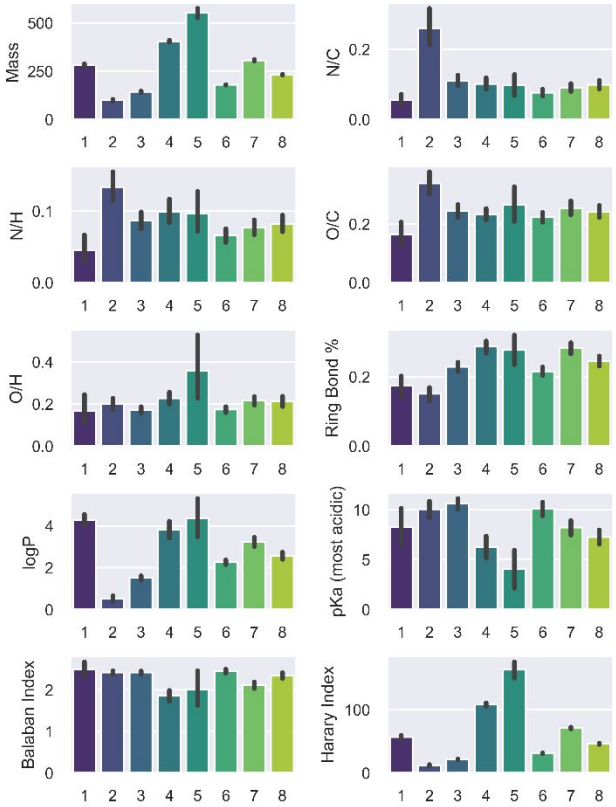

Figure S17. Average properties of compounds in DarkChem clusters, given the full suspect library. Error bars represent the standard deviation.

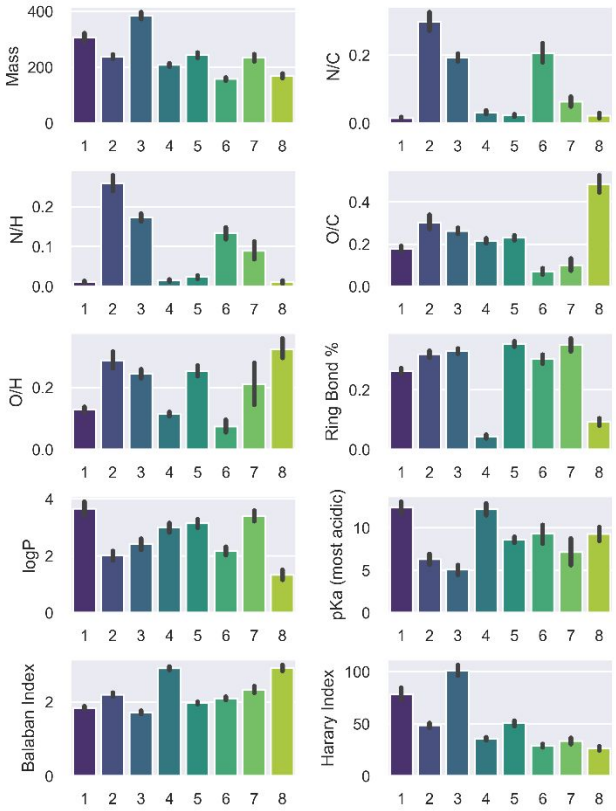

Figure S18. Average properties of compounds in the MACCS substructure clusters, given the full suspect library. Error bars represent the standard deviation.

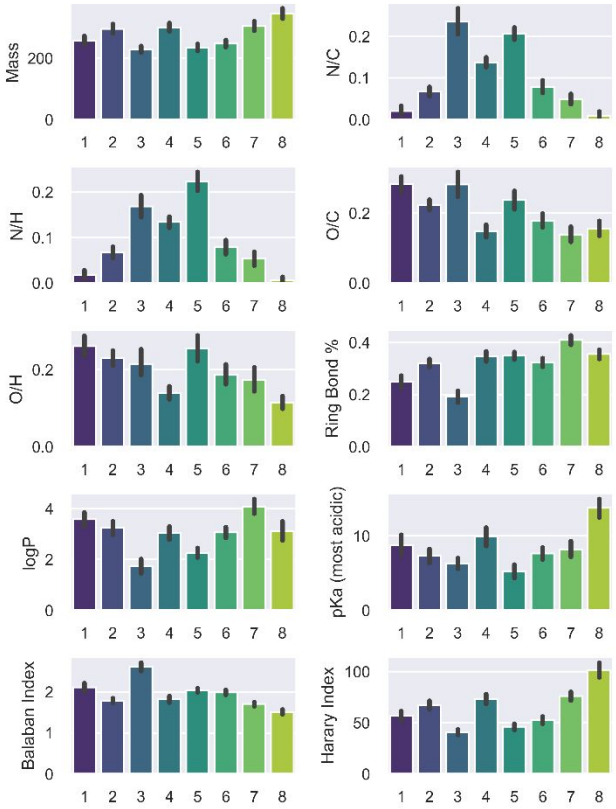

Figure S19. Average properties of compounds in the SPECTRe substructure clusters, given the full suspect library. Error bars represent the standard deviation.

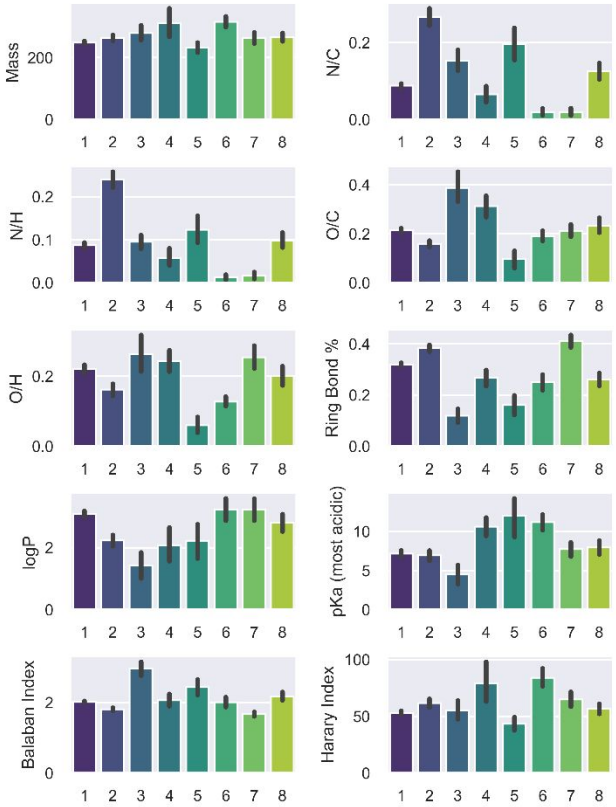

Figure S20. Average properties of compounds in ClassyFire Superclass groups, considering only spiked in compounds. Error bars represent the standard deviation.

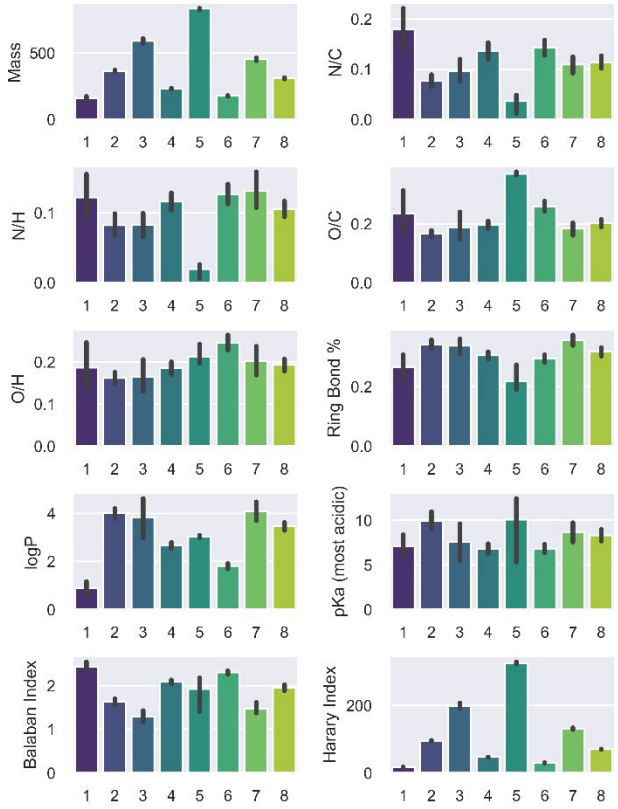

Figure S21. Average properties of compounds in Chemical Space clusters, considering only spiked in compounds. Error bars represent the standard deviation.

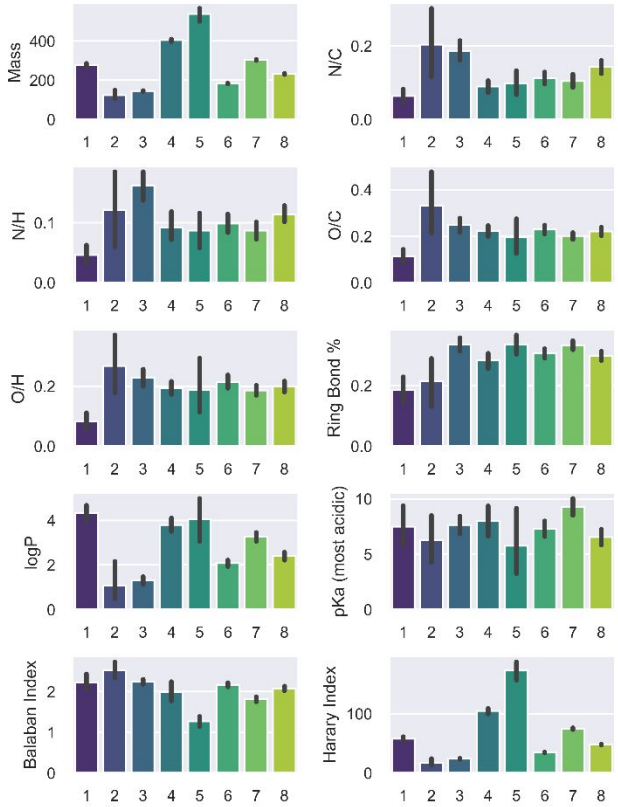

Figure S22. Average properties of compounds in DarkChem clusters, considering only spiked in compounds. Error bars represent the standard deviation.

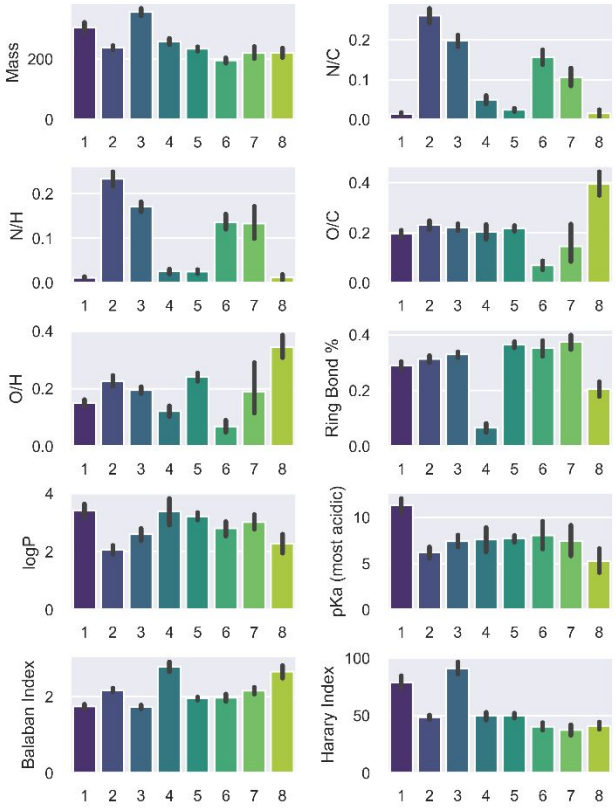

Figure S23. Average properties of compounds in the MACCS substructure clusters, considering only spiked in compounds. Error bars represent the standard deviation.

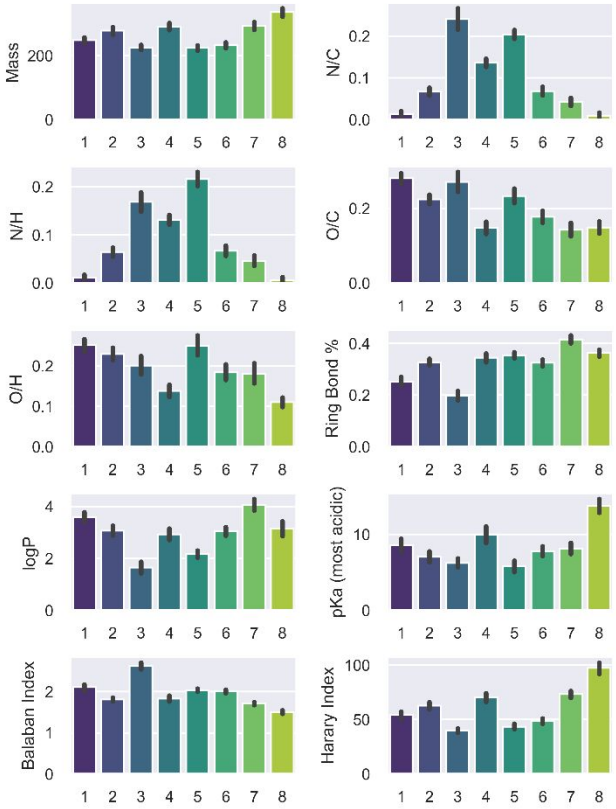

Figure S24. Average properties of compounds in the SPECTRe substructure clusters, considering only spiked in compounds. Error bars represent the standard deviation.

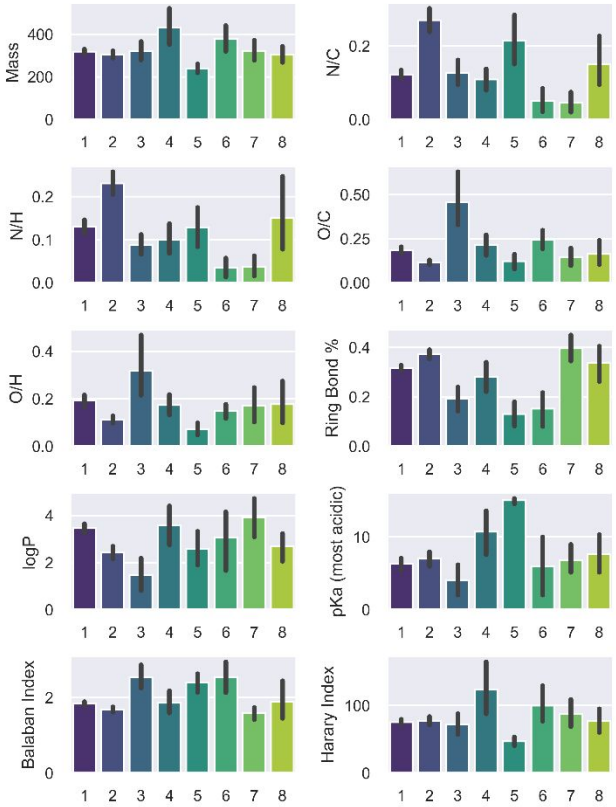

Figure S25. Average properties of compounds in ClassyFire Superclass groups, considering only observed compounds. Error bars represent the standard deviation.

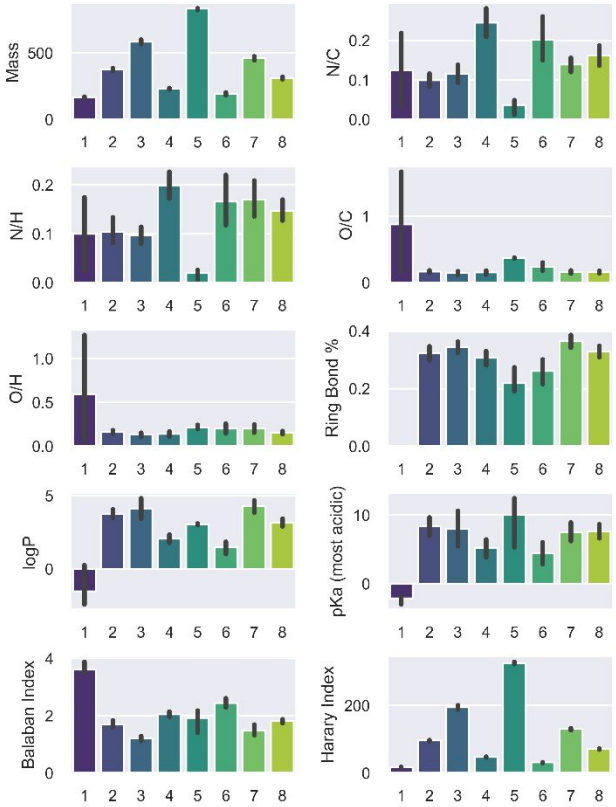

Figure S26. Average properties of compounds in Chemical Space clusters, considering only observed compounds. Error bars represent the standard deviation.

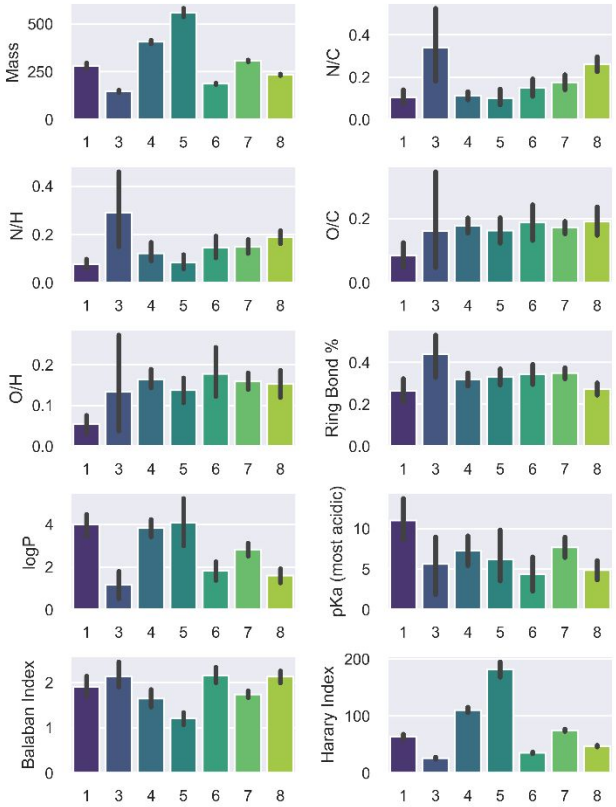

Figure S27. Average properties of compounds in DarkChem clusters, considering only observed compounds. Error bars represent the standard deviation.

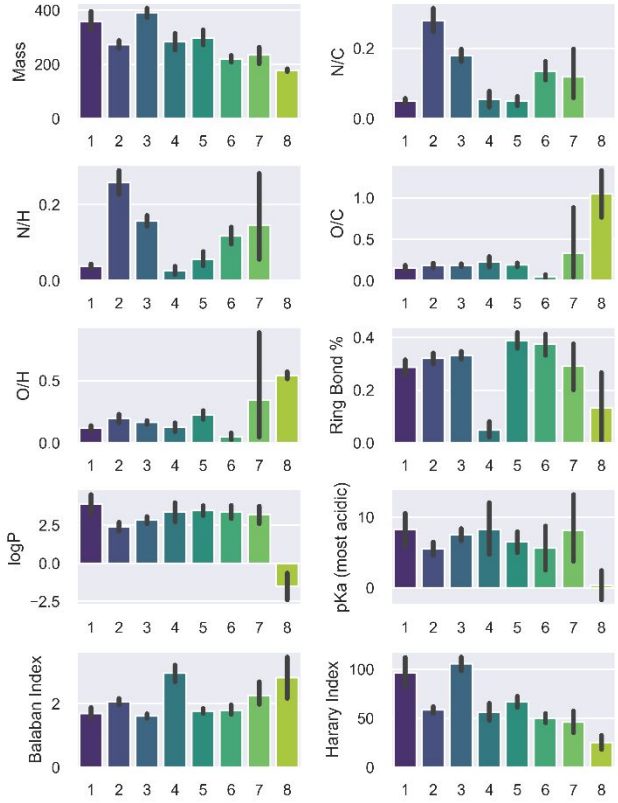

Figure S28. Average properties of compounds in the MACCS substructure clusters, considering only observed compounds. Error bars represent the standard deviation.

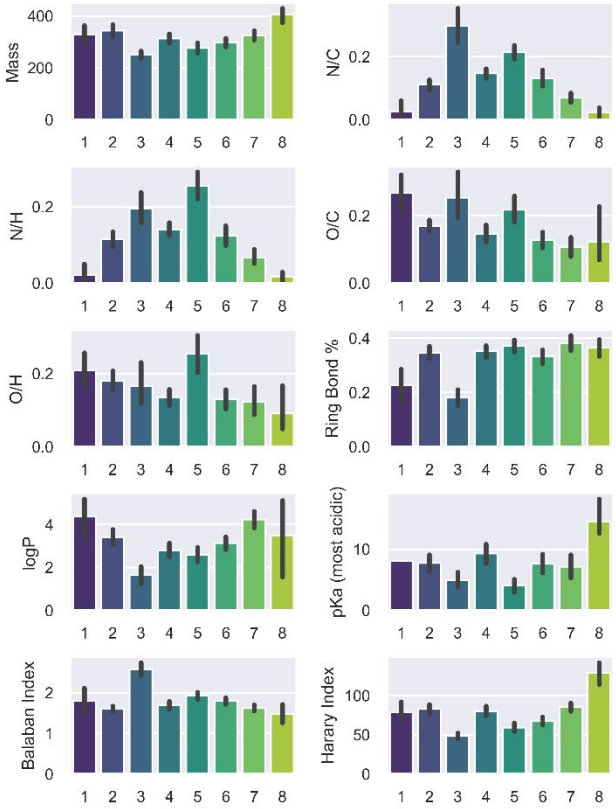

Figure S29. Average properties of compounds in the SPECTRe substructure clusters, considering only observed compounds. Error bars represent the standard deviation.

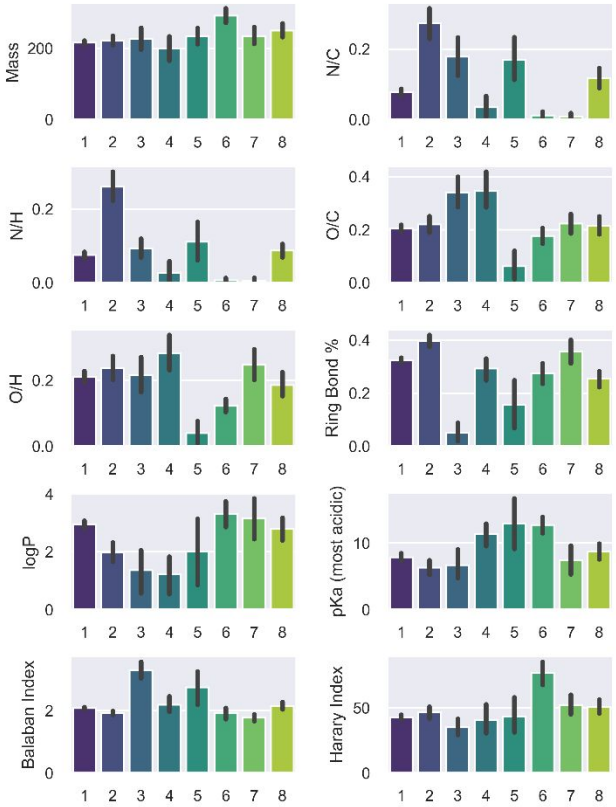

Figure S30. Average properties of compounds in ClassyFire Superclass groups, considering only not observed compounds. Error bars represent the standard deviation.

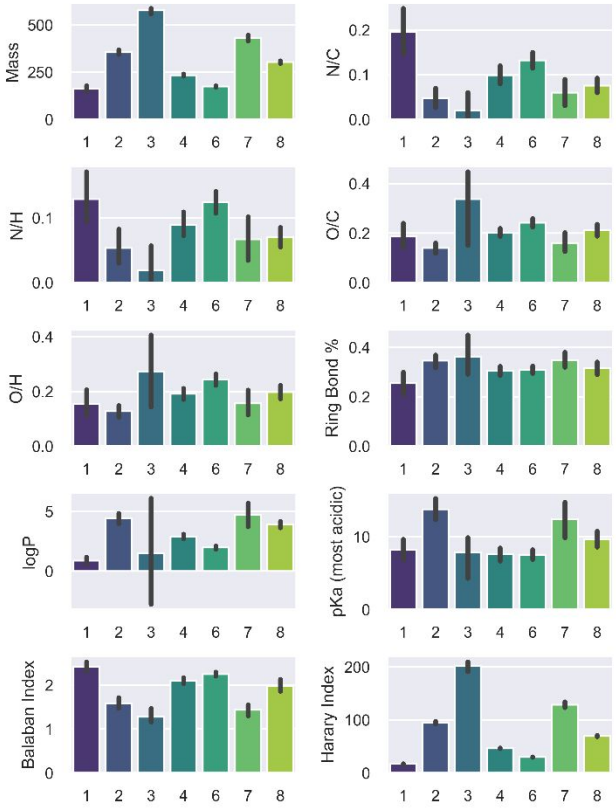

Figure S31. Average properties of compounds in Chemical Space clusters, considering only not observed compounds. Error bars represent the standard deviation.

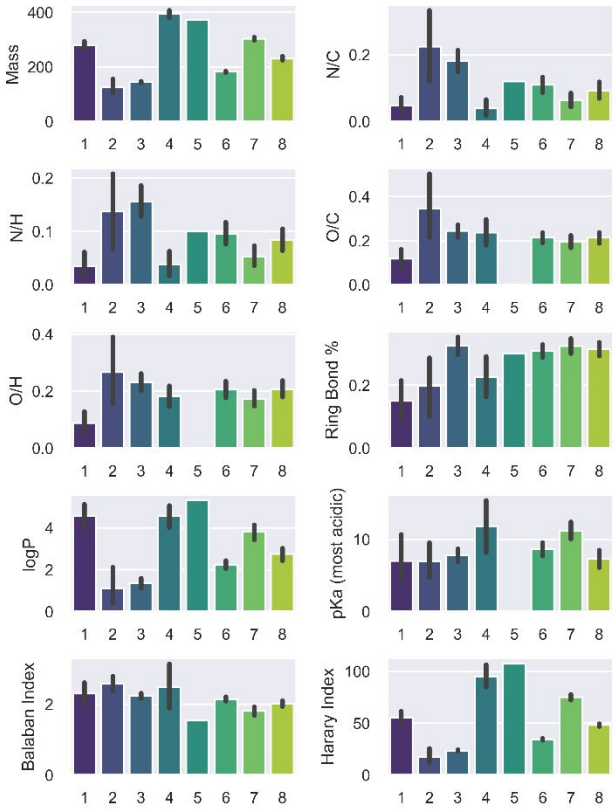

Figure S32. Average properties of compounds in DarkChem clusters, considering only not observed compounds. Error bars represent the standard deviation.

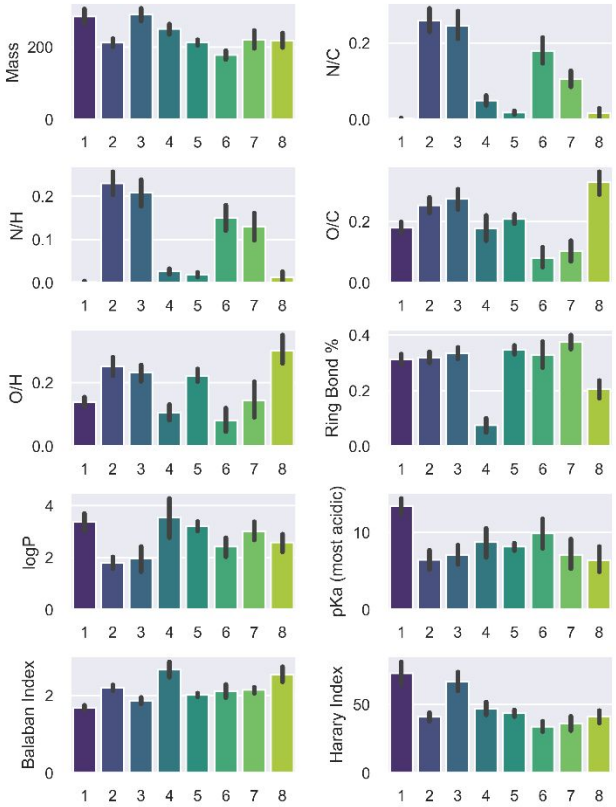

Figure S33. Average properties of compounds in the MACCS substructure clusters, considering only not observed compounds. Error bars represent the standard deviation.

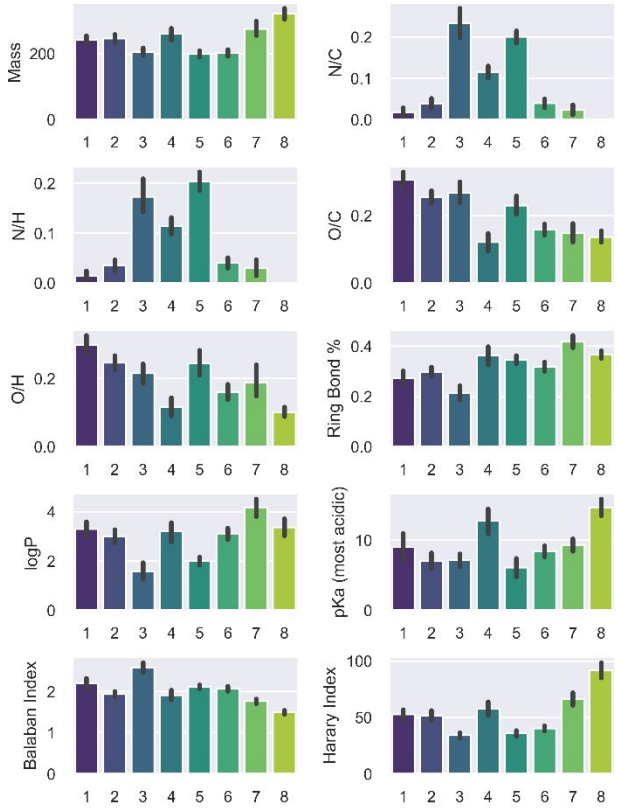

Figure S34. Average properties of compounds in the SPECTRe substructure clusters, considering only not observed compounds. Error bars represent the standard deviation.

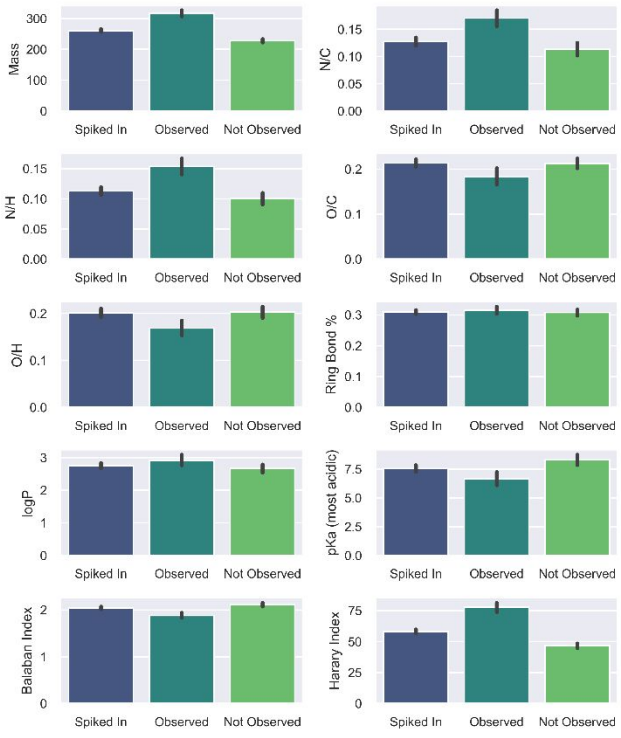

Figure S35. Average properties of compounds when not considering the groups/clusters. Error bars represent the standard deviation.

References

1. Nuñez, J. R.; Colby, S. M.; Thomas, D. G.; Tfaily, M. M.; Tolic, N.; Ulrich, E. M.; Sobus, J. R.; Metz, T. O.; Teeguarden, J. G.; Renslow, R. S., Evaluation of in silico multifeature libraries for providing evidence for the presence of small molecules in synthetic blinded samples. *J. Chem. Inf. Model.* **2019,** *59*, 4052-4060.