Supporting Information for

Leveraging natural language processing to curate the tmCAT, tmPHOTO, tmBIO, and tmSCO datasets of functional transition metal complexes

Ilia Kevlishvili[1], Roland G. St. Michel[1,2], Aaron G. Garrison[1], Jacob W. Toney[1], Husain Adamji[1], Haojun Jia[1,3], Yuriy Román-Leshkov[1,3], and Heather J. Kulik[1,3]*

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[3]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*email: hjkulik@mit.edu

Contents

Table S1. Validation of abstracts and titles obtained from HTML. 100 random samples were inspected manually. Correct refers to cases where relevant text was fully retrieved. Partial refers to cases where text was partially retrieved (1st paragraph of multi-paragraph abstract), incorrect refers to cases where improper text was retrieved (i.e. 1st paragraph of introduction), missing refers to cases when either abstract/title was missing or irrelevant text was (i.e journal name, cookies statement) retrieved. CSV file containing this analysis can be found in the Supplementary Information file.

	Correct	Partial	Incorrect	Missing
Abstract	74	8	1	17
Title	95	0	0	5

Table S2. The five largest clusters of the BERTopic model using text without the introduction of stop words. Key tokens represent the three most common tokens as identified using c-TF-IDF vector. Unsupervised clustering without stop words leads to the formation of transition-metal-based clusters.

Cluster size	Key tokens
1981	Ruthenium complexes, RuCl, cymene
1677	Nickel complexes, Ni(ii), complexes nickel
1646	Palladium complexes, Suzuki Miyaura, Miyaura
1463	Tungsten, Molybdenum, Molybdenum (vi)
1247	Iron complexes, fe(ii), crossover

Table S3. All keywords used to screen for catalysis manuscripts in the title. Keywords screened in the title and associated number of manuscripts retained. Screening was done on a corpus consisting of 28,394 manuscripts.

Keyword	Number of manuscripts
Includes “catal”	4,612
Does not include “uncatal”	4,610
Does not include “acid-catal” or “acid catal”	4,589
Does not include “base-catal” or “base catal”	4,585

Table S4. All keywords used to screen for non-catalysis manuscripts in the abstract and title. Keywords screened in the title and associated number of manuscripts retained. Screening was done on a corpus consisting of 28,394 manuscripts.

Keyword	Number of manuscripts
Does not include “catal”	21,410
Does not include “turnover”	21,398
Does not include “polymer”	20,557

Table S5. Additional stopwords included in the TF-IDF vectorizer for training models without direct catalysis keywords.

Keyword
catalyzed
catalytic
catalyst
catalysts
catalysis
catalyze
catalyse
catalysed

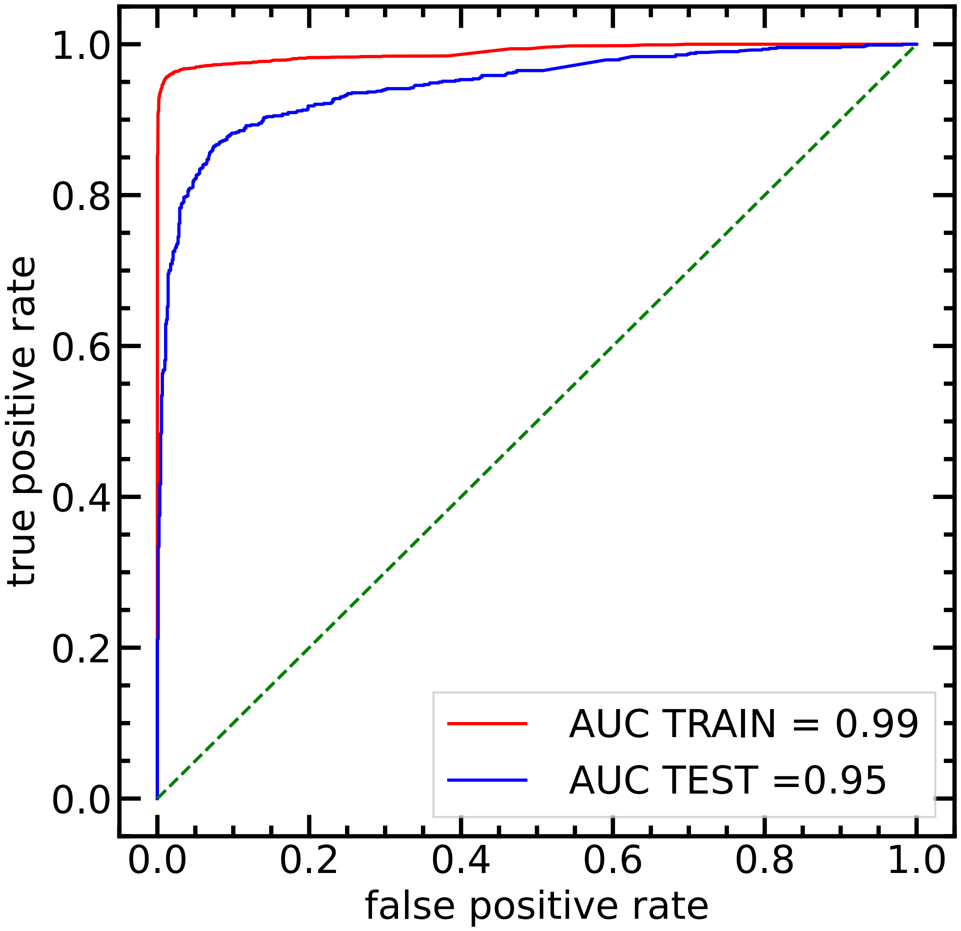


Figure S1. Receiver operating characteristic curve of the secondary catalysis classifier of the training (red, 7,336 datapoints) and test (blue, 1,834 datapoints) sets. The secondary classifier consists of the same training and test set, but the TF-IDF feature vector was constructed after removing direct catalysis keywords. A naïve model is shown as a green dashed line. Areas under the curve are shown in the inset legend.

Table S6. The performance metrics of the random forest catalysis classifier where the feature vector does not include any catalysis keywords shown in Table S3.

	Train	**Test**
Accuracy	0.97	0.89
Precision	0.97	0.89
Recall	0.97	0.89

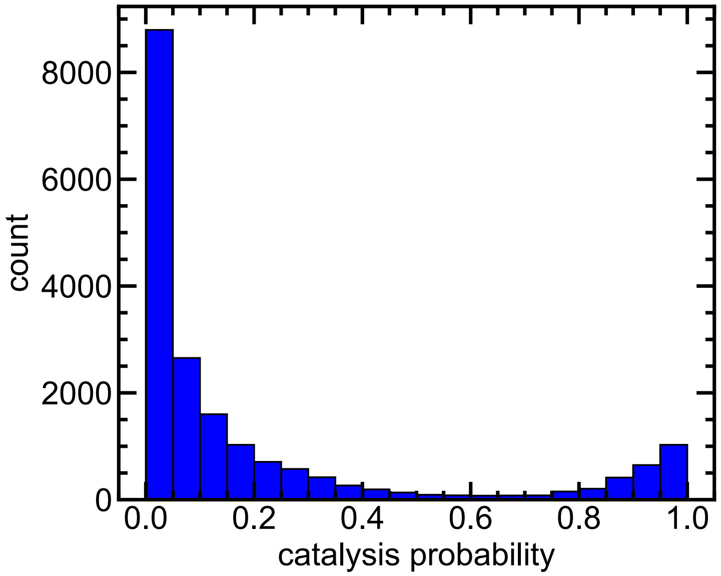


Figure S2. Distribution of the predicted probabilities of the manuscript related to catalysis in the unlabeled set in the corpus. Probabilities were calculated using the best-performing random forest classifier that includes catalysis keywords in the feature vector. The unlabeled set consists of 19,224 manuscripts.

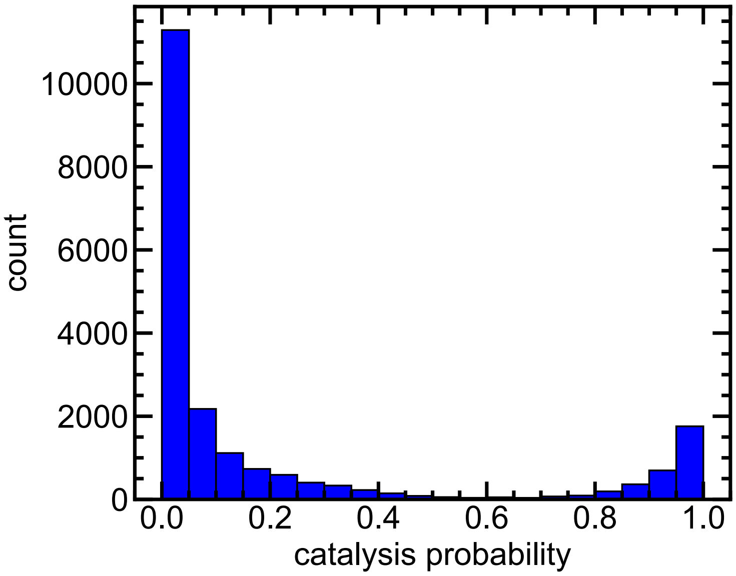


Figure S3. Distribution of the predicted probabilities of the manuscript related to catalysis in the HTML-mined corpus. Probabilities were calculated using the best-performing random forest classifier that includes catalysis keywords in the feature vector. The HTML-mined corpus consists of manuscripts that could not be obtained using the ArticleDownloader package and were directly scraped from the manuscript webpage. The corpus consists of 20,449 manuscript abstracts and titles.

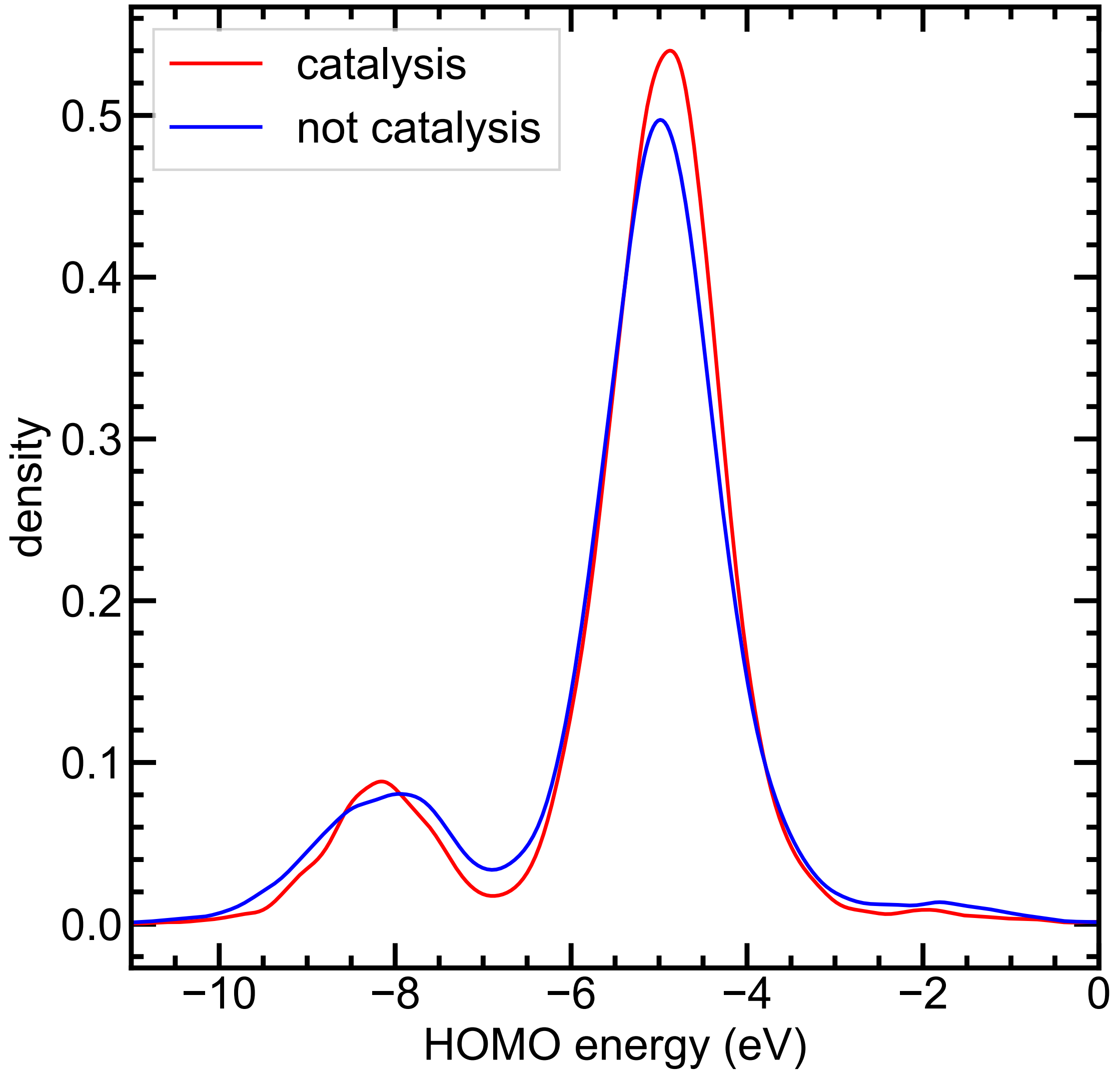


Figure S4. Density distribution of HOMO orbital energies (in eV) in the catalysis (tmCAT) and "not-catalysis" subsets. All orbital energies are obtained from the tmQM dataset. The total density distribution is computed using a bandwidth of 0.1.

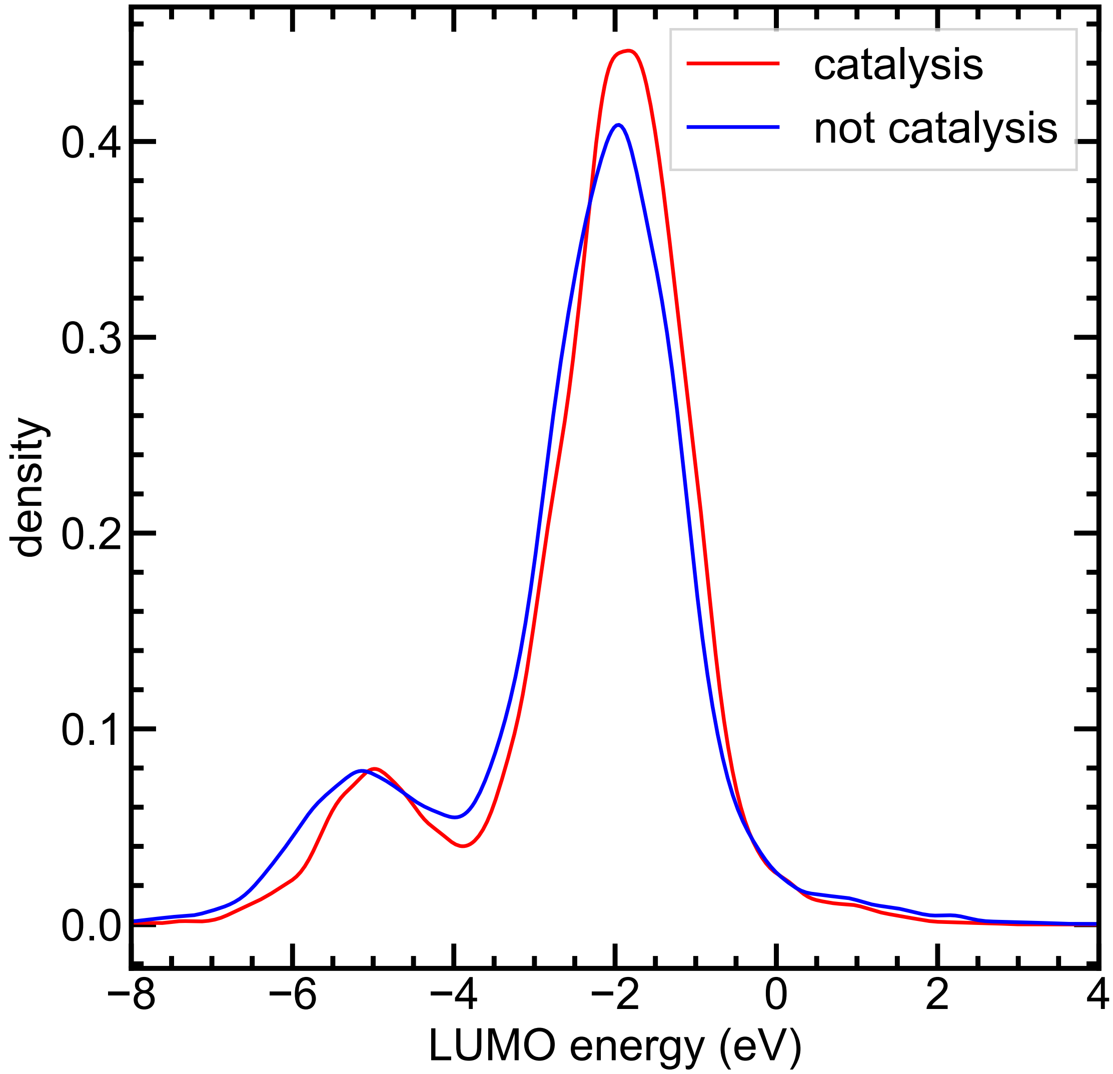


Figure S5. Density distribution of LUMO orbital energies (in eV) in the catalysis (tmCAT) and "not catalysis" sets. All orbital energies are obtained from the tmQM dataset. The total density distribution is computed using a bandwidth of 0.3.

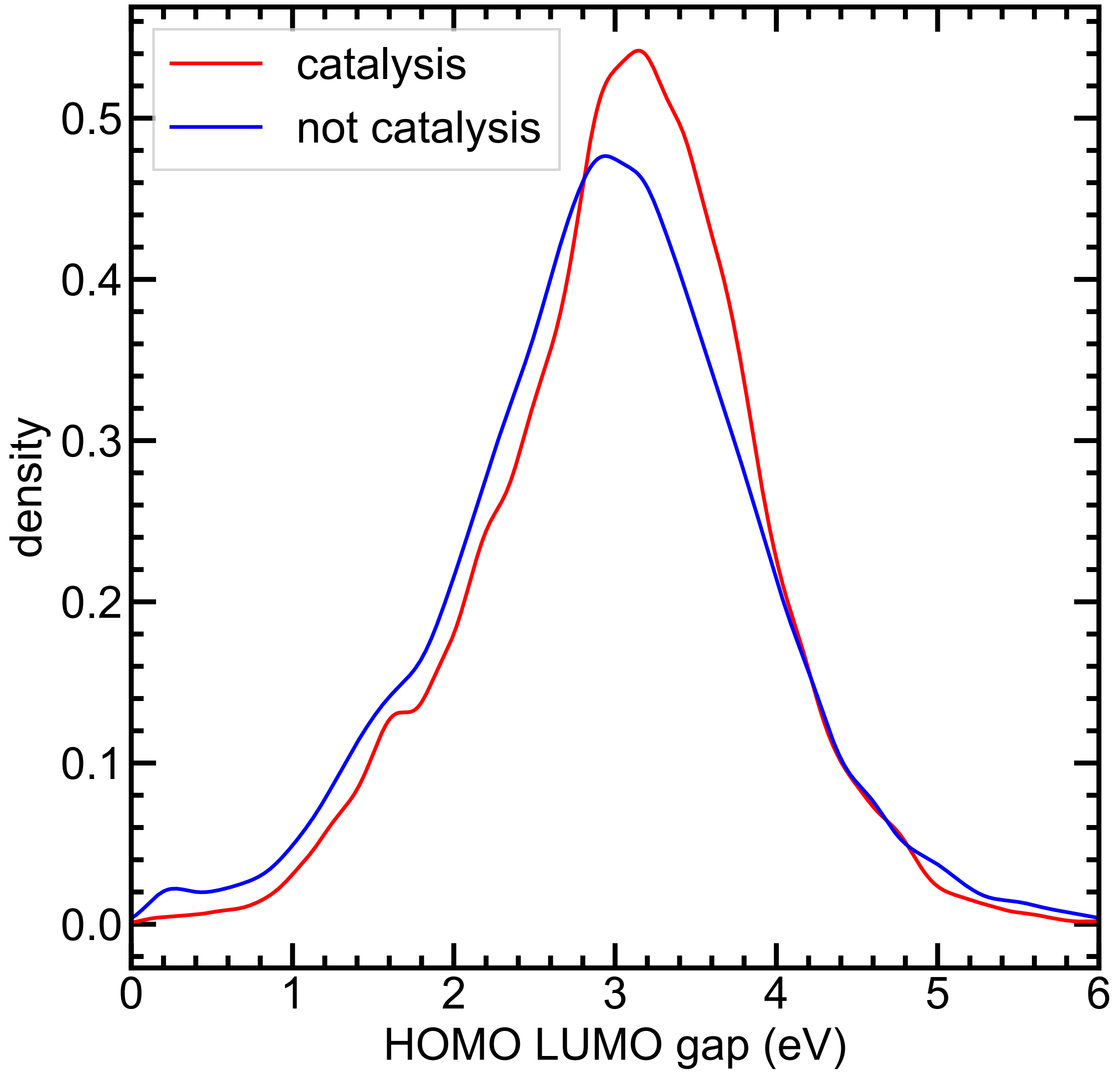


Figure S6. Density distribution of the HOMO-LUMO gap energies (in eV) in the catalysis (tmCAT) and "not catalysis" subsets. All orbital energies are obtained from the tmQM dataset. Densities are computed using the bandwidth of 0.3.

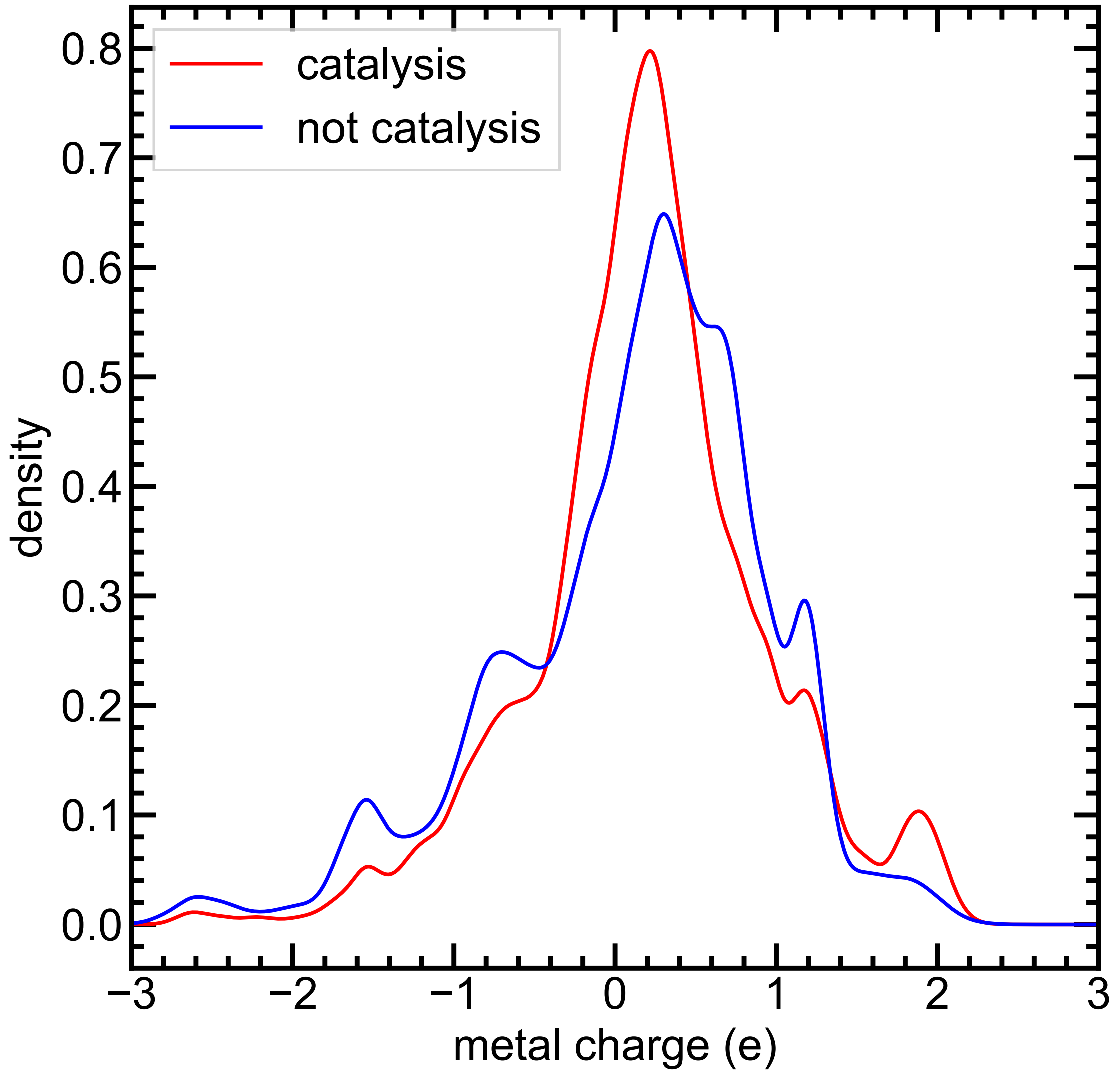


Figure S7. Density distribution of the metal charge (in e) in the catalysis (tmCAT) and "not catalysis" subsets. All orbital energies are obtained from the tmQM dataset. Densities are computed using the bandwidth of 0.1.

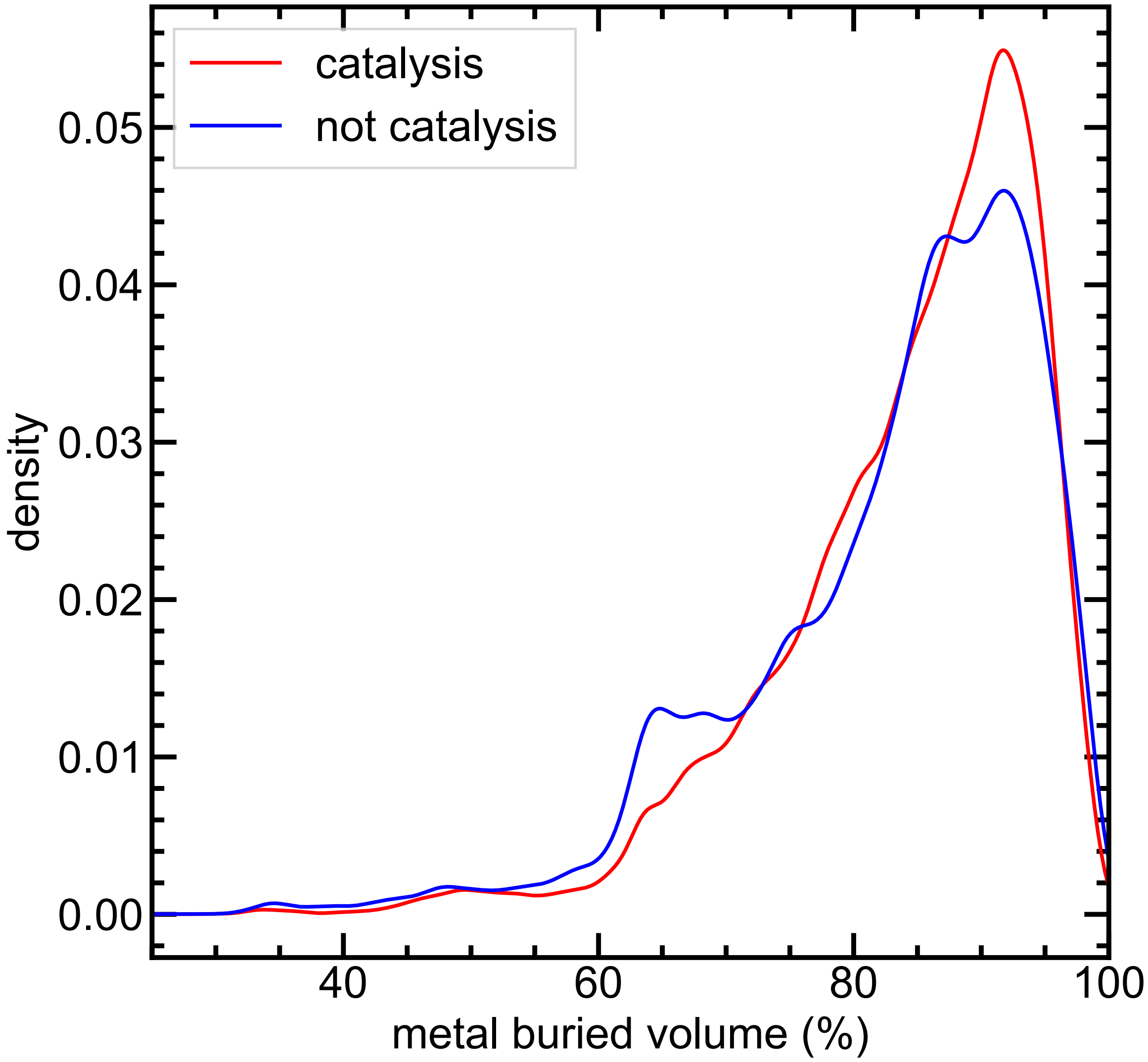


Figure S8. Density distribution of the metal buried volume in the catalysis (tmCAT) and "not catalysis" subsets. All values are computed using the unoptimized transition metal complex structures from the CSD. Densities are computed using the bandwidth of 0.1.

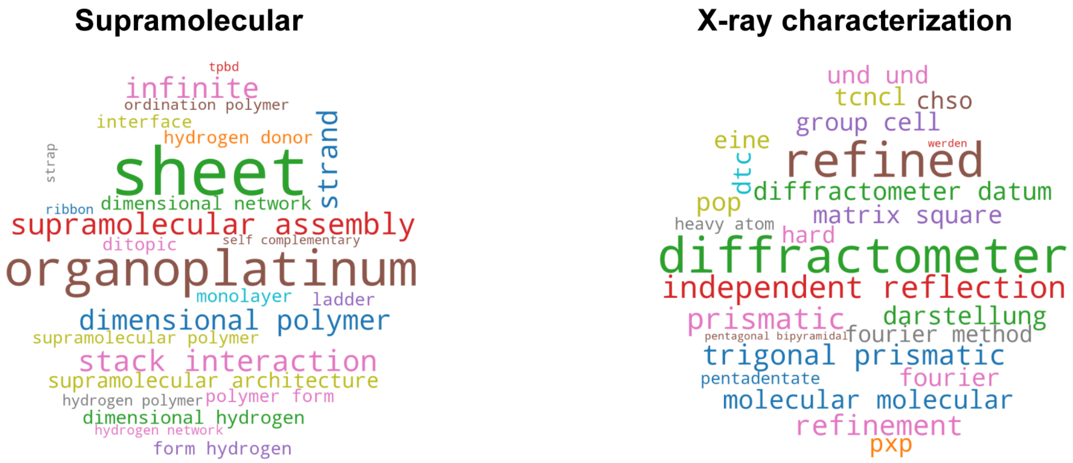


Figure S9. Wordclouds of two additional non-catalysis topics. Word clouds were generated using the 25 most important tokens in the c-TF-IDF vector and scaled based on the given token importance.

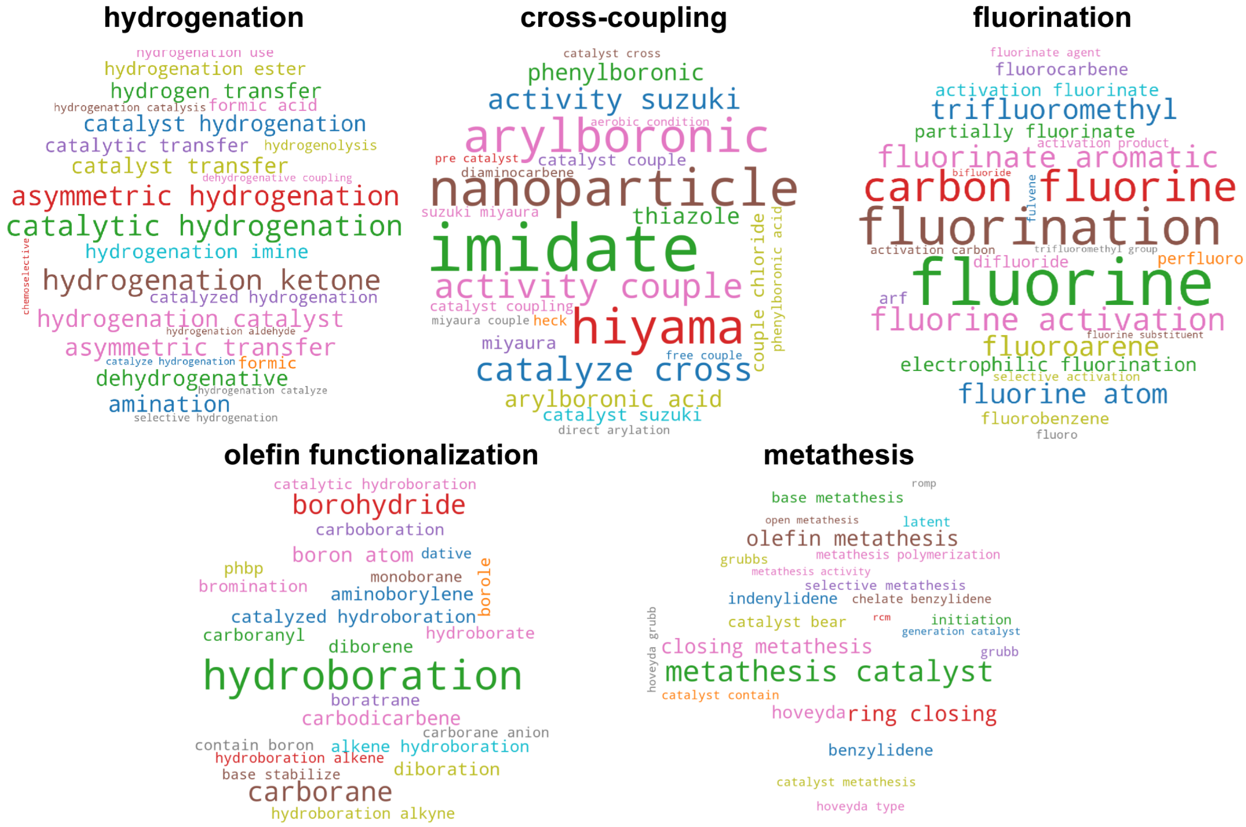


Figure S10. Wordclouds of five additional catalysis-related topics. Word clouds were generated using the 25 most important tokens in the c-TF-IDF vector and scaled based on the given token importance.

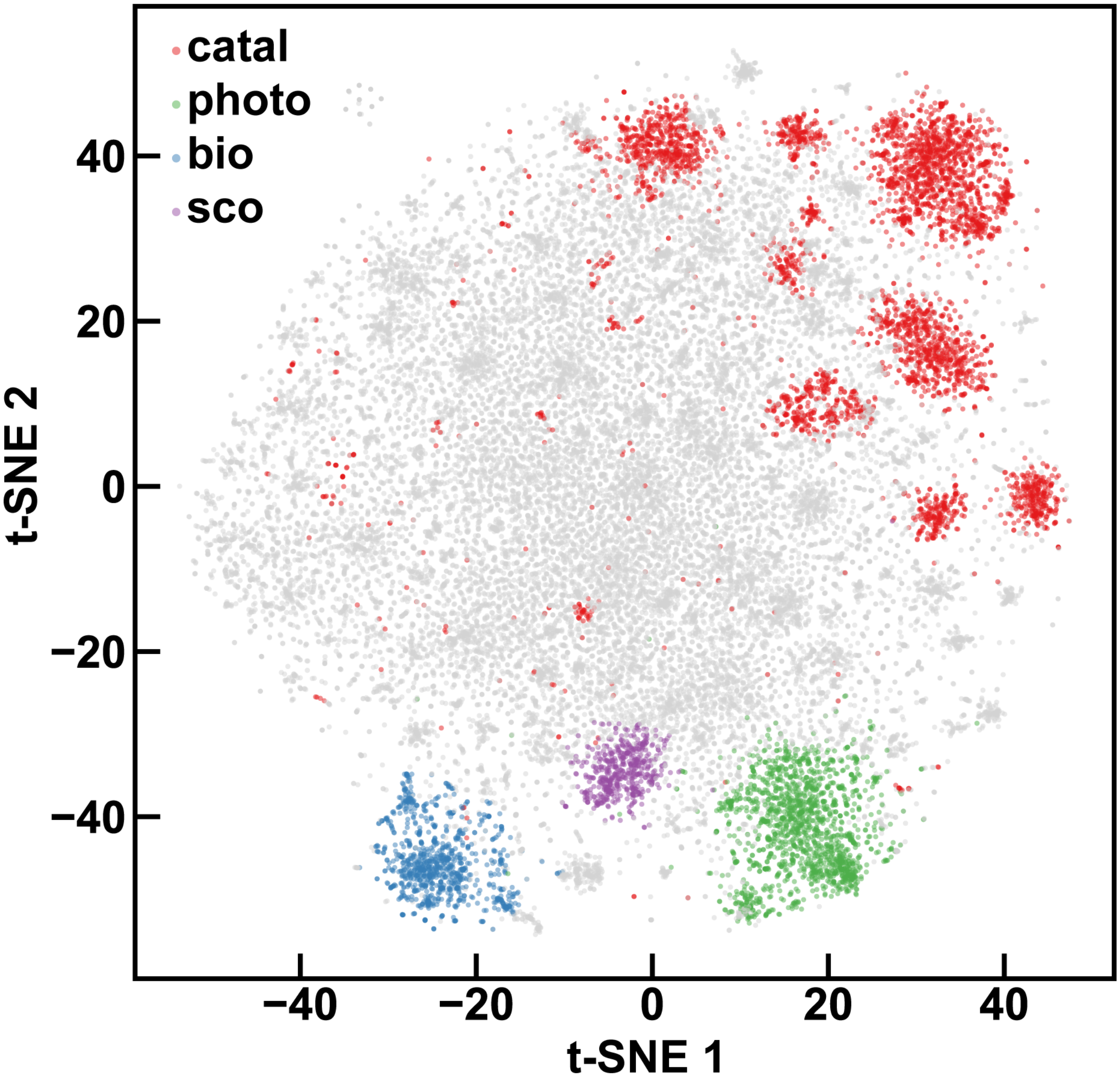


Figure S11. t-SNE embedding of SBERT embedding vectors colored by different cluster topics for different general applications. All catalysis related abstracts are shown in red, photoactivity related abstracts are shown in green, biological activity related abstracts are shown in blue, and magnetism related abstracts are shown in purple.

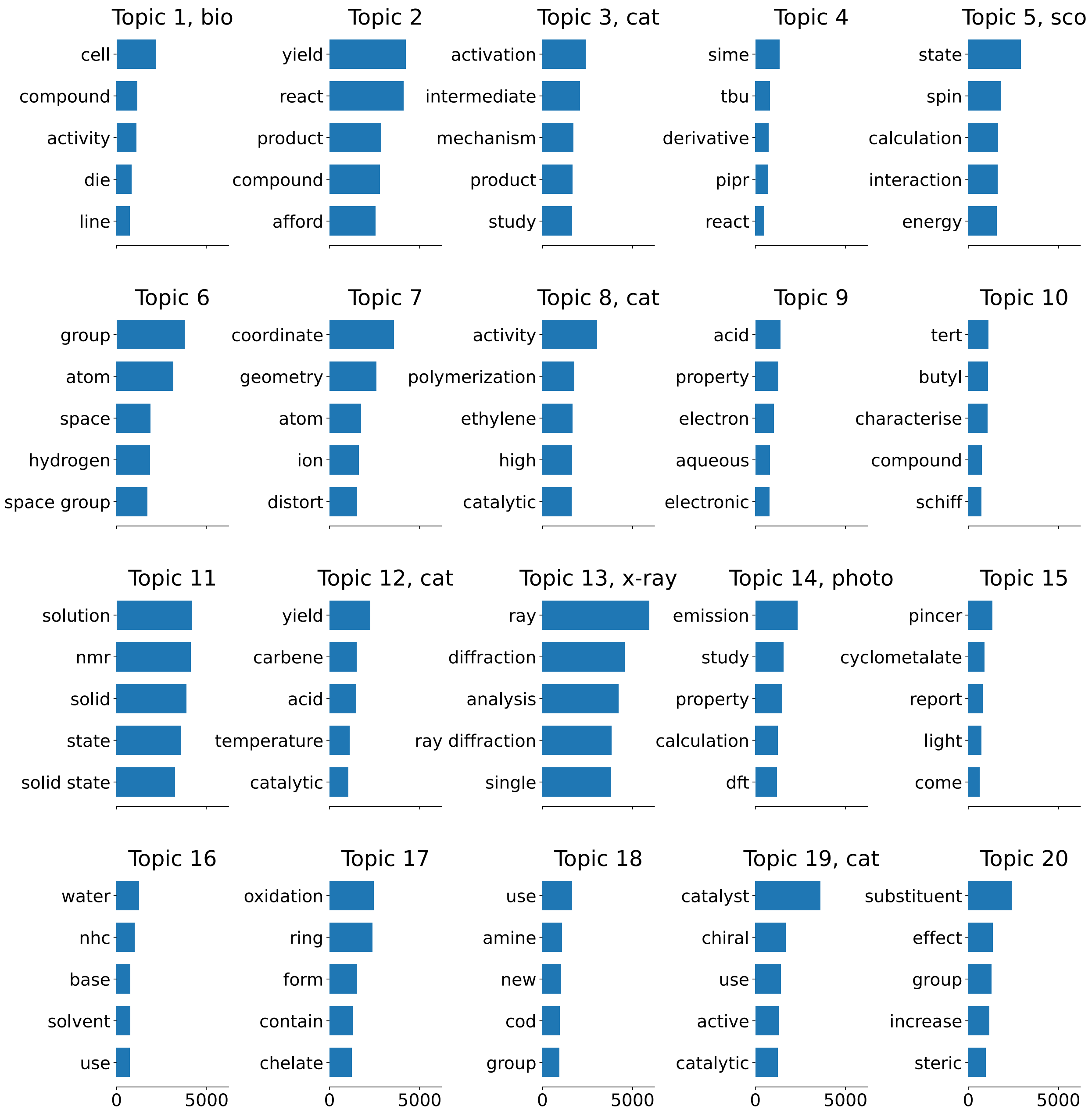


Figure S12. Topic breakdown using Latent Dirichlet Allocation (LDA) analysis and associated five most important tokens. Topics that are interpretable and related to topics also identified by BERTopic are labeled next to the topic index.

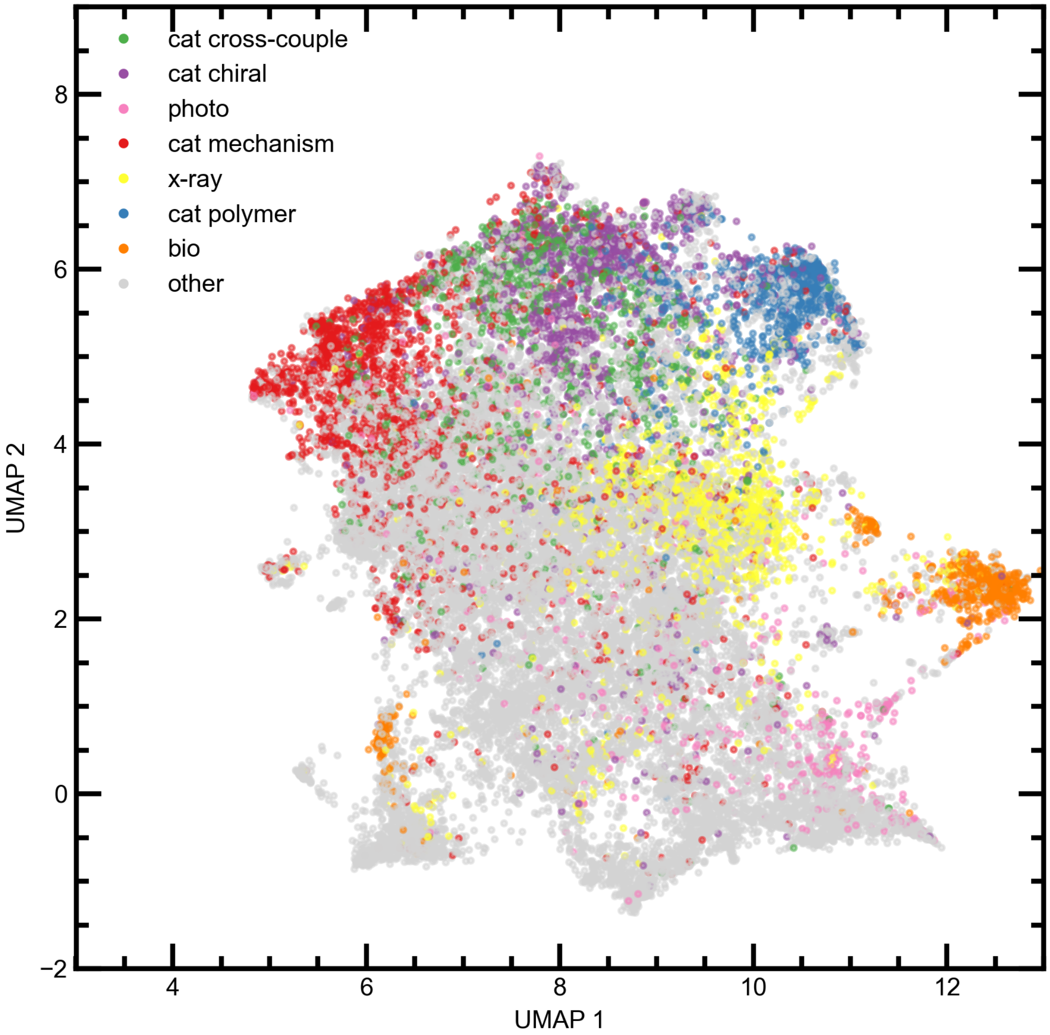


Figure S13. UMAP embedding of count vectorizer feature vector, which was used for training a latent Dirichlet allocation (LDA) model colored by different cluster topics identified from LDA clusters for catalysis subcategories, biological activity, photoactivity, and X-ray characterization.

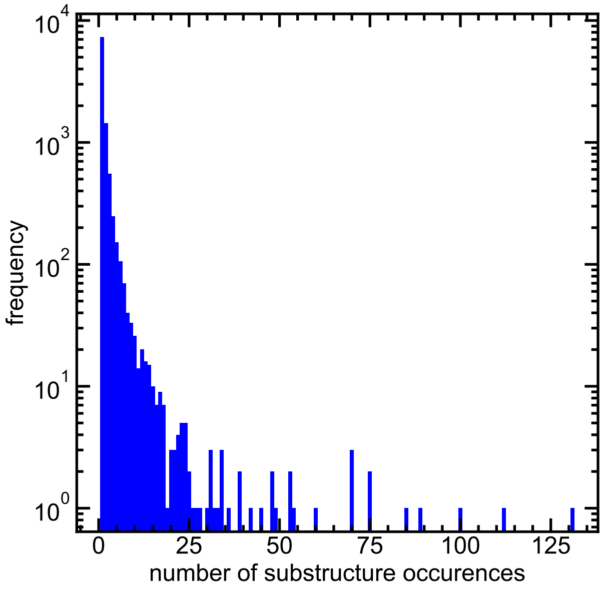


Figure S14. Frequency of molecular subgraph occurrence in the tmCAT dataset.

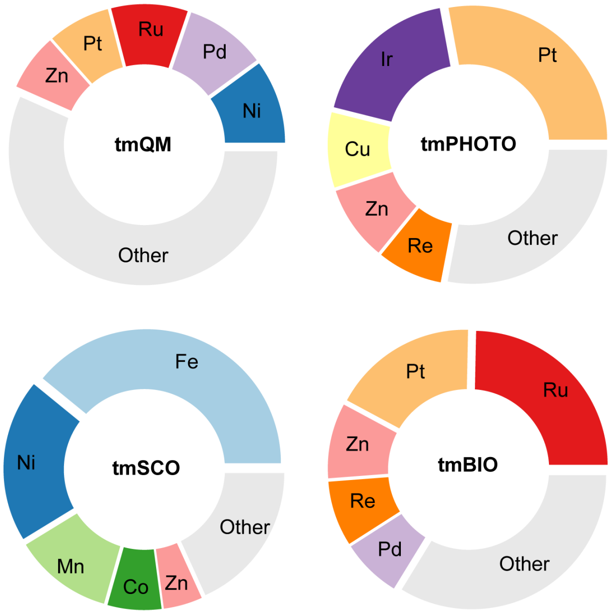


Figure S15. Frequency of the five most common transition metals in tmQM, tmPHOTO, tmSCO and tmBIO datasets.

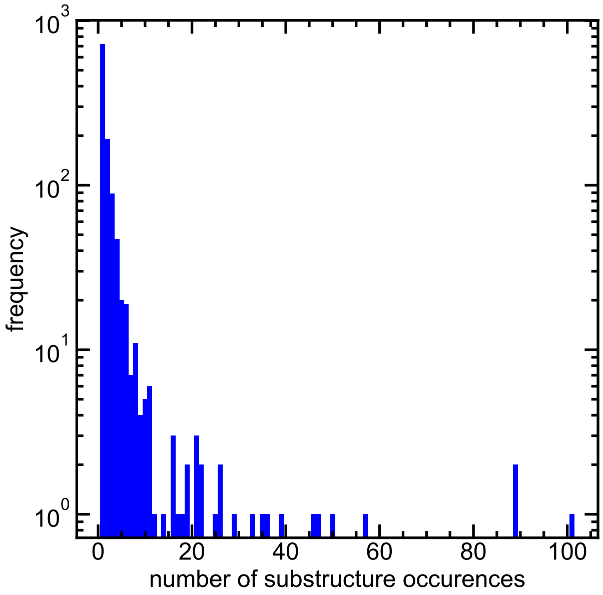
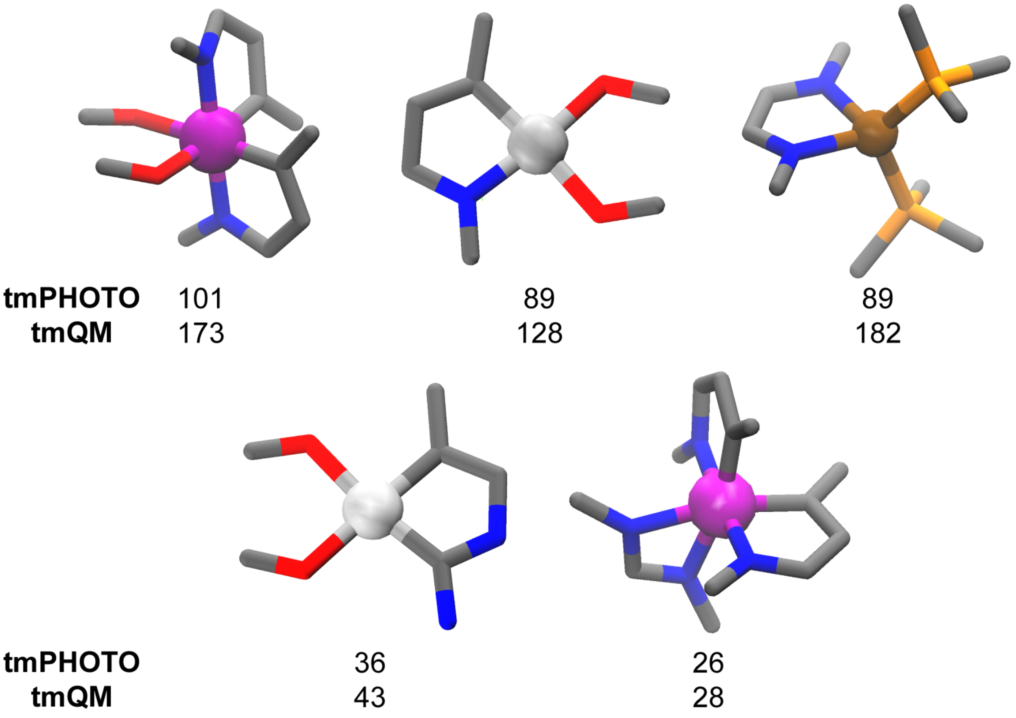


Figure S16. Frequency of molecular subgraph recurrence in the tmPHOTO dataset.



tmPHOTO	101	89	89
tmQM	173	128	182

tmPHOTO	36	26
tmQM	43	28

Figure S17. Representative substructures of the tmPHOTO set. The most common substructures are shown at the top. Structures that appear with high relative frequency in tmPHOTO are shown on the bottom. The recurrence of each substructure in the tmPHOTO set and tmQM superset are displayed. Transition metal centers are shown as spheres. Iridium is shown in pink, platinum in silver, copper in brown, carbon in gray, nitrogen in blue, oxygen in red, and phosphorus in orange.

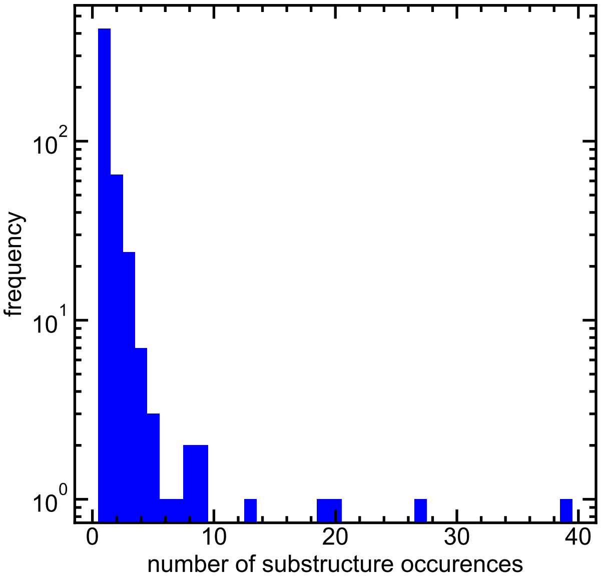


Figure S18. Frequency of molecular subgraph recurrence in the tmSCO dataset.

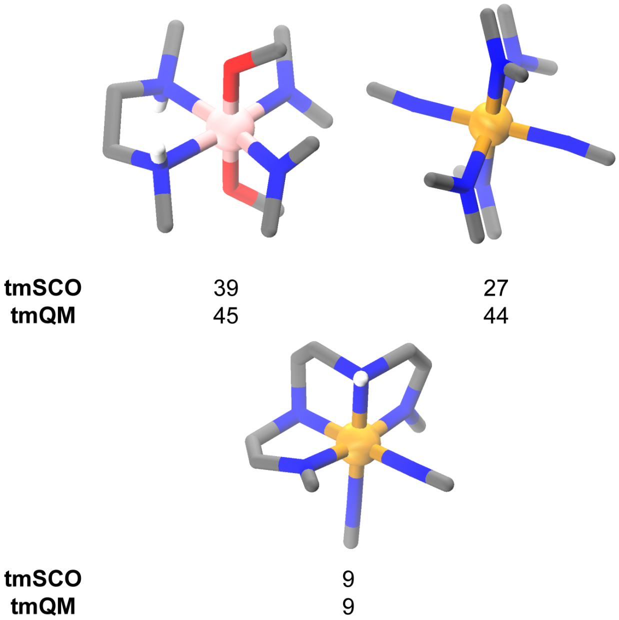


Figure S19. Representative substructures of the tmSCO set. The most common substructures are shown at the top. Structures that appear with high relative frequency in tmSCO are shown on the bottom. The recurrence of each substructure in the tmSCO set and tmQM superset are displayed. Transition metal centers are shown as spheres. Manganese is shown in light pink, iron in orange, carbon in gray, nitrogen in blue, oxygen in red, and hydrogen in white.

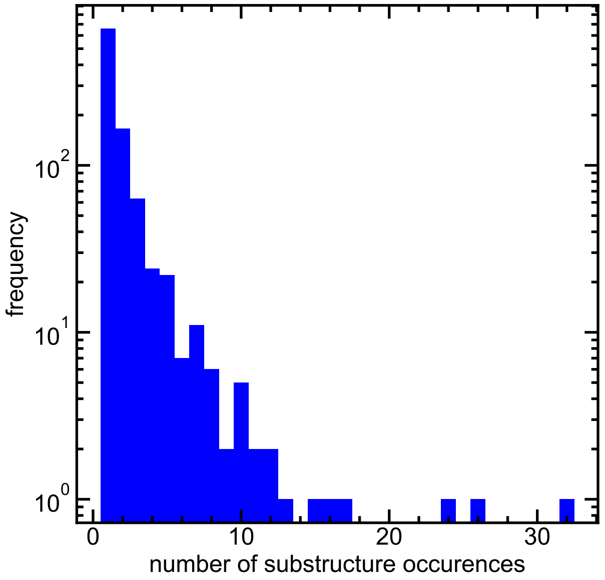


Figure S20. Frequency of molecular subgraph recurrence in the tmBIO dataset.

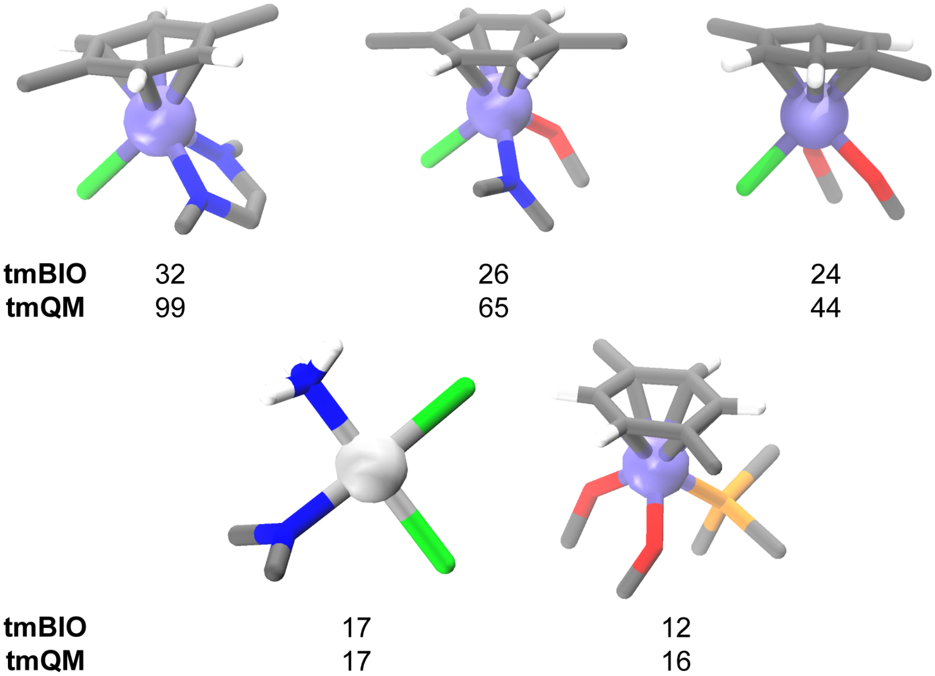


Figure S21. Representative substructures of the tmBIO set. The most common substructures are shown at the top. Structures that appear with high relative frequency in tmBIO are shown on the bottom. The recurrence of each substructure in the tmBIO set and tmQM superset are displayed. Transition metal centers are shown as spheres. Ruthenium is shown in purple, platinum in silver, carbon in gray, nitrogen in blue, oxygen in red, phosphorus in orange, chlorine in green, and hydrogen in white.

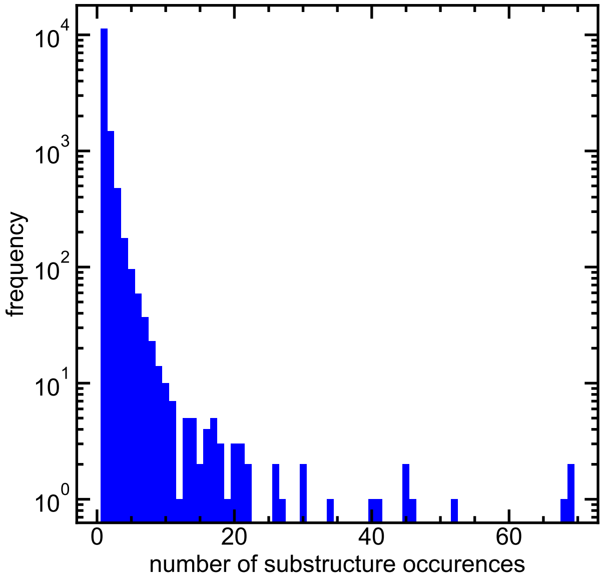


Figure S22. Frequency of d = 3 molecular subgraph recurrence in the tmCAT dataset.