Supporting information for

Deep Learning of Electrochemical CO_2 Conversion Literature Reveals Research Trends and Directions

Jiwoo Choi,[1,2,+] *Kihoon Bang*[1,+], *Suji Jang,*[1] *Jaewoong Choi,*[1] *Juanita Ordonez,*[3] *David Buttler,*[3] *Anna Hiszpanski,*[3] *T. Yong-Jin Han,*[3] *Seok Su Sohn,*[2] *Byungju Lee,*[1] *Kwang-Ryeol Lee,*[1] *Sang Soo Han*[1,*], *Donghun Kim*[1,*]

[1]Computational Science Research Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea

[2]Department of Materials Science and Engineering Korea University, Seoul 02841, Republic of Korea

[3]Materials Science Division, Lawrence Livermore National Laboratory, Livermore, CA, USA

[†]These authors contributed equally.

*Correspondence to: Dr. Donghun Kim(donghun@kist.re.kr), Dr. Sang Soo Han (sangsoo@kist.re.kr)

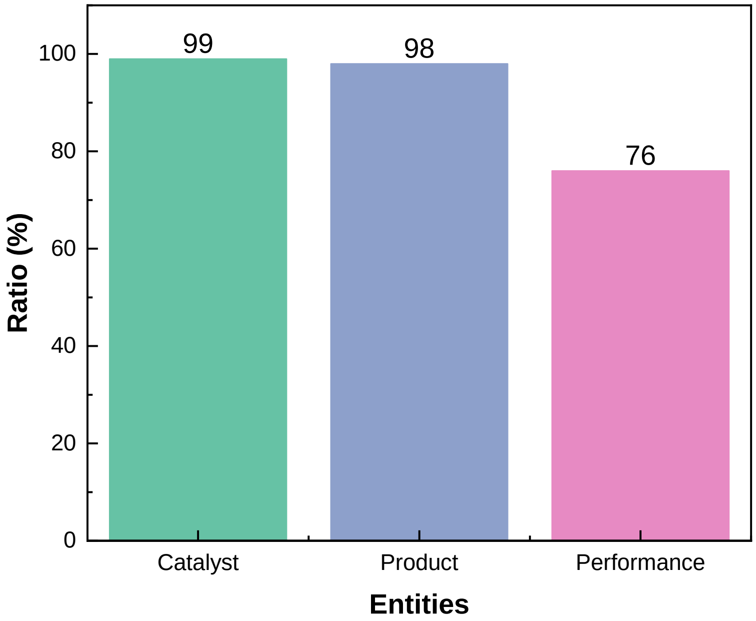

Figure S1. Analysis of major entities present in the abstract. The analysis was done on 100 papers randomly selected. Performance entities include faradaic efficiency, current density, onset potential, overpotential, stability hour and turnover frequency.

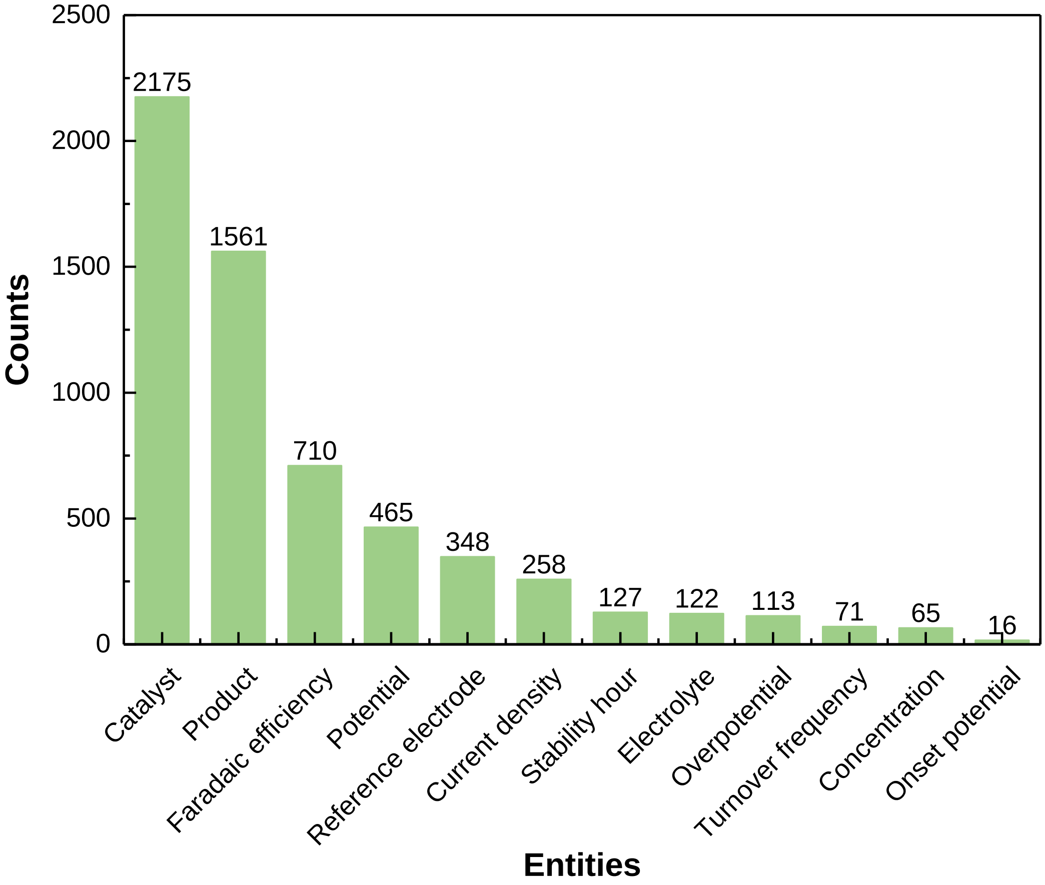

Figure S2. Counts by entity in 500 annotated abstracts used for NER.

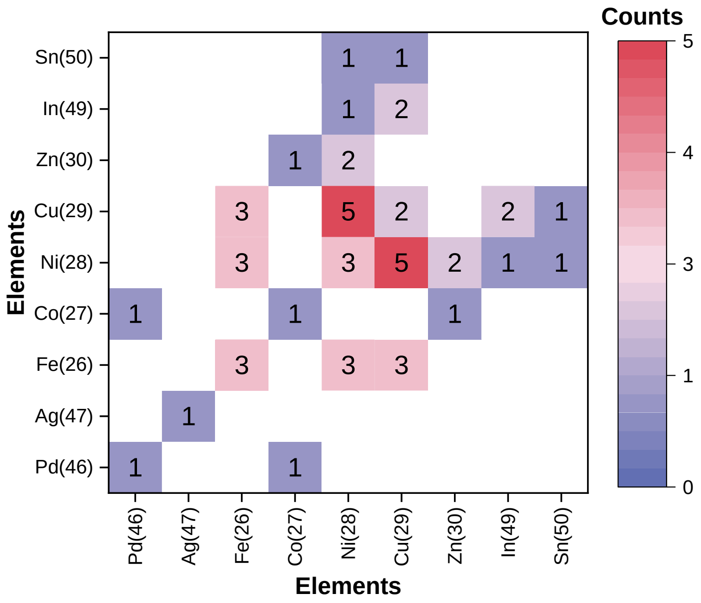

Figure S3. Analysis of catalyst element combinations from the dual-atom catalyst literature. The analysis was done on 32 dual-atom catalyst papers. The redder the color, the more prevalent the combination of that element.

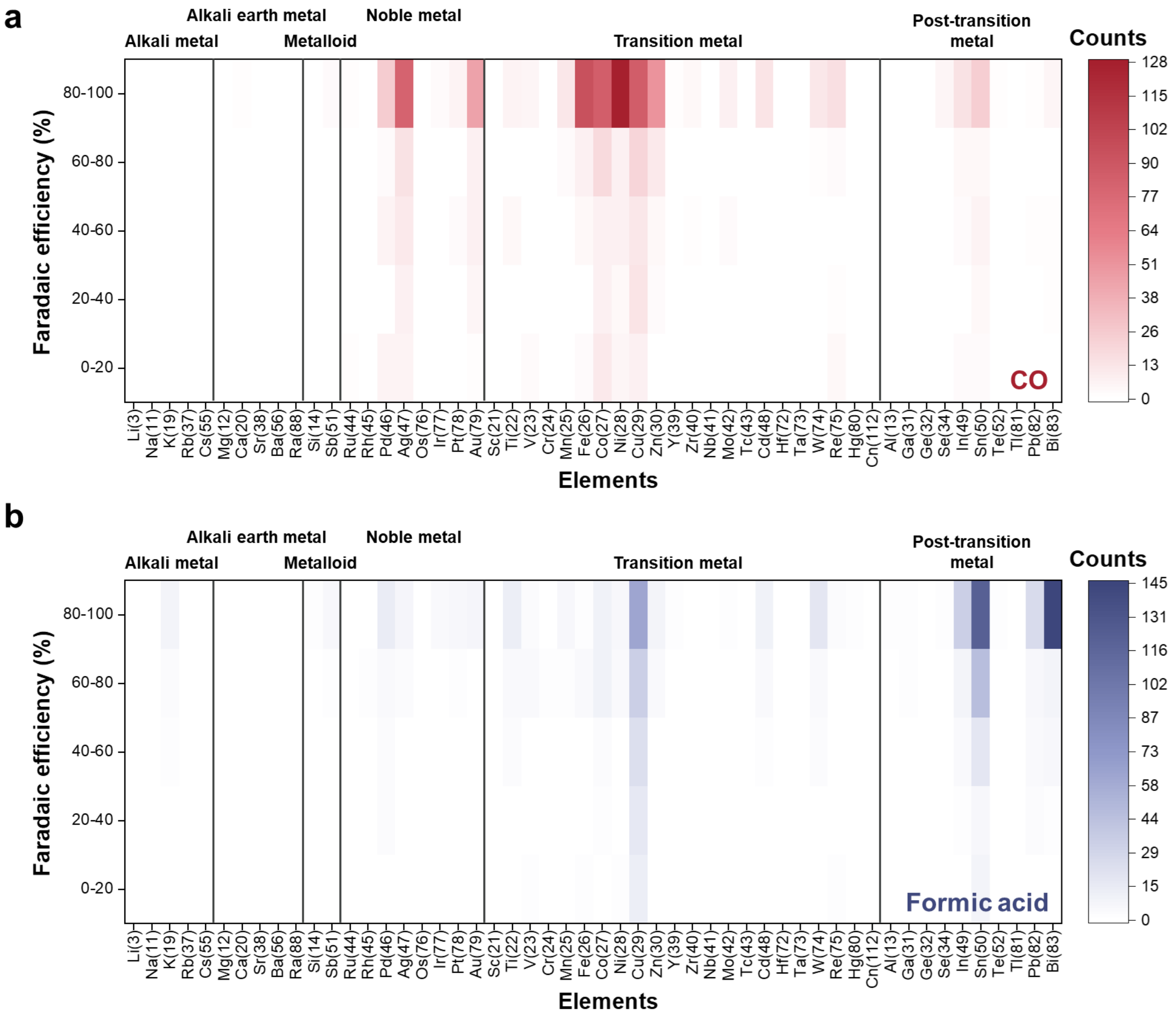

Figure S4. Analysis of Faradaic efficiency values per catalyst element used to generate C1 products. (a) CO (b) Formic acid

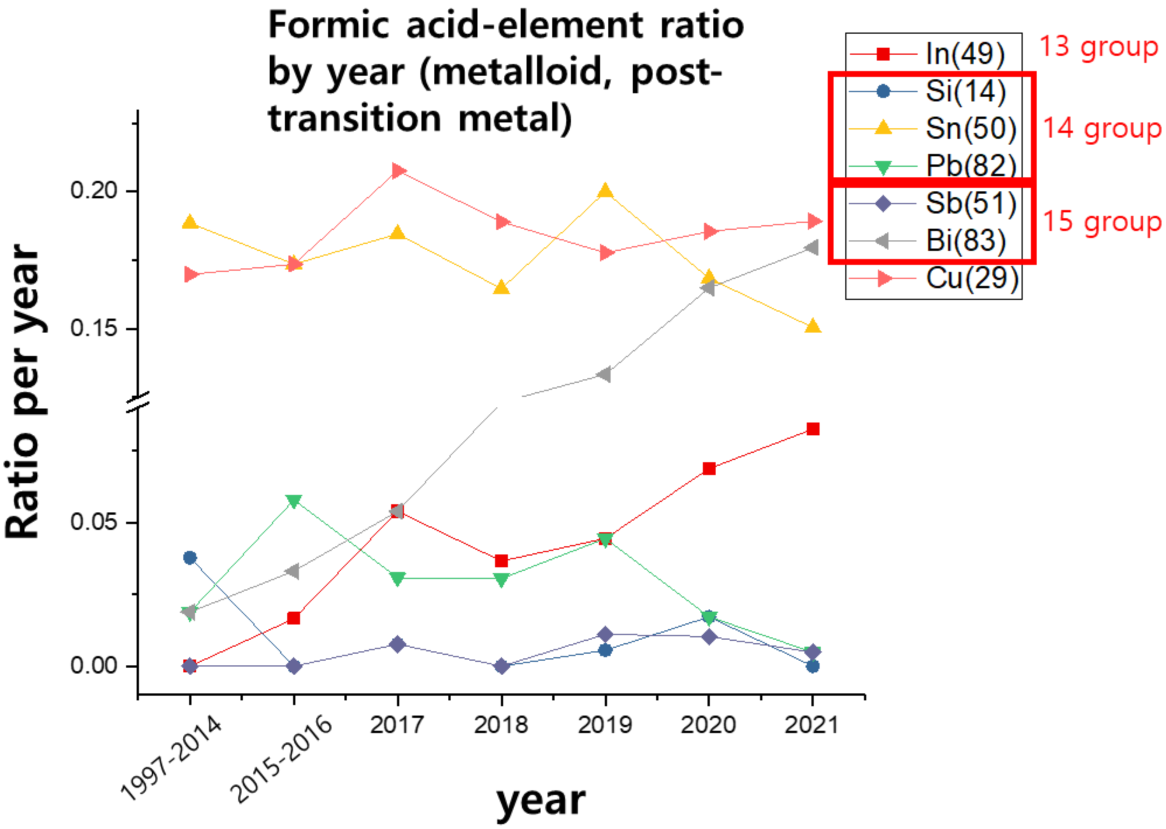

Figure S5. Ratio of metalloids and transition metals that produce formic acid by year

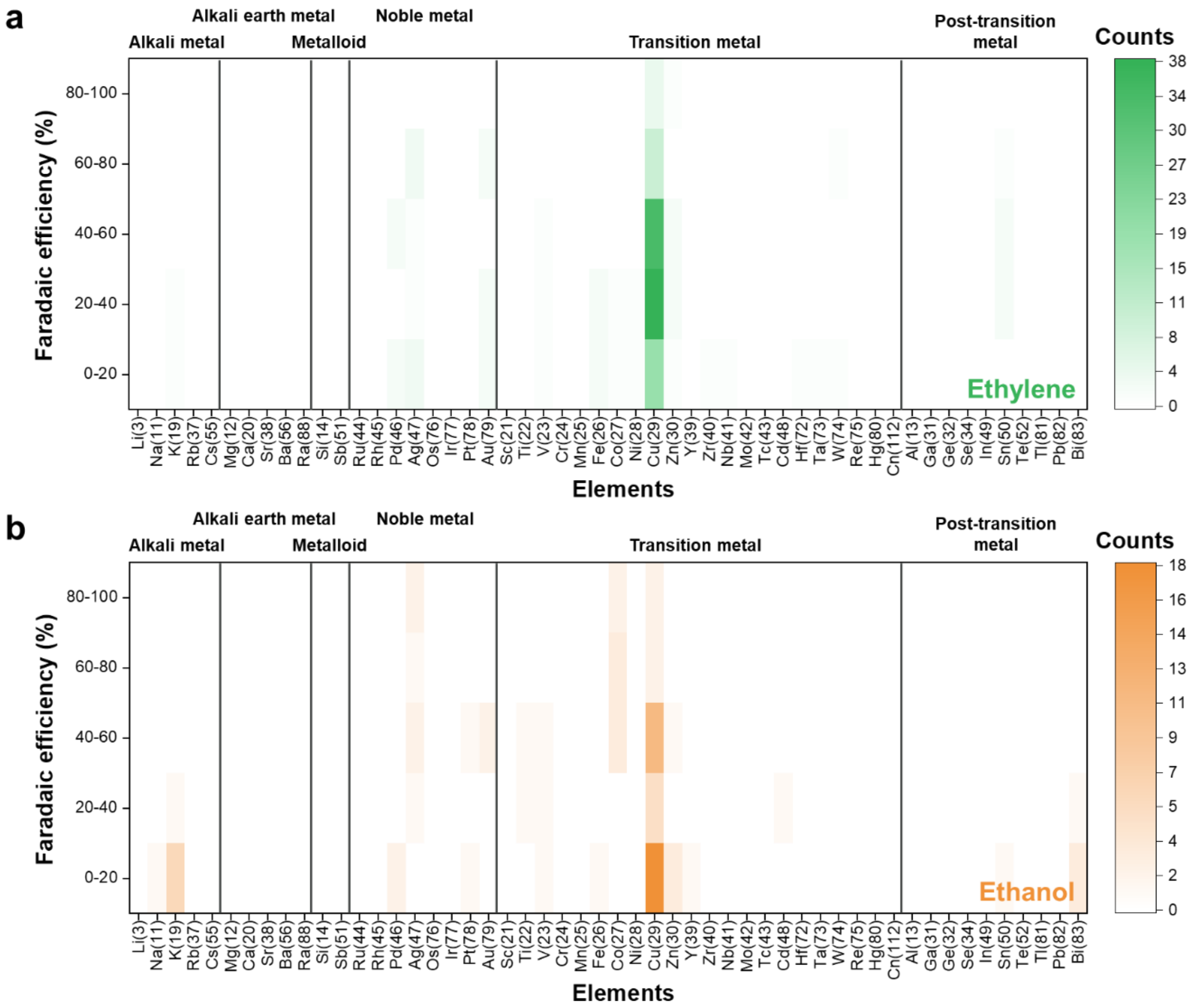

Figure S6. Analysis of Faradaic efficiency values per catalyst element used to generate C2 products. (a) Ethylene (b) Ethanol

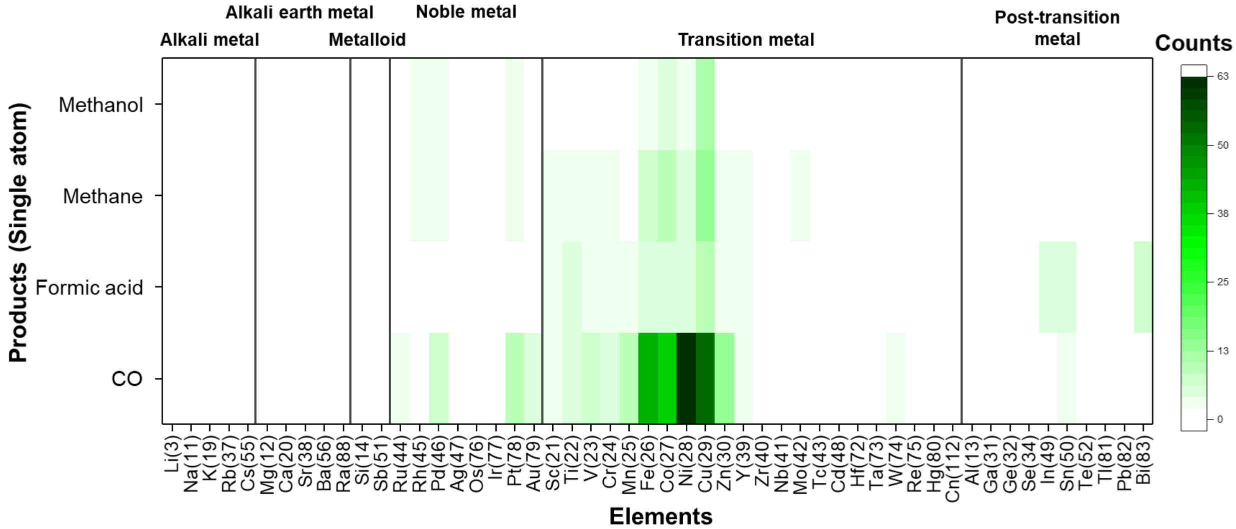

Figure S7. Explorable area heatmaps of products in the case of SAC for each element in CO_2RR catalysts

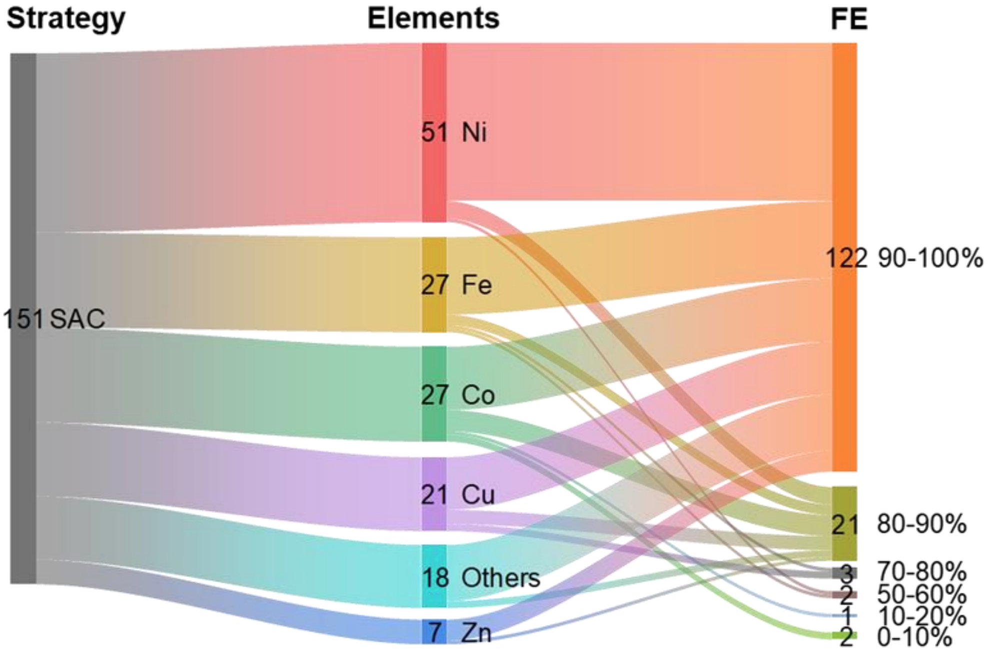

Figure S8. Distribution of 151 SAC papers (for CO productions) connected over used elements and FE ranges. The thicker the connected lines, the higher the corresponding ratio.

Table S1. Examples of annotations by the researcher for each entity in the 500 abstracts used for NER training.

Entity	Example
Catalyst	“copper”, “N-doped hieratically porous carbon”, “MFM-300(In)/carbon-paper electrodes)”, “Pd@Ag cubes”, “Ni-N@NPC”, “Fe, N-co-doped porous carbon nanoparticles”, “Ni-N3-V SAC”, “Cu/Cd composite electrode”, “Porous Pd-In catalyst”, “Bi5Sn60 electrode”
Product	“carbon monoxide”, “CO”, “formic acid”, “C2 products”, “acetic acid”, “acetate”, “CO32-“, “CH3OH”, “HCOOH”
Electrolyte	“KHCO3”, “ChI”, “aqueous ammonia”, “KCl”, “reline”, “K2SO4”, “NaHCO3”, “KOH”
Reference electrode	“RHE”, “reversible hydrogen electrode”, “Ag/Ag+”, “Ag/AgCl”, “FeCp2+/0”, “Fc/Fc+”
Current density	“-9.0 mA/cm2”, “46.1 mAcm-2”, “210 mA・cm-2”, “121.4 mAmg-1”
Faradaic efficiency	“83%”, “over 60%”, “>90%”, “around 81%”, “nearly 96%”, “~85%”, “96 ±8”, “≈95%”, “96 ±8%”, “above 90%”
Stability hour	“20 h”, “30 hours”, “>24h”, “over 35 hours”
Turnover frequency	“1.3 × 104 h-1” , “1622 h-1”, “2100s-1”, “5.7±0.1s-1”, “1.4 × 105 s-1”
Overpotential	“110mv”, “150mv”, “-0.49V”
Onset potential	“-0.8V”, “-0.35”, “-1.0V”
Potential	“-0.9V”, “-1.3V”, “-1.1V”, “-2.15V”
Concentration	“0.1 M”, “0.1m”, “1 molL-1”, “5 mM”

Table S2. 11 LDA topics and words constituting the topic. Topics in red relate to noise, topics in green relate to CO_2RR, topics in blue relate to DFT, and topics in purple relate to analysis.

Topic 1	‘oer', 'mv', 'cm-2', 'evolution', 'water', 'oxygen', '10', 'activity', 'low', 'hydrogen', 'currunt', 'density', 'overpotential', 'metal', 'HER', 'electrocatalysts', 'performance', 'alkaline', 'splitting', 'Co', 'stability', 'efficient', 'catalytic', 'exhibit', 'bifunctional', 'active', 'electrode', 'Ni', 'highly', 'cobalt'
Topic 2	‘atom', 'single', 'metal', 'Ni', 'CO', 'site', 'calculation', 'activity', 'active', 'functional', 'Fe', 'catalytic', 'theory', 'low', 'energy', 'Co', 'coordination', 'hydrogen', 'density', 'potential', 'dft', 'intermediate', 'step', 'overpotential', 'mechanism', 'MoS2', 'ligand', 'transition', 'result', 'evolution'
Topic 3	‘surface', 'site', 'active', 'structure', 'activity', 'oxygen', 'electronic', 'charge', 'performance', 'electron', 'vacancy', 'enhance', 'adsorption', 'improve', 'transfer', 'nanosheets', 'density', 'strategy', 'calculation', 'interface', 'defect', 'result', 'design', 'reveal', 'herein', 'provide', 'functional', 'engineering', 'promote', 'state'
Topic 4	‘energy', 'solar', 'light', 'system', 'TiO2', 'efficiency', 'conversion', 'water', 'photoelectrochemical', 'production', 'chemical', 'fuel', 'PEC', 'process', 'hydrogen', 'use', 'device', 'cell', 'photocathode', 'film', 'H2', 'photocurrent', 'photoanode', 'array', 'photocatalytic', 'visible', 'renewable', 'semiconductor', 'demonstrate', 'produce'
Topic 5	‘carbon', 'metal', 'electrocatalysts', 'activity', 'performance', 'active', 'dope', 'base', 'efficient', 'material', 'strategy', 'site', 'nitrogen', 'porous', 'Zn', 'efficiency', 'catalytic', 'highly', 'herein', 'co2rr', 'framework', 'design', 'work', 'electrocatalytic', 'structure', 'conversion', 'exhibit', 'stability', 'report', 'organic'
Topic 6	‘electrode', 'material', 'g−1', 'carbon', 'battery', 'specific', 'performance', 'energy', 'surface', 'density', 'area', 'capacitance', 'cycle', 'stability', 'composite', 'graphene', 'capacity', 'storage', 'excellent', 'charge', 'porous', 'air', 'rgo', 'structure', 'use', 'exhibit', 'application', 'ion', 'large', 'Ni'
Topic 7	‘complex', 'use', 'oxidation', 'spectroscopy', 'electron', 'electrode', 'study', 'catalytic', 'x-ray', 'show', 'surface', 'water', 'potential', 'solution', 'electrocatalytic', 'oxide', 'result', 'activity', 'base', 'voltammetry', 'film', 'analysis', 'ph', 'formation', 'condition', 'two', 'microscopy', 'investigate', 'cyclic', 'observe'
Topic 8	‘cell', 'electrode', 'use', 'electrolyte', 'Pt', 'carbon', 'gas', 'methanol', 'cathode', 'fuel', 'increase', 'product', 'acid', 'current', 'membrane', 'temperature', 'potential', 'electrolysis', 'study', 'condition', 'result', 'low', 'rate', 'liquid', 'mass', 'concentration', 'process', 'oxidation', 'base', 'CO'
Topic 9	‘efficiency', 'faradaic', 'formate', 'electrode', 'current', 'density', 'CO', 'cm−2', 'selectivity', 'rhe', 'fe', 'electroreduction', 'production', 'potential', 'Ag', 'Sn', 'product', 'bi', 'carbon', 'hydrogen', 'reversible', 'conversion', 'selective', 'achieve', 'low', 'exhibit', 'copper', 'versus', 'efficient', 'dioxide'
Topic 10	‘nrr', 'N2', 'NH3', 'yield', 'ammonia', 'nitrogen', 'ambient', 'condition', 'efficiency', 'h−1', 'faradaic', 'hydrogen', 'process', 'rate', 'synthesis', 'μg', '0.1', 'production', 'energy', 'fixation', 'report', 'electrocatalysts', 'efficient', 'electrocatalytic', 'achieve', 'electrode', 'reversible', 'electrocatalyst', 'rhe', 'Ru'
Topic 11	‘Cu', 'surface', 'CO', 'selectivity', 'co2rr', 'product', 'Au', 'Pd', 'activity', 'show', 'nanoparticles', 'Ag', 'formation', 'np', 'size', 'copper', 'intermediate', 'alloy', 'structure', 'production', 'use', 'potential', 'result', 'ethanol', 'carbon', 'effect', 'CH4', 'low', 'c2', 'catalytic'

Table S3. Comparison of major entities present in the abstract and the main text. Data highlighted in yellow represent the catalysts with the best performance in the paper.

	Type	Catalyst	Product	Performance
Paper 1	Abstract	**N-doped hieratically porous carbon**	**CO**	**83%**
	Main text	**NH3 etched DAPC (NDAPC)**	**CO**	**83%**
		Deasphaltened petroleum pitch-based carbon (DAPC)	CO	52%
		NH3 etched PC (NPC)	CO	53%
		Petroleum pitch-based carbon (PC)	CO	38%
Paper 2	Abstract	**polyoxometalate (SiW11Mn)-assisted metal In**	**acetic acid**	**72.10%**
			formic acid	6.10%
	Main text	**Indium electrode**	**acetic acid**	**72.10%**
			formic acid	6.10%
Paper 3	Abstract	**Catalyst 1-Mn (ortho-)**	**CO**	**901 s^{-1}**
		catalyst Mn(bpy)(CO)3Br	CO	102 s^{-1}
	Main text	**1-Mn**	**CO**	**901.4 s^{-1}**
		2-Mn	CO	245.2 s^{-1}
		3-Mn	CO	296 s^{-1}
		MnBpy	CO	102.1 s^{-1}
Paper 4	Abstract	**Catalyst films with the highest sulfur content of 2.7 at %**	**formate**	**−13.9 mA cm^{-2}**
	Main text	**Cu-5000S**	**formate**	**−13.9 mA cm^{-2}**
		Cu-0S	formate	−1.8 mA cm^{-2}
Paper 5	Abstract	**Cu–Sn alloy**	**CO**	**90%**
		OD-Cu	CO	63%
	Main text	**Cu–Sn**	**CO**	**90%**
		OD-Cu	CO	63%
		Sn deposited on Sn	CO	-

Table S4. Performance change of MatBERT according to the tag format (IOB, IOBE, IOBES). Tenfold crossvalidation was performed. The performance is represented by the F1-score, which is expressed as a %. The best performance is indicated by red text.

Tag format Entity	IOB	IOBE	IOBES
Catalyst	80.16	80.59	80.67
Concentration	91.71	90.52	93.98
Current density	96.35	94.72	94.96
Electrolyte	81.6	82.98	79.72
Faradaic efficiency	96.65	95.94	96.72
Onset potential	62.12	68.08	88.76
Overpotential	93.45	94.15	93.89
Potential	98.34	97.97	98.32
Product	96.19	96.22	96.19
Reference electrode	99.07	99.4	99.58
Stability hour	93.85	90.79	92.79
Turnover frequency	95.32	96.74	92.5
Micro average	90.23	90.28	90.38

Table S5. 10-fold cross-validation result for NER. The value represents F-1 score, which is expressed as a %. 500 abstracts were used as a trainset for NER. A hyphen indicates that the entity does not exist in the test set of the corresponding fold. The lowest performance is indicated by red text.

Fold Entity	1	2	3	4	5	6	7	8	9	10	Average
Catalyst	84.4	81	78.5	79.2	82.8	81.7	77.6	78.9	82.3	80.3	**80.7**
Concentration	100	100	100	100	91	71.4	100	83.3	100	94.1	94
Current density	100	94.1	94.1	97.1	100	91.3	91.7	100	90.4	90.9	945
Electrolyte	71.4	87	96.6	76.2	72.7	92.3	85.7	84.9	62.1	68.4	79.7
Faradaic efficiency	98.6	97.6	99.4	96.6	95	97.8	95.8	96.2	95.2	95	96.7
Onset potential	75	100	66.7	100	100	100	57.1	100	100	-	88.8
Overpotential	93.8	84.6	80	94.1	100	100	100	92.3	100	94.1	93.9
Potential	100	100	93.1	99.2	98.8	100	100	97.3	98.5	96.3	98.3
Product	97.3	98.4	95.6	94.8	96.2	95.8	96.3	96	96.1	95.5	96.2
Reference electrode	100	100	97.4	100	100	100	100	100	100	98.4	99.6
Stability hour	85.7	92.3	90	94.1	87	88.9	93.3	96.6	100	100	92.8
Turnover frequency	100	85.7	88.9	100	87.5	89.7	82.4	90.9	100	100	92.5
Micro average	92.8	91	89.8	91	91.2	89.8	88.9	89.5	90.4	89.6	**90.4**

Table S6. Results of applying our NER model to 100 abstracts not used for training. The value represents F-1 score, which is expressed as a %. The 100 abstracts are selected randomly. Texts that improved performance when boundary relaxation was applied are shown in red.

F1-score Entity	Boundary relaxation X	Boundary relaxation O
Catalyst	**87.5**	**91.2**
Concentration	80	80
Current density	93.6	93.6
Electrolyte	85.7	85.7
Faradaic efficiency	91.5	91.5
Onset potential	100	100
Overpotential	93.3	93.3
Potential	98.1	98.1
Product	**90.9**	**92.1**
Reference electrode	100	100
Stability hour	97.9	97.9
Turnover frequency	100	100
Micro average	**90.7**	**92.5**

Table S7. How to extract the strategy and the words included in the strategy. Based on the high frequency words in the titles of 3,153 abstracts, 7 strategies and words corresponding to each strategy were manually classified by the researcher.

Strategy	**Keywords**	**Total number**
Core shell	**‘shell’, ‘core’**	**130**
Defect engineering	**‘vacancy’, ‘defect’, ‘amorphous’, ‘step’, ‘defective’, ‘polycrystalline’, ‘stepped’**	**190**
Alloy	**‘bimetallic’, ‘alloy’, ‘ordered’, ‘metallic’, ‘intermetaling’, ‘heteroatom’, ‘nanoalloys’, ‘bimetal’, ‘trimetallic’, ‘alloyed’, ‘heterobimetallic’, ‘multimetallic’, ‘bi-metallic’**	**243**
Single atom	**‘single’, ;dispersed’, ‘’atomically’, ‘coordinated’**	**248**
Doping	**‘doped’, ‘doping’, ‘co-doped’, ‘codoped’, ‘co-doping’**	**355**
Architecture engineering	**’embedded’, ‘composite’, ‘decorated’, ‘assembly’, ‘coupling’, ‘nanocomposite’, ‘encapsulated’, ‘coated’, ‘assembled’, ‘heterostructure’, ‘heterostructures’, ‘interfacial’, ‘integrated’, ‘nanocomposites’, ‘coating’, ‘heterostructured’, ‘decoration’, ‘encapsulation’, ‘heterojunction’, ‘compositional’, ‘integration’, ‘incorporation’, ‘hetero’]**	**368**
Shape control	**’nanosheets’, ‘porous’, ‘nanotube’, ‘mesoporous’, ‘nanoporous’, ‘nanowire’, ‘nanosheet’, ‘nanowires’, ‘morphology’, ‘facet’, ‘dendritic’, ‘nanorods’, ‘nanocubes’, ‘dendrite’, ‘nanofibers’, ‘nanorod’, ‘flower’, ‘monolayer’, ‘sheet’, ‘nanofiber’, ‘nanoarrays’, ‘nanoplates’, ‘monolayers’, ‘nanocages’, ‘nanoflake’, ‘nanodendrites’, ‘microporous’, ‘pore’, ‘leaf’, ‘nanoflowers’, ‘honeycomb’, ‘nanocube’, ‘sponge’, ‘nanosponges’, ‘nanoflakes’, ‘nanofibbons’, ‘nanoprisms’, ‘nanoflower’]**	**911**

Table S8. Results of applying our NER model to 32 dual-atom catalyst papers. The value represents F1-score (%). The cases where performances were improved upon boundary relaxation applied were marked in red.

F1-score **Entity**	**Boundary relaxation X**	**Boundary relaxation O**
Catalyst	**79.56**	**85.40**
Concentration	100.00	100.00
Current density	100.00	100.00
Electrolyte	66.67	66.67
Faradaic efficiency	93.67	93.67
Onset potential	-	-
Overpotential	100.00	100.00
Potential	96.20	96.20
Product	**92.86**	**95.71**
Reference electrode	**93.02**	**97.67**
Stability hour	100.00	100.00
Turnover frequency	80.00	80.00
Micro average	**87.95**	**91.15**

Table S9. Performance change of MatBERT according to the batch size. Tenfold cross validation was performed. The performance is represented by the F1-score, which is expressed as a %. The best performance is indicated by red text.

Batch size Entity	Batch 4	Batch8	Batch 16	Batch 32 (best)
Catalyst	66.65	76.05	79.31	80.67
Concentration	83.78	88.73	90.93	93.98
Current density	92.76	93.98	94.89	94.96
Electrolyte	74.03	76.72	77.16	79.72
Faradaic efficiency	92.19	93.59	94.8	96.72
Onset potential	46.91	70.37	83.89	88.76
Overpotential	85.38	90.66	91.78	93.89
Potential	95.78	96.39	97.9	98.32
Product	94.2	94.78	95.83	96.19
Reference electrode	98.45	99.48	99.07	99.58
Stability hour	91.43	91.07	94.57	92.79
Turnover frequency	88.23	92.92	93.95	92.5
Micro average	83.46	87.63	89.5	90.38