scientific data

Check for updates

OPEN

DATA DESCRIPTOR

TheMeCat: Dataset of Thermocatalytic Conversion of CO_2 to Methanol

Árpád I. Toldy[1], Prajwal Pisal[2,3,4], Ondřej Krejčí[2,5], Patrick Rinke[2,3,4,6] & Annukka Santasalo-Aarnio[1]✉

Data-driven materials discovery to accelerate the development of new catalysts for the green transition shows great promise, but requires machine-interpretable experimental data. For this purpose, we compiled the TheMeCat dataset for the thermocatalytic conversion of CO_2 to methanol. TheMeCat is curated from experimental literature sources, and designed for hybrid computational and experimental catalyst workflows. The dataset captures key catalyst features, process parameters, and methanol conversion efficiencies. Upon detailed analysis of the generated dataset and the associated variability in reported values, we perform targeted error analysis to assess consistency and reinforce the need for clear, structured reporting.

Background & Summary

Capturing and converting CO_2 into fuels and chemicals offers a promising route to reduce, or potentially replace, our reliance on fossil fuels[1]. Among the various utilization strategies, the transformation of CO_2 into methanol has attracted the most attention, largely due to methanol's versatility as a platform chemical[2]. While electrochemical and photochemical methods are under active investigation[3–5], nearly all methanol today is produced via thermochemical processes using fossil-derived synthesis gas. Given the maturity of this technology and the existing infrastructure, thermocatalytic conversion is expected to play a central role in the production of low-carbon methanol.

This process involves reacting CO_2 with clean hydrogen, as shown in the reaction: $CO_2 + 3H_2 \rightarrow CH_3OH + H_2O$. Commercial catalysts used for methanol synthesis from syngas ($CO + 2H_2 \rightarrow CH_3OH$) already deliver satisfactory CO conversion rates and methanol yields. However, performance could be significantly enhanced by developing catalysts specifically tailored for CO_2-based methanol synthesis[6]. Despite significant potential, the high costs associated with catalyst development, including design, characterization, and testing, remain a major barrier.

Recently, data-driven material discovery that incorporates artificial intelligence (AI) has led to rapid advances in the development and optimization of new materials for various applications[7–11].

Machine Learning (ML) algorithms play a crucial role in this process by identifying patterns in datasets, thus enabling the prediction of material properties, and guiding experimental efforts. By leveraging predictive models trained on curated data, researchers can efficiently screen candidates, optimize compositions, and accelerate discovery, all while reducing experimental costs.

For example, Takahashi *et al.*[12] used a random forest algorithm trained on data from 1,868 catalysts extracted from the literature to identify key descriptors for optimizing C_2 yield in the oxidative coupling of methane. From more than 330 million ML-generated data points, they selected promising catalyst compositions[12]. A key feature of their study, as well as other similar ones, was the ability to compile high-quality datasets, providing ML tools with machine-readable data on the properties studied[12,13]. Despite several attempts to compile such datasets for CO_2 to methanol thermocatalysis, most existing efforts either do not provide the compiled database[14–16], use

[1]Department of Energy and Mechanical Engineering, Aalto University, P.O. Box 11000, AALTO, FI-00076, Finland. [2]Department of Applied Physics, Aalto University, P.O. Box 11000, AALTO, FI-00076, Finland. [3]Department of Physics, Technical University of Munich, James-Franck-Strasse 1, Garching, 85748, Germany. [4]Munich Center for Machine Learning (MCML), Arcistrasse 21, Munich, 80333, Germany. [5]Department of Mechanical and Materials Engineering, University of Turku, Vesilinnantie 5, Turku, FI-20014, Finland. [6]Atomistic Modeling Center, Munich Data Science Institute, Technical University of Munich, Walther-Von-Dyck Str. 10, Garching, 85748, Germany. ✉e-mail: annukka.santasalo@aalto.fi

only the highest-yield measurements from each source[17], or lack important experimental parameters and the proper documentation of the units used[18]. Suvarna *et al.*[19] provide a well-curated dataset to support ML-driven space-time yield predictions[20]. However, there is a notable lack of a dataset that provides rigorously standardized chemical compositions with complete molar fractions and incorporates all essential catalytic performance indicators. Such a dataset would enable more flexible integration with advanced material descriptors and catalyst analysis.

Recognizing the need for standardized and accessible datasets in catalyst discovery, we collected and curated the TheMeCat dataset[21]. TheMeCat is an open-source, AI-ready dataset focused on the thermocatalytic conversion of CO_2 to methanol. Compiled from experimental literature, it is designed to support high-throughput computational and experimental catalyst screening. As a foundational step toward data standardization in catalysis, TheMeCat aims to promote more consistent and reusable data practices. More broadly, it encourages the catalysis community to adopt publishing standards that enhance the utility of experimental data for accelerated materials discovery. To support this goal, we also identify current limitations and propose best practices for data collection and reporting.

Methods

Data Collection and Processing. TheMeCat was compiled from experimental literature sources. We acquired potential data sources by searching the literature for studies on experimental CO_2-based methanol synthesis, including an initial search on Google Scholar and Web of Science and an ongoing query on PubCrawler (https://pubcrawler.gen.tcd.ie/with the keywords: `(methanol [All Fields] OR CH3OH [All Fields]) AND (CO2 [All Fields] OR carbon dioxide [All Fields])`. Altogether, our search resulted in more than 500 potential sources.

We applied several criteria to select the studies to be included in TheMeCat. First, we restricted the dataset to experiments conducted in purely thermocatalytic single-pass packed-bed reactors without any process-driven enhancements that present confounding factors in the comparison of catalyst materials (e.g., articles on sorption-enhanced or plasma-assisted synthesis were excluded). Second, we only included studies, in which CO_2 and H_2 are the sole reactants - this led to the exclusion of the rich body of experimental literature aimed at kinetic model development due to the presence of CO as a reactant. Lastly, we selected studies where the catalyst composition could be quantified at the elemental level, as this information is essential for descriptor-based machine learning approaches. Elemental composition serves as a fundamental input for model training, allowing for the development of meaningful structure-property relationships that aid in catalyst design.

After applying these criteria, a significant portion of the literature was excluded, resulting in the final dataset being compiled from 44 sources[22–67], resulting in 822 data entries. In order to establish a comprehensive reporting framework for thermocatalytic CO_2-based methanol synthesis, we incorporated all experimental data points from each included study.

Notably, data extraction and processing from literature sources had to be carried out manually, as large language models (LLMs) and natural language processing (NLP) techniques remain limited in this context[68–70]. One major challenge for these tools is interpreting catalyst compositions, as many studies employ ambiguous or inconsistent reporting conventions that hinder reliable extraction and standardization. Additionally, LLMs struggle to generate consistent databases when calculations and unit conversions are involved, as these steps often rely on nuanced methodological details requiring domain-specific reasoning such as the complexities of gas hourly space velocity (GHSV) discussed in the Process Parameters section. Critical experimental data, especially those related to less successful conditions, are frequently presented in graphical formats, necessitating specialized extraction tools. Taken together, these challenges render automated extraction methods unreliable.

We incorporated three primary categories of data into TheMeCat: information on the composition and properties of the catalyst, details of the experimental processing parameters, and records of the experimental results. Each of these categories is presented in sections Catalyst Composition and Properties, Process Parameters, and Catalytic Performance Measures, respectively.

Catalyst Composition and Properties. We included the following data on catalysts into TheMeCat: composition, specific surface area, and mean pore diameter, because a) the composition of the catalyst has by far the largest influence on its activity and b) the specific surface area/mean pore volume are *de facto* process parameters that can limit the availability of active sites. Any further data on catalysts, such as preparation methods and process parameters were excluded.

The extraction of catalyst composition from each source began by identifying the active and support materials. The active-support distinction in TheMeCat was based solely on the claims presented in the sources. Since many catalysts contain multiple materials serving as either active or support components, the dataset allows up to two active and two support components. Within the categories, the order generally reflects the relative mole percentages, placing the more abundant material first. When a reduction step was mentioned in the source, pre-reduction starting materials of the active component were also identified. The composition was then quantified in terms of molar percentage of the entire catalyst material. For example, in a catalyst containing two active materials on a single support, the molar percentage of the active component one, labeled *active_comp*_1, is calculated as

$$active\_comp\_1 = \frac{n_{a_1}}{n_{a_1} + n_{a_2} + n_{s_1}} \times 100\%, \quad (1)$$

where n_{a_1}, n_{a_2}, and n_{s_1} represent the amount of moles of active material 1, active material 2, and support material in one mole of catalyst, respectively.

The availability of compositional information ranged between the following extremes: 1) full elemental and molecular composition was available, typically from a combination of inductive coupled plasma (ICP) spectroscopy, scanning electron microscopy with energy dispersive X-ray spectroscopy (SEM-EDX) and other methods; 2) the catalyst composition was stated without providing any evidence. In the latter case (and similar cases of limited information), we assumed that the synthesis succeeded as intended by the authors of the source, and the claims were taken at face value. Many studies report the presence of various derivative species, specific crystal facets, etc. during catalyst characterization. These were not considered by us.

The specific surface area A_{BET} and the mean pore diameter d_p of the catalyst obtained through physisorption evaluated with the Brunauer-Emmett-Teller (BET) method was usually available from the sources in the units used in the dataset (m^2/g and nm, respectively). In some sources, the total pore volume V_p is provided in cm^3/g instead of d_p. For such cases, d_p can be calculated as[71]:

$$d_p = \frac{4V_p}{A_{BET}} \times 1000 \tag{2}$$

Process Parameters. TheMeCat includes experimental process parameters such as temperature (T), pressure (p), the H_2/CO_2 ratio (based on pressure or molar flow, p_{H_2}/p_{CO_2}), catalyst mass (m_{cat}), and gas hourly space velocity ($GHSV$). Temperature and pressure were typically reported in units that could be readily converted to Kelvin and bar, respectively. When temperature was given in degrees Celsius, a rounded offset of 273 was added for conversion to Kelvin. Catalyst mass was usually provided in grams, which is the unit adopted in TheMeCat. The H_2/CO_2 ratio was commonly specified either as a ratio of partial pressures or molar flow rates.

The extraction of $GHSV$, on the other hand, posed several challenges. $GHSV$ in TheMeCat is defined according to equation (3), which is the most widely used definition in the studies included (e.g.[22,25]):

$$GHSV = \frac{\dot{V}}{m_{cat}} \tag{3}$$

where $\dot{V}$ is the total flow rate of the *reactants* in normal liters per hour (nL/h) and m_{cat} is the mass of the catalyst in the reactor in grams. Thus, $GHSV$ is expressed in units of normal liters per hour per gram of catalyst ($nLh^{-1}g^{-1}$), where normal conditions refer to $T_n = 293.15\,K$ and $p_n = 101325\,Pa$. When $GHSV$ was specified under reaction conditions, we used the ideal gas law $pV = nRT$ to convert it to units of ($nLh^{-1}g^{-1}$). Occasionally, reported $GHSV$ values included the flow rate of the internal standard, an inert gas used for gas chromatography (GC) quantification purposes. In such cases, we corrected $GHSV$ to match equation (3) according to:

$$GHSV = GHSV_R \times (1 - x_{std}), \tag{4}$$

where $GHSV_R$ is the reported $GHSV$ and x_{std} is the molar fraction (equivalent to the volume fraction) of the internal standard in the reactant mixture.

Regardless of its specific formulation, calculating $GHSV$ requires that flow rates (or $GHSV$ itself) are reported in an unambiguous format. Unfortunately, some sources specify $\dot{V}$ in terms of volume/time (e.g., in units of mL/min) without specifying the reference pressure and temperature. When "standard conditions", such as "STP" are specified, the standard (e.g., the most widely used IUPAC definition[72]) is rarely referenced. Moreover, many sources publish $GHSV$ on a volumetric basis (i.e., $GHSV_V$ in units of h^{-1}) according to equation (5):

$$GHSV_V = \frac{\dot{V}}{V_{ref}} \tag{5}$$

where V_{ref} is typically the volume of the reactor or the catalyst bed. However, many sources do not specify which reference volume (V_{ref}) was used to calculate $GHSV_V$, and often omit details about the reference temperature and pressure as well. In many cases, neither $\dot{V}$ nor $GHSV_V$ is sufficiently specified. In such cases, the reference pressures, temperatures, and volumes could mostly be inferred from the reported experimental results to calculate the $GHSV$. Finally, a significant portion of sources define $GHSV$ according to equation (3), but use the incorrect unit of h^{-1}. In those instances, the units were corrected.

Catalytic Performance Measures. Catalytic performance measures included in TheMeCat are CO_2 conversion X, methanol selectivity S, methanol yield Y, and methanol space-time yield STY. The first three are reported as percentages. CO_2 conversion is defined as

$$X = \frac{\dot{n}_{CO_2,in} - \dot{n}_{CO_2,out}}{\dot{n}_{CO_2,in}} \times 100\% \tag{6}$$

where $\dot{n}_{CO_2,in}$ and $\dot{n}_{CO_2,out}$ represent the molar flow rate of CO_2 that enters and exits the reactor, respectively. Methanol selectivity S is given by

$$S = \frac{\dot{n}_{CH_3OH,out}}{\dot{n}_{CO_2,in} - \dot{n}_{CO_2,out}} \times 100\% \tag{7}$$

where $\dot{n}_{CH_3OH,out}$ is the molar flow rate of methanol exiting the reactor. Methanol yield Y is calculated using

$$Y = \frac{\dot{n}_{CH_3OH,out}}{\dot{n}_{CO_2,in}} \times 100\%. \tag{8}$$

Y could also be obtained as a product of X and S

$$Y = \frac{X \times S}{100\%}. \tag{9}$$

Finally, the methanol space-time yield STY is defined as

$$STY = \frac{\dot{m}_{CH_3OH,out}}{m_{cat}} \tag{10}$$

where $\dot{m}_{CH_3OH,out}$ is the mass flow rate of methanol exiting the reactor in units of g/h. Combining equations (3) and (8) and the ideal gas law, STY can be calculated (similarly to, e.g.[51]) as

$$STY = \frac{GHSV \times Y \times M_{CH_3OH} \times p_n}{\left(1 + \frac{p_{H_2}}{p_{CO_2}}\right) \times 100\% \times R \times T_n \times 10^3} \tag{11}$$

where M_{CH_3OH} is the molar mass of methanol (i.e., 32.04 g/mol).

Data on X, S, and Y is readily available in almost all sources, although numerical data is typically only provided for the "best" results, and other data points are relegated to figures. Therefore, extraction of experimental results required the use of plot digitization software such as PlotDigitizer (https://plotdigitizer.com). Very few studies publish any assessment of their quantification methods, such as calibration curves (accuracy) or repeatability analysis (precision) for their gas chromatography (GC) methods. However, the uncertainty introduced by plot digitization was less than a typical GC precision error of 2%, as confirmed by results published both in a numerical and graphical format.

While STY could be calculated according to equation (11) for most data points, it could not be calculated for a few data points due to the complete absence of reported flow rates or $GHSV$s. We left the corresponding field empty in such cases.

Some of the sources reported $GHSV_V$ (eq. (5)) and subsequently STY_V

$$STY_V = \frac{\dot{m}_{CH_3OH,out}}{V_{cat}} \tag{12}$$

where $V_{cat} = V_{ref}$ is the catalyst volume. At the same time, these sources did not specify the bulk density of the catalyst, and thus the standard space-time yield cannot be calculated.

We chose not to include STY_V in this dataset to avoid the following two drawbacks: first, it is impossible to compare such data to STY reported on the basis of catalyst mass (the approach used by the vast majority of sources). Second, catalyst powders consolidate when they are loaded into the reactor and sealed off[73], typically by a layer of inert material. This leads to an increased bulk density and a reduced catalyst volume, which, in turn, results in inaccurate $GHSV_V$ and STY_V figures. From an industrial perspective, STY is arguably the most suitable performance indicator to compare different catalysts. Therefore, the consistent reporting of STY is of utmost importance for data-driven catalyst development.

Data Records

TheMeCat is published in a CSV file[21]. The first row contains the headers of the columns and each further row of the dataset represents an individual experimental result reported in the literature. Table 1 describes each column.

Data Overview

In this section, we analyze the key trends emerging from the dataset to better understand the factors that influence the catalytic performance. We compare the reported yield and STY. With this we consider STY being the more appropriate metric for performance evaluation. We then examine how catalyst composition and experimental parameters such as temperature, pressure, and GHSV collectively shape STY outcomes. Based on these observations, we also discuss limitations inherent to the dataset and propose best practices to guide future data reporting and collection efforts.

Yield vs. STY. TheMeCat covers a wide range of methanol yield values, from 0% to 45%, as illustrated in Fig. 1(a). When normalized to space-time yield (STY, expressed as $g_{CH_3OH}/g_{cat} \cdot h$), the distribution shifts, with all

Keyword	Description	Unit/Data-Type
reference	Source of the data in *Author (year)* format	String
doi_link	Digital Object Identifier (DOI) of the literature source	String
catalyst_name	Human-readable identifier of the catalyst	String
catalyst_composition _available	Indicates whether catalyst composition is available	Boolean
starting_material _active_1	First active material before the reduction step	String
active_comp_1	Chemical formula of the first active component in its reduced state	String
active_1_percent	Mole percentage of the first active component	*mol*%
starting_material _active_2	Second active material before the reduction step	String
active_comp_2	Chemical formula of the second active component in its reduced state	String
active_2_percent	Mole percentage of the second active component	*mol*%
support_comp_1	Chemical formula of the first support component	String
support_1_percent	Mole percentage of the first support component	*mol*%
support_comp_2	Chemical formula of the second support component	String
support_2_percent	Mole percentage of the second support component	*mol*%
catalyst_surface _area_m2_g	Specific surface area of the catalyst	m^2/g
mean_pore_diameter_nm	Reported mean pore diameter.	*nm*
temperature_k	Process temperature	*K*
pressure_bar	Process pressure	*bar*
pH2_pCO2_ratio	Partial pressure ratio of H_2 to CO_2 (or molar flow rates)	Unitless
GHSV_nlph_gcat	Gas hourly space velocity	$NLh^{-1}g_{cat}^{-1}$
catalyst_load_g	Mass of catalyst in the reactor	*g*
CO2_conversion	Observed carbon dioxide conversion	%
selectivity_CH3OH	Methanol selectivity	%
yield_CH3OH	Methanol yield	%
STY_g_per_gcath	Space-time yield of methanol per gram of catalyst per hour	$g_{CH_3OH}g_{cat}^{-1}h^{-1}$
comment_process	Free-form comments about process parameters	String
comment_catalyst	Free-form comments about the catalyst	String

Table 1. Description of dataset columns. If unit information is present, then the data-type is automatically considered to be `float`.

catalysts falling below a value of 1.2, as shown in Fig. 1(b). Both yield and STY distributions are right-skewed, with the majority of entries concentrated at lower values and a few high-performance outliers. Compared to yield, STY exhibits greater dispersion across the dataset, capturing differences in catalytic performance more effectively. Hence, normalizing yield with respect to catalyst mass and time mitigates inconsistencies arising from varying experimental setups.

Catalyst Composition and Its Influence on STY. Figure 2 summarizes the distribution of active and support components across the dataset. The dataset encompasses a wide range of catalysts, but Cu-based catalysts dominate, highlighting their established role in industrial methanol synthesis[74,75].

Following Cu, In_2O_3-based materials are the next most prominent CO_2 catalysts. As shown in Fig. 2(a),(c), In_2O_3 is reported in two roles: as a primary active phase in some studies[27–29,51,67], and as a support in others[25,33,45,46,53,76]. Other frequently occurring active elements include Zn, Zr, and precious metals such as Pd, Pt, Ir, and Au, although less common.

To visualize how different combinations of materials influence catalytic performance, we present a box plot of STY grouped by pairs of active components in Fig. 3. Each box corresponds to a specific combination of Active Component 1 (*x*-axis) and Active Component 2 (color coded), where the orange bars represent catalysts without Active Component 2.

Cu-based catalysts appear in a spectrum of combinations, including ZnO, K_2CO_3, Pd, and CeO_2. Their STY values exhibit moderate dispersion relative to other combinations, with median STY clustered around 0.046 $g_{CH_3OH}/g_{cat} \cdot h$. In_2O_3-based systems show broader STY distributions and often higher median values, particularly when combined with oxides such as ZnO, La_2O_3, or Co_3O_4.

Several outliers are notable in Fig. 3. For example, the Ir-Pd combination achieves the highest STY in the dataset, although its significance is limited by its rarity. A similar case is observed for the $InNi_3O_{0.5}$ catalyst, which has the second highest STY in the dataset but also appears as a statistical outlier. La-based systems including perovskite materials like $LaMn_xCu_{1-x}O_3$ and $LaMnO_3$ appear with low representation and low STY.

Correlation of Experimental Parameters with STY. Catalytic activity is not solely determined by material composition; process parameters such as temperature, pressure, and GHSV play a major role in determining the efficiency of methanol production. Figure 4 visualizes the combined effect of these process parameters on STY.

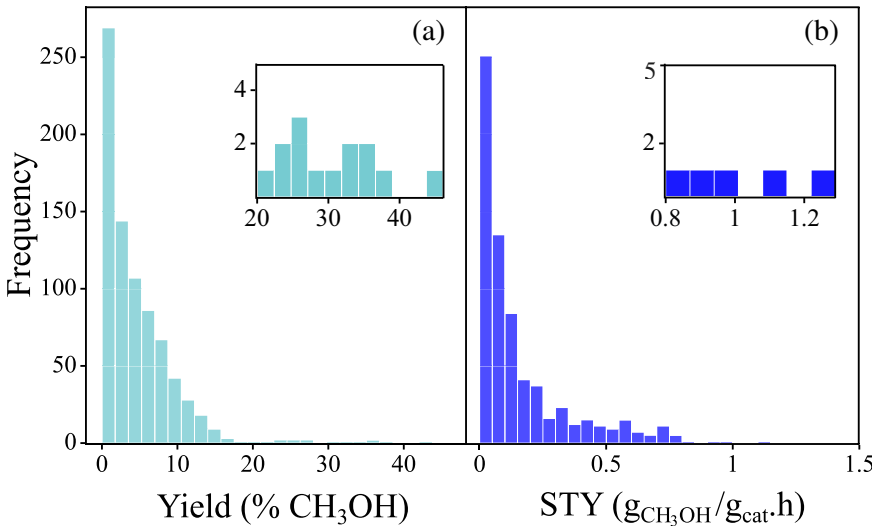

Fig. 1 Distribution of (**a**) the yield and (**b**) the STY across the dataset.

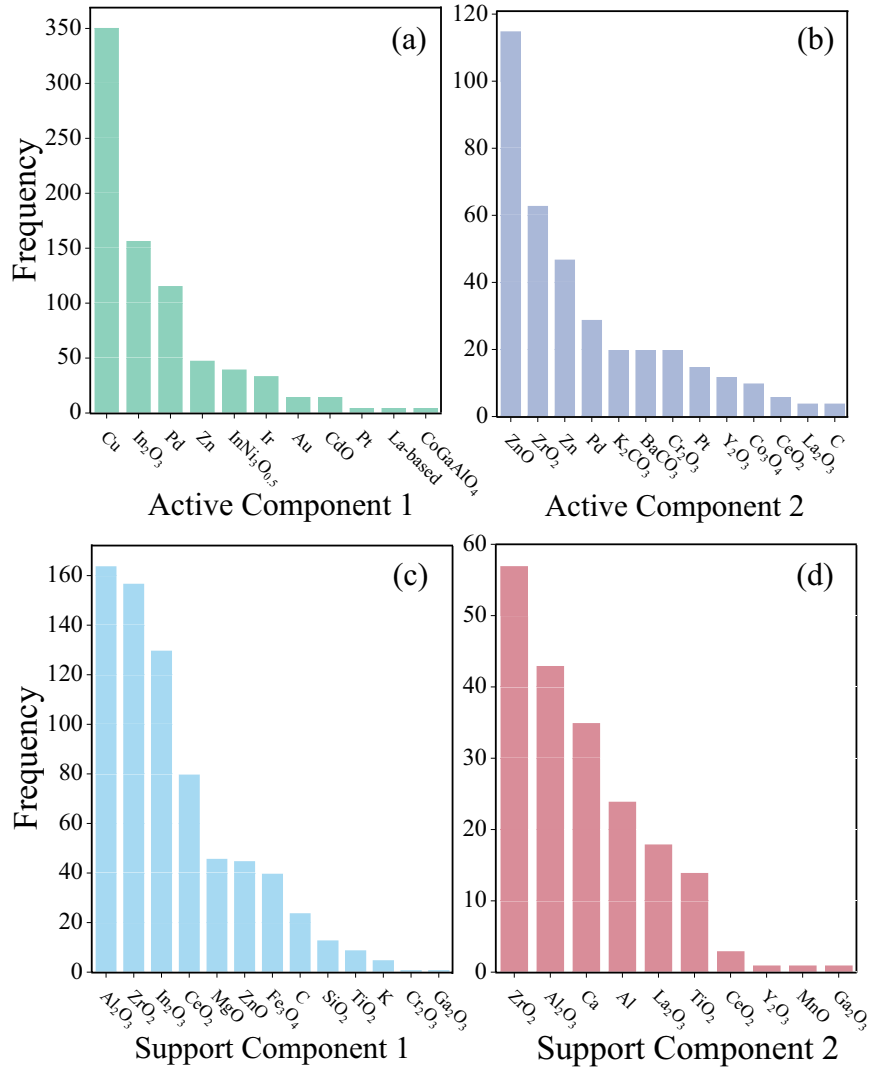

Fig. 2 Distribution of Active and Support components across the dataset. The order (1 or 2) in each category was generally decided, based on Mole percentage.

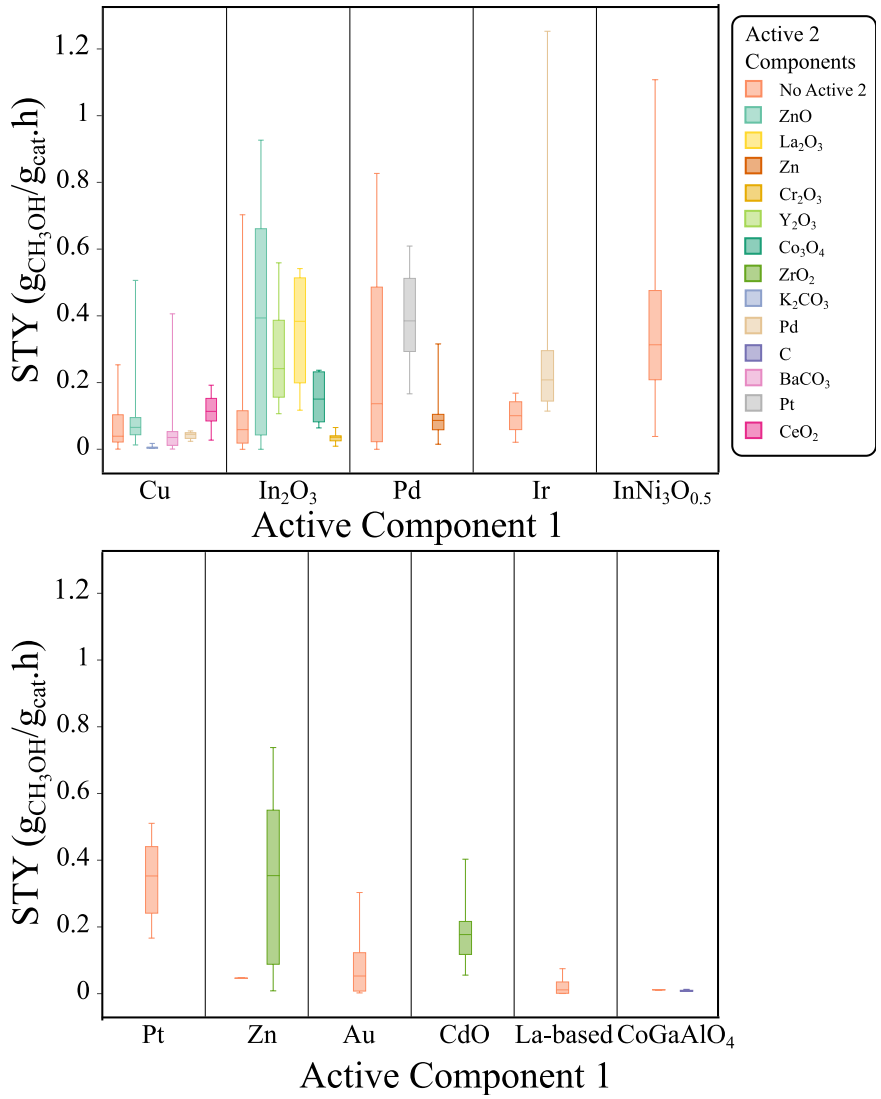

Fig. 3 Box Plot of Space Time Yield (STY) Distribution by Active Components.

Temperature values in the dataset range between 433 K and 673 K, with a denser sampling occurring between 480-540 K. Pressure varies between 0.4 bar and 60 bar, but high-pressure studies above 50 bar are relatively rare. Similarly, GHSV values cover a wide range from 0.87 $nLh^{-1}g^{-1}$ to 240 $nLh^{-1}g^{-1}$, but a significant number of data points cluster below 60 $nLh^{-1}g^{-1}$. Higher STY values are generally observed at moderately elevated temperatures (above 520 K), pressures (above 30 bar), and GHSV values (above 20 $nLh^{-1}g^{-1}$), indicating a positive correlation between these process parameters and catalytic performance.

Technical Validation

To promote consistency and enable robust data-driven analysis, all reported parameters in the dataset were manually extracted and standardized in terms of units and formats. As mentioned in the Data Collection and Processing section, LLM-based methods were not feasible due to the inconsistency of information across sources. Manual extraction ensured alignment with a unified schema, and all entries were independently validated by multiple reviewers to reduce transcription errors and improve data reliability.

As discussed above, the quality of the dataset ultimately depends on the accuracy of the original experimental reports. The accuracy of the reported methanol yield and STY, in particular, is influenced by many experimental factors, including the accuracy of process control and analytical instruments. These sources of uncertainty are rarely disclosed in the literature, yet they can significantly impact reproducibility even under nominally identical conditions.

To assess this variability, we conducted an error analysis based on multiple reported measurements from the same experimental groups, presumably made on the same experimental setup.

Table 2 presents selected datapoints from our corpus where such uncertainty could be estimated. Please note that only a few parameters are listed in the table for clarity and space constraints; other parameters although recorded are omitted here as they remain constant within each experimental group. We calculated the standard

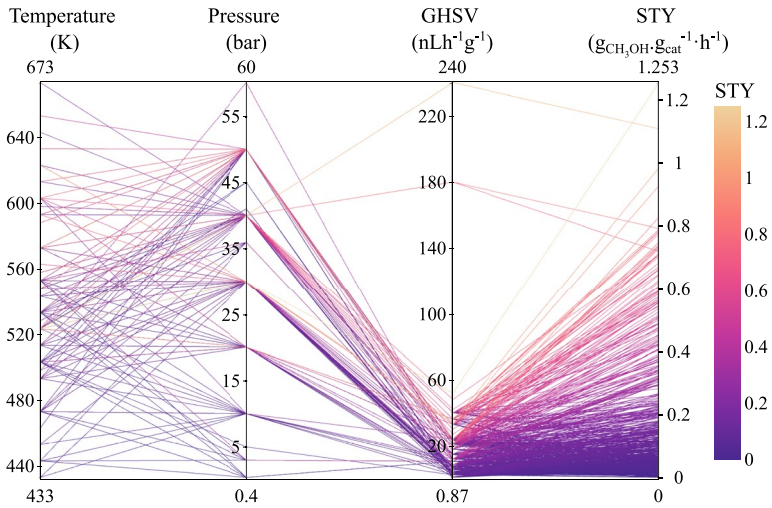

Fig. 4 Parallel Coordinates Plot for Temperature, Pressure and GHSV with STY.

error (σ_Y) of the yields using the formula $\sigma/\sqrt{n}$, where σ represents the standard deviation, and n denotes the number of experiments. The error is often a large fraction of the mean yield ($\overline{Y}$). Possible reasons for the intra-study variability could be either the data-descriptor being agnostic to certain parameters, high sensitivity to experimental parameters or inconsistencies in reporting.

The only instance of inter-laboratory variability is the comparison of In_2O_3(100.0) catalysts from different studies[36,45]. The data shown in Table 3 reveal a discrepancy of approximately 30% in methanol yield under nominally identical temperatures and pressures. The differences in the catalysts' pore sizes and surface area suggest that the preparation procedure may add additional inconsistency to that caused by varying the experimental apparatus and sensitivity to reaction parameters. These aspects highlight the need for standardized experimental protocols and transparent reporting practices in catalysis research.

Overall, the present technical validation provides an initial error analysis and uncertainty structure in the dataset. More approximate comparisons based on coarse groupings could, in principle, provide insight into outliers, which then could be removed for better utilization of the database. However, this falls beyond the scope of this Data Descriptor.

Limitations and Improvement Opportunities. Limitations include those related to the data and those related to the dataset. Assessing the credibility of each literature source was beyond the scope of this study. However, it must be noted that the quality of the dataset is only as good as the quality of the data. Both input data and experimental results suffer from inaccuracies and inconsistencies, as discussed in the Data Collection and Processing section.

The first limitation of the dataset is its limited scope. For example, Cu and In_2O_3-based catalysts are over-represented in the literature. Still, literature on experimental CO_2-based methanol synthesis is growing rapidly, and the dataset could be expanded to several times its current size. As stated above however, the intention of this study is not to gather all data available, but to establish a reporting framework that enables reports in a format conducive to data-driven materials discovery and modern open science practices. Regarding the limited scope of current databases, studies that explore more unconventional material combinations could be particularly valuable additions for data-driven discovery.

Next, data on catalyst synthesis and reduction is not included due to the large variety of methods used and the difficulties in comparing them in a quantitative framework. It is envisioned that a future expansion of this dataset to a *database* could include catalyst *characterization* results in the shape of raw data files (of, e.g., X-ray diffraction spectra) in an open-source format. Similarly, experimental results could be expanded with in-situ analysis data, such as diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS) where available. Finally, catalyst stability, an important factor for industrial applicability, could be included if data existed.

Recommendations. Data-driven materials discovery to accelerate the development of new catalysts for the green transition shows great promise, but requires machine-interpretable experimental data. Therefore, it is encouraged that all experimental data is published in the spirit of open science. This includes data on catalyst preparation, characterization, experimental parameters, and experiment outcomes. Ideally, calibration considerations for key instruments should be stated along with proof of accuracy and precision. For example, many studies claim that the outlet after the back-pressure controller is connected directly to the GC. In such a configuration, the pressure in the GC's sample loop is not fully equilibrated to atmospheric pressure, and a correction factor has to be applied for each pressure and *GHSV*. However, this practicality is not mentioned in any of the sources. In addition, reproducibility of data points should be included in the form of error bars from repeated measurements

Composition	T (K)	P (bar)	n	$\overline{Y}$(%)	σ_Y(%)
Au(1.4)/ In_2O_3(98.6)[45]	473	30	2	0.06	0.01
$CoGaAlO_4$(10.5)/C/ SiO_2(89.5)[48]	543	30	4	1.03	0.18
Cu(50.3)/ In_2O_3(49.7)[53]	493	30	6	5.26	0.29
	513	30	6	5.98	0.37
	533	30	5	6.91	0.60
	553	30	5	7.12	0.58
Cu(63.4)/ ZnO(28.1)/ Al_2O_3(8.5)[59]	513	30	4	29.24	3.10
	533	30	3	28.47	5.08
	523	30	3	25.88	4.98
	513	40	3	35.51	5.18
	513	20	3	16.05	3.04
	503	30	3	20.02	4.15
	493	30	3	15.94	3.67
Cu(18.0)/ Al_2O_3(82.0)[64]	513	30	2	1.79	0.01
Cu(18.1)/ Al_2O_3(78.5)/ ZrO_2(3.4)[64]	513	30	2	1.99	0.11
Cu(18.2)/ Al_2O_3(74.9)/ ZrO_2(6.9)[64]	513	30	2	2.37	0.08
Cu(18.3)/ Al_2O_3(71.3)/ ZrO_2(10.4)[64]	513	30	2	2.26	0.01
In_2O_3/ZnO/C[67]	473	30	3	0.07	0.07
	533	30	3	2.44	0.32
	563	30	3	5.15	0.56
	593	30	3	6.31	0.63
	623	30	3	6.70	0.94
	503	30	3	0.71	0.32
In_2O_3(0.1)/ ZrO_2(99.9)[27]	553	50	2	0.78	0.55
$InNi_3O_{0.5}$ (10.4)/ Fe_3O_4(89.6)[26]	523	40	5	11.70	0.05
	523	40	2	11.10	0.56
Ir(0.65)/ Pd(0.65)/ In_2O_3(98.7)[33]	523	30	4	5.86	0.65
Pd(3.9)/ ZnO(96.1)[23]	523	20	17	2.59	0.65
Pd(0.8)/ ZnO(99.2)[23]	523	20	2	1.00	0.29
Pd(5.5)/Zn(9.0)/ ZrO_2(85.5)[43]	523	30	2	6.96	0.00
Pd(5.4)/Zn(8.9)/ ZrO_2(84.2)/ Ca(1.5)[43]	523	30	2	9.37	0.05
Pd(7.2)/ Zn(11.7)/ CeO_2(79.3)/ Ca(1.8)[42]	503	30	2	9.41	0.01
Zn(13.0)/ ZrO_2(87.0)[56]	573	20	4	1.33	0.55
	603	20	3	4.62	0.03

Table 2. Error analysis of methanol yields for repeated measurements: Catalyst compositions are denoted in the format *Active Component 1*(n_{a_1}) / *Active Component 2*(n_{a_2}) / *Support Component 1*(n_{s_1}) / *Support Component 2*(n_{s_2}), where active components are represented in bold and each component's molar percentage is shown in parentheses. The column T, P and n refers to the reaction temperature, pressure, and the number of measurements, respectively. $\overline{Y}$ is the mean yield of methanol, and σ_Y is the corresponding error in yield.

cit.	Area (m^2/g)	Pore diameter (nm)	GHSV $nLh^{-1}g^{-1}$	Y (%)
[36]	69.4	12.30	20	0.94
[45]	42.3	52.10	15	0.71

Table 3. Inter-laboratory variability for the same catalysts composition (100% In_2O_3), and reaction conditions (Temperature 523 K, pressure 20 bars and H_2 to CO_2 pressure ratio 4). Citation source is abbreviated as cit., area stands for *catalyst surface area*, pore *diameter* is depicting *pore diameter* and Y is *yield*.

under identical conditions. Meta-data should be used to dispel any ambiguity related to units, reference conditions, etc. Data should be complete and supplied in a tabular format, preferably as comma separated value files (CSVs), even for less successful experiments. Experiments with negative outcomes (e.g., no catalyst activity or a catalyst synthesis resulting in an undesired composition) should be reported and enabled by publishers that prioritize scientific rigor over cite-ability.

However, as a first step, colleagues are encouraged to add their experimental results to this dataset, as it is intended to be a living document. Since its publication platform, Zenodo, supports version control, members of

the community are free to add, modify, or correct entries. It can also be expanded with improvement suggestions including but not limited to those listed in the previous section.

Data availability

The TheMeCat dataset was published on Zenodo and is available under https://zenodo.org/records/17085762.

Code availability

This article does not contain code.

Received: 1 October 2025; Accepted: 15 December 2025;
Published online: 03 January 2026

References

1. Aresta, M. & Forti, G. Carbon dioxide as a source of carbon: biochemical and chemical uses, vol. 206 (Springer Science & Business Media, 2012).
2. Olah, G. A., Goeppert, A. & Prakash, G. S. *Beyond oil and gas: the methanol economy* (John Wiley & Sons, 2011).
3. Hu, B., Guild, C. & Suib, S. L. Thermal, electrochemical, and photochemical conversion of CO_2 to fuels and value-added products. *Journal of CO2 Utilization* **1**, 18–27 (2013).
4. Fang, S. *et al*. Photocatalytic CO_2 reduction. *Nature Reviews Methods Primers* **3**, 61 (2023).
5. Tountas, A. A. *et al*. Towards solar methanol: Past, present, and future. *Advanced Science* **6**, 1801903 (2019).
6. Pontzen, F., Liebner, W., Gronemann, V., Rothaemel, M. & Ahlers, B. CO_2-based methanol and DME-efficient technologies for industrial scale production. *Catalysis Today* **171**, 242–250 (2011).
7. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: Status, challenges, and perspectives. *Advanced Science* **6**, 1900808 (2019).
8. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2**, 16028 (2016).
9. Mou, T. *et al*. Bridging the complexity gap in computational heterogeneous catalysis with machine learning. *Nature Catalysis* **6**, 122–136, https://www.nature.com/articles/s41929-023-00911-w (2025).
10. Margraf, J. T., Jung, H., Scheurer, C. & Reuter, K. Exploring catalytic reaction networks with machine learning. *Nature Catalysis* **6**, 112–121, https://www.nature.com/articles/s41929-022-00896-y (2025).
11. Pisal, P., Krejčí, O. & Rinke, P. Machine learning accelerated descriptor design for catalyst discovery in CO_2 to methanol conversion. *npj Computational Materials* **11**, 213 (2025).
12. Takahashi, K., Miyazato, I., Nishimura, S. & Ohyama, J. Unveiling hidden catalysts for the oxidative coupling of methane based on combining machine learning with literature data. *ChemCatChem* **10**, 3223–3228 (2018).
13. Wang, G. *et al*. Accelerated discovery of multi-elemental reverse water-gas shift catalysts using extrapolative machine learning approach. *Nature Communications* **14**, 5861 (2023).
14. Liu, Y. *et al*. Design of CO_2 hydrogenation catalyst by an artificial neural network. *Computers & Chemical Engineering* **25**, 1711–1714 (2001).
15. Bhardwaj, A., Ahluwalia, A. S., Pant, K. K. & Upadhyayula, S. A principal component analysis assisted machine learning modeling and validation of methanol formation over Cu-based catalysts in direct CO_2 hydrogenation. *Separation and Purification Technology* **324**, 124576 (2023).
16. Li, W. *et al*. A machine learning framework for accelerating the development of highly efficient methanol synthesis catalysts. *Journal of Energy Chemistry* **104**, 372–381 (2025).
17. Bahri, S., Pathak, S., Singhahluwalia, A., Malav, P. & Upadhyayula, S. Meta-analysis approach for understanding the characteristics of CO_2 reduction catalysts for renewable fuel production. *Journal of Cleaner Production* **339**, 130653 (2022).
18. Vanjari, P., Kamesh, R. & Rani, K. Machine learning models representing catalytic activity for direct catalytic CO_2 hydrogenation to methanol. *Materials Today: Proceedings* **72**, 524–532 (2023).
19. Suvarna, M., Araújo, T. P. & Pérez-Ramírez, J. A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO_2 hydrogenation. *Applied Catalysis B: Environmental* **315**, 121530 (2022).
20. Suvarna, M., Araujo, T. P. & Perez-Ramirez, J. A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO_2 hydrogenation. *Dataset on Zenodo*, https://doi.org/10.5281/zenodo.6541445 (2022).
21. Toldy, Á. I., Pisal, P., Krejčí, O., Rinke, P. & Santasalo-Aarnio, A. TheMeCat. *Dataset on Zenodo* https://doi.org/10.5281/zenodo.17085762 (2025).
22. Arena, F. *et al*. Synthesis, characterization and activity pattern of cu-zno/ZrO_2 catalysts in the hydrogenation of carbon dioxide to methanol. *Journal of Catalysis* **249**, 185–194 (2007).
23. Bahruji, H. *et al*. Pd/ZnO catalysts for direct CO_2 hydrogenation to methanol. *Journal of Catalysis* **343**, 133–146 (2016).
24. Bansode, A., Tidona, B., von Rohr, P. R. & Urakawa, A. Impact of K and Ba promoters on CO_2 hydrogenation over Cu/Al_2O_3 catalysts at high pressure. *Catal. Sci. Technol.* **3**, 767–778, https://doi.org/10.1039/C2CY20604H (2013).
25. Ding, C. *et al*. Effect of reduction pretreatment on the structure and catalytic performance of Ir-In_2O_3 catalysts for CO_2 hydrogenation to methanol. *Journal of Environmental Sciences* **140**, 2–11 (2024).
26. Meng, C. *et al*. $InNi_3C_{0.5}/Fe_3O_4$ catalyst for the CO_2 hydrogenation to methanol: Co-precipitation preparation, performance and mechanistic insight. *Fuel* **365**, 131111 (2024).
27. Chen, P. *et al*. Morphology-controllable hexagonal-phase indium oxide in situ structured onto a thin-felt Al_2O_3 /al-fiber for the hydrogenation of CO_2 to methanol. *Energy Technology* **7**, 1800747 (2019).
28. Chen, T.-Y. *et al*. Unraveling highly tunable selectivity in CO_2 hydrogenation over bimetallic In-Zr oxide catalysts. *ACS Catalysis* **9**, 8785–8797, https://doi.org/10.1021/acscatal.9b01869 (2019).
29. Chou, C.-Y. & Lobo, R. F. Direct conversion of CO_2 into methanol over promoted indium oxide-based catalysts. *Applied Catalysis A: General* **583**, 117144 (2019).
30. Dasireddy, V. D., Štefančič Neja, S. & Blaž, L. Correlation between synthesis pH, structure and Cu/MgO/Al_2O_3 heterogeneous catalyst activity and selectivity in CO_2 hydrogenation to methanol. *Journal of CO_2 Utilization* **28**, 189–199 (2018).
31. Gao, P. *et al*. Influence of modifier (Mn, La, Ce, Zr and Y) on the performance of Cu/Zn/Al catalysts via hydrotalcite-like precursors for CO_2 hydrogenation to methanol. *Applied Catalysis A: General* **468**, 442–452 (2013).
32. The influence of La doping on the catalytic behavior of Cu/ZrO_2 for methanol synthesis from CO_2 hydrogenation. *Journal of Molecular Catalysis A: Chemical* **345**, 60–68 (2011).
33. Chen, J. *et al*. Photoinduced precise synthesis of diatomic Ir_1Pd_1 – In_2O_3 for CO_2 hydrogenation to methanol via angstrom-scale-distance dependent synergistic catalysis. *Angewandte Chemie International Edition* **63**, e202401168 (2024).
34. Guo, J., Wang, Z., Gao, T. & Wang, Z. Experimental and theoretical study of Pd-Pt/In_2O_3 bimetallic catalysts for enhancing methanol production from CO_2. *Chemical Engineering Journal* **483**, 149370 (2024).

35. Jia, L., Gao, J. & Li, Q. Influence of copper content on structural features and performance of pre-reduced $LaMn_{1-x}Cu_xO_3(0 \leq x < 1)$ catalysts for methanol synthesis from CO_2/H_2. *Journal of Rare Earths* **28**, 747–751 (2010).
36. Jiang, X. *et al*. A combined experimental and dft study of H_2O effect on In_2O_3/ZrO_2 catalyst for CO_2 hydrogenation to methanol. *Journal of Catalysis* **383**, 283–296 (2020).
37. Daifeng, L. *et al*. The Co-In_2O_3 interaction concerning the effect of amorphous Co metal on CO_2 hydrogenation to methanol. *Journal of CO_2 Utilization* **65**, 102209 (2022).
38. Lei, H., Hou, Z. & Xie, J. Hydrogenation of CO_2 to CH_3OH over $CuO/ZnO/Al_2O_3$ catalysts prepared via a solvent-free routine. *Fuel* **164**, 191–198 (2016).
39. Lin, F. *et al*. CO_2 hydrogenation to methanol over bimetallic Pd-Cu catalysts supported on $TiO_2 - CeO_2$ and $TiO_2 - ZrO_2$. *Catalysis Today* **371**, 150–161 (2021).
40. Liu, C. *et al*. Methanol synthesis from CO_2 hydrogenation over copper catalysts supported on MgO-modified TiO_2. *Journal of Molecular Catalysis A: Chemical* **425**, 86–93 (2016).
41. Ma, Y. *et al*. A practical approach for the preparation of high activity $Cu/ZnO/ZrO_2$ catalyst for methanol synthesis from CO_2 hydrogenation. *Applied Catalysis A: General* **171**, 45–55 (1998).
42. Malik, A. S. *et al*. Development of highly selective $PdZn/CeO_2$ and Ca-doped $PdZn/CeO_2$ catalysts for methanol synthesis from CO_2 hydrogenation. *Applied Catalysis A: General* **560**, 42–53 (2018).
43. Malik, A. S. *et al*. Selective hydrogenation of CO_2 to CH_3OH and in-depth DRIFT analysis for $PdZn/ZrO_2$ and $CaPdZn/ZrO_2$ catalysts. *Catalysis Today* **357**, 573–582 (2020).
44. Ota, A. *et al*. Comparative study of hydrotalcite-derived supported Pd_2Ga and PdZn intermetallic nanoparticles as methanol synthesis and methanol steam reforming catalysts. *Journal of Catalysis* **293**, 27–38 (2012).
45. Hou, R., Xiao, J., Wu, Q., Zhang, T. & Wang, Q. Boosting oxygen vacancies by modulating the morphology of Au decorated In_2O_3 with enhanced CO_2 hydrogenation activity to CH_3OH. *Journal of Environmental Sciences* **140**, 91–102 (2024).
46. Rui, N. *et al*. CO_2 hydrogenation to methanol over Pd/In_2O_3: effects of Pd and oxygen vacancy. *Applied Catalysis B: Environmental* **218**, 488–497 (2017).
47. Ghosh, S., Olsson, L. & Creaser, D. Methanol mediated direct CO_2 hydrogenation to hydrocarbons: Experimental and kinetic modeling study. *Chemical Engineering Journal* **435**, 135090 (2022).
48. Zhang, S. *et al*. Carbon modified active pairs of $Co - Co_2C$ for CO_2 hydrogenation to alcohols. *Chemical Engineering Journal* **486**, 150334 (2024).
49. Zaman, S. F. *et al*. Elucidating the promoting role of Ca on $PdZn/CeO_2$ catalyst for CO_2 valorization to methanol. *Fuel* **343**, 127927 (2023).
50. Samson, K. *et al*. Influence of ZrO_2 structure and copper electronic state on activity of Cu/ZrO_2 catalysts in methanol synthesis from CO_2. *ACS Catalysis* **4**, 3730–3741 (2014).
51. Sharma, P., Hoang Ho, P., Shao, J., Creaser, D. & Olsson, L. Role of ZrO_2 and CeO_2 support on the In_2O_3 catalyst activity for CO_2 hydrogenation. *Fuel* **331**, 125878 (2023).
52. Shi, Z., Tan, Q. & Wu, D. Ternary copper-cerium-zirconium mixed metal oxide catalyst for direct CO_2 hydrogenation to methanol. *Materials Chemistry and Physics* **219**, 263–272 (2018).
53. Shi, Z. *et al*. CO_2 hydrogenation to methanol over Cu-In intermetallic catalysts: Effect of reduction temperature. *Journal of Catalysis* **379**, 78–89 (2019).
54. Shi, G. *et al*. Morphology effect of ZnO support on the performance of Cu toward methanol production from CO_2 hydrogenation. *Journal of Saudi Chemical Society* **24**, 42–51 (2020).
55. Song, J. *et al*. The role of Al doping in Pd/ZnO catalyst for CO_2 hydrogenation to methanol. *Applied Catalysis B: Environmental* **263**, 118367 (2020).
56. Wang, J. *et al*. A highly selective and stable $ZnO-ZrO_2$ solid solution catalyst for CO_2 hydrogenation to methanol. *Science Advances* **3**, e1701290, https://doi.org/10.1126/sciadv.1701290 (2017).
57. Wang, Y. H. *et al*. Structure-activity relationships of $Cu-ZrO_2$ catalysts for CO_2 hydrogenation to methanol: interaction effects and reaction mechanism. *RSC Adv.* **7**, 8709–8717, https://doi.org/10.1039/C6RA28305E (2017).
58. Wang, W., Qu, Z., Song, L. & Fu, Q. Effect of the nature of copper species on methanol synthesis from CO_2 hydrogenation reaction over $cuo/Ce_{0.4}Zr_{0.6}O_2$ catalyst. *Molecular Catalysis* **493**, 111105 (2020).
59. Wu, W., Xie, K., Sun, D., Li, X. & Fang, F. $CuO/ZnO/Al_2O_3$ catalyst prepared by mechanical-force-driven solid-state ion exchange and its excellent catalytic activity under internal cooling condition. *Industrial & Engineering Chemistry Research* **56**, 8216–8223, https://doi.org/10.1021/acs.iecr.7b01464 (2017).
60. Ma, X. *et al*. $Zn - CdZrO_x$ solid solution catalysts for hydrogenation of CO_2 to methanol. *Fuel* **346**, 128376 (2023).
61. Xiao, S. *et al*. Highly efficient Cu-based catalysts via hydrotalcite-like precursors for CO_2 hydrogenation to methanol. *Catalysis Today* **281**, 327–336 (2017).
62. Yang, Y., Guo, M. & Zhao, F. Cr_2O_3 promoted In_2O_3 catalysts for CO_2 hydrogenation to methanol. *ChemPhysChem* **25**, e202300530, https://doi.org/10.1002/cphc.202300530 (2024).
63. Yang, C. *et al*. Methanol synthesis from CO_2-rich syngas over a ZrO_2 doped CuZnO catalyst. *Catalysis Today* **115**, 222–227 (2006).
64. Zhang, Y., Fei, J., Yu, Y. & Zheng, X. Methanol synthesis from CO_2 hydrogenation over Cu based catalyst supported on zirconia modified γ-Al_2O_3. *Energy Conversion and Management* **47**, 3360–3367 (2006).
65. Zhang, C. *et al*. Preparation and CO_2 hydrogenation catalytic properties of alumina microsphere supported Cu-based catalyst by deposition-precipitation method. *Journal of CO_2 Utilization* **17**, 263–272 (2017).
66. Zhuang, H., Bai, S., Liu, X. & Yan, Z. Structure and performance of Cu/ZrO_2 catalyst for the synthesis of methanol from CO_2 hydrogenation. *Journal of Fuel Chemistry and Technology* **38**, 462–467 (2010).
67. Shi, Y. *et al*. Carbon coated In_2O_3 hollow tubes embedded with ultra-low content ZnO quantum dots as catalysts for CO_2 hydrogenation to methanol. *Journal of Colloid and Interface Science* **636**, 141–152 (2023).
68. Sipilä, M., Mehryary, F., Pyysalo, S., Ginter, F. & Todorović, M. Annotated textual dataset PV600 of perovskite bandgaps for information extraction from literature. *Scientific Data* **12**, 1401, https://doi.org/10.1038/s41597-025-05637-x (2025).
69. Chen, X. *et al*. Large language model enhanced corpus of CO_2 reduction electrocatalysts and synthesis procedures. *Scientific Data* **11**, 347, https://doi.org/10.1038/s41597-024-03180-9 (2024).
70. Dagdelen, J. *et al*. Structured information extraction from scientific text with large language models. *Nature Communications* **15**, 1418, https://www.nature.com/articles/s41467-024-45563-x (2025).
71. ISO. Determination of the specific surface area of solids by gas adsorption – Bet method. Standard ISO 9277:2022 (E) https://www.iso.org/standard/71014.html (International Organization for Standardization, Geneva, CH, 2022).
72. Of Pure, I. U. & Chemistry, A. Iupac compendium of chemical terminology – the gold book http://goldbook.iupac.org/ (2009).
73. Grey, R. & Beddow, J. On the Hausner ratio and its relationship to some properties of metal powders. *Powder Technology* **2**, 323–326 (1969).
74. Niu, J. *et al*. Comprehensive review of Cu-based CO_2 hydrogenation to CH_3OH: Insights from experimental work and theoretical analysis. *International Journal of Hydrogen Energy* **47**, 9183–9200 (2022).
75. Amann, P. *et al*. The state of zinc in methanol synthesis over a Zn/ZnO/Cu(211) model catalyst. *Science* **376**, 603–608, https://doi.org/10.1126/science.abj7747 (2022).
76. Wang, J. *et al*. CO_2 hydrogenation to methanol over In_2O_3 -based catalysts: From mechanism to catalyst development. *ACS Catalysis* **11**, 1406–1423, https://doi.org/10.1021/acscatal.0c03665 (2021).

Acknowledgements
The authors would like to thank Lotta Järvi, Markus Laitinen, Pawel Winiarski, and Lakshmi Priya Aravamoudane for their help with compiling the dataset. This project has received funding from the European Union - NextGenerationEU instrument and is funded by the Research Council of Finland under grant numbers 348179 and 348180 (AIcon - AI-guided CO_2 conversion).

Author contributions
Á.T.: Conceptualization, Methodology, Investigation, Data Curation, Writing-Original Draft, Project Administration, Funding Acquisition. P.P.: Conceptualization, Methodology, Investigation, Formal Analysis, Data Curation, Validation, Visualization, Writing-Original Draft. O.K.: Conceptualization, Methodology, Supervision, Validation, Writing-Review and Editing. P.R.: Conceptualization, Writing-Review and Editing, Supervision, Project Administration, Funding Acquisition. A.S.-A.: Conceptualization, Supervision, Project Administration, Funding acquisition, Writing-Review and Editing.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.S.-A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.