

Collection, processing, and sharing of a dataset for the selective hydrogenation of benzene to cyclohexene

SUN Chao^{1,2}, ZHANG Bin^{1,*}

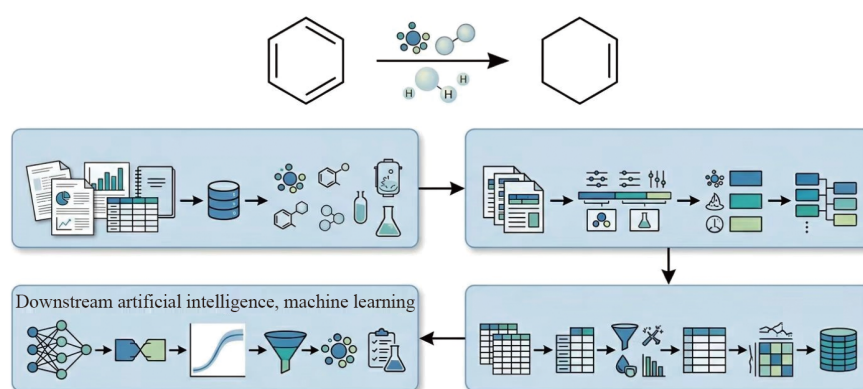
¹ State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan 030001, China;

² Center for Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Cyclohexene is an important raw material for nylon production, and the selective hydrogenation of benzene is a key route for preparing cyclohexene. To promote data sharing and reuse in this field, we collected and standardized experimental data on the hydrogenation of benzene to cyclohexene from

publicly available literature and constructed a comprehensive dataset containing catalyst composition, reaction conditions, and reaction results (conversion, selectivity and yield). This data descriptor details the source, field definitions, generation and processing workflow, quality control, sharing approach and usage recommendations of the dataset, with the aim of providing a reusable data foundation for subsequent statistical analysis, machine learning modeling, experimental design, and catalyst screening.

Key words: selective hydrogenation of benzene; cyclohexene; dataset; catalyst; machine learning



The selective hydrogenation of benzene to cyclohexene is a reaction of considerable industrial value. Given that complete hydrogenation to cyclohexane is more favorable thermodynamically, optimizing catalyst structure and process conditions is practically required to regulate reaction kinetics and improve cyclohexene yield. Existing ruthenium-based catalytic systems often rely on aqueous-phase systems and inorganic salt additives to enhance cyclohexene selectivity, but the cyclohexene yield is still limited. Meanwhile, the benzene hydrogenation system involves complex factors in heterogeneous catalysis, and obtaining a large amount of high-quality and comparable data is costly. Experimental data scattered across the literature provide a basis for data-driven correlation analysis

and predictive modeling, but these data usually lack unified field definitions, unit systems, and organizational formats. Therefore, constructing and publicly releasing a reusable, searchable, clearly structured and quality-controlled dataset is of great significance for advancing research on the selective hydrogenation of benzene to cyclohexene.

1 Basic information on the dataset

The basic information of the dataset is summarized in Table 1. The table provides an overview of the data title, version, authorship, data content, generation time, volume, format, repository location, funding information, and citation format of the data.

Received: 2026-01-20; Revised: 2026-04-03

Foundation items: Supported by Strategic Priority Research Program of the Chinese Academy of Sciences (XDB 1190000) and the General Program of the National Natural Science Foundation of China (22572209).

* Corresponding author. E-mail: zhangbin2009@sxicc.ac.cn.

DOI: 10.1016/S1872-5813(26)60748-7 CSTR: 32157.14.2097-213X.2026.JFCT.0108 Associated Data DOI: 10.57760/sciencedb.27385

本文的英文电子版由 Elsevier 出版社在 ScienceDirect 上出版 (<http://www.sciencedirect.com/science/journal/18725813>)

Table 1 Overview of basic information of the dataset

Item	Description
Data title	Experimental dataset for the selective hydrogenation of benzene to cyclohexene
Data version	V1.0 (compiled and standardized in 2025)
Author's name	SUN Chao, ZHANG Bin
Corresponding author's name	ZHANG Bin
Data content description	Each record corresponds to one experimental combination and contains the catalyst preparation method, catalyst composition (Ru, Cu, ZnO, NaOH, ZnSO ₄ , AlOOH, SiO ₂ , TiO ₂ , etc.), reaction conditions (temperature, pressure, reaction time, feed ratio, etc.), and performance indicators (conversion, selectivity, and yield).
Data generation time	The data were derived from published literature reports, and the present compilation and standardized processing were completed in 2025.
Data volume	178 experimental records covering catalyst preparation methods, catalyst compositions, reaction conditions, pore-structure descriptors, and performance outputs
Data format	Microsoft Excel workbook (.xlsx)
Data source location	GitHub: https://github.com/ttoono-takaki/Code-and-dataset.git ; Science Data Bank: https://doi.org/10.57760/sciencedb.27385
Funding information	Strategic Priority Research Program of the Chinese Academy of Sciences (XDB1190000), and the General Program of the National Natural Science Foundation of China (22572209)
Citation format of data	SUN Chao, ZHANG Bin. Experimental dataset for the selective hydrogenation of benzene to cyclohexene[J]. J Fuel Chem Technol, 2026, 54(7): 20260027. SUN Chao, ZHANG Bin. Experimental dataset for the selective hydrogenation of benzene to cyclohexene[DS/OL]. V1. Science Data Bank, 2025[2026-06-05]. https://doi.org/10.57760/sciencedb.27385 . DOI:10.57760/sciencedb.27385.

2 Value of the dataset

This dataset structurally organizes experimental data on the hydrogenation of benzene to cyclohexene scattered in the literature^[1–10], making it convenient to directly conduct: (1) statistical analysis and visual comparison (Figure 1); (2) mining of key influencing factors and correlation analysis; (3) machine learning/regression modeling; (4) data-driven experimental design and condition screening. Because the data come from different research groups, users should pay attention to the biases caused by differences in experimental apparatus, analytical methods, catalyst characterization methods and related factors when making cross-source comparisons or using the data for scale-up.

3 Description of data generation methods or process

3.1 Literature screening and inclusion criteria

The literature on the selective hydrogenation of benzene to cyclohexene spans a wide time range and shows significant differences in conditions, and some studies introduce relatively uncommon additives or supports to regulate a single indicator. Such “low-frequency factors” may become interfering variables in data-driven modeling. To improve the comparability and stability of the dataset, we systematically reviewed nearly 50 related publications during the preliminary screening stage and preferentially

selected papers that could completely cover the preset descriptor system and contained a sufficient number of data points, thereby ensuring the reliability of the original data and minimizing the bias caused by non-standardizable factors as much as possible. At the same time, in order to keep the distribution of the target variables relatively balanced within the sample space, we further screened the candidate literature and organized the extracted data into a unified table structure. Because this dataset mainly serves subsequent work, we selected only reactions carried out in autoclaves, and the specific reference information for each data entry can be found in the supplementary file dataset.xlsx. According to our needs, we restricted the metal species, support species and additive species of the catalysts in the preliminarily screened papers, then listed all descriptors appearing in these papers, removed descriptors with incomplete information, and finally retained only the literature containing complete data for all descriptors. The final dataset contains $N = 178$ experimental records and 26 descriptors, covering three preparation methods (Impregnation, Co-precipitation and Deposition-precipitation).

3.2 Definition of target variables and unified criteria

To convert performance indicators reported in different publications into unified fields that are computable and comparable, this dataset uses benzene conversion (x), cyclohexene selectivity (s) and cyclohexene yield (Y) as output variables and uniformly calculates and records them according to the definitions given in the original papers.

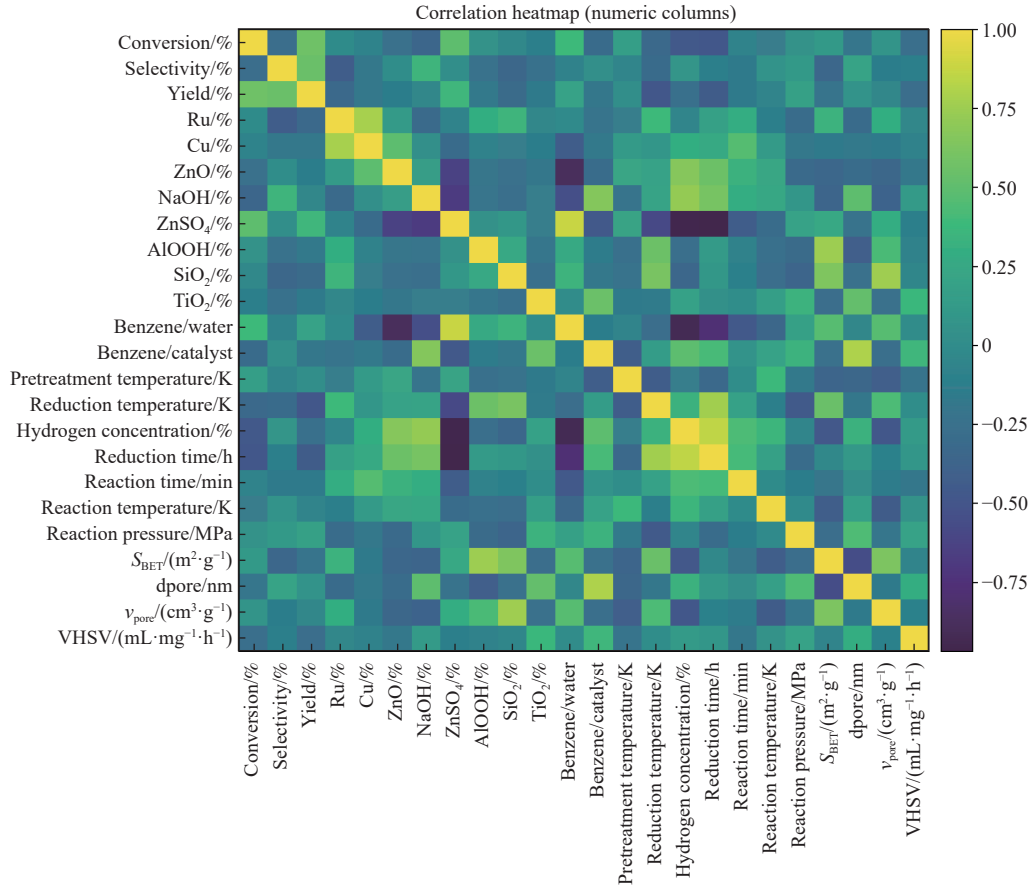


Fig. 1 Correlation heatmap of all numerical columns

$$x_{C_6H_6}(\%) = \frac{n_{C_6H_6, in} - n_{C_6H_6, out}}{n_{C_6H_6, in}} \times 100\% \quad (1)$$

$$s_{C_6H_8}(\%) = \frac{n_{C_6H_8, formed}}{\sum n_{products from C_6H_6}} \times 100\% \quad (2)$$

$$Y_{C_6H_8}(\%) = \frac{n_{C_6H_8, formed}}{n_{C_6H_6, in}} \times 100\% \quad (3)$$

During the data curation process, the following principles were applied:

- (1) If the literature directly reports x , s and Y , they are entered as reported and the source is indicated.
- (2) If the literature reports only some of the indicators (for example, only x and s), Y is calculated according to the above definitions.

3.3 Descriptor system design and field extraction

Before data extraction, we first determined which information was both important and easy to standardize, thereby defining 24 input descriptors covering reaction conditions, pretreatment/reduction conditions, catalyst composition, pore-structure characterization, and related information.

These descriptors include core variables such as reaction temperature, reaction pressure, reaction time, preparation method, and catalyst composition, and they are used as fixed fields of the dataset to enable structured alignment across different publications.

3.4 Unit system and rules for handling missing information

Because recording practices differ across publications, key fields were subjected to unit unification and rule-based filling during data collation:

(1) Temperature was standardized to K, pressure to MPa, reaction time to min, reduction time to h and composition to % (method used for calculating composition: for the catalyst portion, both metals and supports were calculated according to the amount added rather than the actual loading; additives, reactants and solvents were also uniformly calculated according to the amount added). A total of three preparation methods were recorded, and one-hot encoding is recommended when using machine learning and similar methods.

(2) For catalyst samples explicitly reported in the literature as having no reduction step, the reduction temperature was uniformly set to 298 K to ensure field completeness and consistency of the table structure. Regarding hydrogen content, in some publications the catalyst was directly reduced in the autoclave with pure hydrogen, and we manually defined the hydrogen concentration as 99.9%. In some publications no reduction was performed, and we manually defined the hydrogen concentration as 0%. In subsequent machine learning analysis, one-hot encoding can be used to distinguish these cases.

(3) For other fields that are missing in the literature but can reasonably be judged as “not applicable/not

performed,” a consistent placeholder strategy (for example, 0 or NA, depending on the field type) was adopted and the reason was recorded in the remarks column to ensure that subsequent modeling can distinguish “true zero” from “missing/not applicable.”

4 Description of data examples

To facilitate understanding of the data organization, Table 2 gives examples of the variables in the dataset and their value ranges.

The final dataset contains $N = 178$ experimental records for the selective hydrogenation of benzene to cyclohexene, covering three catalyst preparation methods: Impregnation (108 records), Co-precipitation (39 records), and Deposition-precipitation (31 records). In terms of reaction conditions, the reaction temperature ranges from 373 to 453 K, the reaction pressure from 2 to 7 MPa, and the reaction time from 1 to 720 min, with a median of 30 min (interquartile range, 11–58 min) and a mean of 51.55 min, indicating a certain right-skewed distribution. Regarding the catalyst composition, the Ru content ranges from 0.18% to 20%, with a median of 1.32% (interquartile range, 0.95%–2.50%).

Table 2 Examples of the main variables in the dataset and their value ranges

Variable	Unit/Type	Example value range
Benzene/water	mL/mL	0.175–0.5
Benzene/catalyst	mL/mg	0.05–0.3846
Pretreatment temperature	K	298–873
Reduction temperature	K	298–573
Hydrogen concentration	%	0/5/99.9
Reduction time	h	0–3
Reaction temperature	K	373–453
Reaction time	min	1–720
Reaction pressure	MPa	2–7
S_{BET}	m^2/g	8.4–318
Dpore	nm	2.63–42.2
V_{pore}	cm^3/g	0.03–1.05
VHSV	$\text{mL}/(\text{mg}\cdot\text{h})$	0.00058–11.538
Ru	%	0.18–20
Cu	%	0–1.26
ZnO	%	0–97.93
NaOH	%	0–93.26
ZnSO ₄	%	0–90.99
AlOOH	%	0–43.68
SiO ₂	%	0–43.68
TiO ₂	%	0–97.1
Selectivity	%	0.7–89.3
Yield	%	0.3–60.64
Conversion	%	4.04–98.58

As for the performance indicators, benzene conversion ranges from 4.04% to 98.58%, with a median of 40.35% (interquartile range, 30.33%–70.99%); cyclohexene selectivity ranges from 0.7% to 89.3%, with a median of 46.54% (interquartile range, 28.83%–65.60%); and cyclohexene yield ranges from 0.3% to 60.64%, with a median of 20.88% (interquartile range, 11.60%–28.11%). To more intuitively present the data coverage and density, histograms of the target variables and key reaction conditions are provided in Figure S1–S5 and scatter plots of yield versus temperature, pressure, Ru content, and conversion/selectivity are also plotted, together with an overview of correlations among numerical variables, to show the distribution characteristics and coverage of data points under different combinations of variables.

5 Description of data quality control

(1) Completeness check

Records with missing key information or unclear descriptions were removed.

(2) Consistency calibration

The units and expressions of all descriptors were unified. For conversion, selectivity, and yield, because the literature often reports only part of these indicators, we converted them using Eqs. (1), (2) and (3).

(3) Outlier screening

The distributions of key fields were examined, and obvious outliers were checked against the original literature and then retained or removed as appropriate. For example, for the relationship among benzene conversion, cyclohexene selectivity, and cyclohexene yield in the corresponding literature, if calculations showed that the data were not self-consistent, the data could be recalculated and corrected; if key data were missing and calculation was impossible, that set of data could be considered for removal.

6 Methods and suggestions for data use

When carrying out modeling or comparative analysis, users are advised to: (1) clarify the definitions and criteria of inputs and outputs, especially derived indicators such as feed ratio and space velocity; (2) perform necessary standardization/normalization of numerical features and use cross-validation to evaluate model generalization ability; (3) consider the systematic bias that may be introduced by differences in catalyst pretreatment, analytical methods, and apparatus when comparing data across publications; (4) if the dataset needs to be expanded, give priority to supplementing new literature data consistent with the existing field system and record clear source identifiers.

7 Conclusion

This paper presents a comprehensive experimental

dataset for the selective hydrogenation of benzene to cyclohexene and provides a description of the data source, field definitions, standardized processing, and quality-control procedures. The dataset can provide a reusable data foundation for subsequent statistical analysis, data mining, machine learning modeling, and experimental condition screening.

Data Availability Statement

The data files are provided in Excel (.xlsx) format and can be downloaded from the following sources: <https://github.com/ttoono-takaki/Code-and-dataset.git> or <https://doi.org/10.57760/sciencedb.27385>.

Declaration of Competing Interest

The authors declare that there is no conflict of interest affecting the public release and sharing of the data presented in this paper.

References

- [1] LIU H, LIANG S, WANG W, et al. The partial hydrogenation of benzene to cyclohexene over Ru-Cu catalyst supported on ZnO[J]. J Mol Catal A: Chem, 2011, 341: 35–41.
- [2] ZHOU G, DONG Y, JIANG L, et al. Effect of support composition on the structural and catalytic properties of Ru/AlOOH-SiO₂ catalysts for benzene selective hydrogenation[J]. Catal Sci Technol, 2018, 8: 1435–1446.
- [3] WU T, ZHANG P, JIANG T, et al. Enhancing the selective hydrogenation of benzene to cyclohexene over Ru/TiO₂ catalyst in the presence of a very small amount of ZnO[J]. Sci China: Chem, 2014, 58: 93–100.
- [4] HAO F, ZHENG J, OUYANG D, et al. Selective hydrogenation of benzene over Ru supported on surface modified TiO₂[J]. Korean J Chem Eng, 2021, 38: 736–746.
- [5] ZHOU G, WANG F, SHI R. Nanoparticulate Ru on morphology-manipulated and Ti³⁺ defect-riched TiO₂ nanosheets for benzene semi-hydrogenation[J]. J Catal, 2021, 398: 148–160.
- [6] ZHOU G, DOU R, BI H, et al. Ru nanoparticles on rutile/anatase junction of P25 TiO₂: Controlled deposition and synergy in partial hydrogenation of benzene to cyclohexene[J]. J Catal, 2015, 332: 119–126.
- [7] YAN X, ZHANG Q, ZHU M, et al. Selective hydrogenation of benzene to cyclohexene over Ru–Zn/ZrO₂ catalysts prepared by a two-step impregnation method[J]. J Mol Catal A: Chem, 2016, 413: 85–93.
- [8] WANG Z, ZHANG Q, LU X, et al. Ru–Zn catalysts for selective hydrogenation of benzene using coprecipitation in low alkalinity[J]. Chin J Catal, 2015, 36: 400–407.
- [9] SUN H, WANG H, JIANG H, et al. Effect of (Zn(OH)₂)₃(ZnSO₄)(H₂O)₅ on the performance of Ru–Zn catalyst for benzene selective hydrogenation to cyclohexene[J]. Appl Catal A: Gen, 2013, 450: 160–168.
- [10] XUE X, LIU J, RAO D, et al. Double-active site synergistic catalysis in Ru–TiO₂ toward benzene hydrogenation to cyclohexene with largely enhanced selectivity[J]. Catal Sci Technol, 2017, 7: 650–657.

苯选择性加氢制环己烯反应数据集的收集、处理与共享

孙超^{1,2}, 张斌^{1,*}

- (1. 中国科学院山西煤炭化学研究所 煤炭高效低碳利用全国重点实验室, 山西 太原 030001;
2. 中国科学院大学 材料科学与光电工程中心, 北京 100049)

摘要: 环己烯是尼龙生产的重要原料, 而苯的选择性加氢是制备环己烯的关键途径。为促进该领域的数据共享与复用, 本工作从公开发表文献中收集并标准化整理苯加氢生成环己烯反应的实验数据, 构建了一个包含催化剂组成、反应条件与反应结果(转化率、选择性、产率)的综合数据集。本文以数据论文形式介绍数据集的来源、字段定义、生成与处理流程、质量控制、共享方式和使用建议, 旨在为后续的统计分析、机器学习建模、实验设计与催化剂筛选提供可复用的数据基础。

关键词: 苯加氢; 环己烯; 数据集; 催化剂; 机器学习

中图分类号: O643; TP181.1

文献标识码: A

作者简介:



孙超(2001), 硕士研究生, 研究方向为多相催化, sunchao23@mails.ucas.ac.cn。



张斌(1986), 博士, 研究员, 研究方向为原子层沉积及能源催化, zhangbin2009@sxicc.ac.cn。