

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| ☐ | ☒ | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

The experimental data for both lab-scale and pilot-scale were collected from our research group's FCC unit. The data used to train the lab-scale hybrid model were generated by a mechanistic model. All data used for model training and prediction in this study are publicly available at: <https://github.com/chenzhengyuAI/complex-reaction-system-scale-up>.

Data analysis

The AI model was developed using the PyTorch framework in Python, while all data analysis and visualization were performed using MATLAB 2020a. The code of the hybrid model and all scripts used to train and predict product distribution are publicly available at: <https://github.com/chenzhengyuAI/complex-reaction-system-scale-up>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data used for model training and prediction in this study are publicly available at: <https://github.com/chenzhengyuAI/complex-reaction-system-scale-up>.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

N/A

Reporting on race, ethnicity, or other socially relevant groupings

N/A

Population characteristics

N/A

Recruitment

N/A

Ethics oversight

N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences

☐ Behavioural & social sciences

☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Research Design:

1. Develop a molecular-level kinetic model for catalytic cracking based on existing laboratory-scale experimental data.
 2. Generate 10,000 datasets from the mechanistic model to train neural networks
 3. Fine-tune the neural networks using pilot-scale data to enable accurate prediction of product distributions in pilot-scale units.
- All data were obtained in triplicate.

Research sample

The study focuses on naphtha catalytic cracking from a laboratory-scale to a pilot-scale plant.

This sample was selected based on two key considerations:

1. Naphtha represents the C3-C12 petroleum fraction with inherently complex molecular compositions.
 2. There is a significant difference between the laboratory-scale and pilot-scale reactors used in naphtha catalytic cracking. The laboratory-scale and pilot-scale models need to be rebuilt.
- This research sample is suitable to validate our proposed model.

Sampling strategy

1. Investigated typical naphtha catalytic cracking parameters (reaction temperature, catalyst-to-oil ratio, WHSV) to define operational bounds. Generated 10,000 process condition datasets through normal distribution sampling within these ranges.
 2. Analyzed molecular compositions of conventional catalytic cracking naphtha. Introduced controlled noise to representative compositions to produce 10,000 distinct feedstock compositions.
 3. Processed all condition-feedstock pairs (n=10,000) through molecular-level kinetic models to calculate product distributions, obtaining the dataset.
- Some published studies show that 10,000 sample sets are sufficient to train a robust neural network for the reaction system.

Data collection

The data was collected by Zhengyu Chen. The experimental data were recorded by pen and paper, and the data generated by the model were collected in an Excel file.

Timing and spatial scale

The dataset was generated on August 6, 2024, with all 10,000 data points produced in a single batch. No temporal intervals or spatial scales were involved in the data generation process.

Data exclusions	No data were excluded.
Reproducibility	We used the developed model to calculate the results more than three times, and each time we obtained the same results, proving the repeatability of the model's experimental results.
Randomization	This study addresses the challenge of complex reaction systems scale-up by integrating molecular conversion mechanisms with transfer learning, without involving covariates.
Blinding	The data were grouped randomly, and the researchers were blind to the grouping.
Did the study involve field work?	☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involvement in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A