Electronic supplementary information

Can large language models predict the hydrophobicity of metal–organic frameworks?

Xiaoyu Wu and Jianwen Jiang*

Department of Chemical and Biomolecular Engineering, National University of Singapore, 117576, Singapore

*E-mail: chejj@nus.edu.sg

Table of Contents

1. Data distribution

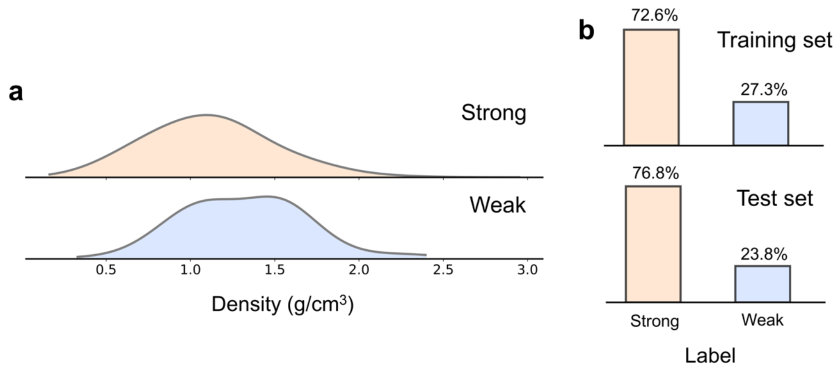

Fig. S1. (a) Kernel density estimations of binary classification versus density. The vertical axis denotes probability. (b) Data distribution in training and test sets.

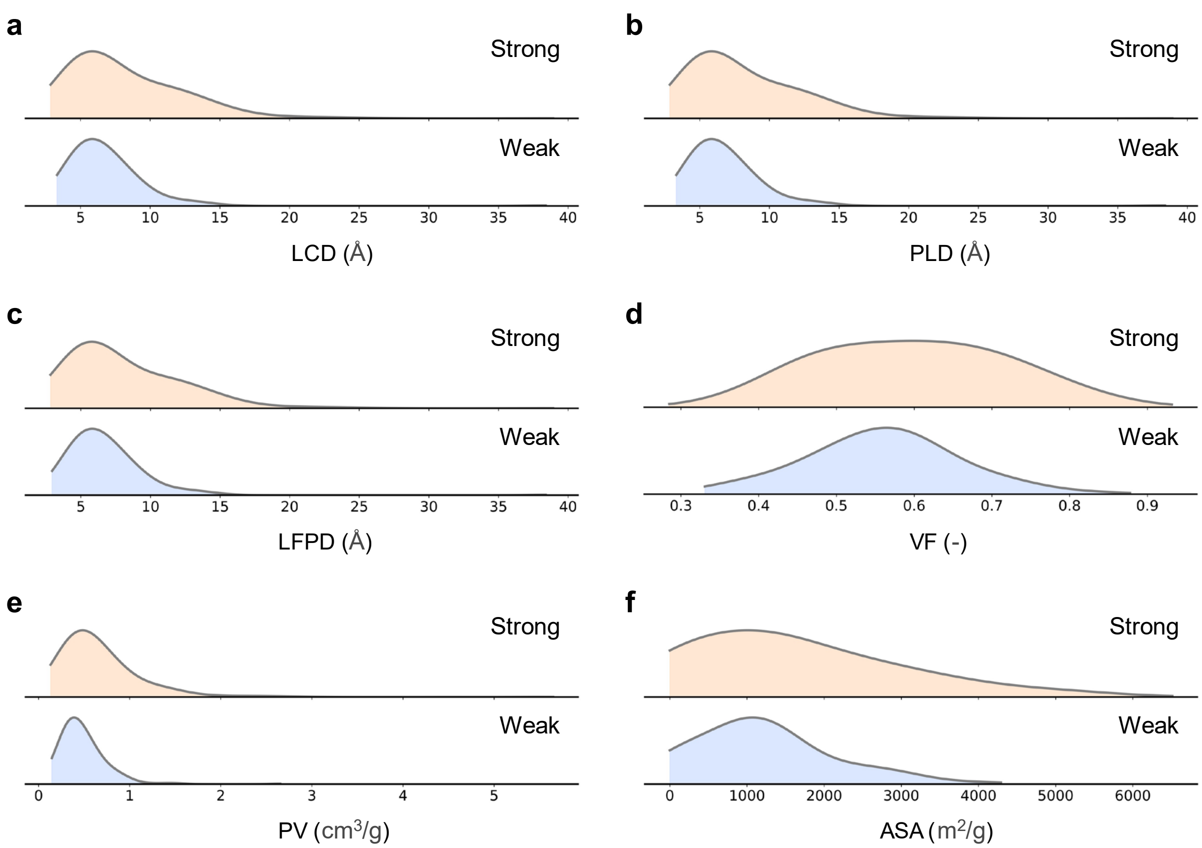

Fig. S2. Kernel density estimations of binary classification versus pore descriptors (LCD, PLD, LFPD, VF, PV and ASA). The vertical axis denotes probability.

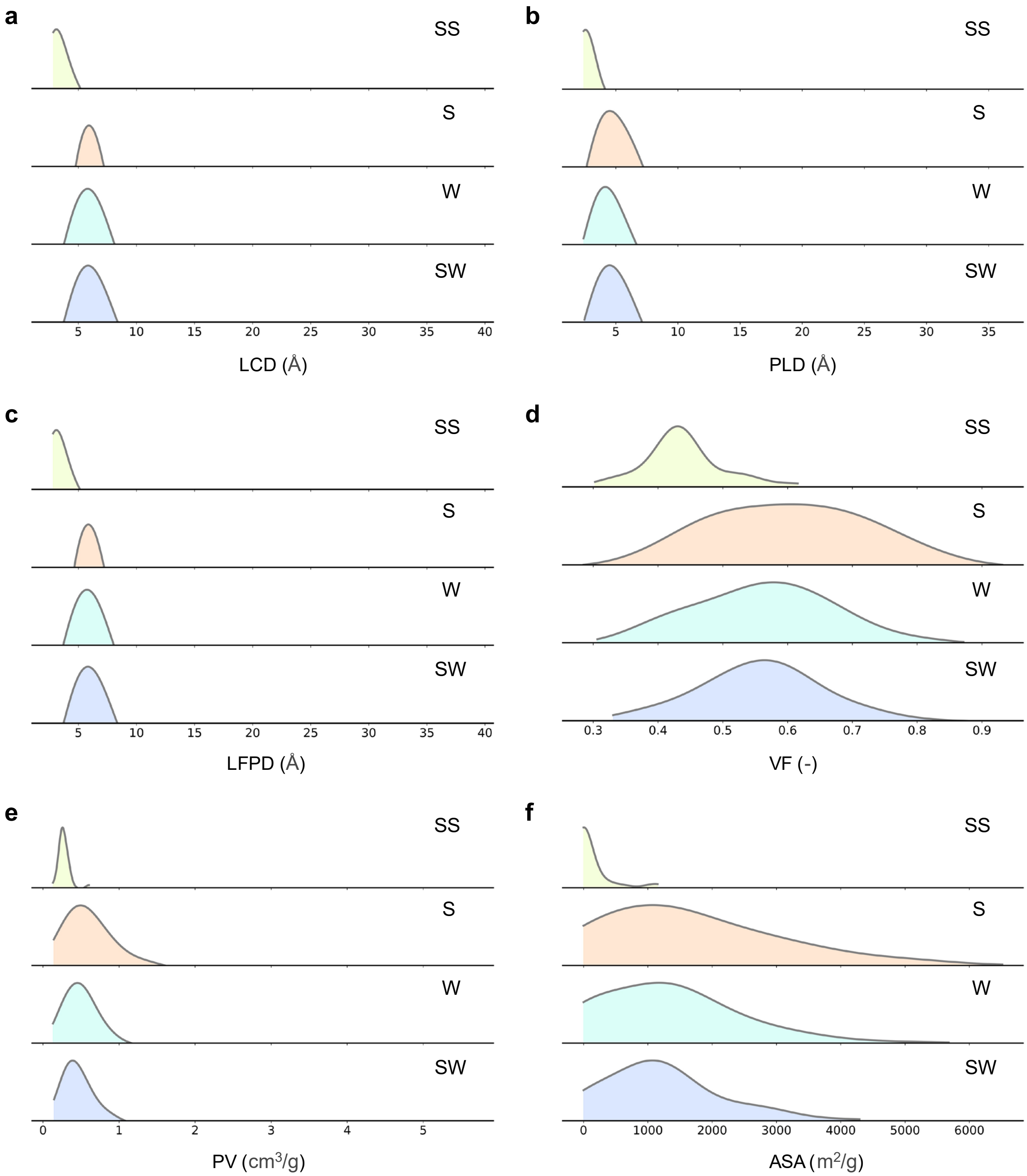

Fig. S3. Kernel density estimations of quaternary classification versus pore descriptors (LCD, PLD, LFPD, VF, PV and ASA). The vertical axis denotes probability.

2. Augmented prompt in the fine-tuned Gemini

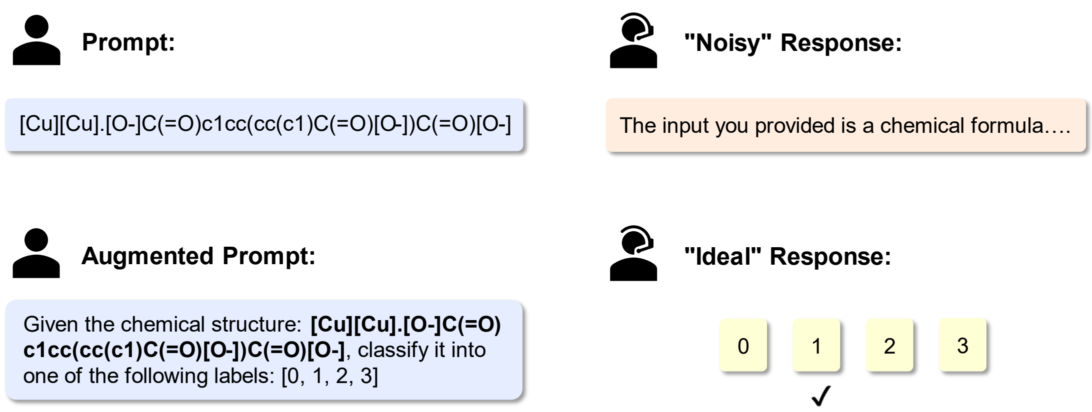

Fig. S4. Examples of an augmented prompt to fix a "Noisy" response.

```
#Python#

import os
import google.generativeai as genai
from google.generativeai.types import HarmCategory, HarmBlockThreshold

os.environ["GOOGLE_API_KEY"]='**your google api key**'
genai.configure(api_key=os.environ["GOOGLE_API_KEY"])

generation_config = {
  "temperature": 0.5,
  "top_p": 0.95,
  "top_k": 40,
  "max_output_tokens": 8192,
  "response_mime_type": "text/plain",
}

model = genai.GenerativeModel(
model_name="**selected fine-tuned model**
generation_config=generation_config,

)

response = model.generate_content([
##  "Input: [Cu][Cu].[O-]C(=O) c1cc(cc(c1)C(=O)[O-])C(=O)[O-]",##
  "Input: Given the chemical structure: [Cu][Cu].[O-]C(=O) c1cc(cc(c1)C(=O)[O-])C(=O)[O-], classify it into one of the following labels: [0, 1, 2, 3]",
  "Output: ",
])

print(response.text)
```

3. Descriptor-based machine learning

Table S1. Pore descriptors.

Descriptor	Elaborations
LCD[a]	Largest cavity diameter
PLD	Pore limiting diameter
LFPD	Global cavity diameter
Density	Density
ASA	Absolute accessible surface area
ASAv	Volumetric accessible surface area
ASAg	Gravimetric accessible surface area
NASA	Absolute non-accessible surface area
NASAv	Volumetric non-accessible surface area
NASAg	Gravimetric non-accessible surface area
PV	Absolute accessible pore volume
VF	Accessible void fraction
PVg	Gravimetric accessible pore volume
NAV	Absolute non-accessible pore volume
NAV_VF	Non-accessible void fraction
NPVg	Gravimetric non-accessible pore volume

[a]All descriptors were calculated using Zeo++[1] using N_2 as a probe with high-accuracy flag.

Table S2. RACs.

Group	Descriptor	Atomic heuristics	Dimension
Product	Metal centre (mc-)	χ, Z, T, I, S	20
	Linker (f-)	χ, Z, T, I, S	20
	Linker-ligand (f-lig-)	χ, Z, T, I, S	20
	Functional group (func-)	χ, Z, T, I, S, α	24
	Linker-connecting (lc-)	χ, Z, T, I, S, α	24
Difference	Metal centre (mc-)	χ, Z, T, I, S	20
	Functional group (func-)	χ, Z, T, I, S, α	24
	Linker-connecting (lc-)	χ, Z, T, I, S, α	24

[a]χ: Pauling electronegativity; Z: nuclear charge; T: atom coordination number; I: atom identity; S: covalent radius; α: polarizability. All descriptors were calculated using molSimplify.[2]

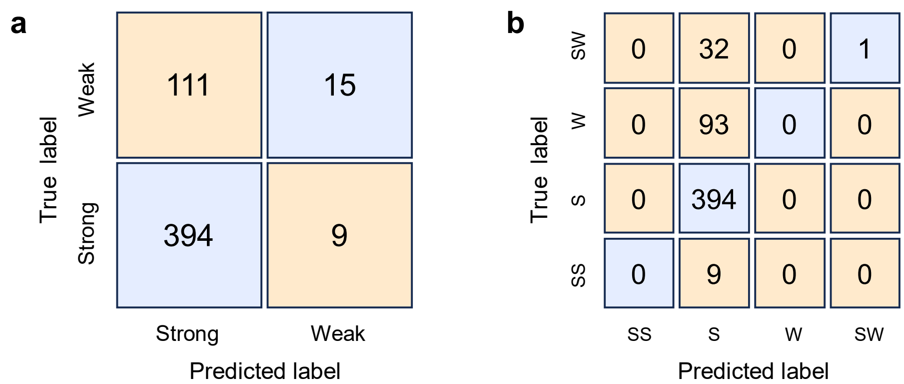

Fig. S5. Confusion matrices on the test set by the ML model with Pore + RACs. (a) binary classification and (b) quaternary classification.

Table S3. Effect of random state in training/test split.

Random state	Binary classification		Quaternary classification	
	Accuracy[a]	F1-score[b]	Accuracy	F1-score
157	0.80	0.72	0.72	0.70
158	0.78	0.73	0.73	0.69
159	0.78	0.73	0.73	0.70
160	0.79	0.72	0.74	0.71
161	0.78	0.72	0.72	0.69

[a] For binary classification, the accuracy metric reflects the model's ability to correctly classify instances into two labels (Strong and Weak). For quaternary classification, the accuracy metric aggregates the overall correct four labels (SS, S, W and SW).
[b] The weighted F1-score combines precision and recall into a single metric, accounting for class imbalance by weighing each class's F1-score by the proportion of instances in that class.

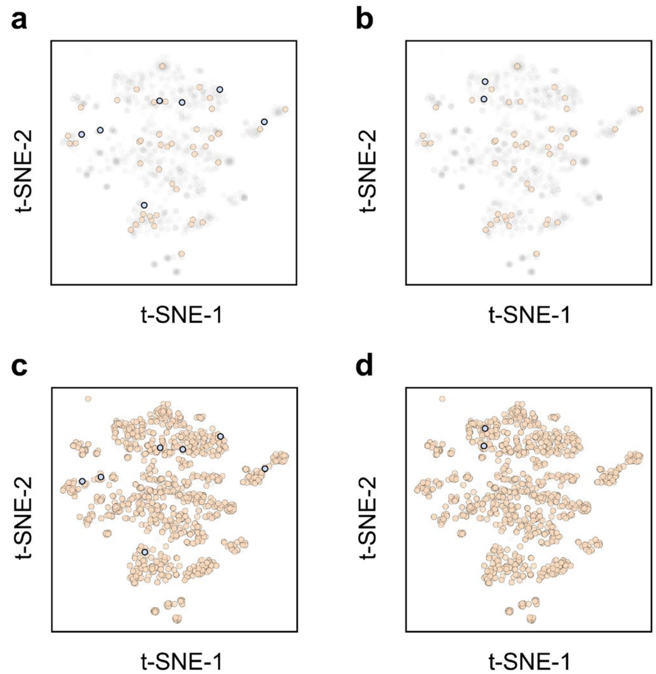

Fig. S6. Spatial variations of chemical space in trained (yellow) and tested (blue) MOFs. (a) Trained SS compared with negatively predicted SS (predicted as S), (b) Trained S compared with negatively predicted SS, (c) Trained SS compared with positively predicted SS, and (d) Trained S compared with positively predicted SS.

4. Blind test

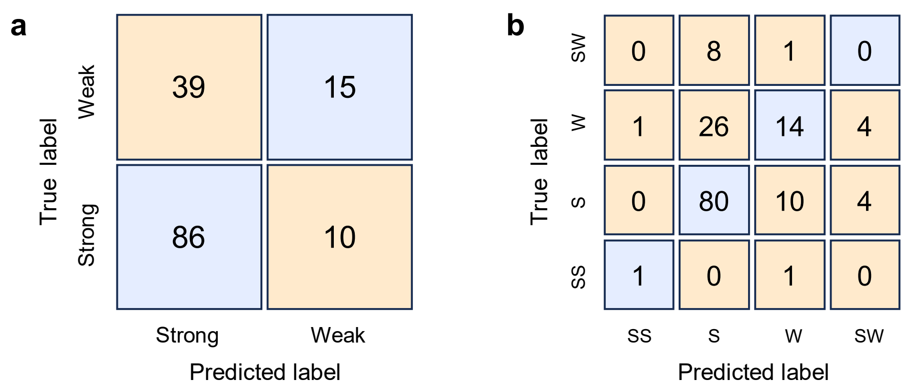

Fig. S7. Confusion matrices by the fine-tuned Gemini on 159 randomly selected MOFs from FSR subset.

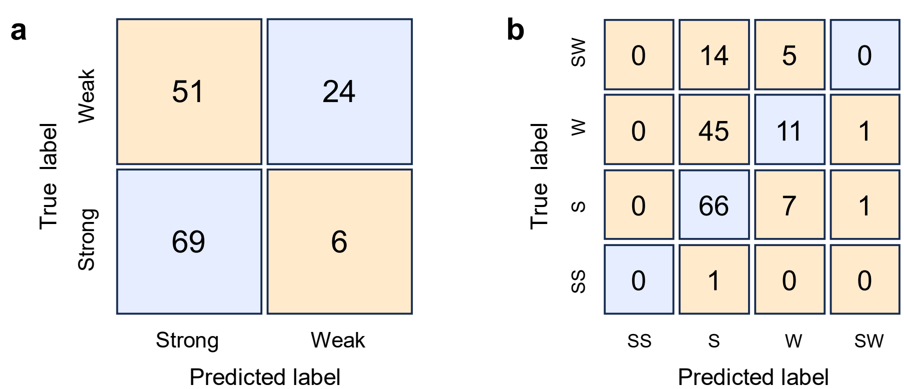

Fig. S8. Confusion matrices by the fine-tuned Gemini on 159 randomly selected MOFs from ION subset.

References

1. T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous Mesoporous Mater.*, 2012, **149**, 134–141.
2. E. I. Ioannidis, T. Z. H. Gani and H. J. Kulik, *J. Comput. Chem.*, 2016, **37**, 2106–2117.