Supplementary Information (SI) for Digital Discovery.

Supporting Information

Large language models in materials science and the need for open-source approaches

Fengxu Yang,[a] Weitong Chen[b] and Jack D. Evans*[a]

[a] *School of Physics, Chemistry and Earth Sciences, Adelaide University, Adelaide 5005, Australia*

[b] *School of Computer and Mathematical Sciences, Adelaide University, Adelaide 5005, Australia*

*Email: j.evans@adelaide.edu.au

Contents

S1 Synthesis Conditions Extraction

S1.1 Model Deployment

All models evaluated for synthesis condition extraction were deployed locally using the SGLang[1] on Phoenix A100 GPU nodes. Inference was conducted under the officially recommended temperature settings for each model. The complete results are available in the Zenodo repository: 10.5281/zenodo.17548056.

S1.2 Extraction Accuracy

Extraction accuracy was evaluated using the benchmark dataset from MOF-ChemUnity comprising manually labelled synthesis conditions from 20 published papers. Due to access restrictions, 15 of these papers were available for our evaluation. For each paper, the full text was provided as input to the model, and the extracted synthesis conditions were compared against the manually curated ground-truth labels. The evaluated fields include metal precursor, organic linker, solvent, temperature, and reaction time. Accuracy was calculated as the ratio of correctly extracted entries to the total number of entries across all evaluated fields:

$$\text{Extraction Accuracy} = \frac{N_{\text{correct entries}}}{N_{\text{total entries}}} \tag{S1}$$

where each entry corresponds to a single field value for a given synthesis record. The comparison between extracted and ground-truth values was performed manually to account for variations in formatting and nomenclature.

S1.3 Hardware Requirements

$$\text{VRAM}_{\text{weights}} \approx 2 \text{ GB} \times (\text{billions of parameters})$$

Beyond the model weights themselves, approximately 20% additional VRAM should be allocated for framework overhead and activation storage.

Table S1: VRAM requirements for models tested on synthesis conditions extraction tasks.

Model	**Params**	**VRAM Estimation (GB)**
GLM-4.5 (355B)	355B	852
Qwen-235B-A30B-Thinking	235B	564
GLM-4.5-Air (106GB)	106GB	254
Qwen3-Next-80B-A3B-Thinking	80B	192
Qwen3-32B	32B	77
DeepSeek-R1-Distill-Qwen-14B	14B	34

Note: This estimate excludes the KV cache. When using inference engines like SGLang, additional VRAM will be required to accommodate the KV cache.

S2 Synthesis Conditions Prediction

S2.1 Recommendation Score

The recommendation score quantifies the accuracy of a model's predicted synthesis recipe against the ground-truth conditions. For each sample, the predicted and true recipes are compared across the following fields, and each individual comparison contributes either a correct or incorrect outcome to the overall tally:

- **Precursors**: Each predicted precursor is independently verified for membership in the ground-truth precursor list.
- **Solvent**: Each predicted solvent is independently verified for membership in the ground-truth solvent list.
- **Temperature**: A prediction is considered correct if the absolute difference between the predicted and true temperatures is within 10 °C.
- **Time**: Both predicted and true durations are first converted to a common unit (hours). A prediction is considered correct if the ratio of the true to predicted value falls within a factor of 2 (i.e., $0.5 < t_{\text{true}}/t_{\text{pred}} < 2$).
- **Pressure**: A prediction is considered correct if the absolute difference between the predicted and true pressures is within 1 atm.
- **Synthesis method, cooling, filtration, and drying**: These fields require an exact match between the predicted and true values.
- **Washing**: Each predicted washing step is independently verified for membership in the ground-truth washing list.

The recommendation score for each sample is then computed as:

$$\text{Recommendation Score} = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{incorrect}}} \tag{S2}$$

where N_{correct} and $N_{\text{incorrect}}$ are the total numbers of correct and incorrect comparisons accumulated across all evaluated fields for that sample. For list-type fields (precursors, solvents, and washing steps), each item in the predicted list is evaluated independently. The recommendation score is first computed independently for every sample in the evaluation set. The median of these per-sample scores is then reported as the overall recommendation score for the model.

S2.2 Data Analysis

We observed that L2M3 training dataset for synthesis recommendation is highly unbalanced, as illustrated in Figure S1 by frequency distributions where most parameters are dominated by a single majority class. For example, 'water' comprises the vast majority of 'Solvent' entries, while 'False' dominates 'pH Adjustment' cases, resulting in significant data sparsity across all other conditions. Although this imbalance is likely to be representative of real-world experimental distributions, it poses a challenge for model training, making it difficult to learn effectively from underrepresented examples and achieve robust generalisation.

S2.3 Model Fine-tuning

The fine-tuning of all models for synthesis condition prediction tasks was conducted using the Hugging Face Transformers[2] and PEFT libraries[3]. This process utilised the AMD Instinct MI250X accelerators on the Setonix supercomputer. The models were fine-tuned for one epoch using a LoRA rank of 32 to produce the results. A higher rank of 128 was also evaluated and it did not produce significantly different performance. All model inferences were generated using a temperature of 0 using vLLM[4]. The full training script is available in the repository cited earlier.

S3 Supplementary References

References

[1] L. Zheng, L. Yin, Z. Xie, C. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez, C. Barrett and Y. Sheng, *SGLang: Efficient Execution of Structured Language Model Programs*, 2024, http://arxiv.org/abs/2312.07104, arXiv:2312.07104.

[2] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest and A. M. Rush, *HuggingFace's Transformers: State-of-the-art Natural Language Processing*, 2020, http://arxiv.org/abs/1910.03771, arXiv:1910.03771.

[3] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao and F. L. Wang, *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*, 2023, http://arxiv.org/abs/2312.12148, arXiv:2312.12148.

[4] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang and I. Stoica, Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.

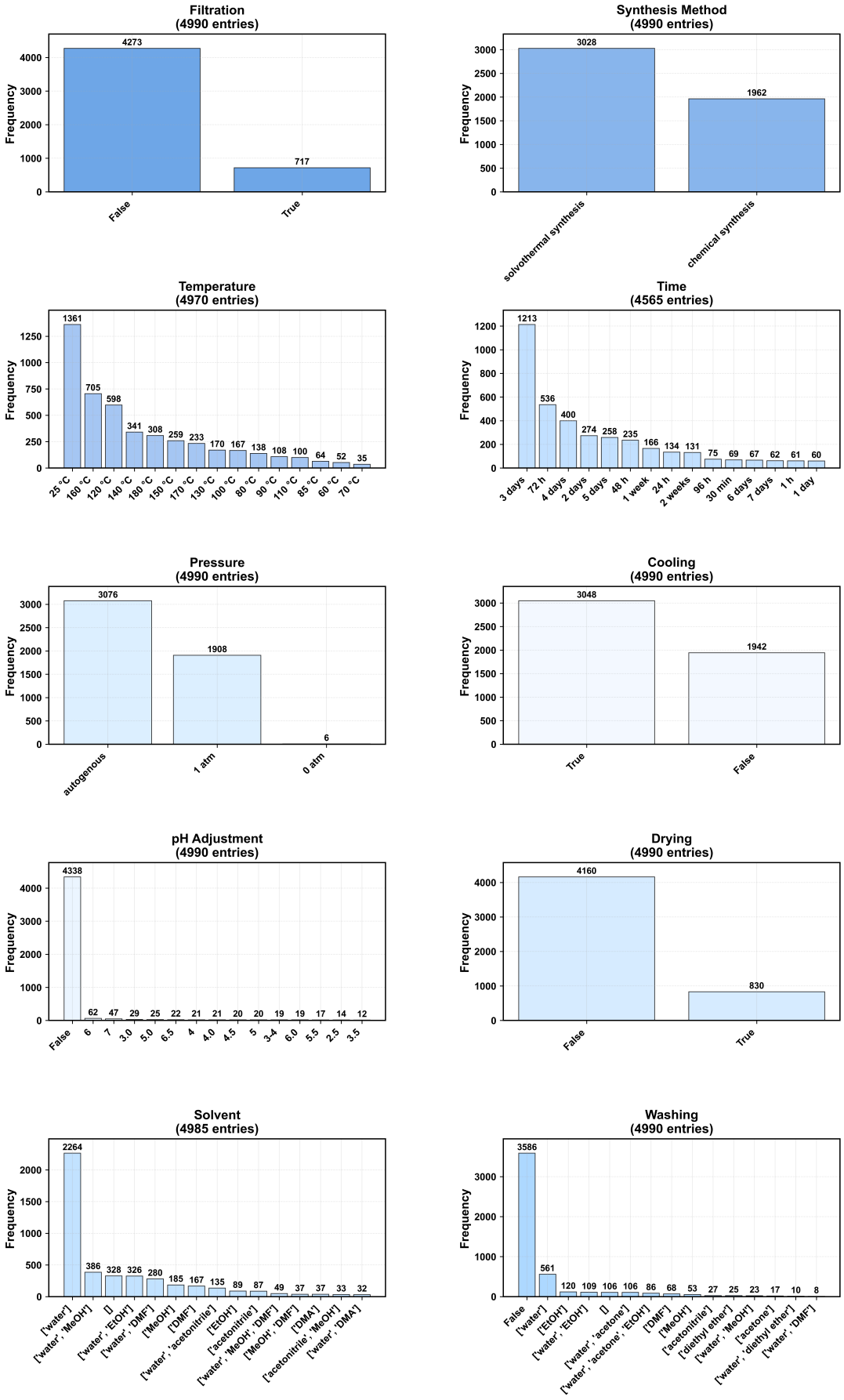

Figure S1: Frequency distribution of the synthesis conditions training dataset from the L2M3 code base.