Model–Data Coevolution as the Basis of Stateful, Problem-Scoped Materials Databases

Fengyu Xie [*1] **Ruoyu Wang** [*1] **Taoyuze Lv** [1] **Yuxiang Gao** [1] **Hongyu Wu** [2] **Zhicheng Zhong** [1 2]

[*]Equal contribution [1]*College of Artificial Intelligence and Data Science, Suzhou Institute of Advanced Research, University of Science and Technology of China, Suzhou, 215123, Jiangsu, China* [2]*Suzhou Laboratory, Suzhou, 215123, Jiangsu, China*. Correspondence to: Zhicheng Zhong zczhong@ustc.edu.cn.

1. Introduction

Computational databases for crystalline materials are foundational infrastructure for modern in silico discovery and screening. However, in the AI era, discovery increasingly proceeds through iterative generative refinement within finite, problem-scoped chemical systems (e.g., a targeted solid electrolyte chemistry) rather than through one-off queries over static repositories. In such workflows, predictive models are repeatedly trained, validated, reused, and extended under bounded elemental constraints, and the learned *model state* itself becomes a key scientific artifact encoding transferable, system-conditioned knowledge. Yet most existing databases[1, 2, 3, 4] remain data-centric: structures and properties are curated and served, while generation logic and predictive models are maintained externally and evolve independently of the database.

We propose a **stateful, model-integrated** formulation of materials databases in which structural data and predictive models jointly constitute the database state and *coevolve* through an endogenous closed loop (Fig. 1). Rather than pursuing a monolithic all-elements repository, we treat each problem-scoped chemical system as an evolvable database node whose growth is formalized as a learnable state transition driven by model–data interaction. This architecture preserves trained model checkpoints alongside curated structures, enabling low-cost continuation of database evolution, reuse of system-conditioned knowledge, and extension to related chemistries via transfer.

2. Approach: endogenous coevolution as a database state transition

Within a selected chemical domain, we implement an integrated generation–evaluation–refinement loop. A deep generative model[5, 6, 7, 8] proposes candidate crystal structures conditioned on composition and stability metrics (E_{hull}). A machine-learned force field (MLFF)[9, 10, 11, 12] provides near–DFT-accuracy energetics for rapid screening and ranking, and a small selected subset is validated by first-principles calculations to update both data and models. Each iteration produces a new database state comprising curated stable–unique–novel structures and the corresponding updated model checkpoints. Because both components are preserved, a node can be exported as a transferable joint model–data state that can be continued, branched, merged, or extended under compositional overlap or elemental expansion, supporting modular development rather than static enumeration.

3. Prototype node: Li–P–S and key results

We prototype this architecture on the chemically intricate Li–P–S ternary system[13, 14], a stringent testbed for solid-state electrolyte chemistry featuring diverse thiophosphate polyanions and coupled lithium configurations. Over seven closed-loop iterations, we generate ∼70,000 candidate structures and obtain >10,000 stable, unique, and novel entries. The node exhibits three operational maturity signals defined within its bounded configuration space:

(i) Stabilization of the thermodynamic distribution. Iterative fine-tuning progressively shifts the generated energy-above-hull distribution toward lower-energy structures while maintaining diversity, indicating improved internal consistency under stability conditioning.

(ii) Rapid saturation of local chemical environments. Although new stable structures continue to appear across iterations, the information entropy of local atomic-environment descriptors rises sharply at early iterations and saturates within the first 2–3 cycles (Fig. 1d), serving as a practical convergence diagnostic that short-range bonding environments are broadly covered in this chemical system.

(iii) Fast MLFF convergence at moderate first-principles cost. Once local environments saturate, the MLFF reaches near–DFT accuracy with a modest number of labeled frames (Fig. 1e), enabling reliable large-scale stability estimation and structure ranking inside the node.

Beyond convergence, the node demonstrates *knowledge extension*: the loop autonomously recovers chemically plausible thiophosphate motifs and phases that are absent from widely used repositories and from the generator's pretraining data, yet are consistent with historical experimental chemistry. Finally, because the mature node stores a coherent model–data state, it supports downstream analyses without task-specific pipeline redesign, including finite-P–T phase stability, high-throughput Li-ion transport screening via MLFF molecular dynamics, and electronic-structure inference by integrating a charge-density model.

4. Implications and outlook

These results support a practical architectural reframing for AI-era materials databases. In contemporary AI-driven workflows, discovery is increasingly realized through iterative generation–evaluation–

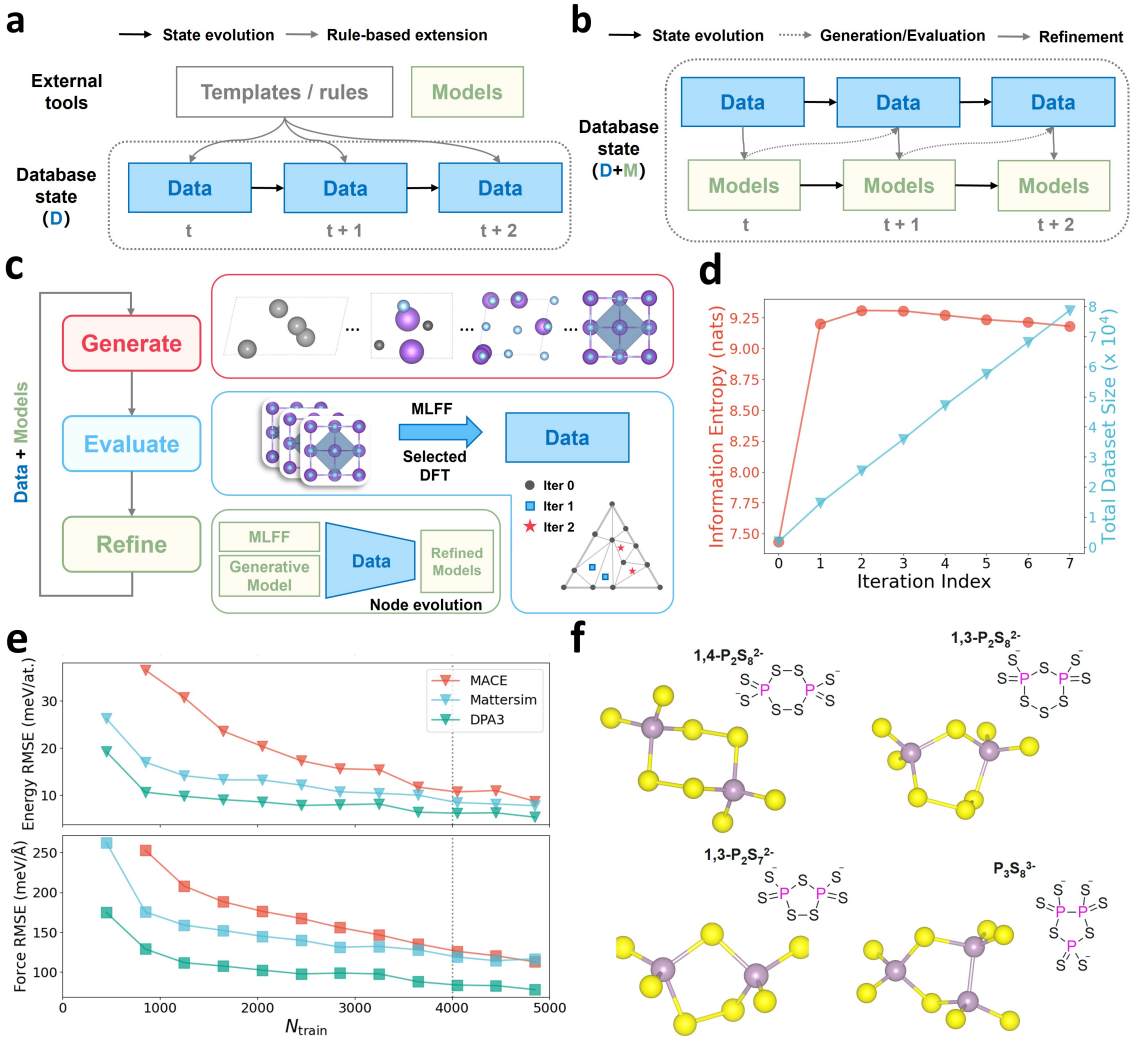

Fig. 1: **Stateful model-integrated database architecture and representative results.** (a) Conventional databases as static repositories expanded by predefined workflows; models remain external. (b) Proposed architecture in which predictive models are persistent components of database state. (c) Endogenous model–data coevolution through a generation–evaluation–refinement loop. (d) Saturation of local atomic environments measured by information entropy versus iteration. (e) Convergence of MLFF accuracy (energy/force RMSE) versus the number of DFT-labeled training frames in Li–P–S. (f) Examples of rediscovered and novel thiophosphate anion motifs.

refinement within *finite, problem-scoped* chemical systems, where both data acquisition rules and predictive models evolve as part of the research process rather than remaining fixed. We therefore formulate database growth as an endogenous *state transition* in which structural data and integrated model checkpoints jointly define and update the database state for a specified chemical domain. Under this formulation, trained model states become preserved, reusable artifacts, enabling continuity of learning within the same system and systematic extension to related systems via transfer, initialization, and compositional expansion. This stateful, model-integrated view *suggests* a scalable pathway for accumulating computational materials knowledge by making model–data coevolution an explicit component of the database, rather than an external procedure applied to it.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grants No. 2021YFA0718900) and National Nature Science Foundation of China (Grants No. 92477114 and No. 12374096). We thank DP Technology for providing computational resources through the Bohrium platform and data hosting services via AIS Square.

References

[1] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013.

[2] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM*, 65(11):1501–1509, November 2013.

[3] Jonathan Schmidt, Hai-Chen Wang, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. A dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals. *Scientific Data*, 9(1):64, March 2022.

[4] Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro J. M. A. Carriço, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. Machine-Learning-Assisted Determination of the Global Zero-Temperature Phase Diagram of Materials. *Advanced Materials*, 35(22):2210788, 2023.

[5] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, pages 1–3, January 2025.

[6] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal Structure Prediction by Joint Equivariant Diffusion on Lattices and Fractional Coordinates. In *Workshop on "Machine Learning for Materials" ICLR 2023*, April 2023.

[7] Chaitanya K. Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary Ward Ulissi. All-atom Diffusion Transformers: Unified generative modelling of molecules and materials. In *Forty-Second International Conference on Machine Learning*, June 2025.

[8] Shuqi Lu, Haowei Lin, Lin Yao, Zhifeng Gao, Xiaohong Ji, Weinan E, Linfeng Zhang, and Guolin Ke. Uni-3DAR: Unified 3D Generation and Understanding via Autoregression on Compressed Spatial Tokens, March 2025.

[9] Duo Zhang, Anyang Peng, Chun Cai, Wentao Li, Yuanchang Zhou, Jinzhe Zeng, Mingyu Guo, Chengqian Zhang, Bowen Li, Hong Jiang, Tong Zhu, Weile Jia, Linfeng Zhang, and Han Wang. Graph neural network model for the era of large atomistic models, June 2025.

[10] Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures, May 2024.

[11] Ilyes Batatia, David P. Kovacs, Gregor Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. In *Advances in Neural Information Processing Systems*, volume 35, pages 11423–11436, December 2022.

[12] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, September 2023.

[13] Atsushi Sakuda, Akitoshi Hayashi, and Masahiro Tatsumisago. Sulfide Solid Electrolyte with Favorable Mechanical Property for All-Solid-State Lithium Battery. *Scientific Reports*, 3(1):2261, July 2013.

[14] Qing Zhang, Daxian Cao, Yi Ma, Avi Natan, Peter Aurora, and Hongli Zhu. Sulfide-Based Solid-State Electrolytes: Synthesis, Stability, and Potential for All-Solid-State Batteries. *Advanced Materials*, 31(44):1901131, 2019.