An Effective Machine Learning Frame for Materials Discovery Structured by a Chemical Concept

Yuanhui Sun
Department of Chemistry and Biochemistry
California State University, Northridge
Northridge, CA 91330, USA
Suzhou Laboratory
Suzhou 215123, China
sunyh@szlab.ac.cn

Austin Ellis
Department of Chemistry and Biochemistry
California State University, Northridge
Northridge, CA 91330, USA
austin.ellis.386@my.csun.edu

Xin Chen
Suzhou Laboratory
Suzhou 215123, China
chenx01@szlab.ac.cn

Maosheng Miao
Department of Chemistry and Biochemistry
California State University, Northridge
Northridge, CA 91330, USA
mmiao@csun.edu

Abstract

Despite extensive studies on binary compounds with high non-metal compositions, there remains a large, unexplored chemical space, particularly regarding non-integer non-metal-to-metal ratios. By integrating the chemical template concept with machine learning algorithms, we developed a specialized structure discovery workflow that significantly enhances the efficiency of predicting stable compounds. Our method led to the identification of 13 new structural prototypes and 31 stable metal superhydrides, representing a 23% increase in discoveries. Metal superhydrides, known for their high hydrogen content and polyhedral hydrogen cages, are promising candidates for high-temperature superconductivity. The method enables us to discover many structures containing over 50 atoms per primitive cell. Additionally, 19 of the newly identified superhydrides exhibit $T_c > 100$ K, highlighting the potential for higher T_c materials within the 3D hydrogen clathrate structures.

1 Introduction

The primary challenge in incorporating non-integer stoichiometry into structural research and density functional theory (DFT)-based stability predictions of superhydrides lies in the significant computational costs required. This is primarily due to the larger unit cells involved, along with the dramatically expanded range of possible metal/non-metal ratios when non-integer values are included. Most previous crystal structure prediction (CSP) studies optimized the positions of both metal and non-metal atoms simultaneously, causing the computational complexity to escalate rapidly with increasing structure size. [1-7]

A recent investigation into the chemical template interactions between metal and hydrogen lattices introduced an effective approach for discovering new metal superhydrides.[8] Guided by the chemical template theory, we first focused on identifying metal lattices with stronger template effects and then constructing corresponding superhydrides by introducing a controlled number of hydrogen atoms into the interstitial sites of selected metal lattices. This two-step approach eliminates many unstable

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat.

or ineffective configurations caused by weak metal-hydrogen interactions, thereby greatly enhancing the efficiency of structural searches.

In this study, we significantly advance this method by developing a specialized structure discovery workflow. First, we used machine learning (ML) algorithms to uncover the relationship between metal lattices from known metal superhydrides and their thermodynamic stability. We then leveraged the trained ML model to efficiently identify metal lattices with strong template effects, which were subsequently used to guide and refine structural searches.

Metal superhydrides, known for their high hydrogen content and polyhedral hydrogen cages, are promising candidates for high-temperature superconductivity.[9-13] Despite extensive studies on binary metal superhydrides, there remains a large, unexplored chemical space, particularly regarding non-integer hydrogen-to-metal ratios. By applying our workflow, we identified 13 new structural prototypes and 31 stable metal superhydrides. Within the compiled dataset used in this work, the newly identified structures account for 24% of the stable set, and the newly identified prototypes correspond to a 23% increase at the prototype level. Among prototypes that host 3D hydrogen clathrates, our results correspond to a 65% increase relative to the baseline in the compiled dataset. Based on superconducting transition temperature estimates for the newly identified structures, 19 of the 31 new stable metal superhydrides exhibit Tc > 100 K.

2 The structure discovery workflow

Instead of simultaneously searching and optimizing the positions of both hydrogen and metal atoms, we first apply ML techniques to identify metal lattices exhibiting significant electron localization at interstitial sites (also called quasi-atoms). Structural searches with varying metal/non-metal ratios are then conducted based on these selected metal lattices to uncover new stable structures. The workflow is schematically illustrated in Figure 1 and comprises six key steps: training set preparation, feature engineering, ML model training, metal lattice candidate preparation, metal lattice screening, and high-throughput computational screening of stable metal superhydrides derived from selected metal lattices.

3 Metal superhydrides of MH_x with non-integer stoichiometry x

The ML workflow enabled us to find several new structural polymorphs of integer-stoichiometry superhydrides, expanding beyond the well-known clathrate phases such as $Fm\bar{3}m$-LaH_{10}. Three new polymorphs of MH_{10} (M = La, Ce, Th) were identified: $P6_3/mmc$-La_4H_{40}, $P6_3/mmc$-La_6H_{60}, and R3m-La_9H_{90}. These structures feature larger and more complex primitive cells (44–66 atoms) compared to the classic LaH_{10} phase (11 atoms). Their metal sublattices adopt diverse symmetries, including double hexagonal close packing and Sm-type arrangements, while the hydrogen networks consist of distorted H_{32} cages, indicating structural flexibility within MH_{10} stoichiometry. Energetically, the polymorphs are comparable to Fm3m phases across 150–300 GPa, and remain favorable when extended to Ce and Th analogues, with stability preserved after accounting for zero-point energy (ZPE) and finite-temperature effects.

Additionally, a new family of M_4H_{52} (M = La, Sr, Ac) compounds with $P6_3/mmc$ symmetry was discovered. Each primitive cell contains 56 atoms and features an AABB-stacked metal sublattice. Hydrogen atoms organize into two distinct motifs: repeating H_{24} clusters within AA layers and a two-dimensional hydrogen sheet in AB regions, constructed from interconnected H_8 cubes. These building units echo motifs previously identified in LaH_{16} and MH_9 clathrates, respectively, but here coexist within a single structure. Thermodynamically, $P6_3/mmc$ M_4H_{52} is stable at target pressures, though its robustness is temperature-dependent: Ac_4H_{52} remains stable at 300 GPa with ZPE and thermal corrections, while La_4H_{52} and Sr_4H_{52} become metastable. Phonon spectra confirm their dynamical stability under pressure.

4 Superconductivity of newly identified metal superhydrides

The superconducting properties of the newly discovered superhydrides were systematically assessed. Using a rapid evaluation method based on key hydride-specific descriptors—three-dimensional electronic connectivity, hydrogen content, and the hydrogen contribution to the electronic density of

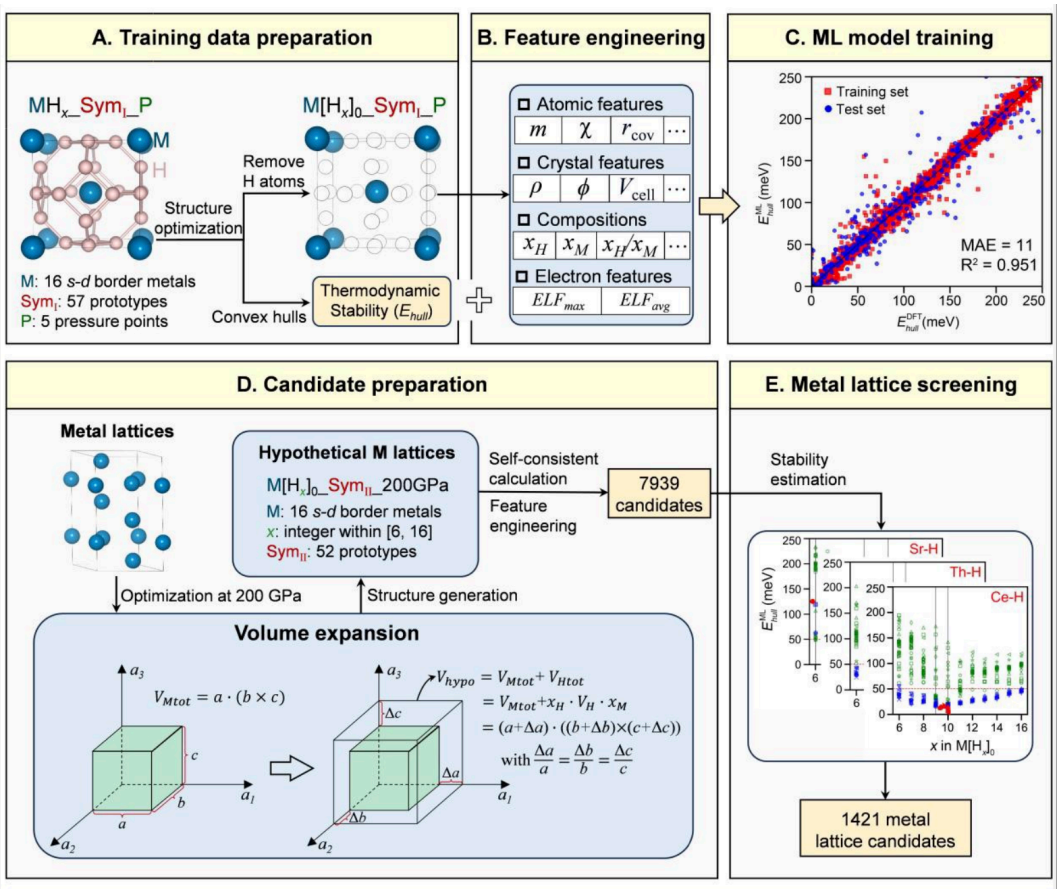

Figure 1: The designed workflow for discovering metal superhydrides. (A) Using 57 collected structural prototypes, metal substitution is carried out with s-d border metals, and structural optimization is performed at various pressure points to obtain the metal lattice and thermodynamic stability (E_{hull}) values for the prepared dataset. (B) Atomic features, lattice features, composition, and electronic features of the metal lattices are extracted to create the final feature set for feature engineering, with 46 features used for model training. (C) A robust relationship between the features and thermodynamic stability of superhydrides is established using a stack-ensembling ML strategy based on the AutoGluon framework, achieving an accuracy of MAE = 11.366 meV/atom and R2 = 0.951. (D) The hypothetical configurations of metal lattices after hydrogen atom filling are considered under different H/Metal ratios, resulting in 7939 metal lattices for screening. (E) The trained ML model is used to predict the stability of the screened metal lattices. In the case of the Ce-H system, the red scatters represent the newly identified Ce-based superhydrides, along with their corresponding $E_{\text{hull}}^{\text{ML}}$ values at various H compositions (shown as blue scatters). The black vertical lines indicate the identified stable CeH_9 and CeH_{10}.

states at the Fermi level—the study estimated transition temperatures (T_c) across the 31 identified structures. Remarkably, 19 compounds (61%) exhibit Tc values above 100 K, underscoring their potential as high-temperature superconductors. Most of these high-Tc phases are associated with La, Ce, and Th systems, aligning with trends observed in previously reported superconducting superhydrides.

To validate the approach, electron–phonon coupling were calculated for Pm3n-Ca_8H_{46} at 200 GPa. [14-16] The phonon spectrum shows low-frequency modes from Ca and high-frequency modes from H. The resulting parameters, including a logarithmic average frequency of 1298 K, an electron–phonon coupling constant (λ) of 1.49, and T_c in the range of 134–147 K ($\mu^\star$ = 0.1–0.13), confirm a strong superconducting potential. Most of the superconductivity is contributed by hydrogen vibrational modes, which contribute 83% of the total coupling.

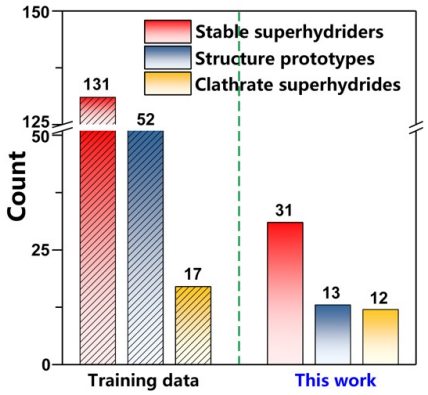

Figure 2: A comparison of newly discovered and known metal superhydrides. The left and right panels show the statistics of the training set and the new superhydrides. The red, blue, and yellow bars represent the number of stable superhydrides, structure prototypes, and clathrate superhydrides, respectively.

5 Discussion

The ML workflow, guided by the concept of chemical templates, proved highly effective in accelerating the discovery of complex superhydrides. Traditionally, crystal structure prediction methods struggle with the vast configurational space of hydrogen-rich systems. By focusing on metal lattices with strong template effects first, the workflow narrows the search to chemically meaningful candidates, eliminating many unstable configurations.

The ML model, trained on 57 known prototypes and expanded with s–d border metal substitutions, achieved excellent predictive accuracy (MAE 11 meV/atom, $R^2 = 0.95$). This allowed rapid screening of nearly 8,000 hypothetical lattices, from which 1,400 promising candidates were selected. High-throughput DFT confirmed 13 new structural prototypes and 31 stable superhydrides, representing 23% growth in prototype diversity and 65% growth in clathrate-type frameworks compared to the training set. Moreover, the workflow captured stability trends consistent with known chemistry, while also revealing overlooked noninteger stoichiometries. Crucially, it balanced efficiency and accuracy, making it feasible to explore large and complex structures (50–94 atoms/cell).

This novel framework provides a conceptual structure for organizing and interpreting complex structural and electronic patterns in superhydrides, enabling more targeted and efficient exploration of the vast compositional space. The workflow operates through a hybrid human-machine learning approach, in which human insight formulates and refines conceptual models, while machine algorithms execute large-scale predictions and optimizations. This methodology exemplifies an early stage attempt to implement concept-level reasoning within an artificial intelligence framework [17-20]. We are expanding our current work to include the direct generation of the charge density and ELF, and the extraction of novel concepts from convolving the 3D images. As such, the current manuscript represents a step towards the broader goal of developing AI systems capable of managing, generating, and applying scientific concepts across domains. The 30 predicted compounds did not receive full electron-phonon calculations, they will need to be performed as the next step. Code is available at https://github.com/hison001/Metal-superhydrides/tree/main. A full version of this work has been published in the Journal of the American Chemical Society, DOI: 10.1021/jacs.5c11731.

Acknowledgments

M.M. and A.E.acknowledge the support of the DoD HBCU/MI Basic Research Funding under grant No. W911NF2310232, NSF funds DMR 1848141 and OAC 2117956, the ACF PRF 59249-UNI6, the Camille and Henry Dreyfus Foundation, and California State University RSCA awards.

References

[1] Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. Crystal Structure Prediction via Particle-Swarm Optimization. *Phys. Rev. B* 2010, 82 (9), 094116. https://doi.org/10.1103/PhysRevB.82.094116.

[2] Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A Method for Crystal Structure Prediction. *Comput. Phys. Commun.* 2012, 183 (10), 2063–2070. https://doi.org/10.1016/j.cpc.2012.05.008.

[3] Avery, P.; Zurek, E. RandSpg: An Open-Source Program for Generating Atomistic Crystal Structures with Specific Spacegroups. *Comput. Phys. Commun.* 2017, 213, 208–216. https://doi.org/10.1016/j.cpc.2016.12.005.

[4] Oganov, A. R.; Glass, C. W. Crystal Structure Prediction Using Ab Initio Evolutionary Techniques: Principles and Applications. *J. Chem. Phys.* 2006, 124 (24), 244704. https://doi.org/10.1063/1.2210932.

[5] Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX—Evolutionary Crystal Structure Prediction. *Comput. Phys. Commun.* 2006, 175 (11), 713–720. https://doi.org/10.1016/j.cpc.2006.07.020.

[6] Pickard, C. J.; Needs, R. J. Structures at High Pressure from Random Searching. *Phys. Status Solidi B-Basic Solid State Phys.* 2009, 246 (3), 536–540. https://doi.org/10.1002/pssb.200880546.

[7] Pickard, C. J.; Needs, R. J. Ab Initiorandom Structure Searching. *J. Phys. Condens. Matter* 2011, 23 (5), 053201. https://doi.org/10.1088/0953-8984/23/5/053201.

[8] Sun, Y.; Miao, M. Chemical Templates That Assemble the Metal Superhydrides. *Chem* 2023, 9 (2), 443–459. https://doi.org/10.1016/j.chempr.2022.10.015.

[9] Ashcroft, N. W. Hydrogen Dominant Metallic Alloys: High Temperature Superconductors? *Phys. Rev. Lett.* 2004, 92 (18), 187002. https://doi.org/10.1103/PhysRevLett.92.187002.

[10] Ashcroft, N. W. Bridgman's High-Pressure Atomic Destructibility and Its Growing Legacy of Ordered States. *J. Phys. Condens. Matter* 2004, 16 (14), S945–S952. https://doi.org/10.1088/0953-8984/16/14/003.

[11] Zurek, E.; Bi, T. High-Temperature Superconductivity in Alkaline and Rare Earth Polyhydrides at High Pressure: A Theoretical Perspective. *J. Chem. Phys.* 2019, 150 (5), 050901. https://doi.org/10.1063/1.5079225.

[12] Pickard, C. J.; Errea, I.; Eremets, M. I. Superconducting Hydrides Under Pressure. *Annu. Rev. Condens. Matter Phys.* 2020, 11 (1), 57–76. https://doi.org/10.1146/annurev-conmatphys-031218-013413.

[13] Errea, I.; Belli, F.; Monacelli, L.; Sanna, A.; Koretsune, T.; Tadano, T.; Bianco, R.; Calandra, M.; Arita, R.; Mauri, F.; Flores-Livas, J. A. Quantum Crystal Structure in the 250-Kelvin Superconducting Lanthanum Hydride. *Nature* 2020, 578 (7793), 66–69. https://doi.org/10.1038/s41586-020-1955-z.

[14] Oliveira, L. N.; Gross, E. K. U.; Kohn, W. Density-Functional Theory for Superconductors. *Phys. Rev. Lett.* 1988, 60 (23), 2430–2433. https://doi.org/10.1103/PhysRevLett.60.2430.

[15] Lüders, M.; Marques, M. A. L.; Lathiotakis, N. N.; Floris, A.; Profeta, G.; Fast, L.; Continenza, A.; Massidda, S.; Gross, E. K. U. Ab Initio Theory of Superconductivity. I. Density Functional Formalism and Approximate Functionals. *Phys. Rev. B* 2005, 72 (2), 024545. https://doi.org/10.1103/PhysRevB.72.024545.

[16] Allen, P. B.; Dynes, R. C. Transition Temperature of Strong-Coupled Superconductors Reanalyzed. *Phys. Rev. B* 1975, 12 (3), 905–922. https://doi.org/10.1103/PhysRevB.12.905.

[17] B. Goertzel, Intensional Inheritance Between Concepts: An Infor-mation-Theoretic Interpretation, arXiv:2501.17393.

[18] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, Building machines that learn and think like people, *Behavioral and Brain Sciences* 40, e253 (2017).

[19] H. Wang et al., Scientific discovery in the age of artificial intelligence, *Nature* 620, 7972 (2023).

[20] L. Messeri and M. J. Crockett, Artificial intelligence and illusions of understanding in scientific research, *Nature* 627, 49 (2024).