Rapid and automated alloy design with graph neural network-powered large language model-driven multi-agent AI

Rapid and automated alloy design with graph neural network-powered large language model-driven multi-agent AI

Alireza Ghafarollahi and Markus J. Buehler*

Received: 4 February 2025 / Revised: 7 June 2025 / Accepted: 9 June 2025

A multi-agent artificial intelligence (AI) model is developed to automate the discovery of new metallic alloys, integrating multimodal data and external knowledge, including insights from physics via atomistic simulations. The system consists of (a) large language models (LLMs) for tasks such as reasoning and planning, (b) AI agents with distinct roles collaborating dynamically, and (c) a newly developed graph neural network (GNN) model for rapid retrieval of physical properties. We chose the ternary NbMoTa body-centered-cubic alloy as our model system and developed the GNN to predict two fundamental materials properties: the Peierls barrier and the solute/screw dislocation interaction energy. Our GNN model efficiently predicts these properties, reducing reliance on costly brute-force calculations and alleviating the computational demands on the multi-agent system. By combining the predictive capabilities of GNNs with the collaborative intelligence of LLM-driven reasoning agents, the system autonomously explores vast alloy design spaces, identifies trends in atomic-scale properties, and predicts macroscale mechanical strength, as demonstrated by several computational experiments. This synergistic approach accelerates the discovery of advanced alloys and holds promise for broader applications in other complex systems, marking a step forward in automated materials discovery and design.

Impact statement

Traditional deep learning models, such as graph neural networks and convolutional neural networks, operate within the confines of their training data sets, making single-step inferences for regression or classification. Our work introduces a multi-agent strategy that transcends these limitations by integrating deep learning with reasoning and decision-making capabilities. This intelligent system actively interprets results, determines subsequent actions, and iteratively refines predictions, accelerating the materials design process. We demonstrate its effectiveness in exploring the vast compositional space of a ternary alloy, where the model dynamically solicits data, analyzes trends, generates visualizations, and derives insights into materials behavior. By enabling accurate predictions of key alloy characteristics, our approach advances the discovery of novel metallic systems and underscores the critical role of solid-solution alloying. More broadly, it represents a major step toward integrating artificial intelligence with scientific reasoning, moving closer to artificial general intelligence in engineering. This paradigm shift has profound implications for materials science, enabling more efficient, autonomous, and intelligent exploration of complex materials spaces.

Introduction

Multi-principal element alloys (MPEAs) represent a relatively new and innovative class of materials comprising three or more elements and showing remarkable mechanical properties superior to their pure or dilute counterparts, such as mechanical strength, fracture toughness, ductility, and resistance to hydrogen embrittlement.[1–4] The body-centered-cubic (bcc) refractory MPEAs of the family Cr–Mo–W–Nb–V–Ta–Ti–Zr–Hf have recently gained particular interest due to their exceptional high-temperature strength retention, exceeding the capabilities of current superalloys.[5–12] With millions of possible compositions within just one single crystal, these complex systems offer significant potential to achieve tailored properties for specialized applications. However, navigating the immense MPEAs compositional space to guide alloy discovery with optimized or desired properties poses significant challenges. Developing mechanistic multiscale theories that connect atomistic-level phenomena to microscopic materials properties, such as temperature-dependent yield stress, have been instrumental in exploring this design space.[13–23]

While theoretical models hold significant potential for predicting the macroscopic properties of MPEAs and navigating their vast design space, a key limitation lies in the computational cost of obtaining the necessary input parameters, which typically

Alireza Ghafarollahi, Laboratory for Atomistic and Molecular Mechanics (LAMM), Massachusetts Institute of Technology, Cambridge, USA
Markus J. Buehler, Laboratory for Atomistic and Molecular Mechanics (LAMM), Center for Computational Science and Engineering, Schwarzman College of Computing, Massachusetts Institute of Technology, Cambridge, USA; mbuehler@MIT.EDU
*Corresponding author
doi:10.1557/s43577-025-00953-4

rely on atomistic simulations. This challenge is particularly acute for bcc materials, where plasticity is controlled by the motion of screw dislocations.[21,24] In contrast to edge dislocation theories, which offer a simplified approach by using misfit volume as a primary parameter,[13,14,18] no such simplification exists for screw dislocation theories. As a result, these models require the computation of critical parameters via expensive atomistic simulations. Two key quantities in screw dislocation theories are the Peierls barrier and the solute-screw interaction energy.[21,24] The Peierls barrier represents the intrinsic lattice resistance to dislocation motion, while the solute-screw interaction energy describes the influence of solute atoms on the dislocation's behavior. Both factors serve as energy barriers to dislocation motion, and their calculation is further complicated by random solute fluctuations around the dislocation. Atomistic simulations are commonly used to determine these parameters in alloys,[25] but applying them to multicomponent alloys presents two significant challenges: (1) the vast design space of these alloys, and (2) the need for numerous realizations to obtain a statistically accurate average due to the random solute environments. To address these limitations, machine learning (ML) and deep learning (DL) models offer a promising alternative by streamlining the computation of these input parameters, potentially reducing the reliance on computationally expensive brute-force methods.

The advent of artificial intelligence (AI) methods have introduced new ways of materials design, physical modeling, and properties measurement.[26–31] These methods can discover hidden patterns in the training data and thus have been merged into all disciplines. The use of DL methods within the field of crystalline materials is wide, from ML interatomic potential development[32–36] to intrinsic properties calculation[37,38] to dynamic crack path prediction.[39] Among all the DL architectures, graph neural network (GNN) models have been developed to deal with graph structures that model a series of objects (nodes) and the relationships between them (edges), making them an ideal solution for crystalline materials where the atoms are nodes and the edges represent the intermediate bond.[40,41]

While ML methods have accelerated the exploration of MPEAs, they often target specific materials properties in isolation and cannot reason over complex relationships, data, and modalities. This narrow focus can limit their ability to incorporate the broader, interdisciplinary knowledge that is critical for true breakthroughs in alloy design and scientific discovery. To overcome these limitations, multi-agent systems have emerged as a transformative approach, facilitating the integration of multimodal data and external knowledge, such as new developments in physics and materials science including theoretical models, into the design process in a more holistic and adaptive way.[42–54] In multi-agent systems, a group of AI agents, each powered by a set of large language models (LLMs) or reasoning models, work together dynamically to solve complex, multifaceted problems. By assigning distinct roles to each agent via targeted prompts, the system can divide and conquer different aspects of the alloy design challenge.

In this article, we present an LLM-based multi-agent system that leverages the power of: (a) a suite of advanced LLMs and reasoning models, responsible for fundamental tasks such as planning, reasoning, and decision-making; (b) a group of specialized agents, each playing distinct roles within the system; and (c) a set of external tools for various tasks, including a newly developed GNN model to predict atomic-level materials properties. This multi-agent model builds upon AtomAgents, a multimodal multi-agent system with advanced simulation capabilities for alloy design and discovery. AtomAgents was capable of retrieving new physics from atomistic simulations, solving complex multifaceted alloy discovery problems. However, as the complexity of the alloy systems increased, the computational cost of atomistic simulations became prohibitive.

To address this, we developed a GNN model that bypasses the limitations of brute-force calculations by providing rapid physics predictions. Although trained on a relatively small portion of the compositional space of the ternary Nb–MoTa family of bcc alloys, the GNN model accurately predicts fundamental quantities, including the Peierls barrier and solute/screw interaction energy parameter. Several examples are provided to demonstrate the proficiency of this multi-agent approach in not only exploring the design space but also tackling complex alloy design challenges, such as predicting macroscopic yield stress.

The plan of the article is as follows. We start by providing details about our GNN model to predict atomic-level materials properties in the Section "GNN model." We then discuss our proposed LLM-driven multi-agent system for automated materials design, guided by the predictions from our GNN model in the Section "Multimodal multi-agent model." We outline the main components of the multi-agent approach and provide several examples to illustrate the proficiency of our model in solving complex alloy design problems. We then present the key findings in the Section "Summary and future perspectives" and discuss the implications of our multi-agent system for future research in materials discovery.

Results and discussion

The main tool implemented in our multi-agent system is the GNN model that accurately predicts the fundamental materials properties in multicomponent alloys (i.e., Peierls barrier and potential energy change due to the screw dislocation movement). We first delve into the basics of development of the GNN model and then discuss our proposed LLM-based multi-agent system to design alloys solicitating physics data from the GNN model.

GNN model

The workflow of the GNN approach for predicting the Peierls barrier and energy changes due to dislocation movement is schematically presented in **Figure** 1. The dislocation is assumed to glide along the $x||[11\bar{2}]$ direction. The dislocated

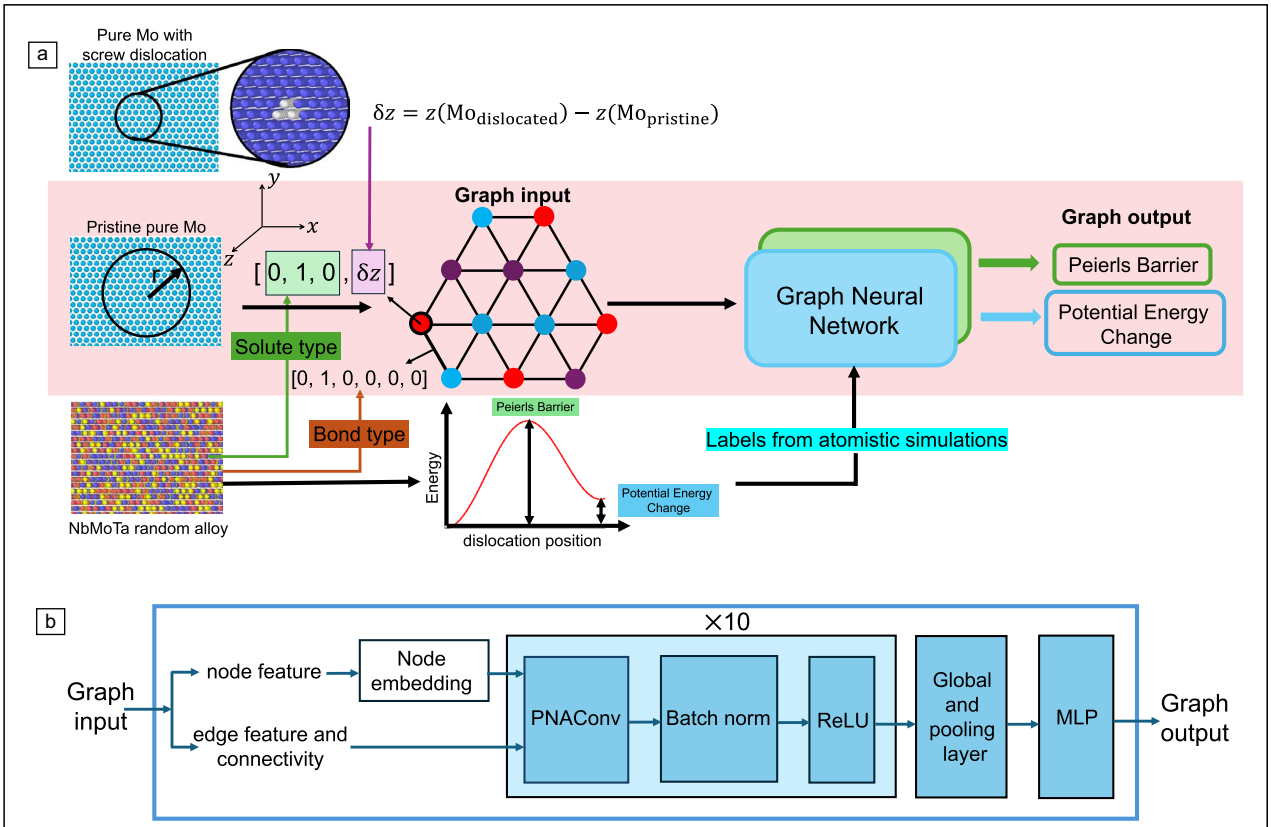

Figure 1. (a) Overview of the workflow used in this work to train graph neural network (GNN) models for an end-to-end prediction of the Peierls barrier and potential energy changes. The process begins with the generation of initial and final dislocation structures in random body-centered-cubic multicomponent alloys, which are then minimized to compute the initial and final potential energies. Next, nudged elastic band simulations are performed to determine the minimum energy path and calculate the Peierls barrier. The input to the machine learning model is the graph representation of the alloys, with nodes encoding spatial and chemical information, and edges representing bond types. Two supervised GNN models are trained to predict graph-level labels: the Peierls barrier and potential energy change. (b) GNN architecture. The graph input is first passed into an input block to enlarge the dimension of node features. The node embedding, edge features, and connectivity of input graphs are then input to the message passing block where the information from the neighbors of each node in the graph is aggregated to update the hidden features of the nodes. The output of the message passing block is then input into a global pooling later, which outputs a graph-level embedding by adding node embeddings across the node, and it is connected to a multilayer perceptron (MLP) that returns a predicted graph label (i.e., Peierls barrier or potential energy change). PNAConv, Principal Neighborhood Aggregation graph convolution operator; ReLU, rectified linear unit.

structures for 200 samples per composition within the training compositional space are generated, as shown in Supplementary information (SI) Figure S1. After minimizing these structures, the potential energy change is calculated by subtracting the final potential energy from the initial one. Additionally, the Peierls barrier for each sample is extracted from the minimum energy path connecting the initial and final states, as derived from nudged elastic band (NEB) simulations. The potential energy changes and the Peierls barriers serve as the ground truth or labels used to train the DL model. In this model, each random configuration is represented as a graph, where nodes correspond to atoms and edges to chemical bonds. Node features consist of chemical (solute types) and configurational features (screw dislocation displacement). The solute types at each site are one-hot encoded, with vectors [1, 0, 0], [0, 1, 0], and [0, 0, 1] representing Nb, Mo, and Ta, respectively. Additionally, the z-component of the screw dislocation displacement, δ_z, is calculated for pure Mo and included as a node feature. Notably, only the chemical features vary across compositions, while the δ_z feature remains constant, thus eliminating the need for atomic relaxation in new random configurations during inference. Furthermore, edge features are constructed based on the solute types of the neighboring atoms. After constructing the graph input, the GNN model is trained to predict either the potential energy change or the Peierls barrier, which serves as the label for the entire graph. Once trained, the model can predict the potential energy change and Peierls barrier for new structures, drastically reducing the need for time-consuming atomistic simulations, particularly NEB calculations. This enables us to explore the full compositional space of NbMoTa ternary alloys to identify candidates with enhanced properties, such as a higher Peierls barrier.

Peierls barrier and potential energy change predictions are graph regression tasks. In this work, we adopt a GNN with the

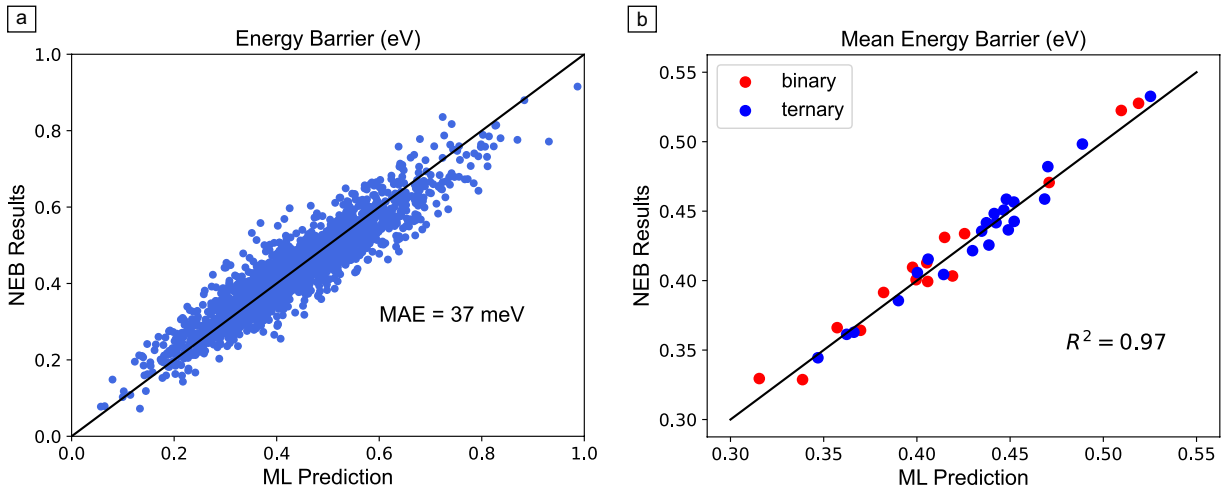

Figure 2. Evaluation of the graph neural network model for Peierls barrier prediction on the test set. Test set contains new compositions that never appeared in the training set. (a) Comparison of machine learning (ML) predictions and nudged elastic band (NEB) results for the Peierls energy barrier. MAE, mean absolute error. (b) Comparison of the ML results and NEB results for the mean energy barriers of binary and ternary compositions in the test set. The correlation coefficient R^2 between the predictions and the ground truth is 0.97.

Principal Neighborhood Aggregation graph convolution operator (PNAConv),[55] which has outperformed many popular GNN models in the literature, such as graph convolutional network,[56] graph attention network,[57] and graph isomorphism network[58] on benchmark graph regression tasks. The PNA model combines multiple aggregators, which affect how messages between nodes are passed, with a degree-scaler that generalizes the sum aggregator. Specifically, the architecture employs four different aggregators: mean, maximum, minimum, and standard deviation. The improved performance of the GNN is attributed to this combination, where the degree-based scalers amplify or attenuate signals in the network based on node degree.

More details about the construction of the data set, the architecture of the GNN model, and the training procedure are provided in the "Methods" section. In the following sections, we evaluate the model's performance in predicting materials properties on the test data set and assess its ability to generalize to unseen configurations. We would like to emphasize that the main goal of this research is not to develop the most accurate DL surrogate model for material property prediction, but rather to create a surrogate model that achieves a reasonable balance between accuracy and computational efficiency, enabling the rapid exploration of the vast compositional space of multicomponent alloys.

Peierls barrier

Random fluctuations in the solute environment surrounding the screw dislocation play a crucial role in determining the Peierls barrier, introducing a degree of complexity that must be captured for accurate predictions. In this section, we assess how well our GNN model predicts the Peierls barrier for various random compositions within the test set. To ensure the model's generalizability, the compositions in the test set are strategically distributed to cover a broad range of the compositional space, including the equimolar composition $Nb_{33}Mo_{33}Ta_{33}$ and binary systems. This comprehensive distribution allows us to verify that the model can accurately capture both the complexity of ternary alloys and the simpler binary ones, which is critical for extending the model to new, unseen compositions. Moreover, each composition consists of 50 realizations of random solute configurations, ensuring a robust evaluation of the model's predictive capabilities across different atomic arrangements.

We evaluate the model's performance by comparing the predicted Peierls barrier values with the ground truth data, as illustrated in **Figure** 2a. The model achieves a low mean absolute error of 37 meV, indicating that it has effectively captured the fundamental physical mechanisms driving dislocation motion. This accuracy, despite the inherent randomness in the solute distribution, demonstrates the model's ability to reliably predict the Peierls barrier across varying atomic environments. This result highlights the model's capacity to grasp complex atomic interactions that influence the Peierls barrier.

From these predictions, we can compute the mean energy barrier for each composition, a key parameter in solute strengthening theories, which rely on these barriers to predict material strength. Figure 2b compares the mean Peierls barrier values predicted by the DL model against the ground truth values, showing a strong agreement across both ternary and binary alloys. This alignment highlights the model's ability to predict individual configurations while also capturing broader trends in solute-strengthening behavior, offering valuable insights into how solute distributions influence materials properties across various compositions.

In summary, the model's low error and consistent performance across diverse alloy compositions affirm its potential as

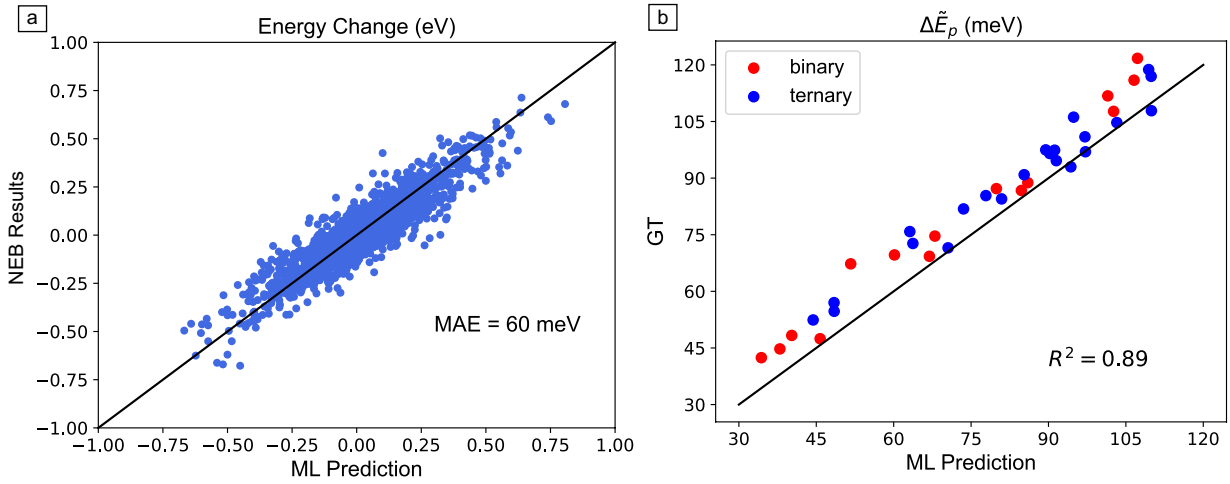

Figure 3. Evaluation of the graph neural network model for the potential energy change prediction on the test set. Test set contains new compositions that never appeared in the training set. (a) Comparison of machine learning (ML) predictions and nudged elastic band (NEB) results for the potential energy change. MAE, mean absolute error. (b) Comparison of the ML results and NEB results for the solute/screw interaction energy parameter, Equation 1 of binary and ternary compositions in the test set. The correlation coefficient R^2 between the predictions and the ground truth (GT) is 0.89.

a powerful tool for predicting material behavior, significantly reducing the need for computationally expensive atomistic simulations such as NEB calculations.

Potential energy change predictions

Another important quantity is the potential energy change as the dislocation moves from one energy minimum to the next. After training the GNN using the potential energy changes as graph labels, we compared the model's predictions with the ground truth values for the graphs in the test set (see SI Figure S1), as shown in **Figure** 3a. The relatively low mean absolute error of 60 meV demonstrates the strong performance of our DL model in accurately predicting potential energy changes.

The random fluctuations in the solute environment lead to a distribution of energy changes. In random alloys, the average of this distribution is expected to be zero, and according to solute strengthening theories, the energy barrier against dislocation motion scales with the standard deviation of this distribution.[24,25] This can be expressed as:

$$\Delta\tilde{E}_p(a) = \left(\frac{b}{\zeta}\right)^{\frac{1}{2}} \sigma_{\Delta U}, \tag{1}$$

where ζ is the dislocation segment length (with $\zeta = 5b$ in our case), and b is the screw dislocation Burgers vector.

Figure 3b shows the values of the solute/screw interaction energy parameter computed by our DL model compared to the ground truth values, demonstrating good agreement for both ternary and binary alloys. This enables the rapid prediction of $\Delta\tilde{E}_p(a)$, a fundamental input parameter in solute-strengthening theories, providing critical insights for predicting material strength in multicomponent alloys.

Multimodal multi-agent model

Figure 4 outlines our multi-agent model for rapid multicomponent alloy design and analysis. Our model builds upon AtomAgents,[53] a multimodal multi-agent system designed to extract new physical insights from atomistic simulations. In this version, AtomAgents has been enhanced with our new GNN model that predicts fundamental physics data. This advancement offers an efficient way to explore the massive compositional space of multicomponent alloys compared to costly atomistic simulations, allowing for solving more challenging problems and rapid design of alloys with enhanced or desired materials properties.

In the current multi-agent framework, the core agents are the "User" and the "AI Assistant," with the User posing the queries and the AI Assistant responding by utilizing various tools. These tools are designed to handle a wide range of tasks, from planning and coding to multimodal data analysis. Additionally, these tools engage a set of autonomous agents that collaborate to respond effectively to the queries provided by the Assistant. A crucial component of this toolset is the physics tools, which empower the model by leveraging both newly developed GNN models and physics-based theoretical frameworks, such as solute-strengthening theories. These theories are essential for connecting atomistic-level features to macroscopic properties such as yield stress. Each agent in our multi-agent system serves a specialized role, defined by a unique profile, and is powered by advanced general-purpose language models from the GPT family, accessed via the OpenAI API. A detailed description of each agent's profile and the implementation details of the AI agents and tools are provided in the "Materials and methods" section.

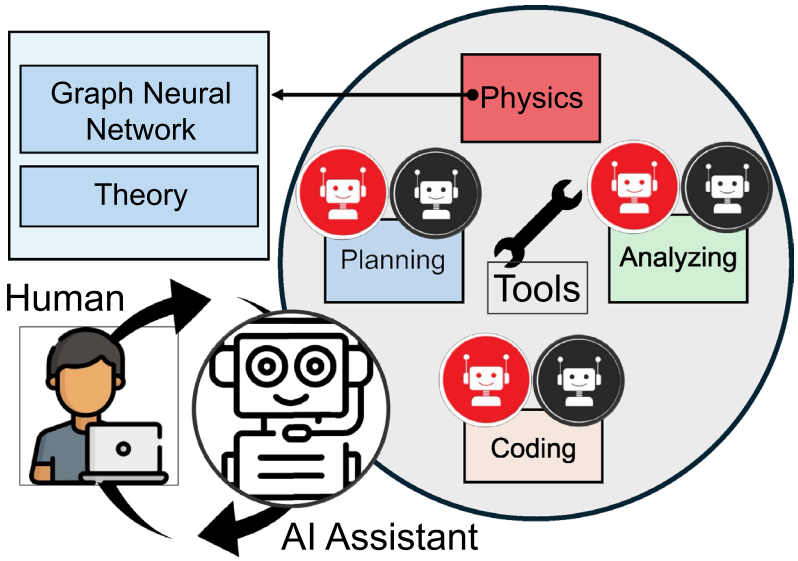

Figure 4. Overview of the graph neural network (GNN)-powered, large language model-based multi-agent system developed here. The system consists of a human and an artificial intelligence (AI) assistant agent at its core, where the human poses queries, and the AI assistant provides responses, seamlessly steering the problem-solving process with the help of integrated tools. These tools are responsible for various tasks, including planning, coding, and multimodal analysis, and each incorporates a set of AI agents that dynamically collaborate to solve complex tasks. A key component is the physics tool, which includes newly developed GNN models to retrieve essential physical parameters (such as the Peierls barrier and potential energy change) as well as physics-based theories (such as solute-strengthening theory). The GNN models enable the rapid prediction of fundamental materials properties, bypassing the need for costly atomistic simulations. The iterative collaboration between agents within the tools and the seamless interaction between the human and AI assistant allows for efficient resolution of complex materials design challenges.

The tools integrated within our multi-agent system enable the resolution of complex challenges in multicomponent alloy design, allowing for automated exploration of the vast compositional design space to identify candidates with superior or targeted performance. These tools are automatically activated by the AI Assistant to perform various tasks throughout the alloy design and analysis process. A general workflow of our model is demonstrated in **Figure** 5. When the User submits a query, the Assistant calls the planning tool to generate a comprehensive plan, detailing the necessary steps, identifying the required tools, and specifying their input parameters. This planning process is driven by two key agents: the planner and the reviewer. The planner drafts the initial version of the plan, while the reviewer critiques and refines it through continuous iterations. This collaborative process ensures that the final plan is precise and thoroughly detailed, which is crucial as the entire problem-solving process depends on this plan.

Once the plan is finalized, the AI Assistant executes it, typically involving tasks such as predicting materials properties using the physics tools, collecting and organizing data, and passing them to the coding tool for visualization and plotting. Finally, the Assistant engages the multimodal agent to conduct a comprehensive analysis of the results.

The advantage of incorporating our newly developed GNN model into the multi-agent system is twofold: (a) it enables the rapid exploration of vast design spaces, providing deeper insights into material behavior across a wide range of compositions, and (b) it significantly accelerates the materials design process by allowing the fast computation of critical input parameters for physics-based theoretical models, such as solute-strengthening theories. To demonstrate the effectiveness of our multi-agent system in these key areas, we conducted two experiments, each detailed next.

Variation of the Peierls barrier in binary and ternary alloys

In this experiment, the multi-agent system was tasked with exploring the Peierls barrier across the ternary compositional space, as illustrated in **Figure** 6a. Upon receiving the query, the assistant agent activates the planning tool to devise a detailed plan for the task. The embedded planner and reviewer agents collaborate, with the planner drafting the initial plan and the reviewer providing feedback. This feedback loop iterates twice to ensure the accuracy and robustness of the plan. The resulting comprehensive plan outlines the following key steps: (1) Composition generation: creating a list of all possible binary and ternary compositions in the Nb–Mo–Ta system at 5% intervals (231 total compositions); (2) Computation of Peierls barrier: using the GNN model to compute the Peierls barrier for all generated compositions; (3) Plotting results: utilizing the coding tool to plot the computed Peierls barrier values on a ternary diagram representing the Nb–Mo–Ta compositional space; and (4) Data analysis: applying the analysis tool to examine the plotted data for trends, correlations, and significant insights. The plan also includes a detailed list of functions and input parameters to ensure consistency and precision during execution.

The Assistant agent then executes the plan, generating the possible compositions in the ternary space and activating the relevant tools, as shown in Figure 6b. We observe that the GNN model enables rapid predictions of the Peierls barrier across the entire compositional space (231 compositions at 5% intervals). Notably, calculating the mean Peierls barrier for a single composition typically requires computationally expensive NEB simulations over numerous random configurations to capture the statistical nature of the problem. The GNN-based approach,

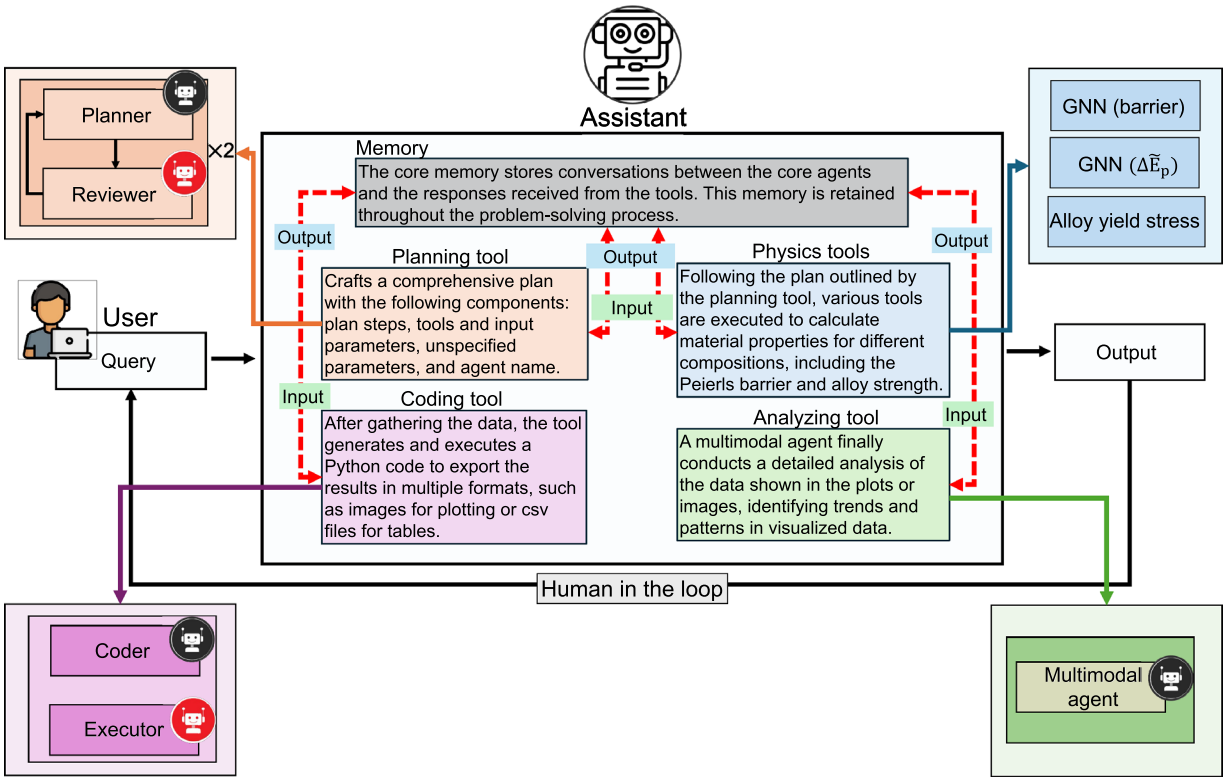

Figure 5. Overview of a typical workflow executed by our multi-agent system for accelerated and automated multicomponent alloy design and analysis. Upon receiving a query from the user, the process begins with the assistant agent, which calls the planning tool to generate a detailed plan. The assistant agent follows this plan by invoking the relevant tools and providing the necessary input functions. These tools are integrated into our system to equip the agent with advanced capabilities, including material property predictions through physics-based retrieval tools (graph neural network [GNN] models), data visualization via code execution (through a code-executor pair of agents), and multimodal analysis (through a multimodal agent-user pair) for enhanced visualization and reasoning. The system incorporates a human-in-the-loop design, fostering dynamic collaboration between humans and AI. This enables continuous refinement of ideas and queries, resulting in a more adaptive and efficient problem-solving process.

however, completes these calculations within seconds per composition, demonstrating significant efficiency. After computing the Peierls barriers, the coding tool is called to write a Python function for plotting the values on a ternary diagram. Despite the complexity of ternary plots, the code generator, powered by the o1-mini reasoning model,[59] produces the plot, as shown in Figure 6c, depicting the variation of the Peierls barrier across Nb–Mo–Ta ternary and binary compositions.

The results of the multi-agent system provide valuable insights into the variation of the Peierls barrier across the entire compositional space. Regions with high Peierls barrier values are identified around compositions with intermediate levels of Nb and Mo and low levels of Ta. As the final step of the workflow, the multimodal agent performs detailed analysis, identifying the highest Peierls barrier in compositions with a high percentage of Mo and the lowest in those with a high percentage of Ta. Additionally, the agent conducts a comparative analysis of the influence of Nb, Mo, and Ta concentrations on the Peierls barrier.

However, some inconsistencies are observed in the agent's analysis. For example, the model suggests that the Peierls barrier decreases with increasing Nb content, which is inaccurate based on the predictions. This highlights challenges in the multi-agent system's ability to detect patterns in complex ternary plots, requiring further experimentation and refinement. The complete results of the ternary plot analysis are presented in SI Figure S5.

The full spectrum of Peierls barrier values across the ternary compositional space offers many opportunities for further insights. For instance, it allows exploration of the barrier's variation within binary systems or specific ternary systems. The human-in-the-loop capability of the multi-agent system makes it possible to issue follow-up queries. Here, we instruct the model to plot the variation of the Peierls barrier with solute concentration for specific binary and ternary systems, including the Nb–Mo and Nb–Ta binary systems, as well as the $(NbMo)_{2x}Ta_{1-2x}$ and $(NbTa)_{2x}Mo_{1-2x}$ ternary systems, as shown in Figure 6e.

The Assistant agent identifies the relevant compositions from prior results and passes their Peierls barrier values to the coding agent for plotting, generating the plots shown in Figure 6f. These plots reveal a nonlinear variation of the Peierls

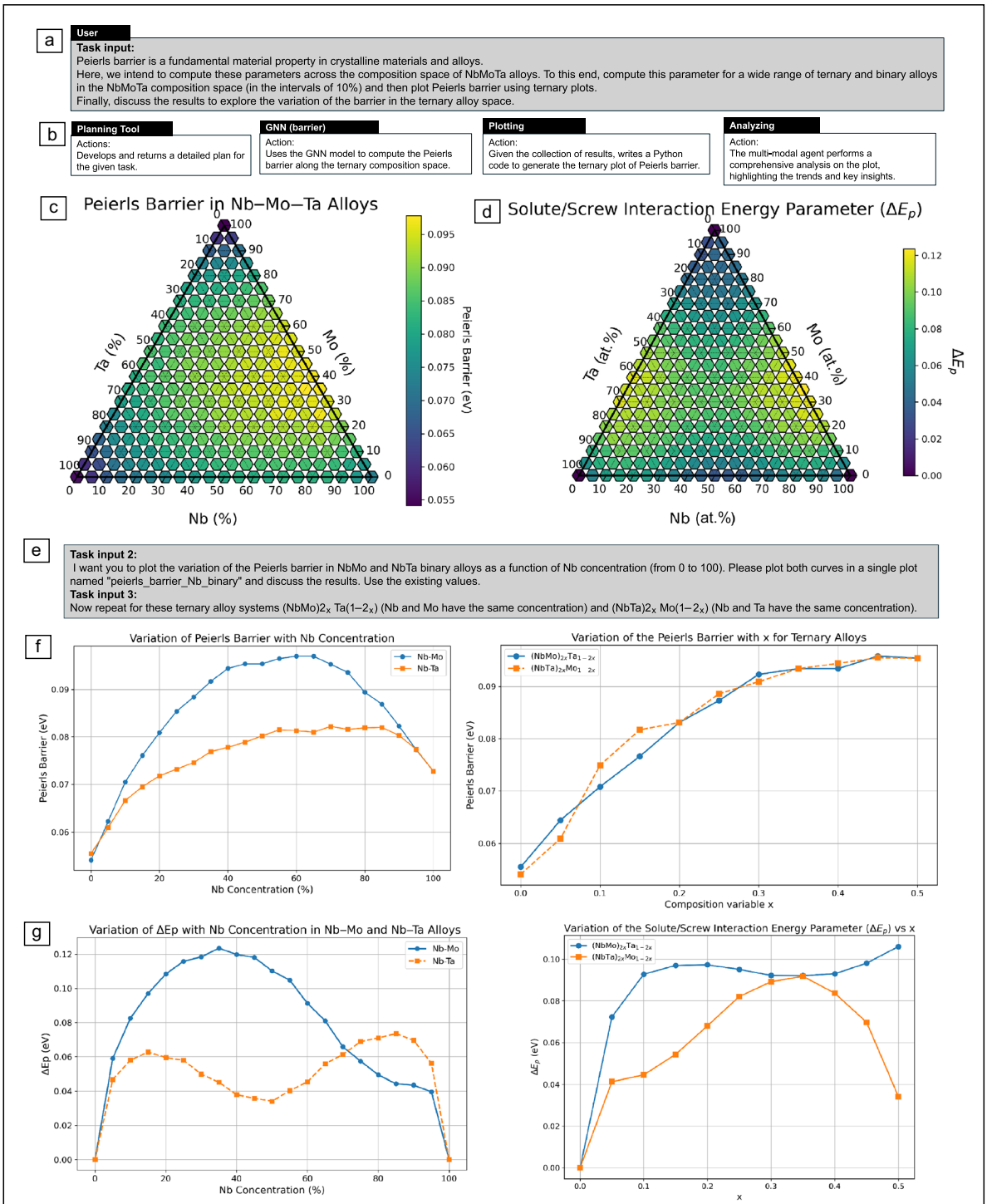

Figure 6. Overview of the multi-agent collaboration to explore the variation of the Peierls barrier and solute/screw interaction energy parameter in binary and ternary alloys. (a) User query requesting exploration of the Peierls barrier across a wide range of compositions, with a similar query repeated for the solute/screw interaction energy parameter. (b) Workflow detailing the computations and analyses performed by the multi-agent system. (c, d) Ternary plots displaying the Peierls barrier and the solute/screw interaction energy parameter, respectively, across the compositional space generated by the multi-agent system. (e) Follow-up tasks requested by the user, involving additional data plotting and analysis. (f, g) Plots illustrating the variation of the Peierls barrier and $\Delta\tilde{E}_p$, respectively, with solute concentration for binary and ternary alloys, generated by the multi-agent model in response to the follow-up tasks.

barrier with solute concentration. Key insights include: (a) In the Nb–Mo system, the Peierls barrier increases with Nb concentration, peaking around 50%, before decreasing as Nb content approaches 100 percent. (b) In the Nb–Ta system, the Peierls barrier also increases with Nb concentration, but at a slower rate. (c) In both ternary systems, the Peierls barrier shows an increasing trend as x rises from 0 to 0.5.

The entire workflow is repeated for the solute/screw interaction energy parameter, $\Delta\tilde{E}p$, a key material property derived from potential energy changes. The problem-solving approach remains identical, but the model now uses the potential energy GNN to compute $\Delta\tilde{E}p$. The results are visualized in the ternary plot of Figure 6d, showing nonlinear variations of the interaction energy parameter with solute concentrations.

Further insights are gained by examining the variation of $\Delta\tilde{E}p$ for selected binary and ternary systems, as shown in Figure 6g. The key findings include: (a) In the Nb–Mo system, $\Delta\tilde{E}p$ increases with Nb concentration, peaking around 30%, and then gradually decreases as Nb content continues to rise. This suggests that the interaction energy parameter is optimized at intermediate Nb concentrations. (b) In the Nb–Ta system, $\Delta\tilde{E}p$ is higher at high Nb concentrations compared to the Nb–Mo system, indicating stronger interactions at these concentrations. (c) In the $(NbTa)_{2x}Mo_{1-2x}$ system, $\Delta\tilde{E}p$ increases steadily with x and decreases as Mo content becomes minimal. (d) In the $(NbMo)_{2x}Ta_{1-2x}$ system, $\Delta\tilde{E}_p$ rises with increasing x, eventually stabilizing, suggesting a saturation point in interaction energy as Ta content diminishes.

These results highlight the complex, nonlinear interactions between solute concentrations and the screw dislocation, which influence both the Peierls barrier and the solute/screw interaction energy parameter. The GNN-powered multi-agent system offers an effective framework for efficiently exploring the vast compositional space of multicomponent alloys. It enables the identification of key trends and intricate behaviors, significantly enhancing the materials design process through faster and more intelligent exploration.

Yield stress in multicomponent bcc alloys

Solute-strengthening theories focus on the atomic-level mechanisms governing dislocation motion and their interactions with solutes, aiming to develop mechanistic models that predict temperature-dependent mechanical strength, specifically the yield stress, of materials. These models typically rely on a number of input parameters derived from first-principles calculations such as density functional theory (DFT) or atomistic simulations based on empirical potentials. A prominent theory for screw dislocation strengthening in bcc alloys is the Maresca–Curtin theory,[24] which requires inputs such as the Peierls potential, the solute/screw interaction energy parameter, the kink formation energy, and vacancy/interstitial formation energies. However, due to the complex energy landscapes in multicomponent systems, accurately computing these parameters poses a significant challenge. The newly developed GNN model addresses this by enabling the rapid and accurate computation of two critical parameters: the Peierls potential and the solute-dislocation interaction energy parameter. This capability allows us to evaluate these parameters in mere seconds for each composition in the NbMoTa alloy space, offering a major breakthrough in efficiently predicting yield stress in multicomponent alloys as a function of temperature.

The objective of this experiment is to demonstrate the potential of our multi-agent system in automating and significantly accelerating the materials design process by seamlessly integrating physics-based theoretical models with advanced DL-driven material prediction tools. In this task, the multi-agent system is assigned the calculation of yield stress for a series of alloy compositions over a broad range of temperatures, as shown in **Figure** 7a. Upon receiving the task from the User, the dynamic collaboration between agents is initiated to accomplish the goal, as illustrated in Figure 7b.

The process begins with the planning tool, which generates a detailed step-by-step plan outlining the required functions and their input parameters. The plan includes the use of GNN-powered tools to predict the Peierls barrier and the solute-dislocation interaction energy—key parameters for yield stress prediction. Additionally, for other input parameters such as the lattice constant, kink formation energy, and vacancy/interstitial formation energies, the system computes their values by averaging over the pure elements. Once all necessary input parameters are obtained, they are fed into the yield stress prediction tool for final calculations. The results are then passed to the coding tool, which generates Python code to plot the predictions. Finally, the multimodal agent, powered by the GPT-4o LLM, analyzes the resulting plots to evaluate trends across compositions.

Despite the complexity of the task, which requires numerous calculations and input parameters, we observe that all sub-tasks are executed seamlessly. The primary outcomes of the multi-agent collaboration are presented in Figure 7c–d, which show the variation of yield stress with temperature for binary and ternary systems, respectively. The binary plot also includes experimental data from the input task for comparison.

Following this, the multimodal agent empowered by GPT-4o conducts a comprehensive analysis of the results, as detailed in SI Figures S6 and S7I, covering binary and ternary alloys, respectively. The analysis indicates a typical decrease in alloy strength with increasing temperature, attributed to enhanced atomic mobility at higher temperatures. Additionally, the AI agent highlights that the Maresca–Curtin model predictions align with experimental data for some alloys, such as Nb81Mo19, but diverge for others, such as Nb95Mo5 and Nb75Mo25. The agent suggests that these discrepancies could result from variations in alloy composition, microstructure, temperature effects, or experimental variability. Furthermore, the AI agent emphasizes the role of solid-solution

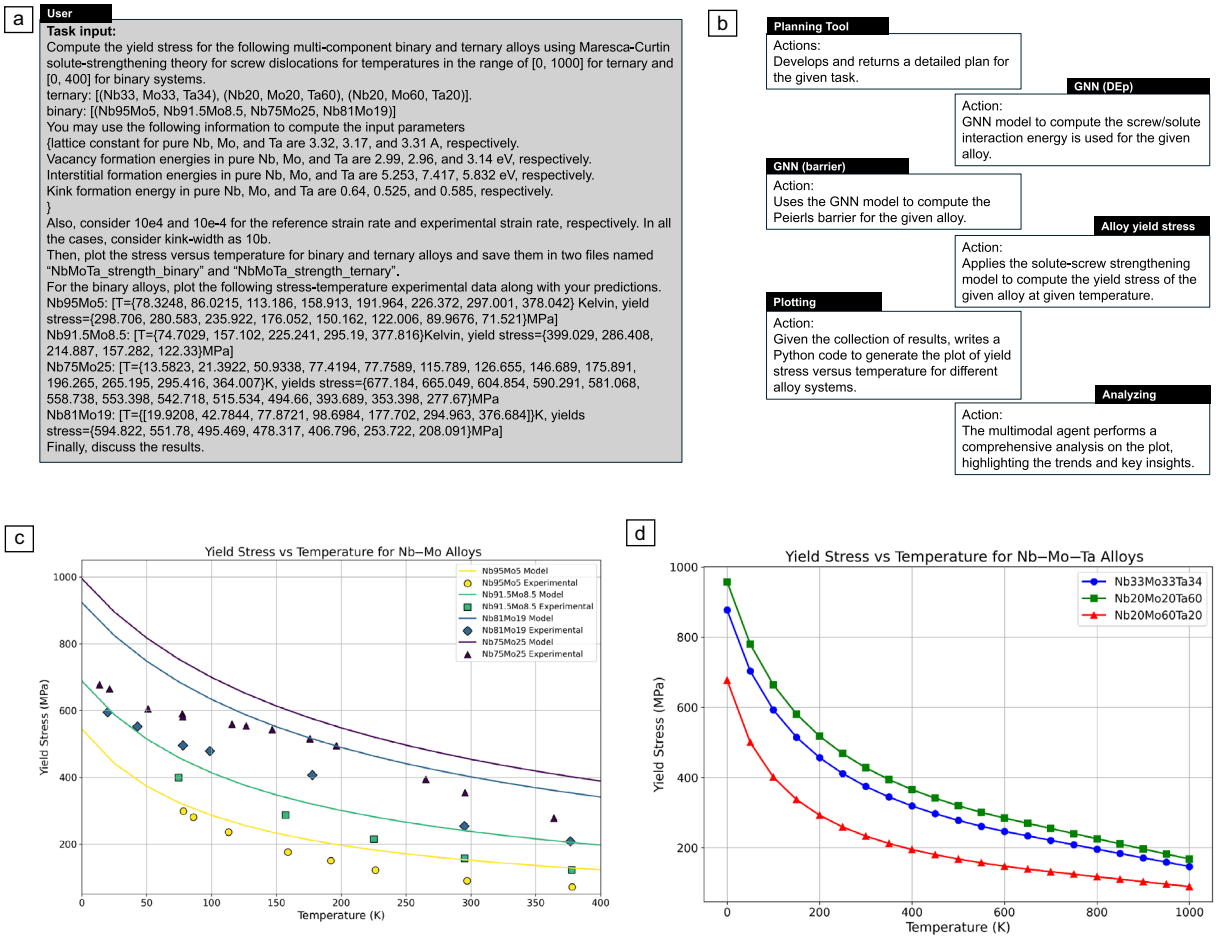

Figure 7. Overview of the multi-agent collaboration to predict the yield stress in binary and ternary alloy body-centered-cubic systems. (a) The input task involving computing the yield stress for a set of compositions. Some materials properties for the pure metals are provided as well as the strain rates. Moreover, the experimental data for binary alloys are provided to compare with the predictions. (b) The workflow of computations performed by our multi-agent system featuring key tasks such as planning, tool calling for material property predictions from the graph neural network (GNN) model, and finally plotting and analyzing the results. (c) Predictions of our model for the ternary alloys and (d) the predictions of our model for binary alloys along with the experimental results.

strengthening, with lattice distortions impeding dislocation motion as a key strengthening mechanism.

It can be observed that the analysis provided by GPT-4o offers a general overview of the topic, focusing on broad trends and general descriptions of behaviors, but it fails to address the intricate details of dislocation behavior in alloys. Its explanation for the discrepancies between experimental results and theoretical predictions is shallow, and the attribution of decreased strength at high temperatures to enhanced atomic mobility, commonly linked to diffusion processes, is inaccurate. A more thorough and detailed analysis is necessary to capture the physical insights essential for advancing scientific research on complex material behaviors.

To further strengthen the multimodal analysis, we reanalyzed the plots using the newly developed o1 model, which has demonstrated advanced reasoning capabilities across domains. The results, detailed in **Figures** 8 and 9, reveal a marked improvement over GPT-4o. For instance, the o1 model provides more accurate explanations for the discrepancies between experimental results and the Maresca–Curtin model, and it better captures the influence of temperature on the reduction in yield strength. These advancements highlight the o1 model's ability to deliver more refined and accurate insights, particularly in domain-specific areas such as solute strengthening.

This multi-agent collaboration exemplifies the rigorous integration of multiple modalities, combining AI-driven physics predictions with established theoretical frameworks, along with plotting and visualization reasoning. This multifaceted approach not only enhances the precision and depth of the materials design process, but also represents a transformative pathway that could revolutionize how materials design is approached and executed.

Plot analysis (o1)

A concise way to view these data is that *all* the alloys show the expected decrease in yield stress as temperature rises, and the model curves faithfully track this overall downward trend. Furthermore, alloys with higher Mo content (e.g., Nb75Mo25) consistently exhibit higher yield strengths across the temperature range than lower-Mo alloys (e.g., Nb95Mo5), and the model correctly captures that ordering.

Where the Maresca–Curtin model appears to diverge is primarily in its **magnitude** of predicted yield stress-especially at lower temperatures and for the higher-Mo compositions. In virtually every alloy, the model curves sit above the experimental points by a fairly systematic offset (ranging from tens of MPa for Nb95Mo5 to as much as a couple hundred MPa for Nb75Mo25 near 50–100 K). The general shape and temperature dependence are largely captured, but the predicted stress levels are higher than measured.

Possible reasons for these discrepancies include:

1. **Assumptions in the Solid-Solution Model**
   Simplifications in the theoretical treatment of dislocation/solute interactions or the temperature dependence of these interactions can lead to systematic overestimates of strengthening.
2. **Microstructural Factors Not Fully Incorporated**
   Real samples may have grain-size effects, impurity levels, or residual stresses that reduce yield strength relative to a "clean" model system.
3. **Experimental Scatter or Material Variability**
   The absence of explicit error bars makes it difficult to judge how much of the offset might be accounted for by data scatter or variations in test specimens.
4. **Temperature-Dependent Mechanisms Not Accounted For**
   Phenomena such as short-range ordering, secondary phase formation, or changes in dislocation substructure at very low temperatures could alter the strength in ways not captured by the baseline model.

Despite these discrepancies, the model does a good job of reproducing the overall trends-namely the strong dependence on both Mo content and temperature. Future refinements that incorporate microstructural and defect-chemistry nuances would likely bring the theoretical and experimental values into better quantitative agreement.

Figure 8. Results of the analysis conducted by the multimodal agent powered by o1 for the plot shown in Figure 7c.

Summary and future perspectives

Our work introduces an innovative framework that seamlessly integrates a GNN into a multimodal, multi-agent artificial intelligence system to form a physics-aware reasoning model that facilitates the design of multi-component bcc alloys. By directly correlating atomic-scale defect structures—specifically those generated by screw dislocations—with key materials properties, our GNN model swiftly predicts critical parameters such as potential energy changes and the Peierls barrier, bypassing the extensive computational cost of traditional atomistic simulations. This rapid predictive capability, combined with an AI system that leverages advanced language models and specialized agents, creates a dynamic and human-in-the-loop process for exploring vast compositional spaces, ultimately paving the way for the accelerated development of alloys with enhanced mechanical performance.

First, we developed a GNN model that directly links dislocated crystalline structures, caused by screw dislocations, to materials properties in multicomponent bcc alloy systems. Despite being trained on a small subset of the massive compositional space in ternary alloys, the model achieves high accuracy in predicting the statistical averages of key properties for each composition, including the average Peierls barrier—which represents the highest energy at the transition state—and the solute/screw interaction energy parameter, a fundamental quantity in solute strengthening theories. Taking the unrelaxed atomic structures as input, the GNN model generates predictions in a matter of seconds, representing a significant improvement over conventional atomistic simulations, which can take days or even months when applied to the large design space of multicomponent alloys. This speed, combined with minimal loss of accuracy, establishes our approach as a feasible alternative to costly simulations for investigating the effects of chemical fluctuations around defects on crystalline materials properties. Furthermore, this method can be easily extended to high-entropy alloys and other crystalline structures, such as face-centered-cubic and hexagonal close-packed systems, as well as to different types of defects, including edge dislocations. Our GNN model enables rapid exploration of the design space, paving the way for the development of alloys

Plot analysis (o1)

From a broad perspective, all three Nb-Mo-Ta alloys exhibit the expected monotonic decrease in yield stress with increasing temperature, consistent with thermally activated dislocation motion. At low temperatures (around 0K), the yield strengths are highest, exceeding 900 MPa for the Nb20Mo20Ta60 alloy, and then smoothly taper off as temperature increases to around 1000 K. A few key observations and insights are as follows:

1. Relative Ordering of Yield Strengths
   - Nb20Mo20Ta60 (Green squares) shows the highest yield stress over the full temperature range. Near 0K it approaches 1000 MPa, and remains above 200 MPa at 1000 K.
   - Nb33Mo33Ta34 (Blue circles) exhibits intermediate yield strength-lower than the high-Ta alloy at all temperatures, but still well above the red data set at lower temperatures.
   - Nb20Mo60Ta20 (Red triangles) consistently has the lowest yield strength of the three, topping out around 600+MPa at cryogenic temperatures and tapering off to about 100 MPa near 1000 K.
2. Effect of Composition
   - The alloy with 60%at. Ta (Nb20Mo20Ta60) clearly benefits from Tantalum's relatively large atomic size and strong lattice-distortion effect, leading to higher solution strengthening.
   - Conversely, the alloy with 60%at. Mo (Nb20Mo60Ta20) sees less net strengthening-even though Mo is also a strong solute, its strengthening contribution in these specific ternary combinations is evidently overshadowed by Ta's effect when Ta is present at higher concentrations.
   - The "intermediate" composition (Nb33Mo33Ta34) lands in between these two extremes in both composition and yield stress.
3. Temperature Dependence and Possible Mechanisms
   - All three curves follow a nonlinear decay in yield stress with temperature-a hallmark of thermally activated dislocation glide in bcc alloys.
   - At low temperatures, strength is dominated by lattice resistance and the elastic strain fields around solute atoms (solute drag). The high yield stresses in these alloys demonstrate the substantial barrier to dislocation motion introduced by the large size mismatch among Nb, Mo, and Ta.
   - As temperature increases, thermal energy aids dislocation movement over or around the solute obstacles, thus reducing the overall yield strength.
   - Tantalum's large atomic radius mismatch relative to Nb gives it a strong solid-solution strengthening effect, while Mo's effect, although still significant, is apparently less dominant under these conditions.
4. Insights Into Solute Strengthening
   - Atomic Size Mismatch: Larger mismatches (as with Ta in Nb) typically yield higher strengthening through greater local lattice distortion and hence stronger dislocation pinning.
   - Modulus Mismatch: Differences in shear moduli among Nb, Mo, and Ta can also contribute to extra dislocation drag.
   - Concentration Effects: Higher solute content increases the density of pinning sites, but the specific solute species matters; the data suggest that Ta is especially potent in raising yield strength compared to Mo at equivalent concentrations.

Overall, these results highlight that Ta content is the key contributor to higher yield strengths in Nb-Mo-Ta alloys. The temperature dependence underscores typical thermally activated flow processes of bcc metals. Going forward, optimizing Nb-Mo-Ta compositions for specific temperature ranges would likely hinge on balancing Ta (stronger strengthening) against factors such as density, cost, and ductility.

Figure 9. Results of the analysis conducted by the multimodal agent powered by o1 for the plot shown in Figure 7d.

with enhanced mechanical properties, particularly those with higher strength.

Next, we constructed a multimodal AI system that integrates three core components: (1) LLMs, which excel at a wide range of tasks, including multimodal reasoning, strategic planning, rational thinking, and even coding; (2) AI agents, each powered by LLMs and external tools, designed with specialized roles and expertise, working collaboratively in a dynamic environment and autonomously tackling multifaceted problems; and (3) a GNN model that provides rapid predictions of fundamental materials properties. This multiagent system is an enhanced version of our previous model, AtomAgents,[53] which relied on direct atomistic simulations for physics-based insights. In the modified version, the GNN model is integrated to mitigate the high computational cost associated with atomistic simulations, thereby enabling faster exploration of the vast compositional design space inherent to multi-component alloys. This new framework offers a comprehensive method for addressing more complex problems in the design and analysis of multicomponent alloys. Its capabilities are demonstrated through a series of experiments, including the exploration of the Peierls barrier and solute/screw interaction energy across the entire ternary composition space, as well as predictions of yield stress in bcc alloys, which are then validated against experimental data.

Additional insights emerging from our experimental results further underscore the transformative potential of our approach. The consistency with which the GNN model captures nonlinear trends in both binary and ternary systems not only validates the underlying graph representation of atomic configurations, but also highlights its robustness across a wide compositional spectrum. Notably, the rapid feedback provided by the model allows for iterative refinement of theoretical frameworks such as the Maresca–Curtin model, pinpointing discrepancies that can guide further model calibration. This synergy between fast, data-driven predictions and established physics-based theories enables a more efficient pathway for inverse design, where desired macroscopic properties can be backtracked to specific compositional windows. Consequently, our system is well-positioned to serve as a versatile platform for hypothesis testing and accelerated materials optimization, bridging the gap between computational discovery and experimental validation for diverse applications in materials research.

Our automated multi-agent system incorporates several key features that make it highly effective for complex materials science challenges. First, it is powered by advanced reasoning models such as o1-preview and o1-mini,[59] which provide exceptional capabilities in reasoning, strategic planning, and coding. These models, having shown a remarkable trajectory of improvements in complex tasks such as coding and logic, excel at handling multimodal data, evaluating emerging trends, and integrating information from diverse sources, making them essential for solving multiscale problems. Second, the system supports the seamless integration of various data modalities, incorporating theoretical models, sophisticated deep learning algorithms, and even generative tools into the design and analysis process. This enables our system to stay at the cutting-edge of scientific and technological advancements, providing a more holistic perspective on materials design. Third, the system is designed to support continuous evaluation and refinement, allowing for iterative problem-solving where errors are identified and corrected through mutual feedback between AI agents, ensuring that the system remains robust and adaptive throughout the problem-solving process. Finally, the human-in-the-loop functionality embedded within the system further enhances its flexibility and usability. This feature allows users to interact with the system directly, providing feedback, refining model parameters, or introducing additional requests. This AI-human collaboration is particularly valuable for complex, multifaceted problems, ensuring that the system remains adaptable and responsive to specific user needs. Together, these advancements position the multi-agent modeling of materials as a powerful and innovative tool for tackling some of the most challenging problems in materials science and engineering, traditionally reserved for human experts.

Furthermore, the multi-agent system's high degree of adaptability makes it suitable for application across a broad range of domains and disciplines, extending its potential impact beyond materials science. One of the primary aspects of this adaptability lies in the underlying LLM, which orchestrates the system's internal operations and interactions. The performance of the entire system is heavily influenced by the capabilities of this LLM, and as LLM technology continues to evolve, we anticipate even greater improvements in system performance. For example, the transition from earlier versions to o1 resulted in significant gains in efficiency and accuracy, as demonstrated here for the plot analysis provided in Figures 8 and 9. This ongoing advancement in LLMs suggests that future iterations of the system could become even more capable, especially as LLMs are further customized to specific tasks and integrated into different agents. Additionally, the system's flexibility extends to its ability to incorporate a wide range of deep learning and generative models, each developed for various disciplines. This capability allows the system to address domain-specific challenges while maintaining a unified problem-solving framework. Another important feature of this system is its capacity to integrate with other multi-agent systems, such as SciAgents,[42,60] which specialize in generating advanced research hypotheses. This creates a cohesive scientific discovery process, where one system is responsible for exploring the hidden space of ideas, generating novel hypotheses, while another system is equipped to test, validate, or refine these ideas using advanced tools. Together, this interconnected approach provides a comprehensive and rigorous pathway for scientific discovery, making the multi-agent system an adaptable and powerful framework for tackling interdisciplinary problems.

Building on these capabilities, future work could pursue several promising directions. For example, the system's knowledge and reasoning abilities could be enhanced by integrating retrieval-augmented generation, providing access to scholarly databases such as arXiv, or incorporating domain-specific LLMs fine-tuned for specialized areas[61,62] such as solute-strengthening theories. In addition, the flexibility of the multi-agent framework naturally allows for the incorporation of established computational methods, such as genetic algorithms and active learning, enabling the system to combine traditional optimization strategies with advanced AI-driven reasoning. While our current approach relies on established, human-developed solute-strengthening theories, a particularly exciting future direction would be to develop foundational scientific multi-agent models capable of autonomously formulating, refining, or improving physics-based theories.[54] This would represent a significant step toward AI-driven scientific discovery and a deeper understanding in materials science and beyond.

Materials and methods

Data set generation

Random compositions of ternary NbMoTa refractory alloys with screw dislocation are generated to compute the potential energy and Peierls barrier as follows. First, pristine crystal structures are generated by creating a rectangular simulation cell using the corresponding lattice constant and then randomly assigning solutes based on the desired concentrations. The simulation cell is oriented with dislocation glide direction x||$[1, 1, \bar{2}]$, glide plane normal direction, y||$[\bar{1}10]$, and dislocation line direction, z||[111] with periodic boundary conditions along x and z and free surface along y. This crystallographic orientation corresponds to the preferential glide system for screw dislocations in bcc alloys, including multicomponent systems, as established by both experiments and simulations. Atomic positions are relaxed by using a combination of the FIRE algorithm[63] and relaxation of the cell dimensions until the convergence is achieved-the norm of the force vector fell below 10e−6 eV/Å and stresses σ_{xx}, σ_{xz}, and σ_{zz} fell below 0.1 MPa. All the atomistic simulations are performed using large-scale atomic/molecular massively parallel simulator[64] at zero temperature. A ML moment tensor potential is used to describe the interatomic interactions in the NbMoTa alloys.[33,65]

We then introduce a screw dislocation in the center of each relaxed simulation cell using the periodic array of dislocations method[16] and minimize the energy along with the relaxation of the pressures. This serves as our initial dislocated configuration. To generate the final dislocated configurations, we use the same initial pristine random structure and insert the dislocation at the adjacent position (next Peierls valley) relative to the initial dislocation at the distance of a along the glide direction where a is the Peierls valley distance. The potential energy changes are computed by subtracting the total potential energies of final and initial configurations. We then perform NEB computations[66–69] as implemented in LAMMPS[64] on these initial and final structures to compute the minimum energy path among these two energy states. The maximum value along this curve is stored as the corresponding Peierls barrier. The entire process is repeated for all the compositions in the training set (200 realizations for each composition) and test set (50 realizations for each composition) as depicted in SI Figure S1.

Graph representation

The results from atomistic simulations of atomic configurations and alloy properties are represented as graphs, which are used as inputs and labels to train the GNN. The graph structure is constructed from the pristine pure Mo configuration, where each Mo atom is a node in the graph, and the connectivity between nodes is determined by the distance between atoms. We consider a cylindrical region with a radius of $r_c = 16$ Å, centered at the screw dislocation core, and only include atoms within this region as graph nodes. Within this graph, two atoms are connected if their distance is below the cutoff distance of 2.8 Å, which is the distance between nearest neighbors in a perfect crystal of molybdenum. The graph representation consists of 705 nodes and 2610 edges.

The node features contain information about chemicals (solute type) and structural defect (screw dislocation). The structural defect feature is constructed from the z-component of the displacement derived from the atomistic simulation in pure Mo. This feature remains identical for all the atomic configurations in the data set, eliminating the need for atomic relaxation for new configurations during inference. The chemical feature of the node is uniquely determined for each composition represented as the solute types converted to one-hot encoded representations ([1,0,0], [0,1,0], and [0,0,1] for Nb, Mo, Ta, respectively). Moreover, the edge features representing the bond type are encoded as one-hot representations based on the neighboring nodes as shown in **Table I**.

GNNs

The GNN model is developed based on the DL framework PyTorch[70] and its geometric extension library PyTorch Geometric.[71] The GNN architecture is depicted in Figure 1b. First of all, the input graph is sent to the input block which uses a combination of PNA convolutional (PNAConv) layer, gated recurrent unit cell (GRUCell), and Batch Normalization layer (BatchNorm) to upscale the dimension of node features to 50. The graph is then passed to the message passing block, which contains 10-time repetition of the combined layers. Within the block, nodes communicate with each other by passing the messages given the node features plus connectivity and update their own node features considering the received messages. The last MLP that returns the predicted Peierls barrier or potential energy change has a structure of an input layer of size 30, a hidden layer of size 20, another hidden layer of size 10, and an output layer of one neuron. ReLU is adopted as the activation function in this MLP. We adopt default weight and bias initialization of all of the layers in the model defined by PyTorch Geometric.

Table I. Edge feature representation for bond type connecting neighboring nodes, *i* and *j*.

Node Type *i*	Node Type *j*	Edge Feature
Nb	Nb	[1,0,0,0,0,0]
Mo	Mo	[0,1,0,0,0,0]
Ta	Ta	[0,0,1,0,0,0]
Nb (Mo)	Mo (Ta)	[0,0,0,1,0,0]
Nb (Ta)	Ta (Nb)	[0,0,0,0,1,0]
Mo (Ta)	Ta (Mo)	[0,0,0,0,0,1]

Model training and evaluation

All data sets created for the training/validation compositions (see SI Figure S1) are split into the train set (90% data) and validation set (10% data). The models were trained with a batch size of 32 using the Adam optimization method[72] for 250 epochs on one NVIDIA Tesla V100s with 32 GB memory. Training starts with a learning rate of 0.0005, and a dynamic learning rate scheduler named ReduceLROnPlateau reduces the learning rate by half if no improvement is seen for 10 epochs to minimize the validation MSE. The learning curves of the GNNs are shown in SI Figure S4 which indicate the convergence of training for both models.

Solute-strengthening theory in bcc alloys

We apply the Maresca–Curtin screw strengthening theory for nondilute to high-entropy alloys to compute the yield stress.[24] The theory is established based on the assumption that the initially straight dislocation becomes spontaneously kinked at zero load and zero temperature so as to lower its total energy. Three mechanisms contribute to the screw dislocation strengthening; (1) Peierls-like mechanism, (2) kink glide mechanism, and (3) cross-kink formation and unpinning and the alloy strength at temperature T is expressed by

$$\tau(\dot{\epsilon}, T) = \tau_{xk}(\dot{\epsilon}, T) + \min[\tau_k(\dot{\epsilon}, T), \tau_p(\dot{\epsilon}, T)], \qquad (2)$$

where $\dot{\epsilon}$ is the experimental strain rate and τ_p, τ_k, and τ_{xk} are the Peierls strength, kink migration strength, and cross-kink unpinning strength, respectively.[21,24]

Agent design

We design AI agents using frontier all-purpose AI models (such as GPT-4o, reasoning models such as o1). Dynamic multi-agent collaboration is implemented in AutoGen framework,[73] an open-source ecosystem for agent-based AI modeling. Additional agents are introduced as described next.

In our multi-agent system, the human *user* agent is constructed using UserProxyAgent class from Autogen, and *Assistant*, *Planner*, *Reviewer*, *coder* agents are created via AssistantAgent class from Autogen, while *multimodal* agent is constructed via MultimodalConversableAgent class. Each agent is assigned a role through a profile description as shown in SI Figures S8–S12, included as *system_message* at their creation.

Function and tool design

All the tools implemented in this work are defined as Python functions. Each function is characterized by a name, a description, and input properties. The full list of tools and their descriptions can be found in the corresponding codes.

Acknowledgments

We acknowledge support from USDA (2021-69012-35978), DOE-SERDP (WP22-S1-3475), ARO (79058LSCSB, W911NF-22-2-0213, and W911NF2120130), and MIT's Generative AI Initiative. A.G. gratefully acknowledges the financial support from the Swiss National Science Foundation (#P500PT_214448).

Author contributions

M.J.B and A.G. conceived the overall concept. A.G and M.J.B developed the GNN model and multi-agent system. A.G. curated the training and testing data for the GNN model, performed the tests for various problems, analyzed the results, and prepared the first draft of the paper. M.J.B supported the analysis, revised and finalized the paper with A.G.

Funding

'Open Access funding provided by the MIT Libraries'.

Data and code availability

All data and codes are available on GitHub at https://github.com/lamm-mit/AlloyAgents. Alternatively, they will be provided by the corresponding author upon request.

Conflict of interest

No conflict of interest to be declared.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1557/s43577-025-00953-4.

Open Access

References

1. Y.F. Ye, Q. Wang, J. Lu, C.T. Liu, Y. Yang, *Mater. Today* **19**(6), 349 (2016)
2. D.B. Miracle, O.N. Senkov, *Acta Mater.* **122**, 448 (2017)
3. E.P. George, D. Raabe, R.O. Ritchie, *Nat. Rev. Mater.* **4**(8), 515 (2019)
4. E.P. George, W.A. Curtin, C.C. Tasan, *Acta Mater.* **188**, 435 (2020)
5. O.N. Senkov, G.B. Wilks, D.B. Miracle, C.P. Chuang, P.K. Liaw, *Intermetallics* **18**(9), 1758 (2010)
6. O.N. Senkov, G.B. Wilks, J.M. Scott, D.B. Miracle, *Intermetallics* **19**(5), 698 (2011)
7. O.N. Senkov, J.M. Scott, S.V. Senkova, D.B. Miracle, C.F. Woodward, *J. Alloys Compd.* **509**(20), 6043 (2011)
8. O.N. Senkov, D.B. Miracle, K.J. Chaput, J.P. Couzinie, *J. Mater. Res.* **33**(19), 3092 (2018)
9. O.N. Senkov, S. Rao, K.J. Chaput, C. Woodward, *Acta Mater.* **151**, 201 (2018)
10. Z.D. Han, H.W. Luan, X. Liu, N. Chen, X.Y. Li, Y. Shao, K.F. Yao, *Mater. Sci. Eng. A* **712**, 380 (2018)
11. O.N. Senkov, S. Gorsse, D.B. Miracle, *Acta Mater.* **175**, 394 (2019)
12. W. Xiong, A.X.Y. Guo, S. Zhan, C.T. Liu, S.C. Cao, *J. Mater. Sci. Technol.* **142**, 196 (2023)
13. C. Varvenne, A. Luque, W.A. Curtin, *Acta Mater.* **118**, 164 (2016)
14. C. Varvenne, G.P.M. Leyson, M. Ghazisaeidi, W.A. Curtin, *Acta Mater.* **124**, 660 (2017)
15. S.I. Rao, E. Antillon, C. Woodward, B. Akdim, T.A. Parthasarathy, O.N. Senkov, *Scr. Mater.* **165**, 103 (2019)
16. A. Ghafarollahi, W.A. Curtin, *Acta Mater.* **196**, 635 (2020)
17. A. Ghafarollahi, W.A. Curtin, *Acta Mater.* **215**, 117078 (2021)
18. F. Maresca, W.A. Curtin, *Acta Mater.* **182**, 235 (2020)
19. S.I. Rao, C. Woodward, B. Akdim, O.N. Senkov, D. Miracle, *Acta Mater.* **209**, 116758 (2021)
20. R.E. Kubilay, A. Ghafarollahi, F. Maresca, W.A. Curtin, *NPJ Comput. Mater.* **7**(1), 112 (2021)
21. A. Ghafarollahi, W.A. Curtin, *Acta Mater.* **226**, 117617 (2022)
22. C. Baruffi, F. Maresca, W.A. Curtin, *MRS Commun.* **12**(6), 1111 (2022)
23. Y. Rao, C. Baruffi, A. De Luca, C. Leinenbach, W.A. Curtin, *Acta Mater.* **237**, 118132 (2022)
24. F. Maresca, W.A. Curtin, *Acta Mater.* **182**, 144 (2020)
25. A. Ghafarollahi, F. Maresca, W.A. Curtin, *Model. Simul. Mater. Sci. Eng.* **27**(8), 085011 (2019)
26. Y. LeCun, Y. Bengio, G. Hinton, *Nature* **521**(7553), 436 (2015)
27. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *NPJ Comput. Mater.* **3**(1), 54 (2017)
28. K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **559**(7715), 547 (2018)
29. J. Wei, X. Chu, X.Y. Sun, K. Xu, H.X. Deng, J. Chen, Z. Wei, M. Lei, *InfoMat* **1**(3), 338 (2019)
30. D. Morgan, R. Jacobs, *Annu. Rev. Mater. Res.* **50**(1), 71 (2020)
31. K. Guo, Z. Yang, C.H. Yu, M.J. Buehler, *Mater. Horiz.* **8**(4), 1153 (2021)
32. J. Behler, *J. Chem. Phys.* **145**(17), 170901 (2016)
33. A.V. Shapeev, *Multiscale Model. Simul.* **14**(3), 1153 (2016)
34. V.L. Deringer, M.A. Caro, G. Csányi, *Adv. Mater.* **31**(46), 1902765 (2019)
35. T. Mueller, A. Hernandez, C. Wang, *J. Chem. Phys.* **152**(5), 050902 (2020)
36. M.J. Buehler, *J. Mech. Phys. Solids* **181**, 105454 (2023)
37. T. Xie, J.C. Grossman, *Phys. Rev. Lett.* **120**(14), 145301 (2018)
38. X. Wang, L. Valdevit, P. Cao, *Acta Mater.* **267**, 119696 (2024)
39. Y.C. Hsu, C.H. Yu, M.J. Buehler, *Matter* **3**(1), 197 (2020)
40. Z. Yang, M.J. Buehler, *NPJ Comput. Mater.* **8**(1), 198 (2022)
41. K. Guo, M.J. Buehler, *Digit. Discov.* **1**(3), 277 (2022)
42. A. Ghafarollahi, M.J. Buehler, *Adv. Mater.* **37**, 2413523 (2024)
43. Q. Pei, L. Wu, K. Gao, J. Zhu, Y. Wang, Z. Wang, T. Qin, R. Yan, Leveraging biomolecule and natural language through multi-modal learning: A survey (2024), Preprint, arXiv:2403.01528
44. T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N.V. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: A survey of progress and challenges (2024), Preprint, arXiv:2402.01680
45. J. Xie, Z. Chen, R. Zhang, X. Wan, G. Li, Large multimodal agents: A survey (2024), Preprint, arXiv:2402.15116
46. Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, X. He, Exploring large language model based intelligent agents: Definitions, methods, and prospects (2024), Preprint, arXiv:2401.03428
47. A.M. Bran, S. Cox, O. Schilter, C. Baldassari, A.D. White, P. Schwaller, *Nat. Mach. Intell.* **6**(5), 525 (2024)
48. M.J. Buehler, *ACS Eng. Au* **4**(2), 241 (2024)
49. B. Ni, M.J. Buehler, *Extreme Mech. Lett.* **67**, 102131 (2024)
50. M.J. Buehler, *NPJ Artif. Intell.* **1**(1), 4 (2025)
51. I. Stewart, M.J. Buehler, *Mol. Syst. Des. Eng.* **10**(4), 314 (2025)
52. A. Ghafarollahi, M.J. Buehler, *Digit. Discov.* **3**, 1389 (2024)
53. A. Ghafarollahi, M.J. Buehler, *Proc. Natl. Acad. Sci. U.S.A.* **122**(4), e2414074122 (2025)
54. A. Ghafarollahi, M.J. Buehler, Sparks: Multi-agent artificial intelligence model discovers protein design principles (2025), Preprint, arXiv:2504.19017
55. G. Corso, L. Cavalleri, D. Beaini, P. Liò, P. Veličković, *Adv. Neural. Inf. Process. Syst.* **33**, 13260 (2020)
56. T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks (2016), Preprint, arXiv:1609.02907
57. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks (2018), Preprint, arXiv:1710.10903
58. K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks? (2018), Preprint, arXiv:1810.00826
59. OpenAI and:, A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A.T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C.M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F.P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H.W. Chung, I. Kivlichan, I. O'Connell, I. Osband, I.C. Gilaberte, I. Akkaya, I. Kostrikov, I. Sutskever, I. Kofman, J. Pachocki, J. Lennon, J. Wei, J. Harb, J. Twore, J. Feng, J. Yu, J. Weng, J. Tang, J. Yu, J.Q. Candela, J. Palermo, J. Parish, J. Heidecke, J. Hallman, J. Rizzo, J. Gordon, J. Uesato, J. Ward, J. Huizinga, J. Wang, K. Chen, K. Xiao, K. Singhal, K. Nguyen, K. Cobbe, K. Shi, K. Wood, K. Rimbach, K. Gu-Lemberg, K. Liu, K. Lu, K. Stone, K. Yu, L. Ahmad, L. Yang, L. Liu, L. Maksin, L. Ho, L. Fedus, L. Weng, L. Li, L. McCallum, L. Held, L. Kuhn, L. Kondraciuk, L. Kaiser, L. Metz, M. Boyd, M. Trebacz, M. Joglekar, M. Chen, M. Tintor, M. Meyer, M. Jones, M. Kaufer, M. Schwarzer, M. Shah, M. Yatbaz, M.Y. Guan, M. Xu, M. Yan, M. Glaese, M. Chen, M. Lampe, M. Malek, M. Wang, M. Fradin, M. McClay, M. Pavlov, M. Wang, M. Wang, M. Murati, M. Bavarian, M. Rohaninejad, N. McAleese, N. Chowdhury, N. Chowdhury, N. Ryder, N. Tezak, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, P. Chao, P. Ashbourne, P. Izmailov, P. Zhokhov, R. Dias, R. Arora, R. Lin, R.G. Lopes, R. Gaon, R. Miyara, R. Leike, R. Hwang, R. Garg, R. Brown, R. James, R. Shu, R. Cheu, R. Greene, S. Jain, S. Altman, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Hernandez, S. Baker, S. McKinney, S. Yan, S. Zhao, S. Hu, S. Santurkar, S.R. Chaudhuri, S. Zhang, S. Fu, S. Papay, S. Lin, S. Balaji, S. Sanjeev, S. Sidor, T. Broda, A. Clark, T. Wang, T. Gordon, T. Sanders, T. Patwardhan, T. Sottiaux, T. Degry, T. Dimson, T. Zheng, T. Garipov, T. Stasi, T. Bansal, T. Creech, T. Peterson, T. Eloundou, V. Qi, V. Kosaraju, V. Monaco, V. Pong, V. Fomenko, W. Zheng, W. Zhou, W. McCabe, W. Zaremba, Y. Dubois, Y. Lu, Y. Chen, Y. Cha, Y. Bai, Y. He, Y. Zhang, Y. Wang, Z. Shao, Z. Li, OpenAI o1 system card (2024), Preprint, arXiv:2412.16720
60. M.J. Buehler, *Mach. Learn. Sci. Technol.* **5**(3), 035083 (2024)
61. M.J. Buehler, *Appl. Mech. Rev.* **10**, 1 (2023)
62. R.K. Luu, M.J. Buehler, *Adv. Sci.* **11**(10), 2306724 (2024)
63. E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, P. Gumbsch, *Phys. Rev. Lett.* **97**(17), 170201 (2006)
64. S. Plimpton, *J. Comput. Phys.* **117**(1), 1 (1995)
65. S. Yin, Y. Zuo, A. Abu-Odeh, H. Zheng, X.G. Li, J. Ding, S.P. Ong, M. Asta, R.O. Ritchie, *Nat. Commun.* **12**(1), 4873 (2021)
66. E. Maras, O. Trushin, A. Stukowski, T. Ala-Nissila, H. Jonsson, *Comput. Phys. Commun.* **205**, 13 (2016)
67. A. Nakano, *Comput. Phys. Commun.* **178**(4), 280 (2008)
68. G. Henkelman, B.P. Uberuaga, H. Jónsson, *J. Chem. Phys.* **113**(22), 9901 (2000)
69. G. Henkelman, H. Jónsson, *J. Chem. Phys.* **113**(22), 9978 (2000)
70. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala. PyTorch: An imperative style, high-performance deep learning library (2019), Preprint, arXiv:1912.01703
71. M. Fey, J.E. Lenssen, Fast graph representation learning with PyTorch Geometric (2019), Preprint, arXiv:1903.02428
72. D.P. Kingma, J.L. Ba, "Adam: A Method for Stochastic Optimization," 3rd International Conference on Learning Representations, in *ICLR 2015-Conference Track Proceedings* (San Diego, 2015)
73. C. Zhu, B. Chen, H. Guo, H. Xu, X. Li, X. Zhao, W. Zhang, Y. Yu, R. Tang, "Autogen: An Automated Dynamic Model Generation Framework for Recommender System," in *WSDM '23: Proceedings of the 16th ACM International Conference on Web Search and Data Mining* (Association for Computing Machinery, New York, 2023), pp. 598–606