

AtomGPT: Atomistic Generative Pre-trained Transformer for Forward and Inverse Materials Design

Kamal Choudhary*

*Material Measurement Laboratory, National Institute of Standards and Technology,
Gaithersburg, MD 20899, USA*

E-mail: kamal.choudhary@nist.gov

Abstract

Large language models (LLMs) such as generative pretrained transformers (GPTs) have shown potential for various commercial applications, but their applicability for materials design remains underexplored. In this article, AtomGPT is introduced as a model specifically developed for materials design based on transformer architectures, demonstrating capabilities for both atomistic property prediction and structure generation. This study shows that a combination of chemical and structural text descriptions can efficiently predict material properties with accuracy comparable to graph neural network models, including formation energies, electronic bandgaps from two different methods, and superconducting transition temperatures. Furthermore, AtomGPT can generate atomic structures for tasks such as designing new superconductors, with the predictions validated through density functional theory calculations. This work paves the way for leveraging LLMs in forward and inverse materials design, offering an efficient approach to the discovery and optimization of materials.

In recent years, machine learning (ML) has emerged as a powerful tool in materials science, offering the potential to accelerate materials discovery by learning from existing data and making predictions for new materials.¹⁻³ Such ML models can be used for forward design tasks such as given an atomic structure predict its properties, or inverse design tasks such as given properties predict its atomic structures. Among various ML approaches such as descriptor-based, graph-based and text-based methods, large language models (LLMs) have shown promises to the exploration of vast materials design space beyond the limitations of traditional methods.⁴⁻⁶

The Generative Pre-trained Transformer (GPT) is a type of large language model (LLM) originally developed for natural language processing and has demonstrated remarkable success in generating coherent and contextually relevant text.⁷⁻⁹ Models such as ChatGPT¹⁰ have been used for code generation/debugging, literature review and numerous other tasks. However, if we attempt to perform forward/inverse materials design tasks, the outcomes can be quite poor.¹¹⁻¹³ Nevertheless, inspired by its simplicity of usage and the massive success of ChatGPT, an alternate model AtomGPT (Atomistic Generative Pre-trained Transformer) is introduced tailored for forward and inverse materials design.

There have been several previous research works for forward prediction material properties such as descriptor-based methods including models in MatMiner,¹⁴ Classical Force-field Inspired Descriptors (CFID);¹⁵ graphs-based methods such as Crystal graph convolutional neural networks (CGCNN),¹⁶ MatErials Graph Network (MEGNet),¹⁷ Atomistic Line Graph Neural Network (ALIGNN),¹⁸ CoGN,¹⁹ M3GNet,²⁰ Matformer,²¹ ComFormer²²; and language-based methods such as LLM-Prop²³ etc. Similarly, some of the earlier inverse design methods are crystal diffusional variational autoencoder (CDVAE),²⁴ Fourier-transformed Crystal Properties (FTCP),²⁵ G-SchNet,²⁶ MatBERT,²⁷ CrystaLLM,²⁸ Crystal-LLM,²⁹ and xyztransformer.³⁰

AtomGPT is a deep learning model⁶ that leverages a few transformer architectures such as GPT2³¹ and quantized Mistral-AI³² to learn the complex relationships between atomic

structures and material properties from dataset such as JARVIS-DFT.³³⁻³⁵ JARVIS-DFT includes a diverse range of structural, electronic, and thermal properties. By learning from this dataset, AtomGPT acquires a rich representation of the materials space, enabling it to predict properties of unseen materials and generate new material structures with desired properties.

One of the key features of AtomGPT is its ability to perform both forward and inverse materials design. In forward design, the model predicts the properties of a given material structure, allowing researchers to screen and evaluate potential materials efficiently. In inverse design, AtomGPT generates material structures that are likely to exhibit specific target properties, guiding the synthesis of materials with desired characteristics. The forward model is intended for pre-screening purposes, where it efficiently predicts properties of known structures. This allows us to quickly filter through a vast number of potential candidates and identify those with desirable properties. This step is crucial in narrowing down the search space and ensuring that only the most promising candidates are considered for further analysis. On the other hand, the inverse models are designed for structure generation based on desired properties. This complementary approach to funnel-type screening allows us to directly generate new structures that meet specific criteria. By focusing on this goal, the inverse models can explore the structural space more effectively and generate novel candidates that may not have been considered otherwise. By maintaining separate models for these tasks, we can leverage the strengths of each approach without compromising their effectiveness. The forward model excels in rapid property prediction, while the inverse models specialize in creative structure generation. This dual capability makes AtomGPT a versatile tool for materials scientists, enabling them to navigate the materials space more effectively and accelerate the discovery of novel materials.

Additionally, as AtomGPT is using transformers and other cutting-edge LLM library models, it can be integrated with newer models as they come into existence. Finally, AtomGPT is closely integrated with NIST-JARVIS infrastructure,^{33,34} which has several existing

databases, streamlined workflows, benchmarks and models to evaluate them using a detailed materials domain knowledge base. The AtomGPT model training scripts will be made available at <https://github.com/usnistgov/atomgpt>.

AtomGPT’s application potential is vast, spanning various domains of materials science. In this work, the focus is going to be on designing Bardeen–Cooper–Schrieffer (BCS)-superconductors^{36–38} with relevant material properties for forward and inverse methods. Finding superconductors with high transition temperatures (T_C) at ambient as well as extreme conditions has been an active area of research for decades.³⁶ Predicting such quantities using experiments and computational methods such as density functional theory³⁷ can be time-consuming and artificial intelligence-based methods^{38,39} can be helpful to accelerate the design process. In addition to T_C , some of the other properties of interest for superconductor design are formation energy (should be negative) and electronic bandgap (should be close to zero to follow BCS formalism), which are also considered in this work.

An overview of the AtomGPT package is shown in Fig. 1a. AtomGPT can be used to predict the property of a material given its information or generate its structure given its property information in terms of text. To illustrate these processes, an example for well-known BCS superconductor MgB_2 (atomic structure shown in Fig. 1b) with an experimental T_C of 39 K,⁴⁰ and DFT prediction of 32.7 K³⁸ is shown in Fig. 1c and Fig. 1d. In order to represent the chemical and structural information of materials ChemNLP describer⁴¹ or Robocrystallographer⁴² can be used. Such a description includes instructions such as chemical formula, space group, bond-distance, bond-angles and explicit fractional coordinates of a material as shown in Fig. 1c.

First, let us discuss the method involved in forward property prediction models with a GPT2 architecture available in the transformers library.⁴³ AtomGPT allows easy integration of any other related huggingface library models available in the transformers library to be easily used in place of the GPT2 model.³¹ GPT2 was developed by OpenAI and its successors GPT3.5 and GPT4 are now commercial and have revolutionised the field of artificial

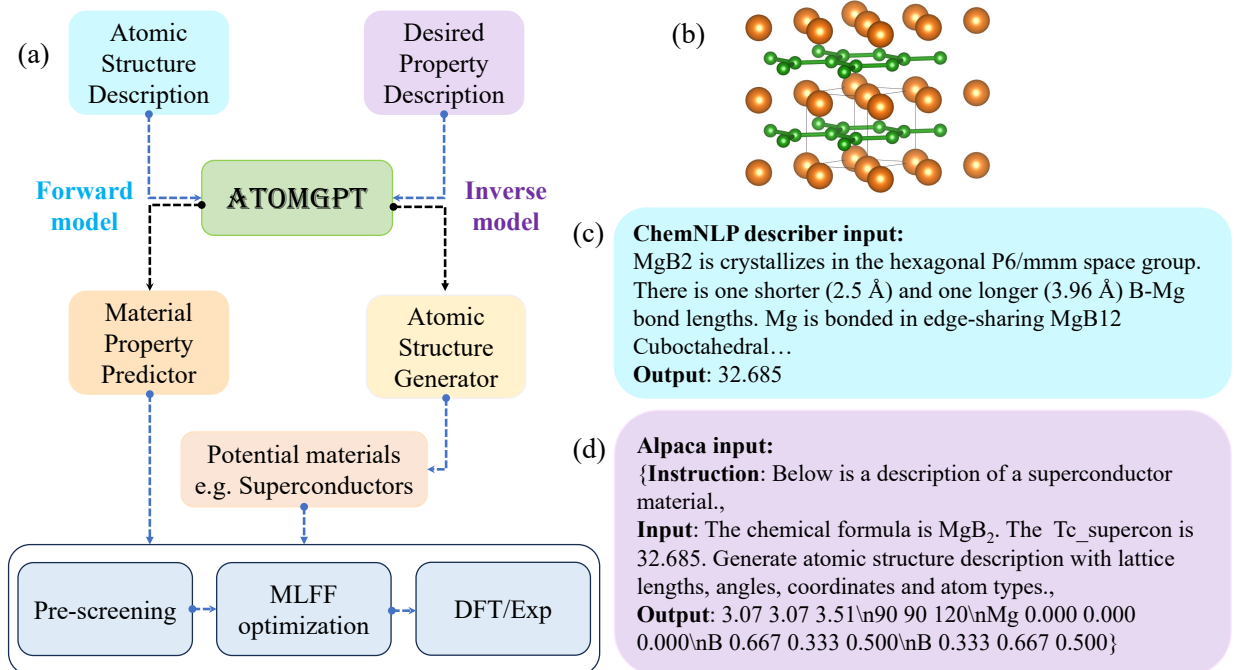


Figure 1: Schematic overview of the AtomGPT workflow. AtomGPT can be used for both forward model (atomic structure to property) predictions as well as inverse design (property to atomic structure generation) using LLMs. a) integrated text to material property predictions, text input to atomic structure generation, pre-screening, unified machine learning force-field (MLFF) optimization and density functional theory (DFT) calculations/experiment (Exp) based validation processes b) an example crystal structure for BCS-superconductor MgB₂ (JVASP-1151), c) text description of atomic structure of MgB₂ including explicit atomic structure as well as chemical information using ChemNLP, d) text prompt to explicit atomic structure generation example using Alpaca format.

intelligence. GPT-2 was primarily trained on CommonCrawl dataset and utilizes a stacked transformer decoder architecture, which consists of multiple layers of transformer blocks.³¹ Each transformer block contains two main components: a multi-head self-attention mechanism and a position-wise feed-forward network. The input to the model is a sequence of tokens, which are first converted into embeddings and then passed through the transformer blocks. Scaled dot-product attention used in a transformer model can be written as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q , K , and V represent the query, key, and value matrices, respectively. Here, d_k is the

dimensionality of the key vectors. The multi-head attention is obtained by concatenating multiple such attention heads. The multi-head self-attention mechanism allows the model to focus on different parts of the input sequence when computing the output for a particular token.

As the LLM models such as GPT2 were primarily designed for text to text models and not explicitly for regression applications, the language model head is modified with a linear neural network model to predict one or several properties as output(s). A pre-trained GPT-2 is loaded as a causal language model (CausalLM), which means it's designed for tasks where the prediction of each word token depends only on the previous tokens in the sequence (e.g., text generation). The new network with a modified language model head is a sequential model consisting of two linear layers. The first linear layer maps the hidden states of the GPT-2 model to a lower-dimensional latent space dimension (here 1024). The second linear layer maps this latent representation to a single output value for a regression task such as formation energy or superconductor T_C . There are various versions of GPT-2 model, but in this work, GPT2-small is used with 124 million parameters for forward design tasks. The models are trained using V100 GPUs with a batch size of 10, L1 loss function, and a learning rate of 10^{-3} for 200 epochs. The hyperparameters such as batch size, learning rate, latent space dimension were chosen by varying them for the smallest superconductor database, and used for other tasks with larger datasize such as formation energy and bandgaps.

The dataset used for this work is taken from the JARVIS-DFT and includes formation energies (55713), OptB88vdW bandgaps (55713), Tran Blaha modified Beck-Johnson potential based bandgaps (18167) and superconducting transition temperatures (1058) at ambient pressures.³³⁻³⁵ The dataset provides atomic structure information such as atomic coordinates, constituent element types and lattice parameters along with DFT calculated properties mentioned above. The dataset is used for 80:10:10 splits for training, validation and testing splits.

As LLM models require text as inputs, we can provide a description of an atomic structure

using ChemNLP⁴¹ which is then fed to the model discussed above. The input of the LLM-based crystal property prediction is carried out with ChemNLP describer⁴¹ which is based on well-defined combinations of structure and chemical features based on JARVIS-Tools. It generates inputs in a deterministic manner, ensuring that the same crystal structure consistently results in the same input description.

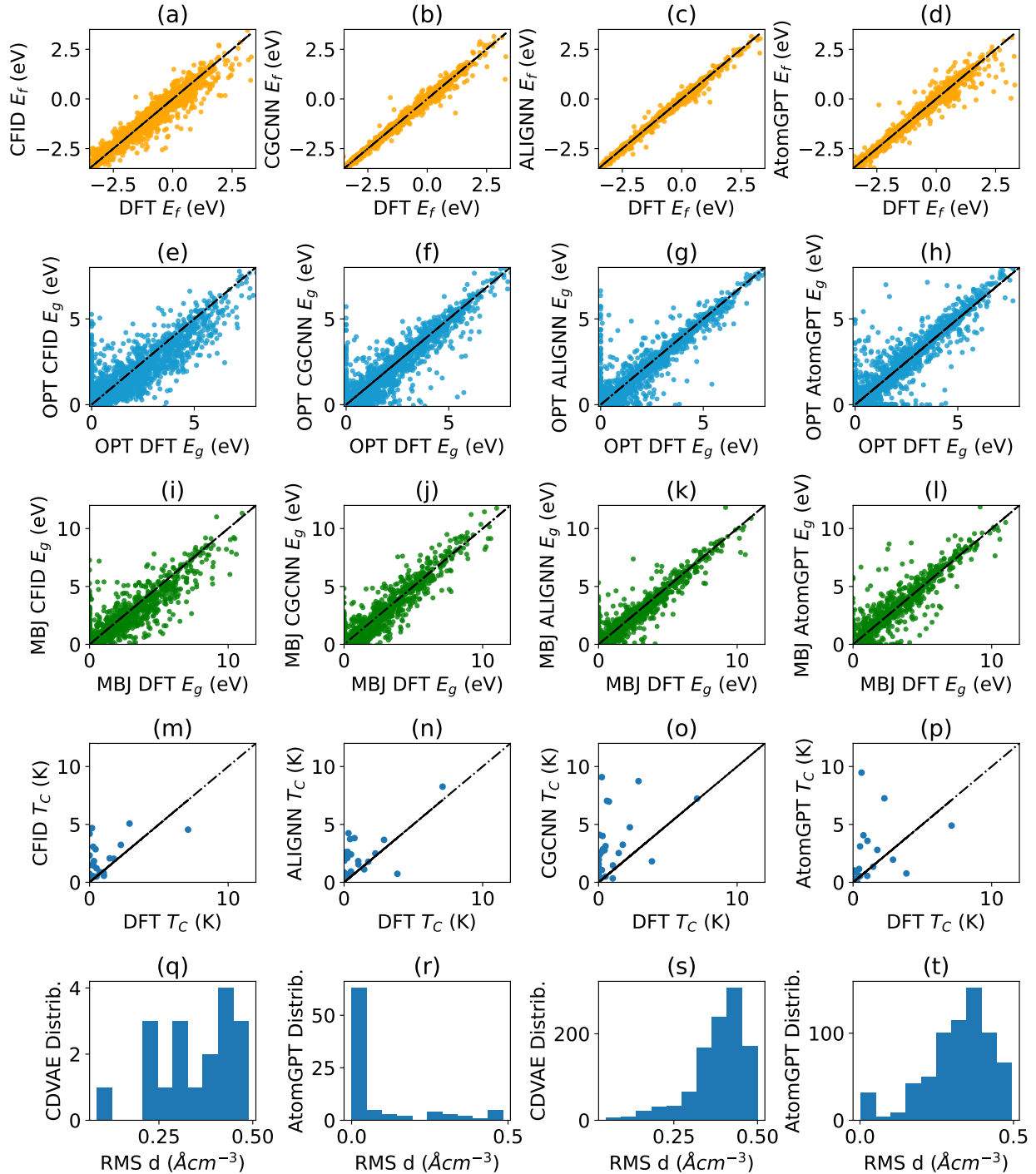


Figure 2: Performance of various models for forward (a-p) and inverse design (q-t) tasks on held out test set. Forward model performance for formation energy (E_f), OptB88vdW (OPT) bandgap (E_g , TBmBJ (MBJ) bandgap, and superconducting transition temperature (T_c) are shown using CFID, CGCNN, ALIGNN and AtomGPT models. Similarly, inverse model performance for superconductor dataset (q-r) and Carbon energy dataset (s-t) are shown using CDVAE and AtomGPT models.

After training such models, the performances are shown in Table 1 and Fig. 2a-2p for the forward design models of formation energy per atom, OptB88vdW bandgaps, TBmBJ bandgaps and superconducting transition temperatures using identical train-test splits (available in the JARVIS-Leaderboard benchmarking platform⁴⁴) for CFID, CGCNN and ALIGNN methods. For formation energy model the mean absolute error (MAE) values for CFID, CGCNN, ALIGNN and AtomGPT are 0.142 eV/atom, 0.063 eV/atom, 0.033 eV/atom and 0.072 eV/atom respectively. The mean absolute deviation (MAD), which can serve as a model baseline, is computed by using the mean of the target values in the training dataset and using it as predictions for all the materials in the test dataset. Usually, MAD:MAE of greater than 1 is desirable and higher values suggest better models. We note that while a MAD:MAE of greater than 1 is desirable, a higher ratio could also indicate a significant impact from outliers and data skewness. The MAD:MAE for the above models are 6.05, 13.74, 25.94 and 11.54 respectively. For OptB88vdW semi-local functional based electronic bandgap, the MAE values for CFID, CGCNN, ALIGNN and AtomGPT are 0.299 eV, 0.199 eV, 0.142 eV and 0.139 eV respectively. The MAD:MAE values are 3.31, 5.00, 6.97 and 7.11. For TBmBJ meta-GGA based electronic bandgap, the MAE values for CFID, CGCNN, ALIGNN and AtomGPT are 0.531 eV, 0.407 eV, 0.310 eV and 0.319 eV with MAD:MAE values of 3.35, 4.37, 5.73, 5.58 respectively. For electron-phonon coupling calculation based superconductor transition temperatures, the MAE values for CFID, CGCNN, ALIGNN and AtomGPT are 1.99 K, 2.06 K, 2.03 K and 1.54 K with MAD:MAE values of 0.890, 0.687, 0.880 and 1.160 respectively. Based on the above results, AtomGPT outperforms ALIGNN for OptB88vdW bandgap and T_C while for TBmBJ bandgaps and formation energies, ALIGNN still dominates other models. Lower MAE values in Table 1 and data along x=y line in scatter plots (Fig. 2a-2p) represent better performance. The superior performance of ALIGNN for formation energy can be attributed to some of the sophisticated information such as bond-angles, directional graphs and atom feature information¹⁸ which are not explicitly included in the AtomGPT model, Nevertheless, AtomGPT performance

for formation energy is closer to CGCNN which is also a GNN model with sophisticated graphs for atomistic systems. Note, it took dedicated efforts to develop such domain specific models such as ALIGNN, while finetuning GPT2 model is very straightforward. For some cases such as TBmBJ, the performances of ALIGNN and AtomGPT are comparable which is promising.

It is important to highlight the distinct design philosophies and underlying mechanisms of ALIGNN and AtomGPT. ALIGNN is a sophisticated graph neural network (GNN) that incorporates intricate features such as bond angles, directional graphs, and specific atom features, enabling it to excel in predicting properties such as the formation energy. The development of ALIGNN required extensive efforts to finely tune these domain-specific features, resulting in its superior performance for formation energy predictions. On the other hand, AtomGPT is based on a transformer architecture, specifically designed to leverage the strengths of large language models (LLMs). While fine-tuning AtomGPT is indeed straightforward compared to the extensive development process of ALIGNN, AtomGPT benefits from the ability to learn complex relationships between atomic structures and material properties through its deep learning capabilities. AtomGPT’s transformer-based architecture allows it to generalize well across diverse datasets. This could be particularly beneficial for predicting T_C , where complex electronic interactions play a significant role. ALIGNN’s explicit incorporation of sophisticated structural features gives it an edge in tasks like formation energy prediction, where detailed geometric and chemical interactions are critical. In contrast, AtomGPT relies on textual and structural descriptions, which, while powerful, may not capture all nuances as effectively for certain properties. We note that using LLM/GPT for materials datasets is a nascent area of research and analyzing the strength/weakness of such models for materials design need further explorations.

Next, the inverse design task is demonstrated by developing a separate model. While the forward prediction model uses GPT2 with a modified large model head, the inverse design model is based on supervised fine-tuning with Mistral-AI 7 billion parameter model³²

using Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning (PEFT).⁴⁵ Mistral is a powerful model with 7.3 billion parameters and has been shown to outperform Large Language Model Meta AI (LLaMA) 2 13B,⁴⁶ LLaMA 1 34B,⁴⁷ and ChatGPT3.5¹⁰ on several publicly available benchmarks. Some of the key features of Mistral are: sliding window attention to adeptly manage sequences of varying length while minimizing inference cost and grouped-query attention for swift inference (decrease inference time), details of which can be found in Ref.³²

Table 1: Performance measurement for forward and inverse design tasks and comparison with Classical Force-field Inspired Descriptors (CFID),¹⁵ Crystal Graph Convolutional Neural Network (CGCNN),¹⁶ Atomistic Line Graph Neural Network (ALIGNN)¹⁸ and Crystal Diffusion Variational Autoencoder (CDVAE)²⁴ and AtomGPT models. Best performing model results are shown as bold texts. The forward property prediction (Prop.) models are shown using mean absolute error (MAE) metric between held out test set and property predictions, while the inverse models are compared using reconstruction loss in normalized root mean square error (RMSE) distance between held out test structures and those using the generative models.

Forward models				
Prop/MAE	CFID	CGCNN	ALIGNN	AtomGPT
E_{form} (eV/atom)	0.142	0.063	0.033	0.072
OPT E_g (eV)	0.299	0.199	0.142	0.139
MBJ E_g (eV)	0.531	0.407	0.310	0.319
T_C (K)	1.99	2.60	2.03	1.54
Inverse models				
Database/RMSE	CDVAE	AtomGPT		
SuperConDB	0.24	0.08		
Carbon24	0.36	0.32		

Let us specifically focus on the superconductor inverse design task. This requires finetuning LLM models such as Mistral AI. Such finetuning requires transforming the instructions into a specialized protocol such as Alpaca.⁴⁸ An example of converting MgB₂ into alpaca is shown in Fig. 1d. It consists of a python dictionary with Instruction, Input and Output keys. As directly finetuning such a model can be computationally expensive, Parameter-Efficient Fine-Tuning (PEFT) method using Hugging Face ecosystem (transformers, PEFT, Transformer Reinforcement Learning (TRL), Rotary Position Embedding (RoPE)⁴⁹ is employed

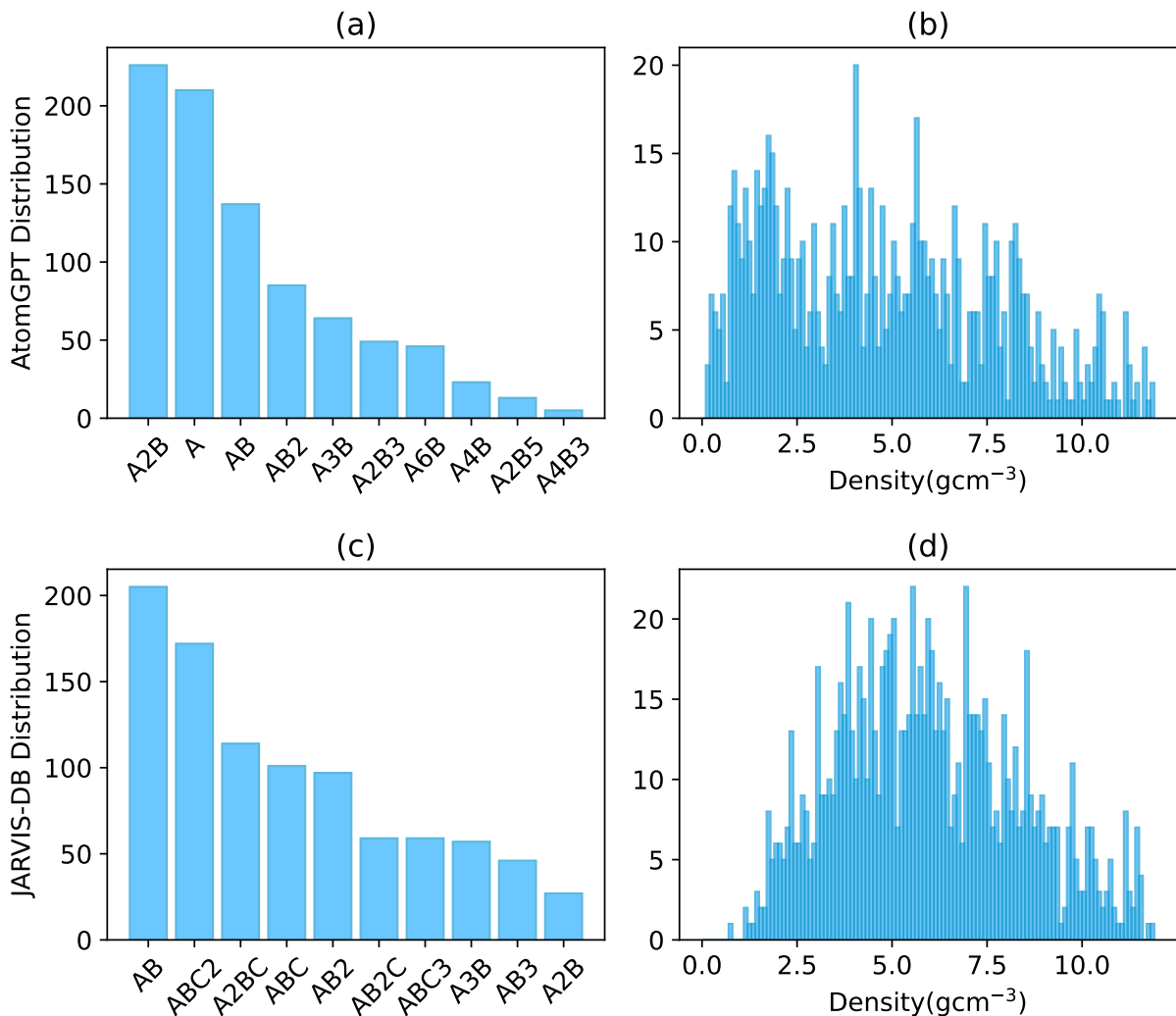


Figure 3: Chemical prototype and density of (a–b) the AtomGPT generated structures and (c–d) the 1058 structures used for training from the JARVIS-DFT database.

in addition to patching a Mistral model with fast LoRA (Locally Reweighted Attention)⁴⁵ weights for reduced memory training. After obtaining the PEFT model, corresponding tokenizer, and alpaca dataset (with 80 % training, 10 % validation and 10 % test splits for the superconductor dataset), supervised fine-tuning tasks are carried out with a batch size of 10, AdamW 8 bit optimizer and cross-entropy loss function. This loss function measures the difference between the predicted probability distribution over the vocabulary and the true distribution (i.e., the one-hot encoded target words). After the model is trained, the model is evaluated on the test set with respect to reconstruction/test performance. To further clarify,

after training the model on the train set, keeping the instruction and input keys in the test set, the trained model is employed to generate outputs. After parsing the outputs to create corresponding crystal structures, the StructureMatcher algorithm⁵⁰ is used to find the best match between two structures considering all invariances of materials. The root mean square squared error (RMSE) is averaged overall matched materials. Because the inter-atomic distances can vary significantly for different materials, the RMSE is normalized following the works in Ref.²⁴ Note that this is just one of the metrics for generative models for atomic structures and there can be numerous other types of metrics. It is important to calculate the T_C using a high fidelity method such as DFT later that acts as a more reliable metric for the generation process. The performance of the model for the JARVIS-SuperconDB (1058) and Carbon24 (10153) datasets⁵¹ are shown in Table 1 and Fig. 2q-2t. We notice from Table 1 that AtomGPT has lower RMSE than CDVAE for both SuperconDB and Carbon24 databases suggesting that given a specific prompt, the model can generate structures very similar to the held out test set. This can be further analyzed in Fig 2q-2t where the root mean square distance differences are lower for AtomGPT model than CDVAE models for both tasks. A higher number of bars at lower RMSE distance (d) values represent close agreement in structures. Moreover, the distributions for Fig. 2q and Fig. 2r are markedly different, while Fig. 2s and Fig. 2t display similar distributions. While there is no obvious explanation for such a behavior, a plausible explanation could be attributed to the inherent nature of the datasets used. The Carbon24 database is composed primarily of carbon-based structures generated via ab-initio random structure search (AIRSS).⁵¹ Many of these entries are thermodynamically unstable and not synthesizable, which introduces higher variability in bond-distribution and low variability in chemical distribution and consequently higher RMSE in bond-distance values in the predictions. On the other hand, the SuperConDB contains a larger variety of diverse chemical compositions and stable structures. The stability and diverse nature of these structures could contribute to more consistent predictions and lower RMSE values, thus possibly explaining the observed difference in distributions. The

similar distributions for carbon dataset (s-t) indicate that both modeling approaches are almost equally effective in predicting the structures within these datasets. Further improvement could be possible with better text prompts, which are subject to future work. Note that only one property T_C and specific type of instructions are used here, and LLMs allow easy addition of other properties and conditional statements as well.

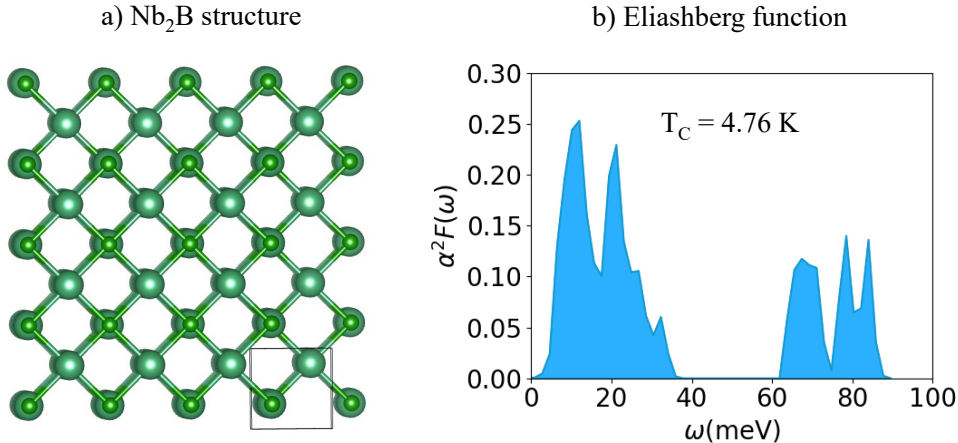


Figure 4: Density functional theory calculation results for one of the identified structures. a) atomic structure visualization, b) Eliashberg function for Nb₂B structure with a predicted T_C of 4.76 K.

Next, materials are generated using such models for identifying potential candidate superconductors. Our focus in his work is on Borides (but are not limited by it) as they are one of the systems of interest for our future experimental synthesis. One of the first computational discoveries and experimentally synthesized superconductors is FeB₄⁵² which further motivates such application. Systems with the formula $X_a B_b$, $X_a Y_b B_c$, $X_a Y_b Z_c B_d$, where X,Y,Z are periodic table elements other than Boron and a,b,c can be integers such as 1,2,3 are generated. After generating such 890 structures, the chemical prototypes and density distributions are shown in Fig. 3a and Fig. 3b. We observe that most of the gen-

erated structures belong to prototypes A2B, A, AB and AB₂. Their densities can attain values upto 15 gcm⁻³. The prototypes generated are similar to those existing in the training dataset i.e., the JARVIS-DFT superconductor dataset as shown in Fig. 3c. The densities of the generated structures also closely resemble distributions of the training set with a peak around 5 gcm⁻³ as shown in Fig. 3d. The best formation energy property model listed above is used to check if they have negative formation energy, bandgap models to make sure they have close to zero bandgaps and T_C greater than 1 K leading to 30 candidate materials. Then, these candidates are subjected to ALIGNN-FF (developed in Ref.⁵³) based pre-optimization. After the pre-optimization, these materials are subjected to JARVIS-DFT workflow using Vienna Ab initio Simulation Package^{54,55} and OptB88vdW functional. Now, electron-phonon calculations are performed for these optimized structures. Note that this step is computationally expensive and can take days if not weeks for each material. One of the potential candidates is found to be Nb₂B with optimized lattice lengths of 3.19 Å, 3.18 Å and 4.52 Å and lattice angles of 93.7°, 90° and 90° crystallizing in the monoclinic spacegroup of P2/m (as shown in Fig. 4). It has a density of 7.1 gcm⁻³ which is almost 3 times higher than Silicon (2.25 gcm⁻³) and aluminum (2.66 gcm⁻³). The formation energy and energy above the convex hull of this system are calculated as -0.13 eV/atom and 0.3 eV respectively. Finally, McMillan-Allen-Dynes formula based superconductor transition temperature calculations are performed using quantum espresso using workflow developed in Ref.³⁸ to calculate its T_C .

The workflow used to generate the training data is based on Ref.,³⁸ where electron phonon calculations (EPC) were performed using non-spin polarized DFT-perturbation theory (PT)^{56,57} (using the interpolated/Gaussian broadening method⁵⁸) with the Quantum Espresso (QE) software package,⁵⁹ PBEsol functional,⁶⁰ and the GBRV⁶¹ pseudopotentials. The EPC parameter is derived from spectral function $\alpha^2F(\omega)$ which is calculated as follows:

$$\alpha^2F(\omega) = \frac{1}{2\pi N(\epsilon_F)} \sum_{qj} \frac{\gamma_{qj}}{\omega_{qj}} \delta(\omega - \omega_{qj}) w(q) \quad (2)$$

where ω_{qj} is the mode frequency, $N(\epsilon_F)$ is the density of states (DOS) at the Fermi level ϵ_F , δ is the Dirac-delta function, $w(q)$ is the weight of the q point, γ_{qj} is the linewidth of a phonon mode j at wave vector q and is given by:

$$\gamma_{qj} = 2\pi\omega_{qj} \sum_{nm} \int \frac{d^3k}{\Omega_{BZ}} |g_{kn,k+qm}^j|^2 \delta(\epsilon_{kn} - \epsilon_F) \delta(\epsilon_{k+qm} - \epsilon_F) \quad (3)$$

Here, the integral is over the first Brillouin zone, ϵ_{kn} and ϵ_{k+qm} are the DFT eigenvalues with wavevector k and $k + q$ within the n th and m th bands respectively, $g_{kn,k+qm}^j$ is the electron-phonon matrix element. γ_{qj} is related to the mode EPC parameter λ_{qj} by:

$$\lambda_{qj} = \frac{\gamma_{qj}}{\pi h N(\epsilon_F) \omega_{qj}^2} \quad (4)$$

Now, the EPC parameter is given by:

$$\lambda = 2 \int \frac{\alpha^2 F(\omega)}{\omega} d\omega = \sum_{qj} \lambda_{qj} w(q) \quad (5)$$

with $w(q)$ as the weight of a q point. The superconducting transition temperature, T_c can then be approximated using McMillan-Allen-Dynes⁶² equation as follows:

$$T_c = \frac{\omega_{log}}{1.2} \exp \left[- \frac{1.04(1 + \lambda)}{\lambda - \mu^*(1 + 0.62\lambda)} \right] \quad (6)$$

where

$$\omega_{log} = \exp \left[\frac{\int d\omega \frac{\alpha^2 F(\omega)}{\omega} \ln \omega}{\int d\omega \frac{\alpha^2 F(\omega)}{\omega}} \right] \quad (7)$$

In Eq. 6, the parameter μ^* is the effective Coulomb potential parameter, which we take as 0.1. It is important to note that the robustness of this workflow was heavily benchmarked against experimental data and higher levels of theory in ref.,^{38,63-65} which indicates that the training data for these deep learning models is of high quality, given the level of theory used to produce the data.

The T_C for Nb₂B is predicted to be 4.76 K with no negative modes in the phonon density

of states curve suggesting dynamical stability. Similar DFT calculations can be performed for other structures after populating more candidates, which can then be added to the JARVIS-DFT Superconductor database.^{38,63-65} Moreover, we can extend this workflow to include other candidate systems, such as nitrides and carbides. As the database of superconductors expands, both forward and inverse models have the potential for further improvement, which is necessary for enhanced predictions. After DFT validation, we can recommend these computationally discovered candidates for experimental synthesis. Superconductor design remains one of the most fascinating areas of research in physics and chemistry. Our work here serves as a proof-of-concept that demonstrates the potential of LLM models, such as GPTs, to accelerate superconductor materials design. Moreover, it can be easily extended to other classes of materials such as alloys, semiconductors, thermoelectrics, dielectrics, piezoelectrics, solar cells and so on. Additionally, AtomGPT’s application of AtomGPT for defect⁶⁶ and interface⁶⁷ materials design can also be interesting areas of research in the future.

In conclusion, this study introduces AtomGPT, an atomistic generative pre-trained transformer model specifically designed for both forward and inverse materials design. The model is capable of predicting material properties with high accuracy and generating novel material structures with desired properties. By leveraging well-defined structural and chemical features from the JARVIS-Tools, AtomGPT ensures consistent and reliable input generation, enhancing the robustness of property predictions. The separation of forward and inverse models allows for efficient pre-screening and targeted structure generation, respectively. This dual approach facilitates the rapid discovery and optimization of new materials, paving the way for accelerated advancements in materials science. Note that, in this study, I present AtomGPT and large language models (LLMs) not as the ultimate solutions for materials science challenges, but as highly adaptable and tunable tools that can be tailored to a wide range of applications. The versatility of these models allows them to be customized for specific datasets and tasks, making them valuable complements to existing methodologies. By focusing on their adaptability and integration with other tools such as DFT and unified

machine learning force-fields, I aim to highlight their potential to enhance and expand the toolkit available to materials scientists, facilitating innovation and discovery across the field. I have demonstrated the application of this workflow by computationally discovering a novel superconductor Nb₂B with T_C of 4.76 K. These findings highlight the potential of large language models in transforming materials design processes, offering a promising direction for future research and applications in various domains of materials science.

Acknowledgements

K.C. thanks computational resources from the National Institute of Standards and Technology (NIST). This work was performed with funding from the CHIPS Metrology Program, part of CHIPS for America, National Institute of Standards and Technology, U.S. Department of Commerce. Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identifications are not intended to imply recommendation or endorsement by NIST, nor it is intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

- (1) Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J.; others Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* **2022**, *8*, 59.
- (2) Vasudevan, R. K.; Choudhary, K.; Mehta, A.; Smith, R.; Kusne, G.; Tavazza, F.; Vlcek, L.; Ziatdinov, M.; Kalinin, S. V.; Hattrick-Simpers, J. Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a

- pathway from correlations to the underpinning physics. *MRS communications* **2019**, *9*, 821–838.
- (3) Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj computational materials* **2019**, *5*, 83.
- (4) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; others 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery* **2023**, *2*, 1233–1250.
- (5) Choudhary, A.; Fox, G. C.; Hey, T. *Artificial Intelligence for Science: A Deep Learning Revolution*; World Scientific, 2023.
- (6) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep learning*; MIT press, 2016.
- (7) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
- (8) Tunstall, L.; Von Werra, L.; Wolf, T. *Natural language processing with transformers*; ” O’Reilly Media, Inc.”, 2022.
- (9) Rothman, D. *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*; Packt Publishing Ltd, 2021.
- (10) Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.-L.; Tang, Y. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* **2023**, *10*, 1122–1136.

- (11) Pimentel, A.; Wagener, A.; da Silveira, E. F.; Picciani, P.; Salles, B.; Follmer, C.; Oliveira Jr, O. N. Challenging ChatGPT with Chemistry-Related Subjects. **2023**,
- (12) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **2024**, 1–9.
- (13) Polak, M. P.; Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications* **2024**, *15*, 1569.
- (14) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; others Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **2018**, *152*, 60–69.
- (15) Choudhary, K.; DeCost, B.; Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Physical review materials* **2018**, *2*, 083801.
- (16) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120*, 145301.
- (17) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials* **2019**, *31*, 3564–3572.
- (18) Choudhary, K.; DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials* **2021**, *7*, 185.
- (19) Reiser, P.; Eberhard, A.; Friederich, P. Graph neural networks in TensorFlow-Keras with RaggedTensor representation (kgcnn). *Software Impacts* **2021**, *9*, 100095.

- (20) Chen, C.; Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science* **2022**, *2*, 718–728.
- (21) Yan, K.; Liu, Y.; Lin, Y.; Ji, S. Periodic graph transformers for crystal material property prediction. *Advances in Neural Information Processing Systems* **2022**, *35*, 15066–15080.
- (22) Yan, K.; Fu, C.; Qian, X.; Qian, X.; Ji, S. Complete and Efficient Graph Transformers for Crystal Material Property Prediction. *arXiv preprint arXiv:2403.11857* **2024**,
- (23) Rubungo, A. N.; Arnold, C.; Rand, B. P.; Dieng, A. B. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029* **2023**,
- (24) Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197* **2021**,
- (25) Ren, Z.; Noh, J.; Tian, S.; Oviedo, F.; Xing, G.; Liang, Q.; Aberle, A.; Liu, Y.; Li, Q.; Jayavelu, S.; others Inverse design of crystals using generalized invertible crystallographic representation. *arXiv preprint arXiv:2005.07609* **2020**, *3*, 7.
- (26) Gebauer, N.; Gastegger, M.; Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems* **2019**, *32*.
- (27) Korolev, V.; Protsenko, P. Accurate, interpretable predictions of materials properties within transformer language models. *Patterns* **2023**, *4*.
- (28) Antunes, L. M.; Butler, K. T.; Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *arXiv preprint arXiv:2307.04340* **2023**,
- (29) Gruver, N.; Sriram, A.; Madotto, A.; Wilson, A. G.; Zitnick, C. L.; Ulissi, Z. Fine-Tuned Language Models Generate Stable Inorganic Materials as Text. *arXiv preprint arXiv:2402.04379* **2024**,

- (30) Flam-Shepherd, D.; Aspuru-Guzik, A. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv preprint arXiv:2305.05708* **2023**,
- (31) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; others Language models are unsupervised multitask learners. *OpenAI blog* **2019**, 1, 9.
- (32) Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; others Mistral 7B. *arXiv preprint arXiv:2310.06825* **2023**,
- (33) Choudhary, K.; Garrity, K. F.; Reid, A. C.; DeCost, B.; Biacchi, A. J.; Hight Walker, A. R.; Trautt, Z.; Hatrick-Simpers, J.; Kusne, A. G.; Centrone, A.; others The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj computational materials* **2020**, 6, 173.
- (34) Wines, D.; Gurunathan, R.; Garrity, K. F.; DeCost, B.; Biacchi, A. J.; Tavazza, F.; Choudhary, K. Recent progress in the JARVIS infrastructure for next-generation data-driven materials design. *Applied Physics Reviews* **2023**, 10.
- (35) Choudhary, K.; Zhang, Q.; Reid, A. C.; Chowdhury, S.; Van Nguyen, N.; Trautt, Z.; Newrock, M. W.; Congo, F. Y.; Tavazza, F. Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms. *Scientific data* **2018**, 5, 1–12.
- (36) Cooper, L. N.; Feldman, D. *BCS: 50 years*; World scientific, 2010.
- (37) Giustino, F. Electron-phonon interactions from first principles. *Reviews of Modern Physics* **2017**, 89, 015003.
- (38) Choudhary, K.; Garrity, K. Designing high-TC superconductors with BCS-inspired

- screening, density functional theory, and deep-learning. *npj Computational Materials* **2022**, *8*, 244.
- (39) Schmidt, J.; Hoffmann, N.; Wang, H.-C.; Borlido, P.; Carriço, P. J.; Cerqueira, T. F.; Botti, S.; Marques, M. A. Machine-Learning-Assisted Determination of the Global Zero-Temperature Phase Diagram of Materials. *Advanced Materials* **2023**, *35*, 2210788.
- (40) Tomsic, M.; Rindfleisch, M.; Yue, J.; McFadden, K.; Phillips, J.; Sumption, M. D.; Bhatia, M.; Bohnenstiehl, S.; Collings, E. W. Overview of MgB₂ superconductor applications. *International journal of applied ceramic technology* **2007**, *4*, 250–259.
- (41) Choudhary, K.; Kelley, M. L. ChemNLP: a natural language-processing-based library for materials chemistry text data. *The Journal of Physical Chemistry C* **2023**, *127*, 17545–17555.
- (42) Ganose, A. M.; Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications* **2019**, *9*, 874–881.
- (43) Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; others Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* **2019**,
- (44) Choudhary, K.; Wines, D.; Li, K.; Garrity, K. F.; Gupta, V.; Romero, A. H.; Krogel, J. T.; Saritas, K.; Fuhr, A.; Ganesh, P.; others JARVIS-Leaderboard: a large scale benchmark of materials design methods. *npj Computational Materials* **2024**, *10*, 93.
- (45) Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* **2021**,
- (46) Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bash-

- lykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; others Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* **2023**,
- (47) Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; others Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* **2023**,
- (48) Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T. B. Stanford alpaca: an instruction-following llama model (2023). *URL* https://github.com/tatsu-lab/stanford_alpaca **2023**,
- (49) Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063.
- (50) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **2013**, *68*, 314–319.
- (51) Pickard, C. J.; Needs, R. Ab initio random structure searching. *Journal of Physics: Condensed Matter* **2011**, *23*, 053201.
- (52) Gou, H.; Dubrovinskaia, N.; Bykova, E.; Tsirlin, A. A.; Kasinathan, D.; Schnelle, W.; Richter, A.; Merlini, M.; Hanfland, M.; Abakumov, A. M.; others Discovery of a superhard iron tetraboride superconductor. *Physical review letters* **2013**, *111*, 157002.
- (53) Choudhary, K.; DeCost, B.; Major, L.; Butler, K.; Thiyagalingam, J.; Tavazza, F. Unified graph neural network force-field for the periodic table: solid state applications. *Digital Discovery* **2023**, *2*, 346–355.
- (54) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical review B* **1996**, *54*, 11169.

- (55) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational materials science* **1996**, *6*, 15–50.
- (56) Baroni, S.; Giannozzi, P.; Testa, A. Green’s-function approach to linear response in solids. *Physical review letters* **1987**, *58*, 1861.
- (57) Gonze, X. Perturbation expansion of variational principles at arbitrary order. *Physical Review A* **1995**, *52*, 1086.
- (58) Wierzbowska, M.; de Gironcoli, S.; Giannozzi, P. Origins of low-and high-pressure discontinuities of T_{c} in niobium. *arXiv preprint cond-mat/0504077* **2005**,
- (59) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; others QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter* **2009**, *21*, 395502.
- (60) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. Restoring the density-gradient expansion for exchange in solids and surfaces. *Physical review letters* **2008**, *100*, 136406.
- (61) Garrity, K. F.; Bennett, J. W.; Rabe, K. M.; Vanderbilt, D. Pseudopotentials for high-throughput DFT calculations. *Computational Materials Science* **2014**, *81*, 446–452.
- (62) McMillan, W. Transition temperature of strong-coupled superconductors. *Physical Review* **1968**, *167*, 331.
- (63) Wines, D.; Choudhary, K.; Biacchi, A. J.; Garrity, K. F.; Tavazza, F. High-throughput DFT-based discovery of next generation two-dimensional (2D) superconductors. *Nano Letters* **2023**, *23*, 969–978.

- (64) Wines, D.; Xie, T.; Choudhary, K. Inverse design of next-generation superconductors using data-driven deep generative models. *The Journal of Physical Chemistry Letters* **2023**, *14*, 6630–6638.
- (65) Wines, D.; Choudhary, K. Data-driven Design of High Pressure Hydride Superconductors using DFT and Deep Learning. *arXiv preprint arXiv:2312.12694* **2023**,
- (66) Choudhary, K.; Sumpter, B. G. Can a deep-learning model make fast predictions of vacancy formation in diverse materials? *AIP Advances* **2023**, *13*.
- (67) Choudhary, K.; Garrity, K. InterMat: Accelerating Band Offset Prediction in Semiconductor Interfaces with DFT and Deep Learning. *Digital Discovery* **2024**, –.

TOC Graphic

