

Soft Specialization via Mixture-of-Experts Enables Accurate Prediction of Superconducting T_c from Crystal Structures

Khac Gia Hoang Nguyen

Posts and Telecommunications Institute of Technology

Hoang Long Dang

Posts and Telecommunications Institute of Technology

Minh Tri Nguyen

`tri.nguyen1@deakin.edu.au`

Deakin University

Article

Keywords:

Posted Date: March 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8567979/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Soft Specialization via Mixture-of-Experts Enables Accurate Prediction of Superconducting T_c from Crystal Structures

Khac Gia Hoang Nguyen¹, Hoang Long Dang¹, and Minh Tri Nguyen^{2*}

¹Posts and Telecommunications Institute of Technology, Viet Nam

²Deakin University, Victoria, Australia

*Corresponding author: tri.nguyen1@deakin.edu.au

Abstract

Predicting superconducting critical temperature (T_c) from crystal structure is challenging due to diverse material families with distinct physical mechanisms and structural complexities like site-occupation disorder. We introduce Superconductor Mixture of Experts (SuMoE), a Mixture-of-Experts (MoE) architecture using multiple Graph Neural Network experts with learned routing, achieving MAE of 0.671 logK on the SuperCon3D dataset, significantly outperforming both a state-of-the-art single generalist Graph Neural Network (0.709 logK) and isolated family-specific specialists (0.683 logK). While traditional intuition suggests models should specialize in rigid material families, our analysis reveals a more sophisticated regime of soft specialization. We find that experts specialize in underlying physical mechanisms rather than arbitrary labels, for instance, a single expert dominates predictions for both Cuprates and Heavy Fermion systems, effectively capturing the physics of strongly correlated electrons across chemically distinct families. Crucially, the model couples this specialized knowledge with a diffuse routing mechanism that synthesizes insights from multiple experts, ensuring robustness through redundancy. Our work demonstrates that for complex structure-property relationships, the most effective paradigm is not specialization but integrative prediction, where the model learns to synthesize complementary insights to form a single, robust prediction.

1 Introduction

The quest for a room-temperature superconductor, a material that conducts electricity with zero resistance under ambient conditions, remains one of the most significant pursuits in modern physics and materials science[1]. Such a discovery would revolutionize energy transmission, computing, transportation, and medical imaging[8]. The critical

temperature (T_c), below which a material becomes superconducting, is the primary figure of merit. However, predicting T_c from a material’s crystal structure is a formidable task, as the mechanisms governing superconductivity, particularly in unconventional high- T_c materials like cuprates and hydrides, are still not fully understood from first principles[2].

Traditional discovery methods, relying on chemical intuition and serendipity, are slow and resource-intensive. Computational approaches, primarily density functional theory (DFT), can provide insights but are too computationally expensive for large-scale screening of candidate materials[7]. This has paved the way for data-driven methods, where machine learning (ML) models are trained on existing experimental data to learn the complex structure-property relationships that determine T_c [3, 10].

Early ML models for T_c prediction relied on hand-crafted features derived from chemical formulas or elemental properties[12]. While successful to an extent, these models often neglect the explicit three-dimensional atomic arrangement that is crucial to superconductivity. More recently, graph neural networks (GNNs) have emerged as the state-of-the-art, representing crystal structures as graphs where atoms are nodes and interatomic bonds are edges[14, 4]. This allows the model to learn material representations directly from the crystal structure, capturing local atomic environments and long-range interactions. Models like CGCNN[14], MEGNet[4], and CrabNet[13] have demonstrated strong performance in predicting various material properties, including T_c . Multi-modal Mixture of Experts [15] combines different type of features as different experts to predict the T_c .

Despite these advances, two significant challenges persist. First, existing models often employ a single, monolithic architecture to predict properties for all material classes. However, the underlying physics of superconductivity can vary drastically between different families (e.g.,

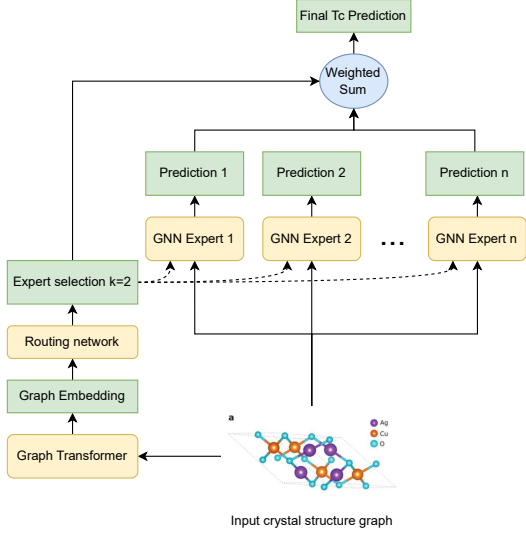


Figure 1: The Mixture of Experts model for T_c prediction.

electron-phonon coupling in conventional superconductors vs. strongly correlated electron effects in cuprates)[6]. A single model may struggle to capture these disparate mechanisms effectively. Second, many important superconductors exhibit site-occupation disorder (SOD), where certain crystallographic sites are occupied by a statistical mixture of different elements. Most GNNs are designed for perfectly ordered structures and cannot naturally represent this crystallographic feature, which can significantly influence T_c .

To address these limitations, we propose the **Superconductor Mixture of Experts (SuMoE)**, the first application of Mixture of Experts (MoE) architecture[11] with learned routing over GNN experts operating on crystal graphs. Our framework employs multiple GNN “experts” with a learned “gating” network that routes materials to appropriate experts, enabling the model to capture family-specific structure-property relationships within a unified, end-to-end trainable architecture. Our key contributions are:

1. We demonstrate that GNN experts trained jointly on all superconductor families significantly outperform both a single generalist model and isolated family-specific specialists. This establishes that shared learning across chemically diverse materials improves predictions for *all* families, including data-scarce ones.
2. Analysis of expert behavior reveals that specialization emerges along physical mechanism boundaries rather than arbitrary family labels. Notably, a single expert dominates both

Cuprate and Heavy Fermion predictions which are two families sharing strongly correlated electron physics despite different chemistries. This suggesting the expert has learned structural signatures of electron correlation that transfer across material classes.

3. We show that combining predictions from multiple experts (top-k routing) consistently outperforms single-expert selection, indicating that robust T_c prediction benefits from synthesizing complementary physical insights rather than relying on perfect routing.

SuMoE achieves state-of-the-art MAE of 0.671 logK on the SuperCon3D dataset[5], providing a powerful tool for computational superconductor discovery. Beyond performance gains, our analysis offers insights into how machine learning models can discover and leverage transferable physical knowledge across material families, a paradigm shift from the common practice of training separate models for each material class.

2 Results

2.1 Overall Model Performance

The SuMoE model was evaluated on 1,578 materials from the SuperCon3D database using K-fold cross-validation. The primary performance metric was the Mean Absolute Error (MAE) and R2 with standard deviation ($\pm$ std) between the predicted and experimental T_c values. All model training and evaluation are performed in logK. Values in Table 3 are reported in Kelvin for interpretability. Unless otherwise stated, all analyses use the best-performing configuration: 7 experts with top-k=3 routing.

As shown in Fig. 2 and Table 1, SuMoE achieves an MAE of 0.671 on K-fold validation. The parity plot of predicted versus actual T_c values demonstrates a strong correlation, with R^2 of 0.641, surpassing the baseline SODNet. The data points cluster tightly along the diagonal identity line across the full temperature range (approximately -6 to 6 logK), indicating that the SuMoE effectively captures the structure-property relationships without significant systematic bias. The parity plot exhibits noticeable heteroscedasticity, with higher variance observed in the low- T_c regime compared to the tight clustering of high- T_c families like Cuprates and Iron-based systems. This dispersion largely corresponds to the ‘Others’ category, which our t-SNE analysis (Fig. 5) reveals as a chemically heterogeneous group with entangled feature representations that prevent the isolation of distinct structural motifs. Physically, this suggests that while high- T_c materials possess strong geometric signals,

such as the layered planes in Cuprates, that GNN experts can easily identify, the physics of low- T_c conductors often relies on subtle electronic structure changes not explicitly encoded in the crystal graph topology. Consequently, the model faces greater difficulty in this regime, where the absence of a dominant structural driver results in less effective expert specialization compared to the signal-rich environments of strongly correlated systems.

As shown in the Table 1, the MoE models utilizing top- $k=3$ routing consistently achieve the best performance. The optimal configuration, using 7 GNN experts, yielded an MAE of 0.6707 and R^2 of 0.6407, representing a notable improvement over the single GNN baseline (MAE=0.709, $R^2=0.614$). Models with 6 and 7 experts performed almost identically (MAE = 0.675), while performance slightly degraded with 8 experts (MAE = 0.6812). This establishes the architectural advantage of the MoE approach for this task but suggests diminishing returns beyond 6-7 experts, as illustrated in Fig. 6. The consistent superiority of top- $k=2$ over top- $k=1$ routing (best $k=1$ MAE=0.686) hints that combining multiple expert outputs is crucial.

Table 1: Overall performance on the SuperCon3D dataset, evaluated using mean absolute error (MAE) and coefficient of determination (R^2) in logK. Results are reported for the SuperCon3D dataset.

Method	MAE	R^2
SODNet	0.709 ± 0.064	0.614 ± 0.067
5 GNN Specialists	0.683 ± 0.094	0.628 ± 0.079
<i>SuMoE</i> (7 Experts, $k=3$)	0.671 ± 0.065	0.641 ± 0.062

2.2 Specialization of Family-based Experts

For the specialization analysis presented in this section, we use a reduced configuration with 5 experts to improve interpretability of routing patterns.

A key hypothesis behind the MoE architecture is that different expert networks will learn to specialize in distinct chemical or structural families of superconductors. To investigate this, we analyzed the routing decisions made by the gating network for different classes of materials in the test set. The gating network assigns each input material to one of the available experts (in our case, 5 experts were used). We note that explicit attempts to promote specialization-including weighted sampling and load balancing losses, proved ineffective or detrimental, reinforcing that the MoE’s strength

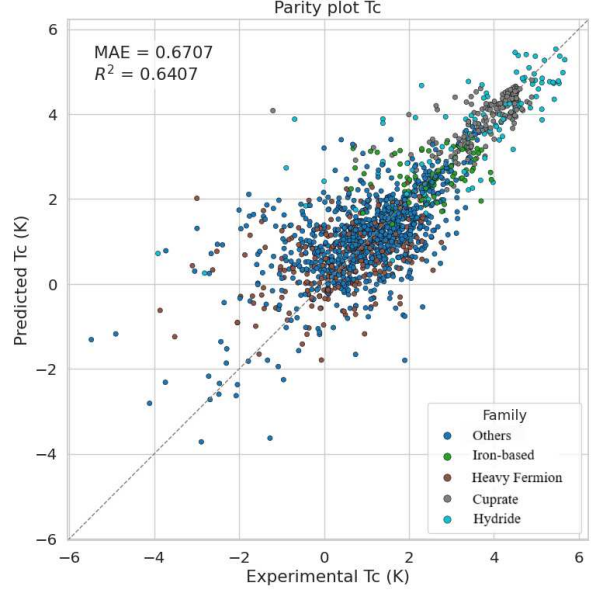


Figure 2: Overall performance of SuMoE. Parity plot of predicted vs. experimental superconducting critical temperature (T_c). The model achieves a Mean Absolute Error (MAE) of 0.671 logK and a Coefficient of determination (R^2) of 0.641. The dashed line represents perfect agreement.

lies in shared learning rather than forced separation.

Fig. 3 shows the distribution of materials from superconductor families across the five expert networks. Rather than routing material families to single experts in a diagonal, one-to-one fashion, the gating network distributes inputs across multiple experts. For instance, while the “Others” family is the majority class, its samples are processed by a diverse set of experts. This suggests the model is employing integrative prediction: it synthesizes outputs from multiple experts to form a prediction, functioning closer to an ensemble than a decision tree.

Analysis of the gating network’s behavior and individual expert performance reveals a nuanced picture of soft specialization. The expert usage matrix in Fig. 3 shows diffuse routing across all experts. The *Others* category, as the majority class, is distributed broadly with Expert 5 receiving the most samples (212) and Expert 1 the fewest (131). Notably, the routing does not mirror the optimal expert assignments at the family level: for *Cuprate*, Expert 2 receives the most samples (66), yet Expert 1 achieves the lowest family-wide MAE (0.423). Similarly, for *Heavy Fermion*, Expert 5 handles the most samples (73), but Expert 1 performs best (0.888). This mismatch arises because the gating network makes per-sample routing decisions based on learned embeddings, not family labels, and as shown in Fig. 5, the embedding

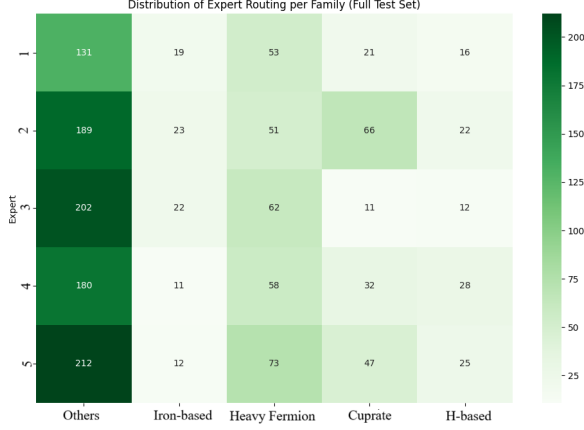


Figure 3: **Experts network specialization.** The distribution of routing assignments by the gating network for five major classes of superconductors.

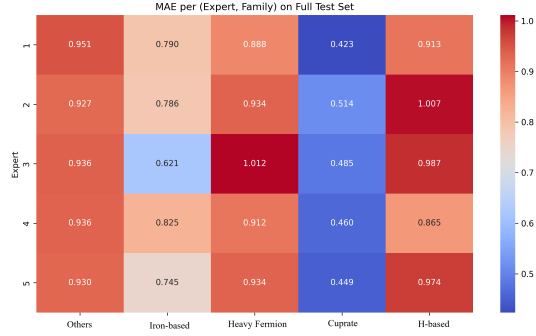


Figure 4: **Expert error distribution** Mean absolute error of each expert across five major superconductor families

space does not cleanly separate material families. This limitation of single-expert routing explains why top- $k > 1$ configurations consistently outperform top- $k = 1$: by consulting multiple experts per prediction, the model compensates for imperfect routing and leverages complementary expert perspectives.

Fig. 4 presents the MAE obtained by passing all test samples from each family through each expert individually (5 experts, top- $k = 1$). This reveals clear competency patterns: Expert 2 achieves the lowest MAE for *Others* (0.927), Expert 3 for *Iron-based* (0.621), Expert 1 for both *Cuprate* (0.423) and *Heavy Fermion* (0.888), and Expert 4 for *H-based* (0.865). While each family has a clear best expert, the performance gaps are often modest—for *Others*, the spread between best (0.927) and worst (0.951) is only 0.024, while for *Iron-based*, the gap between Expert 3 (0.621) and Expert 4 (0.825) is more substantial at 0.204. This indicates that experts develop varying degrees of specialization depending on the distinctiveness of each family’s structural signatures.

The consistent dominance of Expert 1 on both *Cuprate* and *Heavy Fermion* families is particularly striking and physically meaningful. Despite these families having different chemical compositions and crystal structures, they share a common underlying physics: both exhibit unconventional superconductivity driven by strong electron correlations. Cuprates feature strongly correlated electrons in CuO_2 planes, while Heavy Fermion systems involve f-electron hybridization leading to enhanced electron-electron interactions. The emergence of a single expert that excels at both families suggests that Expert 1 has learned to recognize structural features indicative of strongly correlated electron physics, rather than memorizing family-specific patterns. This represents a form of *physics-aware specialization* that transcends arbitrary material classifications. In contrast, the *Others* category shows minimal expert differentiation, consistent with its heterogeneous nature as revealed by the t-SNE embeddings (Fig. 5).

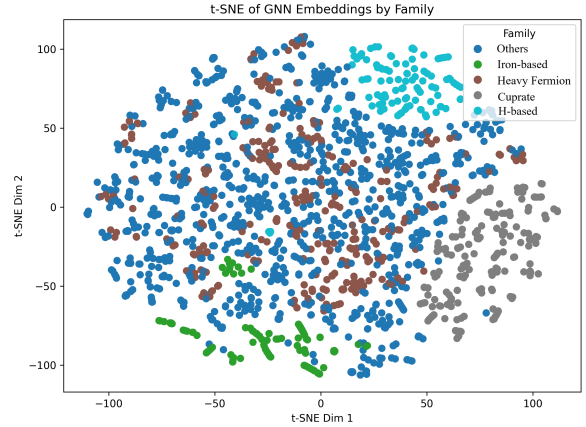


Figure 5: **t-SNE of GNN Embeddings by Family.** t-SNE visualization of graph embeddings colored by superconducting material families.

2.3 Shared learning and Integrative prediction as a Key Mechanism in Mixture-of-Expert

Given the evidence against explicit specialization, we investigated alternative explanations for the MoE’s superior performance.

Our most critical comparison involves the “Forced Specialization” benchmark. The best 5 experts MoE $k = 2$ model (MAE=0.681, R^2 =0.640) outperforms the system of 5 GNNs trained in isolation (MAE=0.683, R^2 =0.628) (Table 2). This result demonstrates that strict expert isolation is suboptimal. GNN experts within the MoE, by training on the full, diverse dataset, benefit immensely from shared learning. Knowledge transferred from data-rich families allows them to build more robust and

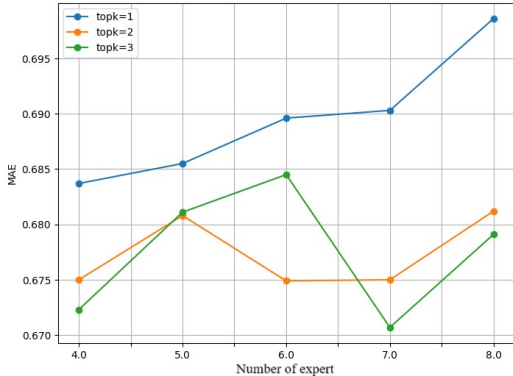


Figure 6: The performance of the Mixture of Experts (MoE) model predicting T_c . The plot illustrates the effect of varying the number of experts and the top k routing value (colored lines) on the final test set MAE.

Table 2: Specialization ablation on the SuperCon3D dataset, evaluated using mean absolute error (MAE) and coefficient of determination (R^2) in logK. Results are reported for the SuperCon3D dataset.

Method	Variant	MAE	R^2
Forced Specialization	5 GNN	0.683 ± 0.094	0.628 ± 0.079
	Specialists		
<i>SuMoE</i>	5 Experts ($k = 1$)	0.686 ± 0.062	0.621 ± 0.071
	5 Experts ($k = 2$)	0.681 ± 0.060	0.640 ± 0.063

generalizable representations of crystal chemistry and physics, making them better predictors even for data-scarce families compared to GNNs trained only on those families’ limited data. This highlights the importance of broad data exposure for training powerful GNNs for materials science.

2.4 Representational Similarity via Centered Kernel Alignment

Having established that the MoE architecture outperforms both single-model and forced-specialization baselines, we now investigate the mechanism underlying this success. We hypothesize that the MoE benefits from a dual mechanism: a **shared representational foundation** that transfers knowledge across families, combined with **subtle, task-relevant divergence** that enables integrative prediction. To test this, we performed a Centered Kernel Alignment (CKA) anal-

ysis on the learned representations of the GNN experts on the best Mixture of Experts config (7 Experts, top- $k=3$). CKA is a robust, scale-invariant metric for comparing the feature spaces learned by different neural networks. We extracted graph embeddings from the output of the final Graph Transformer block for materials from each major superconductor family.

The CKA results (Fig. 7) reveal a representational structure that directly explains why the MoE architecture succeeds: experts maintain high representational overlap while exhibiting systematic, family-dependent divergence. This pattern is precisely what enables effective integrative prediction.

Across all families, pairwise CKA scores remain above 0.89, indicating that all experts converge on a common set of fundamental features encoding crystal chemistry and structural motifs. This shared foundation is the mechanism by which the MoE avoids the pitfalls of forced specialization where each expert benefits from the full dataset rather than being starved of data from other families. The contrast with isolated specialists (which achieve $MAE = 0.683$ compared to MoE’s 0.671) confirms that this representational overlap translates directly to improved generalization.

2.4.1 Family-dependent Divergence Enables Complementary Prediction

Crucially, the degree of expert similarity varies systematically across material families, revealing where soft specialization emerges. The Cuprate family (Fig. 7a) exhibits the *lowest* overall CKA scores, with particularly notable divergence involving Experts 3 and 7. Expert 3 shows reduced similarity with Experts 4, 5, and 6 (CKA = 0.890, 0.899, 0.907 respectively), while the Expert 1-7 pair reaches the global minimum of 0.893. Interestingly, the remaining experts (2, 4, 5, 6) form a more tightly coupled cluster with pairwise scores exceeding 0.95. This suggests the architecture has discovered multiple complementary perspectives for cuprate materials: a majority cluster capturing shared layered-structure features, with Experts 3 and 7 developing alternative representational strategies sensitive to distinct aspects of CuO_2 plane geometry. The learned gating network can leverage these complementary views, synthesizing predictions from experts with genuinely different sensitivities.

For the Iron-based family (Fig. 7c), a striking pattern emerges: Expert 3 stands out as a clear outlier, with CKA scores of 0.917–0.939 against all other experts, while the remaining pairwise comparisons yield very high similarity (> 0.98). Remarkably, this is the *same* expert that shows divergence in the Cuprate family, suggesting Expert 3

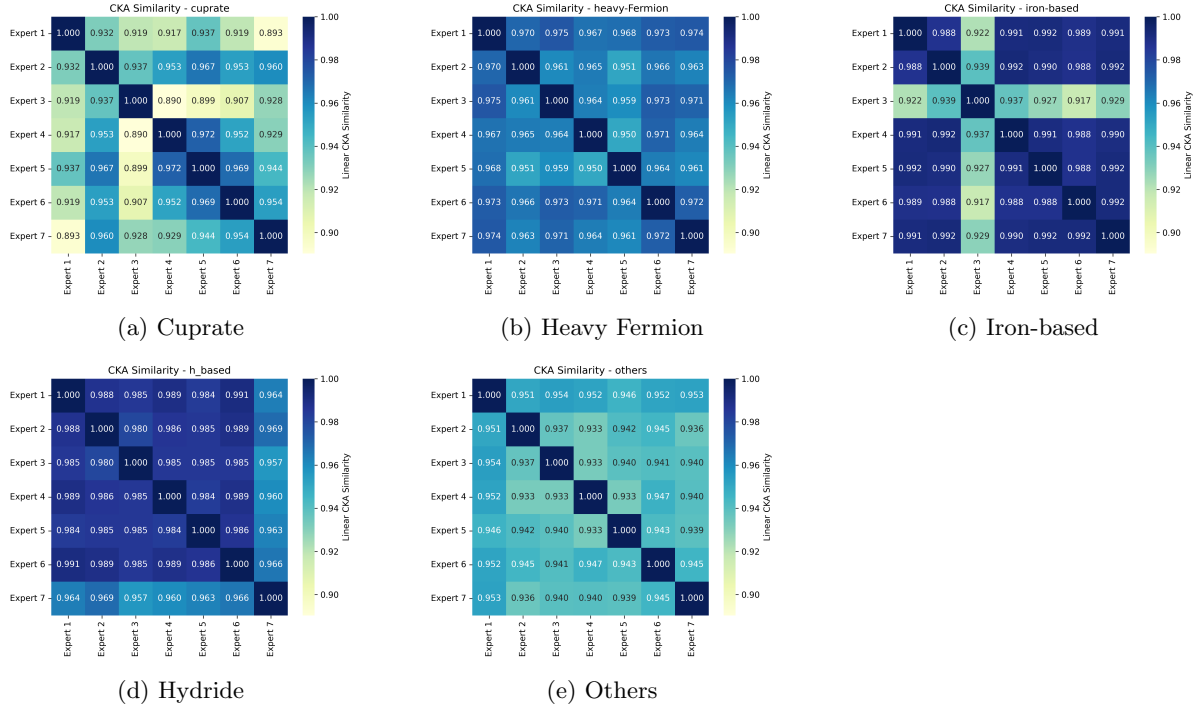


Figure 7: **Centered Kernel Alignment (CKA) Similarity of GNN Expert Representations.** Pairwise CKA similarity scores between the latent representations of the 7 GNN experts for materials from different superconducting families. A score of 1.0 indicates identical representations. The high baseline similarity (> 0.90) confirms shared learning of fundamental crystal chemistry, while the systematic variation across families, with Cuprates showing the greatest divergence (min CKA = 0.893), reveals soft specialization where experts develop complementary sensitivities to family-specific structural motifs.

has developed a specialized representational strategy that captures structural features relevant across multiple high- T_c families. This cross-family consistency provides strong evidence that soft specialization emerges organically from end-to-end training, with certain experts learning to encode physically meaningful features that transfer across chemically distinct material classes.

The Heavy Fermion family (Fig. 7b) shows the most uniform similarity pattern, with all pairwise scores exceeding 0.950 and no clear outliers. Despite the unconventional superconducting mechanisms in these strongly correlated systems, experts converge on a shared representational foundation. This uniformity is physically meaningful: it suggests that the features relevant to Heavy Fermion superconductivity, likely related to f -electron hybridization and Kondo physics, are universally captured by all experts through shared learning, rather than requiring specialized divergence.

For Hydride systems (Fig. 7d), we observe high overall similarity with a subtle pattern: Expert 7 shows slightly reduced CKA scores with all other experts (0.957–0.969), while the remaining pairs cluster above 0.98. This mild divergence may reflect Expert 7 developing sensitivity to the distinct

tive hydrogen-rich coordination environments and high-pressure structural motifs characteristic of hydride superconductors.

The heterogeneous ‘Others’ category (Fig. 7e) provides a critical control: despite spanning diverse chemistries and structures, CKA scores remain uniformly distributed in the range 0.933–0.954 with no clear outlier experts. This confirms that the divergence patterns observed for Cuprates and Iron-based systems are not noise but reflect genuine soft specialization, when clear structural motifs exist, specific experts develop complementary representations; when materials are chemically heterogeneous and lack dominant structural drivers, experts appropriately converge on general-purpose features.

2.4.2 Cross-family Expert Behavior

A particularly compelling finding is the consistent behavior of Expert 3 across families. This expert shows the strongest divergence in both Cuprate and Iron-based systems, the two families with the clearest structural signatures (layered CuO_2 planes and Fe-pnictide/chalcogenide layers, respectively). This suggests that Expert 3 has learned to encode structural features related to layered superconducting motifs that are physically relevant across chem-

ically distinct material classes. Such cross-family specialization would be impossible under forced isolation, where each expert sees only one material family, and demonstrates the power of shared learning to discover transferable physical insights.

The CKA analysis reveals that the MoE’s success arises from a principled balance: sufficient representational overlap to ensure robust generalization through shared learning, combined with targeted divergence that provides complementary predictive signals for the gating network to integrate. This is fundamentally different from both a single monolithic model (which lacks the capacity for complementary predictions) and rigidly specialized models (which sacrifice the benefits of shared learning). The architecture naturally discovers which families benefit from specialized treatment and which are best served by the shared foundation alone.

2.5 Case Studies and New Predictions

To further validate our model, we used it to predict the T_c for several recently discovered superconductors that were not present in the original training data. The results, summarized in Table 3. This demonstrates the model’s potential for extrapolating beyond its training data and its utility as a screening tool for new materials discovery.

The prediction for CaH_6 (217.4 K) is remarkably close to the experimental value (215 K). This highlights the model’s strength in handling hydride materials, which are often high- T_c . For several conventional/low- T_c materials like CsV_3Sb_5 , $\text{Zr}_4\text{Pd}_2\text{O}$, and $\text{Zr}_4\text{Rh}_2\text{O}$, the predictions are in close agreement with the experimental values, with the difference typically being less than 0.2 K.

However, there are some noted failure cases. The prediction for the doped material $\text{Cs}(\text{V}_{0.93}\text{Nb}_{0.07})_3\text{Sb}_5$ (2.79 K) is a considerable underestimation of the experimental value (4.45 K). The model predicted $T_c \approx 2.79$ K, but the experimental value is 4.45 K. The error is ≈ 1.66 K. The model failed to capture the significant T_c increase from the parent compound CsV_3Sb_5 ($T_c \approx 2.3$ K)

3 Methods

3.1 Dataset Preparation

The study utilized the SuperCon3D database[5], which contains experimental data for 1,578 superconducting materials. This database consists of entries with well-defined crystal structures available in CIF format and a single reported T_c value.

3.2 Crystal Graph Representation

Each crystalline material was converted into a graph representation, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of nodes $\mathcal{V}$ corresponds to the atoms in the unit cell and the set of edges $\mathcal{E}$ represents interatomic connections. An edge was created between two atoms if their distance was less than a cutoff radius of 8 Å, chosen based on prior GNN studies in superconductivity [5].

Node (atom) features, $\mathbf{v}_i$, were generated based on intrinsic atomic properties, resulting in a 92-dimensional feature vector for each atom. These features include atomic number, period, group, electronegativity, and covalent radius.

Edge features, $\mathbf{e}_{ij}$, were derived from the distance between atoms i and j , expanded into a 20-dimensional vector using a Gaussian radial basis function (RBF).

3.3 Site-Occupation Disorder Handling

For materials with site-occupation disorder (SOD), where a crystallographic site is occupied by multiple elements with fractional occupancies, we adopt an occupancy-weighted feature representation. For each disordered site, the atom embedding is computed as a weighted sum of the constituent element embeddings, where the weights correspond to the fractional occupancies defined in the CIF file. Specifically, for a site i with elements $\{e_k\}$ and occupancies $\{o_k\}$, the atom feature vector is:

$$\mathbf{v}_i = \sum_k o_k \cdot \mathbf{f}_{e_k} \quad (1)$$

where $\mathbf{f}_{e_k}$ is the embedding vector for element e_k . This deterministic approach encodes the compositional distribution at each site directly into the node features, enabling the model to distinguish between ordered and disordered structures without requiring stochastic sampling or data augmentation.

3.4 SuMoE Architecture

The overall architecture of SuMoE model is presented in Fig. 1. The SuMoE model is built upon a graph attention transformer architecture. The core of the model consists of a stack of interaction blocks that update the atom representations by aggregating information from their local neighborhood. The model architecture comprises three main parts: embedding layers, interaction blocks, and a Mixture of Experts components.

3.4.1 Embedding Layers

Initial atom and edge features are passed through linear layers to create learnable embeddings.

Table 3: Prediction on recently discovered materials not in the training set.

Material Formula	Predicted T_c (K)	Experimental T_c (K)
CaH ₆	217.4	215
CsV ₃ Sb ₅	2.28	2.3
Cs(V _{0.93} Nb _{0.07}) ₃ Sb ₅	2.79	4.45
LaFeSiO	7.54	10
Zr ₄ Pd ₂ O	2.62	2.73
Zr ₄ Rh ₂ O	3.58	3.73

3.4.2 Interaction Blocks

The model uses 6 graph transformer blocks. Each block updates the node embeddings $\mathbf{h}_i$ by attending to the messages from its neighbors. The update process can be described by the following equations:

$$\mathbf{m}_{ij} = \text{MLP}(\mathbf{h}_i \parallel \mathbf{h}_j \parallel \mathbf{e}_{ij}) \quad (2)$$

$$\alpha_{ij} = \text{softmax}_j(\text{attention}(\mathbf{h}_i, \mathbf{h}_j)) \quad (3)$$

$$\mathbf{h}'_i = \text{update}(\mathbf{h}_i, \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{m}_{ij}) \quad (4)$$

where $\parallel$ denotes concatenation, and MLP, attention, and update are learned functions. This architecture employs a multi-head attention mechanism where edge features are integrated into the attention score calculation, allowing the model to dynamically weigh the importance of neighboring atoms based on both their features and their geometric arrangement. Each attention layer is followed by a position-wise feed-forward network, with residual connections and layer normalization applied after each sub-layer.

3.4.3 Mixture of Experts (MoE) Output Layer

The central innovation of our work is the replacement of a standard, monolithic regression head with a Mixture of Experts (MoE) model. This design addresses the challenge of chemical diversity within the superconductor landscape. After the final interaction block, a global representation for the entire crystal, $\mathbf{h}_{\text{graph}}$, is obtained by performing a graph-level mean pooling operation over all atom embeddings. This fixed-size vector, which encodes the structural and chemical information of the material, serves as the input to the MoE layer.

The MoE framework is composed of two main stages. First, an embedding GNN using Graph Transformer architecture processes the input graph to produce a graph-level embedding $\mathbf{h}_{\text{graph}}$. This embedding is fed to a gating network. The gating network, G , is a shallow neural network that acts as a learned router. It takes the graph representation $\mathbf{h}_{\text{graph}}$ and outputs a probability distribution over the available experts. In our implementation, with

N experts, the gating network produces a vector of weights $p \in \mathbb{R}^N$:

$$p = \text{softmax}(W_g \mathbf{h}_{\text{graph}} + b_g) \quad (5)$$

where W_g and b_g are learnable weights and biases of a linear layer, N is the number of experts. Each component p_i of this vector represents the probability that expert i is the most suitable for the given input material.

Second, the original crystal structure graph $\mathbf{G}$ is fed as input to a set of N parallel GNN experts. Each expert is a full Graph Transformer network with independent parameters that processes the crystal graph and outputs a T_c prediction.

Rather than using all experts, we employ top- k routing where only the k experts with the highest gating weights are activated for each input. The final predicted critical temperature, $T_{c,\text{pred}}$, is a weighted average of the outputs from the selected top- k experts, where the weights are renormalized:

$$T_{c,\text{pred}} = \sum_{i \in \mathcal{T}_k} \hat{p}_i \cdot \text{GNNExpert}_i(\mathbf{G}) \quad (6)$$

where $\mathcal{T}_k$ denotes the set of indices corresponding to the top- k experts with highest routing probabilities, and $\hat{p}_i = \text{softmax}_{j \in \mathcal{T}_k}(p_i)$ are the renormalized weights over the selected experts. This sparse routing mechanism reduces computational cost while maintaining model expressiveness. It enables the model to learn a "soft" division of labor, where it can combine the knowledge of multiple specialized experts to make a more robust prediction.

3.5 Training Procedure

The model was trained end-to-end using the AdamW optimizer[9] with a learning rate of 1e-6, a weight decay of 5e-3, and a batch size of 64. The learning rate was scheduled using a cosine annealing schedule with a warm-up phase of 10 epochs. The loss function was the Mean Absolute Error (MAE) between the predicted and experimental T_c values. The model was trained for 150 epochs on NVIDIA V100 GPUs, and the model with the best performance on the validation set was selected for final evaluation on the test set.

Conclusion

Analysis of expert performance reveals a regime of soft specialization, where each expert prioritizes specific superconductor families while maintaining the capacity to capture features beyond their primary domain. This is evidenced by the narrow performance gaps between the best and second-best experts. Such shared learning allows the MoE to integrate complementary information, thereby enhancing overall prediction robustness. These insights suggest that for complex structure-property relationships, integrative prediction with soft specialization offers a superior balance compared to strict isolation or single-family specialization.

Data Availability

The SuperCon dataset used in this study is publicly available at <https://supercon.nims.go.jp/>. The processed data and crystal structures of Supercon3D used for training are available on Github repository <https://github.com/ngminhtri0394/SuMoE>.

Code Availability

The source code for the SuMoE model, along with scripts for data processing, training, and inference, is available on GitHub at <https://github.com/ngminhtri0394/SuMoE>.

Author Contributions

K.G.H.N implemented, ran, and analyzed the experiments. H.L.D wrote, proofread, and revised the paper. M.T.N designed the research program and derived the initial implementation.

Competing Interests

The authors declare no competing interests.

Funding declaration

Funding: Not applicable

References

- [1] J. G. Bednorz and K. A. Müller. "Possible high T_c superconductivity in the Ba-La-Cu-O system". In: *Zeitschrift für Physik B Condensed Matter* 64.2 (June 1986), pp. 189–193.
- [2] Lilia Boeri et al. "The 2021 room-temperature superconductivity roadmap". In: *Journal of Physics: Condensed Matter* 34.18 (Mar. 2022), p. 183002.
- [3] Keith T. Butler et al. "Machine learning for molecular and materials science". In: *Nature* 559.7715 (July 2018), pp. 547–555.
- [4] Chi Chen et al. "Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals". In: *Chemistry of Materials* 31.9 (May 2019), pp. 3564–3572.
- [5] Pin Chen et al. "Learning Superconductivity from Ordered and Disordered Material Structures". In: ().
- [6] Elbio Dagotto. "Correlated electrons in high-temperature superconductors". In: *Reviews of Modern Physics* 66.3 (July 1994), p. 763.
- [7] José A. Flores-Livas et al. "A perspective on conventional high-temperature superconductors at high pressure: Methods and materials". In: *Physics Reports* 856 (Apr. 2020), pp. 1–78.
- [8] B. Keimer et al. "From quantum matter to high-temperature superconductivity in copper oxides". In: *Nature* 518.7538 (Feb. 2015), pp. 179–186.
- [9] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *7th International Conference on Learning Representations, ICLR 2019* (Nov. 2017).
- [10] Atsuto Seko et al. "Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization". In: *Physical Review Letters* 115.20 (Nov. 2015), p. 205901.
- [11] Noam Shazeer et al. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer". In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* (Jan. 2017).
- [12] Valentin Stanev et al. "Machine learning modeling of superconducting critical temperature". In: *npj Computational Materials* 4.1 (Dec. 2018), pp. 1–14.
- [13] Anthony Yu Tung Wang et al. "Compositionally restricted attention-based network for materials property predictions". In: *npj Computational Materials* 7.1 (Dec. 2021), pp. 1–10.
- [14] Tian Xie and Jeffrey C Grossman. "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties". In: (2018).

- [15] Chengquan Zhong et al. “Enhancing Superconductor Critical Temperature Prediction: A Novel Machine Learning Approach Integrating Dopant Recognition”. In: *ACS applied materials & interfaces* 16.44 (Nov. 2024), pp. 60472–60481.