Supplementary Information: Hybrid-LLM-GNN: Integrating Large Language Models and Graph Neural Networks for Enhanced Materials Property Prediction

Youjia Li[1], Vishu Gupta[1,2,3], Muhammed Nur Talha Kilic[4], Kamal Choudhary[5], Daniel Wines[5], Wei-keng Liao[1], Alok Choudhary[1], Ankit Agrawal[1]

[1]*Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA.*

[2]*Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA.*

[3]*Ludwig Institute for Cancer Research, Princeton University, Princeton, NJ, USA.*

[4]*Department of Computer Science, Northwestern University, Evanston, IL, USA.*

[5]*Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD, USA.*

Table 1: Dataset size and baseline test MAE using mean of combined train and validation sets as the prediction on the test set for evaluated material properties

Property	**Train set size**	**Validation set size**	**Test set size**	**Baseline MAE**
Formation Energy (eV/atom)	60639	7580	7580	0.8546
Ehull (eV/atom)	59939	7494	7493	0.2618
Magout (μ_B)	59370	7422	7422	2.1585
$\text{Bandgap}_{\text{mBJ}}$ (eV)	15642	1957	1957	1.6565
Spillage	9041	1131	1131	0.5055
SLME (%)	7810	977	977	10.7721
Tc_supercon (K) (eV)	842	106	106	3.2085

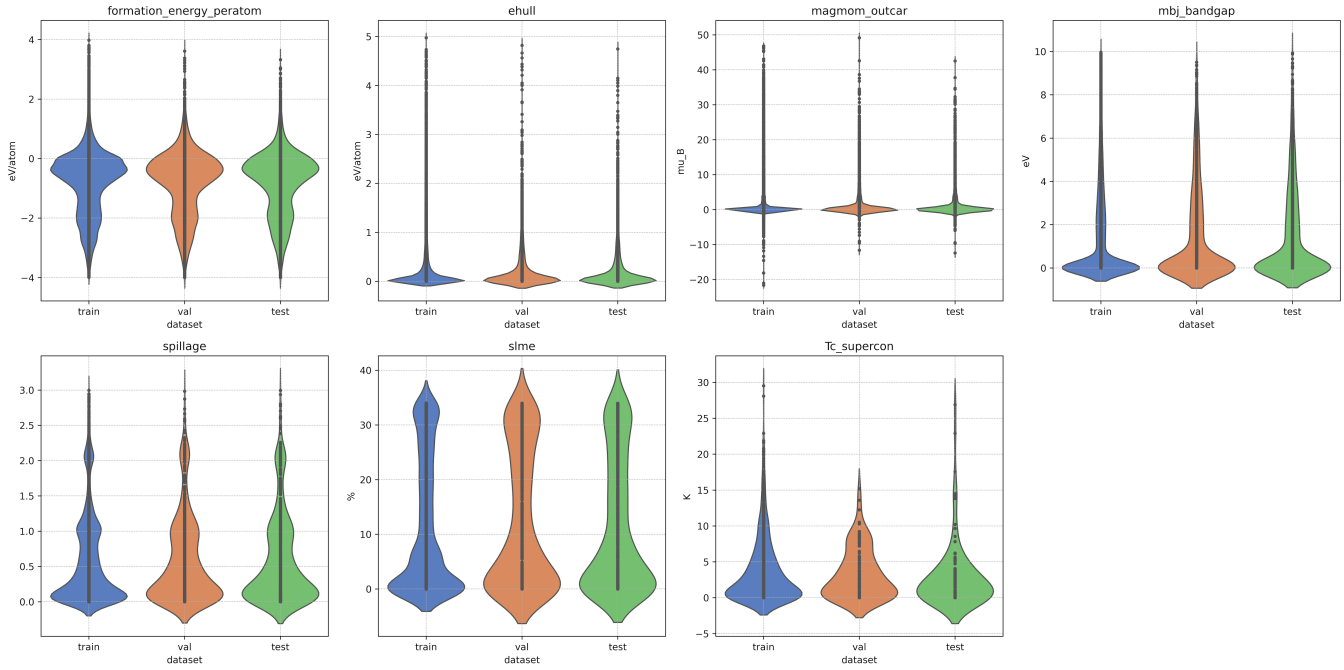

Figure 1: Violin plot for the distribution of DFT-calculated properties; Displayed material properties include: Formation Energy, Ehull, Magout, $\text{Bandgap}_{\text{mBJ}}$, Spillage, SLME and Tc_supercon

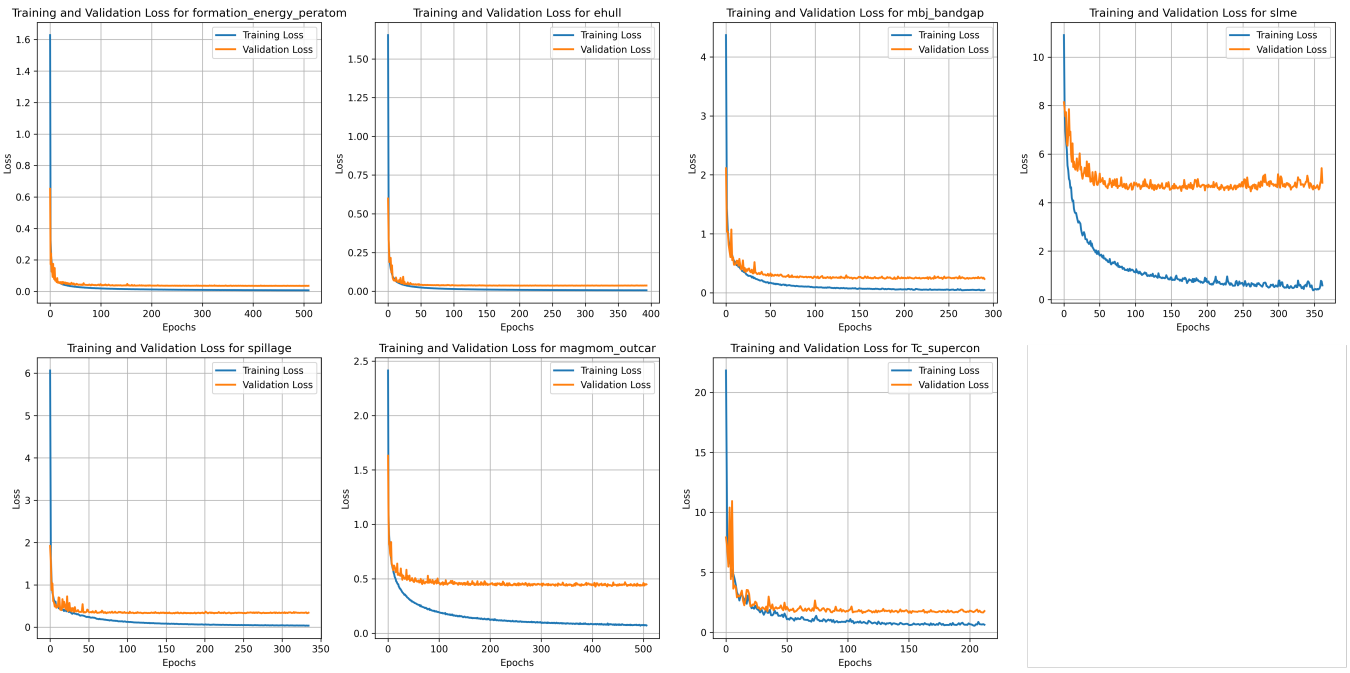

Figure 2: Loss curves for the predictor model of ALIGNN-MatBERT TL approach. Early-stopping is applied in model training, which decides the number of epoches. Displayed material properties include: Formation Energy, Ehull, Magout, $\text{Bandgap}_{\text{mBJ}}$, Spillage, SLME and Tc_supercon

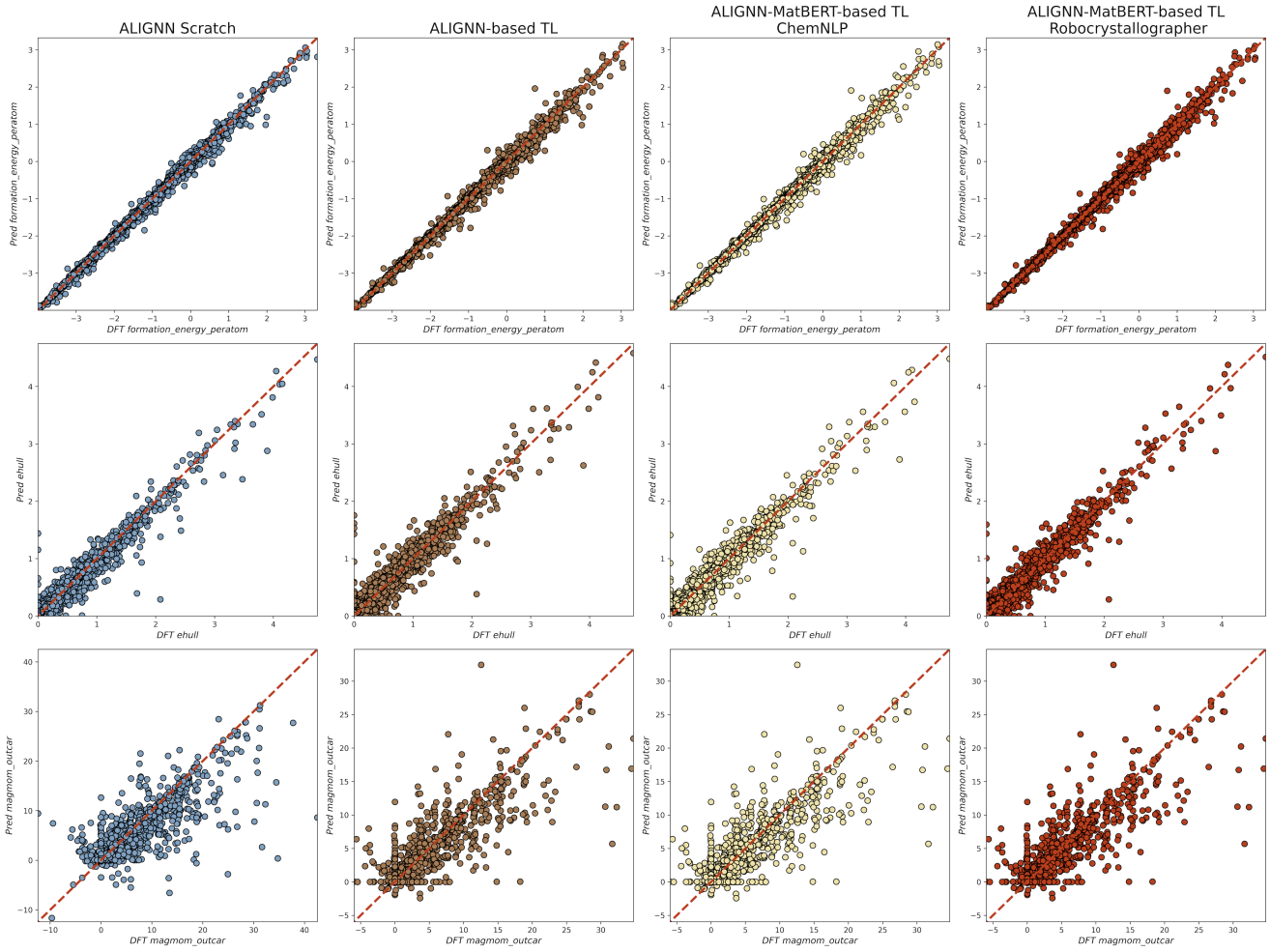

Figure 3: Parity plot of the original ALIGNN model (1st col), ALIGNN-based feature extraction transfer learning (2nd col), and the proposed hybrid transfer learning approach based on ChemNLP (3rd col) and Robocystallographer (4th col). Displayed material properties include: Formation Energy, Ehull, and Magout

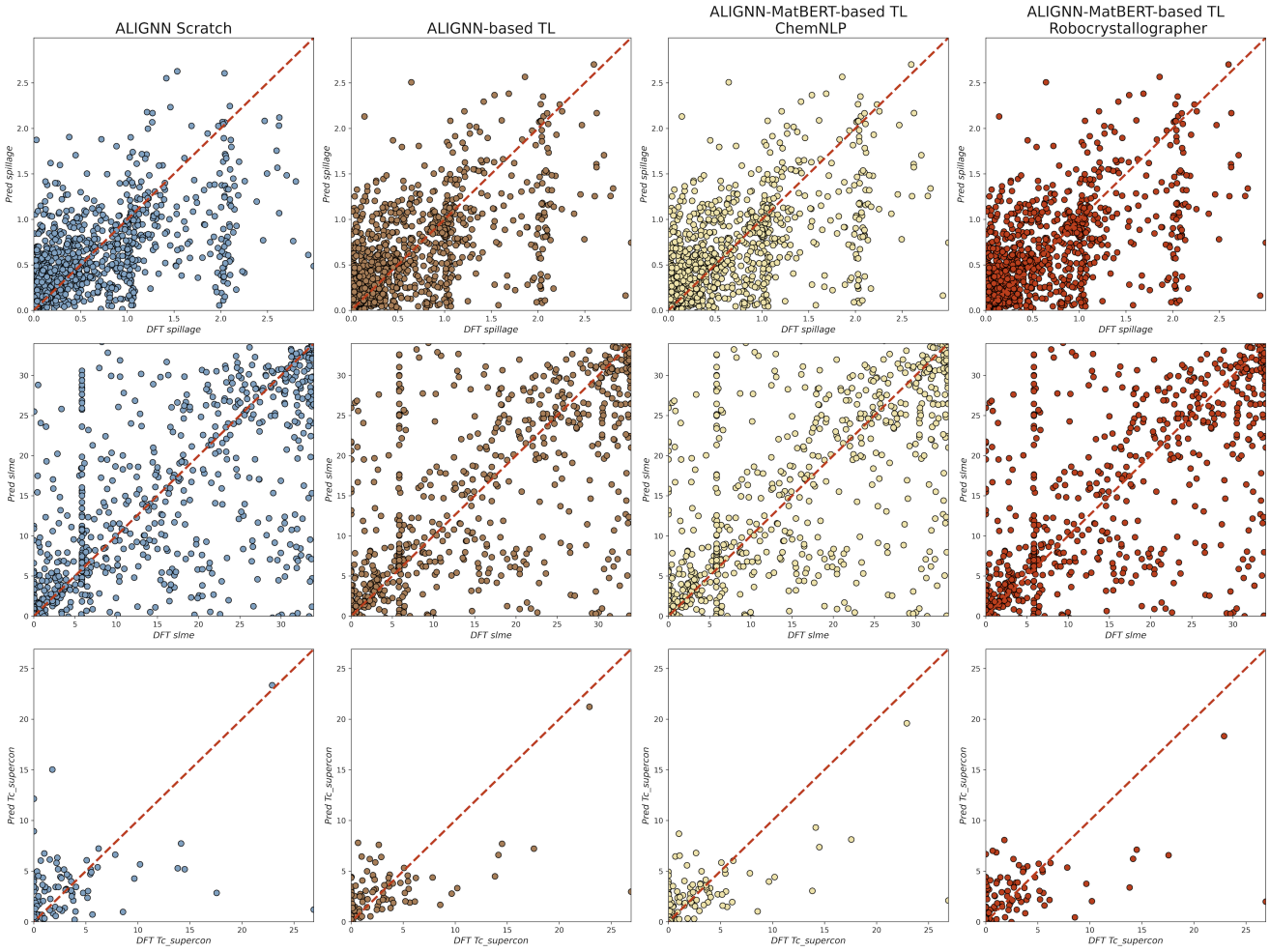

Figure 4: Parity plot of the original ALIGNN model (1st col), ALIGNN-based feature extraction transfer learning (2nd col), and the proposed hybrid transfer learning approach based on ChemNLP (3rd col) and Robocystallographer (4th col). Displayed material properties include: Spillage, SLME and Tc_supercon

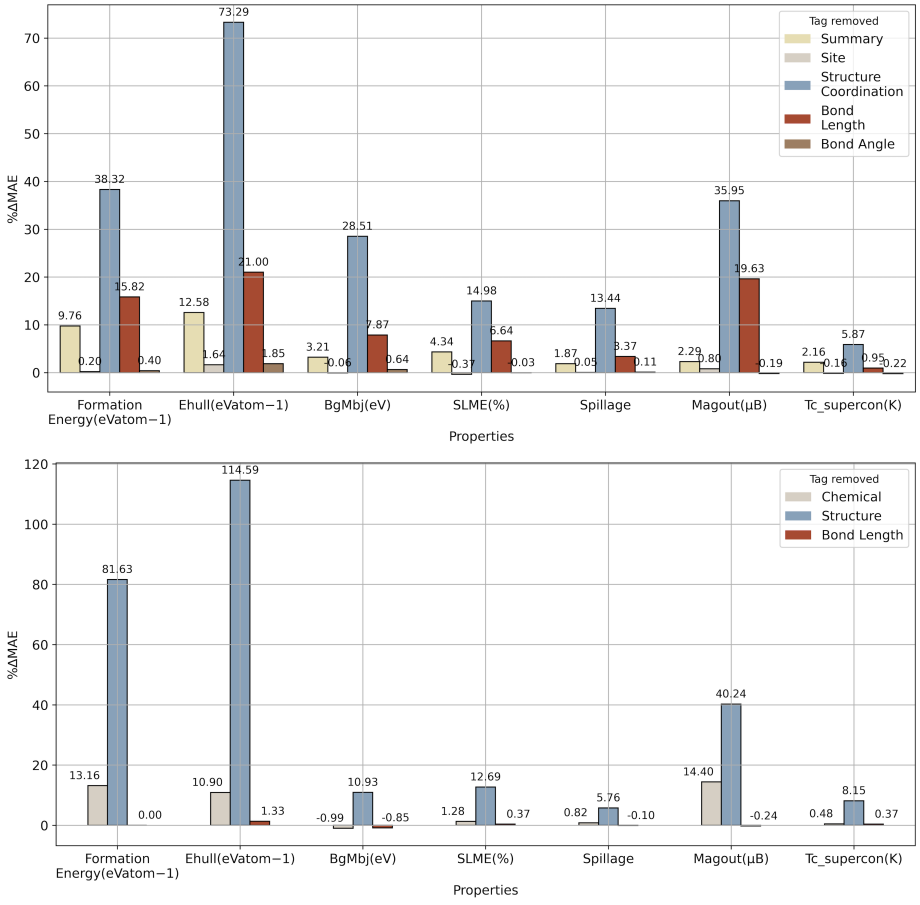

Figure 5: Percentage MAE change for ALIGNN-MatBERT-based TL approach after removal of tagged content from text representation across tested material properties (top: Robocrystallographer; bottom: ChemNLP)