

Machine Learning–Driven Prediction of Graphene Properties from Atomistic Simulations

by

Makhmud Assentay

Submitted to the Department of Data Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

June 2026

© Nazarbayev University 2026. All rights reserved.

Author
Department of Data Science
May 2026

Certified by
Siamac Fazli
Associate professor, School of Engineering and Digital Sciences
Thesis Supervisor

Accepted by
Elizabeth Arkhangelsky
Dean, School of Engineering and Digital Sciences

Machine Learning–Driven Prediction of Graphene Properties from Atomistic Simulations

by

Makhmud Assentay

Submitted to the Department of Data Science
on May 2026, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Abstract

Chemical functionalization of graphene with hydrogen (H) or methyl (CH₃) groups provides a powerful handle for tuning thermomechanical properties, but the combinatorially vast configuration space makes exhaustive simulation-based exploration impractical. This thesis presents a comprehensive machine learning benchmark for predicting four thermomechanical properties—thermal conductivity (κ), Young’s modulus, maximum stress, and maximum strain—of chemically functionalized graphene nanosheets from atomistic molecular dynamics simulations.

Two datasets of 300 configurations each, covering H- and CH₃-functionalized graphene nanosheets of 8,528 carbon atoms at coverages up to 15%, were generated using the AIREBO interatomic potential in LAMMPS. Five model families are evaluated on a common benchmark: classical machine learning regressors (KNN, ridge, tree ensembles, SVR, GPR), dense multi-layer perceptrons trained from scratch, pretrained 2-D convolutional neural networks (ConvNeXt-Tiny, EfficientNetV2-S, ResNet50, MobileNetV2, AlexNet), vision transformers (ViT-B/16, ViT-B/32, Swin-T, Swin-S), and graph neural networks (GCN, GraphSAGE, GIN, GAT) operating on the native honeycomb molecular graph. Four structural representations are compared systematically: a 3-feature atom-count vector, the full 8,528-element atom-type vector, a 2-D lattice image, and the honeycomb molecular graph.

The central finding is that a 3-feature KNN model capturing only global functionalization density achieves $R^2 \geq 0.986$ on κ , stress, and Young’s modulus for the H-functionalized dataset, matching or exceeding all deep learning architectures. Maximum strain remains universally intractable ($R^2 \leq 0.38$ across all families), consistent with fracture initiation being governed by localized structural events not captured by any tested representation. Vision transformers with frozen ImageNet backbones demonstrate the best data efficiency at extreme low-data regimes (mean $R^2 = 0.861 \pm 0.128$ on κ with only 12 training samples), while GNN models achieve competitive accuracy from scratch using only 3-dimensional node features. The

methyl dataset proves substantially harder than the hydrogen dataset for all targets except Young’s modulus, with GNN and ViT models exhibiting greater robustness to this domain shift than fully fine-tuned CNNs.

These results establish that for H-functionalized graphene, computationally cheap surrogates based on global density statistics are sufficient for most engineering targets, while spatial representations offer incremental benefits primarily for strain prediction and for the structurally richer methyl chemistry.

Thesis Supervisor: Siamac Fazli

Title: Associate professor, School of Engineering and Digital Sciences

Acknowledgments

I would like to express my sincere gratitude to my thesis supervisor, Associate Professor Siamac Fazli, for his guidance, patience, and insightful feedback throughout this research. His expertise in machine learning and materials informatics shaped the direction of this work and pushed me to think more rigorously about every experimental design decision.

This work was carried out as part of the DOG-ML research project. I am grateful to Farrokh Yousefi and Narges Vafa for their help in preparing the LAMMPS input scripts used to generate the molecular dynamics trajectories, and to Ravil Ashirmametov for running the simulations and assembling the dataset analyzed in this thesis [1]. The dataset is openly available at the project repository [2].

I am grateful to the School of Engineering and Digital Sciences at Nazarbayev University for providing the computational resources necessary for this project, in particular access to the Shabyt HPC cluster and the RTX 4090 GPU workstation maintained in Prof. Fazli’s lab. The large-scale molecular dynamics simulations and multi-seed deep learning experiments reported in this thesis would not have been feasible without this infrastructure.

I thank the members of my thesis committee and the faculty of the Department of Data Science for their constructive comments during the proposal and defense stages of this work.

I am deeply thankful to my family for their unwavering support and encouragement throughout my graduate studies. Their belief in me sustained my motivation during the most demanding phases of this project.

Finally, I acknowledge the open-source communities behind LAMMPS, PyTorch,

PyTorch Geometric, scikit-learn, and the Hugging Face `timm` library, whose tools formed the computational backbone of this thesis.

Contents

Acknowledgments	5
1 Introduction	15
1.1 Motivation	15
1.2 Problem Statement	17
1.3 Research Objectives	18
1.4 Thesis Contributions	19
1.5 Thesis Outline	19
2 Related Works	21
2.1 Graphene Modification Strategies	22
2.1.1 Chemical Functionalization	22
2.1.2 Substitutional Doping	23
2.2 Computational Methodologies	24
2.3 Machine Learning Models for Property Prediction	24
2.3.1 Gaussian Process Regression	24
2.3.2 Gradient Boosting Methods	25
2.3.3 Artificial Neural Networks	25
2.3.4 Support Vector Methods	26
2.4 Encoding Strategies for Graphene Structures	26
2.4.1 Structural Representation Challenges	26
2.4.2 Geometric and Topological Encodings	27
2.5 Comprehensive Reviews	27

2.6	Research Gaps and Limitations	27
3	Methodology	31
3.1	Nanostructure Generation	31
3.2	Molecular Dynamics Simulations	33
3.2.1	Interatomic Potential	33
3.2.2	Mechanical Properties	33
3.2.3	Thermal Conductivity	34
3.3	Datasets	34
3.3.1	Dataset Structure	34
3.3.2	Atom-Type Encoding	35
3.3.3	Target Properties	36
3.3.4	Exploratory Data Analysis	36
3.3.5	Lattice-to-Grid Mapping	38
3.4	Structural Representations	39
3.5	Machine Learning Models	41
3.5.1	Classical ML and Simple Baselines	41
3.5.2	Dense Multi-Layer Perceptrons	43
3.5.3	Scratch 2D CNNs	44
3.5.4	Pretrained 2D CNNs	44
3.5.5	Vision Transformers	44
3.5.6	Graph Neural Networks	45
3.6	Training Protocol	46
3.6.1	Data Splitting	46
3.6.2	Target Normalisation	46
3.6.3	Multi-Seed Evaluation	46
3.7	Data Efficiency Analysis	47
3.8	Evaluation Metrics	47
4	Results and Discussion	49
4.1	Classical ML and Scratch Deep Learning	49

4.1.1	Hydrogen Dataset	49
4.1.2	Methyl Dataset	52
4.2	Pretrained 2D CNNs	53
4.2.1	Hydrogen Dataset	53
4.2.2	Methyl Dataset	54
4.3	Vision Transformers	55
4.3.1	Hydrogen Dataset	55
4.3.2	Methyl Dataset	56
4.4	Graph Neural Networks	56
4.4.1	Hydrogen Dataset	56
4.4.2	Methyl Dataset	57
4.5	Cross-Domain Comparison	58
4.6	Data Efficiency Analysis	59
4.6.1	Classical ML and Scratch Deep Learning	59
4.6.2	Pretrained 2D CNNs	60
4.6.3	Vision Transformers	61
4.6.4	Graph Neural Networks	62
4.6.5	Cross-Domain Data Efficiency Summary	63
4.7	Summary of Key Findings	64
5	Conclusion	65
5.1	Summary	65
5.2	Key Findings	66
5.3	Broader Implications	68
5.4	Limitations	69
5.5	Future Work	69

List of Figures

3-1	Exploratory analysis of the two datasets. (a) Functionalization coverage distributions: H-functionalized samples span 1–15% (n=300), while the usable CH ₃ samples span 1–8% (n=200, after NaN removal). (b–e) Target property distributions for both datasets. (f) Pearson correlation matrix for the H-functionalized targets.	36
3-2	Coverage versus target properties for both datasets. Each point is one configuration (H: n=300, 1–15%; CH ₃ : n=200, 1–8%); Pearson r values are annotated. (a) κ decreases monotonically with coverage ($r_H = -0.93$, $r_M = -0.92$). (b) Maximum strain shows weak and inconsistent coverage dependence ($r_H = -0.60$, $r_M \approx 0$). (c, d) Stress and Young’s modulus exhibit near-perfect negative correlation with coverage.	38
3-3	Visual overview of the four structural representations for the same H-functionalized graphene configuration ($\sim 8\%$ coverage). (a) 3-feature count vector retaining only global atom-type counts. (b) Full 8,528-element atom-type vector preserving per-site identity. (c) 2D lattice image after resizing to 224×224 pixels. (d) Honeycomb molecular graph with physical C–C bond connectivity (patch view).	42
4-1	Learning curves for the ML and scratch deep learning family on the H-functionalized dataset (κ target). Shaded bands indicate ± 1 SD over three random subsamplings.	59

4-2	Learning curves for pretrained 2D CNN models on the H-functionalized dataset (κ target). Shaded bands indicate ± 1 SD over three random subsamplings.	60
4-3	Learning curves for ViT/Swin models on the H-functionalized dataset (κ target).	61
4-4	Learning curves for GNN models on the H-functionalized dataset (κ target). Shaded bands indicate ± 1 SD over three random subsamplings.	63

List of Tables

3.1	Summary of the two functionalized graphene datasets.	35
3.2	Summary of the four structural representations.	41
3.3	MLP architectures used in this study.	43
3.4	Scratch CNN architectures. Spatial resolution after n pooling stages: 224/ 2^n	44
4.1	Mean R^2 ($\pm$ std, 3 seeds) on the H-functionalized test set. ML mod- els use the full 8,528-element feature vector unless noted; KNN-3feat uses the 3-count representation; SVR and GPR use PCA-50 features. Scratch CNN and MLP models are trained from scratch.	50
4.2	Mean R^2 ($\pm$ std, 3 seeds) on the CH ₃ -functionalized test set for selected models.	52
4.3	Mean R^2 ($\pm$ std, 3 seeds) for pretrained 2D CNNs on the H-functionalized dataset. VGG19 is excluded due to systematic failure ($R^2 \approx 0.2$ – 0.4 , consistent across seeds).	53
4.4	Mean R^2 ($\pm$ std, 3 seeds) for pretrained 2D CNNs on the CH ₃ -functionalized dataset.	54
4.5	Mean R^2 ($\pm$ std, 3 seeds) for ViT/Swin models (head-only fine-tuning) on the H-functionalized dataset.	55
4.6	Mean R^2 ($\pm$ std, 3 seeds) for ViT/Swin models on the CH ₃ -functionalized dataset.	56
4.7	Mean R^2 ($\pm$ std, 3 seeds) for GNN models on the H-functionalized dataset.	56

4.8	Mean R^2 ($\pm$ std, 3 seeds) for GNN models on the CH ₃ -functionalized dataset.	57
4.9	Best R^2 (mean $\pm$ std) per model family and target, H-functionalized dataset.	58
4.10	Best R^2 achieved across all model families, per target and dataset. H = H-functionalized, M = CH ₃ -functionalized.	64

Chapter 1

Introduction

1.1 Motivation

Graphene, a single-atom-thick layer of carbon atoms arranged in a honeycomb lattice, exhibits a remarkable combination of mechanical, thermal, and electronic properties that position it among the most promising nanomaterials of the twenty-first century. Its theoretical Young's modulus of approximately 1 TPa, combined with high intrinsic thermal conductivity and an exceptional surface-to-volume ratio, has driven sustained interest across nanoelectronics, composite reinforcement, thermal management, and sensing platforms [8].

Chemical functionalization of graphene offers a direct handle on these properties. Attaching hydrogen atoms or methyl groups (CH_3) to carbon sites modifies bond hybridization, introduces local lattice strain, and alters phonon scattering pathways, thereby tuning mechanical stiffness, thermal conductivity, and fracture behavior over a wide and controllable range [14, 11].

The ability to predict and control these thermomechanical properties through functionalization has direct engineering relevance. In thermal management, functionalized graphene films are candidate materials for heat spreaders in high-power electronics, where tunable thermal conductivity through surface chemistry could enable device-level thermal regulation without geometric redesign. In structural composites, controlled hydrogenation provides a route to adjusting the stiffness-to-weight ratio

of graphene-reinforced polymers. In protective coatings and membranes, the balance between mechanical robustness and chemical reactivity depends on the precise degree and pattern of surface functionalization. In each of these domains, the central bottleneck is the inability to efficiently navigate the vast configuration space of possible functionalization patterns to identify those that achieve a target property profile. A fast and accurate surrogate model that maps functionalization patterns directly to thermomechanical properties would reduce this design cycle from weeks of simulation to seconds of inference, making systematic property optimization tractable for the first time.

The physical mechanisms underlying these property changes are well understood at the level of individual functionalized sites. When a hydrogen atom bonds to a carbon site, the carbon transitions from sp^2 to sp^3 hybridization: the planar trigonal geometry is replaced by a tetrahedral arrangement that pulls the bonded carbon atom out of the graphene plane. This local puckering disrupts the long-range planarity of the lattice and introduces compressive strain in the surrounding ring structure. Mechanically, sp^3 bonds are weaker in the in-plane direction than sp^2 bonds, so each functionalized site acts as a local reduction in stiffness. Thermally, the mass heterogeneity introduced by the attached group and the bond-stiffness discontinuity at each functionalized site scatter acoustic phonons—the primary heat carriers in graphene—reducing their mean free path and, with it, the bulk thermal conductivity. Methyl groups introduce the same $sp^2 \rightarrow sp^3$ transition but with substantially greater steric bulk, producing larger out-of-plane distortions and stronger phonon scattering per functionalized site. Both functionalization types also create local stress concentrations under tensile loading that govern fracture initiation, though the precise relationship between the spatial distribution of these concentrations and macroscopic failure strain remains poorly understood.

The practical challenge is that each carbon site in a realistically sized sheet can independently carry or lack a functional group, making the space of distinct configurations combinatorially vast. Systematic experimental or simulation-based exploration of this space is impractical without intelligent guidance.

Molecular dynamics (MD) simulation with the AIREBO interatomic potential provides a well-validated route to computing thermomechanical properties of functionalized graphene at the atomic scale [12]. Each simulation, however, is computationally expensive. Building a dataset of several hundred configurations with four simultaneous property targets—thermal conductivity, Young’s modulus, maximum stress, and maximum strain—requires significant high-performance computing (HPC) resources and limits the explorable space to a small fraction of the true configuration landscape.

Machine learning (ML) surrogate models offer a complementary strategy: once a labeled dataset is assembled, a surrogate can predict properties orders of magnitude faster than the simulator, enabling rapid screening of configurations, statistical characterization of property distributions, and systematic comparison of structural representations. The central scientific question addressed by this thesis is: *which ML architectures and structural representations provide the most accurate and data-efficient surrogates for the thermomechanical properties of H- and CH₃-functionalized graphene?*

1.2 Problem Statement

Despite a growing body of work on ML-assisted graphene property prediction [7, 19, 6], several gaps motivate the present study.

Existing benchmarks are predominantly single-family comparisons, testing either classical regressors or a single neural network architecture on the same dataset, making principled cross-family conclusions difficult. Most prior MD-based datasets contain 200–600 samples [13], which is sufficient for classical ML but raises unanswered questions about the data requirements of modern deep architectures. The structural representation—how to encode the spatial arrangement of thousands of atom sites—varies widely across the literature without systematic comparison: some studies use global chemical descriptors [17], others the raw atom-type vector, and recent work employs image-based or graph-based encodings [5], but no study has evaluated all of these on a common dataset.

This thesis addresses all three gaps by constructing a controlled, comprehensive benchmark across five model families, four structural representations, and two functionalization chemistries evaluated on a shared LAMMPS-generated dataset.

1.3 Research Objectives

The specific objectives of this thesis are:

1. To curate, from the LAMMPS+AIREBO molecular-dynamics dataset of Ashirmametov et al. [1], a multi-task benchmark of 300 hydrogen-functionalized and 200 methyl-functionalized graphene nanosheets under stricter Non-Equilibrium Molecular Dynamics (NEMD) steady-state validity filtering, with each configuration characterized by four thermomechanical target properties: thermal conductivity (κ), Young’s modulus, maximum stress, and maximum strain.
2. To evaluate classical machine learning models—including KNN regression on a compact 3-feature atom-count representation, ridge regression, elastic net, random forest, extra trees, gradient boosting, support vector regression, Gaussian process regression, and XGBoost—as surrogate models on both datasets.
3. To design and train deep neural network surrogates spanning five architectural families: (i) dense MLP models of increasing depth (1–5 hidden layers) trained from scratch on the full 8,528-element feature vector; (ii) convolutional networks of 2–5 blocks trained from scratch on the 2D lattice image; (iii) pretrained CNNs (ConvNeXt-Tiny, EfficientNetV2-S, ResNet50, MobileNetV2, AlexNet) fine-tuned via transfer learning; (iv) vision transformers (ViT-B/16, ViT-B/32, Swin-T, Swin-S) with frozen ImageNet backbones; (v) graph neural networks (GCN, GraphSAGE, GIN, GAT) operating on the native honeycomb molecular graph.
4. To quantify the data efficiency of each model family through learning-curve experiments, measuring performance as a function of training set size and providing statistical variance estimates via repeated random subsampling.

1.4 Thesis Contributions

The principal contributions of this work are:

Comprehensive multi-family benchmark. A single-dataset, controlled comparison of 33 model variants across five architectural families, two functionalization chemistries, and four target properties. This constitutes the broadest simultaneous architectural comparison on a single functionalized-graphene dataset reported to date.

Representation comparison. A principled evaluation of four structural representations of the same physical system: a 3-feature atom-count vector, the full 8,528-element atom-type vector, a 2D image derived from the 82×104 honeycomb grid, and the native molecular graph with physically accurate honeycomb connectivity.

Data efficiency analysis. Learning-curve experiments with multiple random subsamplings per training-size point, providing quantitative evidence for the data requirements of each family and establishing where deep learning offers measurable advantages over simpler approaches.

All findings reported in this thesis are conditioned on the specific dataset scale (300 configurations), sheet geometry (8,528 atoms, zigzag loading), and functionalization chemistries (H and CH₃) studied. The dominance of the 3-feature KNN model, in particular, may reflect the smoothness of the property surfaces in this regime rather than a universal feature of functionalized graphene; whether this result generalizes to larger sheets, mixed chemistries, or higher coverages remains an open question.

1.5 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 reviews related literature on graphene functionalization, computational simulation methods, and existing ML approaches for graphene property prediction. Chapter 3 describes the dataset generation, simulation protocols, structural representations, model architectures, training procedures, and evaluation metrics. Chapter 4 presents and discusses experimental results for all model families. Chapter 5 summarizes key findings, states

limitations, and proposes directions for future work.

Chapter 2

Related Works

Functionalized graphene exhibits a combinatorial configuration space that is too vast for exhaustive simulation-based exploration: even a modest 8,528-atom sheet admits more distinct functionalization patterns than can be enumerated by classical molecular dynamics within any practical compute budget. This bottleneck has motivated the emergence of machine learning surrogate models that, once trained on a labeled subset of molecular-dynamics configurations, predict thermomechanical properties orders of magnitude faster than direct simulation. However, most prior work has relied on hand-crafted descriptors evaluated on small datasets and on a single model family, leaving open the central practical question this thesis addresses: *given a fixed MD budget, which combination of structural representation and ML architecture provides the most accurate and data-efficient property surrogate?*

To position this thesis within that landscape, the present chapter reviews three research threads: (i) an understanding of how chemical functionalization alters the lattice and its properties; (ii) simulation methodologies that generate reliable training data; and (iii) machine learning models and structural encodings capable of learning the resulting structure–property maps. Section 2.6 consolidates the four research gaps that emerge from this review and maps each gap to a corresponding contribution of this thesis.

2.1 Graphene Modification Strategies

2.1.1 Chemical Functionalization

Covalent attachment of functional groups converts selected carbon sites from sp^2 to sp^3 hybridization, locally disrupting the planar lattice and introducing both mass heterogeneity and bond-stiffness discontinuities. These local perturbations scatter acoustic phonons (reducing thermal conductivity) and create stress concentrations (reducing mechanical strength), with the magnitude of the effect depending on the type, density, and spatial arrangement of the attached groups [14, 12]. Two categories of covalent functionalization are most directly relevant to the present work: hydrogen and methyl.

For hydrogen functionalization, Li et al. [12] provided foundational understanding of how hydrogen coverage degrades graphene mechanical properties using AIREBO-based MD simulations. Subsequently, Li et al. [11] demonstrated that the spatial arrangement of hydrogen atoms—not only the overall coverage—also matters: strategic placement could increase ultimate strength by approximately 20% compared to random hydrogenation at 15% coverage through stress-concentration shielding. Zhao et al. [18] extended hydrogen-functionalization studies to graphene–copper nanocomposites and reported $R^2 = 0.98$ for the coefficient of thermal expansion using genetic programming coupled with AIREBO simulations.

For methyl functionalization, Pei et al. [14] reported elastic modulus, strength, and fracture strain reductions of 18%, 43%, and 47% respectively at saturation coverage—substantially larger reductions than for hydrogen, consistent with the greater steric bulk of CH_3 .

Among related oxygen-containing systems, Zheng and Gu [19] showed that gradient boosting can predict oxygen-to-carbon ratios in graphene oxide with $R^2 = 0.978$, with total potential energy emerging as the most predictive feature; their observation that random spatial distributions add substantial uncertainty to mechanical-response predictions foreshadows the spatial-distribution question central to the present thesis.

Most recently, Ashirmametov et al. [1] generated and released the dataset used

in this thesis: 600 H-functionalized configurations (1–15% coverage) and 600 CH₃-functionalized configurations (1–12% coverage) on an 8,528-atom sheet, simulated with AIREBO and non-equilibrium molecular dynamics (NEMD) in LAMMPS. Using classical ML on two encoding schemes (label encoding and bag-of-words), they reported $R^2 > 0.9$ for thermal conductivity, Young’s modulus, and maximum stress on a held-out test set, and demonstrated genetic-algorithm-based inverse design over the surrogate. The present thesis builds directly on that dataset and methodology, extending the comparison to four structural representations and five deep model families (Section 2.6).

Taken together, these studies establish two recurring themes: (i) global coverage level is a strong first-order predictor of property changes, but (ii) the spatial distribution of functional groups introduces a second-order effect whose magnitude depends on the property and the functional group—confirmed quantitatively in Ref. [11] (20% strength variation at fixed coverage) and Ref. [1] (Label vs Bag-of-Words encoding comparison). Disentangling these two effects systematically, across multiple representations and model families, is one of the motivations of the present work.

2.1.2 Substitutional Doping

Beyond covalent functionalization, substitutional doping provides an alternative route to property modulation. The most directly relevant precursor to this thesis is Hamed Mashhadzadeh et al. [7], who studied boron-nitrogen co-doped graphene over a 0–15% range using 300 randomly generated structures. Their Gaussian Process Regression model achieved $R^2 = 0.973$ for thermal conductivity prediction with a single-target Matérn 5/2 kernel, but multi-target models performed substantially worse than single-target models. The study also reported strong correlations between failure stress and thermal conductivity, suggesting that dopant-induced structural heterogeneities simultaneously influence phonon dynamics and mechanical stability. Nevertheless, a major research gap remains concerning the precise correlation between the spatial distribution of dopants and their resulting properties—a gap that motivates the cross-representation comparison in Chapter 4.

2.2 Computational Methodologies

The primary tool for generating labeled training data in the graphene-ML literature is classical molecular dynamics simulation, where the choice of interatomic potential determines both the fidelity and the computational cost of each data point. Functionalized systems have been broadly represented using the AIREBO potential (Adaptive Intermolecular Reactive Empirical Bond Order) because it correctly captures sp^2/sp^3 chemistry and the formation/breaking of C–C and C–H bonds. Li et al. [11] used AIREBO to investigate the tensile degradation behavior of hydrogenated graphene, and Lew et al. [10] used it for deep-learning-based graphene fracture studies, reaching 95% prediction accuracy on crack-propagation behavior. AIREBO is also the potential used in the source dataset of this thesis [1].

For doped systems, the Tersoff potential (especially the Lindsay–Broido optimized variant) is shown to be very accurate for both pristine and doped graphene; Hamed Mashhadzadeh et al. [7] validated its use on B/N co-doped graphene against the experimental Young’s modulus of pristine graphene (~ 1.02 TPa). Density functional theory provides higher-fidelity electronic-structure data than classical potentials but at substantially greater computational cost, limiting DFT-based datasets to tens or hundreds of configurations; for the dataset scales required for the deep learning models evaluated in this thesis (300–600 configurations per chemistry), classical MD with reactive bond-order potentials remains the standard choice [3].

2.3 Machine Learning Models for Property Prediction

2.3.1 Gaussian Process Regression

GPR has emerged as one of the strongest models for graphene property prediction in the low-data regime. For functionalized and doped graphene specifically, Hamed Mashhadzadeh et al. [7] achieved $R^2 = 0.973$ for thermal conductivity and $R^2 = 0.954$

for Young’s modulus on a 300-configuration B/N co-doped dataset using Matérn 5/2 kernels, and Ashirmametov et al. [1] confirmed competitive GPR performance on the H- and CH₃-functionalized datasets used in this thesis ($R^2 > 0.9$ on three of four targets). The probabilistic nature of GPR additionally provides calibrated uncertainty estimates, which are valuable for downstream tasks such as inverse design and active learning.

2.3.2 Gradient Boosting Methods

Gradient boosting methods have demonstrated strong performance in graphene compositional prediction. Zheng and Gu [19] showed that gradient boosting outperformed kernel ridge regression and multivariate least squares in predicting oxygen-to-carbon ratios in graphene oxide, achieving $R^2 = 0.978$ and $\text{MSE} = 0.000499$. The model captured non-linear relationships between potential-energy features and functionalization levels that simpler linear models could not represent. However, the present thesis observes (Section 4.1) that tree-based methods struggle on the raw 8,528-element feature vector of functionalized graphene without explicit feature engineering—a dimensionality regime not previously studied in this context.

2.3.3 Artificial Neural Networks

Deep learning architectures are increasingly applied to complex prediction tasks in materials science. Most directly relevant to the present work, Lew et al. [10] used a ConvLSTM-based deep learning model to predict fracture mechanics in graphene from MD trajectories, achieving 95% accuracy on crack propagation, branching patterns, and fractal dimensions. This is the strongest prior result on graphene fracture prediction and provides important context for the universal failure observed on the maximum-strain target across all 33 model variants evaluated in this thesis (Chapter 4): unlike Lew et al., who used full trajectory data as input, the present benchmark uses only static configurations and predicts a single scalar target, which appears to be insufficient for fracture-strain prediction.

2.3.4 Support Vector Methods

Support vector regression has shown consistently strong performance across several graphene property prediction tasks. Fernandez et al. [4] used SVM to predict accuracy differences between DFT and DFTB estimations of hydrogen-passivated graphene nanoflakes, obtaining $R^2 = 0.940$ on Fermi-level estimations. SVR’s ϵ -insensitive loss provides intrinsic robustness to outliers, useful in noisy MD-derived datasets. In this thesis, SVR with an RBF kernel is included as one of nine classical ML baselines (Section 4.1).

Across the studies reviewed above, several patterns emerge. Kernel-based methods (GPR, SVR) and gradient boosting consistently achieve the highest accuracy on small, well-featurized datasets ($R^2 > 0.95$), but their performance is sensitive to feature engineering and they scale poorly to high-dimensional raw representations. Deep learning architectures excel when the input is an image or a trajectory but typically require substantially more training data. Critically, no prior study compares more than two model families on a common functionalized-graphene dataset, making cross-study ranking unreliable—a gap that the present thesis directly addresses.

2.4 Encoding Strategies for Graphene Structures

2.4.1 Structural Representation Challenges

Encoding a graphene nanosheet structure for machine learning is non-trivial because of the large atom count (thousands per sheet) and the sensitivity of thermomechanical properties to spatial arrangement. Conventional molecular descriptors such as SMILES become unstable on large homogeneous structures and cannot natively represent 3D site information, motivating the use of vector, image, and graph-based encodings tailored to the lattice topology. The four representations adopted in this thesis (Section 3.4) span the full design space of dimensionality versus spatial fidelity.

2.4.2 Geometric and Topological Encodings

Fung et al. [5] introduced physically-informed graph neural networks for predicting electronic density of states, demonstrating that physics-aware constraints in GNN architectures improve prediction reliability. Kajita et al. [9] proposed 3D voxel descriptors of solid-state materials, providing a universal encoding scheme compatible with convolutional neural networks. These two studies motivate the GNN and 2D-image representations used in this thesis: the honeycomb molecular graph captures exact bonding topology with minimal node features (Section 3.5.6), while the 224×224 image representation enables transfer learning from pretrained vision models. Crucially, no prior study has compared graph, image, vector, and count representations on a single functionalized-graphene dataset under controlled experimental conditions.

2.5 Comprehensive Reviews

For broader surveys of machine learning in materials science and graphene specifically, the interested reader is referred to Schmidt et al. [15] and Huang et al. [8], which cover interpretability, cross-material transferability, and the bidirectional relationship between ML and graphene-based AI hardware. Shin et al. [16] additionally demonstrated the value of CNN + PCA pipelines for predicting mechanical behavior of nanocrystalline graphene, supporting the dimensionality-reduction strategies adopted in the present benchmark.

2.6 Research Gaps and Limitations

The existing literature exhibits four ongoing gaps that motivate the present work:

(1) Limited cross-family benchmarks. Existing studies typically evaluate one or two model families on a given graphene dataset. No prior work has compared classical ML, dense MLPs, scratch and pretrained CNNs, vision transformers, and graph neural networks on a shared functionalized-graphene dataset under identical experimental conditions. This makes cross-study ranking unreliable and obscures the

practical question of which family a practitioner should choose.

(2) Limited multi-target model performance. Hamed Mashhadzadeh et al. [7] observed that ANN and SVR models predicting multiple properties at once perform poorly relative to single-property models, indicating that capturing cross-property correlations remains challenging. This thesis treats each target with an independent regression head per model, consistent with their finding.

(3) Underexplored effects of spatial distribution. Functional-group or dopant spatial distributions play a role of large magnitude in mechanical properties [11, 14, 19], but systematic studies of distribution patterns are scarce. Ashirmametov et al. [1] compared two encoding schemes (label encoding vs. bag-of-words) and found broadly similar accuracy for κ , stress, and Young’s modulus on H-graphene, hinting at the redundancy of spatial information for these targets—a hypothesis this thesis tests systematically with four representations.

(4) Lack of standardized structural encoding. There is no consensus on the optimal representation for large functionalized-graphene systems. Studies have variously used geometric descriptors [9], physics-informed graphs [5], or chemistry-derived features [19], with no controlled comparison on a shared dataset.

Contribution of this thesis. This thesis directly addresses gaps (1)–(4) above. Chapters 3 and 4 systematically benchmark five model families—classical regressors, dense MLPs, pretrained CNNs, vision transformers, and graph neural networks—against four structural representations (3-feature count vector, full 8,528-element atom-type vector, 224×224 2D lattice image, and honeycomb molecular graph) on the two most commonly studied functionalization chemistries (H and CH_3). To the best of our knowledge, this constitutes the first cross-family, cross-representation comparison of its kind on functionalized graphene, and the first to report negative results on strain prediction across all 33 model variants tested. The benchmark builds directly on the dataset of Ashirmametov et al. [1] and extends their classical-ML inverse-design framework into a controlled architecture-vs-representation ablation that delivers a tiered design strategy (Chapter 5) for practitioners selecting models

under different data and chemistry constraints.

Chapter 3

Methodology

This chapter describes the methodology employed to build and evaluate machine learning surrogate models for the thermomechanical properties of chemically functionalized graphene. The pipeline comprises four stages: nanostructure generation and molecular dynamics simulation, dataset construction and structural representation, model design and training, and evaluation.

3.1 Nanostructure Generation

The molecular-dynamics configurations and simulations described in this section and in Section 3.2 were produced by collaborators as part of the dataset of Ashirmametov et al. [1]; this chapter documents the simulation protocol used to generate the data analyzed in this thesis.

A base graphene nanosheet containing 8,528 carbon atoms with dimensions of $220 \text{ \AA} \times 100 \text{ \AA}$ was constructed as the substrate for all functionalized configurations. These dimensions ensure bulk-like behavior under periodic boundary conditions, minimize artificial periodic interaction effects, and are consistent with established simulation protocols for thermomechanical property prediction [11, 14].

Two distinct functionalization chemistries were studied in this work:

Hydrogen functionalization. Hydrogen atoms were randomly attached to carbon sites on either the top or bottom face of the graphene sheet, with total coverage

ranging from 0 to 15%. A Python script was used to generate 300 unique valid configurations by sampling uniformly within this coverage range while ensuring minimum spacing constraints between adjacent functionalized sites to preserve structural stability.

Methyl functionalization. CH_3 groups were similarly attached to carbon sites at coverages targeted up to 15%; however, NEMD steady-state failed for 100 of the 300 configurations whose coverage exceeded 8%, leaving a usable subset of 200 configurations at 1–8% coverage (see Section 3.3.1 for details). The lower achievable coverage for methyl reflects the greater steric bulk of CH_3 relative to H; a minimum inter-site spacing of 3.5 Å was enforced to avoid steric overlap. A separate set of 300 unique valid configurations was generated using the same scripting approach, of which 200 remained usable after NaN-filtering.

The choice of 300 configurations per dataset reflects a balance between computational cost and statistical coverage. Each configuration requires two separate LAMMPS simulations—one mechanical (tensile loading to failure) and one thermal (NEMD steady-state)—with a combined wall-clock time of approximately 6–8 hours per configuration on the Shabyt HPC cluster, yielding a total simulation budget of $\sim 3,600$ CPU-hours per dataset. This scale is consistent with prior work on the same problem class: Hamed Mashhadzadeh et al. [7] used 300 randomly generated B/N co-doped graphene structures for their GPR-based property prediction study. While larger datasets would benefit data-hungry deep learning models, the learning-curve analysis in Chapter 4 explicitly quantifies the extent to which each model family has saturated at this scale.

The two datasets are treated independently throughout this work, as the different functional groups and their distinct bonding geometries require separate encoding schemes and produce different property ranges.

3.2 Molecular Dynamics Simulations

All MD simulations were performed using the open-source LAMMPS software package [10]. Simulations were executed on two compute clusters: the Shabyt HPC cluster (CPU-based) and a workstation equipped with three RTX 4090 GPUs.

3.2.1 Interatomic Potential

The adaptive intermolecular reactive empirical bond order (AIREBO) potential with a cutoff scaling factor of 2.5 was used for all functionalized systems. AIREBO is well-established for hydrogenated graphene systems and correctly describes C–C, C–H, and H–H interactions, including bond formation and breaking [12]. Periodic boundary conditions were applied along the in-plane directions.

3.2.2 Mechanical Properties

The mechanical simulation protocol for each configuration consisted of:

1. Energy minimization using the conjugate gradient algorithm until forces on all atoms fall below 10^{-4} eV/Å.
2. Relaxation of the simulation box for 3 ns under the isothermal–isobaric (NPT) ensemble at 300 K and zero pressure.
3. Application of uniaxial tensile loading along the zigzag direction at a strain rate of 10^{-3} ps⁻¹ until sample failure.

The per-atom stress was computed using the virial stress formula:

$$S_{ij} = \frac{1}{V} \left[mv_i v_j + \frac{1}{2} \sum_{n=1}^N \left(r_{1i}^{(n)} F_{1j}^{(n)} + r_{2i}^{(n)} F_{2j}^{(n)} \right) \right] \quad (3.1)$$

where i, j denote directional components, V and v correspond to sheet volume and velocity, and F denotes interatomic forces. From the stress–strain curve, Young’s modulus (slope of the linear elastic region), maximum stress (peak stress), and maximum strain (strain at peak stress) were extracted.

3.2.3 Thermal Conductivity

Thermal conductivity was evaluated using Non-Equilibrium Molecular Dynamics (NEMD):

1. Relaxation to a stress-free state at 300 K using the NPT ensemble for 3 ns.
2. Equilibration under the canonical (NVT) ensemble for 3 ns.
3. Establishment of a steady-state temperature gradient ($\Delta T = 40$ K) between hot and cold slabs for 6 ns.

The heat flux along the longitudinal direction is:

$$J_x = \frac{dE/dt}{A_c} \quad (3.2)$$

where A_c is the nanosheet cross-sectional area and E is the accumulated energy exchanged between hot and cold reservoirs. Thermal conductivity was then obtained from Fourier’s law:

$$K = -\frac{J_x}{\partial T/\partial x} \quad (3.3)$$

3.3 Datasets

3.3.1 Dataset Structure

The two functionalized graphene datasets analyzed throughout this thesis are drawn from the MD simulation pool of Ashirmametov et al. [1], with the raw simulation outputs and LAMMPS scripts openly available at the project repository [2]. Each dataset record consists of an 8,528-element integer atom-type vector (the input features) and four real-valued thermomechanical properties (the targets). The atom-type vector encodes the chemical state of each carbon site in the 82×104 honeycomb grid in the order defined by the atomic coordinate file.

Of the 300 CH_3 -functionalized configurations, 100 produced NaN values for thermal conductivity because the NEMD simulation did not reach a steady-state temperature gradient within the 6 ns run time. These failures occur exclusively at coverages

Table 3.1: Summary of the two functionalized graphene datasets.

Property	H-functionalized	CH ₃ -functionalized
Configurations generated	300	300
Usable samples (after NaN removal)	300	200
Feature vector length	8,528	8,528
Coverage range (usable)	1–15%	1–8%
Interatomic potential	AIREBO	AIREBO

above 8%, where the high density of sp^3 sites severely disrupts phonon transport and destabilizes the thermal gradient. To ensure a fair multi-target comparison—so that all four properties are evaluated on exactly the same set of configurations—these 100 samples were removed entirely from the methyl dataset, leaving 200 usable configurations at coverages of 1–8%. All methyl results reported in this thesis, including train/test splits and learning curves, use this reduced 200-sample dataset.

3.3.2 Atom-Type Encoding

The integer encoding scheme differs between the two datasets, reflecting the different bonding geometries:

H-functionalized:

- 1 = plain carbon atom (unfunctionalized)
- 2 = carbon with H attached above the graphene plane (top)
- 3 = carbon with H attached below the graphene plane (bottom)

CH₃-functionalized:

- 1 = plain carbon atom (unfunctionalized)
- 3 = carbon bonded to CH₃ above the graphene plane (top)
- 4 = carbon bonded to CH₃ below the graphene plane (bottom)

The methyl group atoms themselves are stripped from the atom-type vector; only the bonding carbon is retained and encoded, keeping the vector length fixed at 8,528 for both datasets.

3.3.3 Target Properties

All four targets are extracted from the same simulation trajectory:

1. Thermal conductivity, κ (W/m·K)
2. Young's modulus (GPa)
3. Maximum stress (GPa)
4. Maximum strain (dimensionless)

3.3.4 Exploratory Data Analysis

Figures 3-1 and 3-2 summarize the distributional characteristics of both datasets prior to modeling.

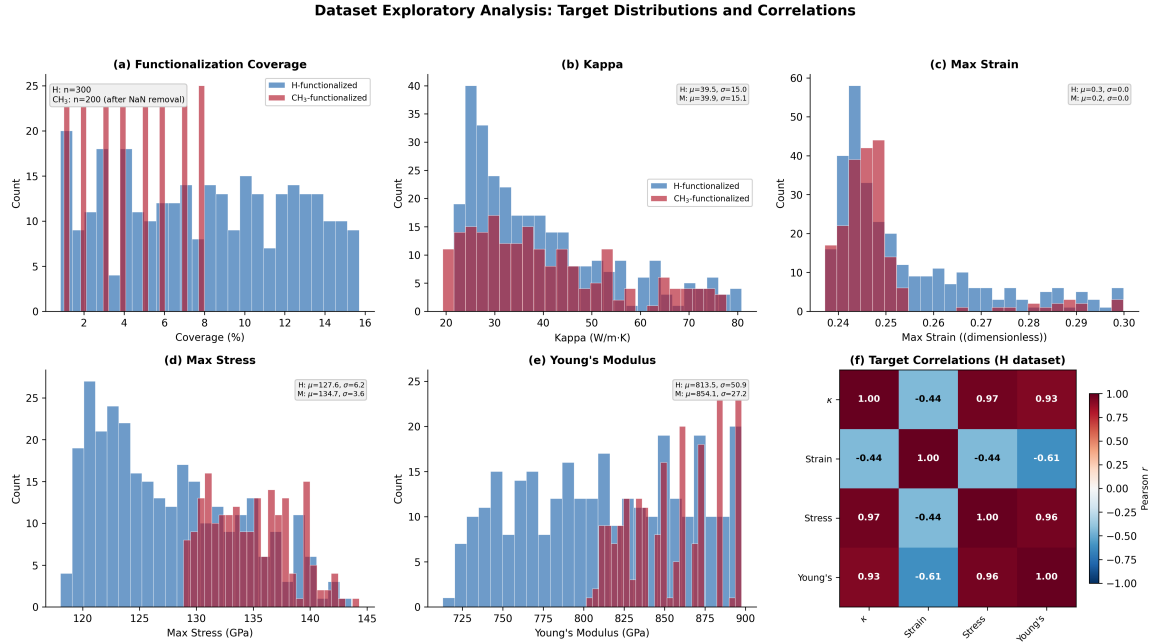


Figure 3-1: Exploratory analysis of the two datasets. (a) Functionalization coverage distributions: H-functionalized samples span 1–15% (n=300), while the usable CH₃ samples span 1–8% (n=200, after NaN removal). (b–e) Target property distributions for both datasets. (f) Pearson correlation matrix for the H-functionalized targets.

Panel (a) of Figure 3-1 shows that coverage is approximately uniformly distributed for hydrogen (1–15%, n=300). The methyl dataset, after removal of the 100 high-

coverage configurations with NaN thermal conductivity, spans 1–8% (n=200), resulting in a narrower and lower-coverage distribution.

Panels (b–e) reveal that κ , stress, and Young’s modulus exhibit broad, roughly symmetric distributions, whereas maximum strain is concentrated in a narrow band (mean ≈ 0.25 , std ≈ 0.015 for H), consistent with strain at fracture being insensitive to global coverage. The narrower methyl coverage range (1–8% vs. 1–15%) also means that methyl target distributions are compressed relative to hydrogen, which contributes to the greater difficulty of the methyl prediction task.

The correlation matrix (panel f) reveals that κ , stress, and Young’s modulus are strongly inter-correlated ($|r| > 0.92$), indicating that they are all governed by the same underlying physical variable—total sp^3 density. Strain shows moderate negative correlation with the other targets for hydrogen ($r \approx -0.44$ to -0.61) but near-zero correlation for methyl, confirming its fundamentally different physical origin.

Figure 3-2 directly visualizes the structure–property relationships that underpin the modeling results in Chapter 4. For κ , stress, and Young’s modulus, both datasets trace smooth, approximately monotonic curves as a function of coverage, explaining why a simple 3-feature KNN model can achieve near-perfect prediction: the mapping from coverage to property is effectively one-dimensional and smooth. The tighter scatter for Young’s modulus ($|r| \approx 1.00$) compared to κ ($|r| \approx 0.93$) is consistent with elastic stiffness being a more strictly additive property of sp^3 bond density than phonon-mediated thermal transport.

Maximum strain (panel b) is the clear outlier: for the hydrogen dataset there is a moderate negative trend ($r = -0.60$), but for methyl the correlation essentially vanishes ($r \approx 0$). This confirms that strain at fracture is not controlled by global functionalization density but by localized structural features—the weakest bond configuration along the loading path—that vary independently of total coverage. This observation provides the physical foundation for the universal failure of all model families on the strain target reported in Chapter 4.

Coverage vs. Target Properties

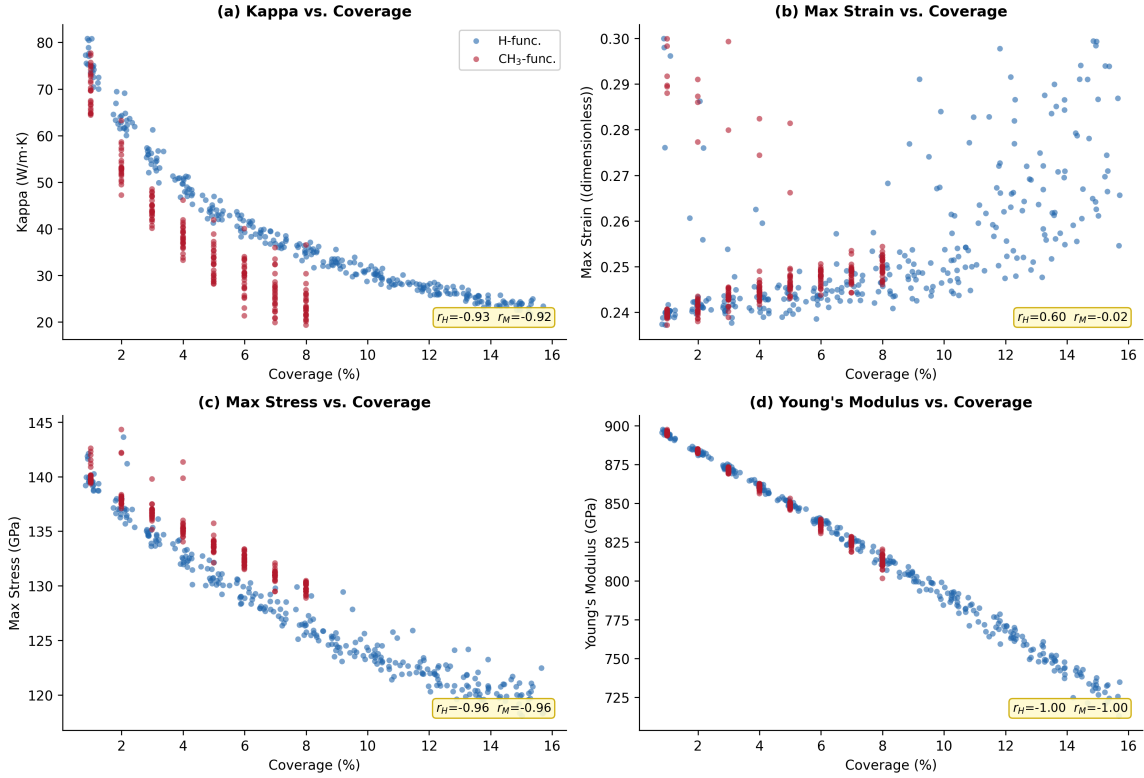


Figure 3-2: Coverage versus target properties for both datasets. Each point is one configuration (H: $n=300$, 1–15%; CH₃: $n=200$, 1–8%); Pearson r values are annotated. (a) κ decreases monotonically with coverage ($r_H = -0.93$, $r_M = -0.92$). (b) Maximum strain shows weak and inconsistent coverage dependence ($r_H = -0.60$, $r_M \approx 0$). (c, d) Stress and Young's modulus exhibit near-perfect negative correlation with coverage.

3.3.5 Lattice-to-Grid Mapping

The 8,528-element vector is laid out according to the physical honeycomb lattice. The mapping follows a zigzag strip pattern: the lattice is divided into 41 horizontal strips of two rows each, each strip containing 208 atoms (2 rows $\times$ 104 columns). Within each strip, even-numbered columns fill top-to-bottom and odd-numbered columns fill bottom-to-top, encoding the alternating sublattice structure of the honeycomb. The grid is assembled bottom-to-top so that row 0 of the grid corresponds to the physical top edge of the sheet.

This mapping is used consistently to construct both the 2D image representation

(Section 3.4) and the honeycomb graph (Section 3.5.6).

3.4 Structural Representations

The choice of structural representation is one of the most consequential design decisions in any materials ML study. A representation must be (i) expressive enough to encode the information relevant to the target property, (ii) compatible with the model family that consumes it, and (iii) computationally tractable to construct and store at the scale of the dataset. For functionalized graphene, the central tension is between representations that preserve full spatial information—at the cost of high dimensionality—and compact representations that discard spatial detail but are far easier for models to exploit with limited data. Four distinct representations of the same physical system were designed and compared in this work.

3-feature count representation. For the K-nearest neighbors model and as a simple baseline, a 3-element feature vector is constructed by counting the occurrences of each atom type in the full vector. For H-functionalized data, the features are (count of type-1, count of type-2, count of type-3); for methyl data they are (count of type-1, count of type-3, count of type-4). This representation discards all spatial information but captures global functionalization density—the total fraction of carbon sites carrying a functional group and the top/bottom asymmetry of the decoration. Its low dimensionality makes it trivially learnable even from very small datasets, and it serves as the minimal baseline against which all spatially-aware representations are compared. The key question it poses is whether global density alone is sufficient to predict thermomechanical properties, or whether spatial arrangement contributes additional predictive signal.

Full 8,528-element vector. The raw atom-type vector, standardized per-feature using a StandardScaler fitted on the training split, is used as input for linear models, tree-based models, and dense MLP networks. This representation preserves the identity of every atom site but encodes position only implicitly, through the fixed ordering of the vector elements. It is the natural input for models that process fixed-

length feature vectors, and its high dimensionality (8,528 features, approximately 240 training samples) makes it a challenging regime for methods that do not impose strong structural priors. Ridge regression and Gaussian process regression handle this regime through regularization and kernel-based smoothing, respectively, while tree-based methods are expected to struggle due to the curse of dimensionality.

2D image (224×224). The atom-type vector is reshaped into the 82×104 grid using the lattice mapping described above, then resized to 224×224 pixels via `scipy zoom`. For the H-functionalized dataset, bilinear interpolation (order=1) is used; for the methyl dataset, nearest-neighbor interpolation (order=0) is used to preserve the discrete structure of the coarser methyl encoding. The resulting grayscale image is min-max normalized using training-split statistics only, then broadcast to 3 channels to match the expected input format of pretrained ImageNet backbones. For ViT models, additional ImageNet channel-wise standardization is applied after min-max normalization. This representation enables the use of the large ecosystem of pretrained computer vision models. Its key advantage is that convolutional and attention operations naturally exploit spatial locality and translational structure—both present in the honeycomb lattice—without requiring explicit graph construction. Its limitation is that the 224×224 resize introduces interpolation artifacts and loses the precise integer identity of atom types that the raw vector preserves.

Honeycomb molecular graph. The 8,528 atoms are represented as nodes and the physical C–C bonds of the honeycomb lattice as undirected edges. Node features are a 3-dimensional vector per atom: normalized atom type (type / max_type), normalized row position (row / 103), and normalized column position (col / 81). The honeycomb connectivity is precomputed once for all samples, as the topology is identical across configurations; only node features differ. Edges encode: right horizontal bonds $(r, c) - (r, c + 1)$; even-column upward bonds $(r, c) - (r - 1, c)$; odd-column downward bonds $(r, c) - (r + 1, c)$. This representation is the most physically faithful of the four: it encodes the exact bonding topology of the honeycomb lattice without any interpolation or discretization artifacts, and it allows graph neural networks to perform message passing along physical bond paths. Its limitation is that graph convolution is

local by construction—each layer aggregates only immediate neighbors—so capturing long-range correlations requires either many layers or a global pooling mechanism. With only 3-dimensional node features and 3 message-passing layers, the GNN representation is also the most parsimonious, making it an important test of how much can be learned from minimal per-atom information.

Table 3.2 summarizes the four representations and their key trade-offs.

Table 3.2: Summary of the four structural representations.

Representation	Dim.	Spatial info	Model families	Key limitation
3-feature count	3	None	KNN	No spatial detail
Full vector	8,528	Implicit	Ridge, GPR, MLP, trees	Curse of dimensionality
2D image	224 ²	Explicit	CNN, ViT	Interpolation artifacts
Honeycomb graph	8,528 nodes	Exact	GNN	Local receptive field

3.5 Machine Learning Models

3.5.1 Classical ML and Simple Baselines

The following scikit-learn models were trained on the full 8,528-element scaled feature vector, except KNN which uses the 3-feature count representation:

- **KNN-3feat:** K-nearest neighbors regression ($k = 5$) on the 3-count feature vector. This model has no knowledge of spatial distribution—it operates purely on global functionalization counts.
- **Ridge:** Ridge regression with L2 penalty ($\alpha = 1.0$).
- **Elastic Net:** Elastic net with combined L1/L2 penalty ($\alpha = 10^{-3}$, ℓ_1 -ratio=0.5).
- **Random Forest:** 100-tree ensemble with bootstrap sampling.
- **Extra Trees:** 100-tree ensemble with random split thresholds.
- **Gradient Boosting:** Gradient boosted trees with 100 estimators.

Structural Representations of H-Functionalized Graphene (sample: 8.0% coverage, 682 functionalized sites)

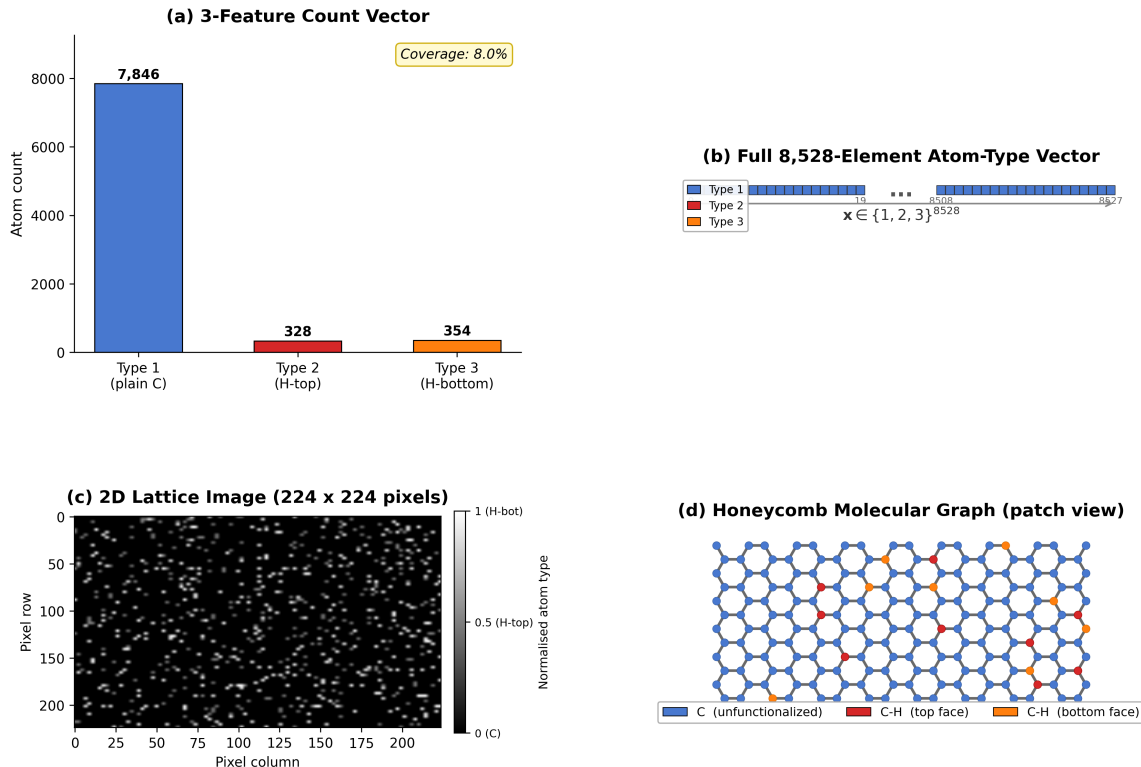


Figure 3-3: Visual overview of the four structural representations for the same H-functionalized graphene configuration ($\sim 8\%$ coverage). (a) 3-feature count vector retaining only global atom-type counts. (b) Full 8,528-element atom-type vector preserving per-site identity. (c) 2D lattice image after resizing to 224×224 pixels. (d) Honeycomb molecular graph with physical C–C bond connectivity (patch view).

- **XGBoost:** XGBoost regressor with 100 estimators.
- **SVR:** Support Vector Regression with RBF kernel ($C = 10$, $\epsilon = 0.1$) on PCA-reduced features (50 components).
- **GPR:** Gaussian Process Regression with Matérn $5/2$ + white noise kernel on PCA-reduced features (50 components). This model additionally provides calibrated uncertainty estimates.

All feature scalers (StandardScaler, PCA) and the target StandardScaler are fitted exclusively on the training fold and applied to validation and test folds without re-fitting, preventing any data leakage.

Hyperparameter selection. Hyperparameters for classical ML models were set to commonly used defaults rather than tuned via cross-validated search. Specifically, KNN uses $k=5$ (the scikit-learn default); ridge uses $\alpha=1.0$; random forest and extra trees use 100 estimators; SVR uses $C=10$ and $\epsilon=0.1$. This choice was deliberate: the goal of the benchmark is to compare model *families* and *representations* under controlled conditions, not to extract maximum performance from any single model. Per-model tuning would introduce a confound—differences in tuning effort—that would complicate cross-family interpretation. The PCA component count for SVR and GPR (50 components) was selected as a round number that retains >95% of training-set variance for both datasets. PCA was not applied to tree-based models (random forest, extra trees, GBM, XGBoost) because their axis-aligned splitting mechanism is designed to operate on raw features; applying PCA would rotate the feature space and remove the sparse structure that tree splits exploit. The poor tree-based results therefore reflect a genuine dimensionality mismatch rather than an unfair experimental setup.

3.5.2 Dense Multi-Layer Perceptrons

Five MLP architectures of increasing depth were designed to study the effect of network depth on regression accuracy when trained from scratch on the full scaled 8,528-element vector:

Table 3.3: MLP architectures used in this study.

Model	Hidden layer widths
MLP-1L	(256)
MLP-2L	(512, 256)
MLP-3L	(1024, 512, 256)
MLP-4L	(1024, 512, 256, 128)
MLP-5L	(2048, 1024, 512, 256, 128)

Each hidden layer is followed by Batch Normalization, ReLU activation, and Dropout (rate=0.3). The output is a single linear unit. All MLP models were trained with Adam optimizer ($\text{lr}=10^{-3}$), MSE loss, ReduceLROnPlateau scheduler

(factor=0.5, patience=5), and early stopping (patience=10).

3.5.3 Scratch 2D CNNs

Four custom CNN architectures were designed and trained from scratch on the $224 \times 224 \times 3$ graphene image, with no transfer of pretrained weights:

Table 3.4: Scratch CNN architectures. Spatial resolution after n pooling stages: $224/2^n$.

Model	Conv blocks	Base channels
CNN-S	2	32
CNN-M	3	32
CNN-L	4	32
CNN-XL	5	16 (narrower to fit GPU memory)

Each convolutional block consists of $\text{Conv2d}(3 \times 3, \text{padding}=1) \rightarrow \text{BatchNorm2d} \rightarrow \text{ReLU} \rightarrow \text{MaxPool2d}(2 \times 2)$. After the convolutional stack, a fully connected head of $\text{Linear}(512) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.5) \rightarrow \text{Linear}(256) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.3) \rightarrow \text{Linear}(1)$ is applied. Training used Adam ($\text{lr}=5 \times 10^{-4}$), MSE loss, and early stopping (patience=10).

3.5.4 Pretrained 2D CNNs

Five ImageNet-pretrained CNN architectures were fine-tuned on the graphene image dataset: ConvNeXt-Tiny, EfficientNetV2-S, ResNet50, MobileNetV2, and AlexNet. In all cases, the original classification head was replaced by a regression head of $\text{Linear}(d \rightarrow 256) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.3) \rightarrow \text{Linear}(256 \rightarrow 1)$, where d is the feature dimension of the backbone. All backbone parameters were updated during fine-tuning (full fine-tune). Training used Adam ($\text{lr}=10^{-3}$), MSE loss, ReduceLROnPlateau, and early stopping (patience=10), with a maximum of 50 epochs and batch size 32.

3.5.5 Vision Transformers

Four pretrained vision transformer architectures were evaluated: ViT-B/16, ViT-B/32, Swin-T, and Swin-S. In all cases, the backbone was fully frozen and only

the regression head was trained (head-only strategy, Strategy A). This decision was motivated by the limited training set size (approximately 240–250 samples), which is insufficient for safely fine-tuning the 86M backbone parameters without catastrophic forgetting.

For ViT-B/16 and ViT-B/32, the `heads.head` module was replaced with a regression head; for Swin-T and Swin-S, the `head` module was replaced. The [CLS] token output (768-dimensional) is passed to $\text{Linear}(768 \rightarrow 256) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.3) \rightarrow \text{Linear}(256 \rightarrow 1)$. Input images receive ImageNet channel-wise normalization (mean $[0.485, 0.456, 0.406]$, std $[0.229, 0.224, 0.225]$) after min-max scaling to $[0, 1]$. Training used Adam ($\text{lr}=10^{-3}$, applied only to the head), ReduceLROnPlateau, and early stopping (patience=10).

3.5.6 Graph Neural Networks

Four GNN architectures were implemented using PyTorch Geometric and trained on the honeycomb molecular graph representation:

- **GCN-3L:** Graph Convolutional Network [7] with 3 GCNConv layers, hidden dimension 64.
- **SAGE-3L:** GraphSAGE with 3 SAGEConv layers, hidden dimension 64, mean aggregation.
- **GIN-3L:** Graph Isomorphism Network with 3 GINConv layers (each using a 2-layer MLP aggregator), hidden dimension 64, trainable ϵ .
- **GAT-3L:** Graph Attention Network with 3 GATConv layers; intermediate layers use 4 attention heads with concatenation (output dim 128), final layer uses 1 head (output dim 64).

All GNN architectures use Batch Normalization after each message-passing layer, ReLU activation, and a shared regression head of $\text{Linear}(64 \rightarrow 256) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.3) \rightarrow \text{Linear}(256 \rightarrow 1)$. Global mean pooling aggregates node embeddings

to a single graph-level vector. Training used Adam ($\text{lr}=5 \times 10^{-4}$, weight decay= 10^{-4}), ReduceLROnPlateau, and early stopping (patience=7), with batch size 8 (each graph contains 8,528 nodes).

3.6 Training Protocol

3.6.1 Data Splitting

All models share the same train/validation/test split. The dataset is partitioned using random seed 42 into:

- Test set: 10% of all samples (held out for final evaluation; never used during training or hyperparameter selection).
- Validation set: approximately 10% of remaining samples (used for early stopping and learning rate scheduling).
- Training set: remaining approximately 80% of samples.

The exact split sizes depend on the dataset size after any NaN removal; for the hydrogen dataset (300 samples) the nominal split is approximately 240/30/30 (train/val/test), while for the methyl dataset (200 samples) the nominal split is approximately 160/20/20 (train/val/test)

3.6.2 Target Normalisation

For deep learning models, target values are normalized using a StandardScaler fitted exclusively on the training fold. For sklearn models, both features and targets are scaled per-fold (training fold only) to prevent data leakage. The same held-out test set targets are used throughout for fair cross-model comparison.

3.6.3 Multi-Seed Evaluation

All benchmark experiments (full training set) were run with three random seeds: 42, 101, and 1256. Mean and standard deviation of R^2 and RMSE over these three runs

are reported for each model and target property.

3.7 Data Efficiency Analysis

To assess how much training data each model family requires, learning-curve experiments were conducted. The test set is held fixed (10% of the full dataset, seed 42). The training set size is varied across seven fractions: 100%, 80%, 60%, 40%, 20%, 10%, and 5% of the available training data. At each fraction, three independent random subsamplings are drawn from the full training pool (using random seeds 0, 1, and 2), providing variance estimates that reflect the stochasticity of small-sample training. Feature scalers and target scalers are always fitted on the full training pool (not on the subsample), so that the only variable is the number of samples shown to the model.

Performance (R^2 and RMSE) at each training size is reported as mean $\pm$ standard deviation over the three subsamplings. Learning curves are plotted with the x-axis inverted (largest training set at the left, smallest at the right) to visualize the saturation behavior as data increases.

3.8 Evaluation Metrics

Model performance was assessed using three complementary metrics:

Coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.4)$$

Root mean square error:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3.5)$$

Mean absolute error:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3.6)$$

where $\hat{y}_i$ is the predicted value, y_i is the true value, $\bar{y}$ is the mean true value, and n is the number of test samples. All metrics are computed in the normalized (StandardScaler-transformed) target space to enable fair cross-model comparison; RMSE therefore represents normalized RMSE.

Chapter 4

Results and Discussion

This chapter presents and discusses experimental results for all model families evaluated on the H-functionalized and CH₃-functionalized (methyl) graphene datasets. Results are organized into five subsections corresponding to the five model families, followed by a cross-domain comparison and a data efficiency analysis. All metrics are computed in the normalized (StandardScaler-transformed) target space on the held-out test set. Performance is reported as mean $\pm$ standard deviation over three random seeds (42, 101, 1256) unless otherwise stated.

4.1 Classical ML and Scratch Deep Learning

4.1.1 Hydrogen Dataset

Table 4.1 summarizes the mean R^2 and RMSE for all classical ML and scratch deep learning models on the H-functionalized dataset.

KNN-3feat dominates three out of four targets. The most striking finding in Table 4.1 is that KNN-3feat, which operates solely on three atom-count features (counts of plain carbon, top-attached H, and bottom-attached H), achieves the highest mean R^2 for κ (0.986 ± 0.004), stress (0.978 ± 0.006), and Young’s modulus (0.993 ± 0.002). This result demonstrates that the global functionalization density, rather

Table 4.1: Mean R^2 ($\pm$ std, 3 seeds) on the H-functionalized test set. ML models use the full 8,528-element feature vector unless noted; KNN-3feat uses the 3-count representation; SVR and GPR use PCA-50 features. Scratch CNN and MLP models are trained from scratch.

Model	κ	Strain	Stress	Young’s mod.
KNN-3feat	0.986 ± 0.004	0.244 ± 0.410	0.978 ± 0.006	0.993 ± 0.002
Linear reg.	0.786 ± 0.100	0.243 ± 0.232	0.943 ± 0.012	0.987 ± 0.004
Ridge	0.786 ± 0.100	0.243 ± 0.232	0.943 ± 0.012	0.987 ± 0.004
Lasso	-0.006 ± 0.451	0.127 ± 0.333	0.306 ± 0.172	0.323 ± 0.084
Elastic net	0.093 ± 0.378	0.170 ± 0.321	0.379 ± 0.170	0.370 ± 0.064
SVR (RBF)	0.895 ± 0.010	-0.066 ± 0.291	0.937 ± 0.014	0.884 ± 0.017
GPR (Matérn)	0.849 ± 0.071	0.118 ± 0.225	0.845 ± 0.012	0.986 ± 0.004
Random forest	0.316 ± 0.158	0.253 ± 0.281	0.457 ± 0.116	0.509 ± 0.070
Extra trees	0.508 ± 0.140	0.125 ± 0.235	0.540 ± 0.126	0.470 ± 0.125
GBM	0.185 ± 0.075	0.131 ± 0.166	0.261 ± 0.185	0.399 ± 0.072
XGBoost	0.021 ± 0.368	-0.106 ± 0.225	0.186 ± 0.337	0.384 ± 0.080
MLP-1L	0.856 ± 0.068	0.313 ± 0.305	0.900 ± 0.060	0.947 ± 0.019
MLP-2L	0.915 ± 0.034	0.335 ± 0.208	0.914 ± 0.005	0.965 ± 0.032
MLP-3L	0.939 ± 0.030	0.297 ± 0.169	0.952 ± 0.017	0.958 ± 0.017
MLP-4L	0.929 ± 0.017	0.295 ± 0.224	0.933 ± 0.018	0.939 ± 0.041
MLP-5L	0.963 ± 0.015	0.383 ± 0.259	0.919 ± 0.033	0.950 ± 0.022
CNN-S (2 blk)	0.849 ± 0.017	0.039 ± 0.791	0.645 ± 0.482	0.610 ± 0.431
CNN-M (3 blk)	0.854 ± 0.034	0.246 ± 0.246	0.893 ± 0.051	0.902 ± 0.038
CNN-L (4 blk)	0.746 ± 0.067	0.347 ± 0.266	0.874 ± 0.009	0.902 ± 0.069
CNN-XL (5 blk)	0.958 ± 0.013	0.355 ± 0.266	0.874 ± 0.043	0.953 ± 0.013

than the precise spatial distribution of functional groups, is the primary determinant of these three thermomechanical properties in the H-functionalized dataset. The Young’s modulus result is particularly striking: a 3-parameter model captures 99.3% of the variance in a target that represents the fundamental elastic response of the lattice.

The physical reason for this dominance is that all three properties are governed by the aggregate density of sp^3 -hybridized sites rather than by their spatial arrangement. Thermal conductivity depends on the phonon mean free path, which is set by the total scattering cross-section—proportional to the number of sp^3 defects—rather than by their mutual distances at coverages below $\sim 15\%$. Young’s modulus and ultimate stress similarly reflect the fraction of weakened sp^3 bonds averaged over the entire loading cross-section. Because the three atom-count features (plain

C, top-H, bottom-H) are a sufficient statistic for the total sp^3 density on each face, KNN in this 3-D space is effectively performing nearest-neighbor interpolation on the one-dimensional curve that maps coverage to each property. The smoothness of these coverage–property curves—confirmed by the low seed-to-seed variance—explains why even $k = 5$ neighbours suffice for near-perfect prediction.

Tree-based methods underperform on the full feature vector. The anomalously poor performance of random forest, extra trees, gradient boosting, and XGBoost stands out, with mean R^2 values in the range 0.02–0.54 across targets. This behavior is explained by the dimensionality of the input: with 8,528 features and only approximately 240 training samples, tree-based methods face severe curse-of-dimensionality effects. Random splitting in decision trees is inefficient in this regime, and the ensemble cannot identify the few relevant features (essentially just the total counts) without explicit feature engineering. This contrasts sharply with ridge regression and GPR, which operate on the same feature vector but impose strong structure (linear model, kernel covariance) that aligns well with the underlying smooth property surfaces.

MLP depth provides consistent but modest gains. The MLP family shows a clear progression: deeper networks improve R^2 for κ from 0.856 (MLP-1L) to 0.963 (MLP-5L). This pattern is consistent with deeper networks gradually learning higher-order interactions between atom positions. However, none of the MLP variants match the trivial KNN-3feat on κ or Young’s modulus, suggesting that the 8,528-dimensional representation contains substantial redundancy relative to the 3-count summary. The largest gap is on stress, where MLP-3L (0.952) approaches but does not exceed KNN-3feat (0.978).

Strain is universally hard. No model achieves $R^2 > 0.40$ on maximum strain. The best result across the entire classical ML and scratch DL family is MLP-5L at 0.383 ± 0.259 , with a large standard deviation indicating high instability across seeds. This behavior is consistent with strain at fracture being governed by localized

rare events—the weakest bond under the applied stress field—rather than by global functionalization statistics. Two pieces of evidence support this interpretation. First, the seed-to-seed variance on strain is $2\text{--}5\times$ larger than on any other target for every model family, indicating that small changes in the train/test partition (and therefore in which rare local configurations appear in the test set) dominate the prediction error. Second, the models that perform best on strain (MLP-5L, ViT-B/32) are those with the largest effective receptive fields, suggesting that strain prediction benefits from capturing long-range spatial context—but the benefit saturates far below useful accuracy, because the critical information is sub-atomic-scale bond geometry that none of the tested representations encode.

4.1.2 Methyl Dataset

Table 4.2: Mean R^2 ($\pm$ std, 3 seeds) on the CH_3 -functionalized test set for selected models.

Model	κ	Strain	Stress	Young’s mod.
KNN-3feat	0.869 $\pm$ 0.051	-0.552 ± 0.168	0.856 ± 0.047	0.990 $\pm$ 0.003
Ridge	0.772 ± 0.080	-0.170 ± 0.089	0.880 $\pm$ 0.064	0.953 ± 0.007
GPR (Matérn)	-0.176 ± 0.440	-0.054 ± 0.049	0.874 ± 0.091	0.680 ± 0.115
MLP-5L	0.533 ± 0.253	-0.201 ± 0.121	0.595 ± 0.139	0.721 ± 0.017
CNN-XL (5 blk)	0.410 ± 0.676	-0.146 ± 0.121	0.836 ± 0.065	0.903 ± 0.034

Table 4.2 shows that performance drops substantially on the methyl dataset compared to hydrogen, particularly for MLP and scratch CNN models. KNN-3feat again leads on κ (0.869) and Young’s modulus (0.990), confirming that total doping density remains the dominant signal even for the more complex CH_3 chemistry. Ridge regression achieves the best stress result (0.880). Strain prediction fails for all models, with the best mean R^2 being -0.054 (GPR), confirming that this target remains intractable across both datasets and all simple representations. GPR itself suffers a dramatic collapse on κ (-0.176), indicating that the Matérn kernel struggles to generalize on the methyl feature space. The performance of MLP-5L drops relative to the hydrogen dataset (κ : 0.533 vs 0.963), and scratch CNN-XL shows extreme

instability (κ : 0.410 ± 0.676), suggesting that deeper models are more sensitive to the increased structural complexity of methyl functionalization.

Three factors contribute to the greater difficulty of the methyl dataset. First, the CH_3 group introduces substantially larger steric perturbations than a single H atom: each methyl group displaces the bonded carbon further out of the graphene plane and distorts a larger neighborhood of C–C bonds, creating a more heterogeneous local environment that is harder to summarize with global statistics. Second, the atom-type encoding used for methyl structures distinguishes only three site types (plain C, C bonded to CH_3 on top, C bonded to CH_3 on bottom), which does not capture the orientational disorder of the methyl group itself; in contrast, the hydrogen encoding is exact because H is a point-like adsorbate. Third, after removing configurations with NaN targets, the effective methyl training set is smaller than the hydrogen set, compounding the difficulty for data-hungry models such as MLP and scratch CNN.

4.2 Pretrained 2D CNNs

4.2.1 Hydrogen Dataset

Table 4.3: Mean R^2 ($\pm$ std, 3 seeds) for pretrained 2D CNNs on the H-functionalized dataset. VGG19 is excluded due to systematic failure ($R^2 \approx 0.2$ – 0.4 , consistent across seeds).

Model	κ	Strain	Stress	Young’s mod.
ConvNeXt-Tiny	0.981 ± 0.008	0.295 ± 0.215	0.975 ± 0.010	0.988 ± 0.005
EfficientNetV2-S	0.979 ± 0.013	0.035 ± 0.254	0.972 ± 0.011	0.983 ± 0.003
ResNet50	0.978 ± 0.016	0.286 ± 0.302	0.964 ± 0.006	0.982 ± 0.006
MobileNetV2	0.973 ± 0.013	0.268 ± 0.152	0.972 ± 0.008	0.989 ± 0.005
AlexNet	0.888 ± 0.043	0.074 ± 0.077	0.936 ± 0.017	0.911 ± 0.121

ConvNeXt-Tiny, EfficientNetV2-S, ResNet50, and MobileNetV2 all achieve $R^2 > 0.97$ on κ and stress, and $R^2 > 0.98$ on Young’s modulus, demonstrating that transfer learning from ImageNet provides strong feature extractors even for graphene lattice images. ConvNeXt-Tiny achieves the highest κ result (0.981 ± 0.008) among CNN models, closely matching KNN-3feat. The small gap between ConvNeXt-Tiny and

KNN-3feat on κ (0.981 vs 0.986) suggests that the spatial information encoded in the image representation provides marginal benefit for thermal conductivity prediction beyond what atom counts alone capture.

AlexNet is a consistent outlier with substantially lower R^2 on κ (0.888) and large variance on Young’s modulus (0.911 ± 0.121), consistent with its shallower architecture being less capable of extracting relevant spatial features at 224×224 resolution.

VGG19 is excluded from Table 4.3 because it consistently produces R^2 in the range 0.21–0.37 across all targets and seeds, worse than a mean predictor for several targets. This is attributed to the interaction between VGG19’s architectural depth (19 layers, large fully-connected blocks), batch normalization behavior, and the small training set size; the model fails to converge reliably within the 50-epoch budget.

Strain prediction improves noticeably over the classical ML and scratch deep learning baselines, with ConvNeXt-Tiny reaching 0.295 ± 0.215 and ResNet50 reaching 0.286 ± 0.302 , though the large standard deviations indicate persistent instability.

4.2.2 Methyl Dataset

Table 4.4: Mean R^2 ($\pm$ std, 3 seeds) for pretrained 2D CNNs on the CH₃-functionalized dataset.

Model	κ	Strain	Stress	Young’s mod.
ConvNeXt-Tiny	0.587 ± 0.547	-0.051 ± 0.056	0.572 ± 0.500	0.986 ± 0.000
EfficientNetV2-S	0.875 ± 0.039	-0.641 ± 1.023	0.780 ± 0.031	0.981 ± 0.009
ResNet50	0.861 ± 0.021	-0.058 ± 0.082	0.829 ± 0.022	0.979 ± 0.009
MobileNetV2	0.847 ± 0.020	-0.452 ± 0.239	0.884 ± 0.044	0.985 ± 0.003
AlexNet	0.739 ± 0.131	-0.059 ± 0.030	0.792 ± 0.100	0.952 ± 0.018

Performance on the methyl dataset degrades markedly for ConvNeXt-Tiny and EfficientNetV2-S, which show extremely high variance (std = 0.547 and 1.023 respectively) on κ and strain, indicating that these models are highly unstable across seeds for methyl functionalization. EfficientNetV2-S achieves the best mean κ (0.875) but its high variance makes this unreliable. ResNet50 provides the most consistent performance (0.861 ± 0.021 for κ), suggesting that its residual connections confer better

optimization stability on the harder methyl dataset. Young’s modulus remains well-predicted by all models ($R^2 > 0.95$), confirming the robustness of this target across both chemistries and model families.

4.3 Vision Transformers

4.3.1 Hydrogen Dataset

Table 4.5: Mean R^2 ($\pm$ std, 3 seeds) for ViT/Swin models (head-only fine-tuning) on the H-functionalized dataset.

Model	κ	Strain	Stress	Young’s mod.
ViT-B/16	0.979 $\pm$ 0.007	0.340 $\pm$ 0.259	0.966 $\pm$ 0.010	0.983 $\pm$ 0.006
ViT-B/32	0.976 $\pm$ 0.006	0.363 $\pm$ 0.244	0.962 $\pm$ 0.010	0.976 $\pm$ 0.005
Swin-S	0.968 $\pm$ 0.022	0.330 $\pm$ 0.381	0.961 $\pm$ 0.005	0.962 $\pm$ 0.014
Swin-T	0.971 $\pm$ 0.015	0.299 $\pm$ 0.296	0.966 $\pm$ 0.006	0.976 $\pm$ 0.002

ViT-B/16 achieves the highest κ R^2 among ViT models at 0.979 ± 0.007 , approaching KNN-3feat (0.986 ± 0.004). Notably, ViT-B/32 achieves the best strain R^2 across all model families at 0.363 ± 0.244 , suggesting that the global self-attention mechanism captures some spatial patterns relevant to fracture behavior that elude local convolutional filters. However, the large standard deviation reflects the fundamental difficulty of this target rather than a reliable improvement.

The comparison between ViT-B/16 (196 patches of 16×16 pixels, each covering approximately 4 carbon atoms) and ViT-B/32 (49 patches of 32×32 pixels, approximately 15 atoms per patch) shows ViT-B/16 marginally outperforming ViT-B/32 on κ (0.979 vs 0.976), while ViT-B/32 achieves the best strain result (0.363 vs 0.340). This suggests that the coarser patch resolution may better capture the longer-range spatial correlations relevant to fracture initiation.

Swin-T and Swin-S achieve results competitive with ViT models, confirming that hierarchical shifted-window attention transfers well to graphene lattice images. Swin-T matches ViT-B/16 on stress (0.966) and shows low variance, consistent with the hierarchical design providing stable training dynamics.

4.3.2 Methyl Dataset

Table 4.6: Mean R^2 ($\pm$ std, 3 seeds) for ViT/Swin models on the CH₃-functionalized dataset.

Model	κ	Strain	Stress	Young’s mod.
ViT-B/16	0.873 ± 0.024	-0.132 ± 0.080	0.846 ± 0.044	0.981 ± 0.007
ViT-B/32	0.852 ± 0.027	-0.129 ± 0.035	0.867 ± 0.045	0.974 ± 0.011
Swin-S	0.877 ± 0.031	-0.055 ± 0.050	0.851 ± 0.053	0.981 ± 0.012
Swin-T	0.886 ± 0.028	-0.123 ± 0.081	0.841 ± 0.053	0.978 ± 0.008

ViT and Swin models show more uniform performance degradation on the methyl dataset compared to the CNN models, with variance staying moderate (std < 0.06 for most). Swin-T achieves the best κ (0.886 ± 0.028), and ViT-B/16 achieves the best Young’s modulus (0.981 ± 0.007). ViT-B/32 achieves the best stress (0.867 ± 0.045). All models produce negative R^2 for strain, confirming the universal failure on this target. The head-only training strategy, which freezes the backbone, appears to provide regularization that limits overfitting on the smaller and harder methyl dataset, as evidenced by the lower variance compared to CNN models that fully fine-tune all parameters.

4.4 Graph Neural Networks

4.4.1 Hydrogen Dataset

Table 4.7: Mean R^2 ($\pm$ std, 3 seeds) for GNN models on the H-functionalized dataset.

Model	κ	Strain	Stress	Young’s mod.
GCN-3L	0.960 ± 0.020	0.164 ± 0.294	0.975 ± 0.014	0.987 ± 0.004
SAGE-3L	0.980 ± 0.013	0.318 ± 0.322	0.973 ± 0.007	0.988 ± 0.006
GIN-3L	0.943 ± 0.039	0.203 ± 0.291	0.967 ± 0.005	0.986 ± 0.011
GAT-3L	0.971 ± 0.016	0.184 ± 0.281	0.971 ± 0.009	0.987 ± 0.007

All four GNN architectures achieve $R^2 > 0.94$ on κ , stress, and Young’s modulus, confirming that the native honeycomb graph representation with the zigzag-corrected

node features provides an effective encoding of the functionalized graphene structure. SAGE-3L achieves the highest κ (0.980 ± 0.013) and Young’s modulus (0.988 ± 0.006), while GCN-3L achieves the highest stress (0.975 ± 0.014). Strain prediction remains poor for all GNN models (best: SAGE-3L, 0.318 ± 0.322), consistent with results from other model families.

The GNN results for κ (0.943–0.980) are competitive with pretrained CNNs (0.888–0.981) despite using only 3-dimensional node features and operating on the exact physical bonding topology. This is notable because GNNs were trained entirely from scratch with no pretrained initialization, and yet match architectures leveraging ImageNet representations. The message-passing mechanism appears to aggregate the local functionalization context around each atom effectively, without requiring the global spatial layout that image representations provide.

4.4.2 Methyl Dataset

Table 4.8: Mean R^2 ($\pm$ std, 3 seeds) for GNN models on the CH₃-functionalized dataset.

Model	κ	Strain	Stress	Young’s mod.
GCN-3L	0.879 ± 0.042	-0.063 ± 0.071	0.847 ± 0.059	0.991 ± 0.002
SAGE-3L	0.864 ± 0.066	-0.070 ± 0.066	0.838 ± 0.059	0.986 ± 0.007
GIN-3L	0.883 ± 0.030	-0.063 ± 0.066	0.832 ± 0.057	0.986 ± 0.009
GAT-3L	0.839 ± 0.057	-0.061 ± 0.056	0.828 ± 0.037	0.987 ± 0.006

On the methyl dataset, GNN models show consistent κ performance (0.839–0.883) with comparatively low variance, making them among the most stable model family on this harder dataset. GIN-3L achieves the highest κ (0.883) and GCN-3L achieves the highest stress (0.847) and Young’s modulus (0.991). The relative ranking of GNN architectures changes between datasets: SAGE-3L, which performed best on hydrogen κ (0.980), drops to 0.864 on methyl, while GIN-3L (sum aggregation) shows the strongest methyl κ performance. Strain prediction remains universally negative for all models.

4.5 Cross-Domain Comparison

Table 4.9 summarizes the best result per model family for each target on the hydrogen dataset. Because each cell reports the highest-scoring model within its family for that specific target, the table presents an optimistic view of each family’s capability; a practitioner selecting a single model *before* seeing test results would not achieve all bolded values simultaneously.

Table 4.9: Best R^2 (mean $\pm$ std) per model family and target, H-functionalized dataset.

Family (best model)	κ	Strain	Stress	Young’s mod.
ML — KNN-3feat	0.986 $\pm$ 0.004	0.244 $\pm$ 0.410	0.978 $\pm$ 0.006	0.993 $\pm$ 0.002
Scratch CNN-XL	0.958 $\pm$ 0.013	0.355 $\pm$ 0.266	0.874 $\pm$ 0.043	0.953 $\pm$ 0.013
Pretrained CNN	0.981 $\pm$ 0.008	0.295 $\pm$ 0.215	0.975 $\pm$ 0.010	0.989 $\pm$ 0.005
ViT	0.979 $\pm$ 0.007	0.363 $\pm$ 0.244	0.966 $\pm$ 0.010	0.983 $\pm$ 0.006
GNN — SAGE-3L	0.980 $\pm$ 0.013	0.318 $\pm$ 0.322	0.975 $\pm$ 0.014	0.988 $\pm$ 0.006

The three most important findings from the cross-domain comparison are as follows.

First, KNN-3feat with only three input dimensions matches or exceeds all deep learning models on κ , stress, and Young’s modulus. This strongly indicates that total functionalization density, rather than spatial arrangement, is the primary structure-property driver for these targets in the H-functionalized system. Deep learning models offer no meaningful advantage on these targets for the hydrogen dataset.

Second, ViT-B/32 achieves the best strain result (0.363), modestly outperforming the best scratch deep learning model (CNN-XL, 0.355) and GNNs (SAGE-3L, 0.318). The global self-attention mechanism appears better suited to capturing the distributed spatial correlations relevant to fracture initiation than local-to-global aggregation (CNNs, GNNs). However, all models remain far from practically useful predictions for strain.

Third, GNN models achieve competitive results on κ and stress despite being trained entirely from scratch and using only 3-dimensional node features. SAGE-3L reaches 0.980 on κ , closely approaching pretrained CNNs (0.981). Compared to

pretrained CNNs with millions of ImageNet-pretrained parameters, the native graph representation provides an efficient encoding that approaches similar accuracy levels without transfer learning.

4.6 Data Efficiency Analysis

Learning curves were computed by varying the training set size across seven fractions of the full training pool (5%–100%), with three random subsamplings per fraction to estimate variance. The held-out test set (10% of the full dataset) was fixed throughout. Training sizes in absolute terms range from 12 to 240 samples (hydrogen) and from 8 to 160 samples (methyl).

4.6.1 Classical ML and Scratch Deep Learning

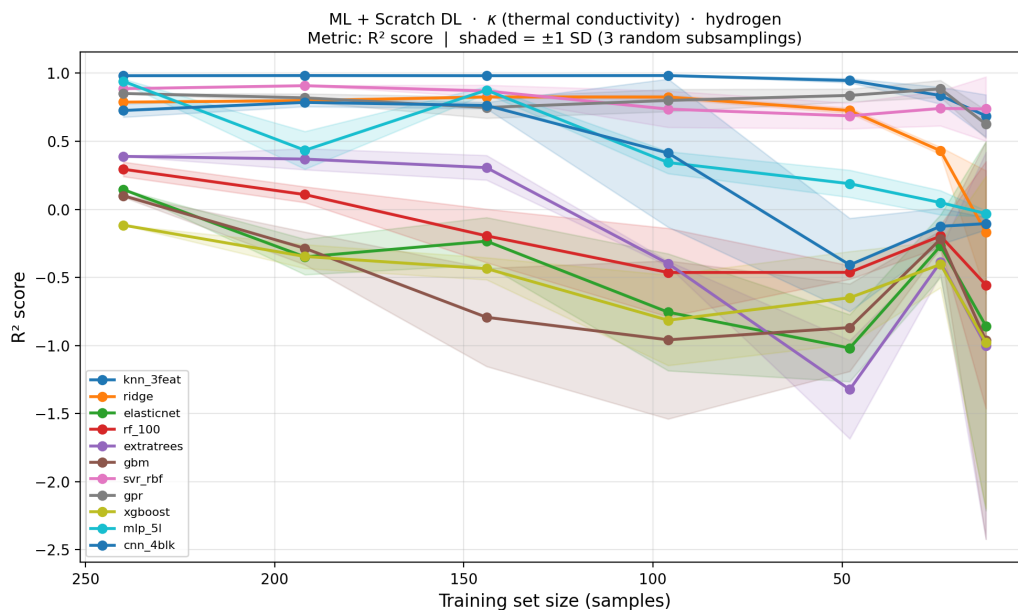


Figure 4-1: Learning curves for the ML and scratch deep learning family on the H-functionalized dataset (κ target). Shaded bands indicate ± 1 SD over three random subsamplings.

KNN-3feat reaches 90% of its full-data R^2 (0.982) at just 48 training samples, demonstrating exceptional data efficiency. This saturation occurs because the three-

dimensional input space is well-covered by ~ 48 points once the doping-density-to-property mapping has been sampled across the functionalization range. In contrast, MLP-5L requires approximately 144 samples to reach 90% of its full-data performance on κ .

Tree-based models (random forest, extra trees, XGBoost, GBM) fail to improve substantially with more training data across all training set sizes, confirming that their poor performance is structural (dimensionality mismatch) rather than a data-quantity issue. These models would likely perform competitively with dimensionality reduction applied before training, but that was not investigated here.

On the methyl dataset, KNN-3feat again saturates early on κ , reaching $R^2 \approx 0.81$ at 16 samples and showing no meaningful improvement with more data up to 160 samples. MLP-5L struggles throughout, reaching only 0.32 at 160 samples, confirming that the MLP cannot extract useful features from the raw methyl feature vector with this dataset size.

4.6.2 Pretrained 2D CNNs

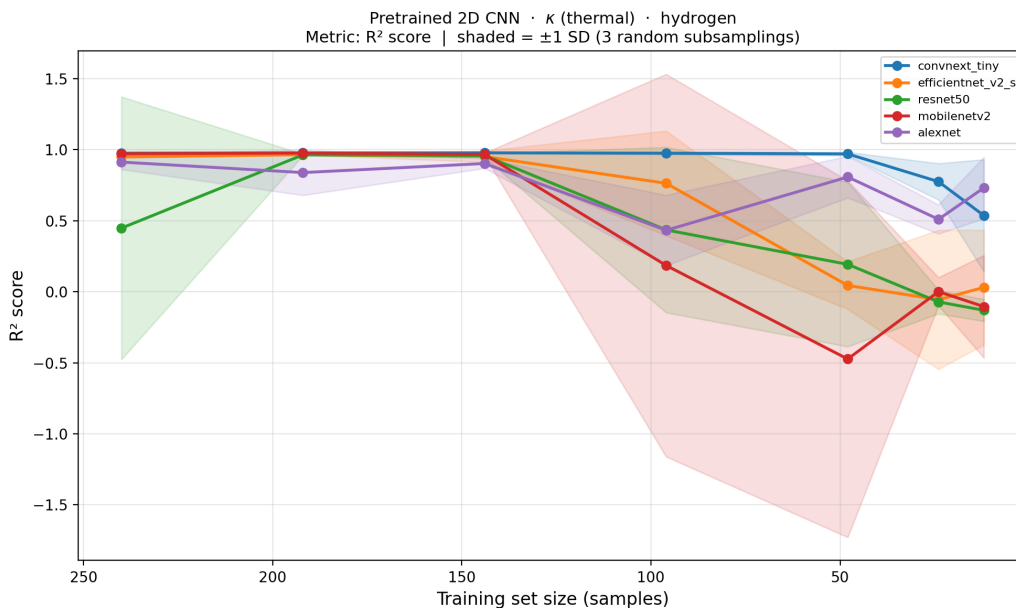


Figure 4-2: Learning curves for pretrained 2D CNN models on the H-functionalized dataset (κ target). Shaded bands indicate ± 1 SD over three random subsamplings.

ConvNeXt-Tiny shows rapid improvement from $R^2 = 0.417$ at 12 samples to $R^2 = 0.960$ at 48 samples for κ on the hydrogen dataset, then plateaus (Figure 4-2). This fast saturation suggests that ImageNet pretrained features transfer effectively to graphene lattice images, requiring only a small number of fine-tuning examples to adapt the regression head. AlexNet shows erratic behavior with high variance across seeds and non-monotonic mean R^2 at larger training sizes (e.g., mean R^2 drops from 0.938 at 96 samples to 0.614 at 192 samples due to occasional seed failures), consistent with its shallower architecture failing to provide stable feature extraction.

MobileNetV2 displays an interesting behavior on Young’s modulus: it achieves $R^2 = 0.960$ at 96 samples after near-zero performance at smaller sizes, suggesting a threshold effect where the model requires sufficient coverage of the property range before the regression head can be correctly fitted.

4.6.3 Vision Transformers

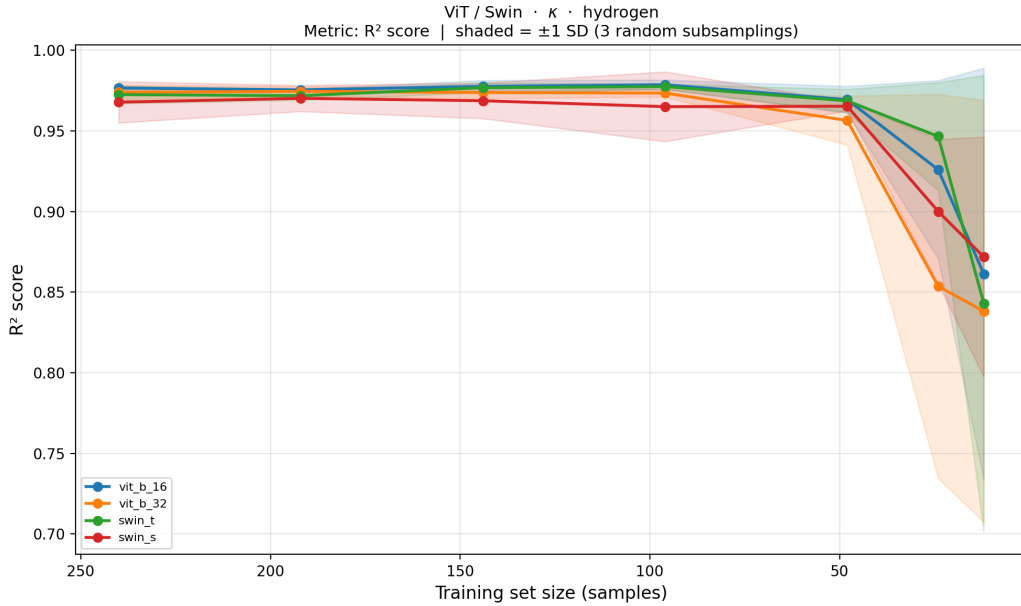


Figure 4-3: Learning curves for ViT/Swin models on the H-functionalized dataset (κ target).

ViT-B/16 and Swin-T display a smooth monotonic improvement from 12 to ~ 96 samples, after which performance stabilizes for κ on the hydrogen dataset ($R^2 = 0.974$

at 96 vs 0.973 at 240). The frozen backbone strategy provides inherent stability: because backbone parameters are fixed, the only trainable degrees of freedom are the $\sim 200\text{K}$ head parameters, which converge quickly. This contrasts with the fine-tuned CNN models, which show more volatile behavior at small training sizes.

Remarkably, ViT-B/16 achieves mean $R^2 = 0.861 \pm 0.128$ on κ with only 12 training samples (individual subsample R^2 values: 0.924, 0.714, 0.946 across the three random subsamplings), outperforming most other model families at this extreme data-scarce regime. This data efficiency stems from the frozen ImageNet backbone providing rich general-purpose features that are immediately useful without any adaptation; the high variance reflects the genuine fragility of 12-sample training rather than instability of the model.

For strain on the hydrogen dataset, ViT models show the most notable improvement with data: Swin-S increases from 0.324 at 12 samples to 0.451 at 240 samples, and ViT-B/16 from 0.399 to 0.487. While still low in absolute terms, this positive trend with training size—absent in most other families—suggests that the global attention mechanism is beginning to capture spatial information relevant to fracture initiation, but requires more data than is available here to fully exploit it.

4.6.4 Graph Neural Networks

All four GNN architectures now have learning curves computed with the corrected zigzag-strip node features (Figure 4-4). SAGE-3L on hydrogen κ shows the strongest data-efficiency profile among GNN models, reaching $R^2 \approx 0.88$ at 48 samples and plateauing near 0.96 at full data, consistent with its leading full-data performance (0.980). All GNN models show rapid improvement from 12 to ~ 48 samples, indicating that the message-passing mechanism quickly identifies the most informative structural patterns once a minimal number of positive examples are available. The plateau below the best CNN and ViT results (~ 0.975 – 0.983) at full data suggests that the 3-dimensional node features and 3-layer depth may be a bottleneck; deeper GNNs or richer edge features may close this gap.

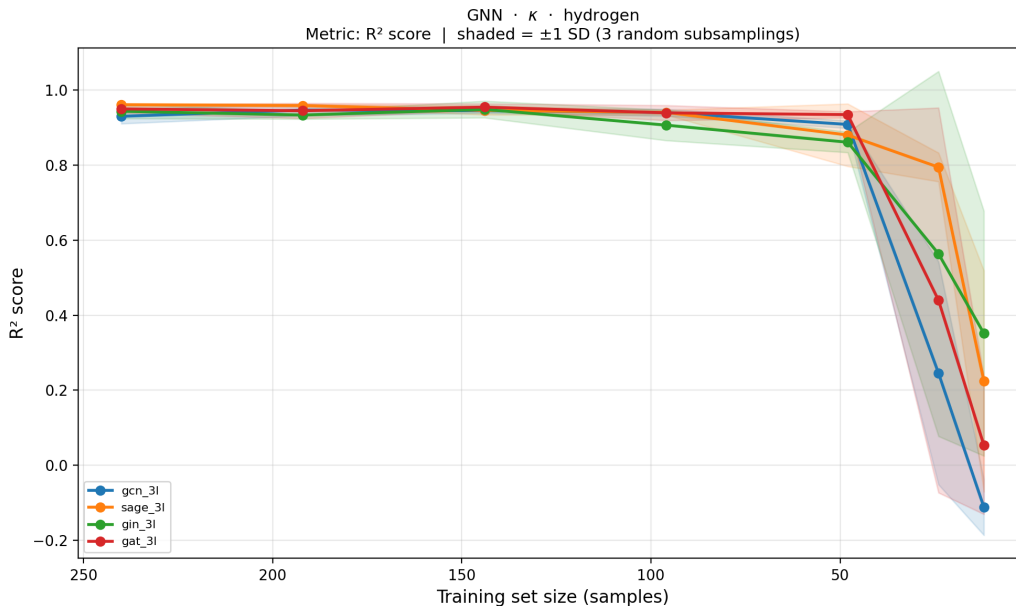


Figure 4-4: Learning curves for GNN models on the H-functionalized dataset (κ target). Shaded bands indicate ± 1 SD over three random subsamplings.

4.6.5 Cross-Domain Data Efficiency Summary

The most data-efficient model overall is KNN-3feat, which saturates at 48 samples (20% of training data) and does not benefit from additional data. ViT models trained with frozen backbones are the second most data-efficient, providing competitive R^2 (> 0.86) at just 12 samples. Pretrained CNNs (specifically ConvNeXt-Tiny) provide the sharpest improvement curve between 12 and 48 samples, achieving near-full-data performance by 20% of the training set. Scratch deep learning models (MLP, scratch CNN) require the most data and show the slowest saturation, consistent with the need to learn all feature extraction from scratch on a limited dataset.

These findings have a practical implication: if the goal is to build an accurate surrogate for κ and Young’s modulus of H-functionalized graphene with minimal MD simulation cost, KNN-3feat trained on as few as 50 configurations provides performance essentially matching that of much more expensive deep learning pipelines.

4.7 Summary of Key Findings

Table 4.10 consolidates the best-performing model per target and dataset across all model families.

Table 4.10: Best R^2 achieved across all model families, per target and dataset. H = H-functionalized, M = CH_3 -functionalized.

Target	Dataset	Best model	R^2	Family	Note
κ	H	KNN-3feat	0.986 ± 0.004	ML	3 features only
κ	M	Swin-T	0.886 ± 0.028	ViT	
Strain	H	ViT-B/32	0.363 ± 0.244	ViT	Universally hard
Strain	M	ConvNeXt-Tiny	-0.051 ± 0.056	CNN	All models fail
Stress	H	KNN-3feat	0.978 ± 0.006	ML	3 features only
Stress	M	MobileNetV2	0.884 ± 0.044	CNN	
Young’s	H	KNN-3feat	0.993 ± 0.002	ML	3 features only
Young’s	M	GCN-3L	0.991 ± 0.002	GNN	

Three conclusions emerge from this study. First, global functionalization density (captured by the 3-count representation) is the dominant driver of κ , stress, and Young’s modulus in H-functionalized graphene; deep learning provides no practical advantage over a 3-parameter KNN model for these targets. On the methyl dataset, however, GCN-3L achieves the highest Young’s modulus (0.991), marginally surpassing KNN-3feat (0.990), showing that graph-based representations can match or exceed count-based features for elastic properties. Second, maximum strain remains an intractable prediction target across all model families and both datasets, with no model exceeding $R^2 \approx 0.38$, suggesting that fracture initiation is governed by local structural configurations not captured by any tested representation. Third, the methyl dataset is substantially harder than the hydrogen dataset for all targets except Young’s modulus, and model stability degrades significantly for CNN models on methyl; GNN and ViT models are more robust to this dataset shift.

Chapter 5

Conclusion

5.1 Summary

This thesis set out to answer a deceptively simple question: for chemically functionalized graphene, which machine learning architecture and structural representation should a practitioner choose? The answer that emerged from the benchmark is both surprising and practically useful. Across 33 model variants spanning five architectural families—classical regressors, dense MLPs, pretrained CNNs, vision transformers, and graph neural networks—evaluated on two datasets of 300 configurations each, a three-parameter K-nearest-neighbors model matched or exceeded the predictive accuracy of far more complex deep learning architectures for three of the four thermomechanical targets studied.

This result does not reflect a failure of deep learning. Rather, it reflects a fundamental property of the H-functionalized graphene system itself: that total functionalization density, not the precise spatial arrangement of functional groups, is the dominant determinant of thermal conductivity, stiffness, and strength. Recognizing this structure of the problem is arguably the most practically useful outcome of the benchmark. It means that for the most common engineering targets, the expensive machinery of transfer learning, graph convolution, and attention mechanisms can be bypassed entirely—at least until the questions being asked demand spatial resolution that global statistics cannot provide.

The methyl dataset complicates this picture in an instructive way. The same three-feature model that dominates on hydrogen retains strong performance on Young’s modulus for methyl functionalization, but struggles for thermal conductivity and stress, where GNN and ViT models prove more robust. This asymmetry suggests that the sufficiency of global density as a predictor is not a universal feature of functionalized graphene but a specific property of hydrogen functionalization, where the small size and symmetric bonding geometry of H atoms make property surfaces particularly smooth over configuration space. The richer steric and bonding complexity of CH₃ groups begins to introduce spatial structure that global statistics miss, hinting at the conditions under which more expressive representations become necessary.

5.2 Key Findings

The experimental results yield four principal findings:

Global density dominates structure–property relationships in H-functionalized graphene. A 3-feature KNN model operating solely on atom-type counts achieves $R^2 \geq 0.986$ for κ , ≥ 0.978 for maximum stress, and ≥ 0.993 for Young’s modulus on the hydrogen dataset. No deep learning model—regardless of architectural complexity or pretraining—provides a meaningful improvement over this three-parameter baseline for these targets. This finding implies that the dominant structure–property signal for thermomechanical properties of H-functionalized graphene is total functionalization density rather than the precise spatial arrangement of functional groups.

Maximum strain is universally intractable. No tested model family achieves $R^2 > 0.38$ on maximum strain across either dataset. The large seed-to-seed variance observed for the best-performing models (ViT-B/32: 0.363 ± 0.244) confirms that this instability is structural rather than incidental. This behavior is consistent with fracture strain being governed by rare, spatially localized failure events—the weakest bond under the applied stress field—whose occurrence cannot be reliably inferred from global or semi-local structural representations of the sizes tested here.

Vision transformers are the most data-efficient deep learning family. ViT-B/16 with a frozen ImageNet backbone achieves mean $R^2 = 0.861 \pm 0.128$ on κ with only 12 training samples, outperforming all other model families in the extreme low-data regime. The frozen backbone strategy additionally provides regularization that limits overfitting on the harder methyl dataset, as evidenced by the markedly lower variance compared to fully fine-tuned CNN models. These results suggest that pretrained visual representations transfer non-trivially to 2-D material lattice images, providing a practical starting point when simulation budgets are severely constrained.

Graph neural networks achieve competitive accuracy without pretraining. GNN models trained entirely from scratch on the native honeycomb molecular graph with only 3-dimensional node features reach $R^2 \geq 0.943$ on κ , stress, and Young’s modulus on the hydrogen dataset, approaching the performance of ImageNet-pretrained CNNs. GNN and ViT models also exhibit greater stability than CNN models when evaluated on the structurally richer methyl dataset, making them preferable choices when cross-chemistry generalization is required.

Taken together, these findings suggest a tiered strategy for surrogate-assisted materials design of functionalized graphene. For rapid, high-throughput screening of H-functionalized configurations, a KNN model trained on as few as 50 MD simulations provides essentially the same predictive guidance as any deep architecture at a fraction of the computational and implementation cost. For data-scarce settings where only a handful of simulations are available, frozen ViT backbones offer robust performance without risk of overfitting. For applications requiring cross-chemistry generalization, where the structural complexity of the methyl system must be handled, GNN and ViT models are the more reliable choice. Strain prediction remains the unsolved problem across all of these regimes, and closing that gap will likely require fundamentally different input representations—ones that encode local bond geometry and stress concentrations—rather than simply more powerful models applied to the same representations tested here. These conclusions are specific to the dataset and conditions studied and should be treated as hypotheses to be tested as datasets grow

and functionalization chemistries diversify.

5.3 Broader Implications

The finding that a trivial model can match deep learning on most targets carries a message that extends beyond functionalized graphene. In any materials-informatics study where the dataset is small and the dominant structure–property driver is a low-dimensional global statistic, the burden of proof should fall on the complex model to demonstrate that it captures information the simple baseline does not. This “simplicity first” principle is consistent with recent calls in the broader ML community for strong baselines before scaling model complexity, and the benchmark methodology developed here—systematic family comparison, learning-curve analysis, representation ablation—can be reused for other 2-D materials systems.

The data-efficiency results have a direct practical consequence for simulation-driven materials design. If a frozen ViT backbone reaches useful accuracy at 12 training samples and KNN-3feat saturates at 50, then a two-stage workflow becomes viable: (i) run a small pilot campaign of ~ 50 MD simulations, (ii) train the cheap surrogate, (iii) use the surrogate to prioritise the next batch of simulations. This is a precursor to full active-learning loops and could substantially reduce the computational cost of exploring new functionalization chemistries.

Finally, the universal failure on strain prediction is itself a finding with broader significance. It suggests that for properties governed by rare, localized events—crack initiation, defect nucleation, phase transformation triggers—global and semi-local representations are fundamentally insufficient regardless of model expressiveness. Closing this gap will require not just larger datasets but qualitatively different input modalities: local stress-field snapshots, bond-topology descriptors, or multi-scale representations that couple atomic detail with continuum mechanics.

5.4 Limitations

Several limitations of the present study should be acknowledged.

Dataset scale. Both datasets contain 300 configurations, placing all models in a data-scarce regime. The learning-curve analysis suggests that deeper architectures (MLP, scratch CNN) have not yet saturated, and their true performance ceiling may be higher with larger datasets.

Single functionalization chemistry per dataset. Each dataset studies one functional group in isolation. Mixed functionalization—simultaneous H and CH₃ coverage, or combinations with substitutional dopants—may exhibit qualitatively different structure–property relationships that are not captured here.

Restricted structural variation. All configurations share the same nanosheet size (8,528 atoms, 220 Å × 100 Å) and the same loading direction (zigzag tensile). Generalization to different sheet sizes, aspect ratios, or loading orientations is not evaluated.

Strain prediction. The universal failure to predict maximum strain across all families and representations indicates a fundamental limitation of the tested input modalities. Addressing this target would likely require representations that explicitly encode local bond-length distributions or stress concentrations.

5.5 Future Work

Based on the findings and limitations of this thesis, the following directions are proposed for future investigation.

Larger and more diverse datasets. Expanding both datasets to $\geq 1,000$ configurations, and introducing mixed functionalization or multi-scale coverage, would allow deeper architectures to exploit their representational capacity and enable more reliable learning-curve analysis.

Active learning for simulation efficiency. Combining the surrogate models developed here with active learning strategies—selecting the next simulation based

on model uncertainty (e.g., GPR posterior variance or Monte Carlo Dropout)—could reduce the number of MD simulations required to achieve a target accuracy by an order of magnitude.

Richer structural representations for strain. Given the universal failure on maximum strain, future work should explore representations that go beyond atom-type assignment: local bond-angle distributions, phonon density of states features, or stress-field maps computed from pre-fracture simulations may provide the localized information necessary to predict fracture initiation.

Deeper GNNs and richer edge features. The plateau of GNN models below the best CNN and ViT results suggests that 3-layer depth and 3-dimensional node features are a bottleneck. Incorporating bond-length or bond-angle edge features, increasing message-passing depth, or employing equivariant GNN architectures may close this gap.

Cross-chemistry transfer learning. Training models on the hydrogen dataset and fine-tuning on a small number of methyl configurations would test whether the representations learned for one functional group transfer to another, potentially reducing the simulation cost for studying new chemistries.

Inverse design. The surrogate models developed here are a prerequisite for gradient-based or evolutionary inverse design: given a target value of κ or Young’s modulus, optimize over the discrete space of functionalization configurations. Combining the GNN surrogate with genetic algorithms or Bayesian optimization represents a natural next step toward automated materials design for functionalized graphene.

Bibliography

- [1] Ravil Ashirmametov, Alexandr Alpatov, Farrokh Yousefi, Narges Vafa, Siamac Fazli, and Konstantinos V. Kostas. Determining the structure of functionalized graphene for tailored thermomechanical properties using ML techniques. *RSC Advances*, 15:44423–44436, 2025.
- [2] Ravil Ashirmametov, Alexandr Alpatov, Farrokh Yousefi, Narges Vafa, Siamac Fazli, and Konstantinos V. Kostas. Functionalized graphene MD dataset. <https://github.com/0-daedalus/graphene>, 2025. GitHub repository, accessed 2026-05-04.
- [3] D. W. Boukhvalov. DFT modeling of the covalent functionalization of graphene: from ideal to realistic models. *RSC Advances*, 3(20):7150, 2013.
- [4] Michael Fernandez, Ante Bilić, and Amanda S Barnard. Machine learning and genetic algorithm prediction of energy differences between electronic calculations of graphene nanoflakes. *Nanotechnology*, 28(38):38LT03, September 2017.
- [5] Victor Fung, P. Ganesh, and Bobby G. Sumpter. Physically Informed Machine Learning Prediction of Electronic Density of States. *Chemistry of Materials*, 34(11):4848–4855, June 2022.
- [6] K.K. Gupta, T. Mukhopadhyay, A. Roy, L. Roy, and S. Dey. Sparse machine learning assisted deep computational insights on the mechanical properties of graphene with intrinsic defects and doping. *Journal of Physics and Chemistry of Solids*, 155:110111, August 2021.
- [7] Amin Hamed Mashhadzadeh, Maryam Zarghami Dehaghani, Amir Hamed Mashhadzadeh, Aidyn Kadyr, Boris Golman, Christos Spitas, and Konstantinos V. Kostas. Machine Learning-Assisted design of boron and nitrogen doped graphene nanosheets with tailored thermomechanical properties. *Computational Materials Science*, 240:112998, May 2024.
- [8] Meirong Huang, Zechen Li, and Hongwei Zhu. Recent Advances of Graphene and Related Materials in Artificial Intelligence. *Advanced Intelligent Systems*, 4(10):2200077, October 2022.
- [9] Seiji Kajita, Nobuko Ohba, Ryosuke Jinnouchi, and Ryoji Asahi. A Universal 3D Voxel Descriptor for Solid-State Material Informatics with Deep Convolutional Neural Networks. *Scientific Reports*, 7(1):16991, December 2017.

- [10] Andrew J. Lew, Chi-Hua Yu, Yu-Chuan Hsu, and Markus J. Buehler. Deep learning model to predict fracture mechanisms of graphene. *npj 2D Materials and Applications*, 5(1):48, April 2021.
- [11] Yinfeng Li, Dibakar Datta, Shihong Li, Zhonghua Li, and Vivek B. Shenoy. Patterned arrangement regulated mechanical properties of hydrogenated graphene. *Computational Materials Science*, 93:68–73, October 2014.
- [12] Yinfeng Li, Dibakar Datta, Zhonghua Li, and Vivek B. Shenoy. Mechanical properties of hydrogen functionalized graphene allotropes. *Computational Materials Science*, 83:212–216, February 2014.
- [13] S.A. Mousavi and A. Montazeri. Predicting mechanical properties of defective h-BN nanosheets using Data-Driven models. *Computational Materials Science*, 228:112380, September 2023.
- [14] Qing-Xiang Pei, Yong-Wei Zhang, and Vivek B Shenoy. Mechanical properties of methyl functionalized graphene: a molecular dynamics study. *Nanotechnology*, 21(11):115709, March 2010.
- [15] Jonathan Schmidt, Mário R. G. Marques, Silvana Botti, and Miguel A. L. Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):83, August 2019.
- [16] Wonjun Shin, Seongwoo Jang, Yunhyoung Hwang, and Jihoon Han. Data-driven prediction of the mechanical behavior of nanocrystalline graphene using a deep convolutional neural network with PCA.
- [17] Tong Wang, Daniel P. Russo, Dimitrios Bitounis, Philip Demokritou, Xuelian Jia, Heng Huang, and Hao Zhu. Integrating structure annotation and machine learning approaches to develop graphene toxicity models. *Carbon*, 204:484–494, February 2023.
- [18] Shaoyu Zhao, Yingyan Zhang, Yihe Zhang, Wei Zhang, Jie Yang, and Sritawat Kitipornchai. Buckling of functionally graded hydrogen-functionalized graphene reinforced beams based on machine learning-assisted micromechanics models. *European Journal of Mechanics - A/Solids*, 96:104675, November 2022.
- [19] Bowen Zheng and Grace X. Gu. Prediction of Graphene Oxide Functionalization Using Gradient Boosting: Implications for Material Chemical Composition Identification. *ACS Applied Nano Materials*, 4(3):3167–3174, March 2021.