# Prediction of Carbon Nanostructure Mechanical Properties and the Role of Defects Using Machine Learning

Jordan J. Winetrout,[1,2#] Zilu Li,[3#] Qi Zhao,[3#] Landon Gaber,[4] Vinu Unnikrishnan,[4] Vikas Varshney,[5] Yanxun Xu,[6] Yusu Wang,[3,7*] Hendrik Heinz[1,2*]

[1] Department of Chemical and Biological Engineering, University of Colorado Boulder, Boulder, CO 80309

[2] Materials Science and Engineering Program, University of Colorado Boulder, Boulder, CO 80309

[3] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093

[4] College of Engineering, West Texas A&M University, Canyon, TX, 79016

[5] Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, 45433

[6] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218

[7] Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA 92093

[#] Equal contribution

[*] Corresponding authors: yusuwang@ucsd.edu and hendrik.heinz@colorado.edu

**Abstract**

Graphene-based nanostructures have extraordinary potential as strong and lightweight materials. A longstanding bottleneck lies in atomic-scale engineering to harness the theoretical limits of modulus and strength, of which only a fraction is routinely achieved today. This contribution introduces a database of over 1,000 pristine and defective nanoscale CNT bundles and graphitic assemblies inspired by microscopy with associated stress-strain curves from reactive molecular dynamics (MD) simulations to analyze structure-stiffness-failure relationships. The 3D structures contain up to 80,000 atoms. Using the database and physics-based machine learning (ML), accurate predictions of elastic moduli and tensile strength with 1,000 to 10,000 faster speeds than efficient MD simulations are demonstrated. Hierarchical Graph Neural Networks with Spatial Information (HS-GNNs) are introduced using chemistry knowledge, and HS-GNNs as well as extreme gradient boosted trees (XGBoost) achieve forecasts of mechanical properties of arbitrary carbon nanostructures with only 3% to 6% mean relative error. Reliability reaches experimental accuracy and is up to 20 times higher than other ML methods. Predictions retain 8% to 18% accuracy for large CNT bundles, CNT junctions, and carbon fiber cross-sections outside the training distribution. The physics-informed HS-GNN works remarkably well for data outside the training range while XGBoost works well with limited training data inside the training range. The databases for carbon nanostructures are suited to integrate multimodal information from diverse experimental and simulation techniques, are extensible beyond 100 nm size, to chemically similar compounds and a wider range of properties. The ML methods can be applied to structural materials, nanoelectronics, and carbon-based catalysts.

Classification: Physical Sciences/Engineering

Keywords: carbon nanostructures, database, simulation, machine learning, mechanical properties

## Significance Statement

We introduce a database of over 1000 dynamic stress-strain curves and failure properties of 3D carbon nanostructures for real-time predictions of elastic and failure properties using machine learning (ML). The models cover defective carbon nanotube bundles, carbon fiber cross-sections, and carbon nanotube junctions up to one hundred thousand atoms and predict mechanical properties up to 10,000 times faster than current molecular dynamics simulations with mean relative errors under 5%. Graph-based machine learning algorithms were developed and compared to existing ML methods, showing promise of physics-based hierarchical graph-based models to perform well for sparse data and structures out of the training range. Potential applications include ML-accelerated defect engineering and design of carbon nanostructures, structural materials, catalysts, and nanoscale electronics.

## 1. Introduction

Since the discovery of graphene-based nanostructures and carbon nanotubes (CNTs), tremendous efforts have been deployed to understand and utilize their remarkable mechanical and functional properties (Figure 1a).(1-4) Examples include applications in aerospace, aviation, and automotive materials, which break down to CNT/polymer composites, CNT paper, CNT yarns, and carbon fiber at the molecular and microscale. In particular, the high modulus and strength of carbon nanostructures in comparison to the low mass density has great potential for applications. The theoretical Young's modulus of CNTs exceeds 1 TPa and the theoretical strength exceeds 100 GPa,(5, 6) however, the actual modulus of CNT-based engineering structures such as CNT bundles and carbon fiber is less than 400 GPa and the strength less than 10 GPa even after decades of development (Figure 1b).(7-9) The losses are attributed to defects, impurities, finite lengths and resulting gaps between non-continuous CNTs,(10-12) random micro-structure orientations, various precursor chemistries for synthetic fiber, gel spinning, and high-temperature graphitization process protocols.(9, 13, 14) A major challenge lies in closing the gap between the current mechanical properties, which are already superior to most natural and engineering materials per

unit weight,(14) and the theoretically possible performance (upper right inset in Figure 1b). Exploration of the large space of improvements could facilitate revolutionary changes in multiple industries.

Quantifying the impact of the performance-limiting features and guiding related synthesis and processing towards customized manufacturing from the atomic scale to the microscopic scale remains difficult in experiments.(15) Characterization of CNT structures relies on X-ray and electron diffraction,(16) microscopy and 3D tomographic reconstruction, Raman spectroscopy, acoustic methods and testing of mechanical properties such as modulus, strength, and toughness.(11, 12) Hereby, the preparation and characterization of samples is time-consuming and expensive, and a comprehensive analysis of the CNT structural, mechanical, and functional characteristics requires a combination of multiple experimental analysis techniques.(17, 18) To complement and accelerate such efforts, simulation methods such as molecular dynamics simulations and density functional theory have been employed.(19-23)

Simulation techniques can provide detailed insight at the scale from atoms to tens of nanometers, recently including entire stress-strain curves from equilibrium to failure in high accuracy.(24-26) It is possible to create and evaluate numerous well-defined model structures inspired by the defects and nanoscale features observed in laboratory characterization to aid experimental design.(10-13, 16, 27, 28) The model structures can then be used to examine the structure-specific failure mechanisms that are difficult to elucidate from experiments alone. While more time-efficient than experiments (hours versus weeks per structure), the computational cost remains considerable, and the typical system size of 1 to 50 nm remains much below the scale of carbon fiber and related devices.

Recently, artificial intelligence (AI) and machine learning (ML) methods have been rapidly developed to interrogate properties of materials (Figure 1c-d).(29, 30) ML has already achieved remarkable success on a wide range of applications, such as image classification and segmentation, language translation, and DNA sequence analysis.(31-37) Here we explore physics-inspired graph neural networks to predict mechanical properties from 3D structures with high acceleration (Figure 1c). Different from conventional approaches, ML algorithms build models and infer knowledge about materials from large, sampled data (Figure 1d). Following feature definitions, hierarchical spatial information, and known target properties in the dataset, the trained models can then predict

materials properties or changes in the dynamics of simulated materials structures. One of the more powerful and commonly used frameworks in chemistry and materials science is deep learning (DL).(38, 39)

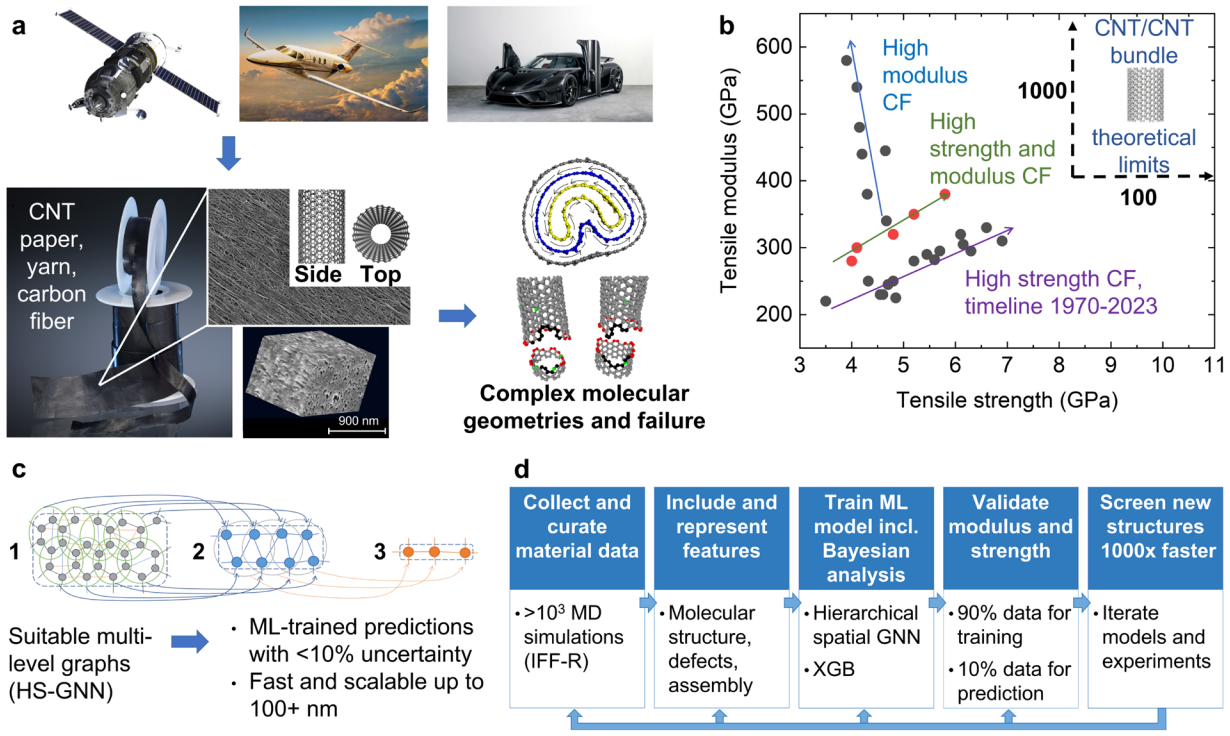


**Figure 1.** Hierarchical structure of carbon-based materials, challenges, and machine learning workflow to predict mechanical properties at the nanoscale. (a) Space, air, and automotive vehicles (top) require strong and lightweight materials, for which CNT yarn, sheet, and carbon fiber composites are a promising choice. At the microscale, CNT yarn and sheet consist of aligned yet highly defective carbon nanotube bundles and irregular graphitic structures (bottom left). 3D imaging by tomography remains limited to the coarse nanometer scale. In this work, we examine stress-strain relationships up to failure for a library of over 1000 structures at the small nanometer scale and feed the data into a ML model to enable faster, real-time predictions (bottom right). (b) The current performance of carbon fiber (CF) remains far lower than the theoretical limits for CNTs and CNT bundles (top-right), which are ca. 1000 GPa in tensile modulus and ca. 100 GPa in tensile strength. Efforts in the last 50 years have brought steady improvements (illustrated by various data points and arrows), and CNTs remain a superior candidate for structural applications.(13) (c) We introduce a Hierarchical Spatial Graph Neural Network (HS-GNN) with

excellent performance in mechanical property predictions for graphitic morphologies up to the large nanometer scale. (d) The workflow includes collection of stress-strain curves by reactive molecular dynamics simulations (validated relative to experiment), representation of the structures for the HS-GNN or extreme gradient boosted trees (XGBoost), and training of the ML model with the data for modulus and strength. The goal is the prediction of mechanical properties of unknown structures $10^3$ times faster than MD simulation, i.e., within seconds, to quantify the role of defects and explore desirable new nanoscale designs.

Deep learning uses multiple neural network layers to map input raw data onto output target properties, performing representation learning and task optimization simultaneously. Neural networks can learn molecule representations and apply them to prediction tasks in an end-to-end and data-driven manner. Depending on the representation of atomic structures used as inputs, we can classify deep learning approaches for molecular or material science applications as follows.

First, we could model a molecular structure as a 3D point cloud where each atom is represented by a point in $R^3$, and then apply the PointNet(40) or Relation-shape Convolutional Neural Network (RS-CNN)(41) architecture to perform property prediction over such point cloud representation.(42) This approach can encode the 3D structural information of molecules, however, simply representing molecules as point clouds can lose significant structural information such as chemical bonds. Xu et al.(43) and Gebauer et al.(44) improved this approach by designing more expressive point cloud convolution networks which align local neighbor information and symmetry factorization of point distributions. Furthermore, fingerprints(45-47) can encode chemical bond relations, which can be viewed as first applications of Graph Neural Networks (GNNs) to the prediction of properties of molecular structures.

Second, a natural way to organize molecules is a representation as graphs (Figure 1c). Atoms are represented as nodes in graphs and two atoms are connected by edges when a chemical bond exists between them. GNNs like Graph Convolution Networks (GCN)(48) and Graph Attention Networks (GAT)(49) can perform convolution over graph-structured data. Message-passing neural networks (MPNNs) are extended frameworks that further introduce message passing into GNNs and support architectures for predicting molecular properties (Section S1 in the SI Appendix). DimeNet(50) and D-MPNN(51) further extended this approach by incorporating bonded

interactions like angles between bonds into models. Although GNNs outperform other approaches mentioned above, for tasks like classification of small molecules, existing approaches have several limitations: (1) In addition to atom connectivity and bonded interactions, a good model should also take non-bonded interactions and features that characterize spatial (geometric) bonded patterns of atomic structures into consideration. (2) Due to over-smoothing or over-squashing phenomena, long-range interactions or signals are hard to capture, making it challenging to process large scale molecular structures and assemblies. (3) The neighbor aggregation scheme in existing GNNs fails to be defect-aware as it smooths out signals of local defects and performance to predict properties of defective structures degrades, e.g., in CNTs with missing atoms after gamma-radiation (Figure 2a-c).(52)

To-date, no study has been reported that determines the influence of the wide array of atomic scale defects on CNT bundle mechanical properties to the best of our knowledge. Also, we are unaware of studies that incorporate machine-learning to quantify how the defects prevent CNT bundles from reaching the theoretically possible performance (Section S2 in the SI Appendix).(53) In this contribution, we generated over 1,000 different 3D atomic structures of CNT bundles and graphitic morphologies, along with stress-strain curves up to failure. We utilized reactive MD simulations with the reactive version of the INTERFACE force field, IFF-R,(25, 26, 54) to generate the stress-strain curves for all 3D structures. IFF-R allows bond breaking, has been extensively validated, and exceeds DFT accuracy up to multiples.(20, 55) Typical deviations of computed moduli and strength from experimental data are ~5%.(26) IFF-R uses interpretable energy terms designed to represent defects and structural variations in comparable accuracy, and includes a wide range of chemistry.(54, 56-59) The 3D carbon nanostructures and associated mechanical properties were used as a ground truth to build a database and explore ML schemes to predict mechanical properties directly from the 3D atomic structure. We introduce a new machine learning framework, the Hierarchical Spatial Graph Neural Network (HS-GNN). HS-GNN is an enhanced GNN designed to process atomic structures of graphitic carbon nanostructures and CNT bundles, and the first reported ML scheme to predict mechanical properties of such structures beyond a size of 10,000 atoms. The input structures are modeled as a heterogeneous graph to include covalent bonds and short-range (non-bonded) interactions among atoms. Persistent summaries were extended by capturing local geometric and topological features, which are necessary to better encode the spatial geometric shape and topological defects, e.g., missing atoms and chemical bonds, that affect the

mechanical properties. A 3-level hierarchy in the neural network captures short range and long-range interactions among different parts of the CNT bundles, providing a means to include multiple length scales. Initial training of the HS-GNN involved 1139 independent configurations of CNT bundles and, separately, a dataset of 958 CNT junctions.(18, 60, 61)

The HS-GNN as well as simpler extreme gradient boosted trees (XGBoost)(62) demonstrate remarkable predictive accuracy and speed. For new atomic arrangements in CNT bundles not present in the training set, HS-GNN predicts the tensile modulus and strength with only 3% to 6% mean relative error (MRE), far exceeding earlier structure-based ML approaches such as DimeNet.(50) HS-GNN also yields the most accurate prediction for mechanical properties of complex CNT junctions, enlarged CNT bundle structures (containing ~$10^5$ atoms), and graphitic carbon fiber cross-sections outside the training distribution at 10% to 20% mean relative error. Computational efficiency is extremely high, requiring less than 0.1% of the time of traditional molecular simulations for individual predictions. XGBoost,(62) using only global features, even surpasses HS-GNN in computation speed, although with much larger errors in property ranges with fewer training data and for geometries outside the training distribution.

## 2. Results

### 2.1. 3D Structure Databases and Training Data for Mechanical Properties

The primary data set contained 1139 periodic models of pristine and defective carbon nanotube (CNT) bundles (Figure 2). Model building utilized available data from X-ray diffraction, microscopy, imaging, spectroscopy, and mechanical analysis.(1-3, 5-13, 15-18, 53) Morphologies in the CNT bundle data set included individual and bundled single wall (Figure 2a), double-wall (Figure 2b), as well as triple-wall CNTs, both defect-free and with many different types of defects (Figure 2c and Figure S1 in the SI Appendix). Target properties for ML predictions were the tensile strength and modulus, which we obtained from reactive molecular dynamics simulations of stress-strain curves up to failure using the reactive Interface Force Field (IFF-R) (Figure 2d and Figure S2 in the SI Appendix).(25, 54) IFF-R quantitatively reproduces the crystal structures of graphite and CNTs within 0.5% deviation from experiments,(20) the structure of flattened CNTs,(63) π-π stacking, surface and interfacial energies within 5% of experimental measurements, as well as tensile moduli and tensile strength of CNTs, graphene, and graphite in 5 to 10% agreement with

experimental data.(20, 24, 25) IFF and IFF-R were also validated for simulations of defects, interfaces of graphitic structures with solvents and polymers, reproducing contact angles and adsorption energies within ~10% of experimental data.(19, 20, 24, 54) IFF and IFF-R also reach high accuracy for chemistries across the periodic table such as metals,(55, 64) oxides,(57) 2D materials,(58, 65, 66) polymers, gases,(67) and associated multi-phase materials.(54, 68)

From the computed stress-strain curves for all structures, the Young's modulus was calculated as the slope of the linear stress-strain response at low strain, typically between 0.00 to 0.01 (Figure 2d). The tensile strength was determined to be the maximum stress before failure. As an example, a double-wall carbon nanotube (DWCNT) with initially broken inner walls had a tensile modulus and tensile strength of 210 GPa and 19.2 GPa, respectively (Figure 2d). We also recorded the strain at break, half strain at break and half the tensile strength to explore causal relationships (Figure 2j).

Coverage of a diverse range of morphologies is advantageous for new physical insight and effective ML-based property predictions. We included over ten different types of defects in approximately two thousand structural variations in our databases (Section S3 and Figure S1 in the SI Appendix). The associated stress-strain curves and mechanical properties display a wide range of trends and takeaways for nanostructure design (Figure S2 in the SI Appendix). In summary, the models of CNTs and CNT bundles reflect experimentally known variations such as pristine structures (Figure 2a, e), missing atoms (Figure 2b-c, 2e, and 2f), bond dissociation defects, Stone-Wales defects, discontinuities (Figure 2d, f), as well as larger-scale defects like plastically deformed, crumpled, and rearranged nanotube forests (Figure 2f). The stress-strain curves from the database for various morphologies, such as defective and crumpled CNTs and CNT bundles, exemplify the structure-property relationships used in machine learning (Figure 2h, i). For a bundle of two SWCNTs of 1.22 nm diameter (the smallest possible diameter would be ~0.7 nm(19, 69)), increasing the size of a vacancy with missing atoms shows a diminishing effect on modulus and strength (Figure 2e, h). The pristine CNT exhibited a tensile strength of 85 GPa, which was lowered to 61 GPa after introducing a monovacancy into each nanotube, a 28% decrease. Di- and tri-vacancies in the CNTs led to a tensile strength of 54 GPa and 52 GPa, respectively, equal to a loss of 36% and 39%. The tensile modulus showed a similar loss, albeit limited to 15% in the same sequence, indicating convergence past a critical flaw size. The role of complex defects is also captured in the database (Figure 2f, i). A compressed, crumpled nanotube

assembly without broken bonds retains a tensile strength larger than 100 GPa (Figure 2i). Other CNT bundles with broken bonds, Stone-Wales defects, or missing atoms exhibit notable losses in tensile strength. Some defects may increase the strain at break and offer a toughening mechanism, however, comparisons require caution due to the coexistence of broken and pristine nanotubes in defective bundles. Defects with broken bonds, i.e., ruptured ends of CNTs or line defects of multiple broken bonds had typically the most detrimental impact on mechanical properties. Stone-Wales and vacancy defects led to properties somewhere between broken and pristine CNTs. Trends in mechanical properties can be explored for further analysis in the full CNT bundle data set of over 1000 different structures with stress-strain curves.

A further, independent data set of 958 non-periodic CNT junctions and associated mechanical properties was developed by Varshney et. al.(70, 71) The dataset on CNT junctions was employed to the test the applicability of HS-GNN and other ML methods to different nanoscale topologies.

As a limitation, we consider nanostructures containing only carbon atoms in our dataset and exclude oxidized states such as graphene oxide. Focusing on data with uniform high quality allows the in-depth assessment of ML methods. The remarkable performance encourages future studies that include heteroatoms, polymers, and heterogeneous reference data, extending the present database.

### 2.2. Feature Definition and Causal Relationships

Machine learning and predictions of mechanical properties from the 3D structural models relies on global features when using XGBoost as well as on global features plus graph analysis for HS-GNN (Figure 2g). We included 7 global features, namely, the number of CNTs, the CNT diameter, the CNT height (length), the number of walls for each CNT (CNT type), i.e., to distinguish SWCNT (1), DWCNT (2), and TWCNT (3), the number of atoms, the number of bonds, and the cross-sectional area of a CNT bundle (calculated as the sum of cross-sectional areas of individual CNTs). These features have some interdependence, and we defined the bond ratio as an additional feature, given by the total number of C-C bonds divided by the total number of atoms. The bond ratio is equal to 1.5 for pristine CNTs and less than 1.5 for a defective CNT bundle, helping to characterize defective bundles. Broken CNTs have the lowest bond ratio.

To analyze relationships between global features and predicted properties (labels), we deployed a causal additive model (CAM) (Figure 2j).(72) Among various interesting correlations such as between the type of CNT and ½ strain at break, and the absence of correlations between the cross-sectional area and some properties (as expected), the CAM revealed a strong correlation between the simulated strength and the simulated modulus (Figure 2j). I.e., if the strength was known, it could be used to improve predictions of the tensile modulus. However, incorporation of the strength-modulus relationship improved the quality of predictions in HS-GNN only if we used the ground truth value for strength and, therefore, our final model of HS-GNN, predicts tensile strength and tensile modulus concurrently without the CAM.

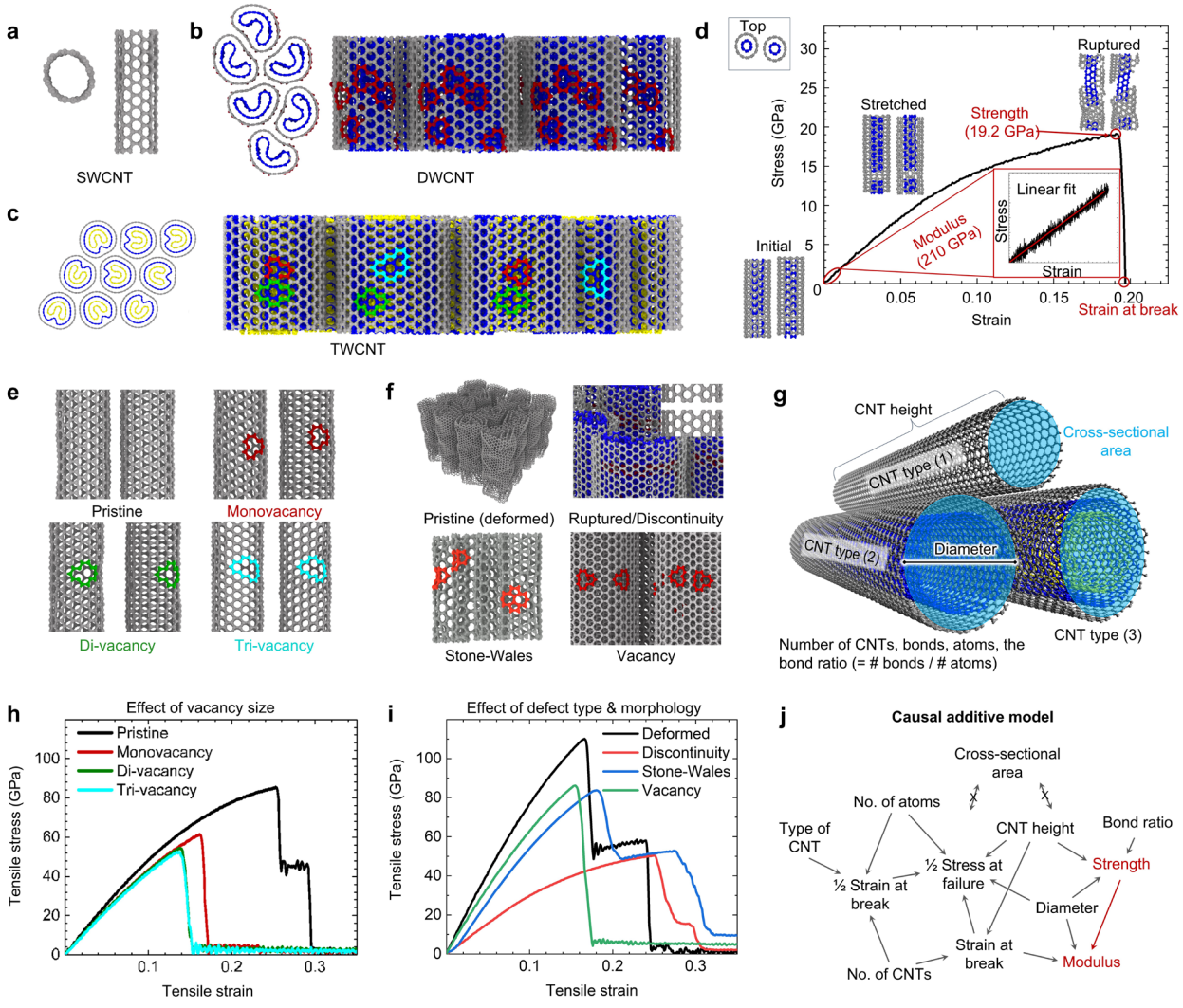

**Figure 2.** Illustration of the primary structure database on CNT bundles and of the methods to collect data for the ML models. (a) Top view (left) and side view (right) of single-walled carbon nanotube (SWCNT) without defects, (b) a bundle of double-walled carbon nanotubes (DWCNT) with monovacancy defects (inner walls highlighted in blue and vacancies highlighted dark red), (c) a bundle of triple-walled carbon nanotubes (TWCNT) with monovacancy, di-vacancy (green), and tri-vacancy defects (innermost walls highlighted in yellow and tri-vacancies highlighted light blue). (d) Sample stress-strain curve from reactive molecular dynamics simulation with IFF-R for two DWCNTs with initially broken inner walls. Snapshots of the CNTs at different strain points are shown. Red highlights indicate mechanical properties used for training the HS-GNN: strength, modulus, and strain at break. The calculated values are listed next to their labels in parentheses. (e) Two SWCNTs show vacancy sizes used in the data set: none (pristine), mono-, di-, and tri-vacancies. (f) Illustration of complex defects in CNT bundles used in the data set. Compressed, deformed CNT bundles, discontinuous inner and outer walls (close-up shown in top-right corner), Stone-Wales defects (pentagonal and heptagonal lattice defects), and vacancies in inner and outer walls are shown. (g) Perspective view of a CNT bundle and illustration of the 7 global features considered in the data set to carry out machine learning. The ratio of bonds to atoms was additionally used to quantify the extent of defects. (h) Sample stress-strain curves for the CNT bundles shown in (e), demonstrating a loss of tensile strength and modulus of elasticity by increasing vacancy size. (i) Sample stress-strain curves for the CNT bundles shown in (f), indicating major differences in stress-strain characteristics. Importantly, mechanical properties of pristine and defective CNTs also depend on the diameter and number of walls. (j) Relationships between structural features from (g) and the mechanical properties as determined by a causal additive model (CAM). For example, the CAM reveals that knowing strength can improve the prediction of the modulus (red highlights). Incorporating this relationship into HS-GNN improved the performance for modulus prediction when using the ground truth for strength but not when using ML-generated values.

### 2.3. Machine Learning Pipeline for Property Prediction

The high-level framework of the Hierarchical Spatial Graph Neural Networks (HS-GNN) is illustrated in Figure 3. HS-GNN uses the initial 3D atomic structure, denoted as *S*, of a CNT-like

or other graphitic structure as an input. HS-GNN takes the importance of covalent bonds as well as of short-range interactions between non-bonded atoms into account, as opposed to employing a plain PointNet-type architecture(40) which would treat the set of atoms as a collection of points as an input, each equipped with a radius. The HS-GNN is a specially designed graph neural network-based architecture to process $S$ and consists of three components: a heterogeneous graph representation of the input structure (Figure 3a), a hierarchical neural network model engineered to capture long-range interaction among different parts of the CNT bundle (Figure 3b), and the integration of local geometric and topological features to enhance the encoding of the shape of CNT bundles and of defect information (Figure 3a). Each of these components is briefly described in the following.

*Heterogeneous Graph.* It is common to use the bond-graph $G_{bondS}$ to represent $S$, where each node corresponds to an atom, and there is an edge between two nodes if the corresponding atoms form a covalent bond (1, 2 bonded). We go beyond the bond-graph and generate a heterogeneous graph, denoted as $G_S$, with one further type of edges among nodes from the initial atomic structure $S$ (Figure 3a). We also connect any two nodes whose corresponding atoms form a dihedral angle, the angle of two planes formed by four sequentially bonded atoms rotated about a central bond (1, 4 bonded). This heterogeneous graph is then fed into a hierarchical GNN.

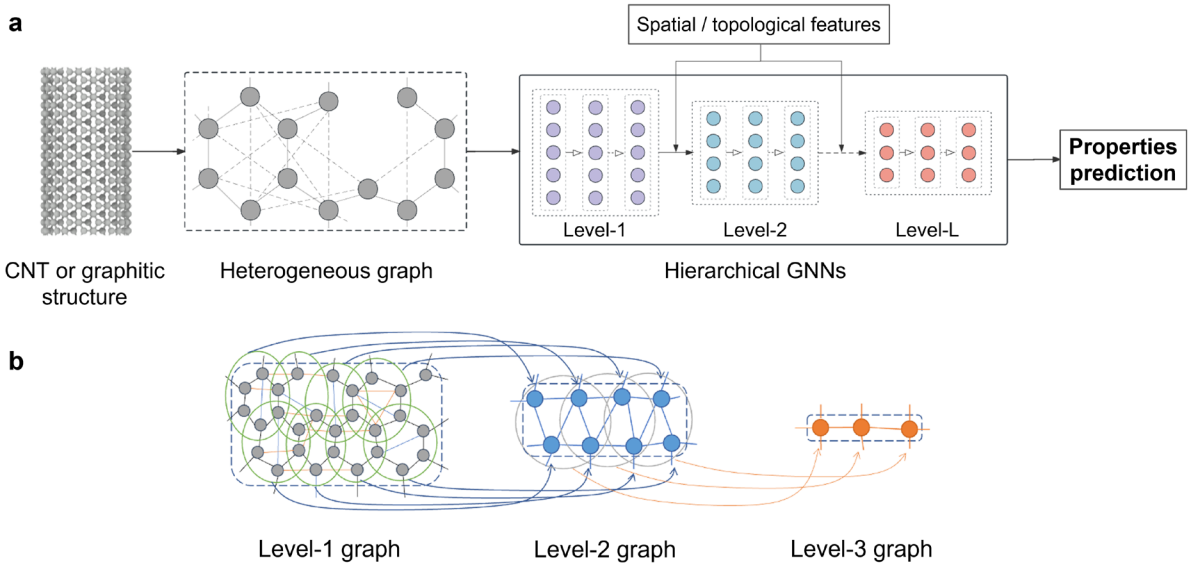

**Figure 3.** Illustration of the machine learning pipeline for HS-GNN. HS-GNN is a carefully designed hierarchical and spatial GNN to predict mechanical properties of CNT-based and other graphitic structures. (a) As an overview, the pipeline begins with the generation of a heterogeneous graph for a given CNT or CNT bundle structure, which is then processed with a hierarchical GNN. A specialized GNN is created within each level of the hierarchy to learn graph representations. The vertical arrays in each level indicate the micro-layers. In addition, topological and geometric features, specifically, persistence summaries of clusters, are computed from the CNT structure. These features encode the local shape and defects and are incorporated into HS-GNN after hierarchical pooling. (b) Construction of the hierarchical graph series. To obtain a higher-level graph, nodes in the lower-level graph that are spatially proximate are grouped into clusters. In the higher-level graph, each node corresponds to a cluster of nodes in the lower level, and nodes in a higher-level are connected by an edge if their corresponding lower-level clusters exhibit any overlap.

*Hierarchical GNN Model.* GNNs are widely used to process graph data and learn graph representations. At a high level, a standard GNN takes a graph with initial node features as an input and processes the graph through $\ell$ layers. In each layer, the GNN updates the feature vector stored at each node of the graph by aggregating features from neighboring nodes and applying a learnable transformation of node features (Section S1 in the SI Appendix). Hereby, GNNs often suffer from over-smoothing(73) which impedes their ability to reach deeper layers. This limitation constrains the effectiveness of capturing information from long-range interactions and the capability of accurately weighing the impact of local defects. We designed a hierarchical GNN that processes the input graph $G_S$ at multiple resolutions to mitigate this issue (three in our algorithm as shown in Figure 3a, b).

Our design finds hierarchies based on spatial geometry information rather than graph topology or node features and distinguishes the HS-GNN from earlier hierarchical GNN approaches.(74, 75) In our approach: (1) The lowest level $L_1$ operates on the input graph $G_S$; (2) At a higher level $L_k$ ($k > 1$), each node corresponds to a cluster of nodes from previous level $L_{k-1}$ and is referred to as a *hyper-node*; (3) The initial feature vector associated to each hyper-node is obtained by applying a pooling layer to the graph at level $L_{k-1}$. Within each resolution level $L_k$, we use several

GNN aggregation layers, specifically, GIN and/or GAT, which we designate as *micro-layers*. Finally, additional pooling layers are applied to the final micro layer of each hierarchy level to obtain a graph-level representation for downstream property prediction.

*Encoding Local Geometry and Topology into HS-GNN Using Persistent Summary.* The shape of carbon nanotubes, graphene-like sheets, and the spatial relation among neighboring structures impact the mechanical properties. To encode such geometric information, namely local interactions among spatially close atoms, from potentially different nanotubes or walls, we use the so-called persistent homology (PH) that characterizes the spatial distribution of points within each cluster in each hierarchical level. PH has been one of the most significant developments in topological data analysis over the last two decades and proven effective in characterizing a variety of intricate shapes.(76, 77) In our context, a set of points (representing atom centers) form a local cluster and PH can provide a concise yet informative summary of the shape formed by these points. Herein, we track the creation and death of topological features with respect to a growing sequence of space around these points, which is effective in capturing local defects using (Figure S3, Section S4 and Section S5 in the SI Appendix).

*Extreme Gradient Boosted Trees (XGBoost).* An alternative to HS-GNN are Extreme Gradient Boosted Trees (Figure S4 and Section S6 of the SI Appendix).(62) XGBoost uses only global features and is faster than HS-GNN. The performance is outstanding in parts of the training range with sufficient local data, however, it decreases sharply in areas of sparse data and for structures outside the training range information.

**Table 1.** Mean relative error (MRE) in the prediction of strength and modulus of CNT bundles and CNT junctions by ML methods. Computing time for predictions and training is also shown. We used 90% of the respective data sets for training and 10% for prediction. Methods are grouped into structure-based approaches and feature-based approaches in the order of decreasing prediction performance from left to right using the primary CNT bundles data set. The highest performing methods XGBoost and HS-GNN reach approximately 5% or less average error on the primary training set. (HS-GNN = hierarchical spatial graph neural network, PH = persistent homology, BWGNN = beta wavelet graph neural network, GCN = graph convolutional network, XGBoost = extreme gradient boosted tree, MLP = multilayer perceptron, LR = linear regression). Prediction times are over 1000 times faster than reactive MD simulations (below 1 second versus 30 min per structure).

| **Data set** | **Property** | **Structure-based methods** | | | | | **Feature-based methods** | | |
|---|---|---|---|---|---|---|---|---|---|
| | | HS-GNN | PH | DimeNet | BWGNN | GCN | XGBoost | MLP | LR |
| CNT bundles | Strength or modulus[a] | **Mean relative error of prediction (%)** | | | | | | | |
| | Strength | **3.4 ± 0.8** | 5.6 ± 0.4 | 13 ± 0.7 | 14 ± 0.03 | 35 ± 1.8 | **3.0 ± 0.1** | 6.6 ± 0.2 | 29 ± 0.7 |
| | Modulus | **5.6 ± 0.1** | 8.7 ± 0.6 | 16 ± 0.5 | 16 ± 0.2 | 39 ± 0.2 | **5.5 ± 0.0** | 7.0 ± 1.3 | 31 ± 0.01 |
| | | **Training time (seconds per 1 A100 GPU)** | | | | | | | |
| | | 5160 ± 150 | 30 ± 60[b] | 6900 ± 690 | 3480 ± 282 | 1080 ± 25 | 30 ± 6 | 30 ± 0.0 | 11 ± 0.0 |
| | | **Prediction time[c] (seconds per 1 A100 GPU)** | | | | | | | |
| | | 0.83 ± 0.10 | 0.03 ± 0.01 | 206 ± 8 | 0.49 ± 0.10 | 0.34 ± 0.20 | 0.20 ± 0.01 | 0.07± 0.01 | 0.001 |

| | | **Prediction time by MD simulation: 1800 seconds per structure (using 64 CPU cores)** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Mean relative error of prediction (%)** | | | | | | | |
| CNT junctions | Strength | **11.8 ± 0.3** | 14.0 ± 1.1 | 12.3 ± 0.9 | 15.0 ± 0.5 | 40 ± 1.8 | **11.3 ± 0.4** | 12.9 ± 0.4 | 20 ± 0.0 |
| | Modulus | **8.3 ± 0.6** | 11.8 ± 1.5 | 10.0 ± 1.8 | 15.5 ± 0.6 | 66 ± 0.1 | **6.2 ± 0.2** | 10.3 ± 0.3 | 56 ± 0.0 |

[a] Computing times refer to training for the prediction of one property, i.e., either strength or modulus. Training of a separate model is necessary to predict both properties simultaneously. [b] The PH based method requires expensive pre-processing time (more than one day). [c] Total computation time to predict strength or modulus for the complete test set of 115 CNT bundles.

### 2.4. Prediction of Tensile Modulus and Strength for CNT Bundles using ML

The tensile strength and the tensile modulus of various CNT configurations were tested for multiple structure-based ML models such as HS-GNN as well as feature-based models such as XGBoost (Table 1 and Figure 4). We first evaluated the HS-GNN using our benchmark dataset of 1139 CNT-bundles, using a random selection of 90% of structures for training and 10% of structures for property prediction (Table 1 and Figure 4).

*Prediction Accuracy.* HS-GNN has a mean relative error (MRE) of 3.4% for the prediction of the tensile strength and 5.6% for the prediction of the modulus of CNT bundles (Table 1, Figure 4a, b, as well as Table S1 and Section S7 in the SI Appendix). Laboratory measurements often have more than ±5% uncertainty so that the reliability of the ML approach is outstanding in comparison. More broadly, 10% uncertainty is valuable for practical purposes. XGBoost reached similar overall accuracy as the HS-GNN (Table 1), although values in certain ranges are much less reliable with errors exceeding 10% (Figure S5 in the SI Appendix versus Figure 4a, b).

Accordingly, the quality of the predictions depends on the range of values and number of available training data (Figure 4a, b and Figure S5 in the SI Appendix). Prediction errors for HS-GNN were low for an intermediate range of tensile strengths of 40 to 100 GPa and for tensile moduli of 200 to 800 GPa (Figure 4a-d), which correlate with over 80% of training data available (Figure 4e, f). The prediction of tensile strengths in the range 0-40 GPa (Figure 4a, c) and moduli in the range 0-200 GPa (Figure 4b, d) remained exceptionally accurate for HS-GNN while several other methods had errors greater than 20%. The associated data in this category represents only 3.7% of the cases in the training set (Figure 4e, f) and includes highly defective CNT bundles that can have dramatically different structures after failure (Figure 2d). Low errors for HS-GNN result from accounting for relevant physical and chemical features as well as for the hierarchical structure, while other ML methods without such attributes contribute higher errors (Figure 4a-d). MLP and XGBoost are among the next best performers. Predictions of mechanical properties for CNT bundles close to the upper limit of performance, 100-120 GPa for strength and close to 800 GPa (or over 800 GPa) for modulus (Figure 4a-d), were dependable in comparison. In these instances, the reduced complexity of near-pristine CNT bundles contributes to predictions of high quality (Figure 2a-c, e), even though the structures contribute only a small portion in the training

set (Figure 4e, f). Specifically, XGBoost has a strong dependence on the number of training data (Figure S5 in the SI Appendix). The performance is extremely good when a large amount of training data is available, or defects are sparse, and it is quite poor when training data are sparse, e.g., for low strength and low modulus. HS-GNN is more balanced and up to ten times more accurate for sparse data (Figure S5 in the SI Appendix versus Figure 4).

In comparison to experimental data, high elastic moduli between 700 and 1000 GPa are predicted by HS-GNN and XGBoost for non-defective CNTs and CNT bundles, as well as tensile strengths between 80 and 120 GPa. These data align with laboratory measurements and high-level ab-initio calculations,(10, 27, 53) in the same range depending on the CNT diameter and number of walls. IFF-R and ML predictions reproduce these trends(26) and are suited to predict the properties of defective structures well which remains a challenge for direct measurements.

*Computational Speed.* ML methods are promising for broad applicability due to less computing time compared to system-specific simulations. The deployment of HS-GNN, XGBoost, and other methods involves three stages: feature analysis and graph creation (or feature analysis only for feature-based methods), training, and prediction. We report the associated times using 1 GPU (A100) on a server with 128 CPU cores. (1) The first step involved the creation of graphs from the 1000+ CNT bundle structures such as the persistent homology (PH) inside HS-GNN. These computations are time-intensive and take 3 to 4 days. In contrast, only minutes were necessary when using XGBoost to prepare a spreadsheet of global features using a suitable script. (2) The second step was initial training of the models on the database (Table 1). For example, 86 minutes were required to train the HS-GNN model, mostly for computing the persistence summaries of CNT bundles in the training set. XGBoost, with simpler input and model, needed only 0.5 minutes for training. (3) The third step involved the prediction of properties, which the algorithms perform simultaneously for all structures (Table 1). The prediction (inference) process to obtain both the tensile strength and the modulus for over one hundred bundled CNT nanostructures simultaneously (10% of the data set) took only 0.83 seconds for HS-GNN. XGBoost completed the same task in 0.2 seconds.

In comparison, an IFF-R MD simulation of the full stress-strain curve for one bundled CNT nanostructure took about 30 minutes using 64 CPU cores, plus analysis time (Table 1). Prediction tasks by HS-GNN and all other ML methods are therefore extremely fast, at least a factor of 1,000

times faster (Table 1). XGBoost leads in terms of speed at useful accuracy due to quick featurization time, training time, and is scalable for millions of structures if more data becomes available. HS-GNN is more dependable than XGBoost and works well with sparse data. The downside is a long lead time for initial graph creation and ~100 times longer training time of the network than XGBoost. Predictions, once trained, occur in under one second, too.

In summary, the speed using HS-GNN for the same nanostructure is more than 1,000 times faster than traditional MD simulations and could be increased up to $10^6$ times faster if MD simulations for training would use lower strain rates. Predictions using XGBoost are 10,000 times to $10^7$ times faster, suitable for real-time feedback.

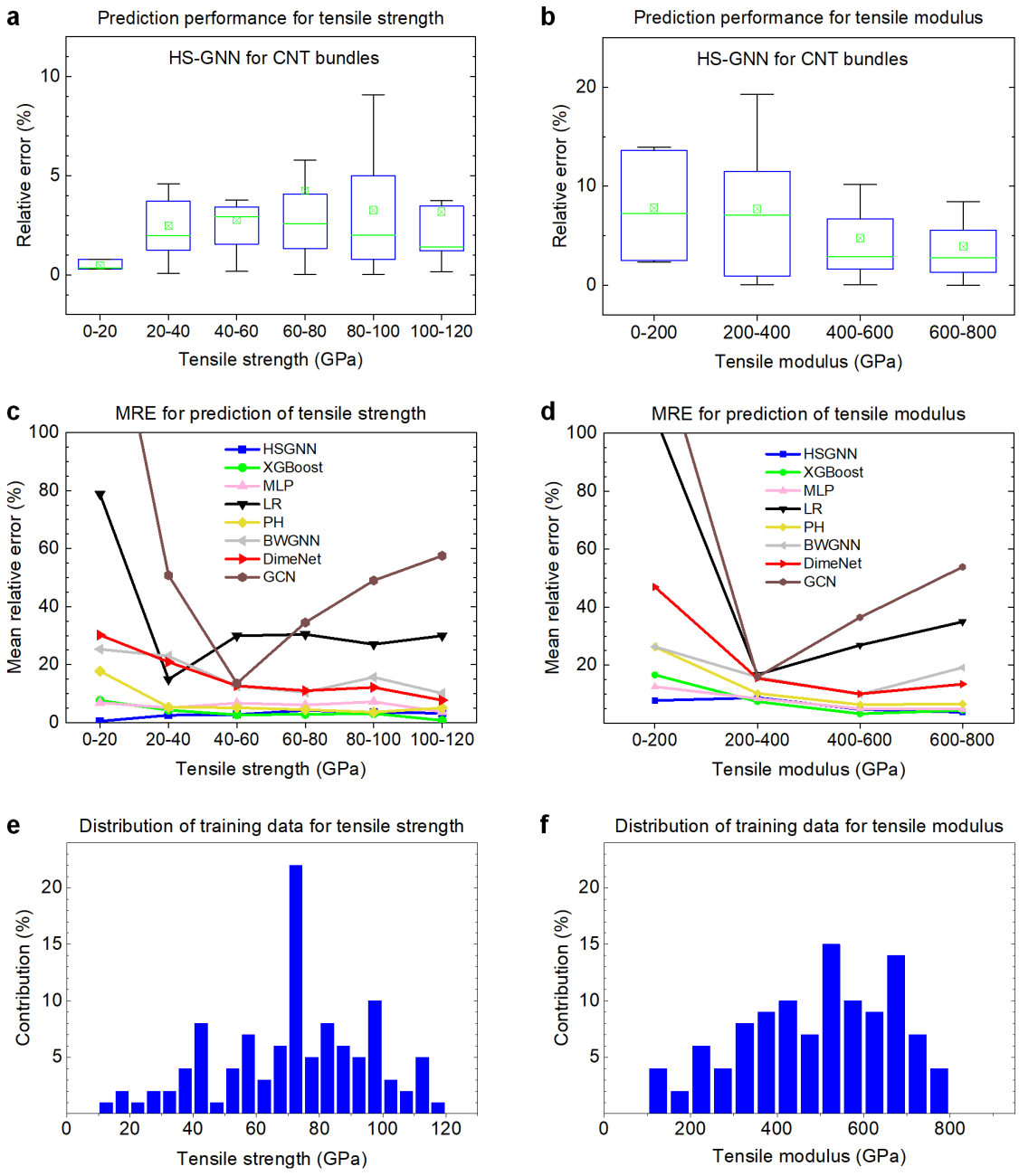


**Figure 4.** Prediction of mechanical properties for CNT bundles using the HS-GNN and other ML methods. (a, b) Box plot of prediction error of tensile strength (a) and the tensile modulus (b) for the data set of CNT bundles using the HS-GNN (see equivalent data for XGBoost in Figure S3 in the SI Appendix). The CNT-based models were divided into six groups of increasing strength and modulus, respectively, to illustrate the sensitivity of errors in property prediction on the range of values. The horizontal lines represent the median prediction error, the open squares the mean relative error (MRE), the box boundaries the 25% and 75% quartiles, and the outside bars the

global minimum and maximum of errors. (c, d) Comparison of MRE baselines for the HS-GNN, extreme gradient boost (XGBoost), and six other methods for tensile strength (c) and tensile modulus (d) using the CNT bundle data set. The baselines include multi-layer perceptron (MLP), linear regression (LR) based on global features of the CNT materials, persistent homology (PH)(77), beta-wavelet graph neural network (BWGNN)(78), DimeNet(50), and graph convolutional network (GCN). HS-GNN or XGBoost outperform many other ML algorithms by an order of magnitude. HS-GNN maintains all MREs under 10% and even 5% for tensile strength. (e, f) Data distribution for tensile strength and tensile modulus in the CNT-bundle data set.

*Comparison with Other ML Approaches.* The HS-GNN is a structure-based method for efficiently predicting mechanical properties of carbon nanostructures, including chemical information and structural hierarchy (Table 1 and Figure 4). (XGBoost)(62) is a feature-based method with a simpler model architecture that requires significantly less training time. A comparison of these two promising ML architectures to six other ML baselines, including structure-based and feature-based models, shows significant differences (Figure 4c, d). Structure-based models take the 3D graph representation of CNTs as an input, whereby the atoms correspond to nodes, and chemical bonds to edges. Vanilla GCN,(48) Persistent Homology (PH),(77) beta wavelet graph neural network (BWGNN)(78), and DimeNet(50) were tested. DimeNet has been prior state-of-the-art in property predictions of molecular structures and underperforms relative to HS-GNN and XGBoost with 2 to 8 times larger errors (Figure 4c, d). PH works notably better, although it still has about double the error compared to HS-GNN and XGBoost, and multiple times the error at low modulus and strength. PH also requires much higher pre-processing time to obtain the structural features (more than one day with our server). The BWGNN specifically captures outlier signals over graphs, which in our case, are local defects and shape distortion. However, the performance is low, like DimeNet, and GCN is clearly unsuited due to large errors (Figure 4c, d). Further feature-based models include linear regressor (LR),(79) multi-layer perceptron (MLP),(80, 81) and extreme gradient boosting (XGBoost),(62) which use global features as inputs (Figure 2g) to predict associated properties. MLP achieves good accuracy, although behind XGBoost and HS-GNN. Linear regression is the fastest method, although with large errors (Table 1 and Figure 4c, d).

HS-GNN outperforms other baselines in predicting tensile strength and Young's moduli of CNT bundles (Table 1 and Figure 4c, d). The mean relative errors (MRE) are approximately half of those of the best prior structure-based method DimeNet. All models indicate a trend of lower prediction error for property ranges with more training cases. HS-GNN prevents large upswings in prediction error towards lower and upper limits of strength and modulus (Figure 4c, d).

*Sensitivity Analysis by Ablation and Reduction of Dataset Size.* An ablation study on HS-GNN, in which we methodically removed different components of the model, helped us to isolate and understand the respective contributions to the overall performance. We examined two ablation setups containing partial components from the new HS-GNN models for our benchmark dataset of CNT-bundles to evaluate the sensitivity. In the first setup, we removed the topological and geometric information from our HS-GNN model to study the influence of hierarchies in the architecture (Figure 3). The prediction error for test datasets for strength and modulus then approximately doubled from 3.4% and 5.6% to 7.4% and 14.6% (Table 1). In the second setup, we used a non-hierarchical GNN, but including topological and geometric information as node features to account for the effect of spatial information in the neural network model. The prediction error for test datasets for strength and modulus increased to 10.7% and 19.5%, respectively, which is more than triple (Table 1). Accordingly, the prediction error significantly decreases by adding both hierarchical designs for long-range interactions as well as local topological and geometric information. Among the two, the impact of hierarchies was more pronounced. In relation to a vanilla GCN without either added feature, the two strategies are complementary and drastically reduce the prediction error from up to 100% to less than 5% (Figure 4c, d).

To evaluate the influence of the size of the dataset, we randomly choose smaller fractions of the dataset of CNT bundles for training and prediction using a stratified sampling split (Figure S6 and Section S8 in SI Appendix). The mean relative errors (MREs) of prediction in mechanical properties remained low for XGBoost, increasing to only ~7% at ¼ of the dataset size. HS-GNN showed no significant changes down to ¾ of the dataset size, followed by major increases in MREs to 15% at ½ dataset size and over 50% at ¼ dataset size. Low MREs over all structures in the prediction set for XGBoost, however, are paired with high MREs in areas with low data density even at 90% dataset size (Figure 4c, d and Figure S5 in the SI Appendix). XGBoost is powerful in areas with relatively few and well distributed training data, whereas HS-GNN has steady performance in all areas of uneven data distributions and requires more training data. The entire

dataset of CNT bundles is large enough to enable accurate predictions of mechanical properties and can be further expanded for future use.

### 2.5. Prediction of Mechanical Properties of Carbon Nanostructures Outside the Training Distribution

We challenged the ML models with structurally more complex configurations to explore the capability to adapt to new, unseen, and out of distribution data, including large CNT bundles, cross-sections of carbon fiber, and CNT junctions (Table 1 and Figures 5 and 6). Specifically, we predicted strength and modulus for 20 larger CNT bundles, 50 carbon-fiber cross-sections, and a dataset of 958 unique carbon-nanotube junctions. Predictions were made from the pretrained HS-GNN framework using training data from 1139 small CNT bundles, i.e., we excluded training data for large CNT bundles or carbon fiber-like structures. Predictions for CNT junctions involved separate training and splitting of a new dataset.

20 out-of-distribution models of large CNT bundles were inspired by TEM images and tomography of cross-sections of CNT yarn and carbon fiber(12, 13, 16, 63, 82) and consisted of more than 20,000 atoms, which is several times larger than in the training set (Figure 5a and Figure 6a-e). Predicted mechanical properties have typical deviations under 20% and the HS-GNN remains 2-3 times more accurate than other structure-based methods, as well as better than feature-based methods such as XGBoost (Figure 5a and Figure 6j). The first example structure was modelled after high resolution TEM images of cross-sections of highly aligned CNT networks (Figure 6a, j).(12) Typical dog-bone formations of close-packed, flattened CNTs were included, as observed for CNTs of diameters larger than 4 nm in microscopy.(82) The models follow the experimental size of diameters of 10 to 12 nm(82) and are 10x larger in comparison to typical training structures. The ground truth mechanical data are from reactive simulations with IFF-R, supported by prior validation of structure, shape, and interaction energies of flattened CNTs against experiments and DFT-D2.(63) ML predictions (HS-GNN) display 4.0% deviation in strength and 7.5% deviation in modulus (Figure 6j). A second example was based on the TEM images of a cross-section of a bundle of double-wall CNTs with hexagonal packing and diameters of 1.45 nm (Figure 6b, j).(16) The structure contained 64 CNTs in the bundle with 78,080 atoms, which is about 16 times larger than average structures in the training set. The HS-GNN predicts ~3%

deviation in strength and ~20% deviation in modulus relative to the IFF-R reference value. The data suggest that the size of the pristine model CNT structures could have more influence on the modulus than on the strength. As a third example, compressed and deformed bundles of over 64 CNTs with numerous fracture defects were tested, which align with recent imaging of dislocations(83) and failure processes in CNTs (Figure 6c, d, j).(84) Deviations of 20% to 25% are found in the ML-predicted mechanical properties, showing that large morphologies with multiple fracture sites may reduce the accuracy of HS-GNN predictions more than other features and benefit from further training data. As a fifth example, a similar morphology without fracture defects (Figure 6e)(10) shows low deviations in the predicted tensile strength and modulus by HS-GNN of 5% and 6%, respectively, relative to IFF-R MD simulations (Figure 6e, j). The computed values for tensile strength of ~75 GPa correlate well with available experimental data of ~80 GPa for comparable CNT forests.(10) Overall, HS-GNN generalizes well for large CNT bundles including flattened CNTs. Little dependence on the larger size of the structures was noticed when nanotube had diameters within the typical training range (0.8 nm to 3 nm) and few fracture-type defects were present. Deviations could be reduced by expanding the training set for large structures with complex fracture defects, large CNT diameters, and lower density.

We tested an out-of-distribution dataset of carbon fiber cross-sections comprising 50 structures of about 20 nm in size. The carbon fiber cross-sections feature intricately curved, cut, and crosslinked graphitic layers, as well as discontinuity defects that are common in fiber morphology and influence mechanical performance (Figure 6f, g).(85) A certain fraction of covalent links between graphitic layers was included, which are known to result from high-temperature graphitization.(13, 86) The creation of 3D atomistic models followed the protocol by Shi et al. (see Methods)(86) and stress-strain curves were computed using IFF-R MD (Figure S7 in the SI Appendix). The prediction of mechanical properties using HS-GNN shows 10% to 20% mean relative error compared to IFF-R MD simulations, which is slightly lower compared to large CNT bundles (Figure 5b, Figure 6f, g, j). XGBoost clearly underperforms and other ML methods have 2 to 3 times larger error than the HS-GNN (Figure 5b). For example, a model structure inspired by high resolution TEM data from Chae et. al.(85) illustrates the unique morphology of irregularly arranged graphitic layers and the translation into 3D atomic models (Figure 6f). The reliability in predicted strength and modulus by the HS-GNN in relation to IFF-R MD was 11% for both quantities (Figure 6j). A second example shows a carbon-fiber cross-section of different

morphology, defects, and distribution of cross-links (Figure 6f). In this instance, deviations were higher in strength (29%) and low in modulus (10%) (Figure 6j). Overall, good predictions for carbon fiber cross-sections demonstrate the robustness and scalability of HS-GNN for carbon nanostructures upon changes in morphology and size (Section S9 in the SI Appendix). Including fiber-like morphologies in the training set and expanding the training set beyond hundreds could drastically reduce current prediction errors (Figure 5b).

Furthermore, we utilized the existing dataset of CNT junctions created by Varshney et al.(70) that includes 958 unique morphologies and mechanical properties (Figure 6h-j). The structures are qualitatively different from CNT bundles and carbon fiber and have potential applications in electronics and thermal management. The structures differ in nanotube size and fusion morphology, including defects near the junctions, as illustrated by two examples (Figure 6h, i and Figure S1j-l in the SI Appendix). A separate HS-GNN model was trained on this dataset and shows mean relative errors of predicted strength and modulus of 11.8% and 8.3% relative to MD simulations, respectively (Table 1 as well as Table S1, Figure S8, and Section S10 in the SI Appendix). Some prediction errors are below 5% (Figure 6j).

In summary, the advanced ML techniques can characterize the structure and associated mechanical properties of out-of-distribution carbon nanostructures on the scale of 1 to 50 nm, including CNT bundles, graphitic layers, defective, random, cross-linked morphologies, as well as CNT junctions, with limited training data. The nanostructures function as building blocks for macroscopic CNT yarn, CNT paper, carbon fiber, and nano-electronic devices.(14) The properties of the associated hierarchical structures critically depend on the multi-scale assembly of nanoscale building blocks via grain boundaries, voids, and macro-defects. The disruptive speed and scalability of ML-driven predictions in comparison to traditional MD simulations (Table 1) open opportunities for integration of information on the building blocks into macroscale modeling and experimental workflows.

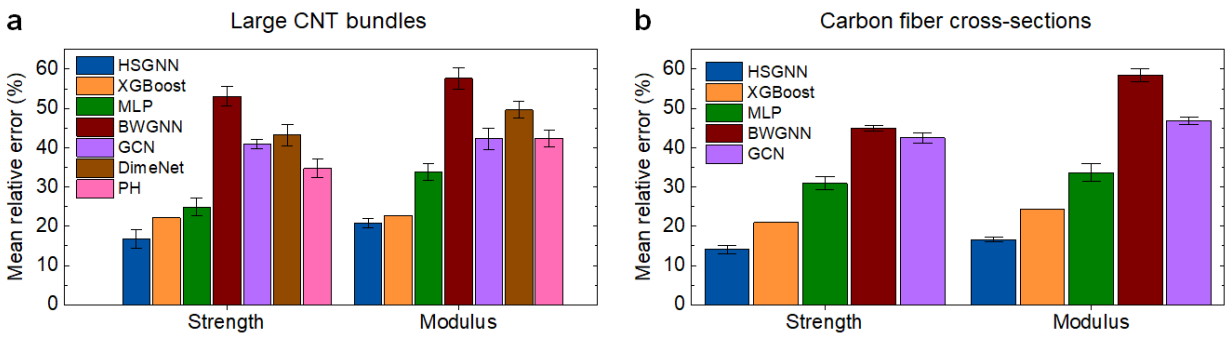


**Figure 5.** Accuracy of predicted mechanical properties for graphitic structures outside the training range with up to ~20 nm size. Examples include large CNT bundles and carbon fiber cross-sections. The mean relative error (MRE) in predicted tensile strength and tensile modulus is shown using various ML methods and the training data set of 1139 (small) CNT bundles. (a) Average prediction error for the dataset of 20 large CNT bundles and (b) for the dataset of 50 carbon-fiber cross-sections. HS-GNN outperforms other methods with average errors between 10% and 20%.

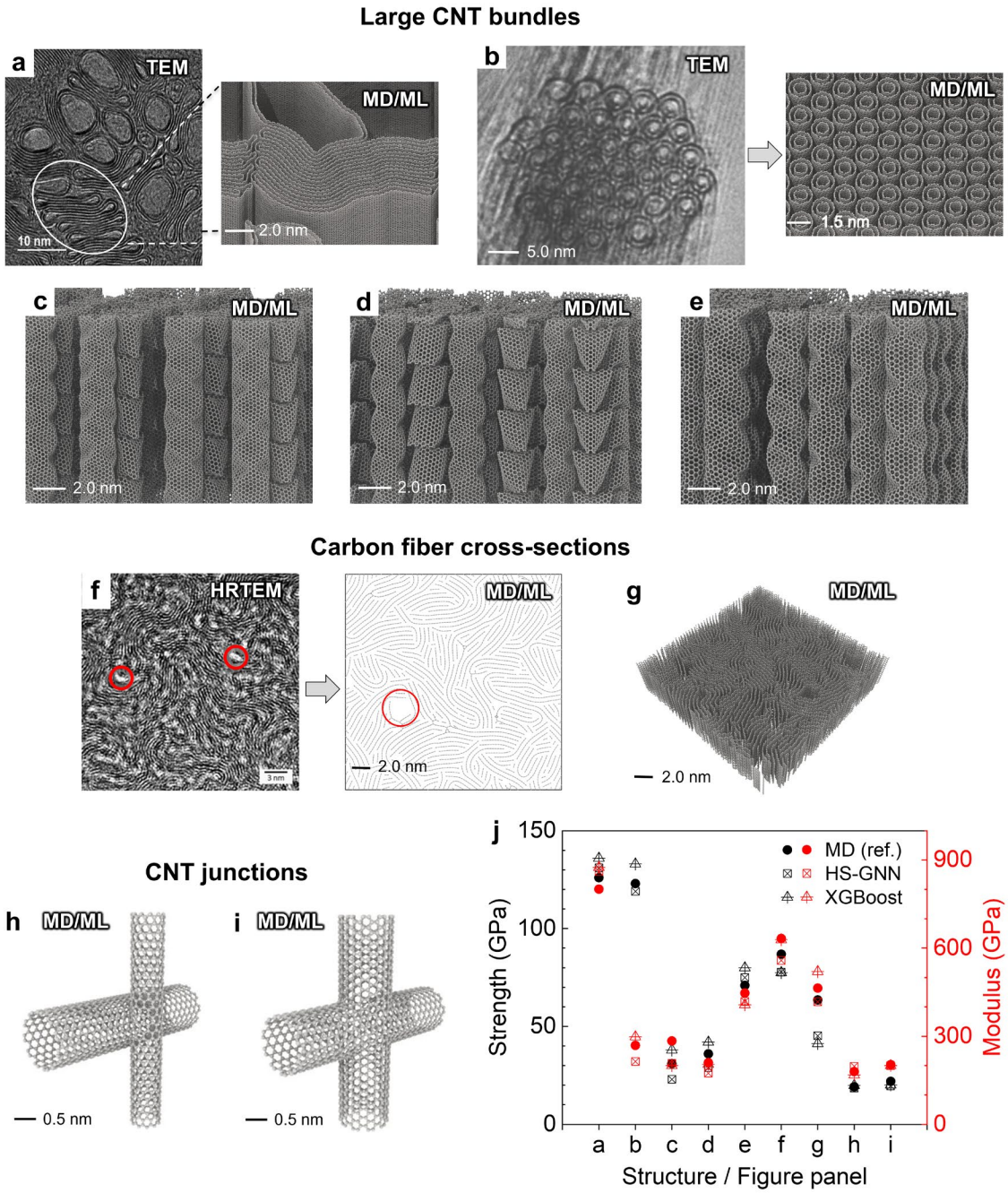


**Figure 6.** Application of ML predictions to larger CNT-based carbon nanostructures and carbon fiber cross-sections, as well as CNT junctions of different geometry and bonding patterns. (a) Transmission electron microscopy (TEM) image of a cross-section of a CNT composite(12) (left) and the corresponding model of flattened DWCNTs for MD simulation and ML prediction (right). (b) Hexagonal packing of a DWCNT bundle observed by TEM(16) (left) and corresponding models for MD/ML prediction of mechanical properties (right). (c-e) Defective structures derived

from compression of SWCNT bundles upon reactive MD simulation. The models include SWCNT bundles of 72 CNTs (c) and 64 CNTs (d), each containing some nanotubes with fracture defects, as well as 64 SWCNTs without defects (e), similar as in ref. (10). (f) High-resolution transmission electron microscopy (HRTEM) image of a carbon fiber cross-section(85) (left) and the model for MD/ML input (right). Red circles represent defective areas and inhomogeneities that result in weaker fibers. (g) A further model of carbon fiber with a large cross-sectional area of 308 $nm^2$ for MD/ML predictions of mechanical properties. (h) Example of a CNT junction. The vertical CNT has a diameter of 1.0 nm and a CNT with a diameter of 1.5 nm was grafted perpendicular to its axis with minimal overlap. (i) A CNT junction of 2 CNTs with 1.5 nm diameter joined with full overlap. (j) Tensile strength and modulus according to IFF-R simulations and predictions using HS-GNN for each of the structures shown in panels (a) to (i). Black symbols refer to strength and red symbols to modulus. Solid spheres indicate the results of MD calculations of stress-strain curves using IFF-R. Open shapes with a cross are ML predictions using HS-GNN (square) or XGBoost (triangle).

### 2.6. Choice of ML Methods, Training Data, and Broader Application

An interesting aspect of the results is that HS-GNN provides a suitable level of accuracy for in-distribution and out-of-distribution data, while XGBoost excels with evenly spread training data inside a distribution that can be relatively sparse. More details on the choice of ML methods and selection of the “right” training data is given in Section S11 in the SI Appendix. Our results also indicate that the heterogeneous graph model, HS-GNN, is a general and powerful methodology (Figure 3). We envision extensions to composites and catalysts (Figure 1a),(1-4) which contain polymers and heteroatoms in addition to carbon nanostructures, by adding new elemental identities and electrostatic interactions to accommodate the added molecular components. Specifics for such broader applications, scalability to larger structures, and limitations are described in Section S12 in the SI Appendix.

## 3. Discussion

We introduce a model database and a machine learning pipeline for accurate and fast predictions of mechanical properties of carbon nanostructures including CNT bundles, cross-sections of CNT yarns, graphitic morphologies in carbon-fiber, and CNT junctions. We utilized experimental data from microscopy, tomography, and other measurements to create a database of 1139 unique morphologies of CNT bundles of up to 10,000 atoms size. Reactive molecular dynamics (MD) simulations of stress-strain curves up to failure with the IFF-R force field were employed to obtain accurate stress-strain curves for the diverse structures in the database and to analyze geometric features that determine the tensile strength and modulus. We developed a hierarchical spatial graph neural network (HS-GNN) that predicts the tensile strength and the tensile modulus of the computer-generated 3D atomistic models of pristine and defective carbon nanotube bundles 1,000 times to 10,000 times faster than MD simulations. The HS-GNN involves persistent summaries with nearest neighbor information and three levels of structural hierarchy to process physical geometric information. Both nearest neighbor information and hierarchy were necessary to achieve high accuracy and outperform other methods by up to two orders of magnitude when operating with sparse data and geometries outside the training range. Predictions of the tensile strength and modulus are within 3% to 6% deviation from the MD training data, which is similar to the typical uncertainty in laboratory experiments. The closest alternative was Extreme Gradient Boosted Trees (XGBoost), which use simpler global features but develop up to ten times higher errors in regions with sparse training data. Both HS-GNN and XGBoost screen mechanical properties of CNT structures 3 or 4 orders of magnitude faster than common molecular dynamics (MD) simulations, respectively, after the networks are trained. Predictions could become even more accurate as the training database expands. Several other structure-based and feature based machine learning methods such as DimeNet, vanilla GCNs, MLP, BWGNN, and Linear Regression are up to an order of magnitude less reliable.

HS-GNN performed well for larger structures and geometries outside the training range, which included 20 large CNT bundles and 50 carbon fiber cross-sections of up to 80,000 nm size with a wide variety of defects, as well as CNT junctions. The deviation from the ground truth by MD simulations was between 10% and 20% while XGBoost and other ML methods scored distinctly lower. The results provide insights into the role of over ten common defect types, alignment of the

graphitic layers, cross-links, partial fracture, irregular deformation, and nanostructure size on mechanical properties.

The database and algorithms can provide guidance towards designing lightweight and ultrastrong carbon nanostructures and composites for aerospace and automotive applications. Quantitative structure-property predictions are feasible in real time (sub-seconds) in near-experimental accuracy towards integration with synthesis, imaging, and generative AI tools. Extensions of the architecture of HS-GNN and XGBoost to larger structures and other properties such as conductivity and catalytic activity are feasible, as well as for other chemistries such as graphene oxide, nitrogen-doped graphitic structures, boron nitride, and polymer composites. The database is openly available and can be extended using multi-modal data from diverse experimental and simulation techniques to reduce residual bias and solve multi-scale problems.

## 4. Methods

The methods are described in detail in Section S13 of the SI Appendix. First, we built the structure database for CNT bundles and graphitic morphologies, chose the force field, and carried out reactive molecular dynamics simulations to create training and validation data. Then, we calculated the mechanical properties for all structures included in the database using reactive molecular dynamics simulations with IFF-R and identified features to inform machine learning of mechanical properties. We developed the HS-GNN network architecture, including the heterogeneous graph representation, hierarchies, and inclusion of spatial information. We utilized the HS-GNN, XGBoost, and other baseline ML methods to train predictions of mechanical properties based on 3D structures, compared to the ground truth, and explored predictions out of the training distribution. We utilized common splitting techniques of data sets and report statistical analyses.

The complete 3D structure databases for training and prediction (~3 GB), the code for HS-GNN, as well as Supporting files with examples of IFF-R reactive molecular dynamics simulations including force field files, run scripts, and 3D structure files are available in the associated figshare database.(87) The code for HS-GNN is also available via GitHub.(88)

## Acknowledgements

The authors acknowledge team discussions as part of the NASA Science and Technology Research Institute, “Ultrastrong Composites by Computational Design”, specifically with Prof. Satish Kumar at Georgia Institute of Technology. We acknowledge support by the National Science Foundation (OAC-1931587, CMMI-1940335, DMREF 2323546), NASA (STRI NNX17AJ32G), as well as by the University of Colorado Boulder. Work by Wang and Zhao was partially supported by NSF under grants OAC-2039794 and CCF-2051197. Work by Xu was partially supported by NSF CMMI-1940107. Unnikrishnan acknowledges partial support of Minority Leaders-Research Collaboration Program (ML-RCP) by the Air Force Research Laboratory under agreement number FA8650-20-2-5853. We further acknowledge the allocation of computational resources at the Argonne Leadership Computing Facility, a DOE Office of Science User Facility supported under contract DE-AC02-06CH11357, the Summit supercomputer supported by the National Science Foundation, ACI-1532235 and ACI-1532236, and High-Performance Computing Resources at the Air Force Research Laboratory.

## References

1. K. M. Li *et al.*, Carbon-Based Fibers: Fabrication, Characterization and Application. *Adv. Fiber Mater.* **4**, 631-682 (2022).
2. Y. Z. Wu *et al.*, Application-Driven Carbon Nanotube Functional Materials. *Acs Nano* **15**, 7946-7974 (2021).
3. L. J. Liu *et al.*, Aligned, High-Density Semiconducting Carbon Nanotube Arrays for High-Performance Electronics. *Science* **368**, 850-856 (2020).
4. A. A. Mamedov *et al.*, Molecular Design of Strong Single-Wall Carbon Nanotube/Polyelectrolyte Multilayer Composites. *Nat. Mater.* **1**, 190-194 (2002).
5. A. Krishnan, E. Dujardin, T. W. Ebbesen, P. N. Yianilos, M. M. J. Treacy, Young's Modulus of Single-walled Nanotubes. *Phys. Rev. B* **58**, 14013-14019 (1998).
6. Z. Shen *et al.*, Theoretical Predictions of Size-dependent Young's and Shear Moduli of Single-walled Carbon Nanotubes. *Physica B* **613**, 412994 (2021).
7. H. Zhan, J. H. Lin, H. L. Shi, J. N. Wang, Construction of carbon nanotubes/bismaleimide composite films with superior tensile strength and toughness. *Compos. Sci. Technol.* **214**, 108975 (2021).
8. C. E. Evers *et al.*, Scalable High Tensile Modulus Composite Laminates Using Continuous Carbon Nanotube Yarns for Aerospace Applications. *ACS Appl. Nano Mater.* **6**, 11260-11268 (2023).
9. A. Jana *et al.*, Atoms to Fibers: Identifying Novel Processing Methods in the Synthesis of Pitch-Based Carbon Fibers. *Sci. Adv.* **8**, eabn1905 (2022).
10. Y. X. Bai *et al.*, Carbon nanotube bundles with tensile strength over 80 GPa. *Nat. Nanotechnol.* **13**, 589-593 (2018).
11. E. R. Meshot *et al.*, High-Speed in Situ X-ray Scattering of Carbon Nanotube Film Nucleation and Self-Organization. *Acs Nano* **6**, 5091-5101 (2012).
12. C. Jolowsky, R. Sweat, J. G. Park, A. Hao, R. Liang, Microstructure Evolution and Self-assembling of CNT Networks During Mechanical Stretching and Mechanical Properties of Highly Aligned CNT Composites. *Compos. Sci. Technol.* **166**, 125-130 (2018).
13. H. B. Chang, J. Luo, P. V. Gulgunje, S. Kumar, Structural and Functional Fibers. *Annu. Rev. Mater. Res.* **47**, 331-359 (2017).

14. D. Nepal *et al.*, Hierarchically Structured Bioinspired Nanocomposites. *Nat. Mater.* **22**, 18-35 (2023).
15. D. G. Papageorgiou, Z. L. Li, M. F. Liu, I. A. Kinloch, R. J. Young, Mechanisms of Mechanical Reinforcement by Graphene and Carbon Nanotubes in Polymer Nanocomposites. *Nanoscale* **12**, 2228-2267 (2020).
16. J. F. Colomer, L. Henrard, G. Van Tendeloo, A. Lucas, P. Lambin, Study of the Packing of Double-walled Carbon Nanotubes into Bundles by Transmission Electron Microscopy and Electron Diffraction. *J. Mater. Chem.* **14**, 603-606 (2004).
17. T. Belin, F. Epron, Characterization Methods of Carbon Nanotubes: A Review. *Mat. Sci. Eng. B-Solid* **119**, 105-118 (2005).
18. F. Banhart, The Formation of a Connection Between Carbon Nanotubes in an Electron Beam. *Nano Lett.* **1**, 329-332 (2001).
19. A. Garley *et al.*, Interaction of Poly(methyl acrylate) with Carbon Nanotubes as a Function of CNT Diameter, Chirality, and Temperature. *J. Phys. Chem. C* **124**, 25632-25644 (2020).
20. C. Pramanik, J. R. Gissinger, S. Kumar, H. Heinz, Carbon Nanotube Dispersion in Solvents and Polymer Solutions: Mechanisms, Assembly, and Preferences. *Acs Nano* **11**, 12805-12816 (2017).
21. J. R. Gissinger, C. Pramanik, B. Newcomb, S. Kumar, H. Heinz, Nanoscale Structure-Property Relationships of Polyacrylonitrile/CNT Composites as a Function of Polymer Crystallinity and CNT Diameter. *ACS Appl. Mater. Interfaces* **10**, 1017-1027 (2018).
22. N. Gupta, J. M. Alred, E. S. Penev, B. I. Yakobson, Universal Strength Scaling in Carbon Nanotube Bundles with Frictional Load Transfer. *Acs Nano* **15**, 1342-1350 (2021).
23. J. M. Alred, K. V. Bets, Y. Xie, B. I. Yakobson, Machine Learning Electron Density in Sulfur Crosslinked Carbon Nanotubes. *Compos. Sci. Tech.* **166**, 3-9 (2018).
24. C. Pramanik *et al.*, Polyacrylonitrile Interactions with Carbon Nanotubes in Solution: Conformations and Binding as a Function of Solvent, Temperature, and Concentration. *Adv. Funct. Mater.* **29**, 1905247(2019).
25. J. J. Winetrout *et al.*, Implementing Reactivity in Molecular Dynamics Simulations with the Interface Force Field (IFF-R) and Other Harmonic Force Fields. http://dx.doi.org/10.48550/arXiv.2107.14418.
26. J. J. Winetrout *et al.*, Implementing Reactivity in Molecular Dynamics Simulations with Harmonic Force Fields. *Nat. Commun.* **15**, 7945 (2024).
27. X. Zhao *et al.*, Smallest Carbon Nanotube Is 3 Å in Diameter. *Phys. Rev. Lett.* **92**, 125502 (2004).
28. L. M. Peng *et al.*, Stability of Carbon Nanotubes: How Small Can They Be? *Phys. Rev. Lett.* **85**, 3249-3252 (2000).
29. G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, A. Fazzio, From DFT to Machine Learning: Recent Approaches to Materials Science-A Review. *J. Phys.-Mater.* **2**, 032001 (2019).
30. D. A. Vecchio, S. H. Mahler, M. D. Hammig, N. A. Kotov, Structural Analysis of Nanoscale Network Materials Using Graph Theory. *Acs Nano* **15**, 12847-12859 (2021).
31. B. Alipanahi, A. Delong, M. T. Weirauch, B. J. Frey, Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning. *Nat. Biotechnol.* **33**, 831-838 (2015).
32. A. Graves, A. R. Mohamed, G. Hinton, Speech Recognition with Deep Recurrent Neural Networks. *IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, 6645 (2013), DOI: 10.1109/ICASSP.2013.6638947.
33. K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, Identity Mappings in Deep Residual Networks. *Lecture Notes in Computer Science* **9908**, 630-645 (2016), DOI: 10.1007/978-3-319-46493-0_38.
34. M. I. Jordan, T. M. Mitchell, Machine Learning: Trends, Perspectives, and Prospects. *Science* **349**, 255-260 (2015).
35. A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks" in Advances in Neural Information Processing Systems. (Curran Associates Inc., Dutchess County, NY, 2012), vol. 25.
36. N. Nguyen *et al.*, DNA Sequence Classification by Convolutional Neural Network. *Journal of Biomedical Science and Engineering* **09**, 280-286 (2016).
37. A. Vaswani *et al.*, Attention Is All You Need. *In Advances in Neural Information Processing Systems (NIPS 2017),* **30** (2017).
38. I. Goodfellow, Y. Bengio, A. Courville, Deep Learning (MIT Press, 2016).
39. Y. LeCun, Y. Bengio, G. Hinton, Deep Learning. *Nature* **521**, 436-444 (2015).

40. C. R. Qi, H. Su, K. C. Mo, L. J. Guibas, PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 77-85 (2017), DOI: 10.1109/CVPR.2017.16.
41. Y. C. Liu, B. Fan, S. M. Xiang, C. H. Pan, Relation-Shape Convolutional Neural Network for Point Cloud Analysis. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8887-8896 (2019), DOI: 10.1109/Cvpr.2019.00910.
42. R. S. DeFever, C. Targonski, S. W. Hall, M. C. Smith, S. Sarupria, A Generalized Deep Learning Approach for Local Structure Identification in Molecular Simulations. *Chem. Sci.* **10**, 7503-7515 (2019).
43. K. Xu, Z. Wang, J. P. Shi, H. S. Li, Q. C. Zhang, A(2)-Net: Molecular Structure Estimation from Cryo-EM Density Volumes. *Proc. AAAI Conf. Artif. Intell.* **33**, 1230-1237 (2019).
44. N. W. A. Gebauer, M. Gastegger, K. T. Schütt (2019) Symmetry-adapted Generation of 3d Point Sets for the Targeted Discovery of Molecules. *NIPS'19: Proc. 33rd Int. Conf. on Neural Information Processing Systems* **680**, 7566 (2019).
45. C. Coley, R. Barzilay, T. Jaakkola, W. Green, K. Jensen, Applying Machine Learning to Synthesis Design: Prediction of Organic Reaction Outcomes. *ACS Cent. Sci.* **3**, 434-443 (2017).
46. D. K. Duvenaud *et al.*, Convolutional Networks on Graphs for Learning Molecular Fingerprints. *ArXiv* Preprint arXiv:1509.09292 (2015).
47. S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput. Aid. Mol. Des.* **30**, 595-608 (2016).
48. T. Kipf, M. Welling, Semi-Supervised Classification with Graph Convolutional Networks. *ArXiv Preprint* arXiv:1609.02907 (2016).
49. P. Velickovic *et al.*, Graph Attention Networks. *ArXiv Preprint* arXiv:1710.10903 (2017).
50. J. Gasteiger, C. Yeshwanth, S. Gunnemann, Directional Message Passing on Molecular Graphs via Synthetic Coordinates. *Advances in Neural Information Processing Systems (NeurIPS)* **34** (2021).
51. K. Yang *et al.*, Analyzing Learned Molecular Representations for Property Prediction (vol 59, pg 3370, 2019). *J. Chem. Inf. Model.* **59**, 5304-5305 (2019).
52. H. Grushevskaya, A. Timoshchenko, I. Lipnevich, Topological Defects Created by Gamma Rays in a Carbon Nanotube Bilayer. *Nanomaterials-Basel* **13**, 410 (2023).
53. A. Ghavamian, M. Rybachuk, A. Ochsner, Defects in Carbon Nanotubes. *Woodhead Publishing Series in Electronic and Oprical Materials*, 87-136 (2018), DOI: 10.1016/B978-0-08-102053-1.00004-1.
54. H. Heinz, T. J. Lin, R. K. Mishra, F. S. Emami, Thermodynamically Consistent Force Fields for the Assembly of Inorganic, Organic, and Biological Nanostructures: The INTERFACE Force Field. *Langmuir* **29**, 1754-1765 (2013).
55. C. Zhu, S. E. Hoff, M. Hémadi, H. Heinz, Accurate and Ultrafast Simulation of Molecular Recognition and Assembly on Metal Surfaces in Four Dimensions. *Acs Nano* **17**, 9938-9952 (2023).
56. C. Pramanik, J. R. Gissinger, S. Kumar, H. Heinz, Carbon Nanotube Dispersion in Solvents and Polymer Solutions: Mechanisms, Assembly, and Preferences. *Acs Nano* **11**, 12805-12816 (2017).
57. K. Kanhaiya *et al.*, Accurate Force Fields for Atomistic Simulations of Oxides, Hydroxides, and Organic Hybrid Materials up to the Micrometer Scale. *J. Chem. Theor. Comput.* **19**, 8293-8322 (2023).
58. S. S. Bamane *et al.*, Boron Nitride Nanotubes: Force Field Parameterization, Epoxy Interactions, and Comparison with Carbon Nanotubes for High-Performance Composite Materials. *ACS Appl. Nano Mater.* **6**, 3513-3524 (2023).
59. J. Liu *et al.*, Understanding Chemical Bonding in Alloys and the Representation in Atomistic Simulations. *J. Phys. Chem. C* **122**, 14996-15009 (2018).
60. W. Z. Li, J. G. Wen, Z. F. Ren, Straight Carbon Nanotube Y Junctions. *Appl. Phys. Lett.* **79**, 1879-1881 (2001).
61. J. M. Ting, C. C. Chang, Multijunction Carbon Nanotube Network. *Appl. Phys. Lett.* **80**, 324-325 (2002).
62. T. Q. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System. *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794 (2016), DOI: 10.1145/2939672.2939785.
63. B. Damirchi *et al.*, ReaxFF Reactive Force Field Study of Polymerization of a Polymer Matrix in a Carbon Nanotube-Composite System. *J. Phys. Chem. C* **124**, 20488-20497 (2020).
64. K. Kanhaiya, S. Kim, W. Im, H. Heinz, Accurate Simulation of Surfaces and Interfaces of Ten FCC Metals and Steel Using Lennard–Jones Potentials. *npj Comput. Mater.* **7**, 17 (2021).

65. J. Liu *et al.*, Interpretable Molecular Models for Molybdenum Disulfide and Insight into Selective Peptide Recognition. *Chem. Sci.* **11**, 8708-8722 (2020).
66. J. Chen *et al.*, Building Two-Dimensional Materials One Row at a Time: Avoiding the Nucleation Barrier. *Science* **362**, 1135-1139 (2018).
67. S. Wang, K. Hou, H. Heinz, Accurate Compatible Force Fields for Molecular Oxygen, Nitrogen and Hydrogen to Simulate Heterogeneous Electrolytes and Interfaces. *J. Chem. Theor. Comput.* **17**, 5198-5213 (2021).
68. Y. K. Choi *et al.*, CHARMM-GUI Nanomaterial Modeler for Modeling and Simulation of Nanomaterial Systems. *J. Chem. Theor. Comput.* **18**, 479-493 (2022).
69. L. Guan, K. Suenaga, S. Iijima, Smallest Carbon Nanotube Assigned with Atomic Resolution Accuracy. *Nano Lett.* **8**, 459-462 (2008).
70. V. Varshney, V. Unnikrishnan, J. Lee, A. K. Roy, Developing Nanotube Junctions with Arbitrary Specifications. *Nanoscale* **10**, 403-415 (2018).
71. S. Bhusal, S. Sihn, V. Varshney, A. K. Roy, A Study on Mechanical Strength and Stability of Partially-Fused Carbon Nanotube Junctions. *Carbon Trends* **3**, 100039 (2021).
72. P. Buhlmann, J. Peters, J. Ernest, CAM: Causal Additive Models, High-Dimensional Order Search and Penalized Regression. *Annals of Statistics* **42**, 2526-2556 (2014).
73. Q. M. Li, Z. C. Han, X. M. Wu, Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. *Thirty-Second Aaai Conference on Artificial Intelligence / Thirtieth Innovative Applications of Artificial Intelligence Conference / Eighth Aaai Symposium on Educational Advances in Artificial Intelligence*, 3538-3545 (2018).
74. J. J. Huang, Z. H. Li, N. N. Li, S. Liu, G. Li, AttPool: Towards Hierarchical Feature Representation in Graph Convolutional Networks via Attention Mechanism. *IEEE/CVF Int. Conf. Comp. Vision (ICCV)*, 6489-6498 (2019), DOI: 10.1109/Iccv.2019.00658.
75. R. Ying *et al.*, Hierarchical Graph Representation Learning with Differentiable Pooling. *Advances in Neural Information Processing Systems (NIPS)* **31**, 4805 (2018), DOI: 10.5555/3327345.3327389.
76. H. Edelsbrunner, J. Harer, *Computational Topology: An Introduction* (American Mathematical Society, 2010).
77. H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological Psistance and Simplification. *Proc. 41st Ann. Symp. Found. Comp. Sci.*, 454-463 (2000), DOI: 10.1109/SFCS.2000.892133.
78. J. H. Tang, J. J. Li, Z. Q. Gao, J. Li, Rethinking Graph Neural Networks for Anomaly Detection. *International Conference on Machine Learning (PMLR), Vol 162*, 21076 (2022).
79. F. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).
80. F. Rosenblatt, The Perceptron - a Probabilistic Model for Information-Storage and Organization in the Brain. *Psychol. Rev.* **65**, 386-408 (1958).
81. P. J. Werbos, Generalization of Backpropagation with Application to a Recurrent Gas Market Model. *Neural Networks* **1**, 339-356 (1988).
82. J. A. Elliott, J. K. W. Sandler, A. H. Windle, R. J. Young, M. S. P. Shaffer, Collapse of single-wall carbon nanotubes is diameter dependent. *Phys. Rev. Lett.* **92**, 095501 (2004).
83. T. Lu, X.-W. Lei, T. Fujii, Nucleation of Disclinations in Carbon Nanotube Bundle Structures Under Twisting Loads. *Carbon* **228**, 119287 (2024).
84. C. Jing, Y. Qin, W. Ouyang, J. Luo, Pseudo-break Imaging of Carbon Nanotubes for Determining Elastic Bending Energies. *Nano Research* **16**, 7443-7451 (2023).
85. H. G. Chae *et al.*, High Strength and High Modulus Carbon Fibers. *Carbon* **93**, 81-87 (2015).
86. L. Y. Shi, M. Sessim, M. R. Tonks, S. R. Phillpot, Generation and Characterization of an Improved Carbon Fiber Model by Molecular Dynamics. *Carbon* **173**, 232-244 (2021).
87. J. J. Winetrout *et al.*, Dataset of Carbon Nanostructures for "Prediction of Carbon Nanostructure Mechanical Properties and the Role of Defects Using Machine Learning" Figshare.com, DOI: 10.6084/m9.figshare.27634290 (2024).
88. HS-GNN Files, https://github.com/hendrikheinz/HS-GNN.