Results in Engineering 32 (2026) 112576

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Research paper

Inverse design of Si/Ge superlattices for targeted phonon filtering using a NEGF-machine learning framework

Julian G. Fernandez [a,b,*], Shubham Tyagi [c], Natalia Seoane [a,b], Antonio Garcia-Loureiro [a,b], Marc Bescond [c,d]

[a] Departamento de Electrónica e Computación, USC, Santiago de Compostela, 15705, Galicia, Spain
[b] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), USC, Santiago de Compostela, 15705, Galicia, Spain
[c] Institut Matériaux Microélectronique Nanosciences de Provence (IM2NP), Aix-Marseille Université, Marseille, UMR CNRS 7334, France
[d] Institute of Industrial Science, University of Tokyo, Tokyo, 4-6-1 Komaba, Meguro-ku, 153-8505, Japan

ARTICLE INFO

Keywords:
Si/Ge superlattice
Inverse design
Phononic engineering
Thermal conductivity
Stop-band filtering
Active cooling
NEGF
Machine learning

ABSTRACT

This work presents a novel inverse design methodology that combines accurate non-equilibrium Green's function (NEGF) simulations with machine learning to engineer Si/Ge superlattices with tailored phononic properties. A multi-layer perceptron architecture (MLP) is trained to predict both the integrated thermal conductivity κ and the full phonon transmission spectra $T(\nu)$ from geometric descriptors, using a hybrid feature representation that combines statistical descriptors and principal component analysis (PCA)-compressed profiles. The model achieves excellent predictive accuracy, with a coefficient of determination of $R^2 \sim 0.99$ for κ and a root mean square error of ~ 0.013 a.u. for $T(\nu)$. The trained model enables rapid screening of 10^5 different structures, identifying optimal configurations that achieve a 10.20% reduction in κ compared to the best configurations in the original dataset. As a proof of concept, the methodology is applied to design a superlattice that selectively suppresses phonon transmission in the 15-36.2 meV (3.63-8.75 THz) frequency window, the characteristic emission range of the EL2 recombination center in GaAs, thereby demonstrating frequency-targeted thermal management at the source. In this use case, the optimized structure achieves a 38.84% reduction in transmission within the targeted frequency window. The proposed workflow is general and can be adapted to any electronic device where specific phonon frequency windows contribute to localized heating, offering a versatile pathway for integrated, at-source thermal management.

1. Motivation

The continuous drive for miniaturization and the increased density of semiconductor based devices in integrated circuits has led to a crisis in thermal management [1]. In high-power electronics and laser power converters, the high operational power densities similarly lead to detrimental local hot spots [2,3]. These hot spots reduce performance, accelerate material degradation, and compromise the reliability and operational lifespan of electronic systems [4,5]. Effectively mitigating these self-induced thermal effects is therefore a cornerstone of advancing semiconductor technology [4,5]. Conventional thermal solutions, such as bulk liquid cooling or forced air convection, operate at the system or package level. These macroscale approaches are inherently inefficient at targeting microscopic hot spots and contribute significantly to the overall energy footprint. Notably, cooling can account for approximately 40 % of a data center's power consumption [6,7]. This underscores the critical need for innovative, in-system cooling paradigms capable of extracting heat directly at its source within the chip architecture [8–10]

Solid-state solutions based on semiconductor physics, such as asymmetric double-barrier heterostructures or superlattices, offer a promising avenue for such integrated thermal management [11–13]. Superlattice heterostructures of alternating material layers provide a highly customizable platform for phonon engineering. By tailoring the layer sequence, thickness, and interface properties, the phonon dispersion and scattering mechanisms can be precisely manipulated, enabling direct control of the thermal conductivity (κ), which is defined as the rate of heat flow per unit temperature difference across the device. Silicon-Germanium (Si/Ge) superlattices are of particular interest due to the material compatibility with mainstream silicon technology and the significant acoustic mismatch between Si and Ge, which provides a strong lever for influencing phonon transport [14].

However, the inverse design of optimal Si/Ge superlattice structures for a target thermal property presents a formidable challenge.

* Corresponding author at: Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), USC, Santiago de Compostela, 15705, Galicia, Spain.
E-mail address: julian.garcia.fernandez2@usc.es (J.G. Fernandez).

https://doi.org/10.1016/j.rineng.2026.112576
Received 22 May 2026; Received in revised form 4 August 2026; Accepted 20 August 2026
Available online 22 August 2026

The design space, defined by parameters such as individual layer thicknesses, number of interfaces, and periodicity conditions, is vast and high-dimensional. The underlying principle relies on selectively filtering detrimental phonon modes by engineering the phononic band structure. Prior studies have shown that phonon transport in Si/Ge multilayers arises from a competition between coherent interference and incoherent scattering mechanisms, highlighting how layer ordering and disorder strongly modulate transmission spectra and thermal conductance [15]. While tuning periodicity and acoustic contrast can create phononic band gaps or suppress specific frequency (ν) ranges, the relationship between structural parameters and thermal conductivity is highly non-linear, governed by complex wave interference and scattering phenomena. Consequently, a trial-and-error optimization reliant on computationally intensive phonon transport simulations, such as methods based on the non-equilibrium Green's function (NEGF) formalism, becomes prohibitively costly and time-consuming for navigating this intricate design space.

This bottleneck highlights the necessity of exploring sophisticated algorithmic or machine-learning (ML) based strategies to guide the design process of these customizable superlattices. ML-based design of aperiodic Si/Ge superlattices has been previously explored using big-data pattern analysis to identify low-conductivity configurations [16], and molecular dynamics ML approaches have also been used to optimize periodic, graded, and aperiodic multilayers [17]. In contrast to these conductivity focused studies, the present work combines NEGF-based coherent phonon transport with ML to predict full frequency resolved transmission spectra. This enables an inverse-design workflow in which the superlattice geometry is optimized to suppress a specific phonon-emission window (15–36.2 meV for EL2 in GaAs), rather than simply minimizing overall thermal conduction. Motivated by the successful application of ML in related nanoelectronic domains [18–21], we propose a computational methodology that leverages NEGF and ML to efficiently explore the design space and identify spectrally selective superlattice configurations.

While the present study focuses on optimizing Si/Ge superlattices, the core inverse design strategy is fundamentally general. It can be directly applied to other material systems to mitigate heating from specific non-radiative recombination processes. To illustrate this critical application, we target the EL2 center (arsenic antisite, As_{Ga}) in gallium arsenide (GaAs), a dominant non-radiative recombination center and a major reliability challenge in GaAs-based high-frequency electronics, optoelectronic devices and laser power converters [22–24]. Its non-radiative recombination process releases a characteristic spectra of high-energy phonons that contribute directly to localized heating, efficiency loss, and long-term degradation. A superlattice engineered to create a phononic stop-band overlapping this specific spectral signature would provide integrated, at-source thermal management for GaAs-based devices. The inverse design workflow demonstrated here, from defining a target phonon spectra to inversely determining the optimal superlattice geometry, is directly transferable to similar challenges in other semiconductor platforms, requiring only the relevant material parameters and the target phonon frequency range as inputs.

The remainder of this work is organized as follows. Section 1.1 describes the Si/Ge superlattice device, its key geometric parameters, and the physical principles governing thermal transport in these structures. Section 1.2 introduces the EL2 recombination center in GaAs as a relevant use case, detailing its non-radiative recombination mechanism and establishing the target frequency window for phonon filtering. Section 2 presents the computational methodology, including the NEGF simulations, dataset generation, feature engineering, and the multi-layer perceptron (MLP) architecture employed. Section 3 presents the main results and their discussion, organized as follows: Section 3.1 validates the ML workflow against NEGF simulations; Section 3.2 applies the methodology to identify superlattice configurations achieving extreme thermal conductivity values and demonstrates the inverse-design for active cooling of the proposed superlattices by filtering the targeted phonon window. Finally, Section 4 summarizes the key findings of this work.

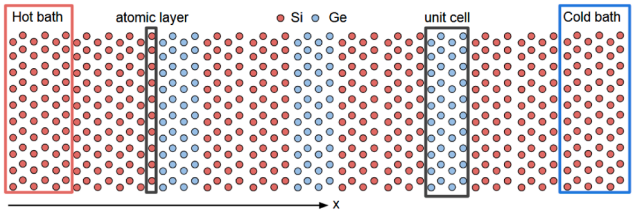

Fig. 1. Schematic of a superlattice with Silicon (Si) and Germanium (Ge) atomic layers. Note that, the contacts are made of Si, corresponding to a hot bath and cold bath contact. The superlattice unit cell (uc) is defined as 4 at. layers.

1.1. Si/Ge superlattice

The inverse design methodology presented in this work is developed and validated using a Si/Ge superlattice. As illustrated in Fig. 1, these semiconductor-based devices consist of a quasi-one-dimensional succession of alternating Si and Ge atomic layers, which selectively transmit or block phonons propagating from a hot Si contact to a cold Si contact, thereby extracting heat from an electronic device. Their phononic response—specifically, the transmission and absorption spectra of thermal phonons—is governed by a set of key geometric parameters: the number of unit cells (uc, defined as the concatenation of 2 at. layers of the same material), the number of interfaces, the total number of layers of the device and the specific sequence and periodicity of the materials. By engineering these parameters, the superlattice's phononic band structure can be tailored to create band gaps or suppress targeted frequency modes, thereby enabling precise control over κ. Recent work has shown that phonon coherence depends not only on layer ordering but also on the total number of layers. Maranets et al. [25] demonstrated that interference features strengthen with increasing device length, underscoring the relevance of finite-size effects and motivating the fixed-length constraint adopted here.

This Si/Ge system serves as an ideal device for several reasons: (i) the significant acoustic impedance contrast between Si and Ge provides a strong mechanism for phonon scattering; (ii) both materials are technologically mature, with well-characterized properties; and (iii) the system's complexity is representative of the broader challenge of designing phononic crystals for thermal management [26,27]. To comprehensively assess the utility of the inverse design methodology, we apply it to two distinct design objectives, both within the constraint of a fixed total device length L_{tot} and guided by practical experimental limits. First, we identify superlattice configurations that achieve the theoretical minimum κ. Second, we move beyond scalar conductivity to perform frequency-targeted phonon engineering. Here, the objective is to select the optimal superlattice geometry that minimizes the contribution of the EL2 recombination centers by designing a phononic stop-band to suppress its specific, detrimental phonon emission window.

1.2. EL2 recombination centers in GaAs

The relevance of the EL2 center as a use case for our inverse design methodology lies in its nature as a primary phonon emission source in GaAs-based applications. The EL2 defect is identified as a deep-level arsenic antisite (As_{Ga}), where an arsenic atom occupies a gallium lattice site. This defect introduces a donor level within the mid-gap, typically located at $E_C - 0.75$ eV (where E_C is the conduction energy band) [28,29]. Native to GaAs, EL2 acts as the dominant non-radiative recombination center, significantly impacting the thermal and electrical performance of high-frequency electronic and optoelectronic devices [30,31].

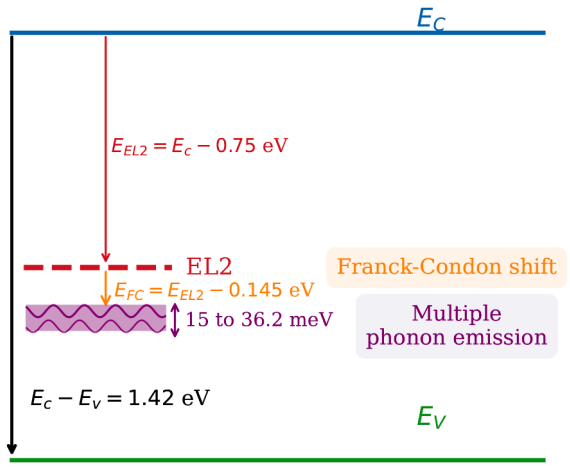

Fig. 2. Schematic representation of non-radiative recombination at the EL2 center in GaAs. The diagram illustrates the conduction band (E_C), valence band (E_V), and the trap level at $E_{EL2} \approx E_C - 0.75$ eV. Following the Franck-Condon principle, lattice relaxation ($E_{FC} \approx E_{EL2} - 0.145$ eV) triggers the emission of a multiple phonon cascade. This energy release, primarily concentrated in the 15-36.2 meV (3.63-8.75 THz) range, encompasses both optical phonons and high-energy acoustic modes. This cascade constitutes the primary source of localized hot spots in GaAs-based devices and defines the target spectral window for our superlattice design.

During the capture of emission of charge carriers at an EL2 site, energy is dissipated through the emission of multiple phonons in a process governed by the Franck-Condon principle and described by multiphonon-assisted tunneling models [32,33]. Electron capture triggers a significant lattice relaxation toward a new equilibrium configuration. The associated Franck-Condon shift, estimated at approximately 145 meV, defines the total energy scale of the resulting phonon cascade [33,34]. The spectral distribution of these emitted phonons is dictated by the interaction between the defect relaxation and the GaAs phonon density of states.

Two distinct spectral regimes are identified for thermal management. The first is a high-energy regime (from 15 to 36.2 meV), corresponding to frequencies between 3.63 and 8.75 THz. This window encompasses the longitudinal and transverse optical (LO/TO) modes of GaAs, which are characterized by low group velocities and lead to the formation of localized hot-spots [35]. Additionally, this range includes high-energy acoustic phonons (from 15 to 25 meV) generated during the initial stages of the relaxation cascade. These phonons exhibit a reduced mean free path as they are highly susceptible to Umklapp scattering, which hinders efficient heat transport away from the defect. In contrast, the low-energy acoustic regime (below 15 meV) consists of the primary heat carriers in GaAs. These modes possess high group velocities and long mean free paths, enabling macroscopic heat dissipation [36].

Consequently, the frequency window for phonon filtering is established between 3.63 and 8.75 THz (15 to 36.2 meV). This target range captures the characteristic phonon emission of the EL2 center while preserving the low-frequency acoustic modes essential for device cooling. The Si/Ge superlattice system employed in this work exhibits a transmission cutoff of up to 16 THz, ensuring that the EL2 target range falls entirely within the tunable window of the material. This provides the necessary spectral flexibility to engineer phononic band structures capable of suppressing harmful frequencies through targeted interference and scattering. A schematic representation of the EL2 emission process and the corresponding energy levels is provided in Fig. 2.

2. Methodology

The combined NEGF and ML methodology developed in this work is summarized in the workflow diagram of Fig. 3. The approach consists

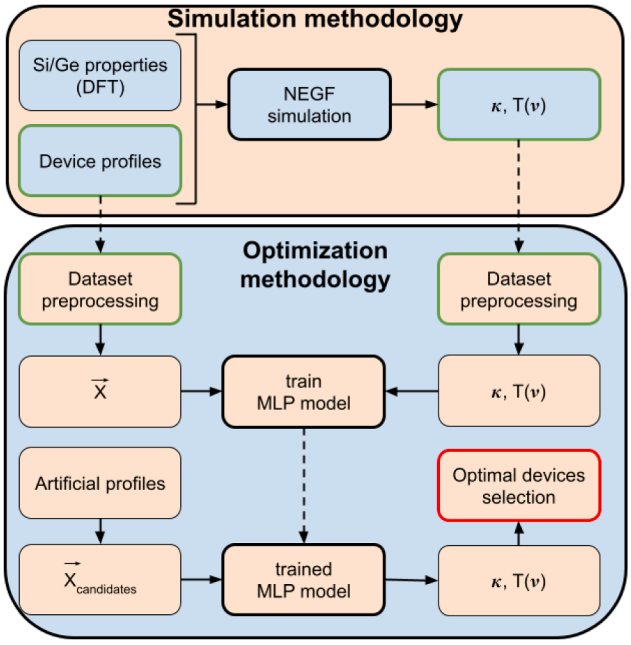

Fig. 3. Schematic overview of the combined NEGF + ML methodology. The workflow comprises two sequential phases: (1) NEGF data generation from Si/Ge material properties and device profiles; (2) model training and optimization for targeted thermal management objectives, with final selection of optimal devices.

of two sequential phases designed to efficiently navigate the vast design space of Si/Ge superlattices while maintaining physical accuracy.

The first phase focuses on generating a reliable dataset through physics-based simulations. Starting from density functional theory (DFT) calculations of Si, a set of device profiles is defined by varying key geometric parameters. NEGF simulations are then performed to compute the thermal conductivity (κ) and the frequency-dependent phonon transmission ($T(\nu)$) for each configuration. This phase provides the initial dataset for the subsequent optimization stage.

The second phase, illustrated in Fig. 3, leverages this dataset to train a model based on a MLP. As depicted in the figure, a preprocessing stage (explained in Section 2.2) is first applied to transform the raw device profiles into the input tensor $\vec{X}$, while the simulation outputs are independently preprocessed to obtain the target properties κ and $T(\nu)$ for training the MLP. Once trained, the model enables rapid prediction of device performance, allowing the exploration of a large number of candidate structures (referred to as artificial profiles in Fig. 3) at minimal computational cost. These artificial profiles are processed to generate the candidates tensor ($\vec{X}_{candidates}$) and used as input into the trained MLP to obtain their respective κ and $T(\nu)$. The optimal device selection then identifies superlattice configurations that meet specific thermal management objectives.

The following subsections describe each phase in greater detail, starting with the simulation methodology and then the optimization methodology.

2.1. Simulation methodology

For the evaluation of material properties of Si, first-principles calculations are carried out within the framework of DFT using the generalized gradient approximation with the Perdew– Burke–Ernzerhof exchange–correlation functional, as implemented in the Quantum ESPRESSO package [37]. A plane-wave basis set with an energy cutoff of 60 Ry is employed. Brillouin zone integrations are performed using a Monkhorst–Pack $8 \times 8 \times 8$ k-point grid. The total energy convergence threshold is set to 10^{-8} Ry, while structural optimization proceeds

until the Hellmann–Feynman forces on each atom are reduced below 10^{-6} Ry/bohr. Harmonic interatomic force constants are obtained using the finite-displacement approach as implemented in Phonopy [38]. For this purpose, a $3 \times 3 \times 3$ supercell is constructed, and the corresponding Brillouin zone is sampled with a $3 \times 3 \times 3$ Monkhorst–Pack k-point mesh. Consistent with prior works, the harmonic force constants of Ge are assumed to be identical to those of Si, with the only distinction arising from the difference in atomic masses [39–41]. Note that using a common harmonic force-constant model for Si and Ge is a controlled approximation. Prior studies of Si/Ge vibrational properties have shown that reduced models reproduce the phonon dispersion and are consistent with experimental benchmarks for bulk Si, Ge, and Si/Ge heterostructures [40,42,43]. However, the mass approximation may overestimate phonon transmission across interfaces, particularly for high-frequency phonons, because it neglects local force-constant disorder arising from alloying and strain in real systems. Since we focus here on clean interfaces, it is expected that the spectral features relevant to the filtering behavior discussed in this work will be retained.

Thermal conductance (σ) and the corresponding thermal conductivity (κ) are computed using the non-equilibrium Green's function (NEGF) formalism within an in-house developed framework [40,41]. The harmonic force constants obtained from Quantum ESPRESSO and Phonopy serve as input for the transport calculations. Phonon transport through the device region is described by the retarded Green's function:

$$G^R(\omega; q_t) = \left[\omega^2 I - \tilde{\Phi}(q_t) - \Sigma^R(\omega; q_t)\right]^{-1}, \tag{1}$$

where $\omega = 2\pi\nu$ denotes the phonon angular frequency, I is the identity matrix, and q_t represents the transverse wave vector. The matrix $\tilde{\Phi}(q_t)$ corresponds to the Fourier-transformed harmonic force constant matrix, while Σ^R captures the interaction between the device and the semi-infinite contacts [44,45]. The total retarded self-energy is expressed as the sum of the contributions from the left and right contacts,

$$\Sigma^R(\omega; q_t) = \Sigma_1^R(\omega; q_t) + \Sigma_2^R(\omega; q_t), \tag{2}$$

which are evaluated using the surface Green's functions of the contacts via the decimation method [46].

The greater and lesser Green's functions are evaluated as [45,47]:

$$G^{>,<}(\omega; q_t) = G^R(\omega; q_t)\, \Sigma^{>,<}(\omega; q_t)\, G^A(\omega; q_t), \tag{3}$$

where the advanced Green's function G^A is given by the Hermitian conjugate of G^R. The corresponding self-energies are written as:

$$\Sigma^{>,<}(\omega; q_t) = \Sigma_1^{>,<}(\omega; q_t) + \Sigma_2^{>,<}(\omega; q_t), \tag{4}$$

and are determined by the Bose–Einstein occupation of the contacts together with their retarded self-energies.

The σ is then obtained using the Landauer formula [48] as:

$$\sigma = \frac{1}{A} \int \frac{\hbar\omega}{N} \sum_{q_t} \mathrm{Tr}\left[\Gamma_1(\omega; q_t)\, G^R(\omega; q_t)\, \Gamma_2(\omega; q_t)\, G^A(\omega; q_t)\right] \times \frac{\partial f_{\mathrm{BE}}(\omega, T)}{\partial T}\, d\omega, \tag{5}$$

where A is the transverse cross-sectional area, f_{BE} denotes the Bose–Einstein distribution function, Tr represents the trace of a square matrix, and N is the number of transverse wave vectors. The coupling matrices $\Gamma_{1(2)}$ are defined as $\Gamma_{1(2)} = i\left(\Sigma^R_{1(2)} - \Sigma^A_{1(2)}\right)$ where the advanced contact self-energy Σ^A is the Hermitian conjugate of the retarded self-energy Σ^R. The κ is then derived from σ using device length L as $\kappa = \sigma$L. The phonon density is calculated using the diagonal of $G^>$ as [49]:

$$\rho(\omega, R_n) = \mathrm{Tr}\left[\frac{1}{N} \sum_{q_t} i G^<_{nn}(\omega; q_t)\right] \times \frac{\omega}{\pi}, \tag{6}$$

where R_n denotes the position of atomic site n.

A discretization of 200 frequency points is used for energy integration, while the transverse Brillouin zone is sampled using a 15×15 q_t-point grid, as verified through convergence analysis. Although our calculations are performed at 200 K to suppress anharmonic scattering and isolate coherent transport effects, we expect the qualitative trends reported here to persist at higher temperatures. Prior anharmonic-phonon NEGF studies of Si/Ge-type interfaces have shown that including three-phonon scattering increases the thermal boundary conductance by ~10% at room temperature and up to ~20% at 600 K relative to the harmonic limit, while leaving the spectral features of transport largely unchanged [40]. This indicates that anharmonicity primarily introduces a quantitative renormalization rather than a qualitative change in transport behavior. Accordingly, the ballistic NEGF results presented here should be interpreted as a coherent-transport baseline, with the identified structural trends expected to remain qualitatively valid at elevated temperatures, although with reduced magnitude due to increased decoherence.

To maintain experimental relevance, the layer thicknesses considered here were chosen within ranges commonly accessible in epitaxial Si/Ge heterostructures. In particular, Ge layers were kept thin enough to remain in the coherent-growth regime and avoid significant strain relaxation [50,51], while Si spacer layers were allowed to vary over a broader range consistent with experimentally realized multilayer structures [52–54]. These constraints ensure that the proposed superlattice designs are physically plausible for fabrication, even though the present work focuses primarily on coherent transport trends. The use of the EL2 phonon-emission window in GaAs is presented here as a theoretical use case to demonstrate how the proposed Si/Ge superlattice can be engineered to target a specific phonon-energy range. The NEGF model employed in this work does not include GaAs or a GaAs/SiGe interface, and therefore no assumption is made regarding the actual phonon-coupling efficiency between GaAs and the Si/Ge structure. A realistic implementation would require explicit modelling of interfacial phonon transmission, acoustic and optical mismatch, interface roughness, and other non-idealities, supported by experimental data. These aspects lie beyond the scope of the present theoretical study but constitute essential steps for future work aimed at integrating the proposed Si/Ge phonon-filtering structures into GaAs-based devices.

2.2. Optimization methodology

The second phase of the workflow, illustrated in Fig. 3, begins with dataset preprocessing. Following the data generation phase, in this step the simulation data is transformed into an efficient input/output representation suitable for the training of the MLP model. This step is further detailed in Fig. 4, which shows the complete transformation from the device profiles to the final input tensor $\vec{X}$ and the stratified data split.

As shown in Fig. 4, the device profiles are first organized into two complementary subsets: one with a fixed device length and another with variable device lengths. This dual-structured dataset is comprised of 2500 unique configurations, ensuring that the model learns both the specific features of length-constrained devices and the general physical relationships between composition, periodicity, and phonon transport.

Next, two parallel feature extraction pathways are applied. The first pathway extracts a set of key statistical descriptors from the ensemble of device profiles, including the distribution's mean value, standard deviation, median, peak-to-peak distance, skewness, kurtosis, 25^{th} and 75^{th} percentiles, profile tendency (mean of the first difference), and tendency variability (standard deviation of the first difference). These statistical descriptors provide a compact, high-level summary of the device profile distribution.

The second pathway begins by normalizing the full high resolution device profile using a Min-Max scaler, ensuring that all layer thicknesses are mapped to a consistent range prior to dimensionality reduction. Subsequently, Principal Component Analysis (PCA) [55] is applied to compress the full material-profile descriptor from its original 896 features to 60 principal components, while preserving 99% of the cumulative variance. This reduction retains the dominant structural variations in a compact and numerically stable representation. The final input tensor

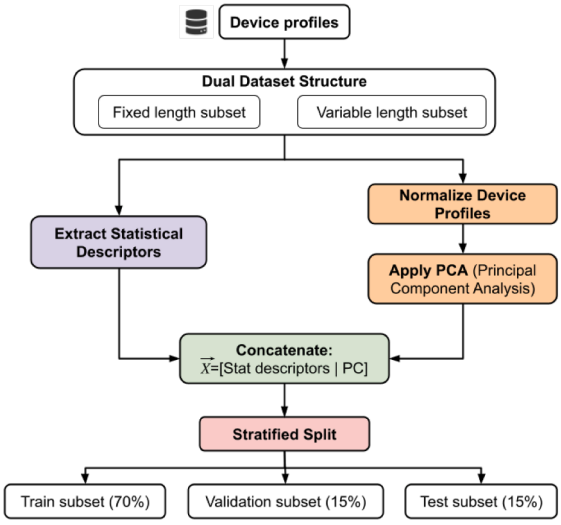

Fig. 4. Detailed view of the input dataset preprocessing workflow.

$\vec{X}$ (dimension 71) is obtained by concatenating the scaled statistical descriptor set (11 features) with the PCA-reduced profile representation (60 principal components).

Following the construction of the input features, the target variables (outputs) are also preprocessed to improve model training stability and performance. For the scalar thermal conductivity κ, we apply a logarithmic transformation ($\log \kappa$) to reduce the dynamic range of the data, followed by a Min-Max scaler to normalize the values to the $[0, 1]$ interval. For the spectral transmission data $T(\nu)$, we directly apply the Min-Max scaler to each frequency point, normalizing the transmission values.

To ensure a representative distribution of the device performance across all data subsets, we perform a stratified split based on the target $\log \kappa$. For the stratification, the dataset is first sorted and binned according to $\log \kappa$ values. Samples are then drawn proportionally from each bin to form the training (70%), validation (15%), and test (15%) subsets. This stratified approach guarantees that each subset contains a similar distribution of low, medium, and high conductivity devices, thereby preventing sampling bias and ensuring robust generalization.

The core of the optimization methodology is an MLP surrogate model implemented using the PyTorch Lightning framework [56]. The network is trained to predict both the integrated κ, and the full phonon transmission spectra $T(\nu)$, enabling simultaneous access to scalar and frequency-resolved transport information. The input descriptor tensor $\vec{X}$ is shared across both prediction tasks. The MLP architecture consists of an input layer followed by two hidden layers and an output layer. Each hidden layer of the MLP applies a linear transformation followed by layer normalization and a ReLU activation function. The linear layers provide affine transformations of the input features, LayerNorm stabilizes the intermediate activations and improves optimization robustness, and the ReLU activation introduces a non-saturating nonlinearity that facilitates gradient flow, reduces the risk of dead neurons, and enhances the surrogate model's ability to capture spectral features. The number of neurons in each layer is 71 (input), 178 (first hidden layer), 142 (second hidden layer), and 35 in the output layer (34 for $T(\nu)$ and 1 for κ). All weights are initialized from a normal distribution with a mean of 0 and a standard deviation of 0.01.

Training is performed using the Adam optimizer [57] with an initial learning rate of 0.1 and a batch size of 32. To prevent stagnation in local minima, we employ the PyTorch ReduceLROnPlateau scheduler that adaptively decreases in a factor of 0.5 the learning rate when the validation loss stops improving. Early stopping based on validation loss is also applied to avoid overfitting. The model is trained to optimize two different loss functions selected according to the nature of each target property. For the scalar thermal conductivity κ, we minimize the Root Mean Squared Error (RMSE), defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\kappa_i - \hat{\kappa}_i)^2}, \tag{7}$$

where κ_i is the simulated thermal conductivity, $\hat{\kappa}_i$ is the predicted thermal conductivity, and n the number of samples. RMSE is the standard choice for scalar regression tasks because it heavily penalizes large deviations, encouraging the model to converge toward the mean of the target distribution. Moreover, its quadratic nature yields gradients that scale with the error, enabling efficient and stable convergence near the optimum. For the phonon transmission spectra $T(\nu)$, the chosen loss function is the Smooth L1 loss (also known as Huber loss) [58], defined as:

$$\mathrm{Smooth\ L1}(y, \hat{y}) = \begin{cases} 0.5 \cdot (y - \hat{y})^2, & \text{if } |y - \hat{y}| < 1 \\ |y - \hat{y}| - 0.5, & \text{otherwise} \end{cases} \tag{8}$$

where y and $\hat{y}$ denote the simulated and predicted transmission values, respectively. This choice is motivated by the distinct nature of spectral prediction: $T(\nu)$ is a high-dimensional output that must simultaneously capture fine spectral features across a wide dynamic range. Smooth L1 combines the advantages of L2 loss (mean square error) for small errors, ensuring smooth convergence near the optimum, with the robustness of L1 loss (mean absolute error) against large errors, providing bounded gradients and outlier insensitivity [59]. Training incorporates an early stopping callback, monitored on the validation loss, and a learning rate scheduler. This combination halts the process when performance plateaus and dynamically adjusts the optimization step size, ensuring efficient convergence and optimal generalization for each property.

To evaluate the accuracy of the predictions, we employ two complementary metrics. For κ, we use the coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\kappa_i - \hat{\kappa}_i)^2}{\sum_{i=1}^{n}(\kappa_i - \bar{\kappa})^2}, \tag{9}$$

where $\bar{\kappa}$ is the mean of the simulated κ_i values. This metric provides a normalized measure of how well the model captures the variance in the target properties, with values close to 1 indicating excellent predictive performance. For $T(\nu)$, we use the RMSE to quantify the absolute prediction error when predicting the spectra in physical units.

To further assess the robustness and reliability of the surrogate model beyond standard accuracy metrics, we employ complementary measures tailored to the nature of each predicted quantity. For the scalar thermal conductivity κ, the predictive uncertainty is quantified by using 95% prediction intervals (PI) derived from the empirical residual distribution ($\kappa_i - \hat{\kappa}_i$), together with the empirical coverage (EC), defined as the fraction of true values lying within their corresponding interval. This metric provides a probabilistic characterization of the model's confidence in its scalar predictions. For the phonon transmission spectra $T(\nu)$, robustness is evaluated using the spectral Pearson correlation (ρ) [60], computed over the frequency axis for each sample. This metric captures the similarity in spectral shape between simulated and predicted transmission curves, complementing RMSE by assessing the model's ability to reproduce fine spectral features across the full frequency domain

Once the MLP model is trained and validated, it is employed to map a vast design space of unobserved configurations. To explore the Si/Ge multilayer design space beyond the configurations available in the NEGF-derived dataset, a large pool of artificial candidate profiles is generated under a set of strict physical and statistical constraints. In total, 10^5 candidates are produced, each consisting of a fixed material profile length of 80 at. layers. The generation process does not enumerate all possible binary sequences, nor does it sample arbitrary configurations. Instead, it follows a controlled procedure designed to ensure fabrication feasibility and consistency with the statistical properties of the initial dataset. Each profile is represented as a discrete sequence of atomic layers, where each layer is assigned a binary material label (1 = Si, 2 =

Ge). To avoid isolated single-layer inclusions, which are not physically fabricable, a minimum block size of two consecutive atomic layers of the same material is imposed. Furthermore, the initial dataset contains structures with 1 to 38 Si/Ge interfaces, and all generated candidates are restricted to this same range to prevent non-representative artificial profiles. The Si volume fraction of each candidate must lie within the 5th–95th percentile of the initial dataset distribution, ensuring that compositions remain representative of the underlying physical system. Candidates identical to any structure in the initial dataset or previously evaluated during optimization are automatically discarded using a hash-based check. To ensure that the surrogate model operates strictly in an interpolation regime, a trust-region constraint based on structural similarity is imposed. For each candidate, the normalized Hamming distance is computed:

$$d_H = \frac{1}{L}\sum_{i=1}^{L}(p_i \neq q_i), \tag{10}$$

where p_i and q_i are the proposed and initial dataset material profile points, respectively. An artificial candidate is accepted only if it has at least 10 neighbors in the initial dataset with $d_H < 0.03$ ($\sim 2.5\%$ difference) and satisfies all other physical constraints. This guaranties that every candidate belongs to a region of the design space that is sufficiently sampled and physically meaningful. In addition to the physical and statistical constraints described above, candidate profiles are sampled using the covariance matrix adaptation evolution strategy (CMA-ES) optimization algorithm [61]. The sampler operates in a continuous latent space $z \in [0,1]^{80}$, and each proposed vector is deterministically mapped to a discrete Si/Ge multilayer by thresholding each component ($z_i < 0.5 \rightarrow Si, z_i \geq 0.5 \rightarrow Ge$) and subsequently enforcing all physical constraints (minimum block size, interface-count limits, composition bounds, and trust-region similarity criteria). Thus, the sampler does not generate arbitrary binary sequences; it explores only the physically admissible region of the design space defined by our constraints. The resulting dataset is then preprocessed using the same scaling and PCA transformations fitted on the original NEGF-derived training data.

Finally, although the surrogate model and CMA-ES sampler are used to efficiently explore the constrained design space, the best candidate structures at the end of the optimization process are independently recalculated using full NEGF simulations. This verification step ensures that the reported phonon-transmission spectra and stop-band characteristics correspond to physically accurate NEGF results rather than surrogate predictions and prevents approximation errors from propagating into the final inverse-designed structures.

The presented methodology constitutes an inverse-design framework because the optimization target is specified a priori. The desired phonon-frequency window to be suppressed (15–36.2 meV for the EL2 defect in GaAs) is defined first, and the structural parameters of the Si/Ge superlattice are subsequently inferred to satisfy this target. The surrogate model is therefore used to accelerate the search within the constrained design space imposed by the inverse problem, rather than performing an unconstrained forward-prediction sweep. This distinction ensures that the workflow follows a target-driven inverse-design paradigm rather than a purely forward screening strategy.

In the present implementation, all Si/Ge superlattices used for optimization share a fixed total length of 80 at. layers. Consequently, the CMA-ES sampler explores only the corresponding fixed-length design space. This constraint, however, is not inherent to the methodology: extending the framework to variable-length multilayers would only require increasing the NEGF-derived dataset with additional profile lengths and retraining the surrogate on a feature representation that explicitly encodes layer count, enabling reliable interpolation across devices of different sizes.

2.3. Structural characterization of superlattices

To characterize the structural properties of the optimized and reference Si/Ge superlattices, we compute a set of descriptors commonly used in signal analysis and elastic-wave physics. The number of interfaces is obtained by counting all material transitions along the profile. Thickness variability is quantified through the mean (μ_t) and standard deviation (σ_t) of the thickness sequence. Global aperiodicity is measured using the Shannon entropy of the thickness distribution (H). For a thickness sequence $t = \{t_1, t_2, \ldots, t_N\}$, let $\mathcal{T}$ denote the set of distinct thickness values and let $p(t_i)$ be the empirical probability of observing thickness $t_i \in \mathcal{T}$, computed as the relative frequency of t in the sequence. The entropy is then defined as:

$$H = -\sum_{t_i \in \mathcal{T}} p(t_i)\,\log p(t_i), \tag{11}$$

which increases when the multilayer contains a broader and more irregular set of thickness values. Short-range structural order is quantified through the characteristic correlation length (ℓ_{corr}) of the thickness sequence. For a profile the discrete autocorrelation ($C(k)$) function is computed and normalized ($\tilde{C}(k)$) as:

$$C(k) = \frac{1}{N-k}\sum_{i=1}^{N-k} t_i\, t_{i+k}, \quad \tilde{C}(k) = \frac{C(k)}{C(0)}. \tag{12}$$

$\tilde{C}(k)$ measures how fast structural correlations decay with layer separation. ℓ_{corr} is defined as the smallest integer shift k for which $\tilde{C}(k)$ falls below a fixed threshold (here $\eta = 0.5$):

$$\ell_{\mathrm{corr}} = \min\{\, k \mid \tilde{C}(k) < \eta \,\}. \tag{13}$$

Spatial modulation is characterized by the dominant frequency of the discrete Fourier transform of the thickness sequence, following classical FFT-based spectral analysis. This metric provides a direct link to the interference mechanisms responsible for frequency-selective phonon filtering: multilayers with higher dominant spatial frequency exhibit stronger high-contrast modulation of acoustic impedance, enabling destructive interference at phonon wavelengths that satisfy an approximate Bragg condition $\lambda_{\mathrm{ph}} \approx 2d_{\mathrm{eff}}$, where d_{eff} is the effective structural period associated with the dominant spatial frequency. Finally, the acoustic-impedance contrast across interfaces is computed using standard definitions from elastic-wave theory [62], providing a measure of the scattering strength that contributes to stop-band formation and phonon localization.

2.4. Computational resources and cost analysis

All DFT and NEGF calculations were performed on SkyLake Dell PowerEdge C6420 nodes equipped with 32-core Intel® Xeon® Gold 6142 processors (2.6 GHz) and 192 GB of RAM. The one-time DFT workflow for Si (geometry optimization, electronic structure, and phonon properties) requires approximately 64 cores × 12 h = 768 core-hours. NEGF simulations for individual structures require between 225 and 900 cores depending on system size; using an average of 550 cores and a wall-time of 20 s per run, each NEGF evaluation costs approximately 3 core-hours. For a system composed of 40 unit cells, the total number of possible Si/Ge configurations exceeds 10^{12}, implying a brute-force cost of more than 3×10^{12} core-hours.

Surrogate-model training, inference, and dataset generation were performed on a local workstation equipped with an Intel® Core™ i9-10850K CPU (10 cores, 20 threads, 3.60–5.20 GHz), using 5 parallel threads. Under these conditions, the MLP requires 6.6 min to train, corresponding to 0.55 core-hours. Loading the trained model for prediction takes 0.02 s. Efficiently sampling the search space with the

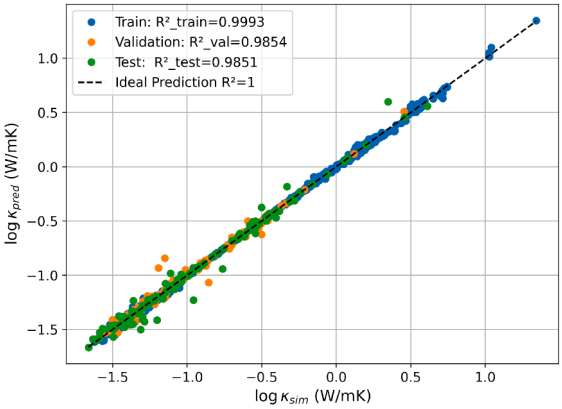

Fig. 5. Comparison between simulated thermal conductivity ($\log \kappa_{sim}$) and predicted thermal conductivity ($\log \kappa_{pred}$) for train, validation test subsets together with their coefficient of determination (R^2). The dashed black line corresponds to the ideal prediction of $R^2 = 1$.

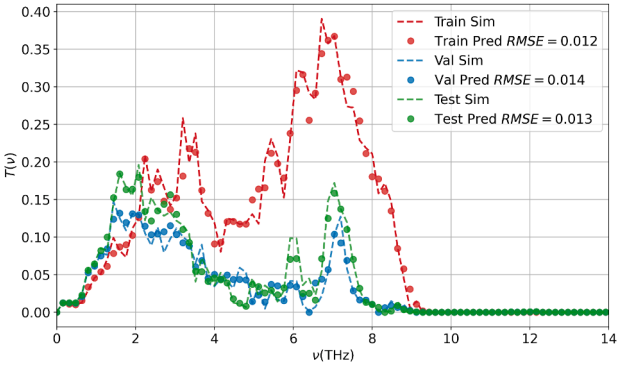

Fig. 6. Comparison between simulated and predicted transmission spectra ($T(\nu)$) as a function of the frequency (ν) for train, validation and test subsets. The root mean square values RMSE of each subset are also shown.

CMA-ES algorithm by generating the candidates and predicting the outputs of 10^5 non-duplicated material profiles at fixed 80-unit-cell length requires 10.5 min, equivalent to 0.17 core-hours. These computational costs are negligible compared to direct NEGF simulations. In practical terms, evaluating 10^5 candidate structures using direct NEGF would require approximately 3×10^5 core-hours, whereas our workflow requires only the 2500 NEGF simulations used for training (~7500 core-hours), achieving a speed-up of ~40×. The computational cost of the ML workflow itself is negligible (0.60 core-hours), enabling efficient exploration of configurational spaces that would otherwise be computationally prohibitive.

With the methodology fully established, the following section presents the validation of the trained models and the results of the inverse design optimization.

3. Results and discussion

3.1. Model validation

The predictive performance of the MLP architecture was evaluated on the independent test sets reserved during the stratified split procedure. Two complementary aspects were assessed: the accuracy of the scalar thermal conductivity predictions and the fidelity of the reconstructed phonon transmission spectra.

Fig. 5 compares the $\log \kappa$ predicted by the MLP model against the values computed via NEGF simulations for the training, validation, and test sets. The model achieves excellent agreement across the full range of $\log \kappa$ values, with a coefficient of determination $R^2 \sim 0.99$ for all three subsets and an empirical coverage of EC = 0.9503 for the validation subset, very close to the nominal 0.95 level. These high R^2 and EC values demonstrate the model's ability to accurately learn the complex relationship between superlattice geometry and heat transport. The consistent performance across training, validation, and test sets confirms that no overfitting has occurred and that the model generalizes well to unseen configurations.

Beyond scalar quantities, the model was also trained to predict the full frequency-dependent phonon transmission spectra $T(\nu)$. Fig. 6 presents representative examples of arbitrarily selected devices for each subset, comparing the predicted spectra against simulated references. The model achieves a RMSE lower than 0.014 a.u. across the training, validation, and test sets, indicating that the predicted transmission values deviate from the true spectra by a maximum of 1.5% on average. This level of accuracy is remarkable given the complexity of the spectral features, including multiple peaks and valleys arising from phononic band-structure effects. In addition to RMSE, we evaluate the consistency of the spectral predictions using the spectral Pearson correlation, which quantifies the similarity in spectral shape between simulated and predicted transmission curves. The model achieves mean correlations of 0.989 (train), 0.983 (validation), and 0.976 (test), confirming that the surrogate reliably reproduces the fine spectral structure of $T(\nu)$ across the entire dataset.

The successful validation of both prediction tasks establishes the reliability of the machine learning models as surrogate solvers for subsequent optimization. The high R^2 for $\log \kappa$ ensures that the model can correctly identify configurations with extreme thermal conductivity values, while the low RMSE for $T(\nu)$ guarantees that frequency-targeted engineering, such as creating stop-bands in a desired frequency window, can be performed with confidence. With this validation completed, we now proceed to the inverse design of optimal superlattice configurations.

3.2. Inverse design optimization

With the validated model, we now proceed to explore the vast design space of Si/Ge superlattices to identify optimal configurations for specific thermal management objectives. The MLP combined with CMA-ES sampler enables rapid screening of a big amount of candidate structures, a task that would be prohibitively expensive using NEGF simulations alone. We first focus on identifying configurations that achieve extreme thermal conductivity values, establishing the performance bounds of the system. For this purpose, the CMA-ES sampler generates 10^5 possible candidates which are evaluated using the trained model, applying the following constraints: a fixed L_{tot} of 80 atomic layers, and a minimum uc of 2 atomic layers per material.

Fig. 7 compares the lowest-κ structure present in the original NEGF dataset, denoted as the κ-reference (top), with the optimal configuration identified by the inverse-design workflow (bottom). The optimal device achieves a κ value ~10.20% lower than the κ-reference, demonstrating the capability of the methodology to discover superior structures beyond the initial training data. Beyond this improvement, the optimized profile exhibits clear structural differences relative to the κ-reference. It contains a higher number of Si/Ge interfaces (26 vs. 24), leading to stronger interface-driven phonon scattering. The layer-thickness distribution is slightly more regular ($\sigma_t = 1.99$ vs. 2.24). Global aperiodicity, quantified through the Shannon entropy of the thickness distribution, is significantly lower in the optimized structure ($H = 1.11$ vs. 1.71), indicating a more repetitive and less disordered thickness pattern. The correlation length is shorter (1 vs. 2), revealing reduced medium-range structural order. Fourier analysis shows that the optimized profile also exhibits a higher dominant spatial frequency (0.190 vs. 0.076), corresponding to a more rapidly varying structural pattern. Finally, the average acoustic-impedance jump across interfaces is larger (2.57 vs. 2.37 MRayl), further reinforcing phonon backscattering. Together, these

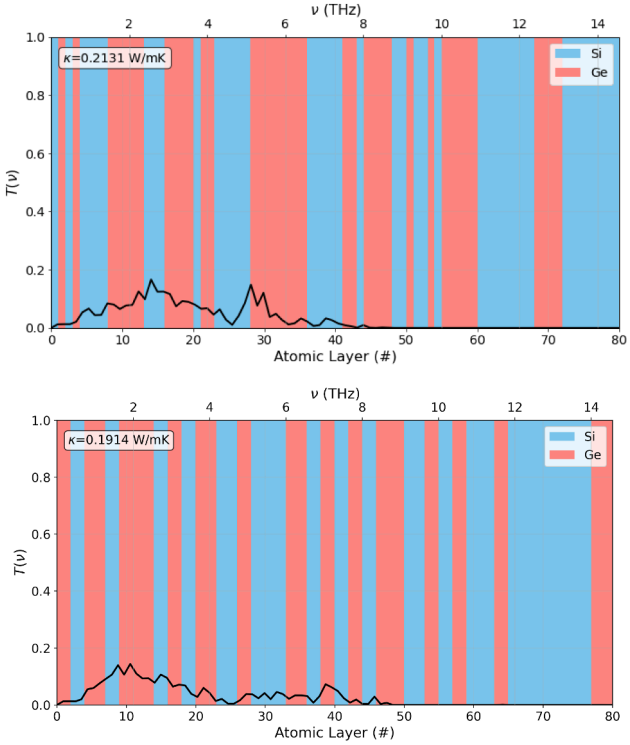

Fig. 7. Transmission spectra and material profile for the device with minimum thermal conductivity in the initial simulated dataset, denoted as κ-reference (top), and the optimal device proposed by the model (bottom).

descriptors explain why the optimized configuration achieves a lower-κ value.

The thermal conductivity κ is directly determined by the area under the transmission spectra, weighted by the phonon energy and thermal population (based on Eq. 5) according to:

$$k \propto \int_0^\infty h\nu \cdot T(\nu) \frac{\partial f_{BE}(\nu, T)}{\partial T} d\nu. \tag{14}$$

The weighting factor ($h\nu$) assigns greater importance to higher frequency modes, which carry more energy per phonon.

Examining the transmission spectra in Fig. 7, the unweighted integral $\int h\nu T(\nu) d\nu$ (the model is not designed to learn the Bose-Einstein distribution) for the optimal device is 11.01% lower than that of the best device in the simulated dataset. This reduction is consistent with the lower thermal conductivity obtained from the full expression in Eq. 14. However, the contribution to κ is not uniform across frequencies: although the $h\nu$ factor increases the weight of high-energy phonons, the derivative of the Bose–Einstein distribution strongly suppresses their thermal occupation at 200 K. As a result, the thermal-conductance integrand is maximized in the mid-frequency range (2–6 THz), where both the phonon energy and the thermal population derivative are significant. When inspecting the spectra above 6 THz, the optimal device exhibits slightly higher transmission in this region. Although these high-frequency modes contribute less to thermal transport at room temperature due to their lower thermal occupation, they are relevant for the presented phononic-stop-band application. Therefore, while the optimized structure achieves lower thermal conductivity, it is comparatively less effective at suppressing transmission above 6 THz, which may reduce its performance as a high-frequency stop-band for the proposed use case.

Having demonstrated the capability of our methodology to identify configurations with minimum κ, we now focus on the second design objective: engineering a superlattice that selectively suppresses phonon propagation in the frequency window corresponding to the EL2 recombination center emission. As established in Section 1.2, non-radiative recombination at arsenic antisites in GaAs releases energy through the emission of multiple optical phonons in the 15 – 36.2 meV range, which corresponds to frequencies of ~3.63-8.75 THz. A superlattice capable of creating a phononic stop-band in this spectral region would provide direct, at-source thermal management for GaAs-based devices.

Fig. 8 presents the optimal device configuration identified by the inverse-design methodology, along with its phonon transmission spectrum $T(\nu)$ and material profile. The CMA-ES sampler targets the integrated transmission within the EL2 frequency window, minimizing the objective function $\mathrm{f} = \int_{\nu_{\min}}^{\nu_{\max}} h\nu T(\nu)\, d\nu$ as a proxy for suppressing phonon propagation in this band. The structure comprises an aperiodic sequence of Si and Ge layers optimized to maximize acoustic-impedance contrast at the interfaces. This architectural arrangement induces strong destructive interference and enhanced phonon scattering, which are the fundamental drivers of phononic band-gap formation.

To better understand the structural features responsible for phononic suppression in the EL2 emission band, we compared the optimized profile with the best-performing configuration in the initial dataset using the structural and spectral descriptors introduced in Section 2.3. The differences between both devices are substantial. The optimized structure contains more Si/Ge interfaces (30 vs. 24), enhancing interface-driven phonon reflection. Its thickness distribution is significantly more regular, with a lower standard deviation ($\sigma_t = 1.01$ vs. 2.24). Global aperiodicity, quantified through the Shannon entropy of the thickness distribution, is markedly reduced in the optimized structure (1.15 vs. 1.71), indicating a more repetitive and less disordered arrangement of layer thicknesses. The correlation length is shorter (1 vs. 2 layers), revealing reduced medium-range structural order. Fourier analysis shows that the optimized profile also exhibits a higher dominant spatial frequency (0.2405 vs. 0.0759), corresponding to a more rapidly varying structural pattern. Finally, the average acoustic-impedance jump across interfaces is larger (2.96 vs. 2.37 MRayl), further reinforcing phonon backscattering. Together, these descriptors explain the strong suppression of phonon transmission in the 3.63–8.75 THz window and the 38.84% reduction in the unweighted spectral integral $\int h\nu\, T(\nu)$ achieved by the inverse-designed profile.

To elucidate the origin of the selective phonon filtering, the participation ratio (PR) together with the spatial distribution of the local phonon density through atomic layers for representative frequency modes is analyzed in Fig. 9, comparing the phononic response of reference device (Ref) and the optimal proposed (Opt). For Fig. 9a) within the 3.63–8.75 THz window, the reference PR exhibits large mode-to-mode fluctuations, with several sharp minima, whereas the optimized PR varies more smoothly and remains at more moderate values across the band.

Notably, the reference structure shows more pronounced PR dips at isolated frequencies, but the optimized structure displays a more systematic suppression of modes in the target window. This suggests that localization in the optimal structure is not simply stronger in a scalar sense; it must instead arise from an asymmetric or spatially confined redistribution of mode energy and that PR alone does not fully capture the filtering mechanism, since a mode may remain only moderately extended yet still be spatially biased toward one side of the device. This motivates the layer-resolved phonon density analysis shown in Fig. 9b–e). At 4.96 THz (Fig. 9b)), the reference and optimized spatial profiles are broadly similar, indicating that this frequency lies outside the most strongly affected part of the spectrum. At 5.28 THz (Fig. 9c)), the optimized structure develops a sharply confined phonon-density peak near layers 60–70, whereas the reference density remains comparatively delocalized across the stack. This behavior is consistent with strong spatial confinement in the optimized structure and correlates with the suppression of transmission in this frequency range. The trend reverses at 6.24 THz (Fig. 9d)) where the reference structure exhibits a strong, sharply localized density peak near layers 10–15, while the optimal structure is essentially flat across the device. This shows that both structures support disorder-induced trap states within the suppression band, but at

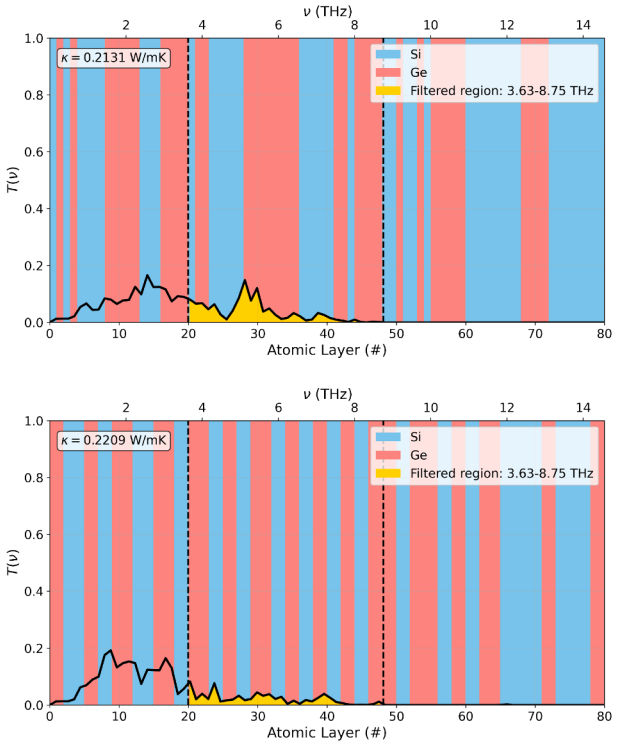

Fig. 8. Material profiles achieving minimal phononic-mode transmission within the target frequency window (3.63–8.75 THz). The top panel shows the best-performing device in the initial simulated dataset, while the bottom panel displays the optimal structure identified by the inverse-design model. The selected frequency bounds correspond to the EL2 defect's multi-phonon emission window, motivating the suppression of transmission in this spectral region.

different frequencies and spatial locations, consistent with distinct short-range-order-driven scattering resonances in each configuration. At 7.84 THz (Fig. 9e)), both profiles remain broad and low in amplitude, with the optimal structure showing a mild upward trend toward layers 70–80 rather than a sharp, isolated peak. This indicates weak, incipient localization rather than a resonant trap comparable to those seen at 5.28 or 6.24 THz.

These results indicate that transmission suppression in the optimized Si/Ge configuration within the 3.63–8.75 THz window does not arise from uniformly stronger disorder scattering, but from frequency-selective spatial confinement of vibrational modes. To connect these to the spatial structure, we examined the underlying Si/Ge stacking sequences directly. The optimal sequence's near-periodic short-range order differs structurally from the reference's larger, more randomly clustered domains, and this coherent ordering can fold the phonon dispersion and generate localized trap states at specific frequencies (e.g., 5.28 THz) that carry negligible net flux. The reference structure, while also hosting its own localized states (e.g., at 6.24 THz), shows this localization distributed differently across frequency and space, resulting in a broader transmission window that is not suppressed as effectively. This suggests that arrangement of Si/Ge layers is a controllable design parameter for engineering frequency-selective phonon filters in Si/Ge heterostructures.

Building on this physical interpretation, these results also validate the core premise of the proposed inverse-design workflow, i.e., the ability to define an arbitrary target frequency window and automatically identify the structural geometry that maximizes phononic suppression within it. Compared with the best configuration in the initial dataset, the optimized device exhibits a clearly reinforced structural signature, characterized by a higher interface density, reduced aperiodicity, a shorter correlation length, stronger impedance contrast, and a higher dominant

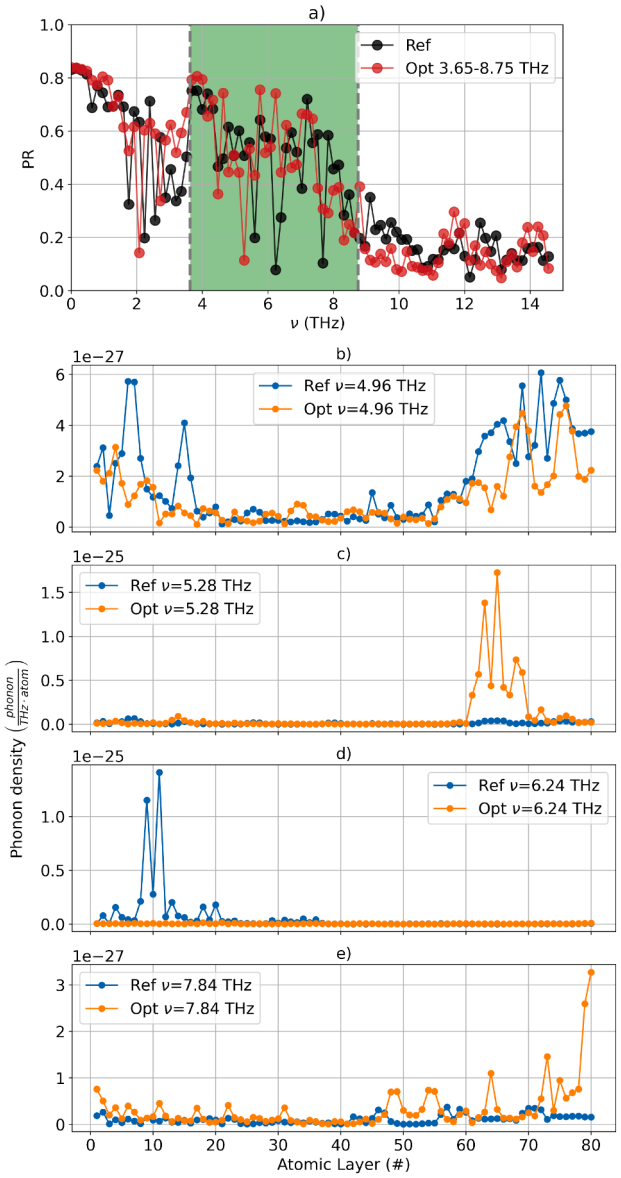

Fig. 9. a) Participation Ratio, PR as a function of phonon frequency, ν, for the reference (Ref) and optimal (Opt) stop-band profile in the defined frequency window. b)-e) Layer-resolved local phonon density for representative phonon modes in desired window: b) ν = 4.96 THz, comparison with similar PR for Ref and Opt; c) ν = 5.28 THz, representing the Opt PR minimum; d) ν = 6.24 THz, representing the Ref minimum; and e) ν = 7.84 THz, comparison at mode with high PR difference between Ref and Opt.

spatial frequency. This convergence toward a distinct architectural motif demonstrates that the model not only discovers a superior solution, but also reveals the underlying structural principles required to engineer a robust stop-band in the 3.63–8.75 THz EL2 emission window. Although the EL2 center in GaAs is used here as a proof of concept, the methodology is entirely general and provides a versatile, scalable framework for any semiconductor platform in which specific vibrational modes drive localized heating, enabling integrated at-source thermal management across a wide range of electronic and optoelectronic devices.

4. Conclusions

In this work, we have developed and validated a combined NEGF and machine learning methodology for the inverse design of Si/Ge

superlattices with tailored phononic properties. The key findings and contributions of this study are summarized as follows.

First, we trained an MLP model with NEGF-generated data that accurately predicts both the scalar thermal conductivity (κ) and the full phonon transmission spectra ($T(\nu)$). The model achieves a coefficient of determination $R^2 \sim 0.99$ for κ and a root mean square error of 0.013 a.u. for $T(\nu)$ across the training, validation, and test datasets, demonstrating its accuracy as a fast strategy for subsequent optimization tasks.

Second, we applied the methodology to identify superlattice configurations with minimum thermal conductivity values. The optimized device achieves a 10.20% reduction in κ compared to the best configuration in the original NEGF-simulated dataset. Analysis of $T(\nu)$ reveals that this reduction is achieved through the strategic suppression of high-frequency phonon modes, which contribute most significantly to heat transport due to their higher energy per phonon.

Third, as a proof of concept, we demonstrated frequency-targeted phonon engineering by designing a superlattice that selectively suppresses transmission in the 15-36.2 meV (3.63-8.75 THz) range, the characteristic phonon emission window of the EL2 recombination center in GaAs. The optimized device exhibits near-complete suppression within this critical region, while maintaining moderate transmission at low-frequencies, illustrating the capability to target specific heat sources without compromising overall thermal dissipation. The GaAs EL2 phonon-emission window is considered as a theoretical example to show the tunability of the proposed Si/Ge superlattice toward selected phonon-energy ranges. Realistic integration in GaAs-based devices will require explicit modelling and experimental validation of interfacial phonon transmission and related non-idealities.

Finally, the methodology presented here is entirely general. While this study focuses on Si/Ge superlattices and the EL2 defect as a use case, the workflow can be directly applied to any semiconductor system where specific phonon frequency windows are identified as being responsible for localized heating. This provides a versatile and scalable approach for integrated, at-source thermal management across a wide range of electronic technologies.

CRediT authorship contribution statement

Julian G. Fernandez: Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation; **Shubham Tyagi:** Writing – original draft, Validation, Investigation, Data curation; **Natalia Seoane:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization; **Antonio Garcia-Loureiro:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization; **Marc Bescond:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

These authors contributed equally: Julian G. Fernandez and Shubham Tyagi. This work has received financial support from Xunta de Galicia (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04 and Reference Competitive Group accreditation ED431C-2022/016), the European Union (European Regional Development Fund - ERDF/EU), and COCONUT project from ANR (ANR-24-CE50- 5682). The Centre de Calcul Intensif d'Aix-Marseille and CINES's Adastra are acknowledged for granting access to its high-performance computing resources.

Data availability

Data will be made available on request.

References

[1] W.-Y. Woon, A. Kasperovich, J.-R. Wen, K.K. Hu, M. Malakoutian, J.-H. Jhang, S. Vaziri, I. Datye, C.C. Shih, J.F. Hsu, X.Y. Bao, Y. Wu, M. Nomura, S. Chowdhury, S.S. Liao, Thermal management materials for 3D-stacked integrated circuits, Nat. Rev. Electr. Eng. 2 (9) (2025) 598–613. https://doi.org/10.1038/s44287-025-00196-0

[2] R. Gaska, A. Osinsky, J.W. Yang, M.S. Shur, Self-heating in high-power AlGaN-GaN HFETs, IEEE Electron Device Lett. 19 (3) (1998) 89–91. https://doi.org/10.1109/55.661174

[3] E. Pop, K.E. Goodson, Thermal phenomena in nanoscale transistors, J. Electron. Packag. 128 (2) (2006) 102–108. https://doi.org/10.1115/1.2188950

[4] A. Bar-Cohen, P. Wang, On-Chip Thermal Management and Hot-Spot Remediation, Springer US, 2009, p. 349–429. https://doi.org/10.1007/978-1-4419-0040-1_12

[5] T. Gong, G. Hou, Y. Wu, L. Li, Y. Wang, M. Shi, L. Kang, J. Zhou, L. Cao, L. Gao, T. Ming, J. Li, W. Su, Co-optimization of electrical-thermal–mechanical behaviors of an on-chip thermoelectric cooling system using response surface method, Appl. Therm. Eng. 244 (2024) 122699. https://doi.org/10.1016/j.applthermaleng.2024.122699

[6] M. Avgerinou, P. Bertoldi, L. Castellazzi, Trends in data centre energy consumption under the european code of conduct for data centre energy efficiency, Energies 10 (10) (2017). https://doi.org/10.3390/en10101470

[7] S.M. Sohel Murshed, C.A. Nieto de Castro, A critical review of traditional and emerging techniques and fluids for electronics cooling, Renew. Sustain. Energy Rev. 78 (2017) 821–833. https://doi.org/10.1016/j.rser.2017.04.112

[8] I. Chowdhury, R. Prasher, K. Lofgreen, G. Chrysler, S. Narasimhan, R. Mahajan, D. Koester, R. Alley, R. Venkatasubramanian, On-chip cooling by superlattice-based thin-film thermoelectrics, Nat. Nanotechnol. 4 (4) (2009) 235–238. https://doi.org/10.1038/nnano.2008.417

[9] A. Ziabari, M. Zebarjadi, D. Vashaee, A. Shakouri, Nanoscale solid-state cooling: a review, Rep. Prog. Phys. 79 (9) (2016) 095901. https://doi.org/10.1088/0034-4885/79/9/095901

[10] T. Gebrael, J. Li, A.R. Gamboa, J. Ma, J. Schaadt, L. Horowitz, R. Pilawa-Podgurski, N. Miljkovic, High-efficiency cooling via the monolithic integration of copper on electronic devices, Nat. Electron. 5 (6) (2022) 394–402. https://doi.org/10.1038/s41928-022-00748-4

[11] M. Bescond, D. Logoteta, F. Michelini, N. Cavassilas, T. Yan, A. Yangui, M. Lannoo, K. Hirakawa, Thermionic cooling devices based on resonant-tunneling AlGaAs/GaAs heterostructure, J. Phys.: Condens. Matter 30 (6) (2018) 064005. https://doi.org/10.1088/1361-648x/aaa4cf

[12] M. Bescond, K. Hirakawa, High-performance thermionic cooling devices based on tilted-Barrier semiconductor heterostructures, Phys. Rev. Appl. 14 (6) (2020). https://doi.org/10.1103/physrevapplied.14.064022

[13] G. Bulman, P. Barletta, J. Lewis, N. Baldasaro, M. Manno, A. Bar-Cohen, B. Yang, Superlattice-based thin-film thermoelectric modules with high cooling fluxes, Nat. Commun. 7 (1) (2016). https://doi.org/10.1038/ncomms10302

[14] M. Settipalli, V.S. Proshchenko, S. Neogi, The effect of electron–phonon and electron-impurity scattering on the electronic transport properties of silicon/germanium superlattices, J. Mater. Chem. C 10 (19) (2022) 7525–7542.

[15] Y. Wang, H. Huang, X. Ruan, Decomposition of coherent and incoherent phonon conduction in superlattices and random multilayers, Phys. Rev. B 90 (16) (2014). https://doi.org/10.1103/physrevb.90.165406

[16] Y. Liu, R. Hu, Y. Wang, J. Ma, Z. Yang, X. Luo, Big-data-accelerated aperiodic Si/Ge superlattice prediction for quenching thermal conduction via pattern analysis, Energy AI 3 (2021) 100046. https://doi.org/10.1016/j.egyai.2020.100046

[17] P. Chakraborty, Y. Liu, T. Ma, X. Guo, L. Cao, R. Hu, Y. Wang, Quenching thermal transport in aperiodic superlattices: a molecular dynamics and machine learning study, ACS Appl. Mater. Interfaces 12 (7) (2020) 8795–8804. https://doi.org/10.1021/acsami.9b18084

[18] R. Butola, Y. Li, S.R. Kola, A machine learning approach to modeling intrinsic parameter fluctuation of gate-All-Around si nanosheet MOSFETs, IEEE Access 10 (2022) 71356–71369. https://doi.org/10.1109/access.2022.3188690

[19] A. García-Loureiro, N. Seoane, J.G. Fernández, E. Comesaña, J.C. Pichel, A machine learning approach to model the impact of line edge roughness on gate-all-around nanowire FETs while reducing the carbon footprint, Plos One 18 (7) (2023) e0288964. https://doi.org/10.1371/journal.pone.0288964

[20] H. Xu, W. Gan, L. Cao, C. Yang, J. Wu, M. Zhou, H. Qu, S. Zhang, H. Yin, Z. Wu, A machine learning approach for optimization of channel geometry and source/drain doping profile of stacked nanosheet transistors, IEEE Trans. Electron Devices 69 (7) (2022) 3568–3574. https://doi.org/10.1109/ted.2022.3175708

[21] J.G. Fernandez, G. Etesse, N. Seoane, E. Comesaña, K. Hirakawa, A. Garcia-Loureiro, M. Bescond, A novel machine learning workflow to optimize cooling devices grounded in solid-state physics, Sci. Rep. 14 (1) (2024). https://doi.org/10.1038/s41598-024-80212-9

[22] S. Fafard, F. Proulx, M.C.A. York, L.S. Richard, P.O. Provost, R. Arès, V. Aimez, D.P. Masson, High-photovoltage GaAs vertical epitaxial monolithic heterostructures with 20 thin p/n junctions and a conversion efficiency of 60%, Appl. Phys. Lett. 109 (13) (2016) 131107. https://doi.org/10.1063/1.4964120

[23] M. Matsuura, H. Nomoto, H. Mamiya, T. Higuchi, D. Masson, S. Fafard, Over 40-W electric power and optical data transmission using an optical fiber, IEEE Trans. Power Electron. 36 (4) (2021) 4532–4539. https://doi.org/10.1109/TPEL.2020.3027551

[24] H. Helmers, E. Lopez, O. Höhn, D. Lackner, J. Schön, M. Schauerte, M. Schachtner, F. Dimroth, A.W. Bett, 68.9% Efficient gaas-Based photonic power conversion enabled by photon recycling and optical resonance, Phys. Status Solidi – Rapid Res. Lett. 15 (7) (2021) 2100113. https://doi.org/10.1002/pssr.202100113

[25] T. Maranets, M. Nasiri, Y. Wang, Influence of spatial coherence on phonon transmission across aperiodically arranged interfaces, Phys. Lett. A 512 (2024) 129572. https://doi.org/10.1016/j.physleta.2024.129572

[26] C. Pattyn, E. Kovacevic, S. Hussain, A. Dias, T. Lecas, J. Berndt, Nanoparticle formation in a low pressure argon/aniline RF plasma, Appl. Phys. Lett. 112 (1) (2018). https://doi.org/10.1063/1.5019926

[27] A.L. Khamets, I.V. Safronov, A.B. Filonov, D.B. Migas, Effects ofSi/Ge superlattice structure with intermixed interfaces on phonon thermal conductivity, Phys. E: Low-dimens. Syst. Nanostructures 165 (2025) 116108. https://doi.org/10.1016/j.physe.2024.116108

[28] D.J. Chadi, K.J. Chang, Theory of the atomic and electronic structure of DX centers in gaas and $Al_xGa_{1-x}As$ alloys, Phys. Rev. Lett. 61 (1988) 873–876. https://doi.org/10.1103/PhysRevLett.61.873

[29] J. Dabrowski, M. Scheffler, Theoretical evidence for an optically inducible structural transition of the isolated as antisite in GaAs: identification and explanation of EL2, Phys. Rev. Lett. 60 (1988) 2183–2186. https://doi.org/10.1103/PhysRevLett.60.2183

[30] H.J. von Bardeleben, D. Stiévenard, D. Deresmes, A. Huber, J.C. Bourgoin, Identification of a defect in a semiconductor: EL2 in gaas, Phys. Rev. B 34 (1986) 7192–7202. https://doi.org/10.1103/PhysRevB.34.7192

[31] M.K. Nissen, A. Villemaire, M.L.W. Thewalt, Photoluminescence studies of the EL2 defect in gallium arsenide under external perturbations, Phys. Rev. Lett. 67 (1) (1991) 112–115. https://doi.org/10.1103/PhysRevLett.67.112

[32] C.H. Henry, D.V. Lang, Nonradiative capture and recombination by multiphonon emission in GaAs and GaP, Phys. Rev. B 15 (1977) 989–1016. https://doi.org/10.1103/PhysRevB.15.989

[33] S. Makram-Ebeid, P. Boher, Defect pairs and clusters related to the EL2 centre in GaAs, Rev. Phys. Appliqu/'Ee 23 (5) (1988) 847–862. https://doi.org/10.1051/rphysap:01988002305084700

[34] Y. Oyama, J. Nishizawa, Excitation photocapacitance study of EL2 in n-GaAs and its relation to non-stoichiometry, Mater. Sci. Semicond. Process. 6 (5-6) (2003) 423–426. https://doi.org/10.1016/j.mssp.2003.08.013

[35] B.K. Ridley, Electron-Phonon interactions in semiconductors, Rep. Prog. Phys. 54 (2) (1991) 169–256. https://doi.org/10.1088/0034-4885/54/2/001

[36] P.Y. Yu, M. Cardona, Fundamentals of Semiconductors: Physics and Materials Properties, Springer, 4 edition, 2010. https://doi.org/10.1007/978-3-642-00710-1

[37] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G.L. Chiarotti, M. Cococcioni, I. Dabo, A.D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A.P. Seitsonen, A. Smogunov, P. Umari, R.M. Wentzcovitch, QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials, J. Phys. Condens. Mater. 21 (2006) 395502.

[38] A. Togo, I. Tanaka, First principles phonon calculations in materials science, Scr. Mater. 108 (2015) 1–5.

[39] Z. Tian, K. Esfarjani, G. Chen, Green'S function studies of phonon transport across si/ge superlattices, Phys. Rev. B 89 (2014) 235307.

[40] Y. Guo, Z. Zhang, M. Bescond, S. Xiong, M. Nomura, S. Volz, Anharmonic phonon-phonon scattering at the interface between two solids by nonequilibrium green's function formalism, Phys. Rev. B 103 (2021) 174306.

[41] Y. Guo, M. Bescond, Z. Zhang, S. Xiong, K. Hirakawa, M. Nomura, S. Volz, Thermal conductivity minimum of graded superlattices due to phonon localization, APL Mater. 9 (2021).

[42] Z. Sui, I.P. Herman, Effect of strain on phonons in Si, Ge, and Si/Ge heterostructures, Phys. Rev. B 48 (24) (1993) 17938–17953. https://doi.org/10.1103/physrevb.48.17938

[43] X. Gu, X. Li, R. Yang, Phonon transmission across $Mg_2Si/Mg_2Si_{1-x}Sn_x$ interfaces: a first principles based atomistic Green's function study, Phys. Rev. B 91 (20) (2015). https://doi.org/10.1103/physrevb.91.205313

[44] W. Zhang, T.S. Fisher, N. Mingo, Simulation of interfacial phonon transport in si–Ge heterostructures using an atomistic green's function method, J. Heat Transf. 129 (2007) 483–491.

[45] Y. Guo, M. Bescond, z. Zhang, M. Luisier, M. Nomura, S. Volz, Quantum mechanical modeling of anharmonic phonon-phonon scattering in nanostructures, Phys. Rev. B 102 (2020) 195412.

[46] F. Guinea, C. Tejedor, F. Flores, E. Louis, Effective two-dimensional hamiltonian at surfaces, Phys. Rev. B 28 (1983) 4397.

[47] J.-S. Wang, J. Wang, J.T. Lü, Quantum thermal transport in nanostructures, Eur. Phys. J. B 62 (2008) 381–404.

[48] N. Mingo, L. Yang, Phonon transport in nanowires coated with an amorphous material: an atomistic Green's function approach, Phys. Rev. B 68 (2003) 245406.

[49] R. Rhyner, M. Luisier, Atomistic modeling of coupled electron-phonon transport in nanowire transistors, Phys. Rev. B 89 (2014) 235311.

[50] M. Bosi, G. Attolini, Germanium: epitaxy and its applications, Prog. Cryst. Growth Charact. Mater. 56 (3) (2010) 146–174. https://doi.org/10.1016/j.pcrysgrow.2010.09.002

[51] D.J. Norris, T. Walther, Stranski–Krastanov growth of (Si)Ge/Si(001): transmission electron microscopy compared with segregation theory, Mater. Sci. Technol. 34 (13) (2018) 1539–1548. https://doi.org/10.1080/02670836.2018.1455013

[52] K. Lyutovich, E. Kasper, F. Ernst, M. Bauer, M. Oehme, Relaxed SiGe buffer layer growth with point defect injection, Mater. Sci. Eng.: B 71 (1) (2000) 14–19. https://doi.org/10.1016/S0921-5107(99)00342-6

[53] T. Ulyanenkova, M. Myronov, A. Benediktovitch, A. Mikhalychev, J. Halpin, A. Ulyanenkov, Characterization of SiGe thin films using a laboratory X-ray instrument, J. Appl. Crystallogr. 46 (4) (2013) 898–902. https://doi.org/10.1107/s0021889813010492

[54] X. Deng, P. Namboodiri, K. Li, X. Wang, G. Stan, A.F. Myers, X. Cheng, T. Li, R.M. Silver, Silicon epitaxy on H-terminated Si (100) surfaces at 250 °c, Appl. Surf. Sci. 378 (2016) 301–307. https://doi.org/10.1016/j.apsusc.2016.03.212

[55] M. Vafakhah, S. Janizadeh, Application of artificial neural network and adaptive neuro-fuzzy inference system in streamflow forecasting, Elsevier, 2021, p. 171–191. https://doi.org/10.1016/b978-0-12-820673-7.00002-0

[56] W. Falcon, K. Cho, PyTorch lightning: the lightweight pytorch wrapper for high-performance AI research, in: NeurIPS 2020 Workshop on PyTorch, 2020.

[57] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014. https://doi.org/10.48550/ARXIV.1412.6980

[58] P.J. Huber, Robust estimation of a location parameter, in: Breakthroughs in Statistics: Methodology and Distribution, Springer, 1992, pp. 492–518.

[59] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[60] L. Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, 2004.

[61] N. Hansen, A. Ostermeier, Completely derandomized self-Adaptation in evolution strategies, Evol. Comput. 9 (2001) 159–195. https://doi.org/10.1162/106365601750190398

[62] L. Brekhovskikh, Waves in layered media, 16, Elsevier, 2012.