Optimization of thermionic cooling semiconductor heterostructures with deep learning techniques

Julian G. Fernandez[a*], Guéric Etesse[b], Enrique Comesaña[c], Natalia Seoane[a], Xiangyu Zhu[d], Kazuhiko Hirakawa[d,e], Antonio Garcia-Loureiro[a], and Marc Bescond[b,d]

[a] CiTIUS, University of Santiago de Compostela, Spain (*e-mail: julian.garcia.fernandez2@usc.es)

[b]Aix Marseille Université, CNRS, IM2NP UMR 7334, 133997 Marseille, France

[c] Escola Politécnica Superior de Enxeñaría, University of Santiago de Compostela, Campus Terra, Lugo, Spain

[d]Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

[e] LIMNS-CNRS, IRL 2820, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

***Abstract*—We present a deep learning neural network model to find the AlGaAs-based thermionic cooling structures with the best trade-off between the cooling of the lattice and the cooling of the electrons. These devices are based on the electron-phonon interactions, and therefore, the computational requirements to perform the non-equilibrium Green's function simulations combined with the heat transport and Poisson equations (NEGF+H+P) are very large. The neural network model used is based on the multi-layer perceptron (MLP) machine learning architecture. The comparison between the NEGF+H+P simulations and the values predicted with the MLP gives accurate estimations for the properties studied: gap between the Fermi level of the emitter and the ground state of the quantum well (W), the electron temperature in the quantum well (T_e), and the cooling power of the lattice (CP). Also, after using the MLP to predict one million of different device configurations we found the heterostructures corresponding to the maximun CP, minimun T_e, and the best trade-off between both.**

***Index Terms*—NEGF, Heat transport, Cooling devices, Machine Learning, Refrigeration, Optimization, GaAs.**

I. Introduction

THE integration of cooling technologies based on solid-state physics is one of the most promising solutions to overcome the issues that appear on nanoelectronic circuits due to the self-heating [1].

Classical refrigeration techniques, based on liquid cooling or fanning, cannot avoid the hot spots that appear on such low-scale devices [2]. In this context, asymmetric double barrier heterostructures based on semiconductors have been found to yield significant electron cooling and they are promising candidates for nanometer scale cooling upon optimization [3].

As these devices are based on electron-phonon interactions, the computational requirements of the simulations are very large. This work aims to find, using deep learning techniques (DL), the heterostructure configurations with the best cooling performance, increasing the speed of the searching process while reducing the computational costs.

The contents of this work are distributed as follows. Section II shows the methodology with the explanation of how the thermionic heterostructures operates (II-A), the description of the simulation process (II-B), and the definition of the multi-layer perceptron (MLP) neural network structure used in this work (II-C). Section III presents the results, showing the prediction performance of the MLP, and the optimal device configurations. Finally, Section IV summarizes the main conclusions of this work.

II. Methodology

A. Device structure

The asymmetric double barrier heterostructure (see Fig. 1) is designed to incorporate a GaAs Quantum Well (QW), separated from the GaAs:Si emitter and collector by two barriers, which are made of AlGaAs with varying aluminium concentration. Applying a bias (V) between the two contacts leads to the resonant tunneling injection of electrons in the QW and, subsequently, the extraction of electrons via thermionic emission above the second barrier (b2). This last, thicker barrier acts as a thermal wall to prevent heat backflow.

Cooling in this structure relies on two interconnected behaviors, the evaporation of hot electrons from the QW that lowers the electron temperature (T_e), and the absorption of phonons by the electrons, cooling the lattice, which is measured with the cooling power (CP).

Each heterostructure is defined (as seen in Fig 1) by the combination of the QW length (L_{qw}), the b2 length (L_{b2}), the height of the b2 (h_{b2}) which is proportional to the fraction of Al in the alloy (γ), and V. The devices studied in this work have a constant first barrier length of 1 nm.

The combination of the design parameters determines W, which corresponds to the gap between the QW ground state energy (E_0) and the Fermi energy of the emitter (E_{Fe}), and defines the injection of electrons in the QW.

B. Simulation methodology

To investigate the electron and heat transport in these semiconductor heterostructures, we use an in-house built simulation software that couples self-consistently the non-equilibrium Green's function formalism for electrons with heat and Poisson equations (NEGF+H+P) [3]. This methodology is able to reproduce key aspects of the physics, taking into account

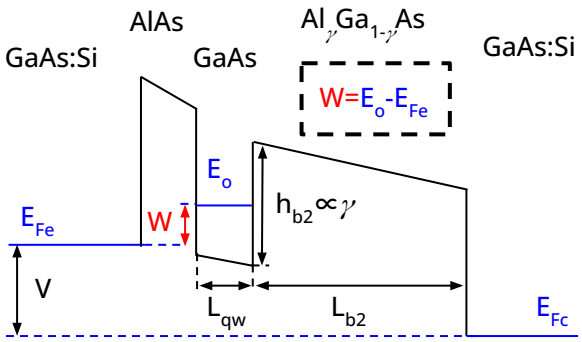

Figure 1: Potential profile of the double barrier heterostructure. L_{qw}, and L_{b2}, are the lenghts of the quantum well (QW) and of the second barrier, respectively. The height of the second barrier (h_{b2}) is proportional to γ, which is the fraction of aluminium in the alloy, and V is the bias between the emitter and the collector. W is defined as the gap between the QW state E_0 and the Fermi level of the emitter E_{Fe}. The AlAs first barrier length is fixed at 1 nm.

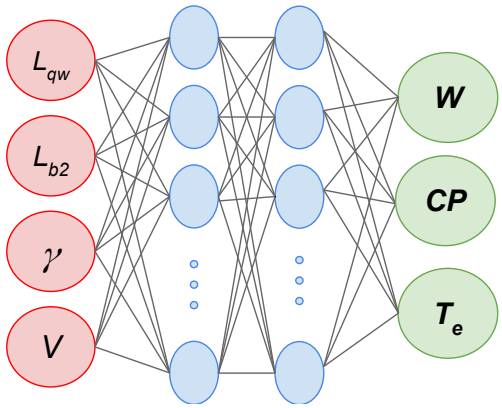

Figure 2: Scheme of the multi-layer perceptron (MLP) neural network. L_{qw}, L_{b2}, γ, and V are the four parameters of the input layer (red). W, T_e, and CP are the 3 parameters of the output layer (green). Also, the MLP has two hidden layers (blue).

thermal, and quantum effects, and the electron transport formalism. Also, the virtual Büttiker probes [4] are used to determine the T_e.

This method relies on the self-consistent calculation of the retarded Green's function at energy E and transverse wavevector k_t that reads:

$$G^r_{k_t} = [(E-U)I - H_{k_t} - \Sigma^r_{L,k_t} - \Sigma^r_{R,k_t} - \Sigma^r_{S,k_t}]^{-1}, \quad (1)$$

where U is the electrostatic potential energy, I is the identity matrix, H_{k_t} is the effective mass Hamiltonian. $\Sigma^r_{L(R),k_t}$ are the self-energies for the left (L) and right (R) semi-infinite device contacts, Σ^r_{S,k_t} is the self-energy calculated within the self-consistent Born approximation (SCBA) that accounts for the interaction between electrons and both the acoustic phonons and polar optical phonons.

The lesser/greater Green's functions are then obtained using the following identities:

$$G^{\lessgtr}_{k_t} = G^r_{k_t}(\Sigma^{\lessgtr}_{L,k_t} + \Sigma^{\lessgtr}_{R,k_t} + \Sigma^{\lessgtr}_{S,k_t})G^{r\dagger}_{k_t}, \quad (2)$$

$$\Sigma^r = \frac{1}{2}[\Sigma^> - \Sigma^<], \quad (3)$$

where the total scattering energy detailed decomposition is detailed in [3]. Obtaining the Green's function then yields many physical properties such as: the electron current density spectrum (in AeV^{-1}m^{-2}) $\mathcal{J}_{j\to j+1}$ from position j to $j+1$:

$$\mathcal{J}_{j\to j+1}(E) = \frac{e}{\hbar}\sum_{k_t}\frac{2n_{k_t}+1}{S}[H_{j,j+1}G^<_{k_t,j+1,j}(E) - G^<_{k_t,j,j+1}(E)H_{j+1,j}], \quad (4)$$

from which we can deduce the electronic energy current that reads:

$$J^E_{j\to j+1} = \int E\mathcal{J}_{j\to j+1}(E)dE, \quad (5)$$

whose first derivative corresponds to the cooling power density (in Wm^{-3}):

$$Q_j = -\nabla_j \cdot J^E \quad (6)$$

Q_j defines the energy transfers between the lattice and the electrons and serves as a source term allowing us to couple electron and heat transport. Finally, integrating the negative part of Q_j over direction of transport yields the Cooling Power (CP), representing the amount of heat removed from the device.

The main drawback of this methodology is the high computational requirements, as the simulation of one device configuration can take a few days on a single CPU core.

C. Neural network calibration

The DL neural network (NN) used in this work is a feed-forward multi-layer perceptron (MLP). The MLP was developed with the Pytorch 1.13.1 [5] and Scikit-learn 1.0.2 [6] libraries on Python 3.8.

The first step to train the neural network is to correctly pre-process the simulation data. One usual and recommended technique is the normalization as part of data preparation before the training process. With the normalization, we change the values of each parameter of the dataset to a common scale, without distorting differences in the ranges of values or losing information. Therefore, it is applied the Scikit-learn standard scaler normalizer separately to inputs and output.

The activation function used in the perceptrons of this NN is the hyperbolic tangent. The selected loss function is the mean square error (MSE), the optimal batch size is 32, and the optimization algorithm is the stochastic gradient descent (SGD) with momentum 0.9 [7]. The adaptative learning rate scheduler technique [8] is applied to avoid the local minimums in minimizing the loss function. The MLP hyperparameters as the previously mentioned or the quantity of hidden layer and their number of neurons, were optimized with the Ray Tune library [9].

The structure of the MLP (see Fig. 2) consists of an input layer with the 4 neurons as the input parameters are L_{qw}, L_{b2}, γ, and V, followed by two hidden layers with 12 and 8 neurons, respectively. The output layer is composed by 3 neurons corresponding to the MLP outputs: CP, T_e and W.

The input NEGF+H+P data to feed the MLP consists of 460 simulated samples, which are split into three datasets: train, validation and test datasets, being their size 294, 74, and 92, respectively.

III. Results

Once the MLP is trained, and before searching for the optimal device configurations, it is crucial to test the MLP performance. Therefore, in Fig 3(a)-(c) the MLP predictions against the NEGF+H+P simulation results for the three output variables are shown for the test dataset, together with their coefficients of determination (R^2). The R^2 is an effective metric for estimating the predictive power of the NN [10]. The test results correspond to R^2 values higher than 0.99 in the case of W (Fig 3(a)) and T_e (Fig 3(b)) and higher than 0.97 for the CP (Fig 3(c)).

With the MLP performance tested, the goal is to find the optimal device configuration that minimizes T_e, and maximizes CP. Hence, a search space is defined with one million different combinations of the input parameters L_{qw}, L_{b2}, γ, and V.

Fig. 4 shows the MLP predictions for the T_e (Y-axis), the W (X-axis), and the CP (colormap) for the search space. The best device configurations are the ones shown in the zoomed inset, corresponding to T_e lower than room temperature ($T_e^{room} = 300$ K, horizontal dashed line) and $CP > 6$ W/mm^2 (blue region). This configurations are obtained for values near the resonance ($W \sim 0$ meV, vertical dashed line).

The time that the trained MLP takes to predict the one million combinations of W, T_e, and CP values is 0.1 s. Here resides the power of using the MLP, as this exhaustive mapping of such large combination of input parameters could not be possible with the NEGF+H+P methodology, taking each single simulation few days.

The best device configurations are displayed in Table I, showing the configuration for $CP^{max} = 6.51$ W/mm^2, and for $T_e^{min} = 264.1$ K. Note that the cooling parameters (T_e, and CP) are not directly correlated (a maximum CP does not implies minimum T_e). Therefore, to find the optimum configuration a trade-off criteria is chosen at the 99% of both, the maximum CP, and the minimum T_e (marked in bold in Table I). This optimum configuration has the following cooling parameters: $CP^{opt} = 6.45$ W/mm^2, and $T_e^{opt} = 266.3$ K.

IV. Conclusions

The thermionic cooling heterostructures based on AlGaAs are promising candidates to be an integral solution to refrigerate nanoelectronic circuits. These devices combine two different cooling mechanisms, the evaporation of hot electrons from the GaAs quantum well, lowering the electron temperature, and the absorption of phonons by the electrons, cooling the lattice.

To simulate the AlGaAs-based heterostructures, we coupled self-consistently the non-equilibrium Green function formalism for electrons with the heat transport equation.

As the computational cost of this simulation metholodogy is very high, we decided to use previously simulated data to find

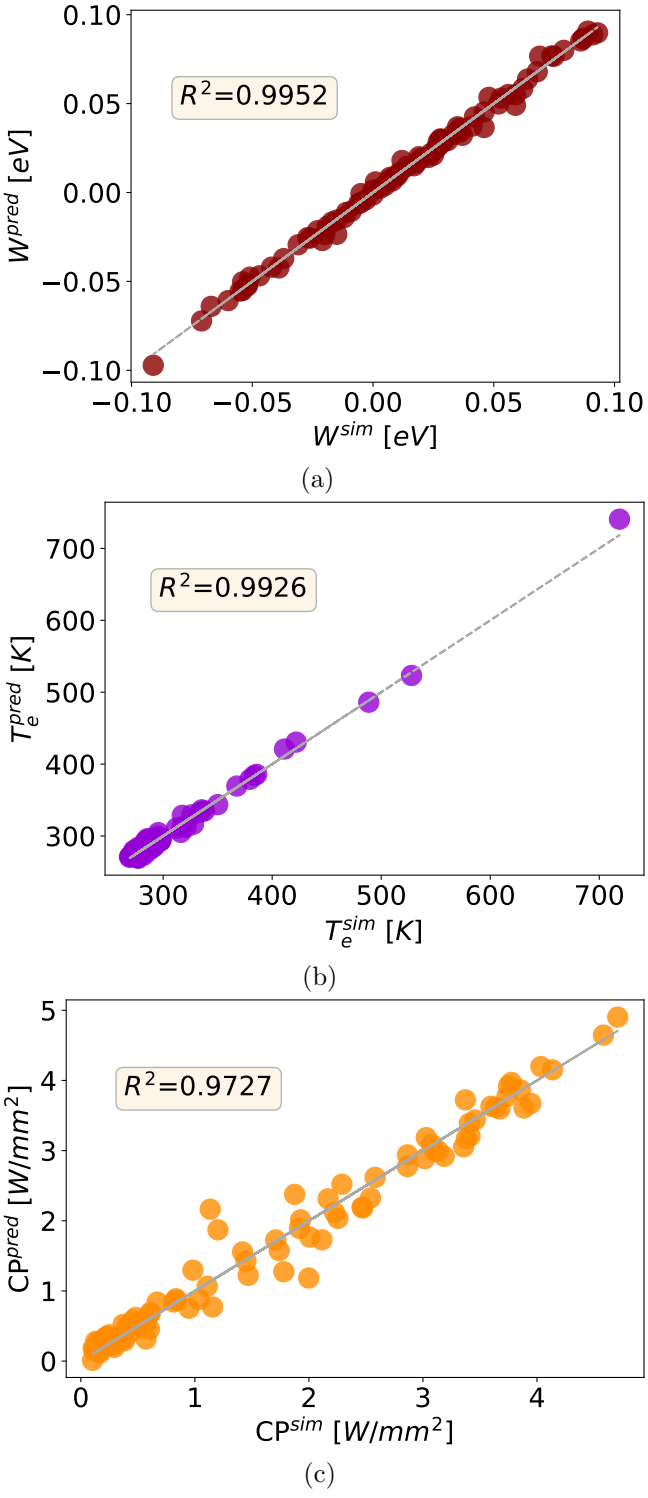

Figure 3: MLP predictions (ordinates) against the simulated NEGF+H+P values (abcissa) with their coefficient of determination R^2 for: (a) W as the energy gap between the Fermi energy of the emitter (E_{Fe}) and the ground state of the quantum well (E_o), (b) the quantum well electron temperature T_e, and (c) the cooling power of the lattice CP.

Configuration	L_{qw} $[nm]$	L_{b2} $[nm]$	γ	V $[V]$	W $[meV]$	CP $[W/mm^2]$	T_e $[K]$
Max(CP)	3.52	50	0.28	0.72	17	6.51	267.7
Min(T_e)	7.20	50	0.17	0.30	3	4.84	264.1
Optimum	**3.36**	**50**	**0.27**	**0.62**	**25**	**6.45**	**266.3**

Table I: Best-performance device configurations. The first row corresponds to the device with maximum lattice CP, the second row to the device with minimum T_e in the QW, and the third row is the optimum configuration that was determined for the best trade-off (99%) between the maximum CP and minimum T_e.

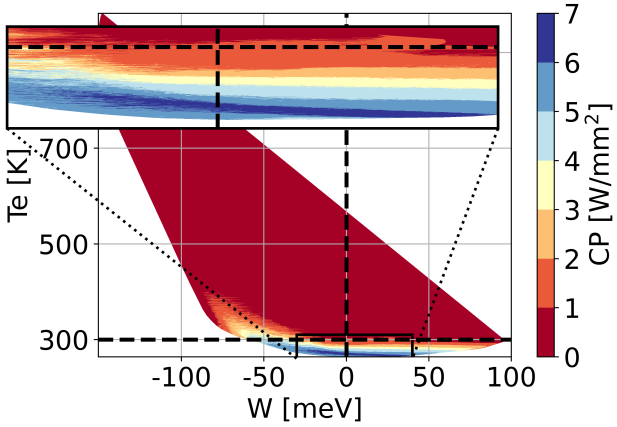

Figure 4: MLP predictions for one million device configurations, where the Y-axis is the T_e, the X-axis is the W, and the colormap is the CP. The zoomed blue region corresponds to the best performance devices, the dashed lines are the room temperature (horizontal) and the resonance injection in the QW (vertical).

the optimal cooling performance for this device by applying neural network techniques. Therefore, we have developed an accurate multi-layer perceptron neural network to find the optimal thermionic cooling heterostructures with the better trade-off between the electron temperature in the quantum well and the cooling power of the lattice.

We have achieved a good prediction accuracy for the thermal device properties T_e, and CP being their coefficient of determination $R^2 = 0.9926$, and $R^2 = 0.9727$, respectively. Also, with the design parameters, we can accurate predict the injection energy in the quantum well as for W we have obtain $R^2 = 0.9952$.

Once the neural network accuracy was demonstrated, we have applied it to one million combinations of the heterostructure design parameters finding the configurations with maximum CP ($L_{qw} = 3.52\ nm$, $L_{b2} = 50\ nm$, $\gamma = 0.28$, $V = 0.72\ V$), minimum T_e ($L_{qw} = 7.20\ nm$, $L_{b2} = 50\ nm$, $\gamma = 0.17$, $V = 0.30\ V$), and the best trade-off between both ($L_{qw} = 3.36\ nm$, $L_{b2} = 50\ nm$, $\gamma = 0.27$, $V = 0.62\ V$).

The combination of the non-equilibrium Green's function and heat transport with machine learning techniques has allowed us to drastically decrease the computational requirements to perform the heterostructure optimization process, as the prediction of one million of device configurations taked 0.1 s in comparison with the few days that each device simulation takes in the NEGF+H+P methodology.

References

[1] A. Ziabari, M. Zebarjadi, D. Vashaee, and A. Shakouri, "Nanoscale solid-state cooling: a review," *Reports on Progress in Physics*, vol. 79, no. 9, p. 095901, aug 2016. doi: 10.1088/0034-4885/79/9/095901. [Online]. Available: https://dx.doi.org/10.1088/0034-4885/79/9/095901

[2] R. Gaska, A. Osinsky, J. Yang, and M. Shur, "Self-heating in high-power AlGaN-GaN HFETs," *IEEE Electron Device Letters*, vol. 19, no. 3, pp. 89–91, 1998. doi: 10.1109/55.661174

[3] M. Bescond, G. Dangoisse, X. Zhu, C. Salhani, and K. Hirakawa, "Comprehensive Analysis of Electron Evaporative Cooling in Double-Barrier Semiconductor Heterostructures," *Phys. Rev. Appl.*, vol. 17, p. 014001, Jan 2022. doi: 10.1103/PhysRevApplied.17.014001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevApplied.17.014001

[4] A. Shastry and C. A. Stafford, "Temperature and voltage measurement in quantum systems far from equilibrium," *Phys. Rev. B*, vol. 94, p. 155433, Oct 2016. doi: 10.1103/PhysRevB.94.155433. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.94.155433

[5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[6] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[7] N. Ketkar, *Stochastic Gradient Descent*. Berkeley, CA: Apress, 2017, pp. 113–132. ISBN 978-1-4842-2766-4. [Online]. Available: https://doi.org/10.1007/978-1-4842-2766-4_8

[8] Z. Xu, A. M. Dai, J. Kemp, and L. Metz, "Learning an adaptive learning rate schedule," *arXiv preprint arXiv:1909.09712*, 2019.

[9] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *arXiv preprint arXiv:1807.05118*, 2018.

[10] D. Chicco, M. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 07 2021. doi: 10.7717/peerj-cs.623

Aknowledgment

This work was supported by the Spanish MICINN, Xunta de Galicia, and FEDER Funds under Grant RYC-2017-23312, Grant PID2019-104834GB-I00, Grant ED431F 2020/008, Grant ED431C 2022/16 and GELATO ANR project (ANR-21-CE50-0017)