Prediction of the band gap in 2D materials using active learning

Author: Alba Quiñones Andrade
Advisor: Dr. Adriana Isabel Figueroa
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Dr. Jose Hugo Garcia
Theoretical and Computational Nanoscience Group,
ICN2 - Catalan Institute of Nanoscience and Nanotechnology,
Campus UAB, 08193 Bellaterra (Barcelona), Spain

Abstract: Accurate band gap prediction of two-dimensional materials holds significant scientific and technological value for the development of electronic and optoelectronic devices. In contrast to the high computational cost associated with traditional first-principles methods, machine learning offers a promising and cost-effective alternative for band gap prediction. In this work, we demonstrate that the combination of artificial neural networks and an active learning algorithm leads to a highly data-efficient method for predicting band gaps of 2D materials while maintaining accuracy, with L1-regularization analyzing feature selection. This approach achieves a computational cost reduction by shrinking the original dataset by 80% compared to traditional training approaches.
Keywords: Materials Science, Computational Physics
SDGs: ODS7 - ODS9 - ODS13 (see page 6)

I. INTRODUCTION

Two-dimensional (2D) materials, first brought to light with the discovery of graphene, have opened up a great deal of scientific and technological possibilities [1]. These materials are composed of a single atomic layer held together by strong covalent bonds. This unique atomic composition, high surface-to-volume ratio, and reduced dimensionality give 2D materials exceptional electrical conductivity, thermal transport, and mechanical strength [2, 3].

Among their extensive range of electronic properties, the customizable band gap plays a crucial role in various applications. However, its theoretical calculation can become computationally expensive, depending on the required accuracy.

Machine learning (ML) has been used to predict various material properties, but it requires training data. Such data is scarcer in the field of 2D materials, where the number of discovered materials is significantly smaller than in traditional three-dimensional systems. Active Learning (AL) is a machine learning strategy that improves model performance by selecting the most informative samples for labeling, thereby reducing the associated computational costs without compromising accuracy. In contrast, Deep Learning (DL) excels at automatically extracting complex features but depends on large labeled datasets for training [4]. Therefore, combining these approaches enables the development of high-performance models trained efficiently on a minimal dataset.

In this project, we employ a computational approach to predict the electronic band gaps of 2D materials using artificial neural networks (NNs) combined with AL to improve prediction efficiency with a limited amount of labeled data. The structure of the work is as follows: first, the fundamental principles of NNs and AL are introduced, including the construction of the network and the training process. Next, a model is developed that integrates AL, NNs, and regularization techniques. Finally, the model's performance is evaluated using a dataset of 2D materials from the Computational 2D Materials Database (C2DB).

II. METHODS

An NN is a ML model inspired by the brain's structure and functionalities, particularly its non-linearity and dense connectivity, features that underlie human cognitive flexibility [5].

A NN is composed of interconnected units called perceptrons. A perceptron is a mathematical function that takes an n_i-dimensional (n_i-D) input vector, $\mathbf{x} \in \mathbb{R}^{n_i}$, and maps it linearly into a scalar output, $y = \sum_{i=0} \omega_i x_i + b$, with $\mathbf{w} \in \mathbb{R}^{n_i}$ and $b \in \mathbb{R}$ being free parameters commonly referred to as the mixing weights and the output bias respectively. This output is then passed through a non-linear function $z = f(y)$, typically a sigmoid due to its resemblance to a neuron's response. Such a description, although simple, serves as the basis for binary classifiers and is the building block of an NN.

A combination of these perceptrons is known to be an effective method for approximating a non-linear function due to the universal approximation theorem [6], which states that any n_o-D vector field, $\mathbf{y} \in \mathbb{R}^{n_o}$, can be represented as a given linear combination of non-linear functions (more details in Appendix V. D). Therefore, each of its perceptron components can be given by $y_k^{m_{\alpha+1}} = \sum_j \omega_{kj} z_j^{m_\alpha} + b_k^{m_{\alpha+1}}$, where the weights ω_{kj} and biases $b_k^{m_{\alpha+1}}$ can be arranged in the form of an $m_{\alpha+1} \times m_\alpha$ matrix, W, and an $m_{\alpha+1}$-D vector, $\mathbf{b}$, respectively. Furthermore, the m_α-D vector $\mathbf{z}^{\mathbf{m}_\alpha}$ is given

by a collection of perceptrons connected to the input and is known as a *layer*. Each of the components of $\mathbf{z}^{\mathbf{m}_\alpha}$ is computed in analogy with the previous discussion as $z_k^{m_\alpha} = f\left(\sum_i \omega_{ki} x_i + b_k^{m_\alpha}\right)$, where now the connecting weights ω_{ij} and biases $b_k^{m_\alpha}$ are arranged as an $m_\alpha \times n_i$ matrix $W_{n_i \to m_\alpha}$ and a m_α-dimensional vector $\mathbf{b}_\alpha^m$, respectively.

The entire discussion above forms the foundation for more complex NNs. To this end, let us start by considering a m_α-D perceptron layer (H) vector, $\mathbf{z}^{m_\alpha}$, connected to a previous $m_{\alpha-1}$-D perceptron layer (H-1) vector, $\mathbf{z}^{m_{\alpha-1}}$, via the weight matrix $W^{H,H-1}_{m_{\alpha-1}\to m_\alpha}$ and the bias vector $\mathbf{b}^{m_\alpha}$, as depicted in Fig. 1. Therefore, each component of that *hidden* layer is given by

$$z_k^{m_\alpha} = f\left(\sum_j \omega_{kj}^{H,H-1} z_j^{m_{\alpha-1}} + b_k^{m_\alpha}\right). \tag{1}$$

This is an iterative relation that connects an arbitrary n_i-D input to an n_0-D output through a sequence of matrix-vector products involving all the weight matrices, $W^{1,0}, W^{2,1}, \ldots, W^{H,H-1}$, associated with each of these H hidden layers. In this project, H=2, $\alpha = 1, 2$.

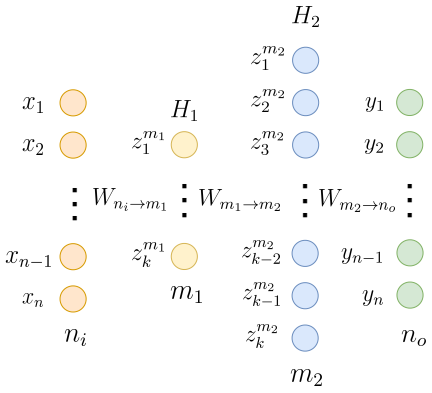


FIG. 1: Architecture of a neural network with two hidden layers, showing the size of each layer (m_α) and the weight matrices ($W_{m_{\alpha-1}\to m_\alpha}$) connecting them with $\mathbf{z}^{m_\alpha} = f(\mathbf{y}^{m_\alpha}) = f(W_{m_{\alpha-1}\to m_\alpha}\mathbf{z}^{m_{\alpha-1}} + \mathbf{b}^{m_\alpha})$.

As mentioned before, the ultimate success of an NN depends on the set of weight matrices, collectively denoted as W (from this point on, we will incorporate both the weights and the bias into the W to ease the notation), used to compute the output. A training process is required to obtain this set of parameters. This involves randomly initializing the parameters and then iteratively minimizing an error or loss function concerning them. In the present work, the error function is the mean squared error (MSE), which is defined as:

$$\mathcal{L}(\mathbf{y}, \mathbf{y}_{\text{exp}}; W) = \frac{1}{n_o}\sum_{i=1}^{n_o}(y_i(W) - y_{\text{exp},i})^2, \tag{2}$$

here, y_i is the i-th component of the network's output vector (which, in this case, is the predicted band gap), and $y_{\text{exp},i}$ is the corresponding true value of the band gap. There are many procedures used to obtain the weights that minimize the loss function. We chose the Stochastic Gradient Descent (SGD) algorithm since it is efficiently implemented in many libraries and can be parallelized to process the dataset in mini-batches rather than as a whole. Once the algorithm has cycled through all batches, it is said that an *epoch* has been completed, and the process is typically repeated over several epochs until the parameters converge [5]. The idea behind this algorithm is that if we start with random weights, we can get closer to the minimum of the loss function by iterating the weights parameters according to the following relation:

$$W_{e+1} = W_e - l_r \nabla_W \mathcal{L}(\mathbf{y}, \mathbf{y}_{\text{exp}}), \tag{3}$$

here, l_r is the learning rate, and $\nabla_W \mathcal{L}(\mathbf{y}, \mathbf{y}_{\text{exp}})$ is the gradient of the loss function, computed using the backpropagation method (explained in Appendix V. A). The differentiability of all operations described so far is crucial for computing the gradient and applying backpropagation. Additionally, the choice of the learning rate is critical to ensure faster convergence; therefore, we used the Adam optimizer, which enables a dynamic adjustment of learning rates for each parameter [5]. This makes it particularly effective in scenarios with sparse or noisy gradients, resulting in faster and more stable convergence.

To evaluate the performance of the NN for band gap prediction, a feedforward NN was trained. This model consisted of two hidden layers with 8 and 32 neurons, respectively. The data was divided into batches $\mathcal{B}_n$, $\mathcal{D} = \{(x_k, y_k)\}_{k=1}^N = \cup_{n=1}^{N_B}\mathcal{B}_n$, with 16 samples per batch.

A. Data preparation and feature engineering

As the main source of our data, we considered the C2DB [7], which contains the band gaps of thousands of 2D materials computed theoretically using density functional theory (DFT) [8]. In DFT, the many-body electronic problem is mapped onto a system of non-interacting electrons moving in an effective potential, allowing the ground-state energy to be found by minimizing it concerning the electron density. All complex many-body effects, such as exchange and correlation, are contained within an approximate functional [9]. One of the most popular approximations is the Perdew–Burke–Ernzerhof (PBE) functional, which is computationally efficient but systematically underestimates band gaps. The Heyd–Scuseria–Ernzerhof (HSE06) hybrid functional offers an improvement by incorporating a fraction of exact exchange, which partially corrects for the self-interaction error inherent in the PBE functional. While HSE06 performs well for bulk semiconductors, its performance for the band gaps of 2D materials can be more variable [10]. Finally, the GW approximation includes many-body effects perturbatively, and

although it is more computationally demanding, it offers significantly higher accuracy [9, 10]. The C2DB contains many band gaps computed using the three methods. To ensure accuracy, we selected only those materials with a finite direct band gap ($E_g > 0$) within the GW method.

The resulting dataset was preprocessed to handle missing values (NaNs), and features that were trivially related to the band gaps or the conduction and valence bands were eliminated.

B. Active learning and regularization

We chose a pool-based active learning framework where an algorithm selects the most informative data to train an ML system based on a continuous feedback loop between the oracle, which is the source of the data labels, and the learner, which is the surrogate model that intends to describe the dataset; in this case, a NN that predicts the band gaps of 2D materials [11, 12].

At each iteration, the NN model identifies the most informative samples, x_N, by estimating its uncertainty over the unlabelled pool $\mathcal{X}$. The discrepancy between the predicted and true band gap values is then computed for each sample. In the next iteration, the sample x_{N+1} to be labeled is selected as the one with the highest error, allowing the model to focus on poorly predicted regions. This process enables the model to gain knowledge about previously unknown regions and distinguish between informative points and outliers. The learner then uses the updated predictions to decide which sample to query next.

These newly labeled samples are added to the training set, refining the NN and progressively improving its accuracy. With each query, the surrogate model incorporates new data, enhancing its approximation of the band gap function across the materials domain. (Pseudocode in Appendix V. C).

While training is performed exclusively on the actively selected subset $\mathcal{L}$, predictions are generated over the entire dataset $\mathcal{X}$. This allows the surrogate model to generalize band gap estimates across all candidate materials, leveraging the information gained from the most informative samples.

Several challenges may arise during model training and evaluation, one of the most common being overfitting. Overfitting occurs when a model learns the training data too well, including its noise and outliers, resulting in poor generalization to unseen data. Developing strategies to reduce overfitting systematically is a key and ongoing area of research in machine learning. We use L1-regularization due to its simple yet efficient implementation [5]. In this method, the loss function is modified in the following way:

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_n w_n \tag{4}$$

where the parameter $\lambda > 0$ includes a penalty over large absolute values of the model parameters, thus encouraging the model to reduce unnecessary complexity by reducing some of the model parameters to zero. Such behavior also makes L1-Regularization a well-suited approach for feature selection, as when applied to the initial layer, it forces the model to train only with the most relevant inputs. In this project, we employ the LASSO (least absolute shrinkage and selection operator) method (Tibshirani, 1995) [5], which combines L1-regularization with a least-squares cost function to generate models that are both interpretable and robust, as it automatically selects a relevant subset of input features while reducing the risk of overfitting.

Our primary goal is to train an NN with a minimal training dataset selected efficiently using AL. We will assume that labeling or generating a dataset is computationally expensive, whereas validation—evaluating the band gap through the surrogative model—is not. To assess the quality of the approach, we compared the predicted value against the real band gap. We consider a perfect fit when the resulting curve matches the identified function ($y = x$), and a reasonable error margin would be 1.5% around the identity.

In parallel, we will apply LASSO to select the most relevant features, and we will analyze their physical significance in the Results section. Since LASSO does not provide a direct way to classify those features, we will use the elbow method (see Pseudocode in Appendix V. B) to determine a threshold based on the trade-off between model complexity and performance. Such threshold is found at the point of maximum curvature in a plotted curve of sorted feature importances. Features above this threshold are retained, resulting in a more compact dataset with both a reduced sample size and a reduced set of informative features.

The refined dataset from the combined methodology would then be used to train the NN. The effectiveness of this approach is again measured by how closely the predicted versus actual band gap values align with the $y = x$, yielding a prediction error of 5 %.

III. RESULTS

To evaluate the potential variability in the effectiveness and accuracy of our AL methodology, we perform a statistical study and consider 100 whole iterations of the model. The NN requires a dataset of 356 2D materials and 77 features as input. The first analysis in Fig. 2 compares the predicted band gaps of all 2D materials in the validation set. The predictions are obtained using AL (Fig. 2. A and Fig. 2. C) and NN (Fig. 2. B and Fig. 2. D) approaches before (Fig. 2. A and Fig. 2. B) and after LASSO (Fig. 2. C and Fig. 2. D) feature selection.

The training based on AL yields a better prediction in both cases before applying feature selection and after selecting the features. On the other hand, the perfor-

mance of the NN—trained on a dataset with the same number of features and samples as the AL training set but composed of randomly selected 2D materials—is also evaluated. This comparison highlights the superior performance of AL, which selects training samples based on uncertainty, leading to more informative and efficient learning.

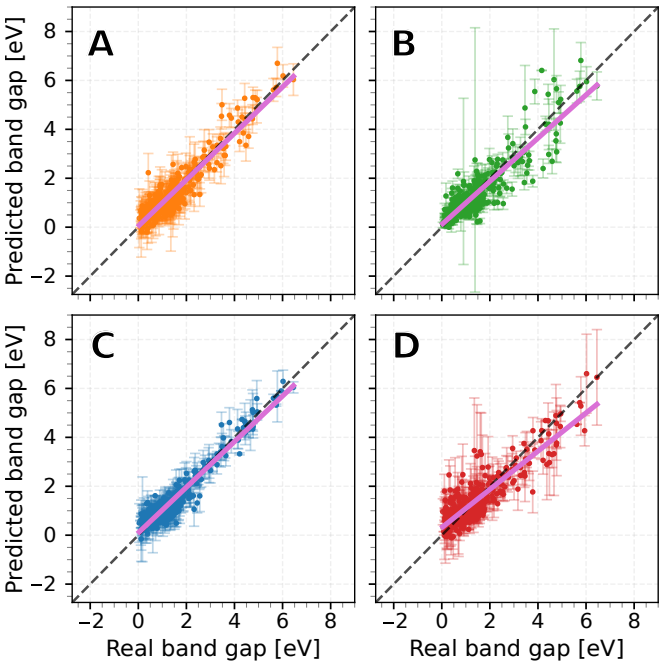


FIG. 2: Linear regression analysis of predicted versus real band gap values obtained from the GW model. Panels A) and B) illustrate the prediction performance before applying LASSO for feature selection. Panels C) and D) present the performance after LASSO-based feature selection. The blue and orange markers show AL results, and the green and red markers show NN results.

The second analysis in Fig. 3 A shows the distribution of the root mean squared error (RMSE) to evaluate the regression performance of the ML models. This is visualized with violin plots, which display the spread of the data and how often values occur at different levels. The AL model achieves a lower RMSE, with a distribution more tightly concentrated around the mean, indicating greater consistency and lower variability across iterations. In contrast, the NN model exhibits a broader RMSE distribution, with higher mean RMSE values, indicating less stable and less accurate predictions. The increased variability observed in the NN, both before and after LASSO, is due to the random selection of training points. When features are reduced, information is lost, and the effect of randomness becomes more pronounced.

During the feature selection process, LASSO is applied in each iteration to identify the most informative features. The analysis, summarized in Fig. 3 B, shows the average selection frequency of each feature across all iterations. The six features with a selection frequency exceeding 50 % are listed and described in Table. I. The

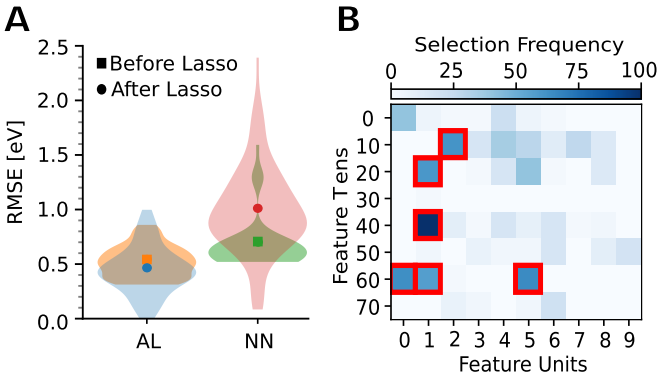


FIG. 3: A. Regression performance of both methodologies, represented by the RMSE. On the left, the performance of the AL algorithm is illustrated through the RMSE distribution and its corresponding mean, with orange representing the results before feature selection and blue representing the results after feature selection. On the right, the RMSE distribution for the NN is displayed, with green indicating performance before feature selection and pink after. B. Feature selection frequency is represented in a matrix format, where the tens digit of each feature index is placed along the vertical axis and the unit digit along the horizontal axis. The color bar indicates the average selection frequency of each feature across all iterations. Features highlighted in red are those with a selection frequency of 50 % or higher.

TABLE I: Selected features used in the model with their descriptions and physical units [7].

Index	*Feature*	*Description*	*Units*
12	**etot**	Total energy	eV
21	**dim_nclusters_2D**	Number of disconnected atomic clusters in the 2D unit cell	1 (dimensionless)
41	**hform**	Heat of formation	eV/atom
60	**alphax_el**	Static interband polarisability along x	Å
61	**alphay_el**	Static interband polarisability along y	Å
65	**E_B**	Binding energy	eV

selected features are physically meaningful descriptors of the properties that govern the band gap in 2D materials (more details in Appendix V. E). Features 12 and 41 characterize electronic stability, where stable configurations exhibit clearer band gap separations. Cluster connectivity (21) directly affects electron delocalization: more disconnected clusters confine electrons locally, widening the bandgap, while connected structures allow extended wavefunctions and narrower gaps. Static polarizabilities (60, 61) quantify how easily electrons respond to external fields—higher polarizability indicates greater electron mobility and smaller bandgaps. Finally, binding energy (65) measures the rigidity of atomic bonds; stronger binding reduces electronic flexibility, which in turn influences

the magnitude of the gap. Since the width of the band gap determines whether a material behaves as a semiconductor or an insulator, these features are crucial for predicting the properties of a broader range of materials.

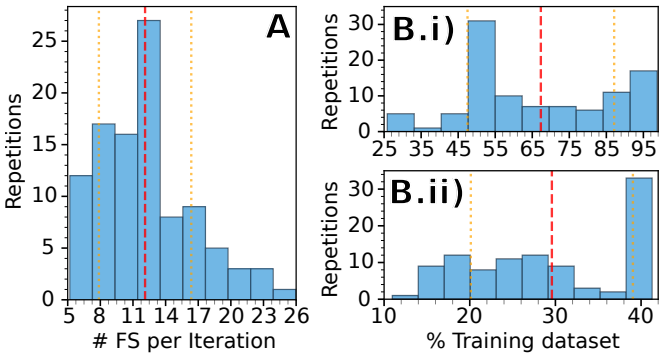


FIG. 4: Distributions of the optimal values found after multiple optimization iterations for different hyperparameters: (A) number of features selected (FS) per iteration; (B.i) percentage of the dataset used for training the second AL; (B.ii) percentage of the dataset used for training the first AL. Dashed lines indicate the mean (red) and quartiles (orange).

Finally, a statistical analysis is presented in Fig. 4, showing that, on average, in Fig. 4 A, 12 features are selected in each iteration (12.11 ± 4.25). Additionally, Fig. 4 B.ii) illustrates that in over 50% of the iterations, the original C2DB dataset is reduced by approximately 40% (the mean is 29.59 ± 9.46). In the second active learning algorithm in Fig. 4 B.i), the dataset is further reduced by about 50% relative to the first reduced dataset (the mean is 67.32 ± 19.79).

IV. CONCLUSIONS

As the results demonstrate, in agreement with the theoretical framework presented, the active learning model developed in this TFG successfully predicts the band gap while using only approximately 40% of the first training data. This reduction is achieved without compromising prediction accuracy and significantly lowers computational costs. On the one hand, AL has proven to outperform a NN trained on a randomly selected subset of the same size by prioritizing the most informative samples. On the other hand, integrating AL with LASSO Feature Selection in the second training enhances performance and efficiency, enabling accurate predictions while minimizing computational requirements. Moreover, the ability to select the most relevant features enables a better understanding of the key characteristics that most significantly influence the band gap of 2D materials. This means that if only these features are experimentally or computationally available, the model can still be effectively trained. This contributes to a more sustainable and time-efficient workflow, guiding future material studies to focus on the most informative descriptors.

Acknowledgments

I would like to thank Marc, my mother and grandparents, and my two advisors for their unwavering support. My advisors not only guided me seamlessly through this project but also became role models for the kind of scientist I aspire to become.

[1] Novoselov KS, Geim AK, Morozov SV, et al. "Electric field effect in atomically thin carbon films". *Science*. **306** (5696): 666-669 (2004)
[2] Butler SZ, Hollen SM, Cao L, et al. "Progress, challenges, and opportunities in two-dimensional materials beyond graphene". *ACS Nano*. **7** (4): 2898-2926 (2013)
[3] Shanmugam V, Mensah RA, Babu K, et al. "A Review of the Synthesis, Properties, and Applications of 2D Materials". *Particle & Particle Systems Characterization*. **39** (6): 2200031 (2022)
[4] Ren P, Xiao Y, Chang X, et al. "A Survey of Deep Active Learning". *ACM Computing Surveys*. **54** (9): 1-40 (2021)
[5] Goodfellow I, Bengio Y, Courville A. "Deep Learning". MIT Press (2016)
[6] Cybenko G. "Approximation by Superpositions of a Sigmoidal Function". *Mathematics of Control, Signals and Systems*. **2** (4): 303-314 (1989)
[7] Haastrup S, Strange M, Pandey M, et al. "The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals". *2D Materials*. **5** (4): 042002 (2018)
[8] Tran F, Doumont J, Kalantari L, et al. "Bandgap of two-dimensional materials: Thorough assessment of modern exchange–correlation functionals". *Journal of Chemical Physics*. **155** (10): 104103 (2021)
[9] Kaczkowski J. "Electronic Structure of Some Wurtzite Semiconductors: Hybrid Functionals vs. Ab Initio Many Body Calculations". *Proceedings of the European Conference Physics of Magnetism*. **121** (5-6): 1020-1024 (2012)
[10] Patra A, Jana S, Samal P, et al. "Efficient Band Structure Calculation of Two-Dimensional Materials from Semilocal Density Functionals". *Journal of Physical Chemistry C*. **125** (20): 11206-11215 (2021)
[11] Burr Settles, *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, Lecture #18. Springer (2022)
[12] Kusne AG, Yu H, Wu C, et al. "On-the-fly closed-loop materials discovery via Bayesian active learning". *Nature Communications*. **11**: 5966 (2020)
[13] Zhang Y, Xu W, Liu G, et al. "Bandgap prediction of two-dimensional materials using machine learning". *PLOS ONE*. **16** (8): e0255637 (2021)

Predició del band gap en materials 2D mitjançant aprenentatge actiu

Author: Alba Quiñones Andrade
Advisor: Dr. Adriana Isabel Figueroa
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

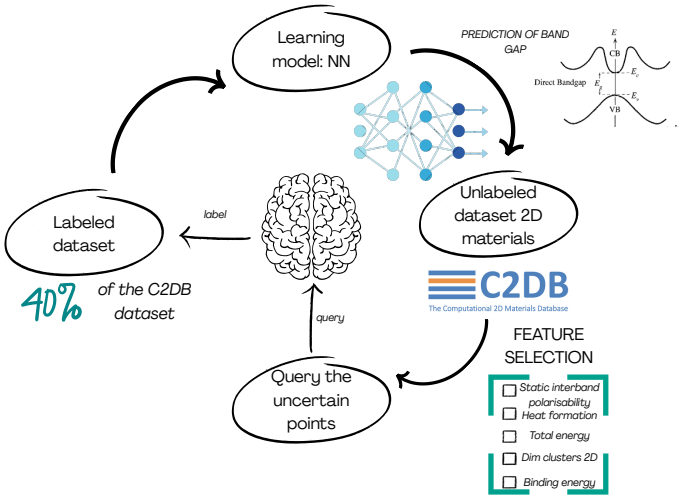
Advisor: Dr. Jose Hugo Garcia
Theoretical and Computational Nanoscience Group,
ICN2 - Catalan Institute of Nanoscience and Nanotechnology,
Campus UAB, 08193 Bellaterra (Barcelona), Spain

Resum: La predicció precisa del band gap en materials bidimensionals té un valor científic i tecnològic significatiu per al desenvolupament de dispositius electrònics. En contrast amb el cost computacional elevat associat als mètodes tradicionals basats en primers principis, l'aprenentatge automàtic ofereix una alternativa prometedora i eficient per a la predicció del band gap. En aquest treball, demostrem que la combinació de xarxes neuronals artificials amb un algoritme d'aprenentatge actiu dona lloc a un mètode altament eficient en dades per predir el band gap de materials 2D mantenint l'exactitud. Aquest enfocament permet una reducció del cost computacional reduint el conjunt de dades original en un 80 % en comparació amb els mètodes d'entrenament tradicionals.
Paraules clau: Ciència de Materials, Física Computacional

Objectius de Desenvolupament Sostenible (ODSs o SDGs)

1. Fi de la es desigualtats		10. Reducció de les desigualtats	
2. Fam zero		11. Ciutats i comunitats sostenibles	
3. Salut i benestar		12. Consum i producció responsables	
4. Educació de qualitat		13. Acció climàtica	X
5. Igualtat de gènere		14. Vida submarina	
6. Aigua neta i sanejament		15. Vida terrestre	
7. Energia neta i sostenible	X	16. Pau, justícia i institucions sòlides	
8. Treball digne i creixement econòmic		17. Aliança pels objectius	
9. Indústria, innovació, infraestructures	X		

El contingut d'aquest Treball de Fi de Grau, emmarcat en un grau universitari de Física, es relaciona amb l'ODS 7 (fita 7.3), ja que la reducció del cost computacional derivada de l'ús de tècniques d'aprenentatge actiu contribueix a una millora de l'eficiència energètica en la recerca científica. També es vincula amb l'ODS 9 (fita 9.5), atès que el desenvolupament de models de ML aplicats a materials 2D representa un avenç significatiu en la investigació científica i impulsa la innovació en àmbits com la física computacional i la ciència de materials. Així mateix, es relaciona amb la fita 9.4, ja que l'estudi eficient de les propietats electròniques d'aquests materials pot afavorir el desenvolupament de tecnologies més netes, sostenibles i eficients. Finalment, es vincula amb l'ODS 13 (fita 13.1) perquè els materials 2D poden ser clau en la generació d'energia renovable, contribuint a la mitigació del canvi climàtic.



V. APPENDIX

A. Theory framework of Backpropagation

Let $z^0 = \mathbf{x}$ be the input and $z^{H+1} = \mathbf{y}$ the final output. Then:

$$z^\alpha = \sigma(W^{\alpha,\alpha-1} z^{\alpha-1} + \mathbf{b}^{(\alpha)}),$$

where the activation function σ is applied element-wise. In index notation:

$$z_m^\alpha = \sigma\left(\sum_{n=1}^{m_{\alpha-1}} W_{mn}^{\alpha,\alpha-1} z_n^{\alpha-1} + b_m^{(\alpha)}\right).$$

Assume a quadratic loss function:

$$L(\mathbf{y}) = \|\mathbf{y}_{\text{exp}} - \mathbf{y}\|^2 = \sum_{i=1}^{n_{H+1}} (y_i^{\text{exp}} - y_i)^2.$$

Then,

$$\frac{\partial L}{\partial y_i} = -2(y_i^{\text{exp}} - y_i).$$

Gradients of the Weights

We aim to compute:

$$\frac{\partial L}{\partial W_{mn}^{\alpha,\alpha-1}} = \frac{\partial L}{\partial z_m^\alpha} \cdot \frac{\partial z_m^\alpha}{\partial u_m^\alpha} \cdot \frac{\partial u_m^\alpha}{\partial W_{mn}^{\alpha,\alpha-1}},$$

where $u_m^\alpha = \sum_n W_{mn}^{\alpha,\alpha-1} z_n^{\alpha-1} + b_m^{(\alpha)}$. Thus,

$$\frac{\partial u_m^\alpha}{\partial W_{mn}^{\alpha,\alpha-1}} = z_n^{\alpha-1}.$$

So:

$$\frac{\partial L}{\partial W_{mn}^{\alpha,\alpha-1}} = \delta_m^\alpha \cdot z_n^{\alpha-1},$$

with:

$$\delta_m^\alpha = \frac{\partial L}{\partial z_m^\alpha} \cdot \sigma'(u_m^\alpha).$$

Recursive Formulation: Backpropagation of Errors

To compute the derivative with respect to z_m^α, we use the chain rule:

$$\frac{\partial L}{\partial z_m^\alpha} = \sum_{l=1}^{m_{\alpha+1}} \frac{\partial L}{\partial z_l^{\alpha+1}} \cdot \frac{\partial z_l^{\alpha+1}}{\partial u_l^{\alpha+1}} \cdot \frac{\partial u_l^{\alpha+1}}{\partial z_m^\alpha}.$$

But:

$$\frac{\partial u_l^{\alpha+1}}{\partial z_m^\alpha} = W_{lm}^{\alpha+1,\alpha}.$$

So:

$$\frac{\partial L}{\partial z_m^\alpha} = \sum_{l=1}^{m_{\alpha+1}} \delta_l^{\alpha+1} W_{lm}^{\alpha+1,\alpha}.$$

Putting it all together:

$$\delta_m^\alpha = \sigma'(u_m^\alpha) \sum_{l=1}^{m_{\alpha+1}} \delta_l^{\alpha+1} W_{lm}^{\alpha+1,\alpha}.$$

Matrix Form of Backpropagation

Define $^\alpha$ as the vector of deltas at layer α, and let $\odot$ denote the Hadamard (element-wise) product. Then:

$$\delta^\alpha = \left((W^{\alpha+1,\alpha})^T \delta^{\alpha+1}\right) \odot \sigma'(u^\alpha).$$

And the gradient of the loss with respect to weights is:

$$\frac{\partial L}{\partial W^{\alpha,\alpha-1}} = \delta^\alpha \cdot (z^{\alpha-1})^T.$$

B. Elbow Method

Algorithm 1 elbow method for feature importance threshold

input: sorted list of importances $\{w_1, w_2, \ldots, w_n\}$ in descending order
define $P = \{(i, w_i)\}_{i=1}^n$ as list of (index, importance) coordinates
let $a = P_1$, the first point
let $b = P_n$, the last point
compute direction vector $\vec{v} = \frac{b-a}{\|b-a\|}$
for each point P_i in P **do**
    compute $\vec{u}_i = P_i - a$
    project onto $\vec{v}$: $\vec{p}_i = (\vec{u}_i \cdot \vec{v}) \cdot \vec{v}$
    compute orthogonal vector: $\vec{d}_i = \vec{u}_i - \vec{p}_i$
    compute distance: $d_i = \|\vec{d}_i\|$
end for
find $i^* = \arg\max_i d_i$
return i^* and w_{i^*} as elbow threshold

C. Pseudocode AL

Algorithm 2 Active Learning Process

Require: Dataset $\mathcal{X}$, labels y, scalers `scaler_X`, `scaler_y`, number of iterations T, initial percentage p_0, increment percentage p, neural network architecture, training parameters

Train a NN model $\mathcal{N}_{fs}$ on $(\mathcal{X}, y)$
Analyze weights of $\mathcal{N}_{fs}$ for feature selection
Initialize labeled set $\mathcal{L} \leftarrow$ random subset of $\mathcal{X}$ with proportion p_0
Set unlabeled pool $\mathcal{U} \leftarrow \mathcal{X} \setminus \mathcal{L}$
Compute real target values y_{real} from y using `scaler_y`
for $t = 1$ to T **do**
    Train a model $\mathcal{N}_t$ on $\mathcal{L}$
    Predict band gap for all $\mathcal{X}$: $\hat{y} \leftarrow \mathcal{N}_t(\mathcal{X})$
    Inverse-transform $\hat{y}$ to get descaled predictions
    Compute RMSE on labeled set $\mathcal{L}$
    if convergence criteria met (slope ≈ 1, intercept ≈ 0) **then**
        break
    end if
    Compute uncertainty for each $x \in \mathcal{U}$ as $|\hat{y}_x - y_x|$
    Select top k most uncertain points $\mathcal{S} \subset \mathcal{U}$ based on uncertainty
    Update labeled set: $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{S}$
    Update unlabeled pool: $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{S}$
end for
Plot final predictions vs real values on labeled set

D. Universal approximation theorem

Demonstration that neural networks with only one internal layer and an arbitrary continuous sigmoidal nonlinearity can approximate any desired function [6]. Cybenko's Universal Approximation Theorem states that for any continuous sigmoidal function σ, finite sums of the form $G(x) = \sum_{j=1}^{N} c_j \sigma(w_j^T x + b_j)$ are dense in $C(I_n)$, the space of continuous functions on I_n, where $I_n = [0,1]^n$ is the n-dimensional unit hypercube. Given any $f \in C(I_n)$ and $\epsilon > 0$, there exists a sum $G(x)$ such that $|G(x) - f(x)| < \epsilon$ for all $x \in I_n$.

While Cybenko's theorem is formally stated for the unit hypercube $[0,1]^n$, it extends naturally to any compact domain $K \subset \mathbb{R}^n$ through linear transformations. For a function defined on an interval $[a,b]$, we can apply the transformation $x' = (x-a)/(b-a)$ to map it to $[0,1]$, approximate the transformed function, and then map back to the original domain. This equivalence ensures that the universal approximation property holds for any bounded interval.

The sigmoid function is defined as $\sigma(x; a, b) = \frac{1}{1+e^{-a(x-b)}}$ where a controls the steepness and b shifts the function horizontally. A function σ is sigmoidal if $\sigma(x) \to 1$ as $x \to +\infty$ and $\sigma(x) \to 0$ as $x \to -\infty$.

The target function is $f(x) = \frac{x}{1+|x|}$ for $x \in [0, 30]$. Although this domain differs from the canonical $[0,1]$ interval, the approximation principle remains valid due to the domain extension property mentioned above. The approximation is constructed as $G(x) = \sum_{j=1}^{N} c_j \sigma(w_j^T x + b_j)$ where $N = 40$ neurons are used. In this one-dimensional case, this simplifies to $G(x) = \sum_{j=1}^{N} h_j \sigma(a \cdot (x - b_j))$ with positions b_j uniformly distributed in $[-20, 20]$ to adequately cover the target domain $[0, 30]$ with sufficient margin, and steepness parameter $a = 100{,}000$ to ensure that sigmoids approximate step functions.

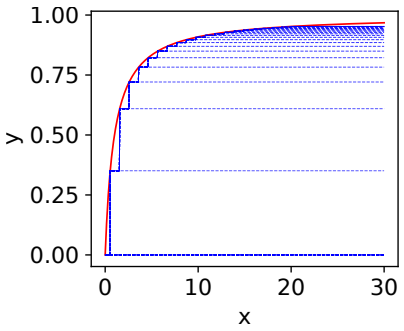


FIG. 5: Approximation of the target function $f(x) = \frac{x}{1+|x|}$ using a single hidden layer neural network with 40 sigmoidal neurons. The red line shows the target function, while the blue vertical lines represent the individual contributions of each neuron $\sigma(100,000(x - b_j))$ positioned uniformly across the interval [-20, 20]. The approximation $G(x) = \sum_{j=1}^{40} h_j \cdot \sigma(100,000(x - b_j))$ empirically demonstrates the Universal Approximation Theorem [6].

E. Dependence of the features

The dependence of the selected features in Table. I is represented in Fig. 6. It is observed that all features exhibit a certain trend concerning the bandgap. Therefore, the static interband polarizability in the x and y directions, Fig. 6 D) and E), shows a trend where materials with high polarizability typically have more mobile electrons and more dispersed bands, which tends to reduce the band gap, following the relationship $\alpha_i^{2D} \propto 1/E_{\text{band gap}}$ [7].

Additionally, the binding energy between layers in 2D materials is shown in Fig. 6 F). A low binding energy can indicate more weakly bound or molecular-like structures, where electronic levels are more confined, resulting in a higher band gap according to $E_B = E_{\text{band gap}}/4$ [7].

Finally, the heat of formation shown in Fig. 6 C) represents the formation energy of the compound from its constituent elements. A material with negative formation energy is thermodynamically stable. As this feature becomes more negative, it tends toward materials with higher band gaps than conductors, such as semiconductors or insulators, following $h_{\text{form}} \propto \ln(E_{\text{band gap}} + 1)$.

Regarding the total system energy, although the band

gap does not have a clear relationship with total energy, more stable materials (those with lower total energy) tend to form more compact structures with greater electronic delocalization, which can correlate with smaller bandgaps. In Fig. 6 A), this has been fitted with a 5^{th}-order polynomial regression.

On the other hand, Fig. 6 B) shows the number of disconnected atomic clusters in the 2D unit cell. This is characterized by the principle that more clusters lead to greater electronic isolation, less delocalization, and consequently a larger bandgap. However, for this feature, the dataset does not capture the actual number of connected clusters, but rather whether this material had them or not, using a binary value of 1 and 0. Therefore, it has not been possible to study the trend with respect to the band gap, despite knowing that cluster connectivity is associated with crystalline samples, which tend to be more semiconducting.

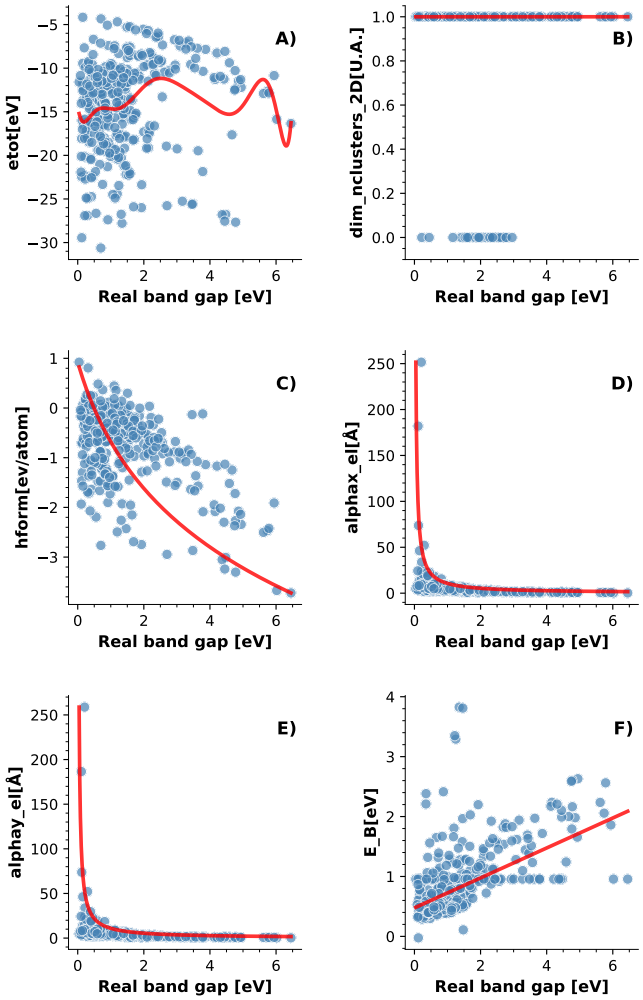


FIG. 6: Study of the dependence of the features selected in Table. I concerning the real band gap [eV]. In all plots, blue dots represent the experimental observations while the red line shows the fitted model for each feature.

This feature set coincides with the framework proposed by Zhang et al. [13], who identified thermodynamic and structural descriptors—such as formation enthalpy (Hform), total energy, volume, and cell area—as key predictors for band gap estimation. While retaining validated thermodynamic features (heat formation and total energy), the present selection incorporates electronic-specific descriptors (static interband polarisability) and 2D-optimized topological measures (Number of disconnected atomic clusters in the 2D unit cell), which indicate the unique structural characteristics of materials that directly influence electronic properties. These descriptors reflect how the atomic arrangement and connectivity within a single or a few atomic layers affect electron confinement and delocalization, key factors that determine the band gap. By quantifying features such as dimensionality and cluster connectivity, they provide insight into the quantum mechanical behavior of electrons confined to two dimensions, which conventional metrics cannot fully describe.