Machine Learning Reconstruction of High-dimensional Electronic Structure from Angle-resolved Photoemission Spectroscopy

Yu Zhang
Department of Physics
University of Florida
yuzhang1@ufl.edu

Yong Zhong
Stanford Institute for Materials and Energy Sciences
Stanford University and SLAC National Laboratory
ylzhong@stanford.edu

Nhat Huy Tran
Department of Physics
University of Florida
tran.n@ufl.edu

Shuyi Li
Department of Physics
University of Florida
lis3@ufl.edu

Kyuho Lee
Stanford Institute for Materials and Energy Sciences
Stanford University and SLAC National Laboratory
kyuho@stanford.edu

Harold Hwang
Stanford Institute for Materials and Energy Sciences
Stanford University and SLAC National Laboratory
hyhwang@stanford.edu

Zhi-Xun Shen
Stanford Institute for Materials and Energy Sciences
Stanford University and SLAC National Laboratory
zxshen@stanford.edu

Chunjing Jia
Department of Physics
University of Florida
cjia1@ufl.edu

Abstract

Extracting key electronic parameters from complex angle-resolved photoemission spectroscopy data is a significant challenge in condensed matter physics. This research introduces an advanced machine learning method, utilizing implicit neural representations, to automate and accelerate this process. Our model is trained to learn the direct relationship between a material's fundamental electronic parameters and its high-dimensional ARPES spectra. Applied to perovskite nickelates, this approach successfully obtained a more precise set of parameters from experimental data, outperforming traditional analytical techniques. This work demonstrates the power of implicit neural representations to bridge the gap between theory and experiment, paving the way for high-throughput discovery in materials science.

1 Introduction

Artificial Intelligence (AI) has driven transformative progress across the life sciences, most notably in protein structure prediction (e.g., AlphaFold) [1] and gene sequence analysis [2]. These breakthroughs underscore AI's profound capability to decode complex, fundamental relationships from large-scale data. In a direct analogy, a material's electronic structure can be viewed as its "gene," as it fundamentally governs the material's macroscopic physical and chemical properties. This perspective motivates a natural hypothesis: AI has the potential to revolutionize materials science in the same way it has reshaped biology. However, the primary bottleneck currently impeding AI's progress in this domain is the extreme scarcity of high-quality, standardized electronic structure datasets, preventing

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

AI models from being trained effectively and limiting their capability to generalize across material classes.

Based on the photoelectric effect first proposed by Albert Einstein, angle-resolved photoemission spectroscopy (ARPES) is the most direct experimental technique for probing the electronic structure of materials [3]. By measuring both energy (E) and momentum ($\mathbf{k}$) information, ARPES generates a four-dimensional dataset (E, k_x, k_y, k_z) that fully describes materials' electronic structure. Despite its unparalleled capability, ARPES suffers from inherent limitations: it requires access to large facilities such as synchrotron light sources, data acquisition is low throughput (often tens of hours per measurement), and downstream data analysis is both labor-intensive and reliant on expert knowledge. These challenges highlight an urgent need for dedicated AI tools to accelerate and standardize the analysis of ARPES data.

Here, we use a real-world example to demonstrate the power of machine learning (ML) methods for ARPES data interpretation and reconstruction. The perovskite nickelate $Nd_{1-x}Sr_xNiO_3$ has emerged as a promising candidate for next-generation electronic devices, owing to its sharp and tunable metal-insulator transition. Its low-energy electronic structure is primarily governed by the two e_g orbitals of the Ni^{3+} ion ($d_{x^2-y^2}$ and d_{z^2}). Recently, systematic ARPES measurements have succeeded in mapping the high-quality, four-dimensional electronic structure of $Nd_{1-x}Sr_xNiO_3$. This rich experimental dataset makes perovskite nickelate an ideal model system for developing and applying machine learning methods to automatically extract the essential electronic parameters that describe its unique properties.

2 Methods

2.1 Two-Band Tight-Binding Model

To capture the essential low-energy physics of perovskite nickelate, we employ a tight-binding model to simulate its electronic structure. This model is designed to reproduce two critical features in ARPES sudy: the band dispersion and the Fermi surface topology. The band dispersion, $E(\mathbf{k})$, describes the fundamental relationship between an electron's energy and its momentum. The Fermi surface—the boundary in momentum space separating occupied from unoccupied states—dictates many of the material's key properties. Based on the electronic configuration of Ni^{3+} ion, a two-band tight-binding model can provide an effective description of the band dispersion and Fermi surface topology for $Nd_{1-x}Sr_xNiO_3$ system:

$$H_{tb} = -\sum_{ij} t_{ij}^{ab} c_{ia}^{\dagger} c_{jb} - \mu \tag{1}$$

Here, i, j are site indices, $a, b = 1, 2$ are orbital indices corresponding to $d_{3z^2-r^2}$ and $d_{x^2-y^2}$ orbitals, and μ is the chemical potential which can be tuned by changing the doping levels. Only the nearest-neighbor hopping t_1 and next-nearest-neighbor hopping t_2 with σ-type bonding are considered in our simulations (More details in Appendices). The goal here is to extract the tight-binding parameters (t_1, t_2, μ) that best fit the experimental ARPES spectra.

The ARPES spectral intensity is calculated through equations (2) and (3), where $\epsilon(\mathbf{k})$ represents the bare band, $\Sigma'(E)$ and $\Sigma''(E)$ are the real and imaginary parts of self-energy, respectively, and $f(E)$ is the Fermi-Dirac function.

$$A(E, \mathbf{k}) = \sum -\frac{1}{\pi} \frac{\Sigma^{''}(E)}{(E - \epsilon(\mathbf{k}) - \Sigma'(E))^2 + (\Sigma''(E))^2} \tag{2}$$

We consider k_z broadening using a Lorentzian convolution.

$$I_{TB}(E, k_x, k_y) = \int_{k_z^0 - \frac{\Delta k_z}{2}}^{k_z^0 + \frac{\Delta k_z}{2}} f(E) \frac{A(E, \mathbf{k})}{(k_z - k_z^0)^2 + (\Delta k_z/2)^2} \tag{3}$$

2.2 Implicit Neural Representations

We developed a machine learning framework to efficiently explore the tight-binding (TB) parameter space using implicit neural representations. Traditional approaches for extracting material parameters

from ARPES data face significant challenges due to the complex dependence on model parameters, the vast momentum-energy configurational space, and the presence of experimental noise. Our method offers an efficient and more accurate alternative for parameter extraction. The machine learning architecture is inspired by the work of Chitturi et al. [4], who employed implicit neural representations to simulate a Heisenberg model and successfully extracted exchange coupling parameters from neutron spectroscopy data. Similarly, we aim to leverage the power of machine learning to deepen our understanding of ARPES spectroscopy.

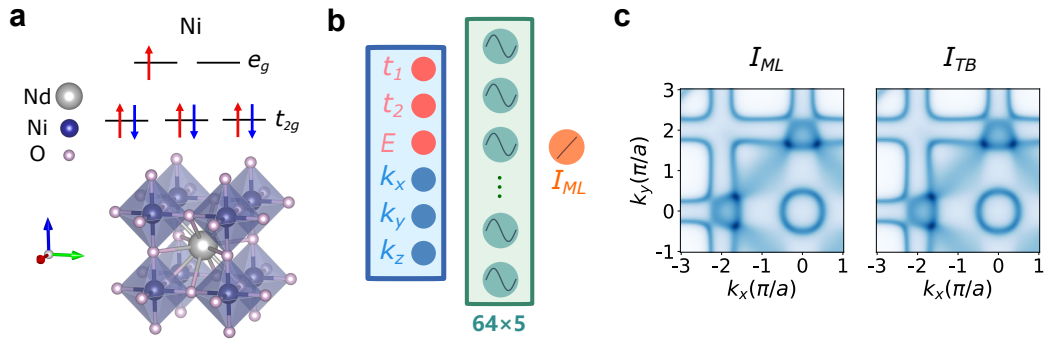

Figure 1: **a**. Crystal structure of $NdNiO_3$ with cubic symmetry, along with the corresponding energy levels of Ni. **b**. Architecture of the SIREN model: five hidden layers are placed between the inputs (three TB parameters and $\mathbf{k}$) and the output (spectral intensity I_{ML}). **c**. Fermi surface obtained from the trained machine learning model (left) and from the tight-binding model (right).

To simulate the tight-binding model for different parameters, we introduce a fully connected neural network with sinusoidal activation functions, specifically the SIREN model [5], which can capture fine details for natural signals. Considering the intrinsic three-dimensional electronic structure of $Nd_{1-x}Sr_xNiO_3$, we chose $(k_x, k_y, k_z, t_1, t_2, \mu)$ as inputs, and the ML model is trained to approximate the ARPES intensity $I_{ARPES}(E, \mathbf{k})$. While discrete data points are used for training, the ML model is capable of interpolation to provide a continuous representation.

Based on the well-trained and differentiable neural network, we optimize the unknown tight-binding parameters using a gradient-based optimization algorithm. More specifically, we use the Pearson correlation coefficient r:

$$r = \frac{Cov(y, y_{pred})}{\sigma_y \sigma_{y_{pred}}} \tag{4}$$

as a metric to assess the similarity between I_{ARPPES} and I_{TB}. We assume that these two values are linearly correlated. Since r is invariant under linear transformation, we can avoid the normalization problem in this case. We define the loss function as $L = 1 - r$, using the SIREN model as the surrogate for the TB model, we minimize L via Adam optimizer, following the workflow as shown in Figure.2a.

Figure.2b shows that the tight-binding parameters obtained from the gradient optimization method lie close to the global minimum, providing a reliable approximation. The small discrepancies can be attributed to two factors: (1) the discrete grid search, which may not capture the exact minimum due to limited resolution, and (2) the trained ML model not perfectly matching the TB model, leading to minor variations.

3 Results

To evaluate the performance of our method, we present a comprehensive comparison of the experimental data, the ML simulations and the traditional analytical fittings in Figure.3. The first five columns show the Fermi surfaces, while the last two present the $E(\mathbf{k})$ dispersion data. We find that a clear improvement is achieved in the Fermi surface topology generated using the ML-optimized parameters, which exhibits a better agreement with the experimental data. This is particularly evident near the (π, π) region, where $k_x = 1, k_y = 1$, at photon energies of 150eV and 155eV. Both the Pearson correlation coefficient and the Mean Squared Error (MSE) indicate a lower error for the

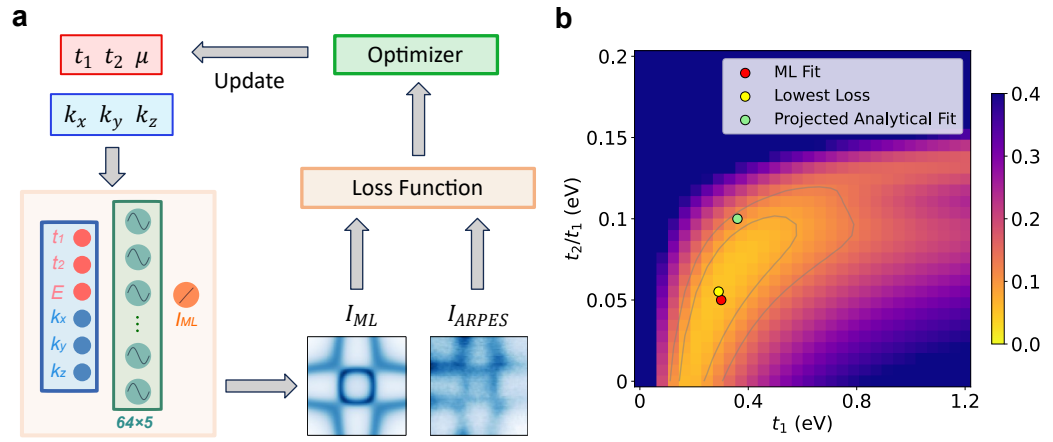

Figure 2: **a**. Our machine learning workflow to obtain the materials parameters. An initial guess of the tight-binding parameters is fed into the well-trained SIREN model. The parameters are iteratively updated using the gradient $\frac{\partial L}{\partial t_i}$ until convergence. **b**. Visualization of the loss landscape at $\mu = 1.9t_1$. A grid of 30 sampled values was employed for both t_1 and t_2. The color map represents the loss calculated by the trained ML model for the corresponding TB parameters. Red, yellow, and green points indicate the results from the ML fitting method, grid search (30×30), and analytical fitting [6], respectively.

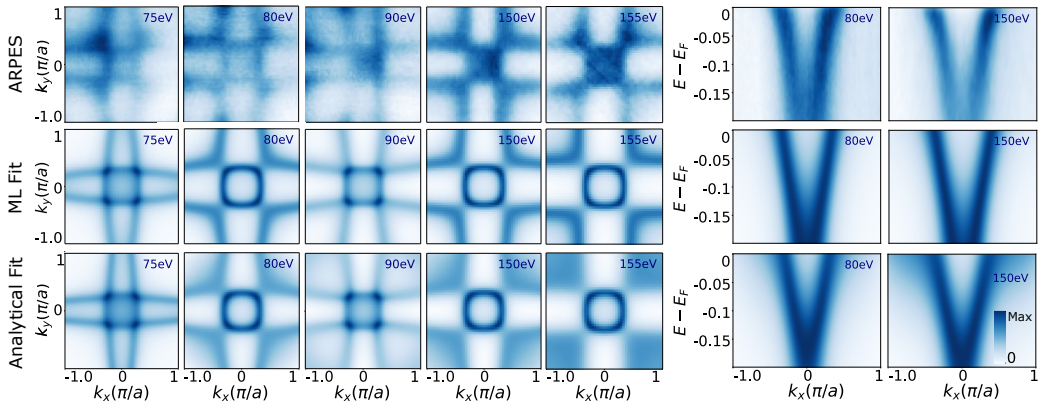

Figure 3: Fermi Surface and E_k dispersion for perovskite nickelates. Comparison between ARPES data (1^{st} Row) and ML predictions (2^{nd} Row) uses ML-fitted parameters with $t_1 = 0.31eV$, $t_2/t_1 = 0.05$, while 3^{rd} Row uses the TB parameters given by analytical fit with $t_1 = 0.36$, $t_2/t_1 = 0.10$. The blue values on each graph represent the photon energy used during measurement.

ML-fitted parameters, with values of 0.046 and 0.009, respectively, compared to the analytical fit, which shows values of 0.090 and 0.235. We want to emphasize that our simulation does not include matrix element effects, which can account for the intensity differences observed symmetric $\mathbf{k}$-points in the experimental data. Regarding the $E(\mathbf{k})$ dispersion, the ML method gives better description of the "waterfall" feature extending to deep energies.

4 Discussion and Outlook

Our work successfully bridges the gap between the complex experimental data and the theoretical models by using a SIREN-based neural network to simulate the electronic structure of perovskite nickelates. The quantitative agreement demonstrates that ML-driven approach can automate the reconstruction of high-dimensional ARPES data with much improved resolution. The primary advantages of this method are its efficiency, accuracy, and transferability. With an affordable initial training cost, the model can be directly applied to similar material systems, such as manganites, without retraining. This technique offers a robust and standardized alternative to labor-intensive fitting procedures, opening a new avenue for the high-throughput analysis of electronic structures in materials science.

Looking ahead, our framework points toward a broader integration of domain-specific machine learning with emerging advances in large language models. One direction is to combine models trained on spectroscopic data with more general reasoning tools, enabling AI systems to autonomously navigate model selection — for example, determining the relevant degrees of freedom, lattice symmetries, or matrix element terms needed for accurate electronic structure reconstruction. Such an integration could evolve into agent-like platforms that not only fit data, but also reason about the appropriate physical model and adapt to new material classes with minimal human input. Ultimately, this vision suggests a path toward autonomous discovery pipelines, where spectroscopic data across diverse quantum materials are interpreted in real time, providing unprecedented scalability and consistency in connecting experiments with theory.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, and Kathryn Tunyasuvunakool et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583, 2021.

[2] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37:2112, 2021.

[3] Jonathan A. Sobota, Yu He, and Zhi-Xun Shen. Angle-resolved photoemission studies of quantum materials. *Reviews of Modern Physics*, 93:025006, 2021.

[4] Sathya R. Chitturi, Zhurun Ji, Alexander N. Petsch, Cheng Peng, Zhantao Chen, Rajan Plumley, Mike Dunne, Sougata Mardanya, Sugata Chowdhury, Hongwei Chen, Arun Bansil, Adrian Feiguin, Alexander I. Kolesnikov, Dharmalingam Prabhakaran, Stephen M. Hayden, Daniel Ratner, Chunjing Jia, Youssef Nashed, and Joshua J. Turner. Capturing dynamical correlations using implicit neural representations. *Nature Communications*, 14(1):5852, September 20 2023.

[5] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions, 2020.

[6] Yong Zhong, Kyuho Lee, Regan Bhatta, Yu Zhang, Yonghun Lee, Martin Gonzalez, Jiarui Li, Ruohan Wang, Makoto Hashimoto, Donghui Lu, Sung-Kwan Mo, Chunjing Jia, Harold Hwang, and Zhi-Xun Shen. Surface preparation method for investigating the three-dimensional electronic structure of perovskite nickelates. *Phys. Rev. B*, 112:035160, 2025.

A Appendices

A.1 Simulate ARPES Intensity through Tight-Binding Model

The tight-binding Hamiltonian is given by:

$$H_{tb} = -\sum_{ij} t_{ij}^{ab} c_{ia}^{\dagger} c_{jb} - \mu$$

The hopping parameters are defined as:

$$t_{i,i\pm\hat{\mu}}^{ab} = t_1 \, \phi_{\mu}^{a} \phi_{\mu}^{b}$$
$$t_{i,i\pm\hat{\mu}\pm\hat{\nu}}^{ab} = t_2 \left(\phi_{\mu}^{a} \phi_{\nu}^{b} + \phi_{\mu}^{b} \phi_{\nu}^{a} \right)$$

where:

$$\phi_x = \left(-\frac{1}{2}, \frac{\sqrt{3}}{2} \right), \quad \phi_y = \left(-\frac{1}{2}, -\frac{\sqrt{3}}{2} \right), \quad \phi_z = (1, 0)$$

These correspond to the wave functions for $d_{3x^2-r^2}$, $d_{3y^2-r^2}$, and $d_{3z^2-r^2}$ σ-bonding orbitals along the three axes. We diagonalize H_{tb} to obtain the bare band dispersion for the system.

A.2 ARPES Data Utilization and Dataset Generation

For this project, we utilize the high-resolution ARPES data collected on $Nd_{1-x}Sr_xNiO_3$. The experimental data include both the Fermi surface and the E(k) dispersion. Specifically, the Fermi surface were measured with incident photon energies of 70, 75, 80, 90, 150, 155eV for the 1^{st} Brillouin zone (BZ). Band dispersion along the k-path from Y to M ($k_x = 1$) was recorded at 80 and 150eV. k_x, k_y are renormalized and have units of π/a, a is the lattice constant. k_y is computed using the following equation:

$$k_z = \sqrt{2m(h\nu - \phi - E_B + V) - k_x^2 - k_y^2}/\hbar \tag{5}$$

where $h\nu$ is the incident photon energy, ϕ is the work function, E_B is the binding energy and V is the inner potential.

Table 1: Parameters Selection for Dataset Generation

Parameter	$k_x(\pi/a)$	$k_y(\pi/a)$	$h\nu(eV)$	$t_1(eV)$	t_2/t_1	μ/t_1
Range	[-3,1]	[-1,3]	[70,160]	[0,1.2]	[0,0.2]	[0.5,2.5]
Quantity	201	201	12	50	12	16

The dataset is generated using the parameters listed in Table.1 and is divided into training, validation, and test sets with a ratio of 8:1:1.

A.3 SIREN Model Training and Tight-binding Parameters Extraction

The SIREN model was trained to predict spectral intensity by minimizing the Mean Squared Error (MSE) between I_{ML} and I_{TB}. We use the Adam optimizer, batch size = 65536, initial learning rate = 0.001, which is decayed exponentially. The model was trained for 55 epochs on an NVIDIA B200 GPU. To accelerate training, only one-fifth of the training dataset was used in each epoch, resulting in an average epoch duration of approximately 150s.

The tight-binding parameters were optimized using the Adam algorithm with an initial learning rate = 0.01 and an exponential decay for 2000 iterations. To integrate all experimental datasets, we employed a weighted loss function, where the weights were determined by the number of data points N_i in each ARPES dataset.

$$L = \frac{1}{N_{total}} \sum_i L_i N_i \tag{6}$$

In our simulation, the matrix element was not included, resulting in a symmetrized Fermi surface in 1^{st} Brilliouin Zone (BZ). For the optimization, we only used 1/4 of the ARPES Fermi surface data in 1^{st} BZ and part of ($E - E_F > -0.10eV$) E(k) dispersion data.