Inverse Design of Mid-IR Quantum Cascade Lasers Active Region via Machine Learning

by
Yunhan Hu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)

at the
University of Wisconsin-Madison
2025

Date of Final Oral Exam: 12/19/2024
The dissertation is approved by the following members of the Final Oral Committee:

Zongfu Yu, Professor, Electrical and Computer Engineering
Luke Mawst, Professor, Electrical and Computer Engineering
Dan Botez, Professor, Electrical and Computer Engineering
Yudong Chen, Associate Professor, Computer Sciences

Inverse Design of Mid-IR Quantum Cascade Lasers Active Region via Machine Learning

Yunhan Hu

Abstract

Quantum Cascade Lasers (QCLs), as pivotal coherent sources in mid-infrared and terahertz spectral regions, have demonstrated significant potential across diverse applications including spectroscopy, free-space optical communications, chemical sensing, and medical diagnostics. However, the performance of QCLs critically depends on the complex active region (AR) structure. Conventional design methodologies heavily rely on expert knowledge and iterative simulations, rendering the efficient exploration of vast design spaces challenging. Machine learning (ML), particularly neural networks, offers a promising alternative to address these challenges by facilitating automated and efficient design processes.

This dissertation presents an ML-based inverse design framework for QCL active regions. The primary objective is to construct a neural network capable of predicting physical structural parameters of QCLs, such as layer thicknesses and compositions, directly from desired performance metrics including operating wavelength, output power, and threshold current. Initially, an automated wavefunction identification algorithm was developed to accurately determine key lasing energy levels from computed band diagrams, significantly enhancing the efficiency and precision of large-scale data generation. Leveraging this method, a comprehensive dataset containing approximately 430,000 QCL structures and associated k·p metrics (energy level differences and carrier transition probabilities/lifetimes) was created.

Subsequently, a forward neural network was trained and validated to accurately predict the k·p metrics from given QCL-AR structures. To tackle the inherent non-uniqueness problem in mapping performance metrics to physical structures, a tandem neural network architecture was introduced. The inverse component of this architecture generates diverse QCL-AR structures from given k·p metrics by incorporating a randomization layer. Experimental results demonstrated that this tandem inverse network achieved excellent performance, with coefficients of determination (R^2) of 0.9153 and 0.9701 for key energy differences (E_{43} and E_{54}), and 0.9568 and 0.9175 for reciprocal lifetimes ($1/\tau_{43}$ and $1/\tau_{54}$), respectively. An example application illustrated the network's capability in generating QCL-AR structures with the potential for reduced threshold-current densities by mitigating shunt-type carrier leakage.

This work systematically details the development and validation of the automated data generation process, forward neural network modeling, and inverse design framework, demonstrating the substantial potential of machine learning to accelerate and optimize the QCL design process. It lays a robust foundation for further advancements toward comprehensive automated inverse design of high-performance quantum cascade lasers.

Keywords: Quantum Cascade Lasers, Machine Learning, Inverse Design, Neural Net-

works, Wavefunction Identification, k·p Metrics, Tandem Network

Dedication

To my beloved parents, your unwavering financial support throughout my master's and doctoral studies as an international student made this journey possible. More than that, thank you for consistently supporting my decisions every step of the way. Your belief in me has been my greatest strength.

To my dear friends, who became my family away from home, especially during the nearly five years I couldn't return due to the pandemic. Your constant companionship and immense social support were crucial in maintaining my well-being and sanity. I am profoundly grateful for each of you.

Declaration

I hereby declare that this dissertation is the result of my own original research. All the work presented in this thesis is my own, except where otherwise acknowledged. This thesis has not been submitted, in whole or in part, for any other degree or qualification at this or any other university.

Statement on Published Works

The core of this dissertation is based on research that has been published. Specifically, the work detailed in this thesis is principally derived from the following publications, of which I am the lead author:

1. **Y. Hu**, S. Suri, J. Kirch, B. Knipfer, S. Jacobs, Z. Yu, D. Botez, L. J. Mawst. *Enhancing quantum cascade laser active region design through inverse neural networks: A machine learning approach to metric-based structure generation.* AIP Advances 1 October 2024; 14 (10): 105033. https://doi.org/10.1063/5.0227270

2. **Y. Hu**, S. Suri, J. Kirch, B. Knipfer, S. Jacobs, S. K. Nair, Z. Zhou, Z. Yu, D. Botez, L. J. Mawst. *Large-scale data generation for quantum cascade laser active-region design with automated wavefunction identification.* Appl. Phys. Lett. 10 June 2024; 124 (24): 241103. https://doi.org/10.1063/5.0209613

In these publications, my primary role was in the conceptualization of the research, leading the investigation, performing the formal analysis, developing the necessary software, and writing the original draft of the manuscripts. The other authors were collaborators who made significant contributions as detailed in the Author Contributions sections of the respective papers, and I gratefully acknowledge their valuable input and support throughout this research.

Declaration on the Use of Generative Artificial Intelligence

In the preparation of this dissertation, generative artificial intelligence (AI) tools were utilized for the sole purpose of language editing, including improving grammar, rephrasing for clarity, and refining the overall readability of the text.

The use of AI was strictly limited to text enhancement. All core scientific concepts, methodologies, data analysis, results, and conclusions presented in this thesis are my own intellectual work. The machine learning models developed and applied as part of the research methodology are a central component of the study itself and are distinct from the use of generative AI in the writing process.

I affirm that the final content of this dissertation is my own, complies with all originality requirements, and adheres to the standards of academic integrity and ethical conduct of **University of Wisconsin-Madison**.

Yunhan Hu
06/10/2025

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my advisor, **Prof. Zongfu Yu**. Throughout my doctoral studies, Prof. Yu has provided me with profound academic inspiration and meticulous guidance. His rigorous approach to research and forward-thinking vision, especially his contributions during the conceptualization phase of the project, laid a solid foundation for the success of this work.

I am equally grateful to my co-advisor and the Principal Investigator of this project, **Prof. Luke Mawst**. Prof. Mawst played a crucial leadership role in the advancement of this entire project, from conceptualization and funding acquisition to daily supervision and guidance. His extensive knowledge and invaluable advice were decisive for the direction and depth of this research.

My heartfelt thanks also go to my other co-advisor, **Prof. Dan Botez**. Prof. Botez provided tremendous help in terms of research resources and academic supervision, and his support was an indispensable part of this study.

The completion of this research would not have been possible without the close collaboration and joint efforts of the entire research team. I would like to extend my special thanks to my collaborators:

- **Suraj Suri** made significant contributions to data curation, software development, visualization, and the co-writing of the original drafts.
- **Jeremy Kirch** provided critical assistance with data curation, software support, and investigation.
- **Benjamin Knipfer** offered support for data curation and software-related tasks.
- **Steve Jacobs** from Intraband, LLC, provided valuable assistance in supervision and validation.
- I also want to thank **Sreenivas Krishna Nair** and **Ziyu Zhou** for their support in the investigation phase.

Finally, I would like to acknowledge the financial support for this research. This work was primarily supported by the U.S. Navy under contract number N6893622C0020, administered by Richard LaMarca. Additional support was provided by Intraband, LLC.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Fundamental Principles of Semiconductor Diode Lasers

Semiconductor lasers, also known as laser diodes, are compact and efficient sources of coherent light that have become indispensable in numerous fields, from telecommunications to medical diagnostics. Their operation relies on the same fundamental principles governing all lasers: achieving population inversion, providing optical feedback via a resonant cavity, and meeting a lasing threshold condition where the optical gain compensates for all the losses within the cavity[22]. The core of a semiconductor laser is a p-n junction formed from a direct bandgap semiconductor material, where the lasing action originates from the radiative recombination of electrons and holes.

The process begins with electrical pumping. When a forward bias is applied across the p-n junction, a large density of electrons from the n-side and holes from the p-side are injected into a thin, undoped region known as the active region. This injection of charge carriers creates a non-equilibrium state where the concentration of electrons in the conduction band (higher energy level) exceeds the concentration of electrons in the valence band (lower energy level). This condition, termed population inversion, is essential for stimulated emission to dominate over absorption[26]. The energy separation between these bands is the bandgap (E_g), which primarily determines the energy, and thus the

wavelength (λ), of the emitted photons, according to the relation:

$$\lambda \approx hc/E_g$$

where h is Planck's constant and c is the speed of light.

To achieve the high carrier densities required for population inversion at manageable currents, modern semiconductor lasers almost universally employ a heterostructure design. A double heterostructure, for instance, sandwiches the lower bandgap active region material between two higher bandgap cladding layers. This structure serves two critical purposes: it efficiently confines the injected carriers to the thin active region due to the potential barriers, and it simultaneously creates a waveguide for the generated photons, as the higher bandgap cladding layers have a lower refractive index than the active region[22]. This dual confinement of both carriers and photons significantly increases the probability of stimulated emission and dramatically reduces the threshold current density required for lasing compared to simple homojunction devices.

The second key requirement, optical feedback, is typically provided by forming a resonant cavity. In the simplest Fabry-Pérot type laser, this is achieved by cleaving the semiconductor crystal along its natural crystal planes to create two parallel, mirror-like facets. The reflectivity at the semiconductor-air interface (typically 30%) is often sufficient to reflect a portion of the generated photons back into the active region, where they can stimulate the emission of further identical photons, leading to optical amplification. For lasing to occur, the optical gain within the active region must be large enough to overcome the total losses in the cavity, which include internal losses (e.g., free-carrier absorption and scattering) and mirror losses from the light escaping the facets[4].

The emission wavelength of a conventional semiconductor laser is fundamentally tied to the bandgap of the material used in the active region. Decades of materials science research have led to the development of mature material systems for specific spectral ranges. For example, the AlGaAs/GaAs system is widely used for lasers in the near-infrared (NIR) range from approximately 750–880 nm, while the InGaAsP/InP system is the workhorse

for fiber-optic communications in the 1.3 μm and 1.55 μm windows[26]. However, this direct reliance on the intrinsic bandgap of available direct-bandgap materials poses a significant limitation. Extending laser operation into spectral regions like the mid-infrared (MIR, 3–5 μm) and beyond has been historically challenging due to a "materials gap"—a lack of suitable direct bandgap semiconductor alloys for fabricating conventional diode lasers[30]. This challenge has spurred the development of alternative laser architectures, such as quantum cascade lasers (QCLs), which circumvent the limitations of band-to-band transitions.

1.2 The Quest for Efficient Mid-Infrared Coherent Sources: Quantum Cascade Lasers

The mid-infrared (mid-IR) spectral region, typically defined as wavelengths from 3 to 20 μm, is of paramount importance for a vast array of scientific and technological applications. This region is often called the molecular "fingerprint" region because the fundamental rotational and vibrational transitions of many molecules occur here. This characteristic makes mid-IR spectroscopy a powerful tool for applications including, but not limited to, the standoff detection of chemicals and explosives, free-space optical communications, medical diagnostics through breath analysis and tissue imaging, and real-time industrial process monitoring[19]. Among the available mid-IR sources, quantum cascade lasers have emerged as a uniquely versatile class of semiconductor emitters, providing powerful and coherent radiation across the mid-IR and into the terahertz (THz) spectral regions[1].

A defining feature of the QCL is its unipolar nature. Unlike conventional diode lasers, which rely on the radiative recombination of electron-hole pairs across a material's bandgap (interband transitions), QCLs generate light through electronic transitions between quantized conduction band subbands within a meticulously engineered semiconductor heterostructure[9]. This stack of alternating thin layers of different semiconductor materials forms a series of quantum wells, creating an artificial "superlattice." The concept was first ingeniously proposed by Kazarinov and Suris in 1971[16], but the first experimental

demonstration was a landmark achievement by Faist, Capasso, and their colleagues at Bell Laboratories in 1994[6]. This event marked a paradigm shift in laser physics.

The revolutionary aspect of the QCL lies in its "bandstructure engineering." The emission wavelength is not determined by the intrinsic bandgap of the constituent materials but rather by the physical thickness of the quantum well and barrier layers. By precisely controlling these thicknesses during epitaxial growth, one can effectively "tailor" the energy spacing between the subbands, thus customizing the laser's emission wavelength with extraordinary flexibility[1]. This design freedom has enabled QCLs to achieve performance metrics that are inaccessible to traditional diode lasers in the mid-IR.

Since their first demonstration, QCL technology has progressed at a remarkable pace. Significant advancements have led to high optical output power (watt-level), room-temperature continuous-wave (RT-CW) operation, and broad spectral coverage, solidifying their role as the dominant coherent source in the mid-infrared[21]. This high performance has, in turn, catalyzed breakthroughs in the applications that first motivated their development. For instance, the high power and spectral purity of QCLs are critical for overcoming atmospheric absorption and scattering in remote sensing and free-space communication[19], while their tunability and room-temperature operation are ideal for compact, field-deployable spectroscopic instruments for medical and industrial use[14]. Consequently, the QCL is not merely a novel laser source but a foundational technology driving innovation across multiple scientific and engineering disciplines.

1.3 The Challenges of Traditional QCL Design and the Dawn of Machine Learning

The core of a Quantum Cascade Laser is its active region, an intricately engineered structure composed of tens or even hundreds of precisely stacked thin semiconductor layers. The thickness and composition of each individual layer critically influence the final laser performance, establishing a complex and often non-intuitive relationship between the device's physical structure and its operational characteristics[1]

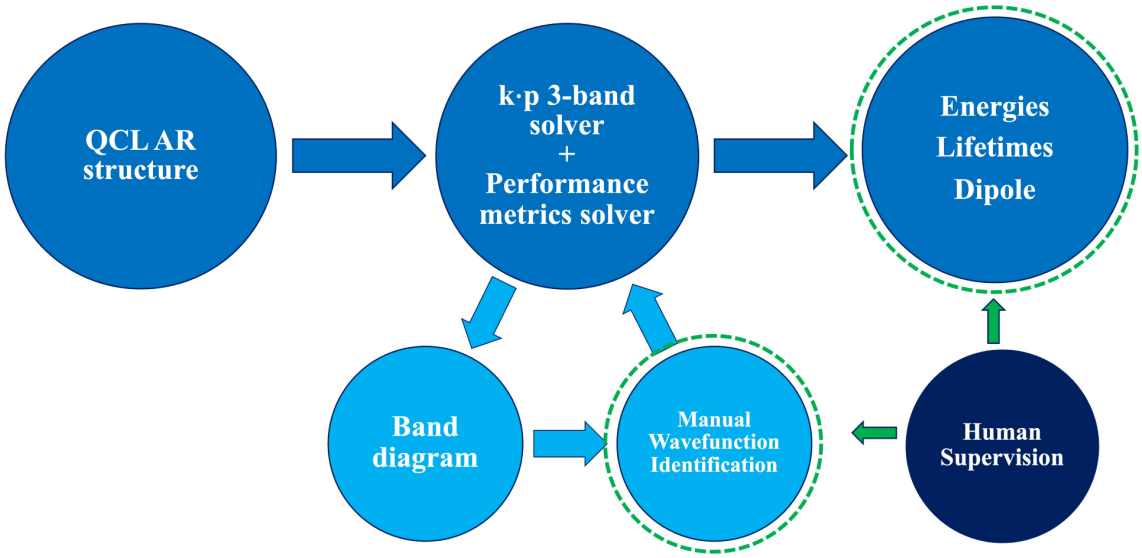


Figure 1.1: Active-region design flow. This loop continues till an acceptable design is achieved.[13]

Traditionally, the design of these complex QCL structures has been a formidable task, relying heavily on the designer's deep physical intuition and extensive experience. The conventional workflow involves a "trial-and-error" iterative process, where designers manually propose a structure and then use physics-based simulation tools—such as solving the Schrödinger equation with the k.p method—to predict its performance[5, 12, 13]. This cycle is repeated until a satisfactory design is achieved. This process is not only labor-intensive and time-consuming, with the development of a single optimized design often taking several weeks or even months, but it is also inherently inefficient[12, 13]. The primary bottleneck in QCL development is the immense, high-dimensional design space defined by the numerous structural parameters. Effectively exploring this vast space to identify optimal or near-optimal configurations that meet specific performance targets remains a significant challenge[10, 12, 13]. A typical conventional design progress is shown in Figure 3.1

The high dimensionality and complex physics inherent in QCL design have naturally driven researchers to explore data-driven, automated approaches like machine learning

(ML). The sheer number of tunable parameters and their nonlinear impact on laser performance make it exceedingly difficult to navigate the design space comprehensively using physical intuition or traditional optimization algorithms alone. Machine learning, particularly deep neural networks, has emerged as a powerful tool to reform and potentially automate the QCL design process[10, 12, 13]. By learning the complex mapping between structural parameters and performance metrics from simulation data, ML models can act as surrogate models that predict a QCL's performance almost instantaneously[13], replacing time-consuming physical simulations.

Beyond supervised learning, other automated design methodologies have also been successfully applied. Optimization algorithms such as Bayesian optimization[8], simulated annealing[28], and genetic algorithms[27] are frequently integrated with ML models or first-principle simulations to efficiently search the design space for novel, high-performance QCL structures. These techniques provide a systematic and automated framework to overcome the limitations of manual design, heralding a new era of rapid and intelligent QCL development.

1.4 Research Objectives: Towards Automated Inverse Design of QCLs

The conventional design process for Quantum Cascade Laser active regions is a significant bottleneck, characterized by its reliance on expert knowledge and extensive, iterative manual adjustments. This traditional "trial and error" approach is often time-consuming and labor-intensive, taking weeks or even months to finalize a single design. The overarching research goal of this thesis is to overcome these limitations by developing a machine learning-based framework for the inverse design of QCLs. In an ideal inverse design scenario, a user would input the desired high-level device performance metrics—such as lasing wavelength, output power, or threshold current—and the model would, in turn, generate the precise physical parameters of the QCL structure (e.g., layer thicknesses and compositions) required to achieve that performance.

However, directly mapping macroscopic performance metrics to microscopic structural parameters is an exceptionally challenging task due to the complex, non-linear interplay of numerous intermediate physical processes. To address this complexity, this thesis adopts a pragmatic, staged strategy. The research focuses on using a set of intermediate physical descriptors, known as **k·p metrics**, as the crucial bridge between the final device performance and the underlying physical structure.

k·p metrics are parameters directly obtained from k·p perturbation theory calculations, which describe the electronic band structure of the QCL's layered heterostructure. These metrics encapsulate critical information about the device's quantum-level behavior, including key energy-level differences (e.g., the transition energy E_{43} and the leakage-related energy separation E_{54}) and electron scattering lifetimes (e.g., τ_{43} and τ_{54}). While not the final device-level metrics themselves, they are more directly related to the layer structure and serve as the essential inputs for calculating performance indicators like threshold-current density, lasing efficiency, and wavelength.

To achieve the ultimate goal of an automated inverse design framework, this thesis sets the following specific research objectives:

1. **Develop an Automated Wavefunction Identification Program.**

   The first critical step is to create a robust, automated program to rapidly and accurately identify key wavefunctions (e.g., upper and lower laser levels, injector states, and leakage paths) from raw QCL band diagrams. This automation is essential to move beyond the slow, manual labeling process and enable the generation of data on a massive scale.

2. **Generate a Large-Scale, Comprehensive Dataset.**

   Using the automated program, the next objective is to generate a vast and diverse dataset containing hundreds of thousands of unique QCL active region structures and their corresponding, accurately calculated k·p metrics. This comprehensive dataset serves as the fundamental training ground for all subsequent machine learning mod-

els.

3. **Train and Validate a Forward Neural Network.**

   With the dataset established, a forward neural network will be trained and validated. This model functions by taking the physical QCL structure (i.e., the sequence of layer thicknesses) as its input and predicting the resulting k·p metrics with high accuracy. The success of this forward model not only confirms the quality and utility of the generated dataset but also provides an essential tool for the final objective, as it can serve as a fast and differentiable evaluator within a more complex model architecture.

4. **Develop, Train, and Test an Inverse (Generative) Neural Network.**

   The capstone objective is to create a generative inverse network that performs the core design task: it takes a set of desired k·p metrics as input and outputs a viable QCL active region structure. A primary challenge in this step is to solve the inherent non-uniqueness problem, where multiple distinct QCL structures can produce nearly identical k·p metrics. The network will be specifically designed to address this by incorporating features, such as a random layer, that allow it to generate a diverse range of valid structural solutions from a single performance-metric input.

By focusing on k·p metrics as an intermediate target, this research strategically deconstructs a highly complex problem into a series of more manageable objectives. This approach lays the foundational groundwork necessary for the future realization of a complete, end-to-end inverse design framework for high-performance QCLs.

1.5 Main Contributions of This Thesis

This thesis presents several key innovative contributions to the field of automated QCL design, forming a cohesive, end-to-end framework for accelerating and enhancing the design process. The primary contributions are[12, 13]:

- **A Robust Automated Wavefunction Identification Program:** We developed

a novel, automated program that efficiently and accurately identifies critical lasing states from complex QCL band diagrams, achieving over 95% accuracy. This program replaces the conventional, time-consuming manual identification process and serves as the cornerstone for large-scale data generation.

- **A Large-Scale, High-Quality QCL Design Dataset:** Leveraging the automated identification program, we generated a comprehensive dataset of approximately 430,000 QCL active region structures and their corresponding k·p metrics. For training the inverse network, this dataset was further refined and filtered to roughly 300,000 entries to remove structures affected by energy-level hybridization, ensuring higher data quality. This dataset provides a solid foundation for training robust machine learning models.

- **High-Accuracy Forward Prediction Models for k·p Metrics:** We successfully trained forward neural networks capable of accurately predicting key k·p metrics—such as the energy differences E_{43} and E_{54}, and lifetimes τ_{43} and τ_{54}—directly from the QCL's physical layer structure. These models demonstrate high fidelity, with coefficients of determination (R^2) exceeding 0.91 for all four key metrics in the final implementation.

- **A Novel Tandem Inverse Network Architecture to Address Non-Uniqueness:** We developed an innovative tandem (or cascaded) inverse network architecture that effectively solves the critical non-uniqueness problem in QCL inverse design, where multiple physical structures can yield the same performance metrics. By using a pre-trained forward network as an advanced, differentiable loss function and incorporating a random layer, the model can generate a diverse portfolio of physically plausible QCL structures for a single set of target metrics.

- **Systematic Validation of Machine Learning's Potential in QCL Design:** Through the successful implementation of the above components, this work systematically demonstrates the immense potential of machine learning to significantly

accelerate and enhance the QCL design workflow. The framework enables the automated generation of optimized QCL structures with predetermined properties in a fraction of the time required by traditional methods.

These contributions are not isolated but are tightly interconnected, forming a dependent workflow. The success of the inverse network is fundamentally reliant on the quality and scale of the dataset generated by the automated wavefunction identification program and on the accuracy of the forward network used as its evaluator within the tandem architecture. The inverse network's task is to learn the complex mapping from performance metrics back to a physical structure. If the training data (i.e., the structure-metric pairs) is flawed or insufficient, the inverse network's predictive capabilities will be severely compromised. The automated identification program is the key enabling technology that makes generating a sufficiently large and accurate dataset feasible.

Furthermore, during the tandem network's training, the forward network acts as a real-time "evaluator" or surrogate for the true physical solver. It provides the essential error gradients needed to guide the inverse network's learning process. Consequently, the accuracy of the forward network is paramount. An inaccurate forward model would provide misleading feedback, fundamentally undermining the training process and leading to a poorly performing inverse design tool.

1.6 Thesis Structure

The structure of this thesis is organized as follows:

- **Chapter 2: Theoretical Background and Literature Review.** This chapter introduces the fundamental working principles of Quantum Cascade Lasers, including the mechanisms of intersubband transitions within engineered superlattices and the role of the active region, which comprises injector, transition, and extractor sections. It covers key material systems, such as InGaAs/AlInAs on InP substrates, and primary performance metrics. The chapter also details the modeling and simu-

lation methods for QCLs, with a particular focus on the application of k·p theory to generate band diagrams. Concurrently, it reviews the progress of machine learning applications in photonics and semiconductor device design, with a focus on existing approaches for QCL design—including Bayesian optimization and genetic algorithms —and their current challenges.

- **Chapter 3: Automated Wavefunction Identification and Large-Scale Dataset Generation for QCL Active Regions.** This chapter addresses the bottleneck of manual wavefunction analysis in traditional QCL design. It details the development of the automated identification program, including its core algorithm based on a refined k-means clustering method and tailored probability formulas , its implementation, and its performance validation, which demonstrates over 95% accuracy in identifying critical energy states. The chapter will then describe the process of using this program to generate the large-scale dataset used for machine learning, which contains approximately 430,000 unique QCL structures mapped to their corresponding k·p metrics.

- **Chapter 4: Forward Prediction of QCL k·p Metrics via Neural Networks.** This chapter presents the design, training process, and performance evaluation of the forward neural network model. This model is trained to predict key k·p metrics, such as energy-level differences (E_{43}, E_{54}) and carrier lifetimes (τ_{43}, τ_{54}), directly from a QCL's structural parameters. The model's high accuracy, validated by a coefficient of determination (R^2) of approximately 90%, is discussed. Finally, the chapter elaborates on the critical role this forward model plays within the broader inverse design framework, where it functions as a fast, differentiable evaluator or an "advanced loss function".

- **Chapter 5: Inverse Design of QCL Active Regions Using a Tandem Neural Network.** This chapter focuses on the core challenge of QCL inverse design: the non-uniqueness problem, where multiple structures can lead to identical performance

metrics. It provides a detailed description of the proposed tandem neural network architecture, which includes the composition of the inverse (generative) network and the forward (evaluator) network. The training strategy and the design of the loss function are explained. The performance of this inverse design framework is demonstrated through experimental results, showcasing its ability to generate diverse and physically plausible QCL structures that satisfy specific target k·p metrics.

- **Chapter 6: Summary, Conclusions, and Future Work.** This chapter summarizes the main research work and the results achieved in this thesis. It distills the core conclusions, analyzes the limitations of the current study, such as inaccuracies in the forward network's predictions due to energy level hybridization, and provides an outlook on promising future research directions. These include improving model accuracy by modifying the k·p solver and expanding the framework's applicability to different material systems and device types, such as THz QCLs.

Chapter 2

Theoretical Framework and Literature Review

2.1 Quantum Cascade Lasers: Principles and Technology

2.1.1 Fundamental Working Principle of Semiconductor Lasers

The core of a semiconductor laser is a specially designed P-N junction, typically in a double heterostructure (DH) configuration, to achieve effective confinement of both charge carriers and the optical field. Its working principle can be broken down into three key physical processes: carrier injection and population inversion, stimulated emission and optical gain, and optical resonance and lasing oscillation[5].

When a forward bias voltage is applied to the laser diode, a large number of electrons from the N-type region and holes from the P-type region are injected into the central low-bandgap active region. Because the active region is sandwiched between high-bandgap cladding layers, these injected non-equilibrium carriers are confined within the active region, leading to a sharp increase in their concentration.

Under such high injection conditions, the statistical distribution of carriers can no longer be described by a single Fermi level. Instead, they are described by an electron quasi-Fermi level (E_{Fn}) and a hole quasi-Fermi level (E_{Fp}), respectively. The rigorous

criterion for achieving population inversion is the Bernard-Duraffourg condition, which states that the separation of the quasi-Fermi levels must exceed the material's bandgap energy (E_g)[5]:

$$E_{Fn} - E_{Fp} > E_g \tag{2.1}$$

This condition implies that the probability of finding an electron near the conduction band edge is much greater than the probability of an electron occupying a state near the valence band edge. Consequently, the rate of downward transitions induced by photons (stimulated emission) surpasses the rate of upward transitions (stimulated absorption), creating the possibility for optical amplification.

Within a system under population inversion, three primary optical processes occur:

- **Spontaneous Emission:** Electrons in the conduction band spontaneously recombine with holes in the valence band, emitting photons randomly. This is the primary mechanism for light emission in LEDs.

- **Stimulated Absorption:** An incident photon is absorbed by an electron in the valence band, causing it to transition to the conduction band, thus consuming the photon.

- **Stimulated Emission:** An incident photon induces an electron at a higher energy level to transition to a lower energy level, generating a new photon that is identical to the incident photon in frequency, phase, polarization, and direction of travel.

When the population inversion condition is met, the stimulated emission process becomes dominant. As a light signal propagates through the active region, its photon number grows exponentially, resulting in optical amplification. This amplification is characterized by the optical gain coefficient, g, which is proportional to the carrier concentration[5].

To achieve lasing oscillation, a positive feedback mechanism is required, which is typically provided by a Fabry-Pérot (F-P) resonant cavity. In semiconductor lasers, the simplest cavity is formed by two parallel facets created by cleaving the crystal along its

natural cleavage planes. Due to the refractive index difference between the semiconductor material and air, these facets act as mirrors with a reflectivity of about 30%[5].

Photons are reflected back and forth within this cavity, passing through the active region multiple times to acquire gain. Lasing oscillation begins when the optical gain accumulated in a single round trip precisely balances all the losses within the cavity. This critical state is known as the lasing threshold. When the injection current exceeds the threshold current, the gain becomes "clamped" at the threshold level, and the excess energy is efficiently converted into a coherent laser output[5].

2.1.2 Intersubband Transitions and the Electron Cascade Mechanism in QCLs

Quantum Cascade Lasers represent a distinct class of unipolar semiconductor light sources, architected from precisely engineered heterostructures. Their operating principle marks a fundamental departure from conventional bipolar diode lasers. Whereas traditional diode lasers generate photons through the recombination of electrons from the conduction band with holes from the valence band (i.e., interband transitions), their emission wavelength is consequently constrained by the intrinsic bandgap of the constituent semiconductor materials[1, 5]. In stark contrast, QCLs leverage an entirely different quantum mechanical process: the radiative transition of electrons between quantized subbands within the conduction band of a quantum well (i.e., an intersubband transition)[1, 5]. As the entire process involves only one type of charge carrier (electrons) traversing the heterostructure, QCLs are classified as unipolar devices.

The core innovation underpinning the QCL is the "cascade" effect. A typical QCL structure consists of tens to hundreds of identical modules, or "periods," connected in series. Each period is meticulously designed and comprises an active region and an injector region[1, 5]. When an external electric field is applied, it drives the electrons through this engineered potential landscape. An electron is injected via resonant tunneling from the injector region into the upper laser level (ULL) of the active region. It then radiatively re-

laxes to the lower laser level (LLL), emitting a single photon in the process. Subsequently, the electron is rapidly depopulated from the LLL, primarily through non-radiative LO-phonon scattering, and is transferred into the injector region of the subsequent period, where the cycle repeats[1, 5]. This process, illustrated for two cascade stages, is schematically shown in Figure 2.1.

This cascading mechanism is the key to the QCL's high power capabilities. In principle, a single injected electron can generate a number of photons equal to the number of cascaded periods as it traverses the structure. This phenomenon allows the quantum efficiency to significantly exceed the conventional limit of unity, leading to high optical output power[1, 5]. The external electric field is not merely a power source; it plays a critical role in precisely aligning the energy subbands across the entire heterostructure. This careful alignment, or "band-structure engineering," is essential to ensure efficient electron injection into the ULL, prevent escape to the continuum, facilitate rapid extraction from the LLL, and suppress parasitic current paths, thereby maximizing the gain and efficiency[1, 5].

The unipolar nature of QCLs and their reliance on band-structure engineering bestow upon them unparalleled flexibility in wavelength design[1, 5]. The energy of the emitted photon, and thus the laser's wavelength, is determined not by the material's innate bandgap but by the physical thicknesses of the quantum well and barrier layers. This allows designers to "artificially tailor" the energy separation between the subbands ($E_{\mathrm{ULL}} - E_{\mathrm{LLL}}$) by controlling layer thicknesses with near-atomic precision during epitaxial growth. It is this extraordinary degree of design freedom that has enabled QCLs to become the dominant coherent light source in the mid-infrared and terahertz spectral ranges, regions where conventional semiconductor lasers are either unavailable or perform poorly.

2.1.3 QCL Heterostructure: Design of the Injector, Active, and Extractor Regions

A single period, or "stage", of a Quantum Cascade Laser is a complex, meticulously engineered heterostructure. It is conventionally divided into three distinct functional regions:

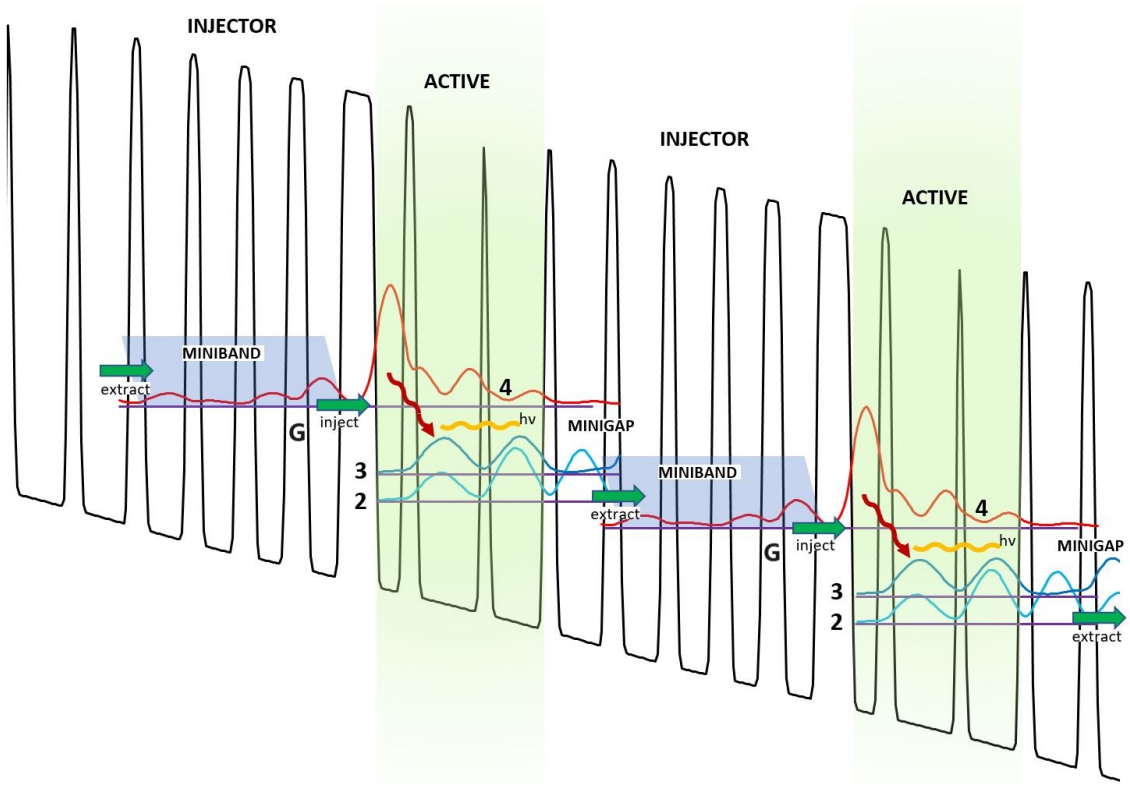


Figure 2.1: A schematic diagram of an AR of a mid-IR QCL with two cascading stages. The ground state, g, of the miniband injects electrons into the upper-laser level, 4. Next, an optical transition occurs from 4 to 3, the lower-laser level, which provides a photon of energy hv. Then, electrons from 3 relax to low-energy levels, 2, and are extracted to the miniband of the next stage.[13]

the injector region, the active region, and the extractor region. As depicted in Figure 2.1, electrons cascade through these stages, with the process repeating over 30-50 cascading stages to achieve the necessary optical gain for lasing[1].

Injector Region

The injector region serves as a carrier reservoir, collecting electrons from the extractor region of the preceding stage and injecting them precisely into the active region of the current stage. It is typically composed of a series of coupled quantum wells, which form a miniband—a quasi-continuum of states. The primary functions of the injector region are[1, 5]:

- **Carrier Reservoir**: It captures and holds electrons that have completed the lasing process in the previous stage, preparing them for the next cycle.
- **Selective Injection**: Through resonant tunneling, it efficiently injects electrons from its ground state (often labeled G_0 or 'g') specifically into the upper laser level (ULL) of the active region. For this to occur, the injector's ground state must be precisely aligned in energy with the ULL under the applied electric field.
- **Parasitic Leakage Suppression**: The energy structure of the miniband acts as a filter, preventing electrons from tunneling back out of the ULL and blocking parasitic transitions that would otherwise bypass the intended lasing process.

Active Region

The active region is the core of the QCL, where photon emission occurs. In this region, electrons undergo an intersubband transition from a higher energy state (the ULL) to a lower energy state (the LLL). In the this study, these levels are consistently labeled E_4(ULL) and E_3(LLL)[12, 13]. The key design objectives for the active region are[1, 5]:

- **Population Inversion**: The primary goal is to achieve and maintain population inversion between the ULL and LLL, meaning there are more electrons in the ULL

than in the LLL. This is essential for net optical gain.

- **Transition Probability**: The design must maximize the radiative transition probability between the ULL and LLL. This is governed by the spatial overlap of the electron wavefunctions of the two states (related to the dipole matrix element) and their energy separation (E_{43}), which determines the emission wavelength.

- **Transition Type**: Based on the spatial distribution of the ULL and LLL wavefunctions, transitions can be classified as "vertical" (occurring mostly within the same quantum well) or "diagonal" (occurring between adjacent quantum wells). Diagonal transitions often feature a longer ULL lifetime (τ_4), which is beneficial for population inversion, but may come at the cost of a smaller dipole matrix element. The choice between these designs represents a fundamental trade-off in AR engineering.

Extractor Region

The extractor region's function is to maintain population inversion by rapidly depopulating the LLL and efficiently transporting these electrons to the injector region of the subsequent stage[1, 5].

- **Rapid Depopulation**: Electrons are quickly removed from the LLL (E3) through scattering into a series of lower-energy extractor states (labeled E2 and E1). This process is dominated by fast, non-radiative Longitudinal Optical(LO) phonon scattering, ensuring the LLL remains empty.

- **Lifetime Management**: For robust population inversion, the design must ensure that the relaxation rate from the LLL to the extractor states is significantly faster than the overall decay rate of the ULL. This is achieved by engineering a large energy separation (approximately the LO-phonon energy) between the LLL and the extractor states to facilitate efficient phonon emission.

The intricate interdependence of these regions makes QCL design exceptionally challenging. Optimizing one region in isolation can adversely affect the others, degrading

overall device performance. Therefore, a successful QCL design must treat the entire stage as a single, coupled quantum system, requiring synergistic optimization of all its components.

2.1.4 Material Systems for Mid-Infrared Quantum Cascade Lasers

The performance of a Quantum Cascade Laser, particularly its operating wavelength, output power, and temperature characteristics, is fundamentally determined by the semiconductor material system used for its active region and waveguide. The choice of materials dictates key band parameters such as the conduction band offset (ΔE_c), electron effective mass (m^*), and phonon energies, which directly influence carrier confinement, intersubband transition efficiency, and non-radiative recombination rates. Over decades of development, several III-V semiconductor material systems have been successfully employed for QCL fabrication. This section reviews and compares several mainstream and emerging QCL material systems.

The InP-based InGaAs/AlInAs Material System

The InP-based material system is currently the most mature and widely used platform for mid-infrared (Mid-IR) QCLs, and it is the system adopted for the research in this dissertation. Its core advantages lie in its ability to be grown lattice-matched to the InP substrate, which offers good thermal conductivity and is supported by mature epitaxial growth technologies (MBE and MOCVD)[1, 21].

Lattice-Matched System The standard lattice-matched system utilizes $In_{0.53}Ga_{0.47}As$ for the quantum wells and $In_{0.52}Al_{0.48}As$ for the barriers. This combination provides a conduction band offset (ΔE_c) of approximately 0.52 eV [1]. This relatively large ΔE_c is sufficient to confine electrons effectively at room temperature, suppressing thermally activated leakage currents. This has enabled high-power, room-temperature continuous-wave (CW) QCLs with wavelengths covering from $\sim$3.8 μm to longer wavelengths[33]. However, the 0.52 eV offset limits its application at shorter wavelengths (< 3.8 μm), as

the higher subband energies reduce the effectiveness of carrier confinement.

Strain-Compensated System To overcome the wavelength limitations of the lattice-matched system, strain-compensation techniques were developed. This approach involves alternately growing tensile-strained barriers (e.g., high Al-content AlInAs) and compressively-strained quantum wells (e.g., high In-content InGaAs) in the active region, such that the net strain of the multilayer stack remains matched to the InP substrate[7]. This design can significantly increase the effective conduction band offset ($\Delta E_c > 0.7$ eV), thereby enhancing electron confinement and extending the emission wavelength of QCLs down to $\sim$3 μm[7]. However, the growth window for strain-compensated systems is narrower, demanding higher precision in epitaxial control.

The GaAs-based GaAs/AlGaAs Material System

GaAs/AlGaAs is one of the most historically studied III-V material systems and was an important platform in early QCL research. Its main advantages are the mature GaAs substrate technology, lower cost, and larger available wafer sizes. This system remains dominant for Terahertz QCLs[32]. For mid-IR applications, however, its conduction band offset is relatively small (e.g., $\Delta E_c \approx 0.33$ eV for an $Al_{0.45}Ga_{0.55}As$ barrier), leading to weaker electron confinement. This makes it difficult to achieve high-performance, short-wavelength mid-IR emission (typically limited to $\lambda > 8$ μm)[24]. Furthermore, the thermal conductivity of GaAs is inferior to that of InP.

The Antimonide-based InAs/AlSb Material System

In the quest for even shorter emission wavelengths, researchers have turned to the InAs/AlSb material system. This system boasts a very large conduction band offset of $\sim$2.1 eV, providing an extremely deep quantum well for electrons[30]. This massive ΔE_c allows carriers to be effectively confined even at very high energy levels, making the system ideal for developing short-wavelength mid-IR QCLs. Lasing down to 2.6 μm has been demonstrated[3]. Additionally, the small electron effective mass in InAs ($m^* \approx 0.023\ m_0$) can theoretically

lead to higher optical gain. The challenges, however, are significant, including the complexity of Sb-based material growth and device processing, as well as the limited availability of high-quality InAs or GaSb substrates.

Emerging Systems and Heteroepitaxy

Beyond the mainstream systems described above, several emerging research directions show great promise. Among them, the growth of III-V QCL structures on heterogeneous substrates like silicon (Si) is a particularly attractive frontier. The primary driver for this research is not merely cost reduction but the long-term vision of mid-infrared photonic integrated circuits (PICs)—the monolithic integration of QCLs with mature Si-based CMOS control circuitry, detectors, and other functional components [25]. Despite facing significant challenges from lattice and thermal mismatches, pioneering progress has been made by research groups, including that of Mawst, who have successfully demonstrated electrically injected QCLs on Si, paving the way for future optoelectronic integration [25].

2.1.5 Key Performance Metrics

The performance of a Quantum Cascade Laser is characterized by a set of interconnected, device-level macroscopic metrics. These metrics are not only the standard for evaluating device quality but also serve as the primary guide for their design and optimization. This section reviews several core performance metrics and elucidates their profound connection to the active region band structure and underlying microscopic physical parameters. These microscopic parameters, such as energy levels, and carrier scattering lifetimes, are the direct optimization targets for the machine learning-based inverse design model in this study.

Lasing Wavelength (λ)

The lasing wavelength is the most fundamental performance parameter of a QCL. It is determined by the energy difference ΔE between the initial state (ULL) and the final

state (LLL) of the optical transition[1, 5]:

$$\lambda = \frac{hc}{\Delta E} = \frac{hc}{E_{\mathrm{ULL}} - E_{\mathrm{LLL}}}$$

where h is Planck's constant and c is the speed of light in vacuum. In our designs, this corresponds to the energy difference E_{43} between levels E_4 and E_3. Through precise control over the thicknesses of the quantum wells and barriers, the band structure can be "engineered" to achieve customized wavelength emission over an extremely broad spectral range, covering from the short-wave mid-infrared (∼2.6 µm) to the terahertz region[1, 5].

B. Output Power (P_{out}) and Wall-Plug Efficiency (WPE)

Optical output power (P_{out}) is a core metric measuring the laser's ability to convert electrical energy into light, and it is critical for many applications. The Wall-Plug Efficiency (WPE, η_{WPE}) is the ultimate standard for evaluating the overall energy efficiency of the device, defined as the ratio of the optical output power to the total electrical input power[1, 5]:

$$\eta_{\mathrm{WPE}} = \frac{P_{\mathrm{out}}}{V \cdot I}$$

where V and I are the total voltage applied across and total current flowing through the device, respectively.

WPE is limited by a combination of efficiency factors, including the injection efficiency (η_{inj}) for transporting electrons into the ULL, the internal quantum efficiency (η_{iqe}) of the ULL, and the photon extraction efficiency (η_{ext}) from the laser facet.

C. Threshold Current Density (J_{th})

The threshold current density (J_{th}) is the minimum drive current density required to achieve lasing action. A low J_{th} is advantageous for reducing device power consumption and thermal load, which is fundamental for achieving continuous-wave (CW) and high-duty-cycle pulsed operation. The value of J_{th} is determined by the gain-loss balance and

can be expressed as[1, 5]:

$$J_{\text{th}} = \frac{\alpha_{\text{w}} + \alpha_{\text{m}}}{g\Gamma} + J_{\text{leak}}$$

where α_{w} and α_{m} are the waveguide and mirror losses, g is the material gain coefficient, Γ is the optical confinement factor, and J_{leak} represents the sum of all leakage current channels that do not contribute to gain. The material gain coefficient g is directly related to several microscopic parameters, including the transition dipole matrix element (z_{ul}), level lifetimes, and lineshape broadening. Concurrently, leakage currents J_{leak}, such as those escaping via high-energy states, are also determined by the band structure (e.g., barrier heights and parasitic state locations).

Slope Efficiency (η_{s}) and Internal Quantum Efficiency (η_{iqe})

The slope efficiency ($\eta_{\text{s}} = dP_{\text{out}}/dI$) is defined as the rate of change of output power with respect to the injection current above threshold (in units of W/A). It is a key factor determining the maximum output power. The slope efficiency is directly related to the internal quantum efficiency (η_{iqe}) and the light extraction efficiency (η_{ext}).

The internal quantum efficiency (η_{iqe}) represents the probability that one photon is generated for each electron injected above threshold. It is affected by all carrier consumption mechanisms that compete with the stimulated emission process. These mechanisms primarily include: (1) non-radiative transitions from the ULL to other non-lasing levels; and (2) carriers that leak directly from the injector/active region to the next stage or into the continuum without participating in the radiative process [4, 6]. It is crucial to note that leakage current often scales with operating current. This implies that even if a device has a low J_{th}, significant leakage at high-power operating points can cause a severe drop in η_{iqe} and slope efficiency, ultimately limiting the WPE[1, 5].

The macroscopic performance metrics discussed above ($\lambda, P_{\text{out}}, J_{\text{th}}$, etc.) are the ultimate goals for a designer. However, they are the result of a combination of multiple microscopic physical processes, and the relationships between them are exceedingly complex. Directly targeting these macroscopic metrics for inverse design would require com-

putationally prohibitive, full-device-level simulations that include thermal effects, optical modes, and carrier transport.

To simplify this highly complex inverse design problem, this study selects a set of microscopic physical metrics, which are intimately related to the aforementioned performance metrics and can be directly calculated by the k.p method, as the direct correlation and optimization targets for the machine learning model. The chosen parameters and the rationale are as follows:

- **Energy Level Spacings (E_{43} and E_{54})**: E_{43} ($E_4 - E_3$) directly determines the lasing wavelength λ and is the primary design target. E_{54} ($E_5 - E_4$) represents the energy separation from the ULL (E_4) to the next higher-lying state (E_5). This separation acts as a critical barrier to suppress the thermal escape of carriers, and is thus strongly correlated with the device's characteristic temperature T_0 and efficiency droop at high currents[1, 5].

- **Electron Scattering Lifetimes/Rates (τ_{43} and τ_{54})**: τ_{43} is the transition lifetime from the ULL (E_4) to the LLL (E_3). A relatively long τ_{43} is beneficial for establishing population inversion. τ_{54} is the scattering lifetime associated with the high-energy leakage state E_5. In the design process, a longer τ_{54} is considered beneficial for mitigating the impact of this parasitic channel. Engineering the band structure to increase this lifetime can, to a certain extent, enhance the stability of the carrier environment near the upper lasing level, thereby positively influencing device performance. These lifetimes collectively determine the efficiency of population inversion and the magnitude of leakage currents, thereby influencing J_{th} and η_{iqe}[1, 5, 12].

By simplifying the design problem into a search for an optimal solution within a multi-dimensional space of microscopic parameters, our approach can efficiently explore and optimize the core quantum design of the QCL active region. However, we must acknowledge that a critical link remains between this set of optimized 'k.p' metrics and the prediction

of the final, fabricable device's macroscopic performance. Bridging this gap requires more sophisticated device-level models that couple optical waveguiding, thermal management, and complex carrier transport effects. Integrating the inverse design methodology of this study with such device-level simulation tools represents a vital and challenging direction for future research.

2.2 Modeling and Simulation of Quantum Cascade Lasers

2.2.1 The k·p Method for Electronic Band Structure Calculations

The design and optimization of Quantum Cascade Lasers are fundamentally rooted in the precise engineering of the electronic band structure. The ability to accurately predict the quantized energy levels (subbands) and corresponding electron wavefunctions within the QCL's complex heterostructure is paramount for determining key device properties, including the emission wavelength, optical gain, and carrier transport dynamics. For this purpose, the k·p perturbation theory stands as a cornerstone modeling tool in semiconductor physics, offering a powerful and computationally efficient semi-empirical framework for these calculations.

The k·p method is particularly effective at describing the energy-momentum ($E - k$) relationship near critical points in the Brillouin zone, such as the Γ-point (k=0), where the fundamental band gap is located in the direct-gap semiconductors typically used for QCLs[29]. However, k·p Hamiltonians are intrinsically formulated for bulk, periodic crystals. To apply them to QCL heterostructures, the theory is almost universally coupled with the Envelope Function Approximation (EFA)[29]. Within the EFA framework, the total electron wavefunction is separated into a rapidly varying, periodic Bloch function (characteristic of the bulk material in each layer) and a slowly varying envelope function that describes the quantum confinement imposed by the potential of the wells and barriers. This powerful abstraction allows the use of established bulk material parameters in a Schrödinger-like equation that governs the envelope function, effectively modeling the

quantum-confined states of the entire heterostructure.

A hierarchy of k·p models exists, each offering a different balance between physical completeness and computational cost. The choice of model is a critical engineering trade-off.

- **Single-Band Model:** The simplest approach, considering a single, non-degenerate band. While computationally trivial, it yields a parabolic energy dispersion and is generally inadequate for QCLs as it entirely neglects non-parabolicity—a dominant effect in these devices—as well as band mixing and spin-orbit interactions.

- **3-Band Kane Model:** This model is a widely used and pragmatic choice for QCL simulations. It explicitly considers the interaction between the s-like conduction band (CB) and the p-like light-hole (LH) and spin-orbit split-off (SO) bands. Its primary advantage is that it naturally and efficiently captures the most critical feature for QCLs: conduction band non-parabolicity[15, 29]. In QCLs, electrons occupy subbands at significant energies above the conduction band minimum, where the band's $E - k$ relationship deviates strongly from a simple parabola; this effect profoundly alters subband energies and spacing. By including the key interactions responsible for this phenomenon, the 3-band model provides the necessary accuracy for wavelength prediction and gain calculations in typical n-type unipolar QCLs, where the valence band plays a secondary role. Its main limitation is an inaccurate description of the heavy-hole (HH) band and overall valence band structure.

- **Multi-Band Models (e.g., 6-Band, 8-Band):** For higher accuracy or for modeling more complex phenomena, more sophisticated models are required. The 6-band Luttinger-Kohn model focuses specifically on providing a detailed description of the valence band manifold (HH, LH, SO)[18], which is crucial for p-type devices or understanding hole transport. The 8-band Kane model is a more comprehensive approach that includes all eight bands (CB, HH, LH, SO, with spin degeneracy) near the band edge[15, 29]. It provides a balanced description of both conduction

and valence bands and is essential when effects like strain, detailed carrier scattering rates (which are sensitive to band mixing), or interband processes are important. For example, 8-band models can predict LO phonon scattering rates more accurately than simpler models by better accounting for the mixing of conduction and valence band character in the wavefunctions. However, the significant increase in computational complexity and the challenge of parameterizing these larger Hamiltonians often make them impractical for large-scale simulations.

For this research, which involves the generation and analysis of a very large dataset of QCL structures, a trade-off between accuracy and computational feasibility is essential. While an 8-band model offers a more complete physical picture, its computational cost is prohibitive for such a task. Therefore, the 3-band Kane model was selected. This model represents an optimal compromise, capturing the essential physics of conduction band non-parabolicity—the most critical factor for determining the lasing transition energy—with sufficient accuracy while maintaining the computational efficiency required for high-throughput simulation. This is a well-established, pragmatic approach for the design of n-type QCLs and provides a robust foundation for the subsequent analysis in this thesis.

2.2.2 Carrier Transport and Leakage Mechanisms

The performance of a Quantum Cascade Laser is fundamentally dependent on the efficient transport of carriers (electrons) within its meticulously engineered active region. Ideally, an electron follows a precise cycle: selective injection into the Upper Laser Level (ULL) via resonant tunneling[1], undergoing a radiative transition within the active region, efficient extraction from the Lower Laser Level (LLL) via mechanisms like fast Longitudinal Optical (LO) phonon scattering[1], and subsequent transport to the injector of the next stage in preparation for another photon emission. In practical devices, however, various competing non-radiative pathways cause electrons to deviate from this ideal cycle. These processes are collectively known as carrier leakage. Carrier leakage is a critical factor limiting QCL performance, as it directly increases the threshold current, reduces the slope efficiency,

and degrades the high-temperature operational capability of the device. Therefore, a deep understanding and effective suppression of the various leakage mechanisms are central challenges in high-performance QCL design.

Carrier leakage in QCLs originates from a variety of physical processes, which can be broadly categorized into thermally activated leakage, parasitic tunneling, and scattering-induced leakage.

- **Thermally Activated Leakage**: This is a primary cause of performance degradation at high temperatures.
  - **Leakage to Higher-Energy States**: Electrons in the ULL can be thermally excited (phonon-assisted) to higher-lying parasitic bound states within the active region or directly into the continuum states above the barriers. Once in these channels, the electrons relax non-radiatively or are incorrectly injected into the next stage, failing to contribute to the laser gain. The magnitude of this leakage type has an exponential dependence on the energy separation between the ULL and the leakage state ($E_{leak} - E_{ULL}$), making the design of high barriers and energetically distant parasitic levels crucial for suppressing thermal leakage[**"cite –botez2023mid"**, 12].
  - **Thermal Backfilling of the LLL**: Electrons from the injector region of the subsequent stage can also be thermally excited and tunnel back into the LLL of the current stage[1]. This process increases the LLL population, thereby directly reducing the population inversion ($N_{ULL} - N_{LLL}$) and harming the optical gain.
- **Parasitic Tunneling Leakage**: Even at low temperatures, electrons can bypass the lasing transition via undesired tunneling paths. This can occur if energy levels in the injector accidentally align with the LLL or other non-lasing states due to improper design or incorrect bias field[1]. Furthermore, at high electric fields, parasitic tunneling paths can allow electrons to tunnel directly from the injector to the

extractor region of the next stage, completely bypassing the current active region and creating a significant shunt leakage current.

- **Scattering-Induced Leakage**: Structural imperfections at the heterointerfaces are another significant source of leakage.
  - **Interface Roughness (IFR) Scattering**: The interfaces between quantum wells and barriers are not atomically flat. This roughness induces elastic scattering, which can transfer electrons from the ULL into parasitic states or directly to the LLL, bypassing the radiative pathway[1]. Recent advanced modeling has suggested that IFR-triggered leakage through high-energy states can be a dominant and previously underappreciated loss channel even in record-performance QCLs[1]. This highlights the paramount importance of achieving high-quality interfaces during epitaxial growth.

Carrier leakage has a direct and detrimental impact on the key performance metrics of a QCL. The leakage current (J_{leak}) acts as a shunt path that contributes no useful gain but increases the total current density (J_{th}) required to reach the lasing threshold [1]. Above threshold, leakage continues to divert injected carriers, reducing the probability that an electron injected into the active region is converted into a photon, thereby severely impacting the internal quantum efficiency and slope efficiency (η_s) [1]. As many leakage mechanisms are thermally activated, their effects are exacerbated with increasing temperature, leading to a rapid rise in J_{th} and a drop in η_s. This is quantified by lower characteristic temperatures, T_0 and T_1, respectively [1]. Consequently, minimizing leakage current across the entire operating current and temperature range is a fundamental prerequisite for achieving high output power, high wall-plug efficiency (WPE), and robust high-temperature operation [1, 12].

2.2.3 Limitations of Traditional Design and Simulation Methods

Although traditional physics-based simulation methods, such as the k.p theory and Non-Equilibrium Green's Function (NEGF), have provided a solid foundation for the understanding and design of Quantum Cascade Lasers, they exhibit several inherent limitations in practice. These challenges not only prolong the development cycle but also constrain the exploration of the vast design space, thus motivating a shift in the research community towards data-driven approaches[12, 13].

High Cost in Time and Computation

The conventional QCL design workflow, heavily reliant on expert knowledge and extensive iterative testing, is an inefficient "trial and error" approach. This process is not only time-consuming and labor-intensive but can often extend over weeks or even months to finalize a single new design. A primary bottleneck is the identification and labeling of wavefunctions. In the traditional flow, after generating a band diagram with a k.p solver, an expert must manually and visually identify crucial wavefunctions, such as the upper and lower laser levels. This manual step becomes a significant bottleneck when analyzing a large number of different structures, severely impeding the efficiency of design optimization.

Significant Challenges in Design Optimization

QCL design optimization faces two major challenges: the "curse of dimensionality" and the difficulty of finding the global optimum.

- **Curse of Dimensionality**: A single QCL period can consist of dozens of layers , where the thickness and even composition of each layer are tunable parameters. This results in an extremely high-dimensional design space where the number of possible design combinations grows exponentially, making it computationally infeasible to find the optimal design through an exhaustive search.
- **Difficulty in Global Optimization**: Due to the high-dimensional and complex nature of the design space, the QCL performance function typically has numerous

local optima. Traditional optimization algorithms, such as Bayesian optimization or genetic algorithms, are prone to getting trapped in these local optima, making it difficult to guarantee that the globally optimal design is found.

Over-reliance on Designer Experience

The traditional design process depends heavily on the designer's physical intuition and accumulated expertise. While experience is valuable for guiding the design direction, this reliance creates two problems: first, it makes the design process difficult to standardize and automate; second, it may lead to overlooking unconventional or counter-intuitive design solutions that could offer superior performance.

These limitations highlight the urgent need for more efficient and automated design tools. In stark contrast to traditional physics solvers, which can only evaluate one design at a time, a well-trained machine learning model can predict or generate design metrics almost instantaneously. For instance, a neural network approach can complete over 10,000 metric evaluations in less than a minute, a task that would take about a day using conventional methods, even when they are integrated with automated wavefunction identification. This dramatic increase in efficiency opens the door to effective exploration and optimization within the vast QCL design space[12, 13].

2.3 Applications of Machine Learning in Photonics and Semiconductor Device Design

2.3.1 Overview of Machine Learning in Scientific Discovery

Machine learning (ML), a branch of artificial intelligence, is driving a profound transformation in scientific discovery by building models that can automatically learn from and improve with data. Unlike traditional programming where humans provide explicit instructions, ML sets up a situation for the program to learn to solve a problem on its own by feeding it input and expected output. This data-driven paradigm can reveal inherent

connections and nonlinear physical correlations hidden within complex, high-dimensional data, thereby significantly accelerating the process of scientific research.

ML algorithms can be broadly categorized into three main types:

- **Supervised Learning:** The model is trained on a labeled dataset, learning the mapping from inputs to outputs. For example, training a model to predict the performance of a Quantum Cascade Laser structure based on a dataset of known structures and their corresponding performance metrics.
- **Unsupervised Learning:** The model works with unlabeled data, attempting to find inherent structures or patterns, such as using clustering algorithms to identify key wavefunctions in a QCL band diagram.
- **Reinforcement Learning:** The model (agent) learns by interacting with an environment, adjusting its strategy based on rewards or penalties to maximize a cumulative reward for a specific goal.

Among these methods, Artificial Neural Networks (ANNs), and particularly Deep Neural Networks (DNNs), have become a key enabling technology for breakthroughs in photonics and numerous other science and engineering fields, owing to their powerful ability to approximate complex non-linear functions.

2.3.2 Data-Driven Forward Prediction in Photonics

In the fields of photonics and semiconductor device design, ML (especially neural networks) has been widely used to build "Forward Models". These models take the physical structural parameters of a device (e.g., layer thicknesses, material compositions, and applied electric field in a QCL) as input to predict its corresponding performance metrics or optical properties (e.g., figure of merit, gain, emission wavelength, energy-level differences)[11, 12, 13].

The fundamental advantage of these models is their exceptional computational efficiency. Traditional device performance evaluation relies on computationally expensive

physics-based simulation methods. For instance, work by Hernandez et al. showed that evaluating 27,000 designs using a traditional 1D Schrödinger solver took 32 hours[10], whereas a trained ML model could predict the performance of approximately 1 billion QCL designs in just 8 hours on a personal computer, representing a speed-up of more than five orders of magnitude. Beyond speed, ML-guided optimization has also led to substantial performance enhancements; for example, Gmachl and collaborators achieved a 1.5 to 2-fold increase in the figure of merit (FoM) of initial QCL designs with ML assistance[10].

Specific application examples are widespread. Researchers commonly use Multi-Layer Perceptrons (MLPs) and other neural networks to predict the laser transition FoM from inputs like layer thicknesses and applied electric field[10, 11]. Additionally, Physics-Guided Neural Networks (PGNNs), a type of hybrid model, have shown performance superior to pure physics-based models in QCL spectroscopy applications by incorporating outputs from physical models as inputs to the neural network[20]. These forward models not only serve as rapid evaluation tools but also often act as crucial evaluators within more complex inverse design architectures.

2.3.3 The Inverse Design Paradigm: Challenges and ML Solutions

Corresponding to forward prediction is "Inverse Design", whose objective is to deduce the physical device structure that can achieve a desired performance or function (e.g., a specific operating wavelength, a low threshold current).

However, inverse design is inherently more challenging than forward modeling, with the most critical challenge being the non-uniqueness problem[12]. Multiple, vastly different physical structures can yield identical or very similar performance metrics. This "one-to-many" mapping makes it exceedingly difficult to directly train a simple inverse neural network, as the model, in trying to minimize its loss, would attempt to learn the "average" of these different structures—an averaged parameter set that is often physically meaningless and fails to accurately reproduce the target performance[12].

To address this fundamental challenge, researchers have developed various ML-based inverse design methodologies.

- **Tandem Networks**: Pioneered by Zongfu Yu's group for nanophotonics, this architecture ingeniously cascades an inverse network with a pre-trained, fixed forward network[17]. During training, the structure generated by the inverse network is fed to the fixed forward network for evaluation, and the difference between the two networks' outputs is used to calculate the loss, updating only the inverse network's weights.

- **Generative Adversarial Networks (GANs):** Comprising a generator and a discriminator, they learn the underlying data distribution through adversarial training. Deep Convolutional GANs (DCGANs), for example, have been used to design high-efficiency photonic power dividers[23].

- **Evolutionary Algorithms (GAs):** Mimicking natural selection, GAs optimize designs through operations like crossover and mutation. GAs have been applied to design QCL active regions for specific target wavelengths and gain[31]. A particularly effective strategy combines GAs with ML models, using a trained forward model as the fitness function, which has reportedly accelerated the search process by a factor of 120[31].

- **Bayesian Optimization (BO):** An efficient global optimization method particularly suited for expensive "black-box" functions. This method has been successfully used with non-equilibrium Green's function (NEGF) transport models to optimize THz QCLs[8], not only achieving a record maximum operating temperature of 210 K but also revealing that a two-well scheme is superior for high-temperature performance.

2.3.4 The Data Generation Bottleneck: The Manual Labeling Dilemma and the Need for Automation

Despite the great potential of machine learning in QCL design, its development and application are severely constrained by a fundamental problem: the general scarcity of large, high-quality public datasets.

The root of this problem lies in a manual bottleneck inherent to traditional data generation workflows[13]. A standard simulation-based data generation process typically involves the following steps: first, a solver (e.g., a k·p solver) is used to compute the band diagram and its corresponding wavefunctions for a given QCL structure; then comes the most critical step: a domain expert must manually inspect the band diagram to identify and label all important wavefunctions, such as the upper and lower laser levels, injector/extractor states, and high-energy leakage paths etc.. Only after this manual labeling can the data be fed into a subsequent metrics solver to calculate meaningful performance parameters like energy differences and lifetimes, thus forming a complete "structure-to-performance" data point suitable for ML.

This reliance on human supervision for labeling is the crux of the data generation problem. It not only requires the operator to have deep domain expertise but is also extremely time-consuming and labor-intensive. When tens of thousands or even more data points are needed to train a robust neural network, performing this manual identification and checking for each one is clearly impractical. This makes the entire data acquisition process inefficient, even with the use of simulation technology, and unable to meet the demands of modern machine learning[13].

In summary, it is this deep reliance on manual labeling, rather than simulation speed alone, that constitutes the largest barrier to large-scale QCL dataset generation. It directly leads to the small size and high acquisition cost of existing datasets. Therefore, developing an automated wavefunction identification program to replace the manual labeling step and enable an end-to-end automated workflow from structure to full performance labels is no longer just a technical improvement; it is an indispensable core prerequisite for the

comprehensive and deep application of machine learning methods in QCL design.

Chapter 3

Automated Wavefunction Identification and Large-Scale Dataset Generation for QCL Active Region Design

3.1 Introduction: the Need for an Automated Solution

The performance of a Quantum Cascade Laser is critically dependent on the meticulous structural design of its active region. The conventional QCL design workflow is a "trial and error" process that relies heavily on expert knowledge and extensive iterative testing. A core component of this workflow is the manual wavefunction analysis performed by the designer after obtaining the electronic states from a k·p simulation or other computational tools. This step requires the designer to leverage deep physics expertise and experience to manually inspect a complex band diagram and identify which states correspond to the upper laser level, lower laser level, injector and extractor states, and potential leakage paths.

This reliance on manual supervision constitutes the primary bottleneck of traditional

design methodologies. The process is not only time-consuming and labor-intensive, with the finalization of a single design often taking weeks or even months, but the repetitive nature of manual identification for numerous different structures is also prone to error. For research aiming to leverage machine learning for QCL design, which requires training sets containing thousands or even hundreds of thousands of structure-to-performance data pairs, depending on manual wavefunction identification and labeling is clearly unrealistic. Therefore, developing a program capable of automatically, accurately, and rapidly identifying these key wavefunctions has become a critical prerequisite for the successful application of ML, particularly data-driven inverse design, to the field of QCLs.

An automated wavefunction identification program is more than a mere efficiency improvement over the existing workflow; it is an enabling technology that lays the foundation for the entire ML-based design paradigm. Advanced ML models, particularly Deep Neural Networks (DNNs), typically require vast amounts of data to learn the highly non-linear mappings between input and output in complex physical systems and to achieve good generalization. If a critical step in the data generation pipeline—namely, wavefunction identification—is manual and slow, creating large-scale datasets to meet the demands of ML becomes impractical. By automating this bottleneck, the generation time for a single data point can be drastically reduced from hours or even days to mere seconds. This speed-up makes it feasible to generate the large datasets, containing hundreds of thousands of data points, that are necessary for machine learning. It fundamentally alters the QCL design paradigm, opening the door for the adoption and development of advanced, data-intensive ML algorithms such as generative inverse design networks.

3.2 Key Energy States in Quantum Cascade Lasers

The performance of a Quantum Cascade Laser is primarily determined by the design of its Active Region. The active region is an engineered superlattice heterostructure composed of multiple layers of semiconductor materials. Within this structure, under an applied electric field, electrons form a series of discrete quantized states in the quantum wells and

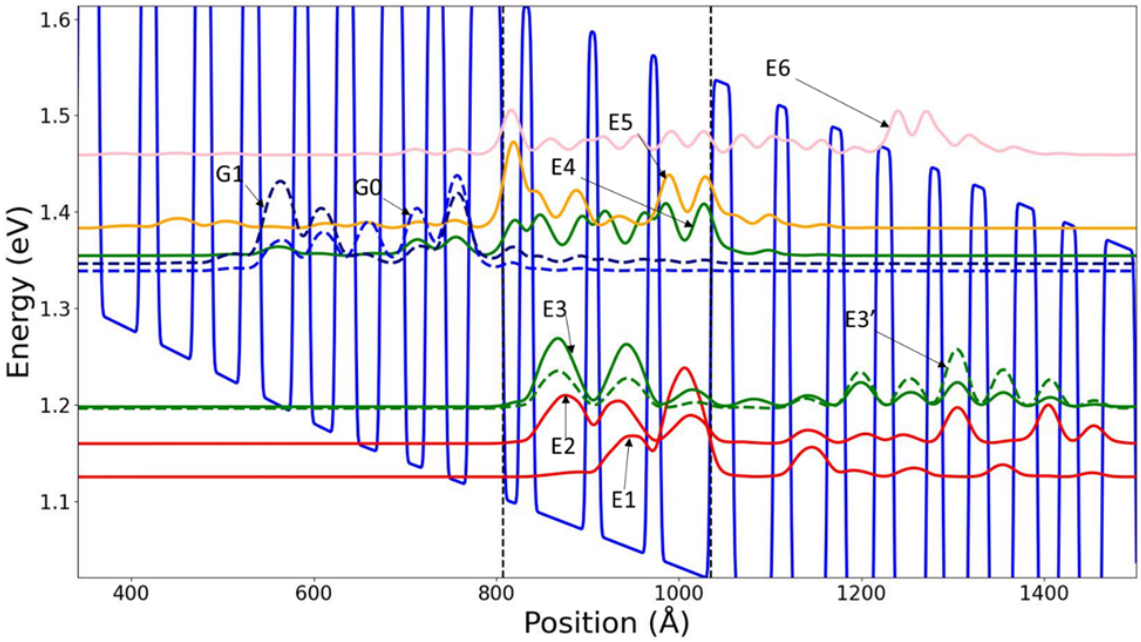


Figure 3.1: An example of the energy states for identifying.

emit photons through intersubband transitions between these levels. For achieving efficient and stable laser emission, the precise identification and engineering of key wavefunctions and their corresponding energy states within the active region are crucial.

These key states, based on their function in the carrier transport and photon emission process, can be categorized into states within the injector, active, and extractor regions, as illustrated in Figure 3.1. The primary states that require precise identification include[1, 13]:

- **Upper and Lower Laser Levels**: Denoted as E_4 (UUL) and E_3 (LLL), respectively. The electron transition from E_4 to E_3 is the core process that generates laser photons. The energy difference between these two levels directly determines the emission wavelength of the QCL, making their accurate identification fundamental to designing lasers for specific wavelength applications.

- **Injector States**: Denoted as G_0 and G_1. Located in the injector miniband, they act as a carrier reservoir, responsible for efficiently injecting electrons into the upper laser level E_4 of the subsequent cascade stage. G_0 typically refers to the ground

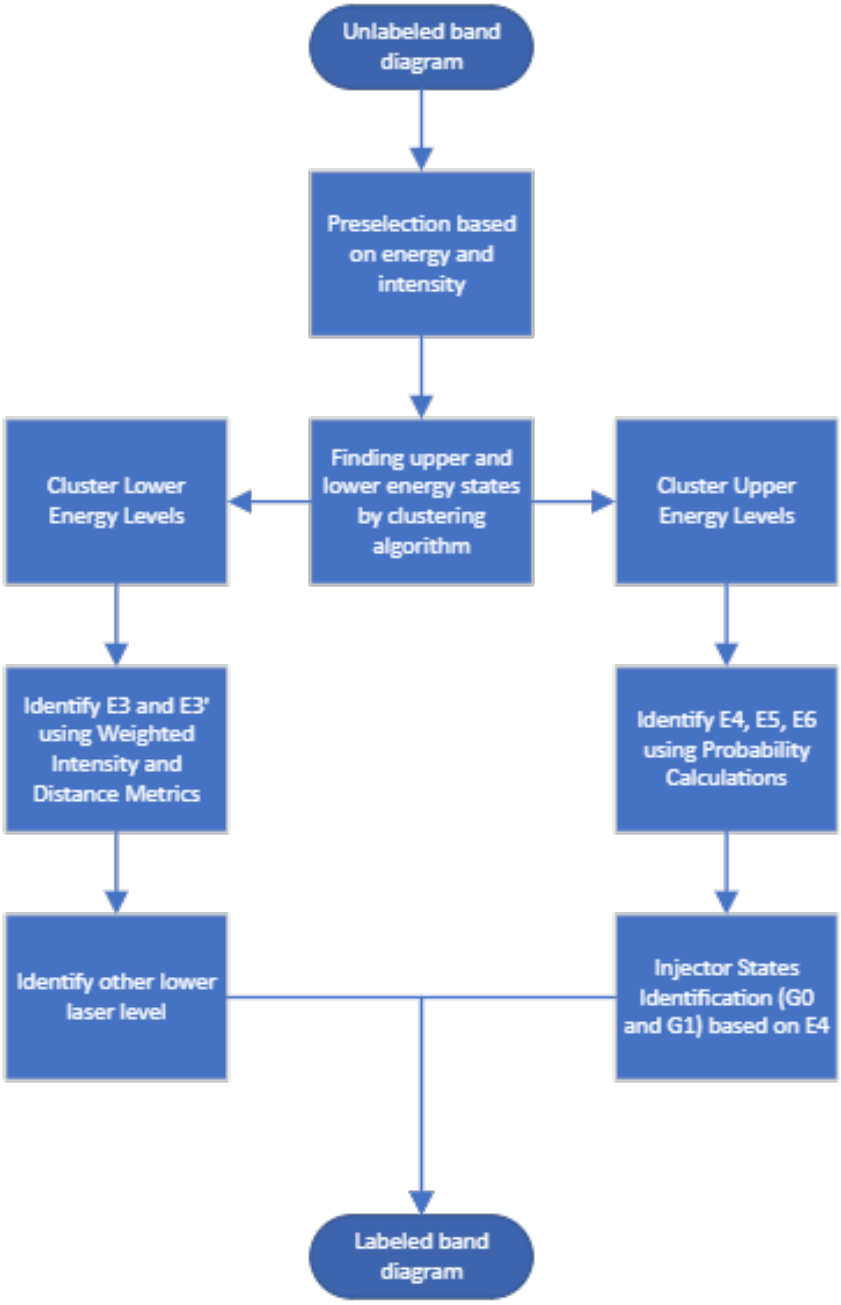


Figure 3.2: The progress of the automated wavefunction identification.

state. A well-designed injector state is critical for maintaining population inversion and enabling continuous-wave operation.

- **Extractor States**: Denoted as E_3'. After an electron transitions from E_4 to E_3 and emits a photon, the function of this state is to quickly and efficiently "extract" the electron from the lower laser level E_3 and transport it to the injector region of the next cascade. Efficient extraction is vital to prevent electron accumulation in the lower level and to maintain population inversion, which is a prerequisite for achieving high output power.

- **Low-Energy AR States**: Denoted as E_1 and E_2. These are low-energy active region states located below the lower laser level E_3. Electrons relax from E_3 to these lower levels before being extracted to the next injector region. While they do not directly participate in light emission, they are part of the electron relaxation pathway and affect carrier extraction efficiency and level lifetimes.

- **Leakage States**: Denoted as E_5 and E_6. These are parasitic states with energies higher than the upper laser level E_4. Electrons in the active region can leak into the continuum via these states without contributing to the laser gain, a phenomenon known as "carrier leakage," which significantly reduces the laser's efficiency and power. Therefore, they must be accurately identified during the design process to optimize the structure and suppress carrier leakage.

Figure 3.1 shows a typical band diagram where the aforementioned key states that need to be identified are labeled.

3.3 Automated Wavefunction Identification

Conventional QCL design relies heavily on expert knowledge, where wavefunctions in the band diagram are manually identified to evaluate a design; this process is not only time-consuming and labor-intensive but also impractical for exploring a vast design space[13].

To address this bottleneck, we have developed an automated wavefunction identification program that can rapidly and accurately identify and label the key states in a vast number of QCL structures. This program is foundational for large-scale data generation and subsequent machine learning applications[13].

The operating principle of this automated identification program is shown in Figure 3.2. Its core combines a series of physics-based filtering rules with an unsupervised machine learning algorithm tailored for this task. The full technical implementation details, including all mathematical formulas, are described in Appendix A. Its main workflow is summarized as follows:

1. **Preliminary Selection**: This is the first data processing step, aimed at filtering out clearly irrelevant wavefunctions. The program first selects wavefunctions based on their energy and intensity within the active region; those with an intensity below a specific threshold (0.1) are filtered out. It then analyzes the relative intensity distribution of a wavefunction across the active, injector, and extractor regions, as well as its center of gravity, to exclude those that primarily belong to adjacent cascade stages.

2. **Clustering into Groups**: After the initial selection, the program employs a refined k-Means clustering algorithm to group the remaining wavefunctions. To make the clustering results more physically meaningful, two key optimizations are implemented: first, weights are assigned to wavefunctions based on their intensity to amplify the influence of stronger states in the clustering process; second, a small, controlled random variation (e.g., ± 0.015eV) is introduced to the energy level of each wavefunction before clustering to effectively prevent the algorithm from incorrectly splitting closely spaced levels. The optimal number of clusters (the k-value) is not preset but is automatically determined by calculating the silhouette score, which ensures the best cohesion and separation of the groups. This step effectively divides the wavefunctions into primary "upper energy" and "lower energy" clusters.

3. **Specific State Identification**: Once the macroscopic grouping is complete, the program performs more detailed identification within each cluster.

   - **Lower-Level Identification (E_1, E_2, E_3, E_3')**: In the lower energy cluster, a further weighting strategy is used to locate E_3, which prioritizes positions that have better wavefunction overlap with E_4. Subsequently, the extractor state E_3' is identified by finding the state "closest" to E_3 based on a composite distance metric that accounts for wavefunction shape difference, energy difference, and intensity. The remaining wavefunctions in this cluster are collectively labeled as the low-energy AR states E_1 and E_2.
   - **Upper-Level Identification (E_4, E_5, E_6)**: In the upper energy cluster, the number of resulting sub-clusters directly indicates the likely types of states present (e.g., one sub-cluster implies only E_4 exists, while two suggest the presence of E_4 and E_5). The final state assignment is determined by a set of probability calculations. These calculations consider factors such as the proximity of a candidate wavefunction's energy to the cluster's base energy and its intensity in the active region. To prevent misidentification of states with very similar energies (e.g., a state with an energy extremely close to E_4 being mistaken for E_5), additional filter factors are introduced.
   - **Injector-State Identification (G_0, G_1)**: Finally, the program searches for injector states outside the initial selection range based on the already identified E_3 and E_4 levels. Each candidate wavefunction is assigned a **possibility score**, which is a product of multiple factors including its energy proximity to the midpoint of E_3 and E_4 and its intensity in the injector region. The two wavefunctions with the highest scores are ultimately designated as G_0 and G_1.

3.4 Performance Validation of the Automated Wavefunction Identification Program

To rigorously evaluate the accuracy and reliability of the developed automated wavefunction identification program, a detailed validation study was conducted. This comprehensive verification process involved the following steps:

First, from our extensive dataset of approximately 430,000 generated data points, we randomly selected 200 quantum cascade laser structure samples. For each of these 200 samples, experienced researchers manually identified and labeled the relevant wavefunctions in their respective band diagrams. This meticulous manual identification served as the "ground truth" or "golden standard" against which the automated program's performance was evaluated. Subsequently, the results of the automated program's identification for these 200 samples were compared one by one with the manual annotations. During this comparison, the number of identification errors for each critical energy level (E_1 to E_6, G_0, G_1, and E_3') was meticulously tallied, and the corresponding accuracy rates were calculated.

The validation results, detailed in Table 3.1, demonstrate that the automated wavefunction identification program achieves an overall accuracy exceeding 95%, indicating its high reliability. Specifically, the highest accuracy is observed for the upper lasing state (E_4) and the lower lasing state (E_3). Due to their distinct characteristics and typical positioning on either side of the superlattice (SL) minigap in the band diagram, the program achieved a remarkable 100% identification accuracy for these states.

However, the accuracy rates for the extractor state E_3' and the high-energy leakage state E_6 were marginally lower. For instance, E_3' showed an accuracy of 96.0%, with 8 errors identified among the 200 samples, while E_6 had an accuracy of 96.5%, accounting for 7 errors. These discrepancies primarily arise because E_3' occasionally gets confused with some other extractor states, and E_6's characteristics can be less distinct due to effects like level hybridization. Nevertheless, the identification accuracy for other energy levels, such as the low-energy states (E_1 and E_2 combined) and the injector states (G_0 and G_1), all

Table 3.1: The accuracies of the states labeled by the automated identification program

States	**Number of Errors**	**Accuracy(%)**
Lower States (incl. E_1, E_2)	1	99.5
E_3	0	100
E_3'	8	96.0
E_4	0	100
E_5	2	99.0
E_6	7	96.5
G_0	2	99.0
G_1	3	98.5

exceeded 98.5%.

Despite the presence of a small number of identification errors, particularly in complex scenarios like distinguishing E_3' from other extractor states, the automated program successfully identifies critical energy levels accurately in the vast majority of cases. These results provide preliminary confirmation that the program can reliably automate the crucial step of wavefunction identification in QCL design, laying a solid data foundation for subsequent QCL research utilizing machine learning methods[13]. Furthermore, the analysis of these slight inaccuracies offers valuable insights into areas where the program might benefit from further refinement. For example, challenges in distinguishing E_3' from extractor states or the less clear characteristics of E_6 suggest the need for enhanced discrimination capabilities in our automated process, possibly by improving the handling of level hybridization effects or introducing more refined criteria for distinguishing states with similar features.

Beyond its application in this study's mid-infrared QCLs (based on the InGaAs/AlInAs InP material system), this automated wavefunction identification method holds potential for adaptation to structurally similar shorter-wavelength QCLs (e.g., in the 3-5 µm range). However, applying it to terahertz QCLs, where carrier injection and photon emission mechanisms differ significantly, might pose greater challenges and necessitate more targeted adjustments and optimizations[13].

3.5 Generation of a Comprehensive Dataset for QCL Neural Network Training

To provide a robust foundation for training neural network models, this study undertook the generation of a large-scale, comprehensive dataset mapping the structural parameters of Quantum Cascade Lasers to their key performance metrics. The process began with creating numerous designs based on a nominal 8 μm-emitting QCL structure, each comprising 24 InGaAs/AlInAs layers lattice-matched to an InP substrate[13]. While a fixed electric field of 40 kV/cm was applied, the thickness of each layer was varied according to multiple random distribution functions. The material compositions were deliberately kept constant to maintain simplicity in the initial training data and reduce the dimensionality of the design space, thereby decreasing the training difficulty. This simplification avoids complexities that could arise from compositional changes, such as lattice strain or impractical alloy ratios[13]. Future research, however, can expand the dataset to include variations in material composition.

Each of these randomly generated structures was then processed through an automated computational pipeline. This workflow starts with a 3-band k·p solver to calculate the electronic band structure and its associated wavefunctions , followed by the aforementioned automated identification program that analyzes the results. This program, utilizing sophisticated clustering and probability formulas, autonomously identifies and labels all critical energy states, such as the upper and lower laser levels (E_4, E_3). This automation is highly efficient; for any given AR structure, the entire process of simulating the band diagram, automatically identifying the wavefunctions, and calculating the corresponding metrics takes only about 30 seconds in total[13].

Once the states were identified, a physics-based solver calculated a comprehensive set of performance metrics for each structure, including energy-level differences, electron lifetimes, and dipole matrix elements. A detailed description of the k·p solver, its graphical user interface for single-structure analysis, and the metric calculation process is provided in Appendix B.

To ensure the highest data quality for neural network training, this large initial dataset underwent a selective filtering process to remove entries adversely affected by physical phenomena such as energy-level hybridization or splitting, which can alter wavefunction distributions and reduce data reliability. This resulted in a refined dataset of approximately 300,000 valid entries. For subsequent model training, the focus was placed on four key metrics: the energy separations E_{43} and E_{54}, and the electron lifetimes τ_{43} and τ_{54}. This choice was deliberate and strategic. Firstly, the automated identification program identifies the associated energy states (E_3, E_4, and E_5) with exceptionally high accuracy—over 99% for E_4 and E_5, and 100% for E_3. Secondly, these four metrics are fundamentally linked to critical device performance: the energy differences (E_{43}, E_{54}) are primary determinants of the lasing wavelength and carrier leakage suppression , while the lifetimes (τ_{43}, τ_{54}) directly influence population inversion and carrier leakage mechanisms. A more detailed description of the specific filtering criteria is referred to Appendix C.

The entire workflow, from initial structure randomization to the final generation of the refined 300,000-entry dataset, represents a significant computational undertaking, with the initial generation phase taking approximately six days on four multi-core computers. The resulting high-quality dataset provides a solid and extensive foundation for training robust neural network models capable of accurately modeling and predicting QCL performance, thereby accelerating future design and optimization cycles.

3.6 Dataset Characteristics and Utility for Machine Learning

The automated generation workflow previously described culminates in a key outcome of this research: a large-scale, comprehensive dataset of Quantum Cascade Laser designs. The initial generation process produced over 430,000 unique active region structures . For specific applications, such as training generative models, this dataset underwent a selective filtering process to remove entries affected by physical phenomena like energy-level hybridization, resulting in a refined, high-quality training set of approximately 300,000

valid entries. The sheer scale of this dataset is one of its most critical features, as it provides the substantial volume of training samples necessary for developing robust, data-intensive machine learning models, such as deep neural networks, and ensuring they have strong generalization capabilities.

Each data point within this dataset provides a direct link between a QCL's structural parameters and its resultant electronic properties. A single entry is composed of the complete active region structure—defined by a sequence of 24 layer thicknesses —and its corresponding set of performance metrics calculated via the k·p solver and the automated wavefunction identification program. These metrics include crucial band structure information, primarily the energy separations between the key laser levels (E_{43} and E_{54}) and the associated electron lifetimes (τ_{43} and τ_{54}). This structured pairing of input (structure) and output (metrics) is precisely what is required to train supervised learning models.

The direct utility of this dataset is demonstrated by its foundational role in the machine learning applications explored in this research. It serves as the direct basis for the training and validation of both forward (predictive) and inverse (generative) neural networks. The forward networks are trained to take a QCL structure as input and accurately predict its performance metrics . Conversely, the inverse networks leverage this data to learn the more complex task of generating novel QCL structures that correspond to a set of desired, user-specified performance metrics. The successful implementation of these distinct models underscores the dataset's richness and versatility as a basis for multifaceted machine learning investigations in QCL design[12, 13].

Beyond its immediate use, the dataset itself represents a valuable scientific contribution. A primary bottleneck in the application of machine learning to complex scientific problems is often the scarcity of large, high-quality, and domain-specific training data . By creating this substantial QCL dataset and making it accessible to the broader research community, as indicated by the provision of structured data files (e.g., `parameters.mat`, `metric_E4E3.mat`, `metric_E5E4.mat`)[12, 13], this work provides an essential resource that can catalyze further research. This facilitates the replication of these findings and enables

other researchers to explore new models and design methodologies, thus fostering wider innovation in the field.

Chapter 4

Forward Prediction of QCL k.p Metrics Based on Neural Networks

4.1 Rationale and Implementation of Forward Modeling in the Design Cycle

In the design and optimization workflow of Quantum Cascade Lasers, the forward model plays an indispensable role. A forward model is defined as a process that predicts corresponding performance metrics or physical properties based on a given set of device physical structure parameters (in our study, the thickness sequence of the 24 quantum wells and barriers in the active region). In the context of this research, the core objective of the forward neural network model is to rapidly and accurately predict key k.p metrics based on the input QCL active region (AR) structural parameters. These metrics primarily include the energy level differences E_{43} (the difference between the upper-laser level E_4 and the lower-laser level E_3) and E_{54} (the difference between the next-higher-energy level E_5 and the upper-laser level E_4), which are crucial for determining the lasing wavelength and influencing carrier leakage, as well as the reciprocals of electron lifetimes (i.e., transition rates) $1/\tau_{43}$ and $1/\tau_{54}$[12, 13]. These k.p metrics directly impact the device's gain, efficiency, and thermal characteristics, serving as the core basis for design evaluation and

optimization. To train these networks, we used a large-scale dataset generated from k.p physical simulations, which was filtered to improve the model's predictive robustness.

The construction and validation of such a high-performance forward model carry multiple critical implications, with its core advantages being speed and compatibility with machine learning frameworks, which address the limitations of traditional physical simulation methods. Firstly, it enables the rapid evaluation of new designs. Once trained, a neural network's prediction time is extremely short. Our work has shown that the model can complete over 10,000 metric evaluations in less than a minute[13], a task that would take approximately a full day using conventional k.p physical simulations. This dramatic speed advantage allows designers to quickly screen and assess a vast number of potential design solutions, thereby greatly enhancing the efficiency and scope of design iterations. More importantly, this forward network serves as an indispensable key component in advanced inverse design frameworks. Traditional physical simulators like k.p solvers, due to their non-differentiable algorithmic nature, are difficult to integrate directly into neural network training pipelines that rely on gradient descent and backpropagation. A trained forward neural network model, being inherently fully differentiable, can be seamlessly embedded into more complex machine learning systems (like a tandem inverse network) to act as an advanced evaluator or a differentiable loss function. Specifically, the inverse network generates a candidate structure, the forward network rapidly predicts its performance, and the error between the prediction and the target value can then be used to guide and optimize the inverse network's parameters via gradient-based backpropagation.

4.2 Forward Prediction Network: Architecture and Training Strategy

To enable rapid and accurate prediction of the key k.p metrics for QCLs, this study developed independent forward neural network models for each metric: the energy level differences E_{43} and E_{54}, and the transition rates $1/\tau_{43}$ and $1/\tau_{54}$, which are related to the lifetimes. This strategy of using independent networks, rather than a single multi-

output model, was chosen to circumvent potential training difficulties. If a combined loss function were used to simultaneously optimize all four parameters, the losses from the different metrics could balance or counteract each other during backpropagation, impeding the network's overall convergence and leading to suboptimal predictive performance for each individual metric. Training separate models ensures that each network can focus exclusively on minimizing the error for its specific target. Additionally, it was found that predicting the reciprocal of the lifetimes (i.e., the transition rates) resulted in greater predictive accuracy than predicting the lifetimes directly.

These forward networks follow similar architectural principles, as schematically depicted in Figure 4.1. The input layer for all networks is designed to accept a 24-dimensional vector, where each element represents the physical thickness of one of the 24 semiconductor layers within a single cascade period of the QCL active region. The core of the network is a deep, fully connected architecture. For the networks predicting the energy differences E_{43} and E_{54}, this comprises three hidden layers with 1000, 1000, and 100 neurons, respectively. These hidden layers act as hierarchical feature extractors, learning progressively more abstract and complex representations from the raw layer thickness data. To effectively model the highly non-linear, quantum-mechanical relationships between layer thicknesses and electronic energy levels, either Sigmoid or Softsign was used as the activation function in the hidden layers. The introduction of this non-linearity is critical for the model's ability to learn complex mappings. Furthermore, to prevent the model from overfitting—that is, "memorizing" the training data and losing its ability to predict new data—Dropout layers were introduced. During training, Dropout layers randomly set the output of a fraction of neurons to zero, which enhances the model's generalization capabilities. A similar design philosophy and training process was used for the networks that predict the lifetime-related parameters.

A meticulous configuration was adopted for the training strategy and hyperparameters to ensure optimal model performance. We employed the ADAM optimizer for its efficiency and adaptive learning rate capabilities when handling large-scale datasets and parameters.

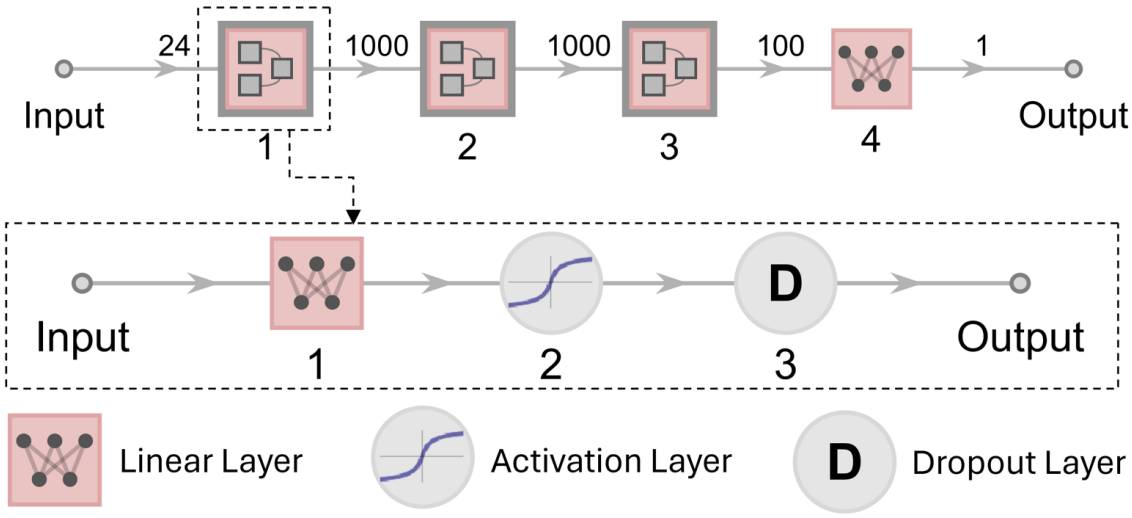


Figure 4.1: The structure of the forward network.

For the loss function, we selected Mean Absolute Error (MAE), which measures the average magnitude of the errors between the model's predictions and the values calculated by the ground-truth k.p solver. The training hyperparameters were also carefully configured: the number of epochs was set to a maximum of 5,000 to 10,000, with an early stopping mechanism triggered based on performance on a validation set (1/10th of the total dataset) to prevent overfitting. The batch size for each training iteration was set to 10,000 to balance stable gradient estimation with training efficiency. The learning rate was fixed at 0.001 to control the step size of the model's weight updates.

4.3 Performance Evaluation and Limitation Analysis of the Forward Prediction Network

To comprehensively evaluate the performance of the trained forward neural network models, they were validated on an independent test set comprising approximately 10,000 data points that were not used in training or validation. The evaluation results indicate that the models exhibit high accuracy in predicting all four key k.p metrics. For the energy level differences E_{43} and E_{54}, the networks achieved high coefficients of determination (R^2), with values of 0.895 and 0.942, respectively. Similarly, the forward networks for pre-

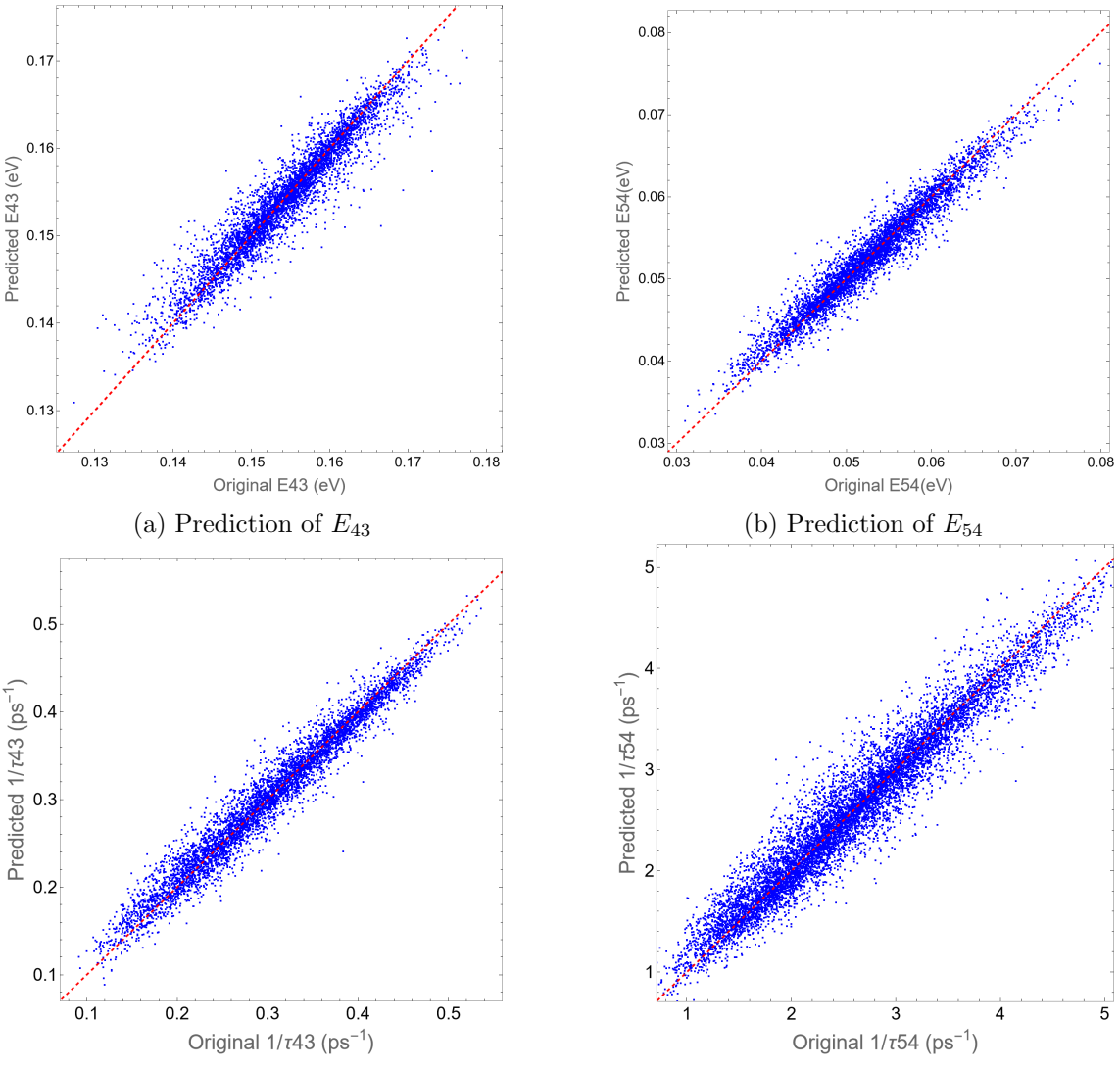
dicting transition rates ($1/\tau_{43}$ and $1/\tau_{54}$) also demonstrated robust performance, with R^2 values reaching 0.957 and 0.937, respectively. For each metric, scatter plots of predicted versus actual values (see, for example, Figure 4.2) visually confirmed the strong agreement between them.

Despite the high overall accuracy achieved, further improving the model's precision remains a challenge, stemming primarily from the inherent limitations of the training data itself. A major reason is that the k.p solver, which serves as the source of the data, can produce energy level hybridization effects for certain structures, thereby introducing unavoidable noise and errors into the training dataset. Although the training dataset was selectively filtered to remove many entries affected by energy-level hybridization or splitting, some subtle, hard-to-distinguish cases of hybridization may still remain. This residual noise can increase the model's prediction error, with lifetime predictions being particularly sensitive to this phenomenon.

The predictive errors of the forward network, even if small, directly impact and can even be amplified in the inverse design process, representing one of the main limitations of the overall methodology. This is because, in the tandem architecture, the forward network serves as the evaluator or "ground truth" during the inverse network's training process. If the forward network has its own prediction bias, the inverse network will learn to generate structures that satisfy this "flawed evaluator" rather than structures that meet the target parameters in true physical reality. Therefore, the accuracy of the forward network directly determines the reliability and fidelity of the final inverse design results.

4.4 The Critical Role of the Forward Network in the Inverse Design Framework

Once established as a fast and accurate predictive tool, the forward model plays an indispensable and critical role within our proposed inverse design framework for QCLs. This framework utilizes a tandem network architecture, where the core idea is to employ the pre-trained forward network as an advanced, physics-informed evaluator to guide the train-



(a) Prediction of E_{43}

(b) Prediction of E_{54}

(c) Prediction of τ_{43}

(d) Prediction of τ_{54}

Figure 4.2: Scatter plots demonstrating the predictive accuracy of forward networks for the k·p metrics

ing of a generative inverse network. This approach is not only efficient but also elegantly overcomes the inherent non-uniqueness problem in mapping from performance metrics to physical structures, where multiple different structures can yield the same performance metrics.

The most central function of the forward network is to serve as a differentiable loss function component within the tandem system. During training, the inverse network generates a candidate QCL structure based on a set of target k.p metrics. This structure is then fed into the pre-trained forward network, which rapidly predicts its corresponding k.p metrics. The difference between the forward network's prediction and the original target values (e.g., the mean absolute error) then constitutes the primary loss signal. Through backpropagation, the gradient of this loss is used to update the weights of the inverse network, guiding it to produce better structures. To ensure a stable training process, the weights of the forward network are kept fixed (i.e., "frozen") throughout the training of the tandem system, providing a consistent optimization target for the inverse network.

The implementation of this architecture hinges on the end-to-end differentiability provided by the forward network. In order for gradient-based optimization algorithms to train the tandem system, the entire computational path—from the structure generation by the inverse network to the final loss calculation—must be differentiable. As previously established, traditional physical simulators like k.p solvers are non-differentiable and computationally intensive, making them incompatible with the backpropagation algorithms central to deep learning. This has been a major bottleneck for conventional methods. The forward neural network, therefore, effectively acts as a "differentiable proxy" for the physical solver. Through its rapid prediction capabilities and inherent differentiability, it successfully bridges the gap between generative models and gradient-based optimization, serving as the key enabling technology for the automated QCL inverse design presented in this study.

Chapter 5

Inverse Design of QCL Active Region Based on Tandem Neural Networks

5.1 The Challenge of QCL Inverse Design: The Non-Uniqueness Problem

The core objective of this chapter is to solve the inverse design problem for Quantum Cascade Lasers, which involves determining the physical QCL structure (specifically, the layer thicknesses of the active region) that can achieve a desired set of performance metrics (in this study, the k.p parameters). The design space of QCLs is exceptionally vast and complex, comprising numerous tunable parameters whose relationships with device performance are highly non-linear and coupled. While the traditional forward design approach (from structure to performance) can analyze a specific design, it is inefficient for searching the vast design space to find an optimal structure that meets specific performance requirements. Inverse design aims to directly answer the fundamental question, "What structure is needed to achieve a certain performance?", and is thus of great significance for accelerating the development and optimization of new devices[12].

A central challenge in the inverse design of QCLs, and photonic devices more broadly, is the so-called "non-uniqueness" or "one-to-many" problem. This means that multiple, distinct physical QCL active region structures (e.g., different combinations of layer thicknesses) can produce an identical or very similar set of k.p parameters[12]. This non-uniqueness is not merely a mathematical coincidence but is rooted in deep physical principles. Quantum mechanics allows for different combinations of potential wells and barriers (i.e., different layer thickness sequences) to produce similar electron wavefunction distributions and energy level arrangements. For example, through coordinated adjustments to the thicknesses of certain layers (e.g., thinning one layer while thickening an adjacent one), it is possible to maintain a key energy difference (like E_{43}) while the effects on other parameters (such as wavefunction overlap or effective lifetimes) compensate for each other, ultimately leading to similar macroscopic k.p metrics. This physical degeneracy, or approximate degeneracy, makes the mapping from performance to structure inherently ill-posed, meaning a unique inverse function does not exist.

This inherent non-uniqueness poses a fundamental difficulty for training an inverse model directly using standard machine learning methods. If a conventional loss function, such as Mean Squared Error (MSE) or Mean Absolute Error (MAE), is used to train a standard inverse neural network (i.e., one that takes performance metrics as input and outputs structural parameters), the network will struggle during the learning process. When the training data contains multiple different structures corresponding to the same set of performance metrics, the loss function will drive the network to output an "average" of these different structures. However, this "averaged" set of structural parameters is often physically meaningless or, when validated with a forward physical simulation, fails to accurately reproduce the target performance metrics it was intended to achieve. Therefore, a directly trained inverse network is often ineffective, necessitating the adoption of special strategies that can navigate this non-uniqueness challenge[12].

5.2 Tandem Neural Network Model

To fundamentally circumvent this issue, we constructed and implemented a Tandem Network architecture. The core idea of this architecture is to cascade the target inverse network with a pre-trained forward network, whose parameters are kept fixed during the tandem training phase[12, 17]. In this design, the forward network functions as a high-order, differentiable loss function. We opted not to use traditional k·p solvers in place of the forward network, primarily because these physics-based simulation tools often require explicit wavefunction identification, are difficult to accelerate efficiently on GPUs, and most critically, they cannot be seamlessly integrated with deep learning frameworks based on gradient descent and backpropagation.

Thus, the role of the forward network in this architecture is not to participate in parameter optimization but to act as a precise and efficient "evaluator". Its workflow is as follows: the inverse network generates a candidate QCL structure based on a set of input k·p metrics. This structure is immediately passed as input to the fixed forward network, which rapidly calculates the corresponding k·p metrics for the generated structure. By comparing the metrics calculated by the forward network with the original input target metrics, the system obtains a loss value that accurately reflects the quality of the generation. This loss value is then backpropagated to exclusively guide the parameter updates of the inverse network, steering it towards generating more accurate structures. This "generate-evaluate-feedback" closed-loop mechanism ensures that the finally generated structures can accurately reproduce the target physical properties, thus elegantly and effectively solving the non-uniqueness problem.

5.3 In-depth Analysis of Network Architecture and Loss Function Design

The overall architecture of the tandem network and the detailed structure of the inverse network are depicted in Figure 5.1. This section provides an in-depth analysis of the

inverse network's design details and its specific mechanism for collaborating with the forward network.

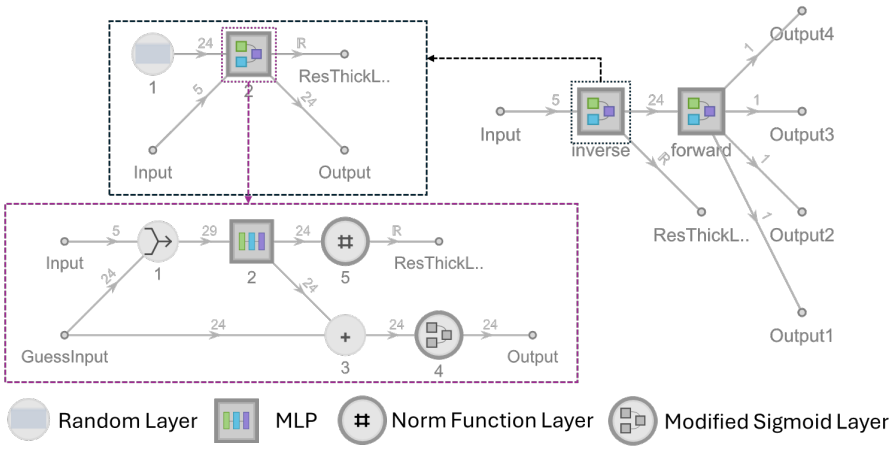


Figure 5.1: Tandem network architecture and a detailed diagram of the inverse network.

As the core of the generative model, the inverse network is designed as a fully connected multilayer perceptron (MLP), with its internal components and data flow meticulously engineered for efficient structure generation and optimization. The inverse network features two functionally distinct input ports: "Input" and "GuessInput".

- **"Input" Port**: This port accepts a one-dimensional vector of length four, which encapsulates the core performance metrics of the target QCL design. These are the energy difference between the upper laser level (E_4) and the lower laser level (E_3), denoted as E_{43}; the energy difference between the next-higher energy level (E_5) and the upper laser level (E_4), denoted as E_{54}; and their corresponding carrier transition rates (reciprocals of lifetimes), $1/\tau_{43}$ and $1/\tau_{54}$.

- **"GuessInput" Port**: This port accepts a one-dimensional vector of length 24, representing the thicknesses of the quantum wells and barriers in an initial, guessed QCL active region. This input can be a randomly generated structure, "speculative estimates," or even an existing design that a user wishes to optimize.

Inside the network, the vectors from the "Input" and "GuessInput" ports are concatenated and then jointly fed into the core MLP module. This MLP consists of two linear layers with neuron counts (widths) of 100 and 50, respectively, with each linear layer being followed by a scaled exponential linear unit (SELU) as the activation function. The final output of the network is also a 24-length vector, but it does not directly represent the final layer thicknesses. Instead, it represents the modification or deviation vector that needs to be applied to the initial structure provided via "GuessInput". Therefore, the final QCL structure is obtained by adding this output modification vector to the input "GuessInput" vector. This design simplifies the learning task by making the network focus on "how to adjust an existing structure based on target metrics" rather than generating from scratch.

To enhance the model's practicality, design generalization, and controllability, several crucial mechanisms were introduced:

1. **Random Layer**: This layer, connected to the "GuessInput" port, has the core function of introducing controllable "variability" into the process. During practical application, activating this layer enables the model to generate multiple different QCL structures for the same set of input k·p metrics, which vastly expands the design possibilities. It allows designers to choose the most physically sound, manufacturable, or error-tolerant solution from a pool of candidates. When precise optimization of a known structure is required, this layer can be bypassed, and the manually designed structure can be fed directly into the "GuessInput" port.

2. **ResThickLoss Term**: The physical meaning of this loss term is to quantify and control the magnitude of deviation between the final output structure and the initial "GuessInput" structure. By carefully tuning its weight in the total loss function, one can flexibly balance the model's "creativity" against its "fidelity" to the initial guess. A higher weight on 'ResThickLoss' will favor outputs that are closer to the initial random guess, whereas a lower weight permits more substantial modifications, yielding a structure that may differ significantly from the guess but aligns more perfectly with the target k·p metrics.

During training, the calculation of the total loss function is the core mechanism of the tandem network. The structure generated by the inverse network is fed into the parameter-fixed forward network, which evaluates each k·p metric independently. The total loss function is composed of two parts: the Mean Absolute Error (MAE) between the metrics calculated by the forward network and the target metrics, and the aforementioned 'ResThickLoss'. We chose MAE because it effectively measures the average magnitude of error without considering direction, making it particularly suitable for this application. Due to the different scales and physical importance of the four k·p metrics and 'ResThickLoss', they were assigned different weights to normalize their contributions and balance the optimization process.

5.4 Detailed Configuration of Training Parameters

To ensure a stable, efficient, and high-performing training process, we meticulously configured all aspects of the training. The ADAM optimizer was chosen, a decision based on its robust performance in handling sparse gradients and its adaptive learning rate capabilities. The entire training dataset contains just under 300,000 valid samples. The training set consisted of 290,000 samples, with an additional 7,000 samples reserved as the test set. During training, 10% of the training data was allocated as a validation set to monitor the model's generalization ability. The batch size was set to 5000, and a total of 60,000 epochs were run to ensure the model could fully learn the complex mappings from the large-scale data. The learning rate was fixed at 0.001 to achieve a steady adjustment of weights without overshooting the minimal loss points.

A critically important setting during the entire tandem network training phase was that only the parameters of the inverse network were updated and optimized, while all parameters of the forward network remained frozen and stable. The purpose of this strategy was to ensure that the forward network always acted as a reliable, consistent, and impartial "judge" or "benchmark" for evaluating the quality of the QCL structures generated by the inverse network at various training stages. The loss convergence during the

training process is illustrated in Figure 5.2.

Figure 5.2: Training loss and validation loss curves during the tandem network training process.

The weights for the individual components of the total loss function were determined after extensive experimentation to achieve the best performance balance, and were set as follows:

- E_{43} Loss Weight: 1000
- E_{54} Loss Weight: 1000
- $1/\tau_{43}$ Loss Weight: 100
- $1/\tau_{54}$ Loss Weight: 10
- ResThickLoss Weight: 1

5.5 Quantitative Evaluation of Inverse Network Performance

Upon completion of the tandem network training, the fully trained inverse network was extracted and subjected to an independent, comprehensive, and quantitative evaluation

of its QCL structure generation capabilities. This evaluation employed a rigorous, multi-step methodology. We selected the first 1,000 data points from the test set, which had never been used for training or validation. For each input k·p metric combination, we ran the inverse network independently 50 times to generate 50 different candidate QCL structures. This step fully leverages the random layer's capability to broadly explore the potential design solution space and increase the probability of finding an optimal solution.

Subsequently, the generated structural parameters (the 24 layer thicknesses) were rounded to the nearest whole number to meet the input requirements of the k·p solver for physical validation. It must be emphasized that this rounding operation must be performed after structure generation. Our attempts to perform rounding directly within the training loop resulted in the tandem network being untrainable with divergent loss values, leading to its abandonment. Because minor changes like rounding can significantly affect sensitive parameters like carrier lifetimes, the strategy of iterating 50 times and selecting the best is particularly necessary. These 50 rounded candidate structures were then individually fed into the forward network for rapid evaluation, and we ultimately selected the structure that yielded the smallest total relative error across all k·p metrics as the best output for that input set. This selection strategy was also adopted to mitigate error propagation, based on our observation that smaller predicted errors from the forward network often correlate with smaller actual discrepancies. The entire evaluation workflow, from generating 50 structures in parallel to their evaluation and error comparison, was fully automated, requiring only about 1 second per input metric set.

The evaluation results are presented in Figure 5.3. The four scatter plots in Figure 5.3 clearly show a high degree of linear agreement between the physical metrics calculated by the k·p solver for the best-generated structures (Y-axis) and the original target metrics (X-axis). The coefficient of determination (R^2) scores for all four metrics were excellent: the R^2 for E_{43} was 0.9153, for E_{54} was 0.9701, for $1/\tau_{43}$ was 0.9568, and for $1/\tau_{54}$ was 0.9175. These high R^2 values provide strong evidence of the inverse network's outstanding generalization capability, demonstrating its ability to accurately generate physically valid

QCL active region structures based on given k·p metrics.

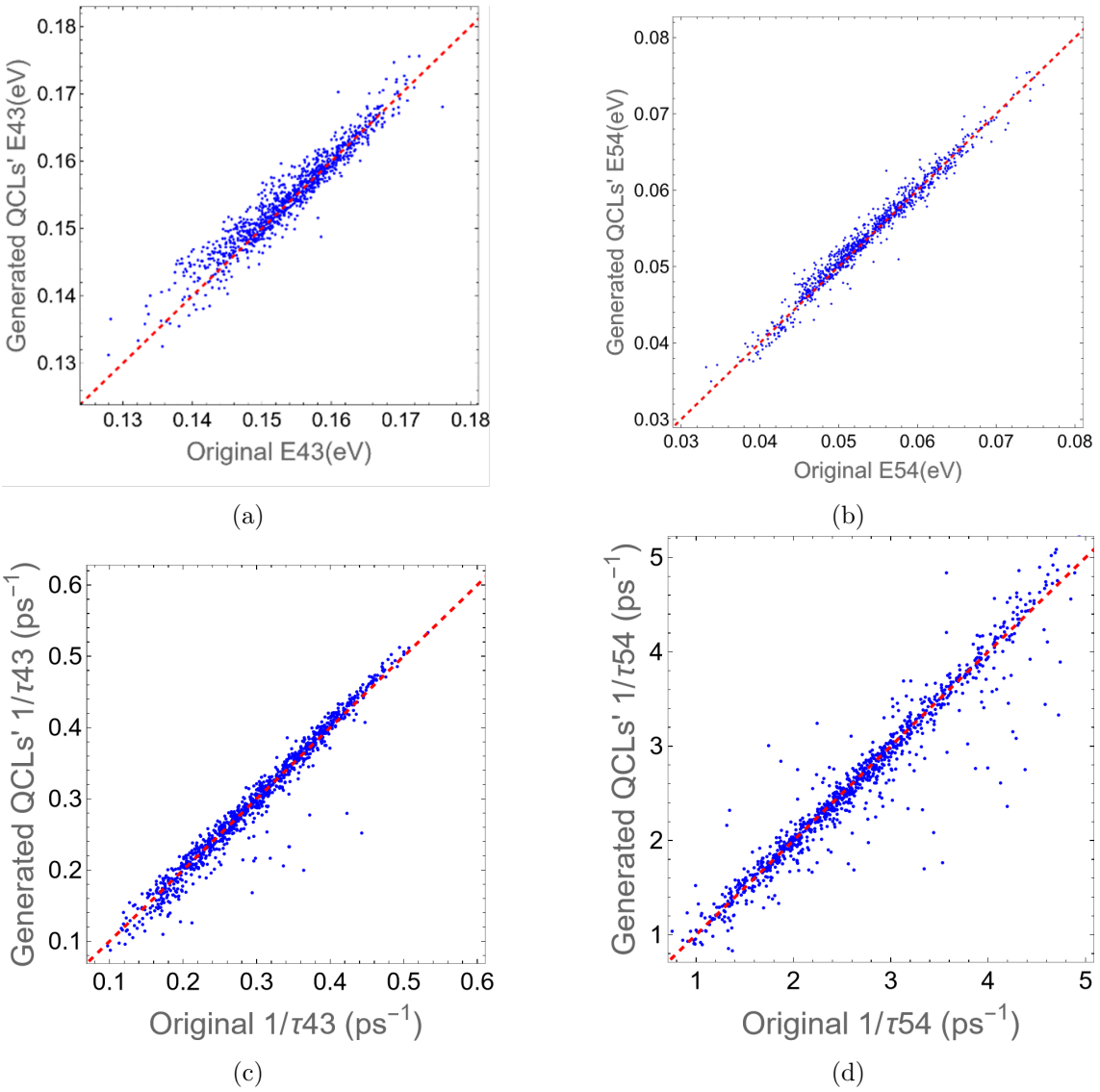


Figure 5.3: The testing results of the generated QCLs by the inverse network for the k·p metrics: (a) E_{43}, (b) E_{54}, (c) $1/\tau_{43}$, and (d) $1/\tau_{54}$.

5.6 QCL Structure Optimization and Quantitative Analysis via the Inverse Network

To validate the potential and precision of the inverse network in practical design optimization scenarios, we conducted a forward-looking test that is quantitatively assessable. This

test was aimed at optimizing QCL performance by actively tuning key k·p metrics, with the explicit goal of suppressing carrier leakage and thus reducing the threshold current density.

The starting point of the test was to establish a design baseline. We selected a set of k·p metrics representing median values from the dataset: E_{43} = 0.155 eV, E_{54} = 0.055 eV, $1/\tau_{43}$ = 0.32 ps^{-1}, and $1/\tau_{54}$ = 2.8 ps^{-1}. Using the 50-iteration selection process described earlier, we first tasked the network with generating the corresponding baseline QCL structure, whose conduction band diagram is shown in Figure 5.4(a). The quantitative evaluation of this baseline generation task is recorded in the first section ("Origin") of Table 5.1. The data show that the network accurately reproduced the baseline design; for instance, the error for E_{43} was only 0.78%. The largest error in this test occurred for the highly sensitive $1/\tau_{54}$ metric at 6.14%, which is still within an acceptable range and demonstrates the network's high-fidelity reproduction capability.

The core of the test lies in the four independent modifications made to the baseline metrics, each with a clear physical objective, to verify if the network can generate structures that meet these more demanding design requirements.

- First, we increased the target E_{43} to 0.165 eV. As shown in the band diagram in Figure 5.4(b), to solely increase E_{43} while keeping other metrics constant, the network intelligently shifted the wavefunction of the E_3 level to the left. This enhances the wavefunction overlap between E_3 and E_4. The data in Table 5.1 provide solid quantitative support for this physical picture: the network successfully generated a structure with a test value of 0.1658 eV, a relative error of only 0.49% from the target. At the same time, the table quantifies the minor "crosstalk" effects, such as a 2.72% error in E_{54}, demonstrating that the network can satisfy the primary optimization goal while keeping unintended impacts on other performance parameters within a small range.

- Second, to suppress carrier leakage, we raised the target E_{54} to 0.065 eV. The physical

motivation for this change stems from the leakage current expression[2]:

$$J_{\text{leak},45} \propto \frac{1}{\tau_{54}} \exp\left(-\frac{E_{54}}{kT_{e4}}\right)$$

where a larger E_{54} can exponentially reduce leakage. Figure 5.4(c) shows that the wavefunction of the E_5 level was significantly shifted to the right to achieve this goal. From Table 5.1, we can see that the structure generated by the network corresponded to an E_{54} of 0.0635 eV, with an error of 2.30% from the target, successfully realizing a key design feature aimed at lowering the device's threshold current.

- Next, we increased the target transition rate $1/\tau_{43}$ to 0.4 ps^{-1}, with the expectation of obtaining higher gain. Figure 5.4(d) again illustrates the fine-tuning of the wavefunctions to augment overlap between E_3 and E_4. The quantitative results, as shown in Table 5.1, reveal that the test result provided by the network (0.4028 ps^{-1}) had an error of only 0.69% from the target. In this case, the crosstalk errors on other metrics were very low, demonstrating excellent decoupling performance.

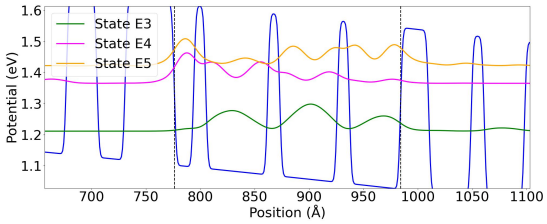
- Finally, also with the aim of reducing leakage, we significantly decreased the target $1/\tau_{54}$ from 2.8 ps^{-1} to 2.0 ps^{-1}. Figure 5.4(e) shows the wavefunction of E_5 moving to the right to reduce its overlap with E_4. The values in Table 5.1 indicate that the network's generated result was 2.0519 ps^{-1}, an error of 2.59%. This modification induced a crosstalk error in E_{54} of 3.57

By synthetically analyzing the physical processes shown in the band diagrams (Figure 5.4) and the detailed data provided in the table (Table 5.1), we can conclude that the inverse network is not just a powerful reproduction tool, but also an efficient and precise design optimization engine. The outcomes indicate that most k·p metrics exhibited errors around 1%, with some exceptions that were slightly higher but still acceptable. This variation is inevitable, as it stems from the inherent physical interdependence among the different k·p metrics, where not all arbitrary combinations of metrics are physically plausible. Despite this intrinsic limitation, these results clearly demonstrate that the inverse

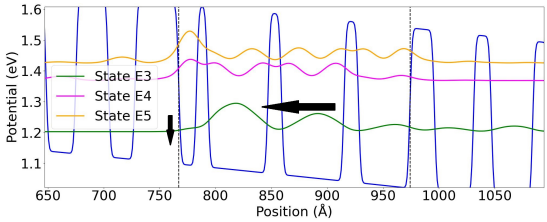
Table 5.1: Test results of the modified k·p metrics

	E_{43} **(eV)**	E_{54} **(eV)**	$1/\tau_{43}$ **(ps^{-1})**	$1/\tau_{54}$ **(ps^{-1})**
Origin	0.155	0.055	0.32	2.8
Test	0.1538	0.0558	0.3152	2.6281
Relative error	0.78%	1.47%	1.51%	6.14%
Modified E_{43}	**0.165**	0.055	0.32	2.8
Test	0.1658	0.0565	0.3169	2.7472
Relative error	0.49%	2.72%	0.98%	1.88%
Modified E_{54}	0.155	**0.065**	0.32	2.8
Test	0.1550	0.0635	0.3253	2.7488
Relative error	0.02%	2.30%	1.67%	1.83%
Modified $1/\tau_{43}$	0.155	0.055	**0.4**	2.8
Test	0.1542	0.0552	0.4028	2.8058
Relative error	0.49%	0.29%	0.69%	0.21%
Modified $1/\tau_{54}$	0.155	0.055	0.32	**2**
Test	0.1549	0.0570	0.3200	2.0519
Relative error	0.04%	3.57%	0.00%	2.59%

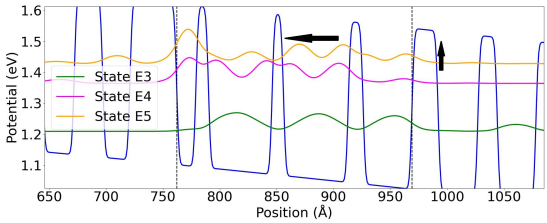
network developed in this study can reliably generate novel QCL structures with corresponding physical properties, according to predetermined, superior performance metrics set by the designer, thereby greatly accelerating the optimization design process for advanced semiconductor devices.



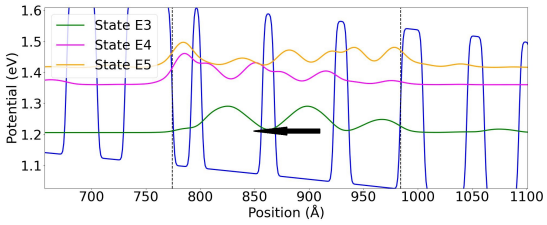
(a) The band diagram of the QCL for original k·p metrics



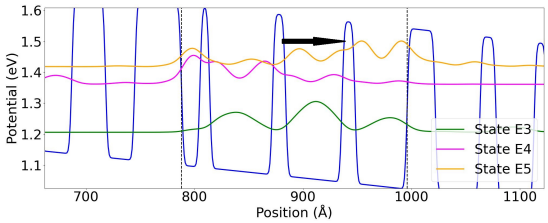
(b) The band diagram of the QCL for modified E_{43}



(c) The band diagram of the QCL for modified E_{54}



(d) The band diagram of the QCL for modified $1/\tau_{43}$



(e) The band diagram of the QCL for modified $1/\tau_{54}$

Figure 5.4: The testing results of the generated QCLs for the modified k·p metrics E_{43}, E_{54}, $1/\tau_{43}$ and $1/\tau_{54}$. Arrows in the diagram indicate the primary changes in the energy states and their directions within the band diagram: vertical arrows show the direction of energy level changes, and horizontal arrows depict the changes in the wavefunction distribution.

Chapter 6

Conclusion

6.1 Summary of Research Objectives and Achievements

The core objective of this dissertation was to develop an automated inverse design framework for the active region of Quantum Cascade Lasers based on machine learning. The primary goal was to construct a neural network model capable of predicting the physical structure parameters of a QCL active region based on a set of desired performance metrics, specifically thek·p metrics.

To achieve this overarching objective, this research has yielded several key achievements:

1. **Development and Validation of an Automated Wavefunction Identification Program:** A program was successfully developed to automatically and accurately identify key lasing energy levels from QCL band diagrams, achieving an accuracy of over 95%. This was accomplished through a multi-stage, rule-based unsupervised learning method, which employs sophisticated K-Means clustering, intensity and positional weighting, and probability calculations to distinguish between different energy states. This program served as the cornerstone for subsequent machine learning model training by enabling efficient data generation.

2. **Construction of a Large-Scale, Filtered QCL Design Dataset:** By leveraging

the automated wavefunction identification program andk·p physical simulations, an initial dataset was generated. This dataset was then carefully filtered to remove entries affected by energy level hybridization or splitting, resulting in a refined training dataset of just under 300,000 QCL active region structures. The dataset contains structures composed of InGaAs and AlInAs layers lattice-matched to InP substrates and their correspondingk·p metrics, such as the energy level differences E_{43}, E_{54}, and the lifetimes τ_{43} and τ_{54}.

3. **Establishment of a High-Accuracy Forward Neural Network Model:** A forward neural network was trained to accurately predictk·p metrics from a given QCL structure. The model's predictions showed excellent agreement with simulation results, achieving coefficients of determination (R^2) of 0.895 for E_{43} and 0.942 for E_{54}. Furthermore, dedicated forward networks for lifetime prediction also demonstrated robust performance, achieving R^2 scores of 0.9573 for τ_{43} and 0.9368 for τ_{54}, respectively. It was found that using the reciprocals of the lifetimes (i.e., transition rates) yielded greater predictive accuracy.

4. **Implementation of a Tandem Network-Based QCL Inverse Design Framework:** A novel tandem neural network architecture was proposed and implemented to generate QCL active region structures from targetk·p metrics. By integrating a generative inverse network with a fixed, pre-trained forward network as an evaluator, this framework effectively addresses the non-uniqueness problem inherent in inverse design. The model was trained using an ADAM optimizer and a Mean Absolute Error loss function, with carefully assigned weights to balance the contributions of different metrics. The resulting inverse network demonstrated high accuracy in reproducing the target metrics, with R^2 values exceeding 0.91 for E_{43}, E_{54}, $1/\tau_{43}$, and $1/\tau_{54}$.

6.2 Main Findings and Their Significance

The principal findings of this research and their scientific significance are summarized as follows:

This study proves that a meticulously designed algorithm, combining physical criteria (such as positional intensity) and clustering techniques, can achieve automated, high-accuracy identification of crucial wavefunctions in complex QCL band diagrams. This not only overcomes the bottleneck of traditional, time-consuming manual analysis but, more importantly, enables the generation of the large-scale, high-quality datasets required for data-driven machine learning approaches.

The results demonstrate that standard forward neural networks can effectively learn the complex, non-linear mapping between QCL structural parameters (layer thicknesses) and their corek·p metrics. This provides a foundation for replacing computationally expensive physics simulations with rapid neural network models within the design workflow.

The proposed tandem inverse network architecture, by employing a forward network as a differentiable evaluator and incorporating a random layer with an initial guess mechanism, successfully navigates the non-uniqueness challenge of mapping performance to structure. This provides a powerful tool for exploring the vast design space and discovering novel QCL structures that meet performance targets.

Through case studies, such as adjusting targetk·p metrics to reduce carrier leakage, this work demonstrates that the developed inverse design framework can not only reproduce known relationships but also generate optimized structures for new, targeted performance goals.

This research systematically showcases the immense potential of machine learning, particularly deep learning, in the field of QCL design. It heralds a paradigm shift from traditional, "trial-and-error" design methodologies towards a more automated, intelligent, and data-driven approach. This shift is poised to significantly accelerate the research and development of novel, high-performance QCL devices and may lead to innovative designs that transcend human intuition.

6.3 Limitations of the Current Study

Despite the significant progress made, this research has several limitations that should be addressed in future work:

1. **Constraints Imposed by Forward Model Accuracy:** The inverse network's performance is inherently limited by the predictive accuracy of the forward network. A key challenge is the quality of the training data, as a filtering process was necessary to reduce the dataset to under 300,000 entries by removing structures affected by energy level hybridization or splitting. These effects, which can stem from physical properties or computational inaccuracies in thek·p solver, were found to significantly impair the predictability of lifetimes. The noise from these anomalies is inevitably propagated to the inverse network, constraining its final accuracy.

2. **Indirect Inverse Design Based onk·p Metrics:** The current framework uses intermediatek·p metrics as the optimization target, rather than the final, device-level macroscopic performance parameters that are of ultimate interest to users (e.g., lasing wavelength, output power, threshold current). The translation fromk·p metrics to full device performance is itself a complex process influenced by many other physical factors (e.g., optical modes, waveguide losses, thermal management) not captured by the model.

3. **Limited Generalizability of the Dataset and Model:** The dataset, while large, is specific. It was generated for the InGaAs/AlInAs on InP material system under a fixed applied electric field. Consequently, the generalizability of models trained on this dataset to other material systems, different operating fields, or broader structural parameter ranges may be limited, potentially requiring retraining or extensive transfer learning.

4. **Insufficient Consideration of Manufacturability Constraints:** While layer thicknesses in the initial dataset generation were empirically constrained and final

outputs were rounded to integers, the neural network does not explicitly or systematically incorporate practical manufacturing constraints (e.g., minimum controllable layer thickness, interface quality, material growth uniformity). The resulting "simulation-to-reality gap" means that a theoretically optimal design may be difficult to fabricate accurately or may be highly sensitive to process variations.

6.4 Future Research and Outlook

Based on the achievements and limitations of this study, several promising directions for future research in ML-based QCL inverse design are envisioned:

1. **Direct Inverse Design for Device-Level Performance Metrics:** Future work should aim to develop models that directly take user-desired macroscopic performance parameters (e.g., wavelength, power, efficiency) as input to generate a complete physical QCL structure. This could be approached through multi-stage ML frameworks or by developing complex end-to-end models that learn the entire mapping.

2. **Enhancing Model Robustness and Generalizability:** To improve model performance, future efforts should focus on refining the automated state identification program to better handle complex cases of level hybridization, which was a primary reason for data filtering. Additionally, expanding the dataset to cover a wider range of materials and operating conditions, and potentially improving the accuracy of the underlying physics simulators, would enhance model robustness.

3. **Exploring Advanced ML Architectures and Physics-Informed Learning:** The applicability of other advanced generative models, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models, should be investigated for QCL inverse design. Furthermore, integrating known physical laws and constraints directly into the neural network's architecture or loss

function (Physics-Informed Machine Learning, PIML) could guide the model to find more physically plausible solutions and reduce its reliance on massive datasets.

4. **Closing the Loop with Experimental Validation:** A critical future direction is to establish a closed-loop design-fabricate-test cycle. Representative designs generated by the inverse network should be fabricated and characterized experimentally. The experimental results must then be fed back to retrain and refine both the machine learning models and the underlying physical simulators. This iterative process is essential for bridging the simulation-to-reality gap and producing practical, high-performance devices.

5. **Integrating Manufacturability and Multi-Objective Optimization:** The design process should explicitly incorporate manufacturing tolerances to find solutions that are not only theoretically optimal but also robust to fabrication errors. Additionally, as QCL design often involves trade-offs between competing metrics (e.g., high power vs. low threshold), future research should explore ML-based multi-objective optimization methods to generate a portfolio of Pareto-optimal designs for designers to choose from.

Appendix A

Automated State Identification for Quantum Cascade Lasers

This appendix details the methodology of the automated state identification program developed for Quantum Cascade Laser (QCL) design, employing a rule-based unsupervised learning method to select pertinent wavefunctions critical for the laser's functioning. The procedure is divided into several stages, from preliminary selection to the final identification of injector states, described below.

A.1 Preliminary Selection

The preliminary selection stage is pivotal in the QCL design process, aimed at identifying essential energy states within the QCL's active region. This region's boundaries are defined early in the structure generation phase, ensuring a focused analysis on the most relevant part of the QCL structure.

Each wavefunction within the active region undergoes evaluation based on its energy level and intensity. Wavefunctions are considered for further analysis only if their energy falls within the potential energy range of the active region and their intensity surpasses a threshold of 0.1, deemed sufficient for identifying lasing and scattering states.

To refine the selection, wavefunctions must meet additional criteria regarding their

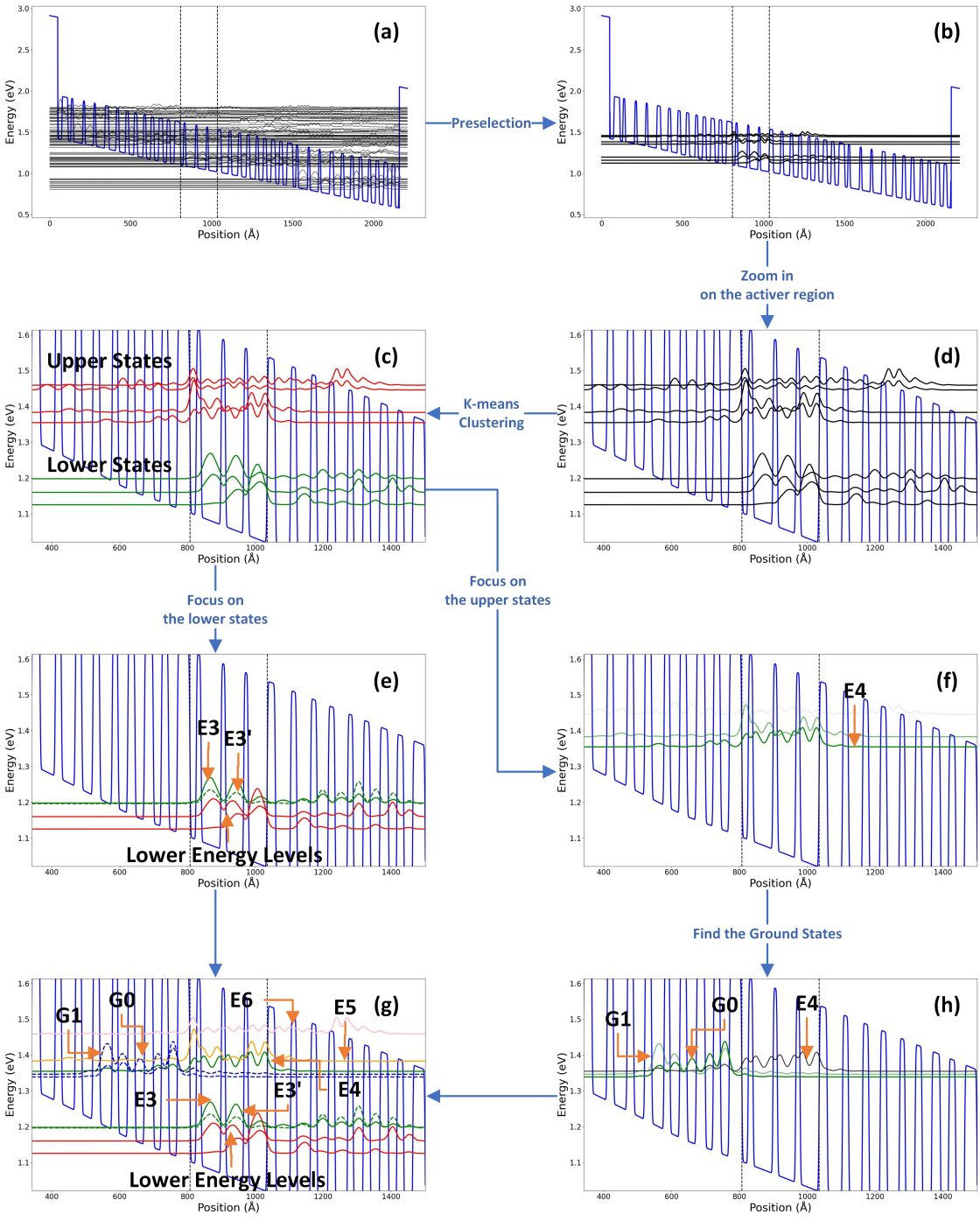


Figure A.1: An example of the progress of automated wavefunction identification.

intensity distribution relative to the injector and extractor regions. Specifically:

Wavefunctions are disqualified if their active region intensity is less than a third of that in the injector region, or less than half of the extractor region's intensity, to account for the partial extension of states like E4 into the injector region and hybridization issues with states such as E5 and E6.

A wavefunction's eligibility is further scrutinized based on its intensity's center of gravity within the active and adjacent regions, disqualifying those likely belonging to a previous or subsequent stage.

This meticulous selection process effectively isolates wavefunctions that significantly influence the QCL's performance, as illustrated in Fig. A.1b.

A.2 Identifying lower energy levels and upper energy levels

In our Quantum Cascade Laser (QCL) design methodology, after the initial selection phase, we employ a sophisticated clustering algorithm to categorize wavefunctions into distinct subgroups based on their energy levels. This process begins by preparing the selected wavefunctions, each characterized by its energy level, for clustering. A key step in this preparation involves assigning weights to each wavefunction proportional to its intensity relative to the minimum intensity observed in the dataset, amplifying the influence of wavefunctions with higher intensity on the clustering outcome.

To accommodate this weighting, we repeat the energy levels of wavefunctions according to their assigned weights, introducing a small, controlled random variation between -0.015 and 0.015 to each. This random variation is crucial for preventing the clustering algorithm from inaccurately dividing closely spaced energy levels into separate groups, thus ensuring a more accurate and cohesive clustering process.

The K-Means clustering algorithm is then applied to these adjusted energy levels, considering two and three possible clusters. The choice between these options is determined by the silhouette score, a measure of how well an object fits within its cluster compared to others, with the aim of achieving the most coherent and representative grouping. The

clustering model that yields the higher silhouette score, whether it be 2-means or 3-means, is selected for further analysis.

Upon finalizing the clustering model, wavefunctions are organized into their respective groups. The lowest energy cluster is identified as comprising the lower laser levels, while the higher cluster contains the upper laser level and any scattering states. Any clusters beyond these primary levels are deemed less relevant for our analysis and are consequently disregarded.

Through this refined clustering technique, we effectively sort wavefunctions into subgroups that mirror their energy states within the QCL structure. This alignment with their functional roles within the QCLs ensures that our automated identification process closely matches the physical properties and operational principles of these lasers, as illustrated in Fig. A.1c. This step is integral to our approach, enabling a more nuanced and precise design and analysis of QCLs.

A.3 Lower Energy levels (Lower Laser Levels and their extractor states)

Following the classification of upper and lower energy states in Quantum Cascade Lasers (QCLs), our automated program advances to accurately identify key lower lasing states, notably E3, alongside its extractor state E3', and additional lower lasing and extractor states as delineated in Fig. A.1e. This involves a refined clustering technique, akin to our initial method but with modifications tailored to these specific energy levels.

To begin, each wavefunction's energy level is prepared for clustering, with an emphasis on the wavefunction's intensity—significant for its impact on QCL performance. We assign weights to each wavefunction based on their intensity relative to the minimum intensity in the dataset. This ensures that wavefunctions with higher intensities have a greater influence on the clustering outcome. Subsequently, energy levels are replicated as per their assigned weights and adjusted with a minor random fluctuation between -0.003eV and 0.003eV to avoid misclassification of adjacent energy levels.

The K-Means clustering algorithm is applied, where we decide on the number of clusters—two or three—based on the silhouette score. This score helps us determine the most coherent grouping, ensuring an accurate representation of the energy states within the QCL structure.

For identifying the E3 state within the highest energy subgroup, we employ a weighting strategy based on the wavefunction's position in the active region. Within this subgroup, each wavefunction's intensity in the active region is weighted based on its position, with the leftmost part (closer to the injector region) assigned a higher weight. This weighting is linear, with a maximum weight of 2 at the leftmost point, decreasing to 1 at the rightmost point. This approach reflects the need for E3 to have a higher wavefunction overlap with E4, which predominantly resides in the left half of the active region. The wavefunction exhibiting the highest weighted intensity in this subgroup is designated as E3.

To discern E3', we calculate a distance metric for each wavefunction in E3's subgroup, taking into account several factors including the norm difference in wavefunction shapes within the active region, closeness in energy levels, and intensity within the active region. This composite metric allows us to pinpoint the wavefunction closest to E3, marking it as E3'.

$$D = F_{\text{act}} \cdot F_{\text{en}} \cdot F_{\text{next}} \cdot F_{\text{ext}} \cdot F_{\text{int}}, \tag{A.1a}$$

$$F_{\text{act}} = \|\psi_{E3}^{\text{act}} - \psi_{\text{cand}}^{\text{act}}\|, \tag{A.1b}$$

$$F_{\text{en}} = \exp\left(-\frac{|\epsilon_{E3} - \epsilon_{\text{cand}}|}{E_{\text{base}}}\right), \tag{A.1c}$$

$$F_{\text{int}} = \frac{1}{\max(I_{\text{cand}}^{\text{act}}, 0.01)}, \tag{A.1d}$$

$$F_{\text{next}} = \exp\left(-\frac{I_{\text{cand}}^{\text{next}}}{10}\right), \tag{A.1e}$$

$$F_{\text{ext}} = \frac{1}{\max(I_{\text{cand}}^{\text{ext}}, 0.00001)}, \tag{A.1f}$$

where D is the distance metric, F factors represent various components of the distance calculation, ϵ denotes the wavefunction, ψ represents energy levels, I is intensity, the value of the factor E_{base} is set as one-twentieth of the range between the lowest and highest

energy levels in the E3's subgroup, superscript act, ext, and next refers to the parameters in active, extractor, and next active region, and subscript cand refers to the candidate wavefunction. There are two minimum cutoff factors of 0.01 and 0.00001. The cutoff is implemented to prevent computational overflow during calculations, ensuring stability and accuracy in the process.

Following E3 and E3' identification, remaining wavefunctions in the lower energy cluster are collectively regarded as lower energy states, including E1, E2, and their corresponding extractor states, without detailing distinctions among these states.

A.4 Upper Energy Levels (Upper Laser Level and Scattering Levels)

In the analysis of the upper energy levels of QCLs, we employ a clustering approach similar to the one used for the lower energy levels. However, there are slight variations tailored to the characteristics of these upper levels. One difference is the range of random variation applied to the energy levels, which is set between -0.005eV and 0.005eV. The process of determining the number of clusters, or subclasses, is conducted by testing a range from one cluster up to the total number of energy states present in the upper energy subgroup. For each potential number of clusters, the silhouette score is calculated. The number of clusters that yields the highest silhouette score is chosen as the optimal partitioning. This approach helps in identifying the existence of various states: a single subclass indicates only E4 is present; two subclasses suggest the existence of both E4 and E5; and three or more subclasses imply the presence of E4, E5, and E6.

In the process of identifying state E4 in the upper energy, we utilize a probability calculation involving two key factors for each wavefunction: the energy factor and the

intensity factor. These calculations are outlined below:

$$P_{E_4} = F_{\text{en}} \cdot F_{\text{int}}, \tag{A.2a}$$

$$F_{\text{en}} = \exp\left(-\frac{|\epsilon_{\text{cand}} - \epsilon_{\text{sub}}|}{E_{\text{base}}}\right), \tag{A.2b}$$

$$F_{\text{int}} = \exp\left(\min\left(I_{\text{cand}}^{\text{act}}, 0.5\right)\right), \tag{A.2c}$$

where ϵ_{sub} is the energy of the bottom of the first subgroup from bottom up, E_{base} is set as a half of the difference between the bottom energy of the first subgroup and of the third subgroup. If there are less than three subgroups, the bottom energy of the third subgroup is replaced by the top energy of the upper energy levels. Here is a cut-off factor of 0.5 for the intensity. This cutoff is implemented to moderate the influence of wavefunction intensity on the overall probability. In the context of identifying E4, the emphasis is primarily on the energy level's proximity to ϵ_{sub}, while the wavefunction's intensity only needs to surpass a certain threshold to be considered significant. Beyond this threshold, further increases in intensity do not disproportionately affect the likelihood of a wavefunction being identified as E4. This approach ensures that the identification of E4 is more sensitive to its energy level rather than its intensity within the active region.

To identify states E5 and E6, the calculations are similar, but with ϵ_{sub} adjusted to the lowest energy level of the 2nd and 3rd subgroups for E5 and E6, respectively. An additional factor F_{E_4} is introduced, which is designed to prevent states that are too close to E4 in terms of energy levels from being erroneously identified as E5. It functions as a filter, reducing the likelihood of a state being classified as E5 if its energy level is very similar to that of E4.

$$P_{E_5} = F_{\text{en}} \cdot F_{\text{int}} \cdot F_{E_4} \tag{A.3a}$$

$$F_{E_4} = \frac{1}{1+\exp(-1000\cdot(\epsilon_{\text{cand}} - \epsilon_{E4} - 0.01))} \tag{A.3b}$$

For E6, the F_{E_4} is replaced by F_{E_5}.

$$F_{E_5} = \frac{1}{1 + \exp\left(-1000 \cdot (\epsilon_{\text{cand}} - \epsilon_{E5} - 0.01)\right)} \tag{A.4}$$

These probability calculations help in accurately identifying each state. The wavefunction with the highest probability in each category is designated as the corresponding state (E4, E5, or E6). The significance of these formulas lies in their ability to consider both the energy proximity to the base energy of each group and the intensity within the active region, thereby enhancing the accuracy of state identification in QCLs. This approach is visually represented in Fig. A.1f, where only the process of identifying E4 is shown.

A.5 Injector States Identification

Following the identification of E4, our focus shifts to determining injector states G0 and G1, pivotal for the Quantum Cascade Laser (QCL) design but not encompassed within the initial selection range. Our methodology targets wavefunctions predominantly situated on the QCL's active region's left side, earmarked for their significant role in carrier injection and, consequently, the laser's performance.

The search for G0 and G1 initiates by examining a specific range, spanning from just above E3 to either E5 or E4, contingent on the presence of E5. This range is methodically sifted through, evaluating each wavefunction for its potential as an injector state based on several criteria encapsulated in a calculated possibility score. Central to this evaluation is the wavefunction's energy proximity to the midpoint between E3 and E4, its intensity distribution favoring the injector region, and its comparative intensity to E4 within the active region. These factors collectively inform the likelihood of a wavefunction being an injector state.

The possibility score for each wavefunction is derived from a multiplicative combination

of distinct factors:

$$P_{\mathrm{G}} = F_{\mathrm{en}}^{\mathrm{G}} \cdot F_{\mathrm{int,\ inj}} \cdot F_{\mathrm{int,\ act}}, \tag{A.5a}$$

$$F_{\mathrm{en}}^{\mathrm{G}} = \exp\left(-\frac{|\epsilon_{\mathrm{avg}} - \epsilon_{\mathrm{cand}}|}{E_{\mathrm{base}}}\right), \tag{A.5b}$$

$$E_{\mathrm{base}} = \frac{\epsilon_{E4} - \epsilon_{E3}}{20}, \tag{A.5c}$$

$$F_{\mathrm{int,\ inj}} = I_{\mathrm{cand}}^{\mathrm{inj}}, \tag{A.5d}$$

$$F_{\mathrm{int,\ act}} = \exp\left(\frac{I_{\mathrm{cand}}^{\mathrm{act}}}{I_{E4}^{\mathrm{act}}}\right), \tag{A.5e}$$

where P_{G} is the possibility score for injector states, $F_{\mathrm{en}}^{\mathrm{G}}$ represents the energy factor for injector states, $F_{\mathrm{int,\ inj}}$ is the intensity in the injector region, and $F_{\mathrm{int,\ act}}$ is the factor for intensity comparison in the active region. ϵ_{avg} is the average energy between E_3 and E_4.

The two wavefunctions with the highest possibility scores are considered as candidates for G0 and G1. This process, illustrated in Fig. A.1h, effectively identifies G0 and G1 from a range of potential candidates. It ensures that the selected states are not only appropriately located within the QCL structure but also have the characteristics that align with the expected behavior of injector states.

Appendix B

The k.p Solver and Performance Metric Calculation Program

The k·p method is a semi-empirical technique widely used for calculating the electronic band structure of crystalline solids, including complex semiconductor heterostructures like Quantum Cascade Lasers. It provides an approximate solution to the Schrödinger equation by modeling the wavefunctions of a periodic crystal. The simulation program used in this research is built upon a 3-band k·p solver, which serves as the core engine for generating the fundamental data connecting a QCL's physical structure to its electronic properties. For analyzing individual QCL designs and visualizing results, a Graphical User Interface (GUI) was developed to streamline the workflow.

The GUI facilitates a step-by-step process for calculating QCL performance metrics. The workflow begins with loading a QCL structure file, which contains the predefined parameters for the design, such as the material composition and thickness of each of the 24 layers in the active region. These parameters are then displayed within the interface, where they can be manually inspected, adjusted, or fine-tuned as needed. This allows for rapid iteration on a single design.

Once the structural parameters are finalized, the k·p solver is executed. It computes the conduction band diagram for the given QCL structure, graphically representing the

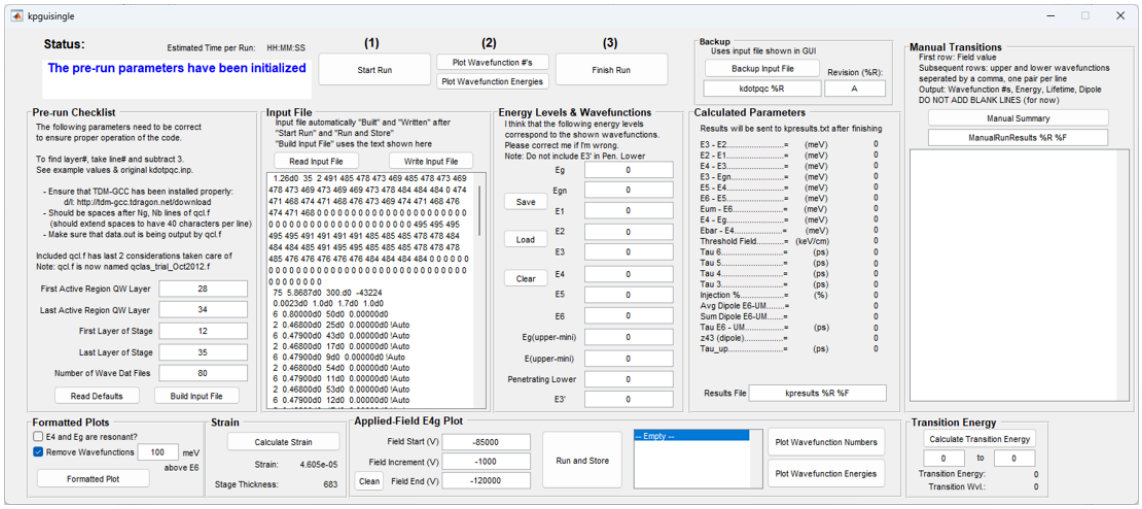


Figure B.1: the Graphical User Interface (GUI) of the k·p solver program.

energy levels and their corresponding wavefunctions. Following this, the crucial step of wavefunction identification takes place. In a conventional workflow, this requires a user to visually inspect the band diagram and manually input the identifying numbers for the key energy states (e.g., the upper and lower laser levels, E_4 and E_3). While effective for a single device, this manual process is prohibitively time-consuming and impractical for generating large-scale datasets. To overcome this bottleneck, the automated wavefunction identification program discussed in the main body of this work was developed to algorithmically perform this step, enabling rapid and consistent labeling across hundreds of thousands of structures.

After the required states are identified (either manually via the GUI or automatically in the large-scale workflow), the final step is the calculation of key performance metrics. The program computes essential parameters directly linked to device performance, such as the energy-level differences between states and the electron scattering lifetimes. While the GUI is an invaluable tool for detailed analysis and small-scale calculations, the generation of the massive dataset for machine learning bypasses the graphical interface entirely. Instead, the k·p solver and the automated wavefunction identification program are directly integrated and called from scripts, allowing for the efficient, high-throughput data generation required to train robust machine learning models.

Appendix C

Dataset and filterings

The dataset used to train the inverse neural network in this study is derived from the comprehensive dataset employed in the previous work, which was generated using a k·p energy band solver[13]. A significant modification to this dataset involved a filtering process that reduced the number of data points to just under 300,000. This reduction was necessitated by the effects of hybridization or energy level splitting, which significantly altered the distribution of wavefunctions, adversely affecting the predictability of lifetimes—a factor not considered in the earlier study focused solely on energy level differences. In cases of energy level hybridization or splitting, the impact on energy level differences is minimal and often obscured by other errors, yet it becomes significant when predicting lifetimes. Manual inspection indicated that these anomalies, potentially due to physical properties or computational inaccuracies in the k·p solver, adversely affected the model's accuracy.

To illustrate the phenomenon of energy level hybridization or splitting, Fig. C.1 provides a clear example. As indicated by the arrow in the image, another energy level exists in close proximity to the E_4 level. The energy difference between these two levels is minimal, and their distribution patterns are remarkably similar, typifying the cases of concern. In the dataset used for training the inverse neural network, efforts were made to exclude such data points, thereby minimizing noise within the dataset. This selective filtering was crucial to enhance the model's robustness by focusing on more stable and distinct

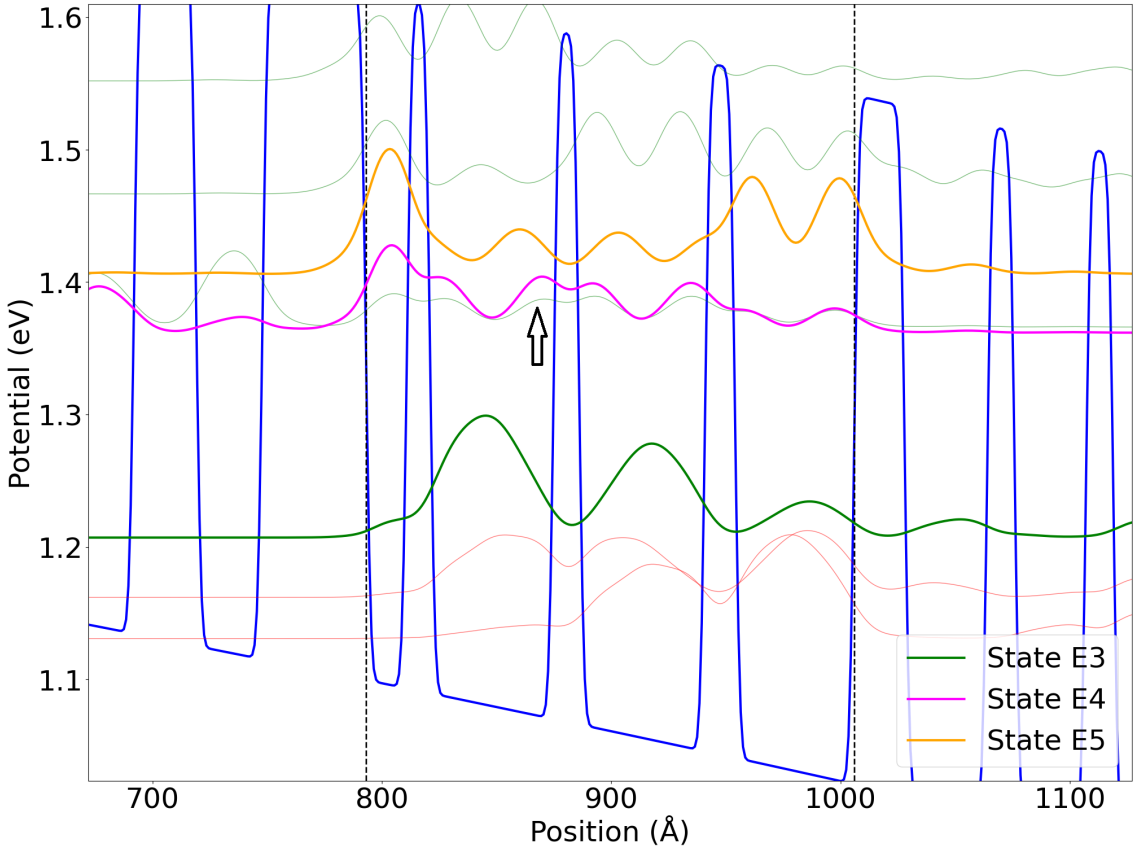


Figure C.1: An example of state hybridization

energy levels, ultimately aiming to improve the predictability of lifetimes without the interference of hybridized or split energy levels, which could introduce significant variability and potential errors in lifetime estimation. This approach underscores the importance of dataset integrity and the need for careful preprocessing to ensure reliable outputs from sophisticated computational models.

The selection of QCL AR structures for inclusion in the dataset was limited to InGaAs and AlInAs layer compositions, lattice-matched to InP substrates, with variations only in layer thickness[13]. The dataset was also limited to the labeling of key energy states (E_3, E_4, E_5)[13]. These states are critical for determining the operational characteristics of QCLs, such as lasing frequency and leakage currents.

In terms of k·p metrics, in addition to the energy level differences (E_{43} and E_{54}), we introduced carrier lifetimes between E_4 and E_3 (τ_{43}) and between E_5 and E_4 (τ_{54})[1]. We found that using the reciprocals of the lifetimes (the transition rates) has resulted in greater predictive accuracy in the forward networks than the use of the lifetimes themselves. However, the transition rate values were exceedingly small in some cases, leading to disproportionately large errors in the calculated lifetimes through inversion, although these instances were rare.

Bibliography

[1] Dan Botez and Mikhail A Belkin. *Mid-Infrared and Terahertz Quantum Cascade Lasers*. Cambridge University Press, 2023.

[2] C Boyle et al. "Carrier leakage via interface-roughness scattering bridges gap between theoretical and experimental internal efficiencies of quantum cascade lasers". In: *Applied Physics Letters* 117.5 (2020), p. 051101.

[3] O Cathabard et al. "Quantum cascade lasers emitting near 2.6 μm". In: *Applied Physics Letters* 96.14 (2010).

[4] Larry A Coldren, Scott W Corzine, and Milan L Mashanovitch. *Diode lasers and photonic integrated circuits*. John Wiley & Sons, 2012.

[5] Jérôme Faist. *Quantum Cascade Lasers*. Oxford University Press, Mar. 2013. ISBN: 9780198528241. DOI: 10.1093/acprof:oso/9780198528241.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780198528241.001.0001.

[6] Jerome Faist et al. "Quantum cascade laser". In: *Science* 264.5158 (1994), pp. 553–556.

[7] Jérôme Faist et al. "Short wavelength (λ 3.4 μm) quantum cascade laser based on strained compensated InGaAs/AlInAs". In: *Applied Physics Letters* 72.6 (1998), p. 680.

[8] Martin Franckié and Jérôme Faist. "Bayesian optimization of terahertz quantum cascade lasers". In: *Physical Review Applied* 13.3 (2020), p. 034025.

[9] Claire Gmachl et al. "Recent progress in quantum cascade lasers and applications". In: *Reports on progress in physics* 64.11 (2001), p. 1533.

[10] Andres Correa Hernandez and Claire F Gmachl. "A Machine Learning Framework for Quantum Cascade Laser Design". In: *arXiv preprint arXiv:2406.07755* (2024).

[11] Andres Correa Hernandez and Claire F Gmachl. "Application of machine learning to quantum cascade laser design". In: *2023 57th Annual Conference on Information Sciences and Systems (CISS)*. IEEE. 2023, pp. 1–6.

[12] Y Hu et al. "Enhancing quantum cascade laser active region design through inverse neural networks: A machine learning approach to metric-based structure generation". In: *AIP Advances* 14.10 (2024).

[13] Y Hu et al. "Large-scale data generation for quantum cascade laser active-region design with automated wavefunction identification". In: *Applied Physics Letters* 124.24 (2024).

[14] Katharina Isensee, Niels Kröger-Lui, and Wolfgang Petrich. "Biomedical applications of mid-infrared quantum cascade lasers–a review". In: *Analyst* 143.24 (2018), pp. 5888–5911.

[15] E.O. Kane. "Chapter 3 The k •p Method". In: *Semiconductors and Semimetals*. Ed. by R.K. Willardson and Albert C. Beer. Vol. 1. Semiconductors and Semimetals. Elsevier, 1966, pp. 75–100. DOI: https://doi.org/10.1016/S0080-8784(08)62376-5. URL: https://www.sciencedirect.com/science/article/pii/S0080878408623765.

[16] RF KazarinovR and A Suris. "Possibility of the amplification of electromagnetic waves in a semiconductor with a superlattice". In: *Sov. phys. semicond* 5.4 (1971), p. 207.

[17] Dianjing Liu et al. "Training Deep Neural Networks for the Inverse Design of Nanophotonic Structures". In: *ACS Photonics* 5.4 (2018), pp. 1365–1369. DOI: 10.1021/acsphotonics.7b01377. eprint: https://doi.org/10.1021/acsphotonics.7b01377. URL: https://doi.org/10.1021/acsphotonics.7b01377.

[18] J. M. Luttinger and W. Kohn. "Motion of Electrons and Holes in Perturbed Periodic Fields". In: *Phys. Rev.* 97 (4 Feb. 1955), pp. 869–883. DOI: 10.1103/PhysRev.97.869. URL: https://link.aps.org/doi/10.1103/PhysRev.97.869.

[19] Colin J Mitchell et al. "Mid-infrared silicon photonics: From benchtop to real-world applications". In: *APL Photonics* 9.8 (2024).

[20] Cara P Murphy and John P Kerekes. "Physics-guided neural network for predicting chemical signatures". In: *Applied Optics* 60.11 (2021), pp. 3176–3181.

[21] Manijeh Razeghi et al. "High power quantum cascade lasers". In: *New Journal of Physics* 11.12 (2009), p. 125017.

[22] Bahaa EA Saleh and Malvin Carl Teich. *Fundamentals of photonics, 2 volume set*. john Wiley & sons, 2019.

[23] Cagatay N Sengor, Feridun Ay, and Cahit Perkgoz. "A deep learning approach for high-resolution and enhanced efficiency in photonic power dividers". In: *Journal of Applied Physics* 137.12 (2025).

[24] Carlo Sirtori et al. "GaAs/AlxGa1xAs quantum cascade lasers". In: *Applied Physics Letters* 73.24 (Dec. 1998), pp. 3486–3488. ISSN: 0003-6951. DOI: 10.1063/1.122812. eprint: https://pubs.aip.org/aip/apl/article-pdf/73/24/3486/18538490/3486_1_online.pdf. URL: https://doi.org/10.1063/1.122812.

[25] Alexander Spott et al. "Quantum cascade laser on silicon". In: *Optica* 3.5 (2016), pp. 545–551.

[26] Simon M Sze, Yiming Li, and Kwok K Ng. *Physics of semiconductor devices*. John wiley & sons, 2021.

[27] Shigeyuki Takagi et al. "Optimization of Active Region of Quantum Cascade Laser (QCL) by Coupled Calculation of Genetic Algorithm and QCL Simulator." In: *PHOTOPTICS*. 2024, pp. 72–77.

[28] Nicolas Villa et al. “Quantum cascade lasers with discrete and non equidistant extended tuning tailored by simulated annealing”. In: *Optics Express* 27.19 (2019), pp. 26701–26707.

[29] Lok C Lew Yan Voon and Morten Willatzen. *The kp method: electronic properties of semiconductors*. Springer Science & Business Media, 2009.

[30] Igor Vurgaftman, JÁR Meyer, and L Ramdas Ram-Mohan. “Band parameters for III–V compound semiconductors and their alloys”. In: *Journal of applied physics* 89.11 (2001), pp. 5815–5875.

[31] Inès Waldmueller et al. “Inverse-quantum-engineering: a new methodology for designing quantum cascade lasers”. In: *IEEE journal of quantum electronics* 46.10 (2010), pp. 1414–1420.

[32] Benjamin S Williams. “Terahertz quantum-cascade lasers”. In: *Nature photonics* 1.9 (2007), pp. 517–525.

[33] Yu Yao, Anthony J Hoffman, and Claire F Gmachl. “Mid-infrared quantum cascade lasers”. In: *Nature Photonics* 6.7 (2012), pp. 432–439.