Optical and Quantum Electronics (2021) 53:287
https://doi.org/10.1007/s11082-021-02885-0

Bayesian optimization of quantum cascade detectors

Johannes Popp[1] · Michael Haider[1] · Martin Franckié[2] · Jérôme Faist[2] · Christian Jirauschek[1]

Received: 31 October 2020 / Accepted: 12 April 2021 / Published online: 6 May 2021

Abstract

A Bayesian optimization algorithm in combination with a scattering based simulation approach is used for the optimization of quantum cascade detectors (QCDs). QCDs operate in the mid-infrared and terahertz regime and are, together with quantum cascade lasers, appropriate for the integration into on-chip applications such as gas sensors. Our modeling approach is based on a rate equation model and a Kirchhoff resistance network for noise modeling, using scattering rates calculated with Fermi's golden rule, or alternatively extracted from an ensemble Monte Carlo transport approach. The appropriate surrogate model of Bayesian optimization is based on Gaussian process regression, which can handle noisy offsets on the objective function evaluations inherent in ensemble Monte Carlo simulations. Here, we focus on the optimization of a matured mid-infrared QCD design detecting at 4.7 μm. For optimization we choose as figure of merit the specific detectivity, which is a measure for the signal-to-noise ratio. As the trade-off between high extraction efficiency and low detector conductance is important for good detection performance, we search for the perfect layer composition and vary the thicknesses of different cascade layers. Due to the high-temperature requirements interesting for cost-effective and mobile on-chip sensing applications, a simulation temperature of 300 K is selected. Our optimization strategy yields an improvement of specific detectivity by a factor of $\sim 2-3$ at room temperature using two different parameter sets. Furthermore, we investigate the sensitivity of our approach to fabrication tolerances, showing robustness of the optimized designs against growth fluctuations under fabrication conditions.

Keywords Bayesian optimization · Gaussian process regression · Quantum cascade detector · Monaco framework · Fermi's golden rule · Ensemble Monte Carlo

This article is part of the Topical Collection on Numerical Simulation of Optoelectronic Devices. Guest edited by Stefan Schulz, Silvano Donati, Karin Hinzer, Weida Hu, Slawek Sujecki, Alex Walker and Yuhrenn Wu.

✉ Johannes Popp
johannes.popp@tum.de

Martin Franckié
martin.franckie@phys.ethz.ch

[1] Department of Electrical and Computer Engineering, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany

[2] Department of Physics, ETH Zurich, Auguste-Piccard-Hof 1, 8093 Zurich, Switzerland

1 Introduction

The quantum cascade detector (QCD)(Hofstetter et al. 2002; Gendron et al. 2004) is based on the matured quantum cascade laser (QCL) design principle (Faist et al. 1994) and consists of a multiple quantum well heterostructure. In contrast to QCLs, where lasing is achieved due to stimulated emission between quantized states, in QCDs the stimulated absorption and thus photodetection is the relevant physical mechanism. Compared to quantum well infrared photodetectors (QWIPs)(Levine et al. 1987; Liu 1993), which utilize a photoconductive operation, QCDs are based on an asymmetric conduction band profile, and thus exhibit photovoltaic detection behavior under zero bias operation. The unipolar character of both designs brings the advantage of high speed operation in comparison to interband devices (Hofstetter et al. 2006). In addition, the QCD exhibits a superior noise behavior accompanying the unbiased operation. In QWIPs, the main noise source is dark current noise, whereas QCDs are mainly limited by Johnson noise (Gendron et al. 2004; Giorgetta et al. 2009; Hofstetter et al. 2010). The reduced noise sensitivity in high-temperature operation is an outstanding feature of QCDs and offers high potential for hand-held mobile applications, e.g. terrestrial staring systems or sensing systems (Hofstetter et al. 2010; Schwarz et al. 2012; Harrer et al. 2016; Schwarz et al. 2014). Furthermore, QCDs have the advantage of simple adaption and integration into the matured processing technique of QCLs for the material systems GaAs/AlGaAs and InGaAs/InAlAs, resulting in increased design freedom and reliability (Schwarz et al. 2012). In order to optimize the detector performance, different designs have been tested, such as vertical (Giorgetta et al. 2009), diagonal (Reininger et al. 2014) or coupled quantum well detectors (Dougakiuchi et al. 2016).

In Fig. 1, a schematic conduction band profile with the subband states of a QCD is represented. The working principle of such a device is based on intersubband transitions. As illustrated in Fig. 1, the absorption transition takes place between the ground state g and the absorption state a in the active well, followed by the extraction through phonon scattering. The unilateral charge transport of photoexcited electrons is ensured by the graded quantum well composition of the extraction cascade.

The relevance of QCDs and thus the on-chip integration with QCLs is growing rapidly. Therefore, the modeling of such devices becomes a valuable tool for the design and

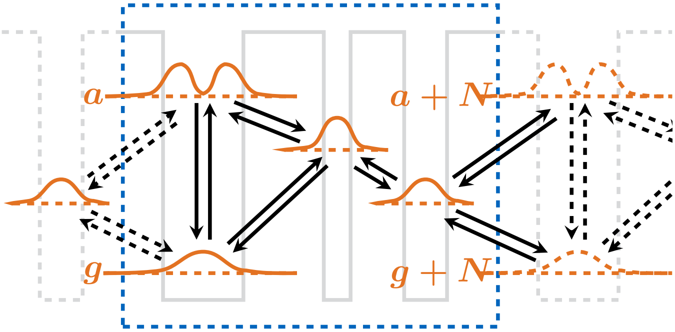

Fig. 1 Schematic conduction band profile of a QCD. Photovoltaic operation is ensured by absorption from ground level g to absorption level a and consecutive scattering through the extractor levels in the quantum cascade

optimization process. Unfortunately, the adoption of QCL modeling methods, e.g. rate equation, ensemble Monte Carlo (EMC) methods, density matrix or non-equilibriums Green's functions (NEGF) (Jirauschek and Kubis 2014; Wacker et al. 2013; Iotti et al. 2005), is rather difficult due to the small photocurrents in QCDs. Models extracting scattering rates from EMC simulations or by calculations using Fermi's golden rule were introduced (Baumgartner et al. 2013; Harrer et al. 2016; Koeniguer et al. 2006; Popp et al. 2020a). Here, a modeling approach based on a rate equation model and noise resistance model for the Kirchhoff equation system is used for the calculation of the figure of merit of QCDs (Popp et al. 2020a, b). An extended description about the methodology of our QCD modeling tool will be published soon.

A systematic design optimization of QCDs is an essential task for the development of highly sensitive detectors. Different optimization strategies have already been applied to QCL design processes. A genetic optimization algorithm in combination with a density matrix transport model was used to improve the wallplug efficiency of a mid-infrared (mid-IR) QCL design (Bismuto et al. 2012). Recently, a Bayesian optimization (BO) algorithm with a NEGF transport model was used to improve the maximum operation temperature of a terahertz (THz) QCL to a new record value of 210 K (Bosco et al. 2019). Franckié and Faist (2020) published results for a comparison of the Bayesian optimization algorithm with an "information algorithm with parallel trials" (IAPT) algorithm and the aforementioned genetic algorithm. The three optimization tools were applied to a Gaussian process (GP) model, which was trained for a THz QCL using the QCL gain as merit function. The BO scheme shows best performance in terms of convergence and robustness.

In this paper, we present the Matlab Monaco framework for the simulation of quantum cascade devices, e.g. detectors and lasers. The Monaco framework is similar to the open source GitHub project aftershoq (Franckié 2019), which focuses on the optimization of QCLs. In contrast, the focus of the Monaco project is laid on QCD optimization. The Matlab framework includes a Bayesian optimization algorithm for quantum cascade devices, utilizing self-consistent simulation tools such as EMC. The paper is organized as follows: Sect. 2 offers a short introduction to Bayesian optimization including Gaussian processes, acquisition functions and the Hilbert curve. In Sect. 3, a small summary about the implemented QCD modeling tool containing the most important detector figures of merit is given. The structure of the Monaco framework is described in Sect. 4. In Sect. 5, results of the Bayesian optimization of a QCD using two different parameter sets are depicted. Here, we focus on the well-established mid-IR QCD design N1022 with a detection wavelength of 4.7 µm (Giorgetta et al. 2009), and discuss the obtained results of improved spectral detectivity.

2 Bayesian optimization

For the optimization of quantum cascade devices using the existing simulation tools (e.g. EMC, density matrix or NEGF), which exhibit advanced complexity and accuracy, an efficient optimization strategy is necessary. The input space in such simulations can span from different layer variations of the active QCD period to changes in the doping density or material compositions. The Bayesian optimization algorithm is an appropriate statistical tool as it is applicable, in general, to an unknown objective function $f(x)$, which can be expensive in sense of time and computational load. The algorithm is characterized by searching for the global minimum of the objective function $f(x)$ on a high-dimensional input space $x \in \mathbb{R}^d$ (Frazier 2018). Bayesian optimization basically consists of two main elements, a surrogate model and an acquisition function. The surrogate model is a Gaussian process, which is trained by function evaluations. The acquisition function acts as a utility function and thus helps to interpret the posterior function distribution, and makes a decision for the next data points to be evaluated (Snoek et al. 2012).

2.1 Gaussian process

A Gaussian process is interpreted as a Gaussian distribution over functions, and fully described by its mean $\mu(x)$ and covariance function $k(x, x')$ (Rasmussen 2004). It specifies a collection of random variables forming a joint Gaussian distribution. For randomly chosen input values x_i, the function values can be drawn by the prior distribution of function values $f(x_i)$, which is a Gaussian distribution. An appropriate choice for a simple covariance function $k(x, x')$ is the squared exponential covariance function

$$k(x, x') = \sigma_0^2 \exp(-\frac{1}{2}\sigma_l^2 |x - x'|^2), \tag{1}$$

where σ_l^2 is the characteristic length scale and σ_0^2 the covariance amplitude. GPs account for noisy function values $f(x) + \epsilon$ by an additional uncorrelated Gaussian noise term ϵ with variance σ_n^2. After the evaluation of the objective function f at some input points $\mathbf{x}^*$, the posterior probability distribution is calculated by conditioning the joint distribution on the function evaluations $\mathbf{y}^* = f(\mathbf{x}^*)$:

$$p(y|x, \mathbf{x}^*, \mathbf{y}^*, \theta) \sim \mathcal{N}(\mu, k) \tag{2}$$

with mean $\mu = k(x, \mathbf{x}^*) \cdot [k(\mathbf{x}^*, \mathbf{x}^*) + \sigma_n^2 \mathbf{I}]^{-1} \cdot \mathbf{y}^*$, covariance matrix $K = k(x, x) - k(x, \mathbf{x}^*) \cdot [k(\mathbf{x}^*, \mathbf{x}^*) + \sigma_n^2 \mathbf{I}]^{-1} \cdot k(\mathbf{x}^*, x)$ and hyperparameters $\theta = (\sigma_0, \sigma_l, \sigma_n)$. The training data from all previous iterations are summarized in $(\mathbf{x}^*, \mathbf{y}^*)$. New function values y can thus be drawn for new random test inputs x. By maximizing the marginal likelihood $p(\mathbf{y}^* | \mathbf{x}^*, \theta)$, the optimal values of hyperparameters θ to describe the training data can be found.

2.2 Acquisition function

The acquisition function $a(x)$ is a measure for the yield of the next evaluation input point x_n and is defined as $x_n = \arg\max_x a(x)$. As an example, we will introduce here the common acquisition function *Expected Improvement* (EI), which gives a sensitive weight for exploration and exploitation. The EI acquisition function is defined as

$$\mathrm{EI}(x) = \langle (t - y)_+ \rangle = \int_{-\infty}^{\infty} (t - y)_+ p(y|x, \theta) \mathrm{d}y \tag{3}$$

with $(t - y)_+ = \max(0, t - y)$, where t is a target value and is usually set to be the maximum of pre-evaluated function values. Here, exploration means to investigate domains in the input space with high uncertainty and exploitation means to evaluate the expected maximum of the predictive mean (Gelbart et al. 2014). To summarize, EI exhibits high function values for inputs x either with high predictive mean $\mu(x)$ or high predictive variance $\sigma^2(x)$, or both.

3 Modeling method

For the characterization of photodetector performance a key figure of merit is the specific detectivity D^*. The Johnson noise limited detectivity for QCDs is given by (Giorgetta et al. 2009)

$$D^* = R_\mathrm{p}\sqrt{\frac{A_\mathrm{d}R_\mathrm{d}}{4k_\mathrm{B}T}}, \tag{4}$$

where R_p is the peak responsivity, A_d the detector area, R_d the detector resistance, T the temperature and k_B the Boltzmann constant.

The detector resistance is dominated by Johnson noise in QCDs and can be calculated by solving the Kirchhoff equations for a network consisting of noise equivalent resistances. The conductance per unit area dominated by Johnson noise, obtained by representing the subbands as nodes of an equivalent circuit (Delga et al. 2013), is given by

$$\sigma_{ij} = \sigma_{ji} = \frac{e^2 n_\mathrm{s}}{2k_\mathrm{B}T}\left(r_{ij}p_i + r_{ji}p_j\right). \tag{5}$$

Here, n_s is the sheet doping density per period, e the elementary charge, and r_{ij} describes the transition rate, taking into account all relevant scattering mechanisms from a state i with the level occupations p_i to state j.

The frequency dependent responsivity, which is defined by the generated detector photocurrent I_out per incident optical power P_in, is calculated by

$$R_\mathrm{p}(\omega) = \frac{I_\mathrm{out}(\omega)}{P_\mathrm{in}(\omega)} = \frac{e}{\hbar\omega}\frac{p_\mathrm{e}}{N_\mathrm{p}}T_\mathrm{f}[1 - \exp(-\alpha n_\mathrm{p} N_\mathrm{p} L_\mathrm{p} \sin\theta)], \tag{6}$$

where T_f is the facet reflectivity, α the absorption coefficient, n_p the number of paths of the optical field through the absorbing region, N_p the number of periods in the active region, L_p the length of one period, $\omega = 2\pi c/\lambda$ the angular frequency and $\hbar$ the reduced Planck constant (Harrer et al. 2016). In general, a mesa-structure with a double-pass waveguide is considered for the characterization of such detector devices. Here, the two facets are polished into 45° wedges. The light beam exhibits a propagation angle $\theta = 45°$ relative to the growth direction to meet the quantum-mechanical polarization rule.

We have developed a robust method to calculate the extraction efficiency p_e, utilizing a rate equation model in analogy to Jirauschek (2018) and calculating the absorption efficiency using transition rates extracted either from Monte Carlo simulations or obtained directly by solving Fermi's golden rule for elastic and inelastic scattering processes (Jirauschek and Kubis 2014; Jirauschek 2017). The wavefunctions and eigenenergies necessary for the calculations of scattering rates in the EMC are generated by a Schrödinger-Poisson solver based on the transfer matrix method (TMM) (Jirauschek 2009). The EMC simulation model (Jirauschek and Kubis 2014) takes into account optical as well as acoustic phonons, interface roughness (IF), impurity, alloy disorder and electron-electron (e-e) scattering mechanisms; evaluated self-consistently based on Fermi's golden rule.

The simulation of photovoltaic QCDs differs from that of QCLs. Here we assume an operation in thermal equilibrium for zero bias and without illumination. The electron sheet density in a subband i at temperature T is then calculated using the 2D density of states $n_i^\mathrm{2D} = m_i/(\pi\hbar^2)$:

$$n_{\mathrm{s},i} = \frac{m_i}{\pi\hbar^2} k_{\mathrm{B}} T \ln\{1 + \exp[(\mu - E_i)/(k_{\mathrm{B}} T)]\} \tag{7}$$

with the quantized subband energy E_i, the effective subband masses m_i containing non-parabolicity effects, and the chemical potential μ.

The level spacing of the active region has to be chosen carefully to optimize the trade-off between maximum extraction efficiency p_{e} and high detector resistance R_{d}, indicating low Johnson noise. Our goal is to enhance the absorption efficiency, while keeping the Johnson noise low.

4 Implementation: Monaco framework

The Monaco framework is a tool for the simulation of quantum cascade devices, here used for the optimization of QCDs employing Bayesian optimization based on Gaussian process regression. For the implementation of this software project, we follow the guidelines of the bertha project skeleton (Riesch et al. 2020). The project is based on MATLAB object-oriented programming, the version control system git together with the appropriate project management tool GitLab are used for the development of new tool features. Furthermore, we use continuous integration for testing and CMake as buildsystem for external codes, such as the EMC code written in Fortran.

The Monaco framework consists of the following modules: optimizer, setup, backend, solver and post-processing. A schematic of the corresponding tool is illustrated in Fig. 2. The optimizer tool is based on the standard MATLAB function bayesopt, which offers a statistical model for the optimization of expensive objective functions. The MATLAB Bayesian optimization tool offers different options for the acquisition function, the kernel function of the GP or parallel function evaluations per step also known as batch Bayesian optimization (González et al. 2016). Additionally, a second choice for a BO toolbox can be integrated such as the open-source software tool GPyOpt from the machine learning group of the

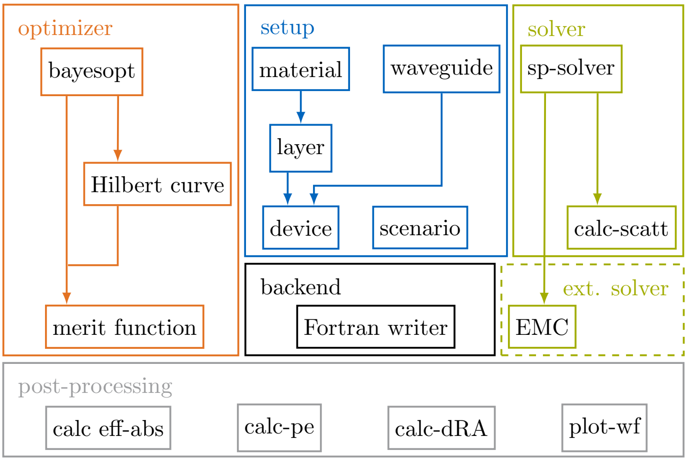

Fig. 2 Overview of the Monaco framework. The project consists of five modules: optimizer, setup, backend, solver and post-processing

University of Sheffield, which is written in Python (GPyOpt 2016). The project is available on GitHub and can easily be added as a submodule. The integration of Python code and interaction with our Matlab based framework is ensured by using the Matlab engine API for Python.

In the module optimizer one can choose between multi-dimensional input vectors and the application of the Hilbert curve for down conversion to a 1-dimensional optimization problem. The objective function can be selected for different figures of merit, e.g. specific detectivity D^* or responsivity R_p. For the evaluation of the figure of merit, the objective function has to be forwarded to the modules setup, solver and post-processing. The module setup is responsible for the generation of a QCD device and a scenario containing information about operation temperature T, simulation end time t_sim and bias V. The class device has the two properties layer and waveguide, representing the simulated design. The two common material systems for quantum cascade devices in the mid-IR and THz regime, InGaAs/InAlAs and AlGaAs/GaAs, are included. Class instances of different materials such as InGaAs or GaAs inherit from their specific material class ternary or binary and represent subclasses of the abstract class material. For the different materials, the interesting material parameters, e.g. band gap E_g, effective masses m_eff and conduction band offset, are calculated depending on the simulation scenario. Here, we consider also parameter changes due to operation temperature variations and the influence of compressive or tensile strain (Vurgaftman et al. 2001; Sugawara et al. 1993).

The generated objects of device and scenario form the investigated setup and serve as inputs for the module solver. For the calculation of wavefunctions and eigenenergies, the implemented Schrödinger-Poisson solver sp-solver is used. Here, we can simulate extended states or generate a tight-binding solution (Jirauschek and Kubis 2014; Jirauschek 2017). The calculated wavefunctions will then be used for the calculation of scattering rates either internally with the function calc-scatt based on Fermi's golden rule, or by the EMC tool. Here, the module backend is used to write the input files containing setup, wavefunctions and nonparabolicity parameters (Jirauschek and Kubis 2014). The given post-processing tools are used for the characterization of the simulated setup and act as inputs for the objective function to determine the given figure of merit to be optimized.

The validity of bayesopt as an appropriate tool for the optimization of QCDs was tested and the results are illustrated in Fig. 3. Here, we executed 50 BO runs of the nominal QCD structure N1022, introduced (see Sect. 5), to characterize the convergence rate and uncertainty of the optimization. The specific detectivity D^* converges quite fast to a global maximum, which makes it also practical using the time-demanding EMC approach for efficient design optimization.

5 Results

In this section, we present a Bayesian optimization of a QCD device using our Monaco framework. The device N1022 detects at a wavelength of 4.7 μm and is based on the lattice-matched material system $\mathrm{In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As}$ grown on an InP substrate (Giorgetta et al. 2009). The conduction band profile and the calculated wavefunctions are illustrated in Fig. 4a for the operation temperature 300 K. The QCD structure consists of multiple periods comprising a doped active quantum well (QW) and an adjacent extraction cascade of QWs with varying thicknesses. Photo-excitation occurs between the ground level g and the two degenerate absorption levels a_1, a_2 in the active QW,

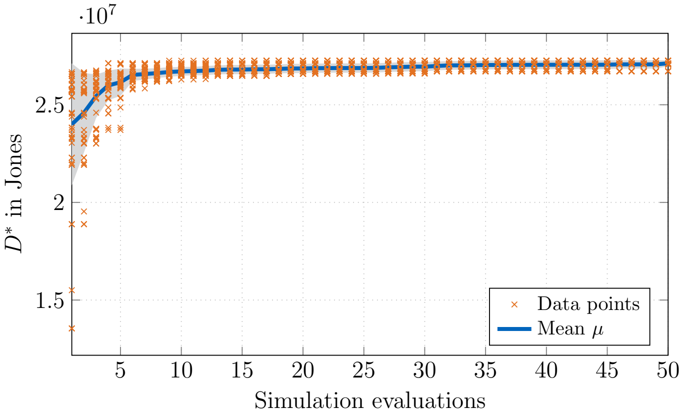

Fig. 3 The optimized specific detectivity D^* of the QCD test structure N1022 introduced in Fig. 4a. Here, the red line exhibits the mean value of 50 BO runs together with the data points (orange crosses). The gray area represents one standard deviation σ from the mean μ. The Bayesian optimization of each run was stopped after 50 evaluations. (Color figure online)

followed by the extraction through the staircase of subbands via longitudinal optical (LO) phonon assisted tunneling to the ground state of the adjacent period.

The structure N1022 was validated both with our scattering rate model and the EMC approach and the experimentally measured results compare well with the simulated ones (Giorgetta et al. 2009). Here, we investigated the specific detectivity D^*, the responsivity R_p, extraction efficiency p_e and resistance R_d in the temperature range $[100\,\mathrm{K}, 300\,\mathrm{K}]$. At 300 K, we obtain a specific detectivity $D^*_{\mathrm{MATLAB}} = 1.36 \times 10^7$ Jones and $D^*_{\mathrm{EMC}} = 1.09 \times 10^7$ Jones, respectively. The simulation values show good agreement with measured values of $D^*_{\mathrm{exp.}} \sim 2 \times 10^7$ Jones (Giorgetta et al. 2009).

The aim of this work is to improve the signal-to-noise ratio of the mid-IR QCD N1022 for the elevated temperature regime. Therefore, we decided to concentrate on the layer sequence $[w_1 b_1 w_2 b_2]$ indicated in Fig. 4a. To reduce the optimization complexity, the sequence was divided into two subsystems with each parameter set consisting of three consecutive layers. Furthermore, we can analyze the impact of the changes in individual layer width on device parameters and thus have more flexibility in the selection of the best layer composition.

For the BO, we chose a stepwidth of 0.1 Å and a testing interval dW ∈ [−2 Å, 2 Å] added to the nominal layer width of each considered layer. As evaluation method, we used the Matlab function calc-scatt. We performed the BOs with 2000 evaluations using multiple cores to get enough training data for a GP, which is used to analyze the optimization results. The conduction band profile and wavefunctions of the most successful scheme in each subset are illustrated in Fig. 4b, c. The optimization scheme is based on the parameter set 1, changing the layer sequence $[b_1 w_2 b_2]$ shown in Fig. 4b, and parameter set 2 by changing the layer sequence $[w_1 b_1 w_2]$ depicted in Fig. 4c. In the following, the results of both optimization results are explained in detail. In the concluding discussion we compare both setups and justify the model accuracy and emerging challenges regarding fabrication tolerances.

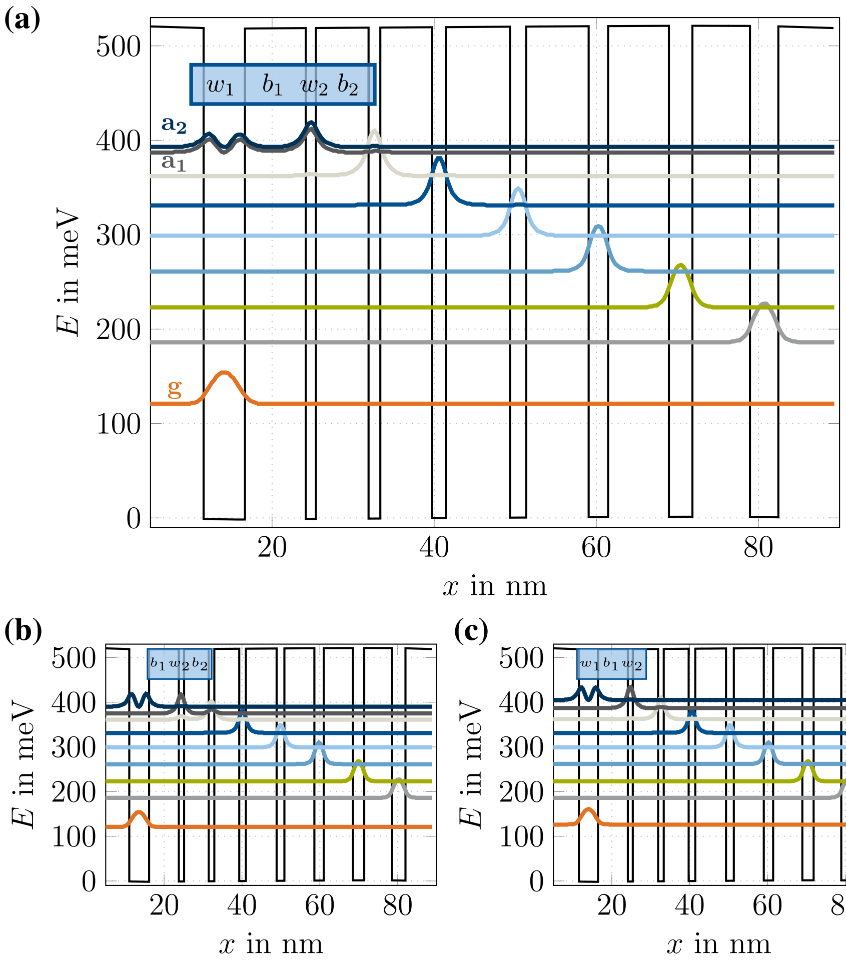

Fig. 4 Calculated conduction band profile and probability densities of **a** the investigated mid-IR QCD structure N1022 (Giorgetta et al. 2009) and the two optimized structures **b** r1_1 and **c** r2_1. The N1022 layer sequence of one period with InAlAs barrier layers in boldface and n-doped layers ($4 \times 10^{17}\text{cm}^{-3}$) underlined is **6.8**/5.1/**7.5**/1.25/**6.5**/1.45/**6.4**/1.7/**7.9**/2/**7.7**/2.4/**7.5**/2.9/**7.1**/3.5. The labeled layers $[w_1 b_1 w_2 b_2]$ in **a** are the investigated parameters for the optimization of the nominal structure N0122. The two optimization schemes **b** $[b_1 w_2 b_2]$ and **c** $[w_1 b_1 w_2]$ are illustrated by blue boxes, respectively

5.1 Parameter set 1

The parameter set 1 consists of the three layers $[b_1 w_2 b_2]$ next to the active well. With the layer sequence we intend to increase the absorption efficiency η_{abs}, while keeping Johnson noise low and the extraction efficiency p_{e} high. The detection wavelength should only be slightly affected by the optimization, since the active well layer and thus the absorbing transition are not directly changed. The optimized structure r1_1, illustrated in Fig. 4b, exhibits a change of the investigated layer sequence by [1.2 Å, 1.2 Å, −0.9 Å]. The simulation results of this structure are given in Table 1. We obtain an improvement

Table 1 Layer sequence with barrier layers in boldface, peak wavelength λ_p, extraction efficiency p_e, peak responsivity R_p, resistance-area product A_dR_d and specific detectivity D^* of the nominal structure N1022 and the optimized structures

ID	layer sequence $[w_1b_1w_2b_2]$ (nm)	λ_p (μm)	p_e (%)	R_p (mA W^{-1})	A_dR_d (Ωcm^2)	D^* ($\cdot 10^7$ Jones)
N1022	5.1/**7.5**/1.25/**6.5**	4.77	19.14	1.22	2.07	1.37
r1_1	5.1/**7.62**/1.37/**6.41**	4.72	17.10	2.41	2.11	2.73
r1_2	5.1/**7.3**/1.15/**6.31**	4.71	14.45	2.19	2.56	2.73
r2_1	4.9/**7.67**/1.27/**6.5**	4.57	24.34	3.03	2.32	3.58
r2_2*	4.8/**7.57**/1.23/**6.5**	4.50	28.18	3.32	2.50	4.07

of the simulated detectivity $D^*_{\text{opt.}} = 2.73 \times 10^7$ Jones by factor ~ 2. As illustrated in Fig. 4b, the change in layer composition results in a displacement of the two absorption levels. Here, the absorption maximum is shifted to the higher lying absorption level a_2, whereas the lower lying absorption level a_1 acts then mainly as an extraction level. In this context, the oscillator strength between the ground level g and the absorbing level a_2 is increased significantly. In summary, the peak responsivity is increased by 100% and accounts for the great improvement of the specific detectivity.

To analyze the obtained optimization results in more detail, we used a GP, trained with the simulation results of the BO run. Using GP regression, we can predict the changes in specific detectivity D^* with variation of the given layer sequence. In Fig. 5, the dependence of the specific detectivity D^*, peak responsivity R_p and resistance area product A_dR_d on pairs of layer thicknesses in the parameter set is shown. The position of the nominal structure N1022 is marked by a red pentagon. The specific detectivity D^* is influenced mostly by the well width w_2, as depicted in Fig. 5a, b. Here, we obtain a maximum at the well width of $w_2 = 11.5$ Å and $w_2 = 13.5$ Å, respectively. The impact of barrier widths b_1 and b_2 on the specific detectivity D^* is rather small. For characterization, we can divide the specific detectivity D^* in two parts: the responsivity, depending on the absorption and extraction efficiency, as a measure for the signal strength, and the detector resistance accounting for the current noise sensitivity. The optimized structures r1_1, representing the group of optimized structures at maximum $w_2 = 13.5$ Å, exhibits a significantly improved peak responsivity due to the increased absorption efficiency. The structure r1_2 given in Table 1 belongs to the other group with maximum $w_2 = 11.5$ Å. Here, both the area resistance product A_dR_d as well as the peak responsivity R_p are increased (Fig. 5c–f). An increased resistance at the cost of reduced extraction efficiency leads to smaller responsivity values, which explains the difference between both maxima in Fig. 5c, d. In summary, both optimized structures listed in Table 1 exhibit similar signal-to-noise behavior and an absorption wavelength of ~ 4.7 μm, which is close to the absorption wavelength of the nominal structure N1022. Structure r1_1 seems to be more robust with respect to fluctuations in layer width w_2 than structures r1_2 (Fig. 5a, b). As the first design r1_1 offers better signal strength and the second design r1_2 favors low noise behavior, one can choose the best-suited design for different applications.

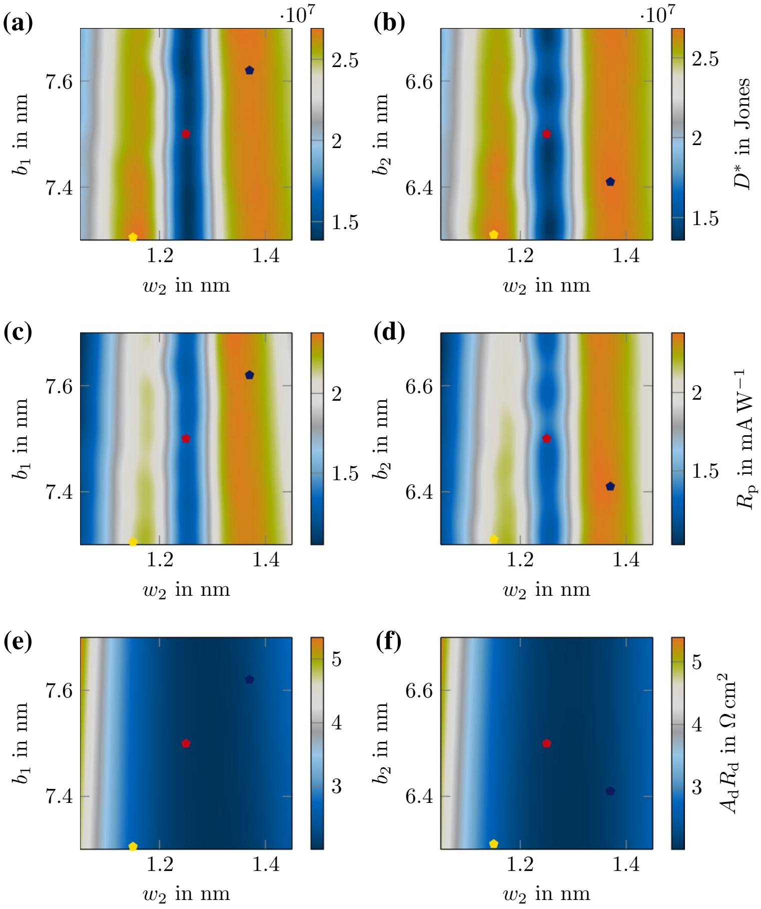

Fig. 5 Dependence of specific detectivity D^* (**a**), (**b**), peak responsivity R_p (**c**), (**d**) and resistance-area product A_dR_d (**e**), (**f**) on pairs of parameters, starting from the nominal structure N1022 and using the BO results of parameter set 1 with layer sequence $[b_1w_2b_2]$. The red, yellow and blue pentagons indicate the layer sequence of the nominal design N1022 and the optimized structures r1_2 and r1_1, respectively. The labels are defined in Fig. 4. (Color figure online)

5.2 Parameter set 2

The parameter set 2 consists of the three layers $[w_1b_1w_2]$ starting with the active well w_1. Here, we are also interested in the influence of the layer width of the active well w_1 on the device performance. In order to keep the the absorption frequency shift small, we introduce a new figure of merit

$$f(x) = D^*(x) \times \left(1 - \frac{|f_0 - f_{\mathrm{p,opt.}}(x)|}{f_0}\right), \tag{8}$$

where the specific detectivity D^* is multiplied by a weighting factor including the peak absorption frequency f_0 of the nominal structure N1022 and the peak absorption frequency of the sampled structure $f_{\mathrm{p,opt.}}$. The value x represents the parameter set consisting of the layer sequence $[w_1 b_1 w_2]$.

For the optimized design r2_1, a specific detectivity $D^* = 3.58 \times 10^7$ Jones is achieved, which implies even better results in absolute values as in BO run 1. The wavelength of 4.57 µm for the optimized design r2_1 is slightly smaller as in BO run 1. Due to the change of well width $w_1 = 49$ Å, both the ground level g and the absorption level a_2 are shifted to higher energy values. The change in energy of absorption level a_2 exceeds that of the ground level g, which results in a lower absorption wavelength. For all investigated structures, the transition rate from the absorption level a_2 to level a_1, as well as to the next extraction level, is dominated by interface roughness scattering. In case of structure r2_1 we observe an increased scattering from $a_2 \rightarrow a_1$ combined with an attenuated extraction from a_1 to the next extraction level. By comparison of Fig. 4b and c, one identifies an increased energy gap between level a_1 and the next extraction level of structure r2_1,

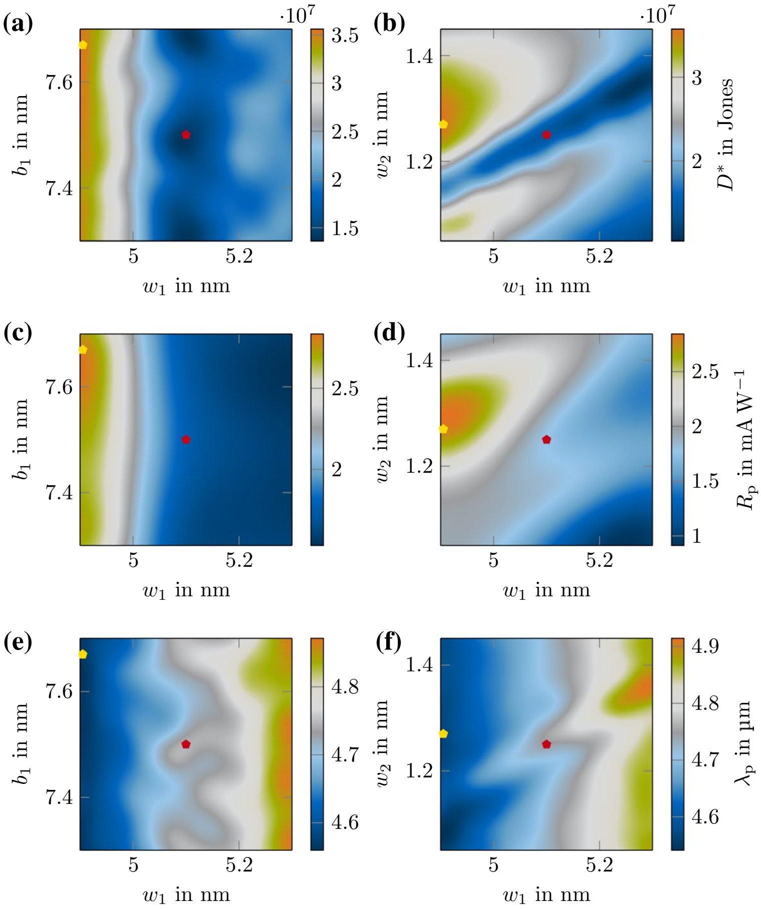

Fig. 6 Dependence of specific detectivity D^* (**a**), (**b**), peak responsivity R_p (**c**), (**d**) and resistance-area product $A_d R_d$ (**e**), (**f**) on pairs of parameters, starting from the nominal structure N1022 and using the BO results of parameter set 2 with layer sequence $[w_1 b_1 w_2]$. The red pentagons indicate the layer sequence of the nominal design N1022 and the optimized structure r2_1. The labels are defined in Fig. 4

which implies a shift of the dominating scattering mechanism from interface roughness to longitudinal optical phonon emission. The extraction efficiency $p_\mathrm{e} = 24.34\%$ and the resistance-area product $A_\mathrm{d}R_\mathrm{d} = 2.32\ \Omega\mathrm{cm}^2$ can thus simultaneously be increased, which results in superior signal-to-noise behavior.

In Fig. 6 the specific detectivity D^*, peak responsivity R_p and wavelength λ_p are shown for variation of pairs of parameters starting with the nominal structure values. Here, we see again a small dependence of the merit function on the layer width of barrier b_1. As explained before, the decrease of well width w_1 results in a significant increase of the specific detectivity D^* at the expense of a detection wavelength shift. For thicker well widths w_1, the opposite effect is observed. As illustrated in Fig. 6b, d and f, a strong correlation between the well widths w_1 and w_2 arises. Here, a balanced choice of these two layer widths is necessary for the optimization.

The optimization went to the edge of the parameter range for well width w_1 (Fig. 6a, c). Therefore, we decided to extend the optimization range and did a third BO run starting from the optimized structure r2_1. By further decreasing the well width $w_1 = 4.8$ Å, the optimized structure r2_2* with a specific detectivity $D^* = 4.07 \times 10^7$ Jones can be found. The simulation parameters of structure r2_2* are detailed in Table 1. By shrinking the well width w_1, the specific detectivity can be substantially enhanced at the expense of an undesired wavelength shift.

5.3 Discussion

In this paper, we focused on two different parameter sets for the Bayesian optimization of the detector design N1022 (Giorgetta et al. 2009; Hofstetter et al. 2010). The parameter set 1 with layer sequence $[b_1 w_2 b_2]$ ensures a stable optimization of the specific detectivity without fluctuations or drifts in the detection wavelength. As a consequence of BO run 1, we identify the influence of barrier width variations (b_1 and b_2) on the simulated device parameter to be rather small. The second parameter set includes the active well w_1 of the QCD. Here, we use specific detectivity multiplied with a weighting factor as new figure of merit to ensure a rather stable detection wavelength. The transitions from absorption level a_2 to the following extractor levels are mainly based on interface roughness. By changing the well widths w_1 and w_2, we can improve the extraction from level a_2 to a_1 and increase the energy gap from level a_1 to the next extraction level, which induces a transition of the dominating scattering mechanism from IF to LO phonon scattering. Here, the reduced scattering leads to an increased detector resistance. Furthermore, we can improve the absorption efficiency of all optimized structures due to the increased oscillator strength of $g \rightarrow a_2$. The important simulation parameters of the optimized structures are listed in Table 1, and the responsivity spectra of the different structures together with the nominal structure N1022 are displayed in Fig. 7a.

As mentioned before, the influence of specific layer widths on the wavefunctions and thus the detector behavior can vary significantly. Therefore, we decided to use smaller parameter sets, such that we can analyze correlations and sensitivities between layer widths and device parameters in more detail. The optimized structures of parameter set 2 implies better results than parameter set 1 in terms of signal to noise ratio. On the other hand, if a specific detection wavelength is crucial, one should concentrate more on parameter set 1. The defined goals of an optimization run are thus strongly dependent on the given constraints and thus the choice of the right input parameters is important.

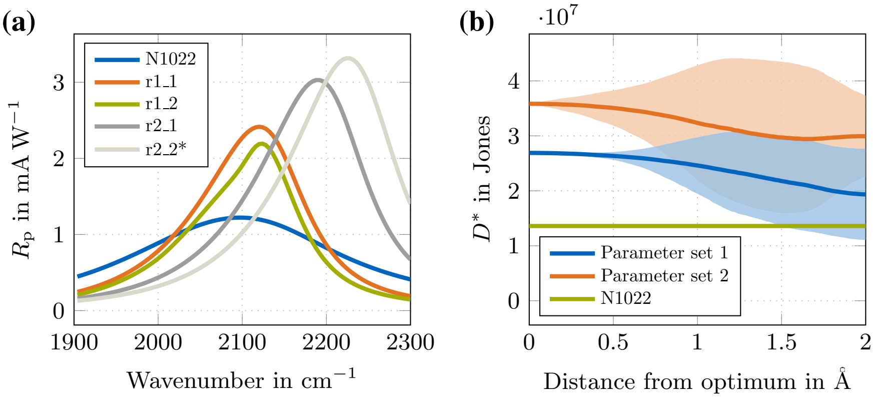

Fig. 7 **a** Simulated responsivity R_p of the nominal QCD structure N1022 and the optimized structures as a function of wavenumber at 300 K. **b** Sensitivity on the specific detectivity D^* over distance from the optimum of parameter set 1 (blue) and 2 (orange). The lines exhibit the mean specific detectivity and the shaded areas represent the 95% prediction interval obtained by samples from the trained GP. (Color figure online)

In this BO, we used the scattering model based on Fermi's golden rule, which does not consider e-e scattering. EMC simulations including e-e scattering exhibit similar results for the mid-infrared detector N1022, which confirms the validity of our optimization. For simulations of terahertz structures in the low temperature regime, e-e scattering becomes more important and has to be considered (Popp et al. 2020a).

Furthermore, we investigated the sensitivity of our optimization results on variations in the parameter set (Franckié and Faist 2020). These variations can arise through growth fluctuations in the fabrication. Starting from the optimal structure of each parameter set, a GP was trained to predict changes in the specific detectivity with respect to fluctuations in the well and barrier widths of the investigated layer sequence. By sampling ~ 500000 points, we can visualize the sensitivity of our model by plotting the predicted specific detectivity D^* over distance from the optimal structure r1_1 and r2_1, respectively. Here, the distance is the radius of a hypersphere in the three-dimensional parameter space. The results are illustrated in Fig. 7b and show the variation of specific detectivity D^* when diverging from the optimal values to a distance of 2 Å. Within a radius of 1 Å the variations of both structures are small, which ensures robustness against fluctuations. Even for longer distances up to 2 Å, both structures promise better results than obtained with the nominal structure N1022. As reported in literature, small deviations of the period thickness in the range of 1% to 2% can be accomplished with the modern molecular-beam-epitaxy technology (Bosco et al. 2019; Beere et al. 2005; Amanti et al. 2009). Applying this to the layer sequence $[w_1 b_1 w_2]$, possible deviations in the range -1.4 Å to 1.4 Å for this layer sequence can occur during the device growth. Within this tolerance our optimization results are still reasonable, and the designs r2_1 and r1_1 show promising alternatives to the nominal design N1022.

6 Conclusion

In this paper, the Monaco framework for simulation and optimization of quantum cascade devices is introduced. The Bayesian optimization approach for QCDs is based on Gaussian process regression and exhibits precise and robust optimization results for the tested QCD design N1022 at 300 K. Investigating two different parameter sets, the specific detectivity D^* of the nominal structure can be improved by a factor of $\sim 2-3$. The oscillator strength between ground level g and absorption level a_2 leads to a significantly increased absorption efficiency η_{abs}, thus resulting in peak responsivities 2-3 times higher than for the nominal structure N1022. Using a GP trained with the simulation results of the BO runs, we can make assumptions about the sensitivity of the optimized designs regarding fabrication tolerances. The optimized structures of both parameter sets appear to be quite robust against growth layer variations. For this optimization approach, we have used a scattering model based on Fermi's golden rule. For further optimizations, we will also use the self-consistent EMC model for the evaluation of QCD figures of merit, and compare them with the scattering rate approach used in this paper. Additionally, an interesting optimization approach to enhance the extraction efficiency, while keeping Johnson noise low, could be the variation of layer compositions at the border of two periods.

In summary, the Bayesian optimization algorithm proves to be an efficient tool for the optimization of QCDs, and can be useful for the design and optimization of on-chip applications for environmental sensing at elevated temperatures based on quantum cascade devices.

Acknowledgements The authors acknowledge financial support by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 820419 – Qombs Project ”Quantum simulation and entanglement engineering in quantum cascade laser frequency combs” (FET Flagship on Quantum Technologies).

Funding Open Access funding enabled and organized by Projekt DEAL. Funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 820419 – Qombs Project ”Quantum simulation and entanglement engineering in quantum cascade laser frequency combs” (FET Flagship on Quantum Technologies).

Declaration

Conflicts of interest The authors declare that they have no conflict of interest.

References

Amanti, M.I., Scalari, G., Terazzi, R., Fischer, M., Beck, M., Faist, J., Rudra, A., Gallo, P., Kapon, E.: Bound-to-continuum terahertz quantum cascade laser with a single-quantum-well phonon extraction/injection stage. New J Phys **11**(12), 125022 (2009)

Baumgartner, O., Stanojevic, Z., Schnass, K., Karner, M., Kosina, H.: VSP-A quantum-electronic simulation framework. J Comput Electron **12**(4), 701–721 (2013). https://doi.org/10.1007/s10825-013-0535-y

Beere, H., Fowler, J., Alton, J., Linfield, E., Ritchie, D., Köhler, R., Tredicucci, A., Scalari, G., Ajili, L., Faist, J., et al.: MBE growth of terahertz quantum cascade lasers. J Cryst Growth **278**(1–4), 756–764 (2005)

Bismuto, A., Terazzi, R., Hinkov, B., Beck, M., Faist, J.: Fully automatized quantum cascade laser design by genetic optimization. Appl Phys Lett **101**(2), 021103 (2012). https://doi.org/10.1063/1.4734389

Bosco, L., Franckié, M., Scalari, G., Beck, M., Wacker, A., Faist, J.: Thermoelectrically cooled THz quantum cascade laser operating up to 210 K. Appl Phys Lett **115**(1), 010601 (2019). https://doi.org/10.1063/1.5110305

Delga, A., Doyennette, L., Carras, M., Trinité, V., Bois, P.: Johnson and shot noises in intersubband detectors. Appl Phys Lett **102**(16), 163507 (2013). https://doi.org/10.1063/1.4803447

Dougakiuchi, T., Fujita, K., Hirohata, T., Ito, A., Hitaka, M., Edamura, T.: High photoresponse in room temperature quantum cascade detector based on coupled quantum well design. Appl Phys Lett **109**(26), 261107 (2016). https://doi.org/10.1063/1.4973582

Faist, J., Capasso, F., Sivco, D.L., Sirtori, C., Hutchinson, A.L., Cho, A.Y.: Quantum cascade laser. Science **264**, 553–556 (1994). https://doi.org/10.1126/science.264.5158.553

Franckié, M.: aftershoq. (2019) https://github.com/mfranckie/aftershoq/

Franckié, M., Faist, J.: Bayesian optimization of terahertz quantum cascade lasers. Phys Rev Appl **13**, 034025 (2020). https://doi.org/10.1103/PhysRevApplied.13.034025

Frazier, P. I.: A tutorial on Bayesian optimization. (2018) arXiv preprint arXiv:1807.02811

Gelbart, MA., Snoek, J., Adams, RP.: Bayesian optimization with unknown constraints. (2014) arXiv preprint arXiv:14035607

Gendron, L., Carras, M., Huynh, A., Ortiz, V., Koeniguer, C., Berger, V.: Quantum cascade photodetector. Appl Phys Lett **85**(14), 2824–2826 (2004). https://doi.org/10.1063/1.1781731

Giorgetta, F.R., Baumann, E., Graf, M., Yang, Q., Manz, C., Kohler, K., Beere, H.E., Ritchie, D.A., Linfield, E., Davies, A.G., et al.: Quantum cascade detectors. IEEE J Quantum Electron **45**(8), 1039–1052 (2009). https://doi.org/10.1109/JQE.2009.2017929

González, J., Dai, Z., Hennig, P., Lawrence, N.: Batch Bayesian optimization via local penalization. In: Artificial Intelligence and Statistics, pp 648–657 (2016)

GPyOpt: Gpyopt: A Bayesian optimization framework in Python. (2016) http://github.com/SheffieldML/GPyOpt

Harrer, A., Schwarz, B., Schuler, S., Reininger, P., Wirthmüller, A., Detz, H., MacFarland, D., Zederbauer, T., Andrews, AM., Rothermund, M., Oppermann, H., Schrenk, W., Strasser, G.: 4.3 μm quantum cascade detector in pixel configuration. Opt Express **24**(15):17041–17049, (2016) https://doi.org/10.1364/OE.24.017041

Hofstetter, D., Beck, M., Faist, J.: Quantum-cascade-laser structures as photodetectors. Appl Phys Lett **81**(15), 2683–2685 (2002)

Hofstetter, D., Graf, M., Aellen, T., Faist, J., Hvozdara, L., Blaser, S.:23 GHz operation of a room temperature photovoltaic quantum cascade detector at 5.35 μm. Appl Phys Lett **89**(6), 061119 (2006) https://doi.org/10.1063/1.2269408

Hofstetter, D., Giorgetta, F.R., Baumann, E., Yang, Q., Manz, C., Köhler, K.: Mid-infrared quantum cascade detectors for applications in spectroscopy and pyrometry. Appl Phys B **100**(2), 313–320 (2010). https://doi.org/10.1007/s00340-010-3965-2

Iotti, R.C., Ciancio, E., Rossi, F.: Quantum transport theory for semiconductor nanostructures: A density-matrix formulation. Phys Rev B **72**(12), 125347 (2005). https://doi.org/10.1103/PhysRevB.72.125347

Jirauschek, C.: Accuracy of transfer matrix approaches for solving the effective mass Schrödinger equation. IEEE J Quantum Electron **45**(9), 1059–1067 (2009). https://doi.org/10.1109/JQE.2009.2020998

Jirauschek, C.: Density matrix Monte Carlo modeling of quantum cascade lasers. J Appl Phys **122**, 133105 (2017). https://doi.org/10.1063/1.5005618

Jirauschek, C.: Universal quasi-level parameter for the characterization of laser operation. IEEE Photonics J **10**(4), 1503209 (2018). https://doi.org/10.1109/JPHOT.2018.2863025

Jirauschek, C., Kubis, T.: Modeling techniques for quantum cascade lasers. Appl Phys Rev **1**(1), 011307 (2014). https://doi.org/10.1063/1.4863665

Koeniguer, C., Dubois, G., Gomez, A., Berger, V.: Electronic transport in quantum cascade structures at equilibrium. Phys Rev B **74**, 235325 (2006). https://doi.org/10.1103/PhysRevB.74.235325

Levine, B.F., Choi, K.K., Bethea, C.G., Walker, J., Malik, R.J.: New 10 μm infrared detector using intersubband absorption in resonant tunneling GaAlAs superlattices. Appl Phys Lett **50**(16), 1092–1094 (1987). https://doi.org/10.1063/1.97928

Liu, H.: Dependence of absorption spectrum and responsivity on the upper state position in quantum well intersubband photodetectors. J Appl Phys **73**(6), 3062–3067 (1993). https://doi.org/10.1063/1.352989

Popp, J., Haider, M., Franckié, M., Faist, J., Jirauschek, C.: Monte Carlo modeling of terahertz quantum cascade detectors. In: 2020 XXXIIIrd General Assembly and Scientific Symposium of the International Union of Radio Science, IEEE, pp 1−3, (2020a). https://doi.org/10.23919/URSIGASS49373.2020.9232167

Popp, J., Haider, M., Franckié, M., Faist, J., Jirauschek, C.: Numerical optimization of quantum cascade detector heterostructures. In: 2020 International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD), IEEE, pp 1–2, (2020b) https://doi.org/10.1109/NUSOD49422.2020.9217784

Rasmussen, C.E.: Gaussian Processes in Machine Learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds.) Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science, vol 3176. Springer, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28650-9_4

Reininger, P., Schwarz, B., Detz, H., MacFarland, D., Zederbauer, T., Andrews, A.M., Schrenk, W., Baumgartner, O., Kosina, H., Strasser, G.: Diagonal-transition quantum cascade detector. Appl Phys Lett **105**(9), 091108 (2014). https://doi.org/10.1063/1.4894767

Riesch, M., Nguyen, T.D., Jirauschek, C.: bertha: Project skeleton for scientific software. PLOS ONE **15**(3), 1–12 (2020). https://doi.org/10.1371/journal.pone.0230557

Schwarz, B., Reininger, P., Detz, H., Zederbauer, T., Maxwell Andrews, A., Kalchmair, S., Schrenk, W., Baumgartner, O., Kosina, H., Strasser, G.: A bi-functional quantum cascade device for same-frequency lasing and detection. Appl Phys Lett **101**(19), 191109 (2012). https://doi.org/10.1063/1.4767128

Schwarz, B., Reininger, P., Ristanić, D., Detz, H., Andrews, A.M., Schrenk, W., Strasser, G.: Monolithically integrated mid-infrared lab-on-a-chip using plasmonics and quantum cascade structures. Nat Commun **5**(1), 1–7 (2014). https://doi.org/10.1038/ncomms5085

Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 2951–2959. Curran Associates Inc., Red Hook (2012)

Sugawara, M., Okazaki, N., Fujii, T., Yamazaki, S.: Conduction-band and valence-band structures in strained $\mathrm{In}_{1-x}\mathrm{Ga}_x\mathrm{As}$/InP quantum wells on (001) InP substrates. Phys Rev B **48**, 8102–8118 (1993). https://doi.org/10.1103/PhysRevB.48.8102

Vurgaftman, I., Meyer, J.R., Ram-Mohan, L.R.: Band parameters for III-V compound semiconductors and their alloys. J Appl Phys **89**(11), 5815–5875 (2001). https://doi.org/10.1063/1.1368156

Wacker, A., Lindskog, M., Winge, D.O.: Nonequilibrium Green's function model for simulation of quantum cascade laser devices under operating conditions. IEEE J Sel Topics in Quantum Electron **19**(5), 1–11 (2013). https://doi.org/10.1109/JSTQE.2013.2239613