SIAM J. APPL. MATH.
Vol. 79, No. 1, pp. 257–283

OPTIMIZATION OF A MULTIPHYSICS PROBLEM IN SEMICONDUCTOR LASER DESIGN*

LUKÁŠ ADAM†, MICHAEL HINTERMÜLLER‡, DIRK PESCHKA§, AND THOMAS M. SUROWIEC¶

Abstract. A multimaterial topology optimization framework using phase fields is suggested for the simultaneous optimization of mechanical and optical properties to be used in the development of optoelectronic devices. The technique provides a means of determining the cross section of the material alignments needed to create a sufficiently large strain profile within an optically active region of a photonic device. Based on the physical aspects of the underlying device, a nonlinear multiphysics model for the elastic and optical properties is proposed in the form of a linear elliptic partial differential equation (elasticity) coupled via the underlying topology to an eigenvalue problem of Helmholtz type (optics). The differential sensitivity of the displacement and eigenfunctions with respect to the changes in the underlying topology is investigated. After proving existence and optimality results, numerical experiments leading to an optimal material distribution for maximizing the strain in a Ge-on-Si microbridge are given. The presence of a net gain at low voltages for the optimal design is demonstrated by solving the steady-state van Roosbroeck (drift-diffusion) system, which proves the viability of the approach for the development of next-generation photonic devices.

Key words. optoelectronics, semiconductor laser, strained germanium microbridges, van Roosbroeck, phase field, design optimization, topology optimization, PDE-constrained optimization

AMS subject classifications. 35J60, 74S05, 35Q93, 49Q10, 90C90, 90C06, 78A60

DOI. 10.1137/18M1179183

1. Introduction.

The rapid miniaturization of microprocessors over the last four decades has been matched by a notable increase in computational performance. For the most part, these developments have followed Moore's law, which predicts a biennial doubling of components per integrated circuit. Nevertheless, there are physical limits to this trend and further improvements will require alternative and innovative approaches. One promising solution is found in silicon photonics, which integrates optical and electronic components into a single microchip. The ultimate goal here is to use optical interconnects to provide faster data transfer both between and within microchips in order to avoid the limitations of electrical wiring; cf., e.g., [25]. This paper is inspired by the promising approach of using strained germanium (Ge) as the optically active medium for an edge-emitting laser, which serves as the light source for silicon photonics; cf. [59, 18, 58, 65].

*Received by the editors April 6, 2018; accepted for publication (in revised form) November 12, 2018; published electronically February 14, 2019.
http://www.siam.org/journals/siap/79-1/M117918.html
Funding: This research was partially carried out within the framework of the research center MATHEON through projects OT1 and OT8 funded by the Einstein Center for Mathematics Berlin. The first author was supported by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (grant 2017ZT07X386).

†Southern University of Science and Technology, 1088 Xueyuan Avenue, Shenzhen, Guangdong, China, 518055, and ÚTIA, Czech Academy of Sciences, Pod Vodárenskou věží 4, 18208, Prague, Czech Republic (adam@utia.cas.cz).
‡Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, and Weierstrass Institute, Mohrenstr. 39, 10117 Berlin, Germany (hint@math.hu-berlin.de, michael.hintermueller@wias-berlin.de).
§Weierstrass Institute, Mohrenstr. 39, 10117 Berlin, Germany (peschka@wias-berlin.de).
¶Philipps-Universität Marburg, FB12 Mathematik und Informatik, Hans-Meerwein-Str. 6, Lahnberge, 35032 Marburg, Germany (surowiec@mathematik.uni-marburg.de).

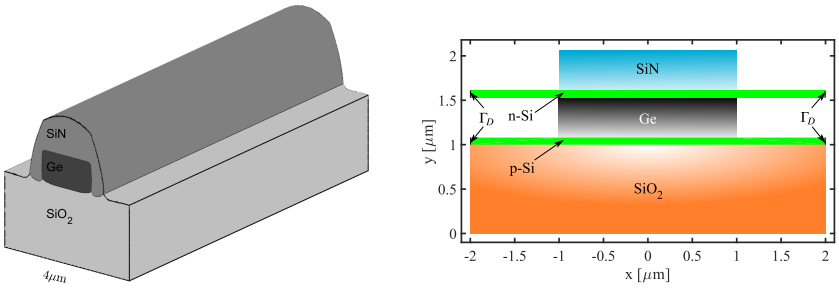

FIG. 1. *(left) A possible prototype strained photonic device exhibiting the microbridge geometry. This configuration was determined in* [1]. *(right) Cross section of a Ge-on-Si microbridge with material distribution and contacts.*

We recall that inside a semiconductor, electrons are restricted to distinct energy levels called energy bands seen as functions of the electron momentum p. Given the electrochemical potential of the system, i.e., the Fermi level E_F, the first band above (below) E_F with lowest (highest) energy is the conduction (valence) band. Letting E_c be the minimum along the conduction band and E_v the maximum along the valence band, we define the band gap by $E_g = E_c - E_v$. This is the minimal amount of energy needed to move an electron from the valence band into the conduction band. If E_c and E_v occur at the same point $p^\star$, then we have a direct band gap semiconductor and otherwise an indirect band gap semiconductor. For example, both silicon (Si) and Ge have indirect band gaps.

Direct band gap semiconductors can be used for light sources in optoelectronic devices, since electrons may pass directly between valence and conduction bands at the same momentum $p^\star$ and potentially emit a photon. This emission is stimulated by another incident photon. When in the process of moving from the conduction band to the valence band an electron emits a photon, we speak of radiative recombination. The newly generated photons travel inside the cavity, in our case a strip of Ge, as in Figure 1 (left), until they are either emitted from the edge or absorbed after traveling for a certain distance within the cavity, a process described by optical gain and losses. The latter is strongly suppressed on its own for indirect band gap materials as the difference in momenta needs to be overcome by a particle-like phenomenon (lattice vibration) known as a phonon. Thus, stimulated emission is unlikely and it would appear that both Si and Ge are unsuited to create an integrated light source.

However, the band structure of Ge can be significantly altered by introducing impurities (dopants) to change the charge carrier concentration, i.e., so-called doping, and even more so by using mechanical strains [24]. Due to the particular band-structure of Ge, tensile strains of 1% to 2% are sufficient to turn Ge into a direct band gap material and drastically enhance stimulated emission [24]. However, it is believed that much lower strains are sufficient to build a functioning laser; see, e.g., [20]. This is highly advantageous from a manufacturing perspective as Ge can be manipulated similarly to Si using standard lithography techniques. For example, the tensile strain can be induced by a silicon nitride (SiN) layer, whereas the photon emission is stimulated via a current introduced through the contacting layers (Si-n, Si-p) (cf. Figure 1 (right)); the silicon dioxide (SiO_2) is merely a substrate.

Whereas an optimal doping profile, i.e., an optimal distribution of dopants, can be determined by optimizing charge transport using nonlinear drift-diffusion models [39, 51], an optimal material configuration, used to create tensile strain in the Ge,

can be found by using techniques from topology optimization applied to linear elastic materials [1].

Various engineering studies have focused on the production of Ge devices, where the light emission is improved by maximizing the strain. This has led to a variety of device designs including suspended bridges [58, 41] and discs [29]. Moreover, the corresponding photoluminescence spectra of these empirical designs support the improvement of the optical properties. However, the manufacturability of these new devices using standard fabrication processes is still active research; see, e.g., [20]. The feasibility of both optically [45] and electrically [18] pumped lasers based on a tensile strained Ge layer has been demonstrated. However, these devices suffer from considerable unwanted Joule heating and ultimately device failure.

The main culprit for this unwanted heating is the overlap of the optically active area with the contacting layers [24]. Note that the optically active area is primarily determined by the bulk of the support of the first fundamental mode (first eigenfunction) of the laser and would therefore be ideally confined inside the Ge.

The final essential component to ensuring that the Ge-on-Si semiconductor device can serve as a light source is the so-called optical gain (g) of the laser, i.e., the ability of the laser to amplify photons by stimulated emission. The gain depends on carrier concentrations and photon energy. For an indirect band gap material such as Ge, the rate of stimulated emission encoded in g is naturally very low and strongly depends on the size of the direct band gap. We note that the smaller the gap, the less energy needed to push electrons into the conduction bands. This allows the laser to operate at lower temperatures and somewhat suppresses the loss mechanisms.

The main parameter influencing gain g is the band gap E_{g}, which itself is a function of the strain e. To see the latter, we refer to [21, Chapter 4.5], where the band energies E_c, E_v and band gap E_g are put into a relationship with strain via a deformation potential $\mathcal{D}$ given by $E_g(e) = E_{g,0} + \mathcal{D} : e(\mathbf{u})$, where $E_{g,0}$ is the band gap of the undeformed crystal and $e(\boldsymbol{u})$ the strain induced by a displacement $\boldsymbol{u}$.

Since the strain distribution has a larger effect on the gain than the dopants, we only consider the mechanical and optical device properties in the forward system of the optimization problem. Nevertheless, the net gain, i.e., photon emission minus losses, is the essential quantity to be maximized.

As a numerical proof-of-concept, we provide a study of the electronic properties of the optimal device at the end of this paper. These results corroborate the observations in the photonics literature, e.g., [24, 58], and thus justify our approach.

Furthermore, the improved strain leads to an improved gain only when the support of the first fundamental mode and large strain regions coincide, a goal we previously phrased as "overlap engineering" [52]. Summarizing these facts, we arrive at the following.

Goal. *Determine a device topology, which simultaneously ensures the support of the first fundamental mode, i.e., the optical cavity, is confined inside the Ge and the strain is maximized within the optical cavity.*

Since the proposed device is static, it is possible to create a permanent strain field only through the shape and topology of the device, where each of the materials supplies a certain amount of strain following the manufacturing process. More specifically, during the manufacturing process, the lattices of the heated components align. However, the different thermal expansion coefficients of the various components result in residual forces along the contact boundaries as the materials "relax" into their final shapes. This is particularly the case for SiN (stressor), Ge (cavity), and SiO_2

(wafer). Therefore, we need to find an optimal composition and placement of the various necessary materials in order to construct a Ge-on-Si laser.

Some ideas for optimal material configurations based on existing empirical, experimental, and analytical studies can be found, e.g., in [45, 18, 20, 52, 53]. We also mention our recent related work [1], in which the optical cavity is assumed to be fixed. The underlying modeling assumption in all of these studies is the usage of a so-called "microbridge" geometry (cf. Figure 1), which can be created by standard manufacturing techniques. As in [52, 1], we again focus on a cross section of an edge-emitter as shown in Figure 1. In the longitudinal direction we assume translation invariance, as indicated in Figure 1. We note here that multimaterial and multidisciplinary topology optimization approaches that take into consideration thermoelastic or piezoelectric properties and their relation to the underlying topologies have been considered in many works; see, e.g., [56, 57]. However, as we will see below in the modeling section, these are fundamentally different applications with distinct goals.

The rest of the paper is organized as follows. In section 2, we motivate the usage of a phase-field approach for the topology optimization. In section 3, we introduce the underlying multiphysics model that appears in the topology optimization problem. In addition, we briefly detail the time-dependent drift-diffusion system, which models the transport of electrons and holes in the device. Afterward, we introduce the optimization framework in section 4, which includes a rigorous analysis of the topology-to-eigenmode mapping in section 4.3. In section 5, we discuss the numerical solution and necessary structural assumptions. Based on our theoretical results, we provide numerical optimization results in section 6, which yield an optimal configuration of materials in the microbridge. Using the optimal configuration, we demonstrate the electronic and optical properties of such a design in section 6.5. We conclude with section 7.

Finally, our notation is more or less standard for PDE-constrained and topology optimization. Nevertheless, we refer the reader to the well-known monographs [3] for Lebesgue and Sobolev spaces, [44, 62, 40] for PDE-constrained optimization, and [34, 4, 11, 49] for a thorough treatment of topology optimization.

2. A phase-field approach for the design parameters. Throughout the text, all functions are assumed to be defined on a fixed domain Ω, which represents the cross section of the microbridge. It is assumed that Ω has a sufficiently smooth boundary $\partial\Omega$. Furthermore, $\{\Omega_i\}_{i=1}^N$ denotes a material distribution, which partitions the domain Ω. Ideally, the partition would be represented by distributed parameters $\{\varphi_i\}_{i=1}^N$, where φ_i serves as the characteristic function for Ω_i. In such a case, we might take $\boldsymbol{\varphi} := (\varphi_1, \dots, \varphi_N) \in BV(\Omega; \{0,1\}^N)$ (a vector of functions of bounded variation (BV) taking discrete values in $\{0,1\}$) along with the condition that $\sum_{i=1}^N \varphi_i = 1$ for almost every (a.e.) $x \in \Omega$. In order to ensure that the sets $\Omega_i := \{\varphi_i = 1\}$ have finite (relative) perimeter $P(\Omega_i, \Omega)$, which is needed to rule out pathological designs and facilitate the mathematical treatment, it suffices that the total variation term $\sum_{i=1}^N TV(\varphi_i, \Omega)$ is finite. Here, we use the total variation of a function $u : \Omega \to \mathbb{R}$ as defined by

$$TV(u,\Omega) = \sup\left\{ \int_\Omega u \operatorname{div} \psi \, \mathrm{d}x : \psi \in C_c^1(\Omega;\mathbb{R}^n),\ |\psi(x)|_{\ell^\infty} \le 1 \text{ a.e. } x \in \Omega \right\};$$

see, e.g., [6]. Note that by [6, Definition 3.35] $P(\Omega_i, \Omega) := TV(\varphi_i, \Omega)$.

Finding $\{\Omega_i\}_{i=1}^N$ that would provide a device satisfying our stated goals would lead to a computationally intractable combinatorial problem. As a remedy, one could

relax the integrality condition on each φ_i, as in [16], and attempt to regain the integrality through other means. Such an approach typically depends on structural assumptions. As in [1, 13, 68, 17, 60], we use a phase-field approach in which we have $\boldsymbol{\varphi} \in H^1(\Omega;\mathbb{R}^N) \subset BV(\Omega,\mathbb{R}^N)$. Here, $H^1(\Omega,\mathbb{R}^N)$ is the space of all vector fields in $\mathbb{R}^N$ with components in the Sobolev space $H^1(\Omega)$; see, e.g., [3]. We utilize this convention throughout the text. Enforcing approximate integrality of $\boldsymbol{\varphi}$ can then be achieved by considering the Ginzburg–Landau-type energy functional

$$f_{\mathrm{GL}}(\boldsymbol{\varphi},\varepsilon) := \int_\Omega \left(\frac{\varepsilon}{2} \nabla\boldsymbol{\varphi} : \nabla\boldsymbol{\varphi} + \frac{1}{2\varepsilon} \boldsymbol{\varphi}\cdot(1-\boldsymbol{\varphi}) \right) \mathrm{d}\mathbf{x} + i_{\mathcal{G}}(\boldsymbol{\varphi}) \tag{2.1}$$

in the associated topology optimization problem. Here, the matrix product is understood as $A : B = \sum_i \sum_j a_{ij} b_{ij}$ and $i_{\mathcal{G}}$ is the usual indicator functional for the well-known Gibbs simplex (a closed convex set)

$$\mathcal{G} := \left\{ \boldsymbol{\varphi} \in H^1(\Omega;\mathbb{R}^N) \,|\, \boldsymbol{\varphi} \geq 0,\ \varphi_1 + \cdots + \varphi_N = 1,\ \text{a.e. in } \Omega \right\}, \tag{2.2}$$

i.e., $i_{\mathcal{G}}(\boldsymbol{\varphi}) = 0$ if $\boldsymbol{\varphi} \in \mathcal{G}$ and $i_{\mathcal{G}}(\boldsymbol{\varphi}) = +\infty$ otherwise. Note that the nonconvex integrand $\frac{1}{2\varepsilon}\boldsymbol{\varphi}\cdot(1-\boldsymbol{\varphi})$ attempts to force pure phases, whereas the first part in (2.1) introduces H^1-regularity of $\boldsymbol{\varphi}$ into the problem.

We note that as $\varepsilon \to 0$, $f_{\mathrm{GL}}(\cdot,\varepsilon)$ Γ-converges to a set functional that is related to the term $\sum_{i=1}^N TV(\varphi_i,\Omega)$ with some modifications. This can be shown by modifying and combining several arguments from [48] and [9]. A detailed discussion would go beyond the scope of this paper.

3. The underlying multiphysics problem. In this section, we introduce the underlying multiphysics (forward) problem as well as the drift-diffusion system, which describes the electronic behavior of the device. The forward problem is a nonlinearly coupled system of linear PDEs, which comprises two kinds of physics: elasticity and optics. The elastic properties follow a standard model of linear elasticity that takes into account eigenstrain and thermal prestress terms. This represents our mathematical description for the interfacial residual forces discussed in the introduction.

The model is dependent on the distributed material parameter $\boldsymbol{\varphi}$, with components $\varphi_1,\ldots,\varphi_N$ or sometimes φ_{SiN}, φ_{Ge}, $\varphi_{\mathrm{SiO_2}}$, φ_{air} for emphasis on the actual material components and their effects on the device. These distributed parameters will act as the design/decision variables in our optimization framework. We focus on optimizing the first/fundamental (eigen)mode of the device, since it has the largest effect on the lasing properties of the Ge-on-Si microbridge. This is done by using a Helmholtz equation, which depends on the material parameters $\boldsymbol{\varphi}$. We provide further physical and mathematical motivations for these models in the subsections below. Finally, in contrast to a standard multiphysics problem, the coupling arises here via the objective function and the fact that the material distribution appears in (E($\boldsymbol{\varphi}$)) and (3.3) and represents an optimization variable.

3.1. Elasticity. Given $\boldsymbol{\varphi} \in \mathcal{G}$, we consider the following model of elasticity, where the solution is a displacement mapping $\mathbf{u} : \Omega \to \mathbb{R}^2$:

$$\text{(E}(\boldsymbol{\varphi})\text{)} \qquad \begin{aligned} -\mathrm{div}\,[\mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{u}) - F(\boldsymbol{\varphi})] &= 0 \quad \text{in} \quad \Omega, \\ \boldsymbol{u} &= 0 \quad \text{on} \quad \partial\Omega. \end{aligned}$$

Here, $e(\mathbf{u}) := \frac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^\top)$ is the symmetric strain of $\mathbf{u}$, $\mathbb{C}(\boldsymbol{\varphi})$ is a fourth-order tensor, and

$$F(\boldsymbol{\varphi}) := e_0(\varphi_{\mathrm{SiO_2}} - 1)\mathbb{C}(\boldsymbol{\varphi})I_{2\times 2} - \sigma_0\varphi_{\mathrm{SiN}}I_{2\times 2} \tag{3.1}$$

incorporates the effect of the eigenstrain e_0 generated by thermal relaxation of Ge on SiO_2 and the (pre)stress σ_0 generated by SiN as discussed in the introduction; $I_{2\times 2}$ is the identity matrix on $\mathbb{R}^{2\times 2}$.

This modeling choice for F is aimed at driving the Ge lattice constant into a tensile region as is needed to alter the band structure. We recall that the lattice constant of a cubic crystal structure such as Ge is the length of a unit cell. For small uniform strains from (E($\boldsymbol{\varphi}$)), this lattice constant can be defined by $a(\mathbf{x}) = a_{\text{bulk}}(1 + e(\boldsymbol{u})(\mathbf{x}) - e_0)$, where a_{bulk} is the lattice constant of unstrained Ge and $a(\mathbf{x}) > a_{\text{bulk}}$ is desired; cf. [19]. The Dirichlet boundary condition implies that the device remains fixed on $\partial\Omega$.

We invoke the following smoothness and ellipticity assumptions throughout.

Assumption A1. $\mathbb{C}$ is a Nemytskii/superposition operator (cf. [7]), induced by a tensor-valued mapping $\widehat{\mathbb{C}} : \mathbb{R}^N \to \mathbb{R}^{2\times 2\times 2\times 2}$ such that for $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$, $\mathbb{C}(\boldsymbol{\varphi})(x) = \widehat{\mathbb{C}}(\boldsymbol{\varphi}(x))$ a.e. on Ω. Moreover, it satisfies that

(i) there exist $c_2 > c_1 > 0$ such that for every $\phi \in \mathbb{R}^N$ and $E_1, E_2 \in \mathbb{R}^{2\times 2} \setminus \{0\}$ we have

$$c_1\|E_1\|^2_{\mathbb{R}^{2\times 2}} \leq \widehat{\mathbb{C}}(\phi)E_1 : E_1, \qquad \widehat{\mathbb{C}}(\phi)E_1 : E_2 \leq c_2\|E_1\|_{\mathbb{R}^{2\times 2}}\|E_2\|_{\mathbb{R}^{2\times 2}};$$

(ii) $\widehat{\mathbb{C}}$ is globally Lipschitz and continuously differentiable with globally Lipschitz derivative.

Consequently, we have the following regularity and sensitivity result for the topology-to-displacement map $S_u(\boldsymbol{\varphi})$. This is essential for the proof of existence of an optimal solution and derivation of optimality conditions for the associated topology optimization problem. In addition, it is needed for the development of gradient-based numerical optimization methods.

PROPOSITION 3.1 (cf. [1]). *Let* A1 *hold. Then there exists $p > 2$ such that for every $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ the unique solution $\boldsymbol{u}$ of* (E($\boldsymbol{\varphi}$)) *lies in $W_0^{1,p}(\Omega, \mathbb{R}^2)$. Finally, the solution mapping $S_u : H^1(\Omega, \mathbb{R}^N) \to W_0^{1,p}(\Omega, \mathbb{R}^2)$, which maps $\boldsymbol{\varphi} \mapsto \boldsymbol{u}$, is continuously Fréchet differentiable. The directional derivative of S_u at $\boldsymbol{\varphi}$ in direction $\delta\boldsymbol{\varphi}$ is given by $S_u'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} = \boldsymbol{q}$, where $\boldsymbol{q} \in H_0^1(\Omega, \mathbb{R}^2)$ is the weak solution of the sensitivity equation*

$$\int_\Omega \mathbb{C}(\boldsymbol{\varphi})e(\boldsymbol{q}) : e(\boldsymbol{v})\mathrm{d}\mathbf{x} = -\int_\Omega [\mathbb{C}'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}]e(\boldsymbol{u}) : e(\boldsymbol{v})\mathrm{d}\mathbf{x} + \int_\Omega F'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi} : e(\boldsymbol{v})\mathrm{d}\mathbf{x} \tag{3.2}$$

for all $\boldsymbol{v} \in H_0^1(\Omega, \mathbb{R}^2)$.

Here, $W_0^{1,p}(\Omega, \mathbb{R}^2)$ is the Sobolev space of two-dimensional vector fields with components in $W_0^{1,p}(\Omega)$. In addition, we note that $p > 2$ in Proposition 3.1 ensures via the Sobolev embedding theorem that $\boldsymbol{u}$ is a continuous vector field over $\overline{\Omega}$. This is useful in the existence proof below.

3.2. Optics. As stated above, we focus our attention on finding a topology that confines the bulk of the support of the fundamental mode within the Ge. In this sense, we assume that the governing optical behavior of the device can be modeled by the following $\boldsymbol{\varphi}$-dependent Helmholtz-type eigenvalue problem in (Θ, λ):

$$\begin{aligned} -\Delta\Theta - g(\boldsymbol{\varphi})\Theta &= \lambda\Theta \qquad \text{in} \quad \Omega, \\ \Theta &= 0 \qquad \text{on} \quad \partial\Omega. \end{aligned} \tag{3.3}$$

The quantity $g(\boldsymbol{\varphi})$ is a topology-dependent term related to the gain of the laser. See section 1 for more information concerning the role of gain in the overall optimization. A full discussion this term can be found in [52].

Here, we assume that the eigenfunction decays exponentially fast approaching the boundary, so that the homogeneous Dirichlet boundary condition on the outer boundary $\partial\Omega$ does not influence (Θ, λ) significantly. This is justified for certain $g(\boldsymbol{\varphi})$ and the eigenmode corresponding to the smallest eigenvalue, the latter often being the most relevant mode for an edge-emitting laser. In particular, we may assume that the properties of air and SiO_2 are such that Θ is effectively zero on these parts of the domain. This is why we focus our discussion primarily on λ_1, the smallest eigenvalue of $-[\Delta + g(\boldsymbol{\varphi})]$, and Θ_1, the corresponding eigenfunction, with $(\Theta_1, \Theta_1) = 1$. Here and below, $(\cdot, \cdot)$ represents the usual $L^2(\Omega)$-inner product given by $(u, v) = \int_\Omega u \cdot v \, \mathrm{d}x$. This leads to the following problem:

(H($\boldsymbol{\varphi}$)) Find the first eigenvalue λ_1 and
corresponding positive eigenfunction Θ_1 of (3.3) with $(\Theta_1, \Theta_1) = 1$.

We henceforth drop the subscripts, whenever it is clear in context. Note that Ω needs to be connected to ensure that λ_1 has multiplicity one; see [35, Remark 1.2.4].

For this model, we make the following standing assumption throughout.

Assumption A2. g is a superposition operator induced by $\hat{g} : \mathbb{R}^N \to \mathbb{R}$ such that $\hat{g}(\boldsymbol{\varphi}(x)) = (g(\boldsymbol{\varphi}))(x)$ a.e. on Ω. Moreover, $|\hat{g}|$ is bounded by M, and $\hat{g}$ is globally Lipschitz with modulus $L > 0$ and continuously differentiable with globally Lipschitz derivative.

We remark on several additional properties implied by A2. First, the smallest eigenvalue in (H($\boldsymbol{\varphi}$)) has multiplicity one and the corresponding eigenfunction can be chosen to be positive almost everywhere; see [31, Theorem 8.38] or [35, Theorem 1.2.5]. Assumption A2 also implies that $g : L^p(\Omega, \mathbb{R}^N) \to L^p(\Omega)$ is globally Lipschitz with modulus L and $g : L^{2p}(\Omega, \mathbb{R}^N) \to L^p(\Omega)$ is continuously differentiable with global Lipschitz derivative for all $p \in [1, \infty]$; see [32]. Moreover, $\|g(\boldsymbol{\varphi})\|_{L^\infty(\Omega)} \leq M$ for all $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$.

Here, we emphasize that $g(\boldsymbol{\varphi})$ is spatially dependent. Thus, the spectrum of $-[\Delta + g(\boldsymbol{\varphi})]$ is not merely the shifted spectrum of the Laplacian. Nevertheless, $|g(\boldsymbol{\varphi})|$ is uniformly bounded, independently of $\boldsymbol{\varphi}$. Consequently, we may take some fixed $c > M$ and consider the equivalent problem

$$(\mathrm{H}_c) \qquad \begin{aligned} -\Delta\Theta + (c - g(\boldsymbol{\varphi}))\Theta &= (c + \lambda)\Theta \qquad \text{in} \quad \Omega, \\ \Theta &= 0 \qquad\qquad\quad \text{on} \quad \partial\Omega. \end{aligned}$$

Indeed, the operators $-[\Delta + g(\boldsymbol{\varphi})]$ and $-[\Delta + (g(\boldsymbol{\varphi}) - c)]$ have the same eigenfunctions corresponding to the same eigenvalues shifted by c. Therefore, we may work with the uniformly elliptic bounded linear operator $-[\Delta + (g(\boldsymbol{\varphi}) - c)]$, which allows us to apply elliptic theory. We postpone the sensitivity analysis of the topology-to-eigenmode mapping $\boldsymbol{\varphi} \mapsto \Theta$, denoted by $S_\Theta(\boldsymbol{\varphi})$, until after we state the optimization problem.

3.3. Electronics. As discussed in the introduction, the optimization procedure needs to alter the electronic properties of Ge in order to ensure sufficient positive net gain, preferably at the lowest possible currents. We emphasize here that the model for the electronic behavior is not directly part of the optimization process. However, we will verify the success of our approach at the end of the paper by simulating the stationary carrier densities associated with an empirical design versus our optimal design.

The following drift-diffusion system forms the so-called van Roosbroeck system

$$-\mathrm{div}\,(\varepsilon_0\varepsilon_\mathrm{r}\nabla\phi) = q(C_\mathrm{dop}+p-n), \tag{3.4a}$$

$$\dot{n} - q^{-1}\mathrm{div}\,(\mathbf{j}_n) = -R_\mathrm{net}, \tag{3.4b}$$

$$\dot{p} + q^{-1}\mathrm{div}\,(\mathbf{j}_p) = -R_\mathrm{net}, \tag{3.4c}$$

which was introduced for semiconductors in [63]; see also [47]. Here, ϕ is the electrostatic potential, ε_0 is the vacuum permittivity and ε_r is the relative permittivity, n and p are the concentration of electrons and holes, q is the elementary charge, and C_dop is the doping profile. The electron and hole fluxes $\mathbf{j}_n$ and $\mathbf{j}_p$ are defined by

$$\mathbf{j}_n = -q\mu_n n\nabla\phi + qD_n\nabla n, \qquad \mathbf{j}_p = -q\mu_p p\nabla\phi - qD_p\nabla p, \tag{3.5}$$

where D_n, D_p denote the diffusion constant of electrons and holes and μ_n, μ_p are the corresponding mobilities, where for $\alpha = n, p$ we have $D_\alpha/\mu_\alpha = gk_BT/q$, where $g \equiv 1$ for Boltzmann statistics, k_B is Boltzmann's constant, and T is temperature. The remaining function R_net is the generation-recombination rate.

In order to solve for stationary solutions of (3.4) and ensure n, p are positive, we employ the following transformation of the charge carrier densities n, p to the so-called quasi-Fermi potentials ϕ_n, ϕ_p:

$$n = N_cF\left(\tfrac{q(\phi-\phi_n)-E_\mathrm{c}}{k_\mathrm{B}T}\right), \qquad p = N_vF\left(\tfrac{q(\phi_p-\phi)+E_\mathrm{v}}{k_\mathrm{B}T}\right), \tag{3.6}$$

where $F(\eta) = \exp(\eta)$ for Boltzmann distributions or $F(\eta) = F_{3/2}(\eta)$ the complete Fermi–Dirac integral with index 3/2 for Fermi–Dirac distributions; N_c, N_v are the material dependent effective density of states. Now, g in D_α above is $g = F/F'$. The choice of F influences the choice of R_net. In the general case, we have

$$R_\mathrm{net} = \left(1-\exp\left(\tfrac{q}{k_\mathrm{B}T}(\phi_n-\phi_p)\right)\right)\frac{np}{\tau_p(n+n_i)+\tau_p(p+n_i)}.$$

For Shockley–Read–Hall recombination terms, see also [54]. In the Boltzmann situation this expression reduces to the well-known form [21].

Most importantly, given $E_g = E_c - E_v$ as well as $E_g = E_{g,0} + \mathcal{D} : e(\mathbf{u}(\boldsymbol{\varphi}))$, (3.6) provides the final link between the topology $\boldsymbol{\varphi}$, the strain $e(\boldsymbol{u})$, the band energies E_c, E_v, and the carrier densities n, p. These all feed into the formula for calculating optical gain g, which we calculate and plot in section 6.5 for the optimal design.

4. The optimization framework. We now derive the optimization problem. We start by introducing the objective function and, following a sensitivity study, we prove existence of solutions and derive optimality conditions.

4.1. Objective function. We identify here a class of objective functions that quantify the goal of maximizing the tensile strain inside the optical cavity. In contrast to [1], we do not consider the optical cavity to be fixed. Instead, we assume that the optical cavity is explicitly determined by $\boldsymbol{\varphi}_\mathrm{Ge}$.

The underlying physics of the optical gain described in section 1 motivates our approach to minimize the functional

$$-\int_\Omega \varphi_\mathrm{Ge}\Theta^2\mathcal{D} : e(\mathbf{u})\mathrm{d}\mathbf{x} = -\int_\Omega j(\boldsymbol{\varphi},\Theta)\mathrm{tr}\, e(\mathbf{u})\,\mathrm{d}\mathbf{x},$$

where we assume that the deformation potential is diagonal, i.e., $\mathcal{D} = DI_{2\times 2}$ and in our case we have $j(\boldsymbol{\varphi}, \Theta) = \varphi_{\mathrm{Ge}}\Theta^2 D$. Recall, in particular, the general relation: $E_{\mathrm{g}}(e) = E_{\mathrm{g},0} + \mathcal{D} : e(\mathbf{u})$.

Note that $\mathcal{D}$ contains material parameters. However, since $\Theta^2 D$ is scaled by φ_{Ge}, which is a relatively smooth approximation of the indicator function for the subset of Ω corresponding to the Ge concentration, we need only consider the values of D for Ge.

As observed in [52] and discussed in section 1, we wish to maximize the region of overlap corresponding to the bulk of support for Θ^2 and the region of high tensile strain in Ge. At an optimal configuration, we expect the nonnegative bilinear relationship $\varphi_{\mathrm{Ge}}\Theta^2$ and $\operatorname{tr} e(\boldsymbol{u})$ to favor large overlap of $\operatorname{supp}\varphi_{\mathrm{Ge}}$ and $\operatorname{supp}\Theta^2$ along with deformations for which $\operatorname{tr} e(\boldsymbol{u})$ is positive on average on $(\operatorname{supp}\varphi_{\mathrm{Ge}}) \cap (\operatorname{supp}\Theta^2)$. We henceforth denote the objective by

$$J(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta) := -\int_\Omega j(\boldsymbol{\varphi}, \Theta)\operatorname{tr} e(\boldsymbol{u})\mathrm{d}\mathbf{x}. \tag{4.1}$$

We allow j to belong to a wide class of functions and make the following assumption.

Assumption A3. j is a superposition operator induced by a polynomial function $\hat{j} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ such that $\hat{j}(\boldsymbol{\varphi}(x), \Theta(x)) = (j(\boldsymbol{\varphi}, \Theta))(x)$ a.e. on Ω.

By admitting higher-order polynomials for $\hat{j}$, we can potentially emphasize regions where $\Theta(x)$ is large.

4.2. The optimization problem. Combining the objectives, constraints, and forward problems from the discussions above, we arrive at the optimization problem:

$$\begin{aligned} &\min\ -\int_\Omega j(\boldsymbol{\varphi}, \Theta)\operatorname{tr} e(\boldsymbol{u})\mathrm{d}\mathbf{x} + \alpha f_{\mathrm{GL}}(\boldsymbol{\varphi}, \varepsilon) \text{ over } (\boldsymbol{\varphi}, \boldsymbol{u}, \Theta, \lambda) \in \mathcal{X} \\ &\text{s.t.}\quad \boldsymbol{u} \text{ solves } (\mathrm{E}(\boldsymbol{\varphi}));\ (\Theta, \lambda) \text{ solves } (\mathrm{H}(\boldsymbol{\varphi})). \end{aligned} \tag{4.2}$$

Here, $\alpha > 0$ is a regularization parameter, the space $\mathcal{X}$ represents the Cartesian product $\mathcal{X} := \mathcal{G}_{ad} \times H_0^1(\Omega; \mathbb{R}^2) \times H_0^1(\Omega) \times \mathbb{R}$, and $\mathcal{G}_{ad} := \{\boldsymbol{\varphi} \in \mathcal{G} |\ \varphi_i = 1 \text{ a.e. on } \Pi_i,\ i = 1, \ldots, N\}$ combines the Gibbs simplex (2.2) and the requirement that material i must be present on $\Pi_i \subset \Omega$. The Π_i may arise due to manufacturing requirements or physical limitations, e.g., some SiN must be above the Ge and SiO_2 must serve as the substrate. We henceforth impose the following assumptions.

Assumption A4. $\Omega \subset \mathbb{R}^2$ and $\Pi_i \subset \Omega$ are open, connected, and bounded sets with Lipschitz boundary and Π_i are strictly separable, i.e., $\operatorname{cl}\Pi_i \cap \operatorname{cl}\Pi_j = \emptyset$ $(i \neq j)$.

The norm on $H_0^1(\Omega)$ is given by $\|u\|^2_{H_0^1(\Omega)} := \int_\Omega |\nabla u|^2\ \mathrm{d}x$. The usual duality pairing between $H_0^1(\Omega)$ and its topological dual $H^{-1}(\Omega)$ will be denoted by $\langle \cdot, \cdot \rangle$.

4.3. The topology-to-eigenmode mapping S_Θ. In this section, we perform a sensitivity analysis for the Helmholtz equation $(\mathrm{H}(\boldsymbol{\varphi}))$. The Lipschitz continuity derived in Lemma 4.1 is necessary for the existence result in Proposition 4.3, whereas the differentiability result in Theorem 4.2 is needed for the first-order necessary optimality conditions in Theorem 4.4. The latter are subsequently used for numerical experiments. Obviously any results providing explicit derivative formulae are ultimately useful in adjoint-based solution algorithms.

Though it is possible that further eigenvalues and eigenfunctions may also be of interest, the nontrivial multiplicity of even the second eigenvalue vastly complicates

any differential sensitivity analysis; see [22] for a method of minimizing eigenvalues with nontrivial multiplicity in the context of topology optimization of mechanical structures. Even with this choice there are still some challenges. For example, it is not possible to write (3.3) or its equivalent formulation (4.4) (below) as an equation (H($\boldsymbol{\varphi}$)) because the former allows for all eigenvalues. Later in the proof of Theorem 4.2 we show that this is possible at least locally around the principal eigenvalue.

We recall that due to $|g(\boldsymbol{\varphi})(x)| \leq M$ for a.e. $x \in \Omega$ independent of $\boldsymbol{\varphi}$, it is possible to shift the operators and obtain a simpler but equivalent eigenvalue problem. Choosing $c > M$, we make the operator on the left-hand side of (4.3) below elliptic,

$$(-\Delta - g(\boldsymbol{\varphi}) + c)\Theta = (\lambda + c)\Theta. \tag{4.3}$$

It then readily follows from [8, Theorem 8.6.1, Remark 8.6.1] that all eigenvalues of $[-\Delta - [g(\boldsymbol{\varphi}) + c]]$ are real and that λ_1 may be computed as the Lagrange multiplier for the normalization constraint in the (nonconvex) Courant–Fisher optimization problem

$$\min \left\{ (\nabla\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\Theta, \Theta) \text{ over } \Theta \in H_0^1(\Omega) \mid (\Theta, \Theta) = 1 \right\}. \tag{4.4}$$

Moreover, the above problem admits an optimal solution and all minimizers are the eigenfunctions corresponding to the smallest eigenvalue.

For notational simplicity, we define the solution mappings $S_u : \boldsymbol{\varphi} \mapsto \boldsymbol{u}$, $S_\lambda : \boldsymbol{\varphi} \mapsto \lambda$ and $S_\Theta : \boldsymbol{\varphi} \mapsto \Theta$ as solutions to (E($\boldsymbol{\varphi}$)) and (H($\boldsymbol{\varphi}$)), respectively. We start with the derivation of the Lipschitz continuity of S_λ.

LEMMA 4.1. *Assume* A2 *and* A4. *Then the following holds true:*

(i) *There exists $\tilde{M} > 0$ such that for all $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$, the corresponding eigenfunction satisfies $\|S_\Theta(\boldsymbol{\varphi})\|_{H_0^1(\Omega)} \leq \tilde{M}$.*

(ii) *The mapping S_λ is globally Lipschitz from $L^\infty(\Omega, \mathbb{R}^N) \to \mathbb{R}$ with modulus L and globally Lipschitz from $L^2(\Omega, \mathbb{R}^N) \to \mathbb{R}$.*

Proof. Let Θ_0 by feasible for (4.4). Then for any $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ and $\Theta = S_\Theta(\boldsymbol{\varphi}) \in H_0^1(\Omega)$ we have $(\nabla\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\Theta, \Theta) \leq (\nabla\Theta_0, \nabla\Theta_0) - (g(\boldsymbol{\varphi})\Theta_0, \Theta_0)$, which due to the normalization condition implies

$$\|\Theta\|^2_{H_0^1(\Omega)} \leq (2M + \|\Theta_0\|^2_{H_0^1(\Omega)}) =: \tilde{M}^2. \tag{4.5}$$

This yields (i). Next, fix $\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}} \in H^1(\Omega, \mathbb{R}^N)$ and let $\Theta, \hat{\Theta} \in H_0^1(\Omega)$ be the corresponding eigenfunctions. Since $\Theta, \hat{\Theta}$ are minimizers of (4.4), we have

$$\begin{aligned} (\nabla\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\Theta, \Theta) &\leq (\nabla\hat{\Theta}, \nabla\hat{\Theta}) - (g(\hat{\boldsymbol{\varphi}})\hat{\Theta}, \hat{\Theta}) + ((g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi}))\hat{\Theta}, \hat{\Theta}) \\ &\leq (\nabla\hat{\Theta}, \nabla\hat{\Theta}) - (g(\hat{\boldsymbol{\varphi}})\hat{\Theta}, \hat{\Theta}) + \|g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi})\|_{L^2} \|\hat{\Theta}\|^2_{L^4} \\ &\leq (\nabla\hat{\Theta}, \nabla\hat{\Theta}) - (g(\hat{\boldsymbol{\varphi}})\hat{\Theta}, \hat{\Theta}) + \tilde{L}\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega, \mathbb{R}^N)}. \end{aligned} \tag{4.6}$$

Here, $\tilde{L}$ combines the Lipschitz modulus L, the embedding constant from $H^1(\Omega)$ into $L^4(\Omega)$ and $\tilde{M}$. Since we may switch the roles of $\boldsymbol{\varphi}$ and $\hat{\boldsymbol{\varphi}}$, we obtain

$$|(\nabla\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\Theta, \Theta) - (\nabla\hat{\Theta}, \nabla\hat{\Theta}) + (g(\hat{\boldsymbol{\varphi}})\hat{\Theta}, \hat{\Theta})| \leq \tilde{L}\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega, \mathbb{R}^N)}. \tag{4.7}$$

As the eigenvalue is the Lagrange multiplier associated with the constraint in (4.4), we obtain from the first-order optimality conditions for (4.4)

$$\begin{aligned} (\nabla\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\Theta, \Theta) &= \lambda(\Theta, \Theta) = \lambda, \\ (\nabla\hat{\Theta}, \nabla\hat{\Theta}) - (g(\hat{\boldsymbol{\varphi}})\hat{\Theta}, \hat{\Theta}) &= \hat{\lambda}(\hat{\Theta}, \hat{\Theta}) = \hat{\lambda}, \end{aligned}$$

where λ and $\hat{\lambda}$ are the corresponding eigenvalues. Plugging this into (4.7), we see that $|\lambda - \hat{\lambda}| \leq \tilde{L}\|\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}\|_{L^2(\Omega,\mathbb{R}^N)}$, which proves the second statement in (ii). The proof of the first statement in (ii) follows from (4.6) using the upper bound:

$$\begin{aligned}((g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi}))\hat{\Theta}, \hat{\Theta}) &\leq \|g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi})\|_{L^\infty(\Omega)}\|\hat{\Theta}\|_{L^2}^2 \\ &= \|g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi})\|_{L^\infty(\Omega)} \leq L\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^\infty(\Omega,\mathbb{R}^N)}. \quad \square\end{aligned}$$

THEOREM 4.2. *Under* A2 *and* A4 *the solution mapping* $S := (S_\lambda, S_\Theta)$ *is Fréchet differentiable at any* $\boldsymbol{\varphi} \in H^1(\Omega,\mathbb{R}^N)$ *and, given a direction* $\delta\varphi \in H^1(\Omega,\mathbb{R}^N)$*, its directional derivative* $S'(\boldsymbol{\varphi})(\delta\boldsymbol{\varphi}) = (\delta\lambda, \delta\Theta)$ *can be computed as the unique solution* $(\delta\lambda, \delta\Theta) \in \mathbb{R} \times H_0^1(\Omega)$ *of the system*

$$\begin{aligned}-[\Delta + g(\boldsymbol{\varphi}) + \lambda]\delta\Theta &= \delta\lambda\Theta + [g'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}]\Theta, \\ (\Theta, \delta\Theta) &= 0.\end{aligned} \tag{4.8}$$

Proof. The proof and subsequent statement of the theorem are highly reminiscent of a similar result from classical shape and topology optimization in a sharp interface regime; we refer the reader to [36]. In the interest of completeness, we include a proof for the current (phase-field) setting in the appendix. $\square$

4.4. Existence of an optimal topology. We define the reduced objective by

$$\mathcal{J}(\boldsymbol{\varphi}) := -\int_\Omega j(\boldsymbol{\varphi}, S_\Theta(\boldsymbol{\varphi}))\mathrm{tr}\, e(\boldsymbol{S}_u(\boldsymbol{\varphi}))\mathrm{d}\mathbf{x}. \tag{4.9}$$

We now prove the existence result.

PROPOSITION 4.3. *Under Assumptions* A1–A4, (4.2) *has an optimal solution.*

Proof. We first show that $\mathcal{J}$ is bounded on $\mathcal{G}_{ad}$. Using standard arguments we obtain the boundedness of S_u in $H_0^1(\Omega,\mathbb{R}^2)$ on $\mathcal{G}_{ad}$. The boundedness of $S_\Theta(\boldsymbol{\varphi})$ follows from Lemma 4.1(i). Using $\boldsymbol{u} := S_u(\boldsymbol{\varphi})$ and $\Theta := S_\Theta(\boldsymbol{\varphi})$, we infer

$$|\mathcal{J}(\boldsymbol{\varphi})| = |J(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta)| \leq \int_\Omega |j(\boldsymbol{\varphi}, \Theta)\mathrm{tr}\, e(\boldsymbol{u})|\mathrm{d}\mathbf{x} \leq \|j(\boldsymbol{\varphi}, \Theta)\|_{L^2(\Omega)}\|\boldsymbol{u}\|_{H_0^1(\Omega,\mathbb{R}^2)}.$$

By A3–A4, $j(\boldsymbol{\varphi}, \Theta)$ is bounded in the $L^2(\Omega)$-norm.

Next, we consider an infimizing sequence $\{\boldsymbol{\varphi}^k\}$ of (4.2). Given the form of $\mathcal{G}$, this sequence is bounded in $L^\infty(\Omega,\mathbb{R}^N)$. Based on the previous argument and using the form of the Ginzburg–Landau energy we see that $\boldsymbol{\varphi}^k$ is bounded in $H^1(\Omega,\mathbb{R}^N)$. This allows us to select subsequences, denoted by the same indices, such that $\boldsymbol{\varphi}^k \rightharpoonup \boldsymbol{\varphi}$ in $H^1(\Omega,\mathbb{R}^N)$, $\boldsymbol{u}^k := S_u(\boldsymbol{\varphi}^k) \rightharpoonup \boldsymbol{u}$ in $H_0^1(\Omega,\mathbb{R}^2)$, $\Theta^k := S_\Theta(\boldsymbol{\varphi}^k) \rightharpoonup \Theta$ in $H_0^1(\Omega)$ for some $\boldsymbol{\varphi} \in H^1(\Omega,\mathbb{R}^N)$, $\boldsymbol{u} \in H_0^1(\Omega,\mathbb{R}^2)$, and $\Theta \in H_0^1(\Omega)$. From [1, Lemma 3.2] we obtain $\boldsymbol{\varphi} \in \mathcal{G}_{\mathrm{ad}}$ and $\boldsymbol{u} = S_u(\boldsymbol{\varphi})$. Moreover, for any $\hat{\Theta} \in H_0^1(\Omega)$ with $(\hat{\Theta}, \hat{\Theta}) = 1$ we have

$$\begin{aligned}(\nabla\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\Theta, \Theta) &\leq \mathrm{liminf}_k\left[(\nabla\Theta^k, \nabla\Theta^k) - (g(\boldsymbol{\varphi}^k)\Theta^k, \Theta^k)\right] \\ &\leq \mathrm{liminf}_k\left[(\nabla\hat{\Theta}, \nabla\hat{\Theta}) - (g(\boldsymbol{\varphi}^k)\hat{\Theta}, \hat{\Theta})\right] \\ &= (\nabla\hat{\Theta}, \nabla\hat{\Theta}) - (g(\boldsymbol{\varphi})\hat{\Theta}, \hat{\Theta}),\end{aligned}$$

where in the second inequality we have used that Θ^k globally minimizes (4.4) for $\boldsymbol{\varphi}^k$. Since the minimizers of (4.4) are the vectors corresponding to the smallest eigenvalue,

we have $\Theta = S_\Theta(\boldsymbol{\varphi})$. Finally, we obtain

$$\lim_k \int_\Omega j(\boldsymbol{\varphi}^k, \Theta^k) \operatorname{tr} e(\boldsymbol{S}_u(\boldsymbol{\varphi}^k)) \mathrm{d}x = \int_\Omega j(\boldsymbol{\varphi}, \Theta) \operatorname{tr} e(\boldsymbol{u}) \mathrm{d}x, \text{ and}$$
$$\liminf_k f_{GL}(\boldsymbol{\varphi}^k) \geq f_{GL}(\boldsymbol{\varphi}).$$

Since $\{\boldsymbol{\varphi}^k\}$ is a minimizing sequence and $\boldsymbol{\varphi}$ is feasible, $\boldsymbol{\varphi}$ is optimal for (4.2). ∎

4.5. First-order optimality conditions. We now derive first-order necessary optimality conditions. This implictly yields useful adjoint formulae.

THEOREM 4.4. *Assume that* A1–A4 *are satisfied. If $\boldsymbol{\varphi}$ is an optimal solution to* (4.2)*, with the corresponding $\boldsymbol{u} = S_u(\boldsymbol{\varphi})$ and $\Theta = S_\Theta(\boldsymbol{\varphi})$, then the following first-order necessary optimality conditions are satisfied:*

$$\begin{aligned} &\alpha\varepsilon(\nabla\boldsymbol{\varphi}, \nabla(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})) + \frac{\alpha}{2\varepsilon}(1 - 2\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) + \int_\Omega [\mathbb{C}'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})] e(\boldsymbol{u}) : e(\boldsymbol{p}) \mathrm{d}\mathbf{x} \\ &- \int_\Omega F'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) : e(\boldsymbol{p}) \mathrm{d}\mathbf{x} - \int_\Omega [g'(\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})] \Theta w \mathrm{d}\mathbf{x} \geq 0 \ \forall \hat{\boldsymbol{\varphi}} \in \mathcal{G}_{ad}, \end{aligned} \tag{4.10}$$

where $\boldsymbol{p} \in H_0^1(\Omega, \mathbb{R}^2)$ is the adjoint state associated with the elasticity equation

$$-\operatorname{div} \mathbb{C}(\boldsymbol{\varphi}) e(\boldsymbol{p}) = -J_u'(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta) \qquad \text{in} \quad \Omega, \tag{4.11}$$

and $w \in H_0^1(\Omega)$ is the adjoint state associated with the Helmholtz equation

$$\begin{aligned} -\Delta w - g(\boldsymbol{\varphi}) w - \lambda w &= \langle J_\Theta'(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta), \Theta \rangle \Theta - J_\Theta'(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta) \qquad \text{in} \quad \Omega, \\ (w, \Theta) &= 0. \end{aligned} \tag{4.12}$$

Here, J_u', J_Θ' denote the partial derivatives of J with respect to u and Θ, respectively.

Proof. For the first part of the objective of (4.2) we have $\mathcal{J} = J_2 \circ J_1$, where $J_1 : H^1(\Omega, \mathbb{R}^N) \to L^q(\Omega, \mathbb{R}^N) \times L^2(\Omega) \times L^q(\Omega)$ and $J_2 : L^q(\Omega, \mathbb{R}^N) \times L^2(\Omega) \times L^q(\Omega) \to \mathbb{R}$ are defined by

$$J_1(\boldsymbol{\varphi}) := (\boldsymbol{\varphi}, \operatorname{tr} e(S_u(\boldsymbol{\varphi})), S_\Theta(\boldsymbol{\varphi})), \quad J_2(\boldsymbol{\varphi}, v, \Theta) := -\int_\Omega \hat{j}(\boldsymbol{\varphi}(\cdot), \Theta(\cdot)) v(\cdot) \mathrm{d}\mathbf{x}.$$

Then J_1 is differentiable for all $q \in [1, \infty)$ due to Proposition 3.1 and Theorem 4.2. Since j is a polynomial due to A3, by direct computation it can be shown that J_2 is differentiable, as well. Consequently, the reduced objective of (4.2) is differentiable.

By a standard technique (see, e.g., [40, section 1.6.2]), we obtain

$$\mathcal{J}'(\boldsymbol{\varphi}) = J_\varphi'(\boldsymbol{\varphi}, \boldsymbol{u}) + E_\varphi'(\boldsymbol{\varphi}, \boldsymbol{u})^* \boldsymbol{p} + G_\varphi'(\boldsymbol{\varphi}, \boldsymbol{u})^*(w_\Theta, w_\lambda), \tag{4.13}$$

where E denotes the operator on the left-hand side of the elasticity equation (E($\boldsymbol{\varphi}$)) and G is defined as in the proof of Theorem 4.2. Here $\boldsymbol{p} \in H_0^1(\Omega, \mathbb{R}^2)$ is the solution of the adjoint equation $E_u'(\boldsymbol{\varphi}, \boldsymbol{u})^* \boldsymbol{p} = -J_u'(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta)$ and similarly $(w_\lambda, w_\Theta) \in \mathbb{R} \times H_0^1(\Omega)$ solves the second adjoint equation $G_{\lambda,\Theta}'(\boldsymbol{\varphi}, \boldsymbol{u})^* \boldsymbol{p} = -J_{\lambda,\Theta}'(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta)$. While the first adjoint equation amounts to (4.11), the second adjoint equation is given by

$$\begin{aligned} -\Delta w_\Theta - g(\boldsymbol{\varphi}) w_\Theta &= \lambda w_\Theta + w_\lambda \Theta - J_\Theta'(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta) \qquad \text{in} \quad \Omega, \\ (w_\Theta, \Theta) &= 0. \end{aligned} \tag{4.14}$$

Using Θ as a test function in the first equation, the boundary condition and along with the fact that (λ, Θ) is an eigenpair implies $w_\lambda = \langle J'_\Theta(\boldsymbol{\varphi}, \boldsymbol{u}, \Theta), \Theta\rangle$. Plugging this back into (4.14) and setting $w := w_\Theta$, we obtain (4.12). The rest of the proof follows from standard optimality theory; see, e.g., [14]. ∎

Remark 4.5. As in [1], we cannot guarantee the existence of Lagrange multipliers for $\mathcal{G}_{\text{ad}}$ and consequently, we only have variational optimality conditions.

5. Solution method for the optimization problem.

5.1. Data assumptions. As explained in section 1, the choice of integrand $j(\boldsymbol{\varphi}, \Theta)$ is crucial for forcing an overlap of the first eigenmode with the Ge experiencing the highest levels of strain. For our numerical experiments, we choose

$$j(\boldsymbol{\varphi}, \Theta) := \varphi_{\text{Ge}}\Theta^2. \tag{5.1}$$

This is justified since, on the one hand, $\varphi_{\text{Ge}} \in [0, 1]$ is in effect a smoothed characteristic function for the region Ω_{Ge} occupied by Ge. On the other hand, since we are optimizing for tensile strain, we expect $\operatorname{tr} e(\mathbf{u}) \geq 0$ (at least on average over the domain Ω). Therefore, a minimization procedure should force Θ^2 to be as large as possible on Ω_{Ge}. Since $(\Theta, \Theta) = 1$, this ultimately confines the bulk of $\operatorname{supp}\Theta$ to Ω_{Ge}.

Since the electronics are modeled externally, we need to add some small but non-negligible restrictions on the possible configurations. We do so by using the fixed regions $\Pi_i \subset \Omega$. Without them, the optimization method might suggest designs that are infeasible from a manufacturing perspective, e.g., there must be an SiO_4 substrate or it might omit Ge.

5.2. Optimization algorithm. The optimization algorithm is based on a standard projected gradient step as in [33, 43]. Denoting the reduced objective by $\hat{\mathcal{J}} := \mathcal{J} + \alpha f_{GL}$, we thus obtain at each step

$$\boldsymbol{\varphi}^{k+1}(t) = \operatorname{Proj}_{\mathcal{G}}\big(\boldsymbol{\varphi}^k - tR^{-1}_{\text{Riesz}}(\hat{\mathcal{J}}'(\boldsymbol{\varphi}^k))\big), \quad t > 0, \tag{5.2}$$

where $\operatorname{Proj}_{\mathcal{G}}(v)$ is the usual projection of v on the closed convex set $\mathcal{G}_{\text{ad}}$. Since each $\boldsymbol{\varphi}^k \in H^1(\Omega; \mathbb{R}^N)$, we have $\hat{\mathcal{J}}'(\boldsymbol{\varphi}^k) \in H^1(\Omega; \mathbb{R}^N)^*$. Therefore, we need to obtain the Riesz representation $R^{-1}_{\text{Riesz}}(\hat{\mathcal{J}}'(\boldsymbol{\varphi}^k)) \in H^1(\Omega; \mathbb{R}^N)$. Failure to do so may result in a theoretical inconsistency on the continuous level as well as a drastically reduced convergence rate or even lack of convergence in the discrete setting (asymptotically, assuming conforming discretizations). Fortunately, the Riesz representation $\boldsymbol{\xi} = R^{-1}_{\text{Riesz}}(\hat{\mathcal{J}}'(\boldsymbol{\varphi}^k))$ can be easily calculated by solving a linear elliptic PDE:

$$\begin{aligned} -\Delta\boldsymbol{\xi} + \boldsymbol{\xi} &= \hat{\mathcal{J}}'(\boldsymbol{\varphi}^k) && \text{in} \quad \Omega, \\ \partial_n\boldsymbol{\xi} &= 0 && \text{on} \quad \partial\Omega. \end{aligned} \tag{5.3}$$

As suggested in [12], we use a generalized Armijo rule in order to select the step size $t^k = t$ in (5.2) and set $\boldsymbol{\varphi}^{k+1} := \boldsymbol{\varphi}^{k+1}(t^k)$. Here, for a given $\sigma > 0$ we use a simple backtracking strategy to find the largest $t^k > 0$ such that

$$\hat{\mathcal{J}}(\boldsymbol{\varphi}^k) - \hat{\mathcal{J}}(\boldsymbol{\varphi}^{k+1}) \geq \sigma(t^k)^{-1}\|\boldsymbol{\varphi}^k - \boldsymbol{\varphi}^{k+1}\|^2_{H^1(\Omega,\mathbb{R}^N)}. \tag{5.4}$$

We then iterate until

$$\|\boldsymbol{\varphi}^k - \operatorname{Proj}_{\mathcal{G}}\left(\boldsymbol{\varphi}^k - R^{-1}_{\text{Riesz}}(\hat{\mathcal{J}}'(\boldsymbol{\varphi}^k)\right)\|_{H^1(\Omega,\mathbb{R}^N)} \leq \text{tol}_{PG} \tag{5.5}$$

is satisfied for some tolerance $\text{tol}_{PG} > 0$, as suggested in, e.g., [12].

The choice of a first-order numerical method is motivated by the nature of the constraints in our problem. In particular, a direct application of second-order optimization techniques as in [1] is not possible here. Indeed, if we were to write the Helmholtz equation in the form of optimality conditions (3.3), there would be no guarantee that the steps generated by a second-order method are related to the smallest eigenvalue.

For the projection onto the Gibbs simplex $\mathcal{G}$, we use the potentially mesh-dependent semismooth Newton method as suggested in [38], where it is shown to be equivalent to a primal-dual active set strategy with warm start. This strategy is efficient provided the active sets are stable over mesh refinements. Another possibility would be to use the path-following method from [2], as it is mesh-independent.

Finally, in order to calculate $\hat{\mathcal{J}}'$ we need to solve both forward equations and both adjoint equations (see Theorem 4.4). For the solution of the Helmholtz equation (H($\boldsymbol{\varphi}$)), we first apply the shift as described in section 5.3 below, then (for the discretized problem) we solve the resulting eigenvalue problem via MATLAB function `eigs` (which is built on top of ARPACK; cf. [42]) and finally apply a shift back. Since the directional derivative from (4.8) has a unique solution, it can be simply solved as a system of linear equations.

5.3. Estimating the shift parameter c. The choice of the shift parameter c in (4.3) is a delicate matter as it has a major impact on the computation of the smallest eigenvalue. Several methods such as the inverse method [50] find the eigenvalue closest to zero and the rate of convergence equals to the ratio of the two eigenvalues closest to zero. Thus, if the shift is too small, a different eigenvalue may be found, while if the shift is too big, the convergence will be slow.

To keep positivity of the smallest eigenvalue, it is always possible to choose $c = M$, where M is the bounding constant from A2. However, this choice may be suboptimal. Here, we present two possibilities for a shift which ensures positivity of the smallest eigenvalue.

LEMMA 5.1. *Set $\Omega = (0,a) \times (0,b)$ and assume that* A2 *and* A4 *hold. Consider the shift*

$$c := L(M + 2\|g(0)\|_{L^\infty(\Omega)}) - 2\pi^2 / ab\,. \tag{5.6}$$

Then the smallest eigenvalue of $-\Delta - g(\boldsymbol{\varphi}) + cI$ is nonnegative for all $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$.

Proof. Denote by λ_1 the smallest eigenvalue of the operator $-\Delta - g(\boldsymbol{\varphi})$ and by $\lambda_1(\Omega)$ the smallest eigenvalue of the operator $-\Delta$. From Lemma 4.1 we infer

$$|\lambda_1 - \lambda_1(\Omega)| \le L\|g(\boldsymbol{\varphi}) - g(0)\|_{L^\infty(\Omega)} \le L(M + 2\|g(0)\|_{L^\infty(\Omega)}).$$

From [8, Proposition 8.5.2] we obtain $\lambda_1(\Omega) = \frac{2\pi^2}{ab}$. The assertion follows. ∎

Let $\boldsymbol{\varphi}$ be the current iterate of a procedure for solving our optimization problem. If c is the shift and the computed smallest eigenvalue of operator $-\Delta - g(\boldsymbol{\varphi}) + cI$ equals λ_1, then the optimal shift is $c - \lambda_1$. Even though we cannot use this information for determining $\boldsymbol{\varphi}$, it is of use for determining the next iterate. In fact, in this case we may use the shift in (5.7). In what follows, C_P denotes the Poincaré constant, i.e., for every $\Theta \in H_0^1(\Omega)$ one has $\|\Theta\|_{L^2(\Omega)} \le C_P \|\nabla\Theta\|_{L^2(\Omega)}$.

LEMMA 5.2. *Assume that* A2 *and* A4 *hold and that for $\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$ and for some shift c we know the eigenvalue λ_1 of the operator $-\Delta + g(\boldsymbol{\varphi}) + cI$. Consider*

$\delta\boldsymbol{\varphi} \in H^1(\Omega, \mathbb{R}^N)$, define

$$\hat{c} := c - \lambda_1 + 2^{-3/4} C_P^{-2} (C_P^2 + 1)(2MC_P^2 + 1) L \|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)}, \tag{5.7}$$

and denote by $\hat{\lambda}_1$ the smallest eigenvalue of operator $-\Delta + g(\boldsymbol{\varphi} + \delta\boldsymbol{\varphi}) + \hat{c}I$. Then $\hat{\lambda}_1 \geq 0$. Moreover, denote the second smallest eigenvalues of the previous two operators by λ_2 and $\hat{\lambda}_2$, respectively. Let $\kappa := 2^{-3/4} C_P^{-2}(C_P^2+1)(2MC_P^2+1)L$. If $\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)} < (\lambda_2 - \lambda_1)/(2\kappa)$, then

$$0 \leq \hat{\lambda}_1 \hat{\lambda}_2^{-1} \leq 2\kappa(\lambda_2 - \lambda_1)^{-1} \|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)}. \tag{5.8}$$

Proof. Due to [61, Chapter 3, Lemma 3.3], for any $\Theta \in H_0^1(\Omega)$ we have

$$\begin{aligned} \|\Theta\|_{L^4(\Omega)}^2 &\leq 2^{\frac{1}{4}} \|\nabla\Theta\|_{L^2(\Omega)} \|\Theta\|_{L^2(\Omega)} \leq 2^{-\frac{3}{4}} \big(\|\nabla\Theta\|_{L^2(\Omega)}^2 + \|\Theta\|_{L^2(\Omega)}^2\big) \\ &\leq 2^{-\frac{3}{4}} (1 + C_P^2) \|\nabla\Theta\|_{L^2(\Omega)}^2. \end{aligned}$$

By definition, $C_P^{-1} = \inf\{\|\nabla u\|_{L^2(\Omega)} / \|u\|_{L^2(\Omega)} : u \in H_0^1(\Omega)\}$. This optimization problem can be reformulated as $\inf\{\|\nabla u\|_{L^2(\Omega)} : u \in H_0^1(\Omega),\ \|u\|_{L^2(\Omega)} = 1\}$, which has a solution based on our analysis of (4.4). Therefore, there exists some $\hat{\Theta}_0 \in H_0^1(\Omega)$ with $1 = \|\hat{\Theta}_0\|_{L^2(\Omega)} = C_P \|\nabla\hat{\Theta}_0\|_{L^2(\Omega)}$. Now we provide an estimate for the constant $\tilde{L}$ in (4.6). Fix any $\hat{\boldsymbol{\varphi}} \in H^1(\Omega, \mathbb{R}^N)$ and let $\hat{\Theta} \in H_0^1(\Omega)$ be the corresponding minimizer of (4.4). Then we have

$$\begin{aligned} \|g(\hat{\boldsymbol{\varphi}}) - g(\boldsymbol{\varphi})\|_{L^2(\Omega)} \|\hat{\Theta}\|_{L^4(\Omega)}^2 &\leq L \|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega)} \|\hat{\Theta}\|_{L^4(\Omega)}^2 \\ &\leq 2^{-\frac{3}{4}} L (C_P^2 + 1) \|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega)} \|\nabla\hat{\Theta}\|_{L^2(\Omega)}^2 \\ &\leq 2^{-\frac{3}{4}} L (C_P^2 + 1) \|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega)} (2M + \|\nabla\hat{\Theta}_0\|_{L^2(\Omega)}^2) \\ &= 2^{-\frac{3}{4}} C_P^{-2} (C_P^2 + 1)(2MC_P^2 + 1) L \|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{L^2(\Omega)}, \end{aligned}$$

where the third inequality is due to (4.5) and L is the Lipschitz constant of $\hat{g}$. Thus, we have $\tilde{L} = 2^{-\frac{3}{4}} C_P^{-2}(C_P^2+1)(2MC_P^2+1)L$. Then the first two eigenvalues of the operator $-\Delta + g(\boldsymbol{\varphi}) + \hat{c}I$ are equal to $\tilde{\lambda}_1 := \tilde{L}\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)}$ and $\tilde{\lambda}_2 := \tilde{L}\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)} + \lambda_2 - \lambda_1$, respectively. From Lemma 4.1 we then obtain $0 \leq \hat{\lambda}_1 \leq 2\tilde{L}\|\delta\boldsymbol{\varphi}\|_{L^2(\Omega)}$, $\lambda_2 - \lambda_1 \leq \hat{\lambda}_2$. ∎

Note that the shift $c = M$ and the shift from Lemma 5.1 are independent of $\boldsymbol{\varphi}$, where the one from Lemma 5.2 depends on the perturbation. Observe, furthermore, that an iterative scheme for solving the optimization problem (4.2) in reduced form yields $\delta\boldsymbol{\varphi} \to 0$ in $L^2(\Omega, \mathbb{R}^N)$. Hence, the shift in (5.7) converges to the optimal shift $c - \lambda_1$ and the ratio in (5.8) tends to zero.

6. Calculating the optimal design. In this section, we present the results of numerical optimization experiments. The optimal solution is then used in the final section below to demonstrate the electronic properties of the associated microbridge design.

6.1. Structural assumptions: Elasticity and optics. For the elasticity equation, we primarily follow the setting in [1]. The $\boldsymbol{\varphi}$-dependent elasticity tensor is of the form

$$\mathbb{C}(\boldsymbol{\varphi}) := \widehat{\mathrm{cut}}(\varphi_1)\mathbb{C}_1 + \cdots + \widehat{\mathrm{cut}}(\varphi_N)\mathbb{C}_N,$$

where $\mathbb{C}_i$ is a standard elasticity tensor associated with material i. Thus, for $E_1, E_2 \in \mathbb{R}^{2\times 2}$ we have $\mathbb{C}_i E_1 : E_2 = \lambda_i \operatorname{tr} E_1 \operatorname{tr} E_2 + 2\mu_i E_1 : E_2$, where λ_i and μ_i are Lamé constants of individual materials and $\widehat{\operatorname{cut}} : \mathbb{R} \to \mathbb{R}$ is the cutoff function

$$\widehat{\operatorname{cut}}(x) := \begin{cases} \operatorname{arctg}(x-\delta_2)+\delta_2 & \text{if } x \geq \delta_2, \\ x & \text{if } x \in [\delta_1, \delta_2), \\ x - 2\delta_1(x-\delta_1)^3 - (x-\delta_1)^4 & \text{if } x \in [0, \delta_1), \\ a \operatorname{arctg}(bx) + \delta_1^4 & \text{if } x < 0 \end{cases} \tag{6.1}$$

for some small $\delta_1 > 0$, $\delta_2 > 0$ with $\delta_2 \gg \delta_1$, $a = \delta_1^4/\pi$, and $b = (1 - 2\delta_1^3)\pi/\delta_1^4$. Note that the cutoff function is a twice continuously differentiable increasing function with the property $\widehat{\operatorname{cut}}(x) \geq \delta_1^4/2$ for all $x \in \mathbb{R}$ and thus A1 is satisfied. As in [1], where we employed second-order optimization methods, $\widehat{\operatorname{cut}}$ is chosen to ensure that $\mathbb{C}$ is sufficiently smooth and the resulting differentiable operator remains elliptic. Note that as $\delta_1 \to 0$ and $\delta_2 \to 1$, $\widehat{\operatorname{cut}}$ approaches the identity on $[0, 1]$.

Concerning the Helmholtz equation, we define g by

$$g(\boldsymbol{\varphi}) := 2\pi^2 \lambda^{-2} (\varepsilon_1 \operatorname{cut}(\varphi_1) + \cdots + \varepsilon_N \operatorname{cut}(\varphi_N)). \tag{6.2}$$

Here, $\lambda > 0$ is the desired wavelength and $\varepsilon_i > 0$, $i = 1, \ldots, N$, are the relative permittivities of the individual materials. For some small $\delta_3 > 0$, the cutoff function $\operatorname{cut} : \mathbb{R} \to \mathbb{R}$

$$\operatorname{cut}(x) := \begin{cases} 1 + \delta_3 \operatorname{arctg}(\frac{x-1}{\delta_3}) & \text{if } x > 1, \\ x & \text{if } x \in [0, 1], \\ \delta_3 \operatorname{arctg}(\frac{x}{\delta_3}) & \text{if } x < 0 \end{cases} \tag{6.3}$$

is necessary for Assumption A2 to hold true. Since the requirements on $\mathbb{C}$ and g are different, we work with different cutoff functions.

6.2. Discretization and refinement strategies. For the numerical implementation, we discretize the underlying function spaces using P1-finite elements. All numerical experiments are carried out using MATLAB. In order to increase the computational efficiency of the scheme, we use an adaptivity heuristic to generate new meshes following the "red" refinement strategy (cf. [15]), which is implemented in the package P1-AFEM; see [28]. The marking heuristic is as follows: After solving (4.2) on a given mesh, every element on which the phases are not pure or where there is a transition between two materials is refined. Otherwise, we coarsen or leave the element unchanged if there exist pure phases and no transition. It would go beyond the scope of this paper to develop a proper AFEM scheme for the given problem. However, we believe that it should be possible to extend several ideas from the literature to the current setting, e.g., as found in [27, 37, 5].

In addition to the role of the various phases in the refinement strategy, we need to take into account the interfacial thickness parameter ε, which appears in the Ginzburg–Landau term f_{GL}. Since ε corresponds to the interfacial thickness, the initial ε is chosen to be twice the length of the largest element. Subsequently, we divide ε by 2 upon every mesh refinement. We refine the mesh in our experiments five times.

6.3. Parameters and starting values. As mentioned above, we consider three possible materials, Ge, SiN, and SiO_2, as well as air. In Table 1 we summarize

TABLE 1
List of material properties for elasticity.

	λ [GPa]	μ [GPa]	ε [1]	σ_0 [GPa]	ε_0 [1]
Ge	44.279	27.249	17.64	·	·
SiN	110.369	57.813	4	-3.8	·
SiO_2	16.071	20.798	2.25	·	$2.6{\cdot}10^{-3}$

TABLE 2
List of parameters.

α	N	$h_{\min}$	$\varepsilon_{\min}$	δ_1	δ_2	δ_3	tol_{PG}	σ
$4{\cdot}10^{-4}$	4	2^{-8}	2^{-7}	10^{-3}	10^{16}	10^{-3}	10^{-6}	10^{-4}

their physical properties (see [46, 64, 66]) and the fixed domains Π_i are given by $\Pi_{\mathrm{Ge}} := [-0.125, 0.125]\times[1, 1.49]$, $\Pi_{\mathrm{SiN}} := [-0.75, 0.75]\times[1.5, 1.75]$, $\Pi_{\mathrm{SiO_2}} := [-2, 2]\times[0, 0.99]$, $\Pi_{\mathrm{air}} := [-2, 2]\times[2.5, 3]$ (in μm). Since the general model contains a number of parameters, we list them here for convenience:

- N: Number of phases.
- α: Weights in the objective for the Ginzburg–Landau energy f_{GL}.
- ε: Parameter corresponding to interfacial thickness.
- δ_1, δ_2, δ_3: Cutoff parameters from (6.1) and (6.3).
- ϵ_0, δ_0: Constants for the eigenstrain generated by SiO_2 and the thermal (pre-)stress generated by SiN; see (3.1).
- λ, ε_i: The wavelength and the relative permittivities of materials; see (6.2).
- tol_{PG}: Stopping tolerance for first-order system (5.5).
- $h_{\min}$, $\varepsilon_{\min}$: Width of the smallest triangle and value of ε on the finest mesh.

The parameter values are summarized in Table 2. The cutoff parameters δ_1, δ_2, and δ_3 were chosen so that the cutoff has a negligible effect on the interval $(0, 1)$. Since $\Omega = (-2, 2)\times(0, 3)$ (in μm), the values $h_{\min} = \frac{1}{256}\mu$m and $\varepsilon_{\min} = \frac{1}{128}\mu$m give rise to a rather fine mesh along the interface. For the wavelength we choose $\lambda = 1.64\mu$m.

6.4. Numerical results. In Figure 2 we depict the optimal $\boldsymbol{\varphi}$ (left) and the corresponding strain field (right). For efficiency, we employ the mesh refinement strategy described above. The meshes after the first, third, and fifth refinements are shown in Figure 3. Since we are able to drive ε to a small value, the final design has a rather sharp interface; see Figure 2.

The number of active nodes (where no material is prescribed) is depicted in the left-hand side of Table 3. The small increase from the penultimate to the final mesh is caused by the disappearance of an artifact above the structure, whose presence can be inferred from the structure of the refined mesh in Figure 3. The refined region above the structure in the final mesh is a remanent of this artifact, which disappears in the final phase field $\boldsymbol{\varphi}$; see Figure 2. The number of iterations is shown on the right-hand side of Table 3. Note that on the intermediate meshes 3, 4, and 5, we accepted a suboptimal, i.e., substationary, solution after reaching 500 iterations. Nevertheless, on Mesh 6, the algorithm only needs 223 iterations to reach a tolerance below 10^{-7}.

6.5. Optoelectronic properties of the optimal design. We conclude with a numerical study investigating the optoelectronic properties of the optimal design. As discussed in the introduction and later in section 3.3, we can expect the optimal design to be successful only if it exhibits a positive net gain. We manually introduce the contacting layers by adding two new phase fields φ_i with $i = n$-Si or $i = p$-Si

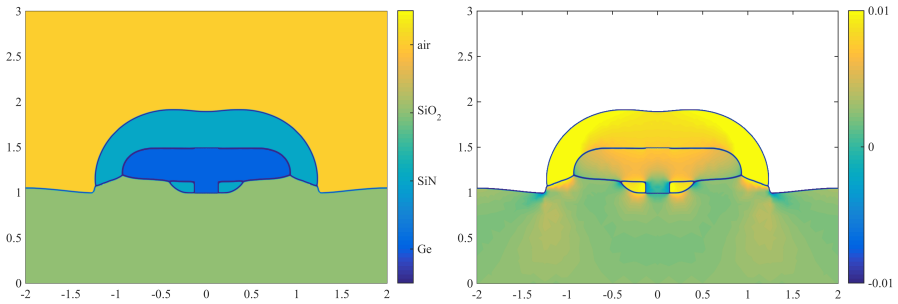

FIG. 2. *Optimal $\boldsymbol{\varphi}$ (left) and its corresponding strain field (right).*

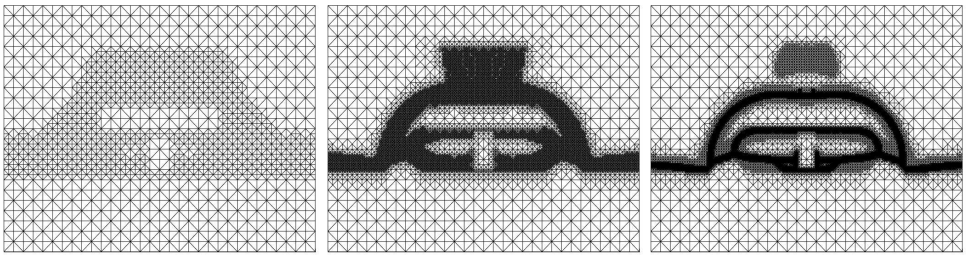

FIG. 3. *Adaptively updated mesh (both refined and coarsened) after first, third, and fifth refinements.*

TABLE 3

`active`: Number of active nodes (with no material prescribed). `iter`: number of iterations. `res`: the best residual (5.5). *`GL`: the value of the Ginzburg–Landau energy for all meshes.*

	Mesh 1	Mesh 2	Mesh 3	Mesh 4	Mesh 5	Mesh 6
`active`	345	1043	3027	9902	25824	36273
`iter`	37	427	500+	500+	500+	223
`res`	$9.33{\cdot}10^{-8}$	$8.95{\cdot}10^{-8}$	$8.42{\cdot}10^{-5}$	$9.95{\cdot}10^{-5}$	$3.26{\cdot}10^{-7}$	$7.94{\cdot}10^{-8}$
`GL`	9.298	8.821	8.659	9.154	8.630	8.580

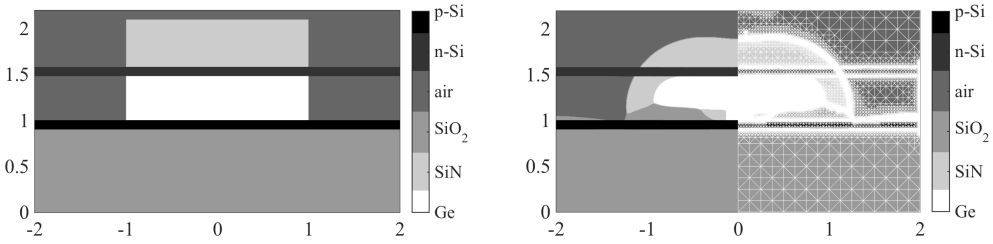

FIG. 4. *(left) Original phase fields φ_i with inserted Si contacts and (right) optimized phase fields φ_i with inserted Si contacts and mesh with shading indicating $i \in \{Ge, SiN, SiO_2, air, n\text{-}Si, p\text{-}Si\}$.*

representing thin highly n- and p-doped Si layers above and below the Ge; cf. Figure 4. For the simulation, we consider the stationary version of (3.4) after reformulating it in terms of the quasi-Fermi potentials (ϕ_n, ϕ_p) in (3.6). We enforce inhomogeneous Dirichlet boundary conditions for the potentials at the two Ohmic contacts Γ_{D_i} $(i = 1, 2)$ corresponding to $\varphi_{n\text{-Si}}$ and $\varphi_{p\text{-Si}}$ on $\partial\Omega$; otherwise we have natural boundary

TABLE 4

Spatial interpolation $\pi(x) = \sum_i \varphi_i(x)\pi_i$ for electronic simulation given phase fields $\varphi_i(x)$ and pure phase material parameters π_i. The global parameters $\tau_n = \tau_p = 10\,\mathrm{ns}$ and $n_i = 10^6\,\mathrm{cm}^{-3}$ are used for the recombination. ($\star$) Given a strain distribution $e(\mathbf{u})$, the band gaps are modified by deformation potentials $\mathcal{D}^\alpha_{kl} = \mathcal{D}^\alpha \delta_{kl,xx}$ via $E_\alpha(x) = \sum_i (E_i + \mathcal{D}^\alpha e_{xx})\varphi_i(x)$ with $\alpha \in \{c,v\}$ and in-plane biaxial strain e_{xx}. The electronic parameters and deformation potentials are from [52]; the values for $\mu_n, \mu_p, N_c, N_v, E_c, E_v$ for SiN, SiO_2, and air are chosen to prevent existence and transport of carriers.

Param.	Phys. unit	Ge	SiN	SiO_2	Air	Si^{top}	Si^{bottom}
ε_r	[1]	16.2	7.5	3.8	1	11.9	11.9
μ_n	$[m^2V^{-1}s^{-1}]$	0.39	10^{-4}	10^{-4}	10^{-4}	0.14	0.14
μ_p	$[m^2V^{-1}s^{-1}]$	0.19	10^{-4}	10^{-4}	10^{-4}	0.045	0.045
N_c	$[10^{19}cm^{-3}]$	1.256	10^{-2}	10^{-2}	10^{-2}	3.2	3.2
N_v	$[10^{19}cm^{-3}]$	0.118	10^{-2}	10^{-2}	10^{-2}	1.8	1.8
C_{dop}	$[10^{19}cm^{-3}]$	5	0	0	0	+20	−20
E_c	[eV]	$0.76^\star$	1	1	1	1.169	1.169
E_v	[eV]	$0.09^\star$	0	0	0	1.169	1.169
$\mathcal{D}^c$	[eV]	−3.5	0	0	0	0	0
$\mathcal{D}^v$	[eV]	+1.4	0	0	0	0	0

conditions. In particular, on each Γ_{D_i} we have

$$\phi = \bar{\phi} + V^i_{\mathrm{ext}},\ \phi_n = V^i_{\mathrm{ext}},\ \phi_p = V^i_{\mathrm{ext}},$$

where $\bar{\phi}$ is the built-in potential and the voltage biases are given by $V^1_{\mathrm{ext}} = 0$, $V^2_{\mathrm{ext}} = V_{\mathrm{ext}}$, and $\Gamma_{D_1} \cap \Gamma_{D_2} = \emptyset$.

Given the optimal distribution of materials $\boldsymbol{\varphi}$, the material data μ_n, μ_p, N_c, N_v,E_c, E_v, ε_r, C_{dop} are made to depend on space through the phase fields via interpolation as introduced in Table 4, e.g., $\mu_n(x) = \sum_i \mu^i_n \varphi_i(x)$. We treat the Fermi–Dirac integral $F_{3/2}$, used in the transformation into quasi-Fermi potentials, via a standard closed-form approximation as in [10].

In order to easily fit the solver to the optimal design and associated refined mesh, we use P1-finite elements to discretize the stationary van Roosbroeck system. The nonlinearities are treated using a standard seven-point Gauss quadrature and the inhomogeneous boundary conditions are enforced numerically using Lagrange multipliers. The use of finite elements stands in contrast to the usual Scharfetter–Gummel finite volume methods [55]. However, several studies do advocate for the benefits of using finite elements; see, e.g., [23].

The resulting discretized system of equations is solved by using Newton's method. In order to ensure its convergence for large applied biases V_{ext}, an initial value is calculated by solving this system at thermal equilibrium. This is done by setting V_{ext}, ϕ_n, and ϕ_p to zero and solving the remaining system for ϕ, which amounts to an elliptic equation of Poisson–Boltzmann-type. Given this initial value, we can then calculate the stationary solutions at each desired bias V_{ext} by a standard continuation step. Finally, recalling the equations for the flux terms (3.5), we can calculate the total current at a given bias V_{ext} using the formula

$$J = \int_{\Gamma_{D_i}} (\mathbf{j}_p + \mathbf{j}_n) \cdot \mathbf{n}\, \mathrm{d}a,$$

We compute the currents, current densities, and the modal net gains, which are defined (pointwise) by subtracting optical losses from optical gain and scaling the

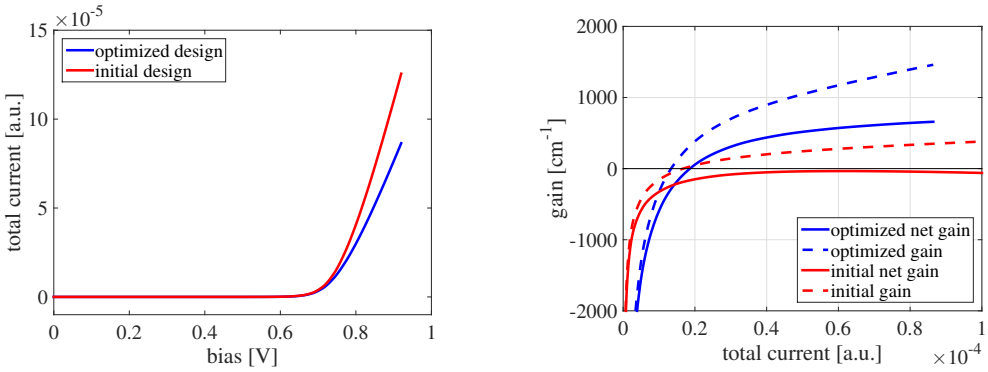

FIG. 5. *(left) Current-voltage characteristic of initial (red) and optimized (blue) device (right) current gain (solid) and current-net gain (dashed) characteristics of initial and optimized device showing that the optimized configuration yields considerably higher gain and net gain compared to the initial design.*

result by the optical mode. The resulting gain model is the same as published in [51], and generally higher net gains for a given current is desired. We refer to Figure 5, where we see that the optimal design clearly exhibits positive net gain, as opposed to the empirical design. We comment on this further in the conclusion below.

7. Conclusions and outlook. As desired, the topology optimization delivers a rather smooth material distribution, which increases the in-plane biaxial strain in the Ge phase for the initial design from an average strain $\bar{e}_{xx} = 2 \cdot 10^{-4}$ to an average strain of $\bar{e}_{xx} = 9 \cdot 10^{-4}$ for the improved design; see Figure 2. While loss mechanisms due to low confinement or recombination are not included in the optimization, the cost functional in (4.2) is designed to optimize the overlap of the optical mode and regions of large tensile strain. Therefore, the optimal designs exhibit overall improvements for the integrated strain (on average) versus the maximal/peak in-plane strains. For the latter, we see here that the maximal (pointwise) in-plane strain in the Ge cavity only features an increase by a factor of $\times 1.2$.

Another interesting feature of the optimal designs is that the Ge phase is surrounded by an SiN stressor. This is very similar to the all-around stressor designs considered for germanium microdiscs in [30].

The optimized design also features an aperture, which, as we showed previously, can be highly beneficial for lowering the threshold current of an edge-emitting laser. The main idea of the aperture is visible in the hole currents in Figure 6, where the currents in the optimized microbridge (right) are guided efficiently into the optical mode to recombine without creating a shortcut pathway around the center of the optical mode, as is the case for the initial microbridge (left). For better interpretation we also indicate the material boundaries between the phases by plotting regions where $\varphi_i \varphi_j > 0$ between material i and j in white. However, while a doping optimization can produce such an aperture geometry from a suitably defined cost functional, the aperture of the optimized design is more likely created artificially due to the location of the highly doped Si contacts above the cavity.

Nevertheless, due to the improved strain and the better overlap of the hole current (see Figure 5) and the optical mode (see Figure 7), the Ge phase also features much higher modal gain at the prescribed external bias (see Figure 8). Also, the characteristic curve in Figure 5 features a lower current, certainly due to higher Ohmic

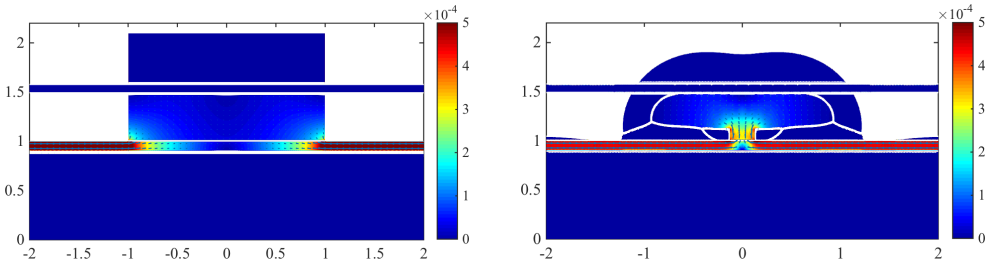

FIG. 6. *Hole currents for (left) initial design and (right) optimized design. Material boundaries are indicated in white.*

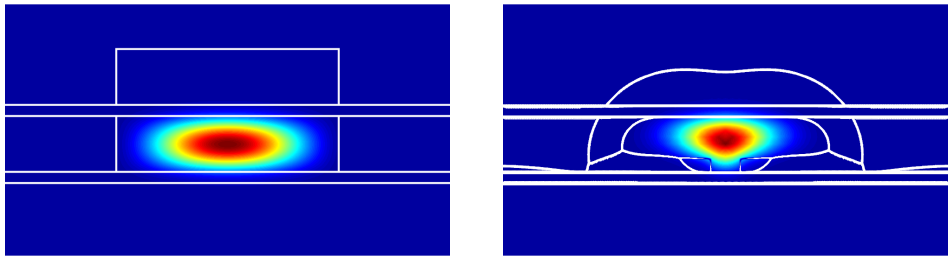

FIG. 7. *Optical mode* $|\Theta|^2$ *(shading) and material boundaries indicated in white (left) for initial design and (right) for optimized design.*

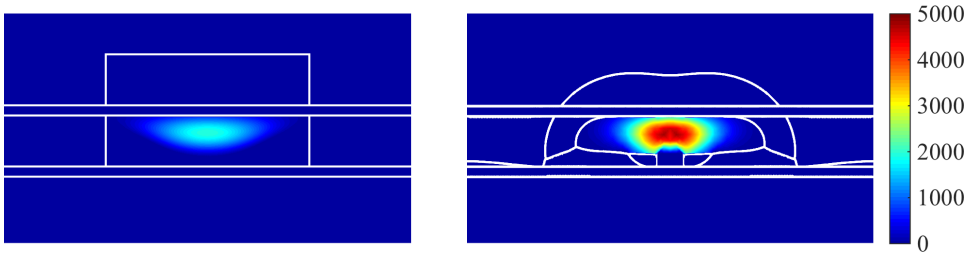

FIG. 8. *Modal gain* $g|\Theta|^2$ *[*cm^{-1}*] for (left) initial design and (right) optimized design shows almost threefold increase in gain due to optimized design. Material boundaries are indicated in white.*

resistance based on the implementation of the aperture. Most noticeable, however, is that the modal gain as well as the net gain show significant improvement of the optimized design as compared to the initial design. For a recent study containing a thorough explanation of the calculation of the gain curves, we refer the interested reader to [51].

This allows us to conclude that even though not yet fully coupled, topology optimization for optoelectronic devices can improve device designs significantly. The optimized designs are similar to what is considered by engineers. The optoelectronic simulations prove the feasibility of the optimization strategy. Nevertheless, since optoelectronic devices also suffer from loss mechanisms due to recombination, future optimization studies might even consider the fully coupled optoelectronic system.

Appendix. The following is a proof of Theorem 4.2, which we provide for the sake of completeness.

Proof. Based on (3.3) and (H($\boldsymbol{\varphi}$)) we consider the following system of equations in strong form:

$$\begin{aligned} (-\Delta - g(\boldsymbol{\varphi}))\Theta - \lambda\Theta = 0 \qquad & \text{in} \quad \Omega, \\ \Theta = 0 \qquad & \text{on} \quad \partial\Omega, \\ (\Theta,\Theta) - 1 = 0. & \end{aligned} \tag{7.1}$$

Multiplying $(-\Delta - g(\boldsymbol{\varphi}))\Theta$ by $\psi \in H_0^1(\Omega)$ and integrating over Ω, it follows from Green's theorem that

$$\int_\Omega (-\Delta - g(\boldsymbol{\varphi}))\Theta\psi \mathrm{d}\mathbf{x} = (\nabla\Theta, \nabla\psi) - (g(\boldsymbol{\varphi})\Theta, \psi).$$

Therefore, there exists a unique coercive bounded linear operator $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ such that $\langle A\Theta, \psi\rangle = (\nabla\Theta, \nabla\psi)$. Nevertheless, we allow a slight abuse of notation and denote A by $-\Delta$. The boundary condition in (7.1) is therefore "absorbed" by the operator.

Continuing, we denote the solution mapping of (7.1) by $\hat{S} : \boldsymbol{\varphi} \mapsto (\lambda, \Theta)$. Note that $\hat{S}$ is in fact multivalued (for every $\boldsymbol{\varphi}$, $\hat{S}(\boldsymbol{\varphi})$ is the set of all eigenpairs). Nevertheless, since S_λ is single-valued, the Lipschitz continuity of S_λ from Lemma 4.1 implies that there exists an open ball around $\boldsymbol{\varphi}$ and a selection of $\hat{S}$ that coincides locally with S. To derive differentiability of S it suffices then to apply the implicit function theorem [67, Theorem 4.B] to (7.1).

Denote the function on the left-hand side of (7.1) by $G(\boldsymbol{\varphi}; \lambda, \Theta)$. Clearly, G is continuous. By formally differentiating this mapping in direction $(\delta\boldsymbol{\varphi}, \delta\lambda, \delta\Theta)$, we obtain the formula

$$G'(\boldsymbol{\varphi}, \lambda, \Theta)(\delta\boldsymbol{\varphi}, \delta\lambda, \delta\Theta) = \begin{pmatrix} -\Delta\delta\Theta - [g'(\boldsymbol{\varphi})\delta\boldsymbol{\varphi}]\Theta - g(\boldsymbol{\varphi})\delta\Theta - \delta\lambda\Theta - \lambda\delta\Theta \\ 2(\Theta, \delta\Theta) \end{pmatrix}.$$

Furthermore, by substituting this formula into the usual difference quotient, it is not difficult to verify that $G : H^1(\Omega, \mathbb{R}^N) \times \mathbb{R} \times H_0^1(\Omega) \to H^{-1}(\Omega) \times \mathbb{R}$ is in fact continuously Fréchet differentiable. Finally, we show that the partial derivative $G'_{\lambda,\Theta}(\boldsymbol{\varphi}, \lambda, \Theta)$ is bijective. We have

$$G'_{\lambda,\Theta}(\boldsymbol{\varphi}, \lambda, \Theta)(\delta\lambda, \delta\Theta) = \begin{pmatrix} -\Delta\delta\Theta - g(\boldsymbol{\varphi})\delta\Theta - \delta\lambda\Theta - \lambda\delta\Theta \\ 2(\Theta, \delta\Theta) \end{pmatrix}. \tag{7.2}$$

To demonstrate injectivity, we need to show that

$$\begin{aligned} -\Delta\delta\Theta - g(\boldsymbol{\varphi})\delta\Theta - \delta\lambda\Theta - \lambda\delta\Theta = 0, \\ (\Theta, \delta\Theta) = 0, \end{aligned} \tag{7.3}$$

admits only the trivial solution $(\delta\lambda, \delta\Theta) = (0, 0) \in \mathbb{R} \times H_0^1(\Omega)$. To this aim, suppose $(\delta\lambda, \delta\Theta)$ is some solution pair. Using Θ as a test function in the first equation in (7.3) we obtain

$$(\nabla\delta\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\delta\Theta, \Theta) - \delta\lambda(\Theta, \Theta) - \lambda(\delta\Theta, \Theta) = 0. \tag{7.4}$$

Realizing that $(\nabla\delta\Theta, \nabla\Theta) - (g(\boldsymbol{\varphi})\delta\Theta, \Theta) - \lambda(\delta\Theta, \Theta) = 0$ due to symmetry and the definition of the eigenvalue, relation (7.4) reduces to $0 = \delta\lambda(\Theta, \Theta) = \delta\lambda$. Plugging

this back into (7.3) we see that $(\lambda, \delta\Theta)$ is an eigenpair. But this implies $\delta\Theta = 0$ because the multiplicity of λ is one and $\delta\Theta$ is orthogonal to Θ. Thus, we have shown injectivity.

For surjectivity, we need to show that for any $v \in H^{-1}(\Omega)$ and $\mu \in \mathbb{R}$ the system

$$\begin{aligned} -\Delta\delta\Theta - g(\boldsymbol{\varphi})\delta\Theta - \delta\lambda\Theta - \lambda\delta\Theta &= v, \\ (\Theta, \delta\Theta) &= \mu \end{aligned} \tag{7.5}$$

has a solution $(\delta\lambda, \delta\Theta)$. In what follows, we will construct a solution pair $(\delta\Theta, \delta\lambda)$ associated with (v, μ). We use aspects of the proof of [26, section 6.2, Theorem 4]. Fix some $\gamma > M + \lambda$ and define the mappping $\mathcal{L}_\gamma := -\Delta - g(\boldsymbol{\varphi}) - \lambda I + \gamma I$. Since $\gamma > M + \lambda$, the operator $\mathcal{L}_\gamma$ is $H_0^1(\Omega)$-coercive, bounded, and linear. In what follows, we let $\mathcal{L} := \mathcal{L}_0$. Hence, $\mathcal{L}_\gamma^{-1}$ exists. Moreover, since the canonical embedding $E_{1,-1}$ of $H_0^1(\Omega)$ into $H^{-1}(\Omega)$ is compact, the operator $K := (E_{1,-1} \circ \mathcal{L}_\gamma^{-1})$ is a compact linear operator from $H^{-1}(\Omega)$ into itself.

Note that for the sake of making the compact embedding of $H_0^1(\Omega)$ into $H^{-1}(\Omega)$ explicit, we include the embedding operator $E_{1,-1}$. However, we have left this out of the notation on many other occasions for the sake of readability, e.g., in the definition of $\mathcal{L}_\gamma$.

The dual operator of K, denoted by K', is given by $K' = \mathcal{L}_\gamma^{-1} E_{1,-1}$. This is a mapping from $H_0^1(\Omega)$ into itself. The latter follows from the fact that $E_{1,-1} : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by $E_{1,-1}\psi = (\psi, \cdot)_{L^2}$, where $\psi \in H_0^1(\Omega)$. Therefore, for any $\xi \in H_0^1(\Omega)$, we have $\langle E_{1,-1}\psi, \xi\rangle = (\psi, \xi)_{L^2} = \langle \psi, E_{1,-1}\xi\rangle$. Hence, $E_{1,-1}$ coincides with its dual operator. Similarly, for some $h \in H^{-1}(\Omega)$, there exists a unique $z_h := \mathcal{L}_\gamma^{-1} h \in H_0^1(\Omega)$. Then given an arbitrary $k \in H^{-1}(\Omega)$ we have

$$\langle \mathcal{L}_\gamma^{-1} h, k\rangle = \langle z_h, k\rangle = \langle z_h, \mathcal{L}_\gamma \mathcal{L}_\gamma^{-1} k\rangle = \langle \mathcal{L}_\gamma' z_h, \mathcal{L}_\gamma^{-1} k\rangle = \langle \mathcal{L}_\gamma z_h, \mathcal{L}_\gamma^{-1} k\rangle = \langle h, \mathcal{L}_\gamma^{-1} k\rangle.$$

The second-to-last equality follows from the specific form of $\mathcal{L}_\gamma$. Hence, $\mathcal{L}_\gamma^{-1}$ also coincides with its dual operator.

Next, using $R : H_0^1(\Omega) \to H^{-1}(\Omega)$ with $R = -\Delta$ as the Riesz isometry, we define the adjoint $K^* : H^{-1}(\Omega) \to H^{-1}(\Omega)$ of K by $K^* = RK'R^{-1} = -\Delta\mathcal{L}_\gamma^{-1} E_{1,-1}(-\Delta)^{-1}$. In addition, we observe that $K'\Theta = \gamma^{-1}\Theta$, since $z = K'\Theta = \mathcal{L}_\gamma^{-1} E_{1,-1}\Theta$ means

$$\mathcal{L}_\gamma z = E_{1,-1}\Theta \Leftrightarrow [\mathcal{L} + \gamma] z = E_{1,-1}\Theta \Rightarrow z = \gamma^{-1}\Theta. \tag{7.6}$$

This property carries over to the adjoint as well since $K^* R\Theta = RK'R^{-1}R\Theta = RK'\Theta = \gamma^{-1}R\Theta$, i.e., $K^* R\Theta = \gamma^{-1}R\Theta$.

Continuing, we use the Fredholm alternative (see, e.g., [26, Appendix D, Theorem 5]), which implies

$$\mathrm{Rng}(\gamma^{-1}I - K) = \mathrm{Ker}(\gamma^{-1}I - K^*)^\perp. \tag{7.7}$$

Here, I is the identity on $H^{-1}(\Omega)$ and the orthogonal complement is defined using the inner product on $H^{-1}(\Omega)$.

Next consider that for $w \in \mathrm{Ker}(\gamma^{-1}I - K^*)$, $w \in H^{-1}(\Omega)$, we have

$$\begin{aligned} \gamma^{-1}w - K^*w = 0 &\Leftrightarrow \gamma^{-1}w - RK'R^{-1}w = 0 \\ &\Leftrightarrow \gamma^{-1}R^{-1}w - \mathcal{L}_\gamma^{-1}E_{1,-1}R^{-1}w = 0. \end{aligned}$$

But then $\mathcal{L}_\gamma R^{-1}w = \gamma E_{1,-1}R^{-1}w \Rightarrow [\mathcal{L}+\gamma]R^{-1}w = \gamma E_{1,-1}R^{-1}w$, which furthermore implies $\mathcal{L}R^{-1}w = 0 \Rightarrow R^{-1}w = t\Theta$ for $t \in \mathbb{R}$. Hence, $w = tR\Theta$.

Conversely, for any $t \in \mathbb{R}$, we can show that $tR\Theta \in \mathrm{Ker}(\gamma^{-1}I - K^*)$ using an analogous argument. Hence, it follows from this and (7.7) that

$$\mathrm{span}(R\Theta)^{\perp} = \mathrm{Rng}(\gamma^{-1}I - K). \tag{7.8}$$

In fact, for any for any $h \in \mathrm{span}(R\Theta)^{\perp}$ it follows from (7.6) that

$$(Kh, R\Theta) = (h, K^*R\Theta) = \gamma^{-1}(h, R\Theta) = 0, \tag{7.9}$$

where $(\cdot,\cdot)$ represents the inner product on $H^{-1}(\Omega)$, i.e., for $\xi, \eta \in H^{-1}(\Omega)$ we have $(\xi,\eta)_{H^{-1}(\Omega)} = (\nabla R^{-1}\xi, \nabla R^{-1}\eta)_{L^2(\Omega)}$. Hence,

$$Kh \in \mathrm{span}(R\Theta)^{\perp}, \tag{7.10}$$

as well. Then, taking v from (7.5), we observe that

$$\begin{aligned}\langle v - (v, R\Theta)_{H^{-1}} E_{1,-1}\Theta, \Theta\rangle_{H^{-1},H_0^1} &= \langle v, \Theta\rangle_{H^{-1},H_0^1} - (v, R\Theta)_{H^{-1}}(\Theta,\Theta)_{L^2} \\ &= (v, R\Theta)_{H^{-1}} - (v, R\Theta)_{H^{-1}} \cdot 1 = 0,\end{aligned}$$

where we once again make use of the Riesz representation theorem. It follows that $(v - (v, R\Theta)E_{1,-1}\Theta) \in \mathrm{span}(R\Theta)^{\perp}$. Furthermore, by (7.10), we also have $K(v - (v, R\Theta)E_{1,-1}\Theta) \in \mathrm{span}(R\Theta)^{\perp}$. Then by the Fredholm alternative theorem, in particular due to (7.8), there exists a $h \in H^{-1}(\Omega)$ such that

$$\gamma^{-1}h - Kh = K(v - (v, R\Theta)E_{1,-1}\Theta).$$

In fact, as the above equality implies $h = E_{1,-1}(\gamma\mathcal{L}_\gamma^{-1}(h + v - (v, R\Theta)E_{1,-1}\Theta))$ we readily infer that $h \in H_0^1(\Omega)$. Furthermore, it follows that for any $\psi \in H_0^1(\Omega)$

$$\langle \gamma^{-1}\mathcal{L}h, \psi\rangle = \langle v, \psi\rangle - \langle v, \Theta\rangle(\Theta, \psi). \tag{7.11}$$

Now define $\delta\Theta := \gamma^{-1}h + (\mu - (\gamma^{-1}h, \Theta))\Theta,\ \delta\lambda := -\langle v, \Theta\rangle$. Then we have for any $\psi \in H_0^1(\Omega)$ that

$$\langle \mathcal{L}\delta\Theta - \delta\lambda\Theta, \psi\rangle = \langle \gamma^{-1}\mathcal{L}h, \psi\rangle + (\mu - (\gamma^{-1}h, \Theta))\langle \mathcal{L}\Theta, \psi\rangle + \langle v, \Theta\rangle\langle \Theta, \psi\rangle = \langle v, \psi\rangle$$

due to (7.11) and $\mathcal{L}\Theta = 0$ due to the definition of eigenfunction. But this means that $(\delta\Theta, \delta\lambda)$ solves the first equation in (7.5). Since obviously $(\Theta, \delta\Theta) = \mu$, the second equality holds true as well. Thus, we have verified the assumptions of the implicit function theorem. ∎

Acknowledgments. The authors would like to cordially thank Marita Thomas for the many helpful discussions and comments. In addition, we are indebted to the careful reading and helpful comments by the two anonymous reviewers.

REFERENCES

[1] L. Adam, M. Hintermüller, and T. M. Surowiec, *A PDE-constrained optimization approach for topology optimization of strained photonic devices*, Optim. Eng., 19 (2018), pp. 521–557.

[2] L. Adam, M. Hintermüller, and T. M. Surowiec, *A semismooth Newton method with analytical path-following for the H^1-projection onto the Gibbs simplex*, IMA J. Numer. Anal. (2018), https://doi.org/10.1093/imanum/dry034.

[3] R. A. Adams and J. J.-F. Fournier, *Sobolev Spaces*, 2nd ed., Elsevier, Amsterdam, 2008.

[4] G. Allaire, *Shape Optimization by the Homogenization Method*, Appl. Math. Sci. 146, Springer, New York, 2002.

[5] G. Allaire, C. Dapogny, and P. Frey, *Shape optimization with a level set based mesh evolution method*, Comput. Methods Appl. Mech. Engrg., 282 (2014), pp. 22–53.
[6] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., Oxford University Press, New York, 2000.
[7] J. Appell and P. P. Zabrejko, *Nonlinear Superposition Operators*, Cambridge Tracts in Math. 95, Cambridge University Press, Cambridge, UK, 1990.
[8] H. Attouch, G. Buttazzo, and G. Michaille, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*, MOS-SIAM Ser. Optim. 17, SIAM, Philadelphia, 2006.
[9] S. Baldo, *Minimal interface criterion for phase transitions in mixtures of Cahn-Hilliard fluids*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 67–90.
[10] D. Bednarczyk and J. Bednarczyk, *The approximation of the Fermi-Dirac integral $F_{1/2}(\eta)$*, Phys. Lett. A, 64 (1978), pp. 409–410.
[11] M. Bendsøe and O. Sigmund, *Topology Optimization: Theory, Methods, and Applications*, Springer, Berlin, 2003.
[12] D. P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automat. Control, 21 (1976), pp. 174–184.
[13] L. Blank, H. Garcke, M. H. Farshbaf-Shaker, and V. Styles, *Relating phase field and sharp interface approaches to structural topology optimization*, ESAIM Control. Optim. Calc. Var., 20 (2014), pp. 1025–1058.
[14] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.
[15] S. C. Brenner and C. Carstensen, *Finite element methods*, in Encyclopedia of Computational Mechanics Second Edition, Wiley Online Library, 2017, pp. 1–47.
[16] M. Burger and M. Hintermüller, *Projected gradient flows for BV/level set relaxation*, PAMM, 5 (2005), pp. 11–14.
[17] M. Burger and R. Stainko, *Phase-field relaxation of topology optimization with local stress constraints*, SIAM J. Control Optim., 45 (2006), pp. 1447–1466.
[18] R. E. Camacho-Aguilera, Y. Cai, N. Patel, J. T. Bessette, M. Romagnoli, L. C. Kimerling, and J. Michel, *An electrically pumped germanium laser*, Opt. Express, 20 (2012), pp. 11316–11320.
[19] G. Capellini, M. De Seta, P. Zaumseil, G. Kozlowski, and T. Schroeder, *High temperature x ray diffraction measurements on Ge/Si* (001) *heterostructures: A study on the residual tensile strain*, J. Appl. Phys., 111 (2012), 073518.
[20] G. Capellini, C. Reich, S. Guha, Y. Yamamoto, M. Lisker, M. Virgilio, A. Ghrib, M. E. Kurdi, P. Boucaud, B. Tillack, and T. Schroeder, *Tensile Ge microstructures for lasing fabricated by means of a silicon complementary metal-oxide-semiconductor process*, Opt. Express, 22 (2014), pp. 399–410.
[21] S. Chuang, *Physics of Optoelectronic Devices*, Wiley Ser. Pure Appl. Optics, Wiley, New York, 1995.
[22] F. de Gournay, *Velocity extension for the level-set method and multiple eigenvalues in shape optimization*, SIAM J. Control Optim., 45 (2006), pp. 343–367.
[23] M. A. der Maur, G. Penazzi, G. Romano, F. Sacconi, A. Pecchia, and A. Di Carlo, *The multiscale paradigm in electronic device simulation*, IEEE Trans. Electron. Devices, 58 (2011), pp. 1425–1432.
[24] B. Dutt, D. S. Sukhdeo, D. Nam, B. M. Vulovic, Z. Yuan, and K. C. Saraswat, *Roadmap to an efficient germanium-on-silicon laser: Strain vs. n-type doping*, IEEE Photon. J., 4 (2012), pp. 2002–2009.
[25] M. El Kurdi, G. Fishman, S. Sauvage, and P. Boucaud, *Band structure and optical gain of tensile-strained germanium based on a* 30 *band $k \cdot p$ formalism*, J. Appl. Phys., 107 (2010), 013710.
[26] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1994.
[27] P. J. Frey and P.-L. George, *Mesh Generation: Application to Finite Elements*, 2nd ed., ISTE, London, 2008.
[28] S. Funken, D. Praetorius, and P. Wissgott, *Efficient implementation of adaptive P*1*-FEM in Matlab*, Comput. Methods Appl. Math., 11 (2011), pp. 460–490.
[29] A. Ghrib, M. El Kurdi, M. De Kersauson, M. Prost, S. Sauvage, X. Checoury, G. Beaudoin, I. Sagnes, and P. Boucaud, *Tensile-strained germanium microdisks*, Appl. Phys. Lett., 102 (2013), 221112.
[30] A. Ghrib, M. El Kurdi, M. Prost, S. Sauvage, X. Checoury, G. Beaudoin, M. Chaigneau, R. Ossikovski, I. Sagnes, and P. Boucaud, *All-around SiN stressor for high and ho-*

mogeneous tensile strain in germanium microdisk cavities, Adv. Opt. Mater., 3 (2015), pp. 353–358.
[31] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 3rd ed., Springer, Berlin, 2001.
[32] H. Goldberg, W. Kampowsky, and F. Tröltzsch, *On Nemytskij operators in L^p-spaces of abstract functions*, Math. Nachr., 155 (1992), pp. 127–140.
[33] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.
[34] J. Haslinger and P. Neittaanmäki, *Finite Element Approximation for Optimal Shape Design: Theory and Applications*, Wiley, New York, 1988.
[35] A. Henrot, *Extremum problems for eigenvalues of elliptic operators*, Frontiers in Math., Birkhäuser, Basel, 2006.
[36] A. Henrot and M. Pierre, *Variation et optimisation de formes: Une analyse géométrique*, Math. Appl. (Berlin) 48, Springer, Berlin, 2005.
[37] M. Hintermüller, M. Hinze, and M. H. Tber, *An adaptive finite-element Moreau-Yosida-based solver for a non-smooth Cahn-Hilliard problem*, Optim. Methods Softw., 26 (2011), pp. 777–811.
[38] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
[39] M. Hinze and R. Pinnau, *An optimal control approach to semiconductor design*, Math. Models Methods Appl. Sci., 12 (2002), pp. 89–107.
[40] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Math. Model. Theory Appl. 23, Springer, New York, 2009.
[41] J. R. Jain, A. Hryciw, T. M. Baer, D. A. Miller, M. L. Brongersma, and R. T. Howe, *A micromachining-based technology for enhancing germanium light emission via tensile strain*, Nat. Photonics, 6 (2012), 398.
[42] R. Lehoucq, D. Sorensen, and C. Yang, *ARPACK Users' Guide*, SIAM, Philadelphia, 1998.
[43] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, Zh. Vychisl. Mat. Mat. Fiz., 6 (1966), pp. 787–823.
[44] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Grundlehren Math. Wiss. 170, Springer, Berlin, 1971.
[45] J. Liu, X. Sun, R. Camacho-Aguilera, L. C. Kimerling, and J. Michel, *Ge-on-Si laser operating at room temperature*, Opt. Lett., 35 (2010), pp. 679–681.
[46] Z. Lu, *Dynamics of Wing Cracks and Nanoscale Damage in Silica Glass*, Ph.D. thesis, University of Southern California, 2007.
[47] P. A. Markowich, *The Stationary Semiconductor Device Equations*, Springer, New York, 1986.
[48] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Ration. Mech. Anal., 98 (1987), pp. 123–142.
[49] A. A. Novotny and J. Sokołowski, *Topological Derivatives in Shape Optimization*, Interact. Mech. Math., Springer, Heidelberg, 2013.
[50] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[51] D. Peschka, N. Rotundo, and M. Thomas, *Doping optimization for optoelectronic devices*, Opt. Quant. Electron., 50 (2018), 125.
[52] D. Peschka, M. Thomas, A. Glitzky, R. Nurnberg, K. Gartner, M. Virgilio, S. Guha, T. Schroeder, G. Capellini, and T. Koprucki, *Modeling of edge-emitting lasers based on tensile strained germanium microstrips*, IEEE Photon. J., 7 (2015).
[53] D. Peschka, M. Thomas, A. Glitzky, R. Nürnberg, M. Virgilio, S. Guha, T. Schroeder, G. Capellini, and T. Koprucki, *Robustness analysis of a device concept for edge-emitting lasers based on strained germanium*, Opt. Quant. Electron., 48 (2016).
[54] J. Piprek, *Nitride Semiconductor Devices: Principles and Simulation*, Wiley, New York, 2007.
[55] D. L. Scharfetter and H. K. Gummel, *Large-signal analysis of a silicon read diode oscillator*, IEEE Trans. Electron. Devices, 16 (1969), pp. 64–77.
[56] O. Sigmund and S. Torquato, *Design of materials with extreme thermal expansion using a three-phase topology optimization method*, J. Mech. Phys. Solids, 45 (1997), pp. 1037–1067.
[57] O. Sigmund and S. Torquato, *Design of smart composite materials using topology optimization*, Smart Mater. Struct., 8 (1999), pp. 365–379.
[58] M. J. Suess, R. Geiger, R. A. Minamisawa, G. Schiefler, J. Frigerio, D. Chrastina, G. Isella, R. Spolenak, J. Faist, and H. Sigg, *Analysis of enhanced light emission from highly strained germanium microbridges*, Nat. Photon, 7 (2013), pp. 466–472.

[59] X. Sun, L. Jifeng, L. Kimerling, and J. Michel, *Toward a germanium laser for integrated silicon photonics*, IEEE J. Sel. Top. Quantum Electron., 16 (2010), pp. 124–131.
[60] A. Takezawa, S. Nishiwaki, and M. Kitamura, *Shape and topology optimization based on the phase field method and sensitivity analysis*, J. Comput. Phys., 229 (2010), pp. 2697–2718.
[61] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis*, North-Holland, Amsterdam, 1977.
[62] F. Tröltzsch, *Optimal Control of Partial Differential Equations*, Grad. Stud. Math. 112, AMS, Providence, RI, 2010.
[63] W. van Roosbroeck, *Theory of the flow of electrons and holes in germanium and other semiconductors*, Bell. Syst. Tech. J, 29 (1950), pp. 560–607.
[64] J. J. Vlassak and W. D. Nix, *A new bulge test technique for the determination of Young's modulus and Poisson's ratio of thin films*, J. Mat. Res., 7 (1992), pp. 3242–3249.
[65] S. Wirths, R. Geiger, N. von den Driesch, G. Mussler, T. Stoica, S. Mantl, Z. Ikonic, M. Luysberg, S. Chiussi, J. M. Hartmann, H. Sigg, J. Faist, D. Buca, and D. Grutzmacher, *Lasing in direct-bandgap GeSn alloy grown on Si*, Nat. Photon., 9 (2015), pp. 88–92.
[66] J. J. Wortman and R. A. Evans, *Young's modulus, shear modulus, and Poisson's ratio in silicon and germanium*, J. Appl. Phys., 36 (1965), pp. 153–156.
[67] E. Zeidler, *Nonlinear Functional Analysis and its Applications* I: *Fixed-Point Theorems*, Springer, New York, 1986.
[68] S. Zhou and M. Y. Wang, *Multimaterial structural topology optimization with a generalized Cahn–Hilliard model of multiphase transition*, Struct. Multidiscip. Optim., 33 (2006), pp. 89–111.