Quantum transport for the gate-length scaling limit of Si nanowire field-effect transistors based on calibrated k · p Hamiltonian parameters

Guohui Zhan,[1,2,3,*] Tongshuai Zhu,[4,5] Jiaxin Yao,[3] Kun Luo,[3] Huaixiang Yin,[2,3] Shengli Zhang,[6] and Zhenhua Wu[1,2,†]

[1] *Center for Quantum Matter, School of Physics, Zhejiang University, Hangzhou, 310058 Zhejiang, China*
[2] *University of Chinese Academy of Sciences, 100049 Beijing, China*
[3] *Institute of Microelectronics, Chinese Academy of Sciences, 100029 Beijing, China*
[4] *College of Science, China University of Petroleum (East China), Qingdao, 266580 Shandong, China*
[5] *School of Materials Science and Engineering, China University of Petroleum (East China), Qingdao, 266580 Shandong, China*
[6] *Key Laboratory of Advanced Display Materials and Devices, College of Material Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094 Jiangsu, China*

(Received 28 February 2024; revised 22 November 2024; accepted 21 February 2025; published 20 March 2025)

We present a comprehensive investigation of quantum transport in silicon nanowire field-effect transistors (SiNWFETs) at the scaling limit. The Si bulk **k** · **p** Hamiltonian parameters are rendered invalid at smaller scales due to pronounced quantum confinement effects. Consequently, nanowire **k** · **p** Hamiltonian parameters are meticulously calibrated with the use of first-principles HSE06 band structures through the fast least-squares method. On the basis of the nonequilibrium Green's function formalism, we then systematically analyze the performance limits of gate-all-around SiNWFETs under various gate lengths. With diminishing gate lengths, quantum tunneling from source to drain intensifies, leading to a degradation in subthreshold swing. Our findings reveal that the gate-length scaling limit for *n*-type devices is numerically larger than for *p*-type devices, and distinct gate scaling limits are elucidated for various cross-section sizes; for instance, the ultimate scaling limit of Si nanowires of 4×4 nm^2 cross section is approximately 10 nm.

DOI: 10.1103/PhysRevApplied.23.034049

I. INTRODUCTION

With the evolution of complementary metal oxide semiconductor (CMOS) technology, effective suppression of the short-channel effect becomes increasingly imperative [1,2]. Gate-all-around (GAA) nanowire field-effect transistors (NWFETs) have emerged as the predominant device architecture due to their superior gate control compared with fin FETs. NWFETs, including lateral and stacked GAA configurations [3–6], are considered promising candidates for the next generation of transistors. According to the 2022 edition of the International Roadmap for Devices and Systems (IRDS) [6], CMOS technology is progressing toward the "0.5 nm eq" technology node by 2037, where the physical gate length is 12 nm. At this scale, and with a smaller nanowire cross section, the classical approach based on the drift-diffusion equation becomes invalid. Instead, a quantization model based on solving the Schrödinger equation becomes necessary, considering that the transport is quasiballistic [7,8]. A significant quantum effect arises, making quantum tunneling from source to drain non-negligible. Thus, an advanced quantum transport analysis of NWFETs is crucial to ascertain optimal performance. In such cases, nonequilibrium Green's function (NEGF) simulations are preferred, as they accurately capture both quantum confinement and source-to-drain tunneling.

In the early years, the single-band parabolic effective mass approximation was commonly used to describe electronic properties in materials for simulating quantum ballistic transport in various devices, such as the quantum ballistic transport of a resonant tunneling diode [9,10], a double-gate MOS [11,12], and NWFETs [13–15]. However, this approach often fails without one considering some key quantum effects, such as inaccurate higher energy interval, or the strong coupling between heavy holes and light holes in the valence band. The bands of Si nanowires, on the basis of first-principles calculations, and tight-binding (TB) and **k** · **p** methods, have been studied extensively, providing an explicit physical picture to understand the transport behavior of NWFETs.

*Contact author: zhanguohui@ime.ac.cn
†Contact author: wuzhenhua@zju.edu.cn

Liu *et al.* [16] explored the performance limits of GAA NWFETs through *ab initio* quantum transport simulations. Luisier *et al.* [17] performed atomistic simulations of nanowires using the $sp^3d^5s^*$ TB formalism. Shin *et al.* [18–21] developed a highly efficient computational simulator by reducing the Hamiltonian size through transformation to the mode-space basis. Full quantum transport simulations of large-scale NWFETs based on the $\mathbf{k}\cdot\mathbf{p}$ mode-space method often effectively reduce computational costs [20,22,23]. As the device is miniaturized and the material approaches the nanoscale, many parameters of the bulk material fail due to significant quantum effects, surface effects, lattice distortions, defects, etc. It is urgently required to refine these bulk parameters for nanowires with differing widths and heights. Several attempts have been made to optimize these parameters [21,24], determining the nanowire $\mathbf{k}\cdot\mathbf{p}$ parameters by achieving the best match between the band structures obtained from the $\mathbf{k}\cdot\mathbf{p}$ method and the TB method, Wang *et al.* [25] proposed a machine learning method to fit Slater-Koster TB parameters from the *ab initio* band structure.

In this work, we adapt more general first-principles band structures as reference data, ensuring Si nanowire surfaces are passivated and relaxed. Using a fast least-squares curve-fitting approach, we identify a suitable parameter space for the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian, calibrating bulk parameters for different nanowire cross-section sizes. As Moore's law continues to drive scaling down, would it continue to shrink to sub-5-nm physical gate length? [26,27] To address this, on the basis of these calibrated parameters, we conduct a comprehensive simulation to explore the performance limits of Si nanowire FETs (SiNWFETs) with gate lengths ranging from 4 to 12 nm. The key quantum tunneling aspects are theoretically investigated with the use of the NEGF formalism. As the gate length decreases, the tunneling ratio of total current increases, and the subthreshold swing (SS) degrades. Notably, we observe that the gate scaling limit for *n*-type devices surpasses that of *p*-type devices, and the ultimate scaling gate length for a 4 nm × 4 nm cross section is approximately 10 nm.

The rest of this paper is organized as follows. The computational method and simulation details are provided in Sec. II. In Secs. III A and III B, we report first-principles calculations and the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian of Si nanowires, respectively. Calibration of the bulk $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian parameters is discussed in Sec. III C, and Secs. III D and III E explore the device performance limits of GAA SiNWFETs. A summary is provided in Sec. IV.

II. COMPUTATIONAL METHODS

The electronic properties of Si nanowires are performed by Synopsys QuantumATK T-2022.03 [28]. The optimal geometric structures are relaxed, our focusing primarily on the Si and H atoms near the nanowire edges. We set the force tolerance to 0.05 eV/Å and the stress error tolerance to 0.1 GPa, using the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) optimizer method. The generalized gradient approximation [29] in conjunction with the Heyd-Scuseria-Ernzerhof (HSE) hybrid functional [30] is used for all band structure simulations. We use $1\times1\times7$ **k**-point meshes for relaxation and self-consistent calculations. The complex band structures of Si nanowires are calculated by a semiempirical method with the use of the Hückel basis set [31].

The calibration of the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian parameters is done with our in-house MATLAB codes. Using the first-principles HSE06 calculations as a reference, we optimize the bulk $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian parameters via the least-squares curve-fitting method. In addition, the machine learning algorithm and neural network architecture can be conveniently achieved with PyTorch platforms, which use its built-in Adam optimizer. Our codes are publicly available for downloading [32].

The quantum transport simulations based on the NEGF formalism are performed with the use of the QTX module in Synopsys Sentaurus TCAD [33]. This module accounts for quantum mechanical effects such as confinement, source-to-drain tunneling, and coherence effects such as resonant tunneling. The calibrated $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian parameters are incorporated into the input files. For simplicity, the spin-orbit coupling effect is ignored, and the valley model of the conduction and valence bands is described with the use of two-band and three-band $\mathbf{k}\cdot\mathbf{p}$ models. A lumped series resistance of 2000 Ω is included in the simulations to model the source (drain) contact resistance and the source (drain) epidiffused resistance. The simulation process began with the Sentaurus Process to generate the geometry and doping profile. Key formulas of the NEGF framework [34,35] were used, where the density matrix ρ was calculated as

$$\rho = \frac{1}{2\pi}\int G^r(\Gamma_L f_L + \Gamma_R f_R)G^a dE. \tag{1}$$

The retarded G^r and advanced G^a Green's functions are defined as $G^r = [E - H - \Sigma_L - \Sigma_R]^{-1}$ and its Hermitian conjugate for G^a, where H is the Hamiltonian of the nanowire system, and Σ_L and Σ_R denote the self-energy of the left and right electrodes. Once the self-consistent potential is obtained by our solving the coupled Poisson equation, the current I passing through the SiNWFETs is calculated:

$$I = \frac{2e}{h}\int \mathrm{Tr}[G^r\Gamma_L G^a\Gamma_R](f_L - f_R)dE, \tag{2}$$

where $\mathrm{Tr}[G^r\Gamma_L G^a\Gamma_R]$ is the transmission spectrum, and Γ_L and Γ_R represent the broadening functions of the left and right electrodes, e and h denote the electron charge

and the Planck constant, respectively, and f_L and f_R are the Fermi-Dirac distribution of the left and right electrodes. Besides the above ballistic simulation framework, dissipative transport of SiNWFETs was also modeled by our incorporating phonon, Coulomb, and surface roughness (SR) scattering mechanisms in the QTX module [33]. The Dyson and Keldysh equations were solved iteratively with open boundary conditions to capture nonequilibrium carrier dynamics.

III. RESULTS AND DISCUSSION

A. First-principles calculations of Si nanowires

For large-scale Si nanowires, the band structures can be efficiently obtained with the use of an effective mass approximation model. In addition, the Hamiltonian based on semiempirical TB parameters is well suited for certain bulk semiconductors [36–39]. Nonetheless, this approach has limitations when one is dealing with surface passivation in nanowire structures and is highly dependent on TB parameters. To accurately calculate the band structure of sub-10-nm Si nanowires, first-principles density functional theory (DFT) calculations are preferable. The realistic HSE06 hybrid functional calculations are used for all simulations. Taking Si nanowires along the [100] direction as an example [as illustrated in Fig. 1(b)], we construct Si nanowires with various rectangular cross sections ranging from 2×2 nm^2 to 6×6 nm^2. The surface dangling bonds are passivated with H atoms, and the resulting structures are optimized to attain dynamical stability. The HSE06 band structures for different cross-section sizes are depicted in Fig. 2. The conduction edge at the Γ point is formed from the four equivalent Δ valleys (Δ_4), and the next-higher valleys are formed from the two equivalent Δ valleys (Δ_2). Additionally, the complex band structures of Si nanowires are calculated by a semiempirical method, as shown in Fig. S5 [40]. The imaginary part κ of the complex wave vector dominates the transport properties, and its product with the barrier thickness defines the exponential decay of the wave functions within the barrier and determines essentially the transmission probability. As the cross-section size (nanowire thickness) increases, the bands around the Fermi surface become denser, and the band gap approaches that of the bulk gap [Fig. 2(f)]. It is noteworthy to observe that Si nanowires with smaller cross-section areas exhibit a significantly higher proportion of surface-passivated hydrogen atoms. After optimization, the change in positions of the surface atoms differs significantly from that of the interior atoms of the nanowires. This observation indicates that the default bulk parameters may no longer be valid for these nanoscale materials, necessitating a modification of parameters.

B. k · p Hamiltonian of Si nanowires

Si is the fundamental material in the realm of modern nanoelectronics, possessing a diamond lattice structure with a face-centered-cubic Bravais lattice, comprising two identical atoms per unit cell. In this work, the $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian of bulk Si is simplified by our ignoring the spin-orbit coupling effect.

The lowest conduction band edge of bulk Si is situated around the X point, and there are six equivalent X points along the k_x, k_y, and k_z axes in the Brillouin zone. Taking the X point of the [001] valley as an example, we can express the effective two-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian in the vicinity of the X point as follows [41]:

$$\mathcal{H}_c(\mathbf{k}) = \left\{ \frac{\hbar^2 k_z^2}{2m_l} + \frac{\hbar^2 (k_x^2 + k_y^2)}{2m_t} + U \right\} \mathbf{I} + \begin{bmatrix} -\frac{\hbar^2 k_0 k_z}{m_l} & \frac{\hbar^2 k_x k_y}{M} \\ \frac{\hbar^2 k_x k_y}{M} & \frac{\hbar^2 k_0 k_z}{m_l} \end{bmatrix}, \tag{3}$$

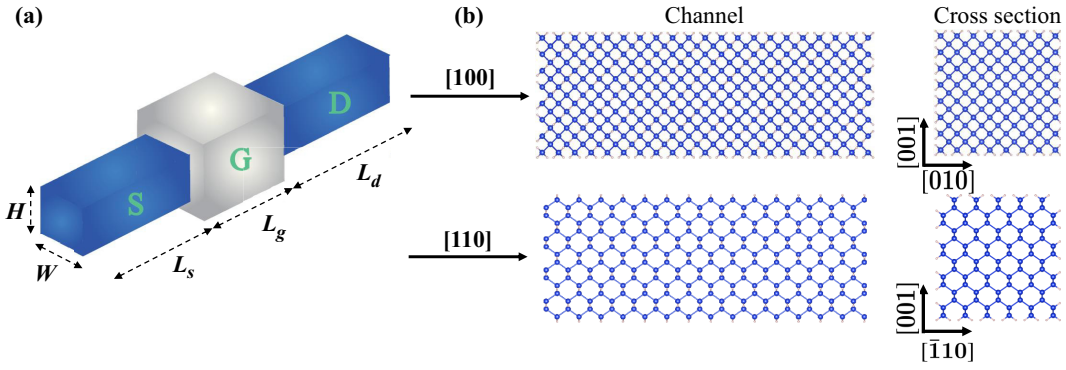

FIG. 1. Schematic representation of GAA SiNWFETs. (a) A simple model of GAA transistors, where H denotes the height, W denotes the width, L_g denotes the gate length, and L_s and L_d denote the length of the source and drain extensions. (b) Atomistic structure along different channel orientations ([100] and [110]). D, drain; G, gate; S, source.

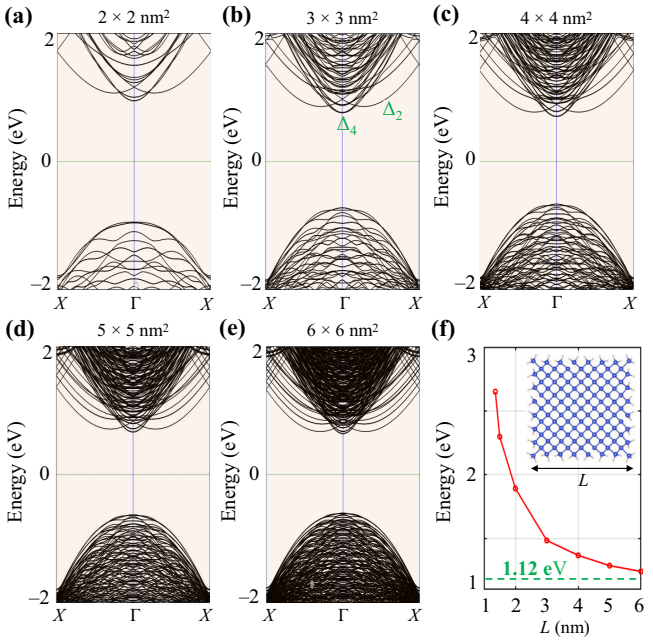

FIG. 2. Band structures of Si nanowires of different cross-section size: (a)–(e) from 2×2 nm^2 to 6×6 nm^2. (f) Band gap as a function of different cross-section size L. The inset shows the relaxed atomic structure of 3×3 nm^2 Si nanowires.

where m_l and m_t denote the longitudinal and transverse effective masses, $M^{-1} \approx m_t^{-1} - m_0^{-1}$, k_0 represents the position of the valley minimum relative to the X point, and $U(z)$ is the confinement potential. $\mathbf{I}$ denotes the identity matrix. Other valleys, such as the [100] and [010] valleys, follow similar forms.

For the valence band, characterized by the Γ_{15} representation based on group theory, the effective three-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian around the Γ point is described by the model [42–45]

$$\mathcal{H}_v(\mathbf{k}) = M\mathbf{k}^2\mathbf{I} + \begin{bmatrix} (L-M)k_x^2 & Nk_xk_y & Nk_xk_z \\ Nk_xk_y & (L-M)k_y^2 & Nk_yk_z \\ Nk_xk_z & Nk_yk_z & (L-M)k_z^2 \end{bmatrix}, \tag{4}$$

$$L = -\frac{\hbar^2}{2m_0}(\gamma_1 + 4\gamma_2), \quad M = -\frac{\hbar^2}{2m_0}(\gamma_1 - 2\gamma_2),$$
$$N = -\frac{\hbar^2}{2m_0}6\gamma_3, \tag{5}$$

where γ_1, γ_2, and γ_3 are Luttinger parameters. It is worth noting that the Hamiltonian forms for some common semiconductor materials, such as Ge and III-V compounds (e.g., GaAs), are analogous [46,47].

Using the bulk $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian, one can derive the Hamiltonian for Si nanowires or nanosheets by discretizing the bulk Hamiltonian along the two open boundaries. Specifically, for nanowires along the [100] direction, open boundaries are applied to k_y and k_z, while k_x remains periodic. The discretization of the bulk Hamiltonian in real space involves the conversion $k_y = -i(\partial/\partial y)$ and $k_z = -i(\partial/\partial z)$. Consequently, $\mathcal{H}_c(\mathbf{k})$ and $\mathcal{H}_v(\mathbf{k})$ are transformed into $\mathcal{H}_c(k_x, y_j, z_k)$ and $\mathcal{H}_v(k_x, y_j, z_k)$, where j and k represent real-space grid indices. For a general orientation [48,49], a rotation of the $\mathbf{k}$-space coordinates is required:

$$\begin{pmatrix} k_x' \\ k_y' \\ k_z' \end{pmatrix} = \mathcal{R} \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix}, \tag{6}$$

where k' denotes the new coordinate system and $\mathcal{R}$ is the rotation matrix. The matrix size of $\mathcal{H}_c(k_x, y_j, z_k)$ and $\mathcal{H}_v(k_x, y_j, z_k)$ depends on the y-z-plane grid of the cross section, and these matrices have the block tridiagonal form (see Appendix).

C. Calibration of bulk k · p Hamiltonian parameters

The real-space matrices $\mathcal{H}_c(k_x)$ and $\mathcal{H}_v(k_x)$ are large, making solving eigenvalue problems computationally demanding. For instance, in the case of 5×5 nm^2 Si nanowires, the matrix size of the Hamiltonian is approximately $10\,000 \times 10\,000$. Following the method presented in Ref. [20], the real-space Hamiltonian $\mathcal{H}$ is transformed to the $\mathbf{k}$ space by Fourier transform, reducing the matrix size through a simple reordering scheme. We focus only on the energy range near the band edge, which dominates the transport properties of the devices. Subsequently, a mode-space approach is used, resulting in a smaller Hamiltonian $\mathcal{H}_M(\mathbf{k})$ without precision being compromised. The $\mathbf{k} \cdot \mathbf{p}$ band structures of Si nanowires are then calculated from the mode-space Hamiltonian $\mathcal{H}_M(\mathbf{k})\Psi_k = E_k\Psi_k$.

Figure 3 illustrates a flowchart for calibrating $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian parameters of nanowires. The initial step involves obtaining the band structures of bulk Si and realistic nanowires (with H atoms adsorbed on the surface) from first-principles calculations. Subsequently, the bulk $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian is constructed on the basis of group theory, and the $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian of Si nanowires is derived according to the crystal orientation. The next step involves dimension reduction of the real-space Hamiltonian matrix $\mathcal{H}_R$, achieved through transformations to $\mathbf{k}$ space or a mode-space approach. Finally, the band of the $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian with $\mathbf{k} \cdot \mathbf{p}$ parameters is compared with that calculated from first-principles HSE06 results. The $\mathbf{k} \cdot \mathbf{p}$ parameters are then adjusted until the desired accuracy is achieved, essentially solving a least-squares problem for function parameters. While similar to the methods in Ref. [25], where optimized $\mathbf{k} \cdot \mathbf{p}$ parameters can be trained

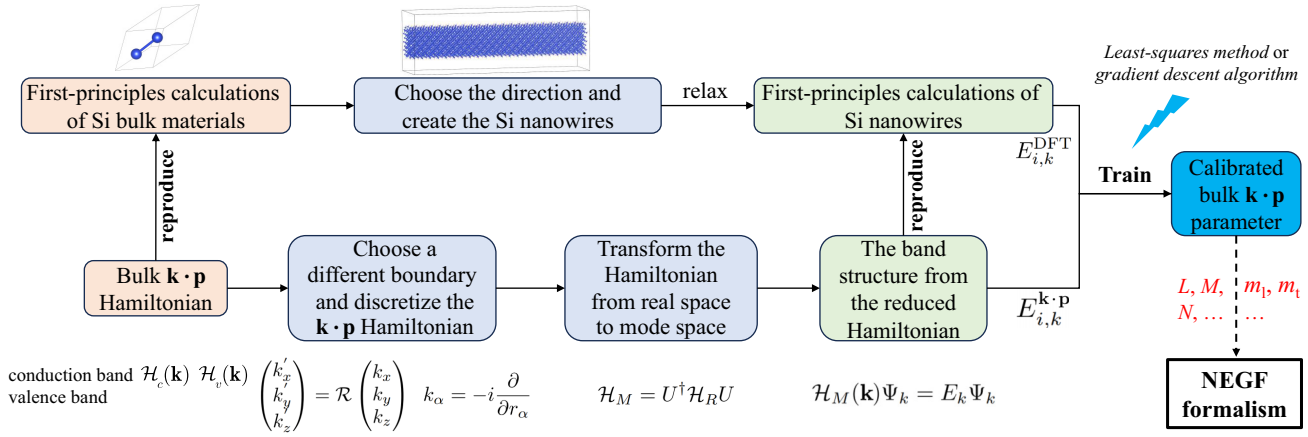

FIG. 3. Workflow for calibrating the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian parameters based on realistic first-principles calculations of Si nanowires. Starting from first principles and with the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian, the parameters are optimized through a least squares curve-fitting approach or the gradient descent algorithm.

using machine learning methods, least-squares method is used in this work for calibration. The optimized $\mathbf{k}\cdot\mathbf{p}$ parameters are trained by our defining a tolerance between $\mathbf{k}\cdot\mathbf{p}$ band $E_{i,k}^{\mathbf{k}\cdot\mathbf{p}}$ and $E_{i,k}^{\mathrm{DFT}}$ from first-principles calculations:

$$\delta = \frac{1}{N_b N_k} \sum_{i=1}^{N_b} \sum_{k=1}^{N_k} w_{i,k} (E_{i,k}^{\mathbf{k}\cdot\mathbf{p}} - E_{i,k}^{\mathrm{DFT}})^2, \tag{7}$$

where N_b and N_k are the number of bands and $\mathbf{k}$ points, respectively, and $w_{i,k}$ represents the weight (0–1) for an energy point for the ith band and kth $\mathbf{k}$ point. The closer E_k is to the band edge (band minimum in the conduction band, band maximum in the valence band), the greater the weight $w_{i,k}$. The tolerance δ is set to 10^{-8}–10^{-6}, and the parameter fitting is accomplished with built-in MATLAB functions such as lsqcurvefit, lsqnonlin, and fminsearch. Additionally, we use a gradient descent algorithm to optimize the $\mathbf{k}\cdot\mathbf{p}$ parameters based on the PyTorch machine learning framework.

Since the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian is derived and reduced by the mode-space method, our fitting process is particularly fast. Tables I and II present the calibrated $\mathbf{k}\cdot\mathbf{p}$ parameters of the conduction and valence bands for different cross sections of Si nanowires, and the bulk $\mathbf{k}\cdot\mathbf{p}$ parameters are also listed. As the cross-section size decreases, the quantum confinement effect becomes more significant, resulting in greater differences in $\mathbf{k}\cdot\mathbf{p}$ parameters. As the cross-section size increases, the calibrated $\mathbf{k}\cdot\mathbf{p}$ parameters of the conduction band approach the bulk parameters. Figure 4 provides a comparison of the conduction and valence bands of 4×4 nm^2 Si nanowires (the results for 3×3 nm^2 and 5×5 nm^2 Si nanowires are depicted in Fig. S3 and S4 [40]). The black lines denote the HSE06 band structures, and Figs. 4(a) and 4(c) and Figs. 4(b) and 4(d) show the band differences based on bulk and calibrated $\mathbf{k}\cdot\mathbf{p}$ parameters, respectively. The band edges are well fitted, and the band shape and band gap are modified. However, it is worth noting that, unlike in Ref. [21], our reference data are based on realistic and atomic first-principles band structures with different Si nanowires. Figure 4 reveals that the band difference is significant, especially for the valence band from the three-band Hamiltonian, emphasizing the inaccuracy of calculating electronic and quantum transport properties using default bulk $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian parameters.

D. Quantum transport properties of GAA SiNWFETs

To investigate the impact of $\mathbf{k}\cdot\mathbf{p}$ parameters on quantum transport characteristics, the calibrated parameters are used as inputs for Sentaurus QTX. A simplified GAA SiNWFET is depicted in Fig. 1(a). The transport direction aligns with the x axis (oriented to the [100] crystal orientation), while the confinement directions align with the

TABLE I. Calibrated $\mathbf{k}\cdot\mathbf{p}$ parameters of the conduction band for [100] direction Si nanowires.

	m_t	m_l	k_0	E_{c0}
3×3 nm^2	0.2438	0.7803	0.1720	0.5127
4×4 nm^2	0.2182	0.8580	0.1725	0.5531
5×5 nm^2	0.196	0.910	0.1730	0.5569
Bulk	0.196	0.916	0.1735	0.562

TABLE II. Calibrated $\mathbf{k}\cdot\mathbf{p}$ parameters of the valence band for [100] direction Si nanowires.

	L	M	N	E_{v0}
3×3 nm^2	−3.5131	−1.4958	−4.3062	−0.5976
4×4 nm^2	−3.5091	−1.2730	−4.0057	−0.6097
5×5 nm^2	−3.5030	−1.2676	−3.9899	−0.5931
Bulk	−5.53	−3.64	−8.32	−0.562

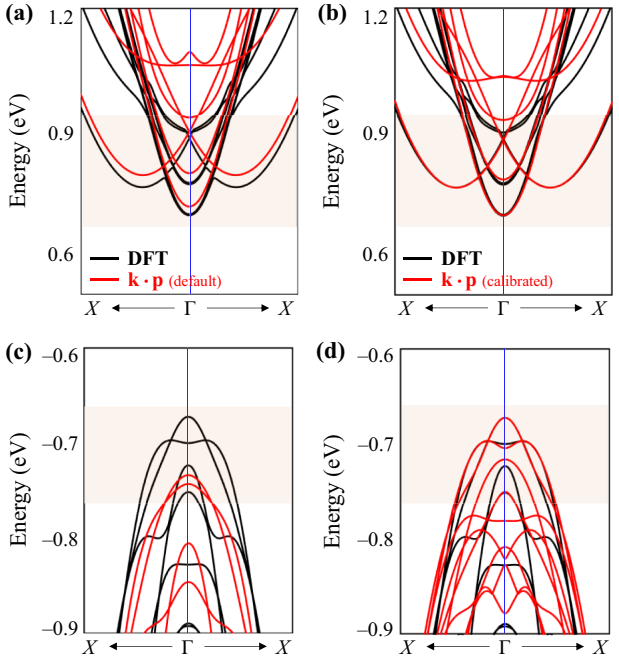

FIG. 4. Comparison of band structures based on **k** · **p** Hamiltonian and first-principles calculations. (a),(b) Conduction band and (c),(d) valence band. Black and red lines represent DFT (HSE06) data and **k** · **p** data, respectively. The Si nanowires have a cross-section size of 4 × 4 nm^2.

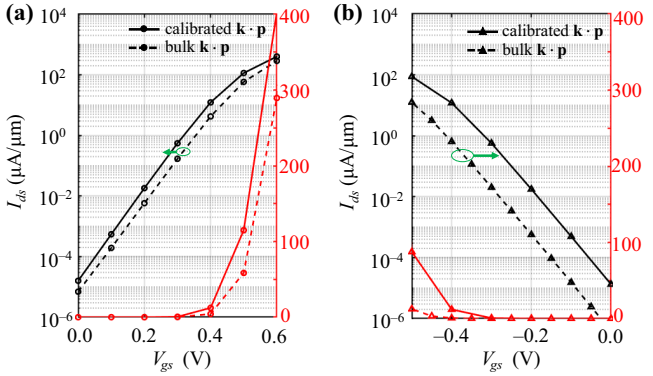

FIG. 5. Comparison of I-V characteristics based on default bulk and calibrated **k** · **p** Hamiltonian parameters. (a) n-type devices and (b) p-type devices. The Si nanowires have a cross-section size of 3 × 3 nm^2, and the gate length of the SiNWFETs is 8 nm.

y and z axes. The channel is undoped (intrinsic), and the source and drain extension doping is set at 10^{20} cm^{-3}. The source and drain doping profiles are created through diffusion at 1100 °C. The source and drain extension length is 10 nm, with the region surrounded by low-k materials, the dielectric constant is $2.5\varepsilon_r$, and the oxide thickness is set to 0.8 nm. The supply voltage V_{dd} is fixed at 0.6 V, following the IRDS 2022 edition requirements [6]. The specified OFF-state current target (I_{off} = 10 nA/μm) and the fixed supply voltage are used to evaluate I_{on}. Figure 5 presents a comparison of the I-V characteristics based on default bulk and calibrated **k** · **p** Hamiltonian parameters. Noticeable threshold voltage V_T drift is observed, especially for p-type devices, which is attributed to the significant variation in the valence band edge of Si nanowires. This observation underscores the unreliability of default bulk **k** · **p** parameters for sub-10-nm nanoscale devices.

E. Gate-length scaling limit of GAA SiNWFETs

In the upcoming "0.5 nm eq" technology node from IRDS 2022 edition [6], the gate length is set at 12 nm. The question arises whether further shrinkage is feasible. To address this, a comprehensive quantum transport simulation of SiNWFETs with cross-section sizes of 3 × 3 nm^2, 4 × 4 nm^2, and 5 × 5 nm^2 under various gate lengths L_g is conducted. The current is normalized by the perimeter of the cross section (12, 16, and 20 nm). By our adjusting the gate voltage, I_{off} corresponds to $V_g = 0$, and I_{on} corresponds to $V_g = V_{dd} = 0.6$ V. More detailed I-V results for SiNWFETs are presented in Figs. S6–S13 [40]. The SS, a crucial metric for measuring the switching efficiency of MOSFETs, is defined as

$$\frac{dV_{gs}}{d\log_{10}(I_{ds})}. \tag{8}$$

It quantifies the gate control capability. Figures 6(a) and 6(b) illustrate the SS of 3 × 3 nm^2, 4 × 4 nm^2, and 5 × 5 nm^2 SiNWFETs. The results indicate that for smaller cross sections, the SS is small, signifying strong gate control. However, as the gate length increases to 12 nm and beyond, the SS values approach the theoretical thermal limit of approximately 60 mV/dec at room temperature. This implies that the tunneling current contribution to the total current in the subthreshold region is minimal in long-channel devices. Nevertheless, as the channel length decreases, the SS degrades, reaching 160 mV/dec at L_g = 4 nm for 5 × 5 nm^2 Si nanowires. Since the OFF-state current is fixed, larger SS values indicate a smaller ON-state current [Fig. 6(c)]. In other words, the source-to-drain tunneling degrades the on:off ratio by increasing the SS, and it becomes particularly significant for devices with short gate lengths. This underscores the necessity for a higher gate work function in the design of devices at the scaling limit.

Beyond pure ballistic transport, we also conducted simulations of dissipative transport of SiNWFETs, considering different scattering mechanisms such as phonon scattering, SR scattering, and Coulomb scattering. Specifically, when SR is included in the QTX module, SR scattering at the silicon–SiO_2 interface is modeled as an intravalley elastic process for both electrons and holes. To capture the characteristics, the exponential power-spectral density function is selected to model the SR . As depicted

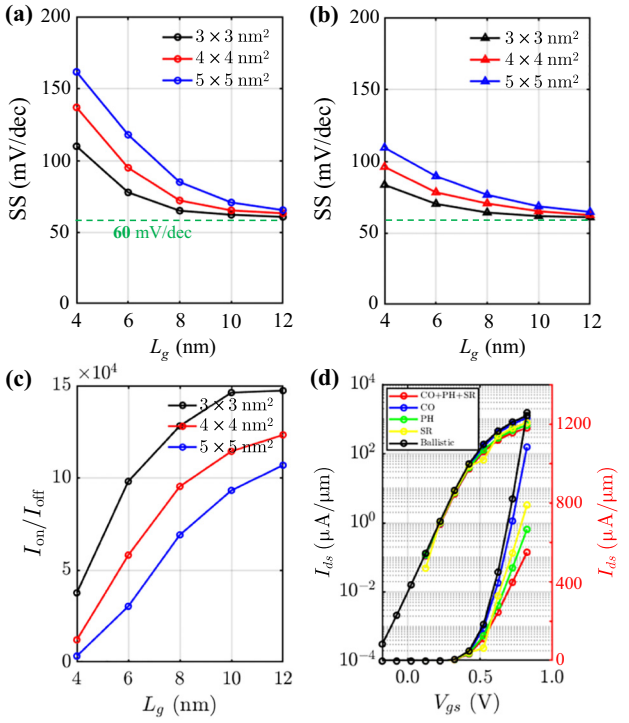

FIG. 6. SS under different gate lengths for (a) *n*-type and (b) *p*-type NWFETs, for which the black, red, and blue lines represent cross-section sizes of 3×3 nm^2, 4×4 nm^2, and 5×5 nm^2, respectively. (c) I_{on}/I_{off} ratio under different gate lengths. (d) I-V characteristics of 3×3 nm^2 SiNWFETs under different scattering mechanisms; the gate length is 4 nm. CO, Coulomb; PH, phonon.

in Fig. 6(d) and Figs. S16–S18 [40], this reveals that scattering would reduce the device current, and more specially, phonon scattering has the most significant impact on device transport, followed by SR scattering and then Coulomb scattering. The impact of scattering is more pronounced in SiNWFETs with longer gate lengths and is even greater in those with smaller cross section. These results provide insights into the negative impacts of scattering mechanisms and offer guidance on how to mitigate these effects and optimize device performance.

Tunneling is a quantum mechanical phenomenon that often occurs in short-channel devices, resulting in a leakage current that is generally considered to be the end of Moore's law. Since scattering effects are more significant for the ON-state current, it is not considered in the calculation of the off-state tunneling current. Figure 7 depicts the band diagram of the OFF state for both long and short channels. The transmission spectrum is the result of the joint contribution of tunneling and thermionic emission: $T = T_{\text{tunneling}} + T_{\text{thermionic}}$. The energy levels above the barriers exhibit ballistic (thermal) transport, while the energy levels below the barriers exhibit tunneling transport. In the long-channel limit, the source-to-drain tunneling can be ignored

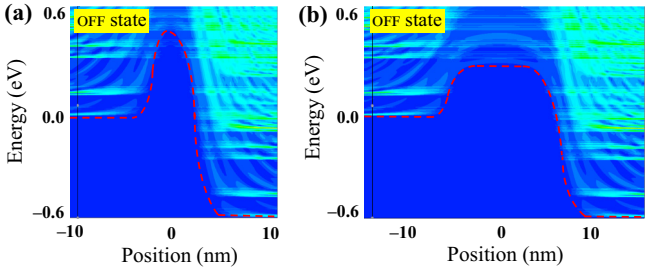

FIG. 7. Band diagram of the OFF state. Projected local density of states of SiNWFETs with (a) a short-channel limit and (b) a long-channel limit. Short-channel devices exhibit both thermionic emission and tunneling, but long-channel devices exhibit only thermionic emission. The channel length is 4 and 12 nm, respectively.

due to near-zero tunneling probability, whereas in the short-channel limit, the source-to-drain tunneling dominates. Figure 7 presents the projected local density of states for the OFF state of short-channel (4 nm) and long-channel (12 nm) devices, revealing a significant source-to-drain tunneling effect in the short-channel device.

To have better insight into quantum tunneling, we conducted simulations to assess the OFF-state tunneling current. To simplify the physical model, we did not consider the scattering mechanism because it would be crucial for the ON-state current (I_{on}) rather than the OFF-state current (I_{off}) limits. The tunneling ratio, denoted as $I_{\text{tunneling}}/I_{\text{total}}$, represents the proportion of the tunneling current with respect to the overall current. In Fig. 8(a), a short gate length ($L_g = 4$ nm)increases tunneling effectiveness, where tunneling predominates, significantly impacting the OFF-state current. The tunneling ratios are 99.3%, 99.86%, and 99.90% for SiNWFETs with cross sections of 3×3 nm^2, 4×4 nm^2, and 5×5 nm^2, respectively. The influence of tunneling weakens for devices with longer channel lengths. As the channel length decreases, the tunneling component in the total current increases, leading to an increase in SS as depicted in Figs. 6(a) and 6(b).

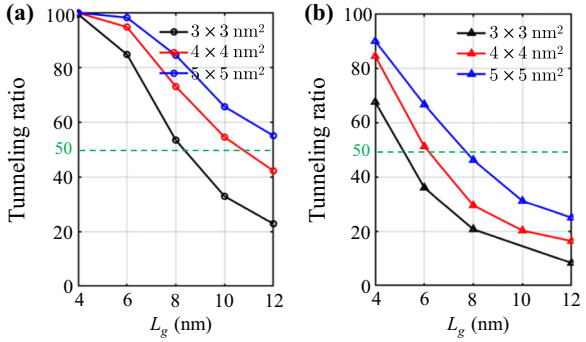

FIG. 8. Tunneling ratio of the OFF-state current I_{off} under different gate lengths of (a) *n*-type devices and (b) *p*-type devices.

Figure 8 provides a summary of the tunneling ratios for both *n*-type and *p*-type devices under various gate lengths. The results distinctly indicate that when the gate length is below 8 nm, the tunneling current starts to dominate the OFF state. Specifically, if we define 50% as the limit of the tunneling proportion of the current, if this value is exceeded, the tunneling current dominates the device, and it may face the risk of failure. Our results reveal that different cross sections and SiNWFET types exhibit differing gate-length scaling limits. For *n*-type devices [Fig. 8(a)], 3×3 nm^2 devices reach a limit at approximately 8 nm, while 4×4 nm^2 devices achieve this at 10 nm, which is consistent with findings in Ref. [26]. Moreover, *p*-type and *n*-type devices have different ultimate scaling limits, with *p*-type devices having a shorter scaling limit. For instance, the gate-length limit of 3×3 nm^2 devices is about 5 nm. In the upcoming technology node with a 4×4 nm^2 cross section, the gate-length scaling limit is approximately 10 nm on the basis of combined *n*-type and *p*-type device quantum simulation results. Consequently, source-to-drain tunneling emerges as a crucial consideration in SiNWFETs.

IV. CONCLUSION

In conclusion, we present a comprehensive study of quantum transport in GAA SiNWFETs at the scaling limit. The $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian parameters of the Si nanowires were calibrated from first-principles band structures with the use of the fast least-squares method or a machine learning approach. Using the NEGF formalism, we systematically analyzed the quantum transport characteristics under various gate lengths and provided gate scaling limits for different types of devices. We hope that the results will be helpful for the study and design of future logic devices.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant No.12174423 and No. 62331009, the Ministry of Science and Technology under Grant No. 2021YFA1200502, the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA0330401), and the Fundamental Research Funds for the Central Universities (Grant No. 23CX06063A).

The authors declare that they have no conflict of interest.

APPENDIX: REAL-SPACE k · p HAMILTONIAN OF Si NANOWIRES

Once one obtains the bulk $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian, the Hamiltonian of the Si nanowires can be derived by the finite difference method, and the cross section of the Si nanowires along the [100] orientation is split into an $N_y \times N_z$ grid mesh. The real-space discretized tight-binding Hamiltonian of Si nanowires is as follows:

$$
\begin{array}{c|cccccccccc}
 & (1,1) & (1,2) & (1,3) & \cdots & (1,N_z) & (2,1) & (2,2) & (2,3) & \cdots & \cdots \\
\hline
(1,1) & H_{i,j} & H_{i,j+1} & & & & H_{i+1,j} & H_{i+1,j+1} & & & \cdots \\
(1,2) & H^{\dagger}_{i,j+1} & H_{i,j} & H_{i,j+1} & & & H_{i+1,j-1} & H_{i+1,j} & H_{i+1,j+1} & & \cdots \\
(1,3) & & H^{\dagger}_{i,j+1} & H_{i,j} & H_{i,j+1} & & & H_{i+1,j-1} & H_{i+1,j} & H_{i+1,j+1} & \cdots \\
\vdots & & & \ddots & \ddots & \ddots & & & \ddots & \ddots & \ddots \\
(1,N_z) & & & & H^{\dagger}_{i,j+1} & H_{i,j} & & & & \cdots & \cdots \\
(2,1) & H^{\dagger}_{i+1,j} & H^{\dagger}_{i+1,j-1} & & & & H_{i,j} & H_{i,j+1} & & & \cdots \\
(2,2) & H^{\dagger}_{i+1,j+1} & H^{\dagger}_{i+1,j} & H^{\dagger}_{i+1,j-1} & & & H^{\dagger}_{i,j+1} & H_{i,j} & H_{i,j+1} & & \cdots \\
(2,3) & & H^{\dagger}_{i+1,j+1} & H^{\dagger}_{i+1,j} & H^{\dagger}_{i+1,j-1} & & & H^{\dagger}_{i,j+1} & H_{i,j} & H_{i,j+1} & \cdots \\
\vdots & & & \ddots & \ddots & & & & \ddots & \ddots & \ddots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}
, \tag{A1}
$$

where $H_{i,j}$ is the on-site energy of the (i,j) grid of the y-z plane and $H_{i\pm1,j\pm1}$ is the hopping energy between the (i,j) and $(i \pm 1, j \pm 1)$ grids. The real-space Hamiltonian size is $N_b N_y N_z \times N_b N_y N_z$, where N_b is the number of bands of the bulk $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian. Obviously, it is a large sparse block tridiagonal matrix. The matrix can be solved by the

Lanczos algorithm, or directly by MATLAB's built-in eigs function for solving sparse matrices.

[1] Victor Moroz, Lee Smith, Joanne Huang, Munkang Choi, Terry Ma, Jie Liu, Yunqiang Zhang, Xi-Wei Lin, Jamil Kawa, and Yves Saad, in *2014 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, USA, 2014), p. 7.4.1.

[2] Lee Smith, Munkang Choi, Martin Frey, Victor Moroz, Anne Ziegler, and Mathieu Luisier, in *2015 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)* (IEEE, Washington, DC, USA, 2015), p. 254.

[3] Katsuhiro Tomioka, Masatoshi Yoshimura, and Takashi Fukui, A III–V nanowire channel on silicon for high-performance vertical transistors, Nature **488**, 189 (2012).

[4] H. Mertens, *et al.*, in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, USA, 2016), p. 19.7.1.

[5] R. Ritzenthaler, *et al.*, in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, USA, 2018), p. 21.5.1.

[6] International Roadmap for Devices and Systems (IRDS) 2022 Edition (2022), https://irds.ieee.org/editions/2022.

[7] Seonghoon Jin, Sung-Min Hong, Woosung Choi, Keun-Ho Lee, and Youngkwan Park, in *2013 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)* (IEEE, Glasgow, UK, 2013), p. 348.

[8] Seonghoon Jin, Anh-Tuan Pham, Woosung Choi, Yutaka Nishizawa, Young-Tae Kim, Keun-Ho Lee, Youngkwan Park, and Eun Seung Jung, in *2014 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, USA, 2014), p. 7.5.1.

[9] Gerhard Klimeck, Roger Lake, R. Chris Bowen, William R. Frensley, and Ted S. Moise, Quantum device simulation with a generalized tunneling formula, Appl. Phys. Lett. **67**, 2539 (1995).

[10] Roger Lake, Gerhard Klimeck, R. Chris Bowen, and Dejan Jovanovic, Single and multiband modeling of quantum electron transport through layered semiconductor devices, J. Appl. Phys. **81**, 7845 (1997).

[11] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches, J. Appl. Phys. **92**, 3730 (2002).

[12] A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal, Two-dimensional quantum mechanical modeling of nanotransistors, J. Appl. Phys. **91**, 2343 (2002).

[13] Jing Wang, Eric Polizzi, and Mark Lundstrom, A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation, J. Appl. Phys. **96**, 2192 (2004).

[14] Jing Wang, Eric Polizzi, Avik Ghosh, Supriyo Datta, and Mark Lundstrom, Theoretical investigation of surface roughness scattering in silicon nanowire transistors, Appl. Phys. Lett. **87**, 043101 (2005).

[15] Mathieu Luisier, Andreas Schenk, and Wolfgang Fichtner, Quantum transport in two- and three-dimensional nanoscale transistors: Coupled mode effects in the nonequilibrium Green's function formalism, J. Appl. Phys. **100**, 043713 (2006).

[16] Shiqi Liu, Qiuhui Li, Chen Yang, Jie Yang, Lin Xu, Linqiang Xu, Jiachen Ma, Ying Li, Shibo Fang, Baochun Wu, Jichao Dong, Jinbo Yang, and Jing Lu, Performance limit of gate-all-around Si nanowire field-effect transistors: An ab initio quantum transport simulation, Phys. Rev. Appl. **18**, 054089 (2022).

[17] Mathieu Luisier, Andreas Schenk, Wolfgang Fichtner, and Gerhard Klimeck, Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations, Phys. Rev. B **74**, 205323 (2006).

[18] Mincheol Shin, Efficient simulation of silicon nanowire field effect transistors and their scaling behavior, J. Appl. Phys. **101**, 024510 (2007).

[19] Mincheol Shin, Quantum simulation of device characteristics of silicon nanowire FETs, IEEE Trans. Nanotechnol. **6**, 230 (2007).

[20] Mincheol Shin, Full-quantum simulation of hole transport and band-to-band tunneling in nanowires using the $k \cdot p$ method, J. Appl. Phys. **106**, 054505 (2009).

[21] Mincheol Shin, Sunhee Lee, and Gerhard Klimeck, Computational study on the performance of Si nanowire pMOSFETs based on the $k \cdot p$ method, IEEE Trans. Electron. Devices **57**, 2274 (2010).

[22] Shuo Zhang, Jun Z. Huang, Hao Xie, Afshan Khaliq, Dawei Wang, Wenchao Chen, Kai Miao, Hongsheng Chen, and Wen-Yan Yin, Design considerations for Si- and Ge-stacked nanosheet pMOSFETs based on quantum transport simulations, IEEE Trans. Electron. Devices **67**, 26 (2020).

[23] Shuo Zhang, Hao Xie, Jun Z. Huang, Wenchao Chen, Jie Liao, and Wen-Yan Yin, Rigorous modeling and investigation of low-field hole mobility in silicon and germanium gate-all-around nanosheet transistors, IEEE Trans. Electron. Devices **69**, 4777 (2022).

[24] Oves Badami, Cristina Medina-Bailon, Salim Berrada, Hamilton Carrillo-Nunez, Jaeyhun Lee, Vihar Georgiev, and Asen Asenov, Comprehensive study of cross-section dependent effective masses for silicon based gate-all-around transistors, Appl. Sci. **9**, 1895 (2019).

[25] Zifeng Wang, Shizhuo Ye, Hao Wang, Jin He, Qijun Huang, and Sheng Chang, Machine learning method for tight-binding Hamiltonian parameterization from ab-initio band structure, npj Comput. Mater. **7**, 11 (2021).

[26] Jing Wang and M. Lundstrom, in *International Electron Devices Meeting. Technical Digest (IEDM)* (IEEE, San Francisco, CA, USA, 2002), p. 707.

[27] Mathieu Luisier, Mark Lundstrom, Dimitri A. Antoniadis, and Jeffrey Bokor, in *2011 International Electron Devices Meeting (IEDM)* (IEEE, Washington, DC, USA, 2011), p. 11.2.1.

[28] Søren Smidstrup, *et al.*, QuantumATK: An integrated platform of electronic and atomic-scale modelling tools, J. Phys: Condens. Matter **32**, 015901 (2020).

[29] John P. Perdew, Kieron Burke, and Matthias Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. **77**, 3865 (1996).
[30] Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof, Hybrid functionals based on a screened Coulomb potential, J. Chem. Phys. **118**, 8207 (2003).
[31] J. Cerdá and F. Soria, Accurate and transferable extended Hückel-type tight-binding parameters, Phys. Rev. B **61**, 7965 (2000).
[32] ghzphy, https://github.com/ghzphy/nanowires_kp_parameters.
[33] Synopsys, Sentaurus Device QTX User Guide (2022).
[34] Supriyo Datta, *Quantum Transport: Atom to Transistor* (Cambridge University Press, Cambridge, England, 2005).
[35] Supriyo Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, Cambridge, England, 1997).
[36] P. Vogl, Harold P. Hjalmarson, and John D. Dow, A semi-empirical tight-binding theory of the electronic structure of semiconductors, J. Phys. Chem. Solids **44**, 365 (1983).
[37] Jean-Marc Jancu, Reinhard Scholz, Fabio Beltram, and Franco Bassani, Empirical spds* tight-binding calculation for cubic semiconductors: General method and material parameters, Phys. Rev. B **57**, 6493 (1998).
[38] Timothy B. Boykin, Gerhard Klimeck, and Fabiano Oyafuso, Valence band effective-mass expressions in the $sp^3d^5s^*$ empirical tight-binding model applied to a Si and Ge parametrization, Phys. Rev. B **69**, 115201 (2004).
[39] Timothy B. Boykin, Mathieu Luisier, Mehdi Salmani-Jelodar, and Gerhard Klimeck, Strain-induced, off-diagonal, same-atom parameters in empirical tight-binding theory suitable for [110] uniaxial strain applied to a silicon parametrization, Phys. Rev. B **81**, 125202 (2010).
[40] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevApplied.23.034049 for more details of the computational results obtained in the present work, including the complex band structure of Si nanowires calculated by quantumATK and the *I*-*V* characteristics of *n*-type SiNWFET devices under different scattering mechanisms.
[41] J. C. Hensel, H. Hasegawa, and M. Nakayama, Cyclotron resonance in uniaxially stressed silicon. II. Nature of the covalent bond, Phys. Rev. **138**, A225 (1965).
[42] G. Dresselhaus, A. F. Kip, and C. Kittel, Cyclotron resonance of electrons and holes in silicon and germanium crystals, Phys. Rev. **98**, 368 (1955).
[43] J. M. Luttinger and W. Kohn, Motion of electrons and holes in perturbed periodic fields, Phys. Rev. **97**, 869 (1955).
[44] E. O. Kane, Energy band structure in p-type germanium and silicon, J. Phys. Chem. Solids **1**, 82 (1956).
[45] P. Enders and M. Woerner, Eight-band k · p hamilton matrix for strained tetrahedral semiconductors: 4 × 4 block diagonalization for symmetric k-directions, Phys. Status Solidi (b) **194**, 585 (1996).
[46] Manuel Cardona and Fred H. Pollak, Energy-band structure of germanium and silicon: The $\mathbf{k}\cdot\mathbf{p}$ method, Phys. Rev. **142**, 530 (1966).
[47] P. Lawaetz, Valence-band parameters in cubic semiconductors, Phys. Rev. B **4**, 3460 (1971).
[48] D. Mamaluy, D. Vasileska, M. Sabathil, T. Zibold, and P. Vogl, Contact block reduction method for ballistic transport and carrier densities of open nanostructures, Phys. Rev. B **71**, 245321 (2005).
[49] Mincheol Shin, Quantum transport of holes in 1D, 2D, and 3D devices: the $k \cdot p$ method, J. Comput. Electron. **10**, 44 (2011).