ARTICLE   OPEN

Check for updates

Ab initio perspective of ultra-scaled CMOS from 2D-material fundamentals to dynamically doped transistors

Aryan Afzalian [ID][1][✉]

Using accurate dissipative DFT-NEGF atomistic-simulation techniques within the Wannier-Function formalism, we give a fresh look at the possibility of sub-10-nm scaling for high-performance complementary metal oxide semiconductor (CMOS) applications. We show that a combination of good electrostatic control together with high mobility is paramount to meet the stringent roadmap targets. Such requirements typically play against each other at sub-10-nm gate length for MOS transistors made of conventional semiconductor materials like Si, Ge, or III–V and dimensional scaling is expected to end ~12 nm gate-length (pitch of 40 nm). We demonstrate that using alternative 2D channel materials, such as the less-explored HfS_2 or ZrS_2, high-drive current down to ~6 nm is, however, achievable. We also propose a dynamically doped field-effect transistor concept, that scales better than its MOSFET counterpart. Used in combination with a high-mobility material such as HfS_2, it allows for keeping the stringent high-performance CMOS on current and competitive energy-delay performance, when scaling down to virtually 0 nm gate length using a single-gate architecture and an ultra-compact design (pitch of 22 nm). The dynamically doped field-effect transistor further addresses the grand-challenge of doping in ultra-scaled devices and 2D materials in particular.

INTRODUCTION

Scaling and Moore's law, that sets the footprint area of a transistor to scale by a factor 2, that is the transistor gate length L to scale by a factor $\sqrt{2}$, every 2 years, have been the driving force of the electronic industry[1]. Today L has been scaled well below 20 nm and further scaling has become increasingly difficult owing to short-channel effects (SCE) that degrade the subthreshold slope (SS) of a transistor (i.e., the efficiency with which the current is switched from off to on state by changing the gate bias). SCE leads to an increased off-state leakage current, I_{OFF}. To mitigate SCE, i.e., to keep good electrostatic control of the gate over the transistor channel, its thickness, t_s, has to be scaled as well[1,2]. Also, transistors have evolved from planar single-gate transistors to 3D multi-gate devices such as FINFETs[3], nanowires[4,5], and nanosheets[6]. As a rule of thumb, in a multi-gate device, the channel thickness t_s has to be of the order of ½ L in order to keep the electrostatic integrity leading to t_s of a few nm only in modern advanced nanoscale technologies[7]. At such value of t_s, conventional 3D semiconductors, like Si, or possible high-mobility channel-replacement materials like Ge[3] or III–V[5], suffer from quantum-confinement effects that strongly deteriorate their performance (e.g., current drive, gate coupling, mobilities…)[8–10], as well as, lead to increased variability (e.g., strong threshold-voltage variations with surface roughness for instance)[2,10,11]. It is commonly accepted that conventional dimensional scaling will stop for L of the order of 10 nm. The current international roadmap for device and system (IRDS) predictions have actually forecast that gate-length scaling will stop for a L of 12 nm (https://irds.ieee.org/editions/2018).

As an attempt to further push the scaling, transistors made of novel 2D materials[12], i.e., an atomically thin layer of material that does not create strong atomic bonds in the 3rd dimension, such as transition-metal dichalcogenide (TMD)[13–15] or black phosphorus (P_4)[13,16], are being actively investigated as future replacement of Si as channel material. These materials would offer the ultimate electrostatic control and are free from quantum confinement due to their 2D nature. In principle also, their thicknesses could be well controlled, which would remove the variability issue. The research development on 2D material is still at an early stage today. Despite the ever-growing list, including several thousands of newly discovered such materials[17], an ideal complementary metal oxide semiconductor (CMOS) candidate for the sub-10-nm channel has not yet emerged. Experimentally, only a subset of 2D-material transistors, such as those using MoS_2, WS_2, WSe_2, and P_4[13,14,16], have been explored and the current drive is typically too low for high-performance (HP)-CMOS applications (https://irds.ieee.org/editions/2018). Although the low drive-current is, at least in part, related to the immaturity of the technology, the fundamental physics and performance of these transistors are not yet fully elucidated. Even using 2D materials, scaling L below 10 nm is further complicated by an additional quantum-mechanical SCE. This effect, called source-to-drain tunneling (SDT), which is the ability of the electronic quantum-mechanical waves to evanescently leak through the channel barrier, further degrades SS and I_{OFF}.

Here, using our state-of-the-art density functional theory–non-equilibrium Green's function formalism (DFT-NEGF) ATOmistic modeling Solver (ATOMOS)[18], we offer unique insights on several 2D-material physics and performance, including those of the less-explored HfS_2 and ZrS_2 that feature appealing performance for ultra-scaled CMOS. We demonstrate the possibility of $L = 6$ nm high-performance devices, providing that high doping can be achieved. Finally, we demonstrate that further geometrical scaling, down to a virtually 0 nm gate length footprint, is possible using the dynamically doped field-effect transistor (D_2-FET). The D_2-FET concept further addresses the difficult challenge of doping in nanoscale devices (https://irds.ieee.org/editions/2018) and 2D materials in particular[19], and the need for chemical doping could be suppressed.

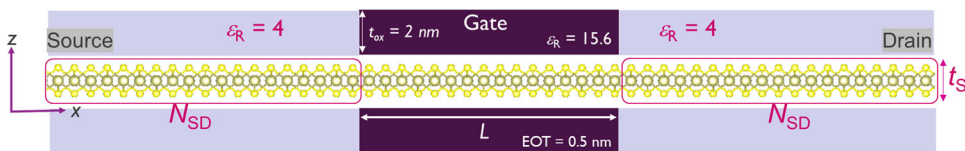[1]Imec, Kapeldreef 75, 3001 Leuven, Belgium. ✉email: Aryan.Afzalian@imec.be

Fig. 1 Schematic view of a double-gate MOSFET, where the channel is made of one of the monolayer 2D materials studied here. The doped contact, oxides (gate and spacer regions), as well as the main device parameters, are shown on the figure. The atomic structure that is depicted in this figure is that of a TMD, here it is HfS$_2$, with the metallic atom (Hf) in the center, sandwiched between the two chalcogen (S) atoms at the top and bottom.
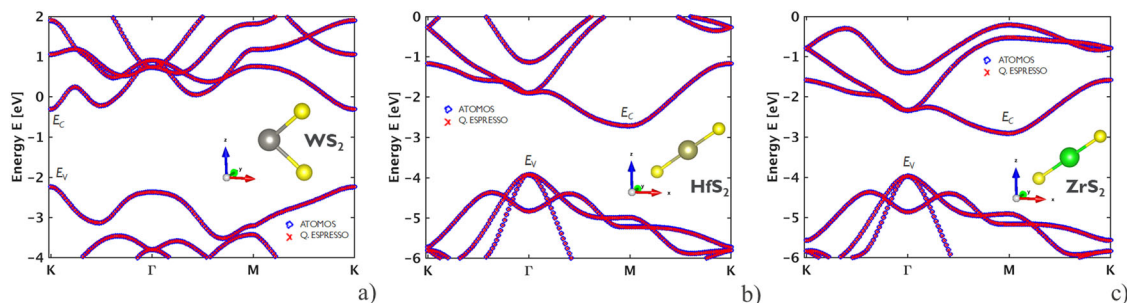

Fig. 2 Band structure computed with QUANTUM ESSPRESSO using plane-wave DFT and with ATOMOS using the Wannierized Hamiltonian. Band structure for **a** a monolayer WS$_2$ (2H-phase) **b** a monolayer HfS$_2$ (1T-phase) and **c** a monolayer ZrS$_2$ (1T-phase). The insets also show the atomic structure and chosen cartesian-axes directions for the various supercells.

RESULTS

Device structure and methodology

The schematic of the studied monolayer (1 ML) double-gate (DG) MOSFETs is shown in Fig. 1. An intrinsic channel of length L is used. The source- and drain- (S&D) extension regions are doped with a concentration N_{SD}. An abrupt junction profile is assumed. The 2 nm thick HfO$_2$ gate oxide has a relative permittivity $\varepsilon_R = 15.6$ and an equivalent oxide thickness (EOT) $= 0.5$ nm. The work function of the gate-voltage bias, V_G, is typically adjusted to achieve a fixed I_{OFF} value at $V_G = 0$ V. A low-K spacer oxide with $\varepsilon_R = 4$ surrounds the S&D extensions. Ohmic contacts are assumed with S&D bias $V_S = 0$ V and V_D, respectively.

The 1st step toward transport simulations of a given material is a first-principle geometry relaxation of its primitive unit cell, followed by an electronic structure calculation. We used the DFT package Quantum ESPRESSO[20] and the generalized gradient approximation with the optB86b exchange-correlation functional[21]. The Bloch wavefunctions are then transformed into maximally-localized Wannier functions (MLWF) typically centered on the ions using the wannier90 package[22]. Figure 2 demonstrates the validity of our MLWF representation for the case of 3 of the 2D materials studied here. The resulting supercell information, including atoms and MLWF positions, lattice vectors, as well as the localized Hamiltonian-matrix elements, are used by ATOMOS as building blocks to create the full-device atomic structure and Hamiltonian-matrix. Transport calculations are then performed using a real-space NEGF[23,24] formalism including electron–phonon (e-ph) scattering within the self-consistent Born approximation[25]. More details can be found in the method section.

Transport model requirement for 2D-material screening

We first focus on evaluating intrinsic monolayer (1 ML) 2D-material device physics and performance using our DFT-NEGF model. The goal is to find a meaningful upper limit to identify the trends and screen the most promising candidates for scaled CMOS applications. Looking at the list of existing TMDs and other 2D materials, we have pre-selected five TMDs[13,15,17], as well as P$_4$[16,17], owing to their relevant electronic and transport properties (band structure,

phonon properties, material stability) and/or experimental relevance. For the TMDs, we focus here on MoS$_2$, WS$_2$, and WSe$_2$ in the trigonal prismatic (2H) phase, as these materials are among the most studied and mature experimentally. We also focus on the less-explored HfS$_2$ and ZrS$_2$ TMD's (in their most stable octahedral, 1T, phase), as their band structures hint for better transport properties (higher drive current), whereas retaining a sufficiently high bandgap and balanced properties[15,17] to expect good off-state currents at scaled gate lengths. Finally, current-voltage characteristics for P$_4$ will be investigated here as well, owing to the strong attention and expectation this material has stirred in the recent literature[16].

Non-atomistic models, such as effective-masses (EM) and derived simplified two-bands NEGF models, although widely used[26–28] owing to their wider availability and strongly reduced computational cost, are typically inaccurate to model 2D materials. In Fig. 3, it can be seen that although a reasonable fitting of the lowest part of the band structure can be obtained for WS$_2$ using an effective-mass Hamiltonian model (Fig. 3b), the current is strongly over-estimated (Fig. 3a). This is owing to the combination of two facts. First, owing to their specific band structures, e.g., many TMD's (in particular in the conduction band of that made of W or Mo atoms) have narrow energy valleys with discontinuous density-of-state, DoS, profiles that are not captured with simplified band models (e.g., see Fig. 3b, where the 1st WS$_2$ conduction-band valley is a narrow valley with an energy extend that is <0.6 eV in the KΓ direction)[29]. Second, in an extremely thin material, a full atomistic treatment of how the charge is distributed within the 2D layers is required to accurately capture the charge-centroid position. For the case shown in Fig. 3a, as typically the case in TMDs, ~90% of the charge is located on the metallic (W) atom, which is in the middle, not on the surface chalcogenide (S) atoms (Fig. 1). This information is lost in a non-atomistic model, and a homogeneous charge distribution with a centroid closer to the surface is obtained. The effective-mass model is also not able to predict accurately SDT, hence the subthreshold characteristics of the device (Fig. 3a), a crucial effect for the sub-10 nm gate-length regime where 2D materials are envisioned to be used for CMOS. Finally, using these approximate EM and derived 2-bands NEGF models, it was assessed that a
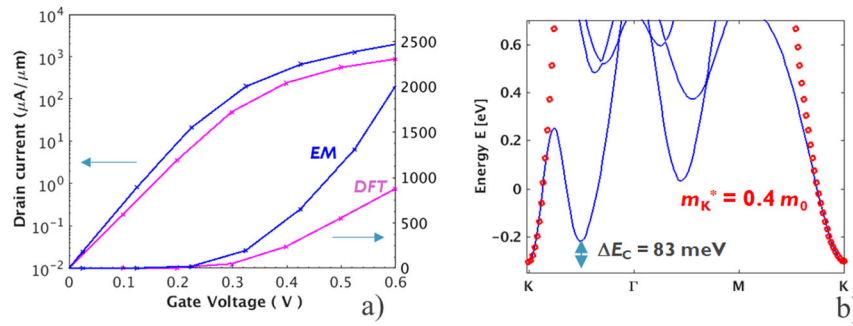
Fig. 3 DFT vs fitted effective-mass (EM) model. a Drain-current–gate-voltage, $I_D(V_G)$, characteristics of a $L = 5$ nm 1ML-WS$_2$-DG nMOSFET computed with DFT and a fitted EM model (including the two first valleys seen in Fig. 3b). $V_D = 0.6$ V. **b** DFT-computed band structure around E_C (blue line) and that fitted with the 1st valley (centered at K-point) isotropic effective mass $= 0.4\,m_0$ (red stars), m_0 being the free electron mass. Despite that a good agreement between the EM- and the DFT-band structure model is achieved in the vicinity of E_C, the EM-NEGF model strongly overestimates the current drive in the device. The second valley (located between K and Γ) was also fitted with an EM of $0.9\,m_0$ and included in the effective-mass NEGF model.
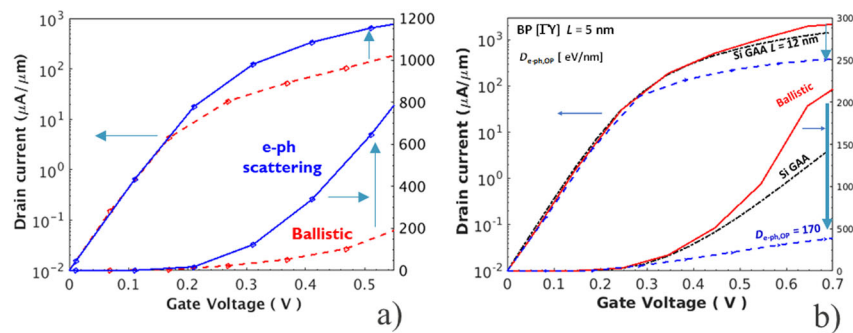

Fig. 4 Impact of electron–phonon scattering in 2D-material transistors. a $I_D(V_G)$ characteristics of a $L = 14$ nm bi-layer WS$_2$-DG nMOSFET computed with DFT-NEGF with and without e-ph. $V_D = 0.7$ V. As q.V_D is larger than the width of the bi-layer first valley, the 1st valley electrons cannot travel ballistically from the channel to the drain and the ballistic current is reduced compared with the case with scattering. **b** $I_D(V_G)$ characteristics of a $L = 5$ nm, 1ML-P$_4$-DG nMOSFET computed with DFT-NEGF with and without e-ph. $V_D = 0.6$ V. Owing to the strong optical-phonon coupling ($D_{OP} = 170$ meV/nm, $\hbar\omega_0 = 32$ meV)[43], the drive current drops significantly when e-ph is included, despite the very-short-channel length. $I_{OFF} = 10$ nA/μm. ΓY (zigzag) channel orientation.

bi-layer DG device could deliver more drive-current than that of a 1 ML device for the sub-10 nm HP -CMOS application[27,28]. Our DFT-NEGF results, however, show that a 1 ML material, which is the main focus of this paper, is preferred (more details can be found in Supplementary Note 1). For accurate results, full-band atomistic-transport simulations, such as the DFT-NEGF results presented here, are therefore needed.

It, however, turns out that for 2D materials, ballistic full-band transport simulations are not accurate enough, even at a gate length as short as 5 nm. The argument often used of very-short L to justify ballistic transport in conventional 3D-material transistors does not hold true here. In 2D-materials with narrow valleys, a valley-filtering effects typically strongly reduce the ballistic current and inelastic-scattering effects need to be included to recover the current in the device[29] as shown in Fig. 4a for a bi-layer WS$_2$ transistor. In other cases, like for black phosphorus, strong optical-phonon modes can significantly reduce the current compared with the ballistic case (Fig. 4b). Thus, a full-band dissipative atomistic treatment, as presented here is needed to get an accurate and meaningful upper limit of 2D-material devices. This upper limit may still be far from today's reality, as we are neglecting interactions with the environment (e.g., contact resistance, surrounding oxides…), and defects that are usually strongly present in nowadays experimental devices[13,16,19]. It, however, gives an insight of what is to be the fundamental potential of such a technology, as it matures[19,30].

DFT-computed material parameters and properties

Supplementary Table 1 (in Supplementary Note 2) gives the relaxed unit-cell dimensions and bandgaps we obtained for the TMDs studied here. They are in good agreement with other DFT calculations in the literature[15,17] and experimental results (http://www.hqgraphene.com/All-Semiconductors.php). From our NEGF simulations, we have also extracted the electron or hole concentrations vs the Fermi level, E_F, position with regards to conduction- or valence-band edges, E_C or E_V, respectively, i.e., $E_F - E_C$ or $E_V - E_F$. By fitting those to an analytical 2D-DoS model, the conduction- or valence-band DoS, N_{2D}, as well as an equivalent DoS mass, m_{DoS}, can be computed. Both values are reported in Supplementary Table 1 for the TMDs studied here. This m_{DoS} folds the DFT-computed non-parabolicity of the occupied bands close to the conduction- or valence-band edges into a simplified, equivalent, single-band, parabolic effective-mass model (the details are in the method section). Supplementary Fig. 4 shows, for a representative sample, the good level of agreement that can be achieved between the analytical charge model and the DFT-NEGF-simulated data. N_{2D}, or equivalently m_{DoS}, as well as the mobilities, that we will extract next, are useful quantities for developing simplified TCAD or compact models and benchmarking 2D-material performance[31]. Note that, in this paper, densities, as well as doping concentrations, are given per unit of volume. Those can be converted to the per unit of area, often used for 2D materials, by multiplying by the 2D-film thickness, t_S, ~0.6 nm for a
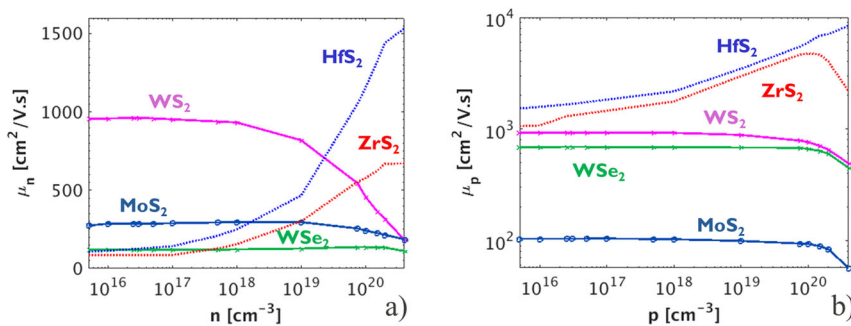
Fig. 5 Electron–phonon-limited mobility. Intrinsic electron–phonon-limited mobility vs carrier concentration in **a** n- and **b** p-type TMD transistors. The mobilities were extracted from our DFT-NEGF simulations using long-channel devices (for several channel L ranging from 100 nm to 1 μm typically) at $V_D = 50$ mV. The L-independent ballistic resistance was removed from the extraction by using the dR/dL method[32]. HfS$_2$ and ZrS$_2$ results are shown along the ΓK channel orientation and include polar-optical phonons using a Fröhlich model.

monolayer TMD (the exact value used for each studied monolayer can be found in Supplementary Table 1).

Figure 5 shows the long-channel low-field intrinsic electron and hole mobilities. Each mobility curve was extracted from 4 to 7 DFT-NEGF simulations using devices with various channel lengths, typically ranging from 100 nm to 1 μm, using the dR/dL method[32] to correct for the ballistic resistance. The intrinsic mobility is a convoluted value, resulting from band structure (intrinsic transport properties) and electron–phonon (e-ph) scattering. It is a key-performance metrics for long-channel devices. For electron-phonons, we used the DFT-computed values of ref. [15] and included the dominant acoustic, and optical modes. For the 1T TMDs that are polar materials, we also include the polar-optical phonons (POP) using a Fröhlich model[33,34] that considers the electronic screening and the long-range interaction up to 3 nm (more details can be found in Supplementary Note 3). We note that the long-range LO polar component can be directly computed using DFT calculations, as it was done in ref. [35] and as can be seen in Supplementary Fig. 14b. However, a Fröhlich model provides a direct way to consider the screening, which is important and was neglected in ref. [35]. 1T TMD material mobilities could also suffer, in principle, from a non-vanishing ZA phonon term owing to the lack of horizontal mirror symmetry[36]. In order to verify this, we have computed the HfS$_2$ electron–phonon matrix elements from DFT, using a similar approach that the one used in ref. [35]. We have found that ZA phonons do contribute in HfS$_2$, but that their contribution is small compared with that of the other acoustic phonon modes (TA and LA, see Supplementary Fig. 14a). A similar conclusion was found for both the mobilities of HfS$_2$ and ZrS$_2$ in ref. [35].

For the 2H TMDs, the mobility curves typically present a plateau region dominated by intra-band low-energy acoustic e-ph-scattering. At higher carrier concentrations, when the position of the energy band with regards to E_F is sufficiently degenerated so that satellite energy valleys start to be populated, higher-energy intervalley e-ph scattering mechanisms further degrade the mobilities. The carrier concentration at which this degradation happens depends on the energy separation between the 1st and the satellite valleys and the m_{DoS} (a larger m_{DoS} leads to less degeneracy at a given concentration as illustrated in Supplementary Fig. 4). The NEGF-computed mobility values for the 2H TMDs are in qualitative agreement with mobilities calculated in the literature with various methods[35,37,38], showing the same order of magnitude and ordering. WS$_2$ has the highest mobility. For p devices, WSe$_2$ also features an interesting value.

For the 1T TMDs, the plateau region is not observed in the mobility curves. Their mobility rather increases for increasing carrier concentration. Supplementary Fig. 7 compares the n- and p-type HfS$_2$ total mobilities including POP and screening to that without POP and that with POP but neglecting the screening.

The total mobility value is limited by the strong, and nearly unscreened, POP interaction at low carrier concentration. As the carrier concentration increases, however, screening renders POP scattering less efficient and the mobility increases toward the limit without POP. It is to be noted that the n-type mobility value of 1896 cm^2/V s that we obtain in the plateau region for the case without POP well agrees with the ~1800 cm^2/V s acoustic phonon-limited value computed in refs [35,37]. The n-type low mobility value of ~60 cm^2/V s obtained for the case that included POP, but neglecting the screening, is consistent with the results and hypothesis presented in ref. [35].

Finally, as presented in Supplementary Fig. 8 for HfS$_2$, for scaled nanoscale devices, the impact of POP scattering is not significant, owing to the short-channel lengths and the strong screening related to the high carrier concentration in on-state. The impact of high-energy optical phonons is typically rather limited in the subthreshold regime of a scaled transistor. Despite the weak screening in the channel, related to the low subthreshold-carrier concentration, most of the electrons are injected at an energy in thermal equilibrium with the top-of-the-channel barrier (e.g., see Fig. 6). The majority of empty states, in which electrons could scatter to, are, however, localized at the same energy. Hence, low-energy acoustic phonons are rather the dominant scattering mechanism in the subthreshold regime. In on-regime, the situation is different (e.g., see Fig. 7a) but POP is effectively screened. From the above discussion and results, one concludes that for nanoscale devices with strong polar interactions, the high-density screened mobility is likely to be the relevant one. The rather high mobilities obtained for HfS$_2$ and ZrS$_2$, at high carrier concentration (in on-state carrier densities of several 10^{20} cm^{-3} are typical), highlights their interesting transport properties.

By definition, the current is the product of the number of mobile charge carriers, that is proportional to m_{DoS}, times their velocity, that is proportional to the mobility. Hence the $m_{DoS} \times$ mobility product of a given material is an indication of its MOSFET drive-current potential. This product, normalized so that it is equal to 1 for the nWS$_2$-case, is reported in Supplementary Table 1 for n- and p-type conduction and allows a relative comparison between the different TMDs reported here. Again, the drive potential of HfS$_2$ and ZrS$_2$ stands out, while WS$_2$ comes in 3rd position.

Sub-10 nm fundamentals

2D materials are, however, envisioned to be used in the sub-10-nm gate-length regime as potential replacement for Si. At such L, the mobility $\times m_{DoS}$ product alone is not a sufficient metrics to compare performance. Other effects such as SDT, that become important owing to the narrow channel barrier, typically penalize more high-mobility materials (as the ability of quantum-mechanical tunneling is enhanced in low-effective mass materials)
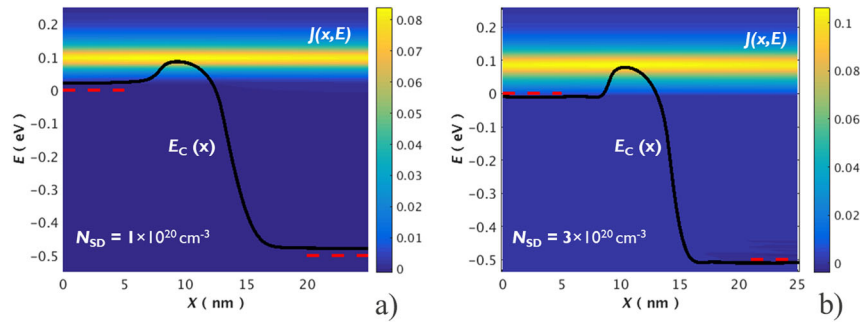
Fig. 6 Impact of source and drain doping on internal properties of a 2D MOSFET in off-state. Current spectrum $J(E, x)$ (surface plot), as well as top conduction-band (E_C) (-) edge along the channel direction, x, of a HfS$_2$ nMOS with $L = 5$ nm, in off-state. **a** For $N_{SD} = 1 \times 10^{20}$ cm^{-3}. In this case, the lower N_{SD} value is not sufficient to ensure degeneracy of the extension as can be seen at the source and drain-sides, where the conduction band is above the Fermi levels, E_{FS} and E_{FD}, respectively. Both Fermi levels are indicated by a red dashed line. **b** For $N_{SD} = 3 \times 10^{20}$ cm^{-3}. The resulting narrower channel barrier allows for a larger part of the current spectrum to tunnel under the channel barrier (SDT). This effect is enhanced for larger N_{SD} values as the effective-channel length is reduced. $V_D = 0.5$ V.
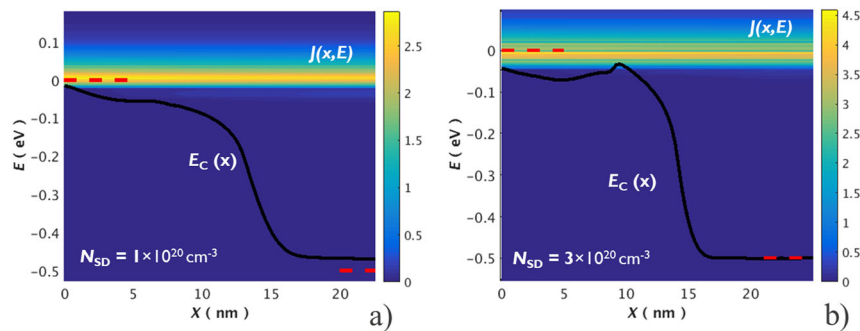

Fig. 7 Impact of source and drain doping on internal properties of a 2D MOSFET in on-state. Current spectrum $J(E, x)$ (surface plot), as well as top conduction-band (E_C) (-) edge along the channel direction, x, of the $L = 5$ nm HfS$_2$ nMOS in on-state. **a** For $N_{SD} = 1 \times 10^{20}$ cm^{-3}. In this case, due to the lower N_{SD} value, the current flow is limited by the source conduction band and source starvation is observed in the $I_D(V_G)$ characteristics. **b** For $N_{SD} = 3 \times 10^{20}$ cm^{-3}. In this case, the current is still limited by the channel barrier that is well controlled by the gate. Current saturation is not yet observed in the characteristics at this gate voltage. $V_D = 0.5$ V. $V_G = 0.6$ V.
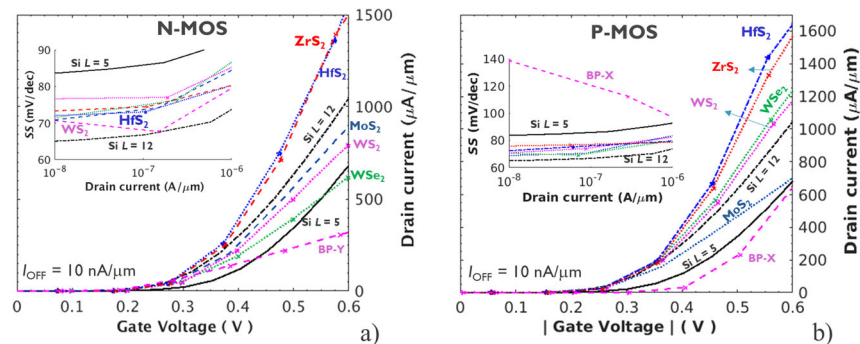

Fig. 8 $I_D(V_G)$ characteristics and $SS(V_G)$ (inset) of the optimized $L = 5$ nm 2D materials DG-MOSFETs, and the $L = 5$ and 12 nm Si optimized GAA. a n- and **b** pMOS transistors. $|V_D| = 0.6$ V. $I_{OFF} = 10$ nA/μm. e-ph scattering is included. The devices are optimized in terms of source- and drain-doping concentration (N_{SD}), channel orientation (if anisotropic) and thickness (t_S) for the GAA. HfS$_2$ and ZrS$_2$ IV's are shown along the ΓK channel orientation. The current is normalized by the gate perimeter.

and a trade-off exists. The case of P$_4$ that we will further discuss below is a good example. Similarly, a recent publication has used DFT-NEGF simulations to screen 100 2D materials and found 13 potential candidates with very high-drive current potential at $L = 15$ nm (pending a detailed study on the impact of e-ph scattering)[39]. When scaling L down to 5 nm, however, the SS of these devices were all degraded and ranging in the 110–275 mV/decade using a DG architecture. It is to be noted that the dynamically doped-transistor concept, that will be studied in the

last part of this manuscript, might be a way to utilize the strong drive potential of such materials at further scaled dimensions.

Figure 8 compares the $I_D(V_G)$ characteristics of $L = 5$ nm DG n- and p-MOSFETs made of six different 2D materials (Fig. 1), the five TMDs previously studied as well as P$_4$, at a typical HP off-state leakage of $I_{OFF} = 10$ nA/μm (https://irds.ieee.org/editions/2018). We further benchmark their performance against that of an optimized Si n-type gate-all-around nanowire (GAA) with a square cross-section, but with a relaxed gate length ($L = 12$ nm)[40].
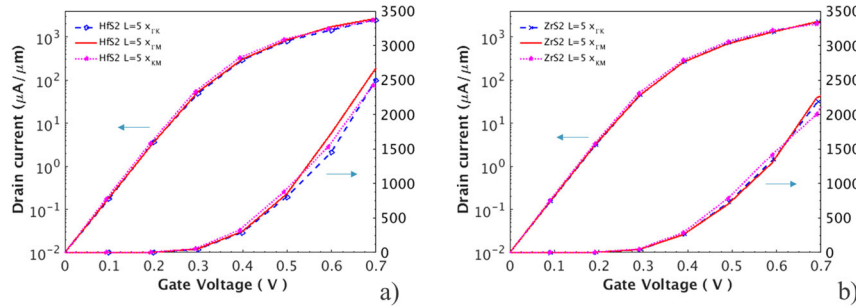
Fig. 9 $I_D(V_G)$ **characteristics of group-IV TMD along principal crystallographic axes.** $I_D(V_G)$ characteristics of group-IV TMD nMOSFETs for three different channel orientations along principal crystallographic axes (ΓK, ΓM, and KM), **a** for HfS$_2$ and **b** for ZrS$_2$. V_D = 0.6 V. L = 5 nm. I_{OFF} = 10 nA/μm. e-ph scattering is included.

The optimized L = 5 nm Si-GAA $I_D(V_G)$ is shown as well. The GAA were simulated using cleaned mode-space sp$_3$d$_5$s* tight-binding NEGF models[34,40].

For Si, scaling L below 10 nm typically results in SS and I_{ON} degradation. This is due to electrostatic control losses, quantum confinement, and SDT. SDT and QC become significant for L < 10 nm and t_S < 4 nm, respectively. It was observed that e-ph scattering was significantly enhanced, even at short L, for Si-GAA with t_S < 4 nm[8,10,25], one reason being the increase of the electron–phonon wave-function overlap in strongly confined wires due to volume inversion[10]. This is one of the fundamental reasons behind the strong mobility reduction observed in thin-film 3D materials. For the 5 nm long Si device, the narrow t_S of 3.5 nm used in the simulations will further result in a strong threshold-voltage variability related to surface roughness induced bandgap change with QC[2,10,11].

For all the 2D materials shown on Fig. 8, excepted for the p-type P$_4$-device case that will be discussed below, we observed less I_{ON} and SS degradation than for Si, when scaling L down to 5 nm. This is related to their excellent electrostatic control (a better electrostatic control enables a larger effective-channel length at same nominal L, hence less SDT) and QC-free characteristics stemming from their 2D nature.

The outstanding performance of the two group-IV TMDs, i.e., of HfS$_2$ and ZrS$_2$, that feature, by far, the highest on-current levels is also highlighted on the plot. Besides the afore-mentioned excellent electrostatic control and confinement-free characteristics, this is related to HfS$_2$ and ZrS$_2$ well-balanced transport properties that allow for high I_{ON} with limited SDT. The closely matched characteristics of both HfS$_2$ and ZrS$_2$ can be understood in light of their similar band structure (Fig. 2b, c) and the mitigation of the mobility × m_{DoS} product (that is better for HfS$_2$) at very-short L. It is to be noted that HfS$_2$ and ZrS$_2$ have anisotropic band structure properties. Figure 9 shows the impact of crystal orientation on the performance of the HfS$_2$ and ZrS$_2$ n-type devices. Their performance is not varying much along the principal-axis directions shown here. Overall, group-IV TMDs show promise for scaled HP CMOS.

From the more studied group-VI TMD family, WS$_2$ emerges as the best candidate for n and p on average, i.e., second best and close to MoS$_2$ for n, and best with WSe$_2$ for p. MoS$_2$ performs poorly for p-type, whereas WSe$_2$ performs poorly for n-type conduction. This can be correlated to the intrinsic transport properties (e.g., see mobility × m_{DoS} in Supplementary Table 1) of these materials for p-type. For n-type, an additional factor has to be considered. Excepted for MoS$_2$ that has very poor p-type performance, group-VI TMDs have a markedly stronger p-type drive-current than that of the n-type. This is related, at least to a great extent, to the narrow valleys that are present in the conduction band of the Mo- or W-based TMDs, as discussed above. These prevent, at least partially, direct ballistic current from

the channel to the drain at V_{DS} ≥ 0.6 V, so that a less-efficient phonon-assisted transport is required. 1ML-WS$_2$ or MoS$_2$ feature a relatively wide 1st conduction-band valley (for WS$_2$, for instance, its width is actually not isotropic and can be especially large in certain orientation such as the KM orientation shown on Figs. 2a and 3b). The combination of this fact with its higher electron mobility × m_{DoS} product explains WS$_2$ good position for n-type transport in this group.

The P$_4$ device shows the worst performance, which might come as a surprise. P4 is a strongly anisotropic material with a low-effective mass in the ΓX direction (i.e., high mobility) and a high effective mass in the ΓY direction (i.e., low mobility). This is true both for n- and p-type conduction (more information can be found in Supplementary Note 4). Hence, this material has a very strong drive current for longer channel in the ballistic regime and the ΓX transport direction. In the sub-10 nm regime, however, the armchair (ΓX) oriented P$_4$ transistor, strongly suffers from SDT and is difficult to switch-off due to its very-low transport effective mass. For L < 10 nm, the strongly anisotropic P$_4$ material was shown to perform best in the zigzag (ΓY) orientation for the n-type transistor[18]. Still, the ΓY P$_4$ n-device shows a strong drive current in the ballistic regime (Fig. 4b). Most theoretical studies on scaled P$_4$ devices have looked either at ballistic performance and simplified band models[41], or have neglected the optical-phonon coupling[42]. The ΓY P$_4$ nMOSFET I_{ON} is strongly degraded by its optical-phonon (OP) coupling ($D_{e-ph,OP}$ = 170 eV/nm for a single monolayer)[18,43] (Fig. 4b). Concerning the pMOS, the ΓX transistor still performs the best at L = 5 nm but its drive-current is severely degraded by SDT (Fig. 8b). The ΓY drive-current is indeed even lower than in the n-case, while the SDT-related sub-10 nm SS degradation in the ΓX direction is not as strong as for the n-case (more details are available in the Supplementary Note 4).

For all the 2D materials studied here, we found that a high-doping density in the source and drain extensions, N_{SD}, (N_{SD} = 2 to 4 × 10^{20} cm^{-3}) had to be used to reach their fundamental current level. Figure 10 shows the impact of N_{SD} on the current for the L = 5 nm HfS$_2$ and WS$_2$ nMOS devices. Figures 6 and 7 give details on the current spectrum flow and the conduction-band position in the HfS$_2$ transistor for N_{SD} = 1 and 3 × 10^{20} cm^{-3} in off- and on-state, respectively. Owing to the typically high density of states in these materials (~30× (10×) that of Si for HfS$_2$ (WS$_2$, respectively)) (see Supplementary Table 1), a high N_{SD} value is required to fully degenerate S&D extensions (Fig. 6) and avoid a saturation in the $I_D(V_G)$ characteristics in on-regime related to a source-starvation effect[44]. In the case of source starvation, as the gate voltage is increased in on-regime, the current is limited by the availability of carrier in the source. On Fig. 7a, it can be seen that, when the source-starvation regime is reached, the current is not limited by the channel barrier. It is rather limited by the energy band at source side. The latter is only indirectly and weakly affected, when switching on V_G, by the increase of the non-equilibrium transport
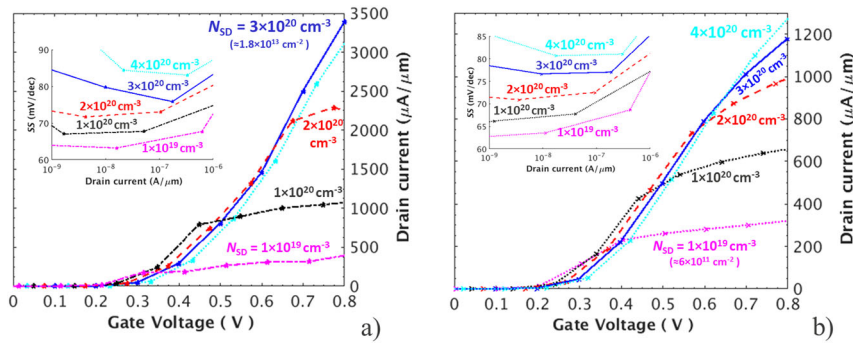
Fig. 10 Impact of source and drain doping on transfer characteristics of 2D MOSFETs. $I_D(V_G)$ characteristics and $SS(V_G)$ (inset) of 2D nMOSFETs vs N_{SD} **a** for HfS$_2$ and **b** for WS$_2$. $V_D = 0.6$ V. $L = 5$ nm. $I_{OFF} = 10$ nA/μm. e-ph scattering is included. The trade-off between on-state (better for higher N_{SD} values, owing to a reduction of source starvation) vs off-state (better for lower N_{SD} values, owing to a reduction of SDT) is observed for both materials.
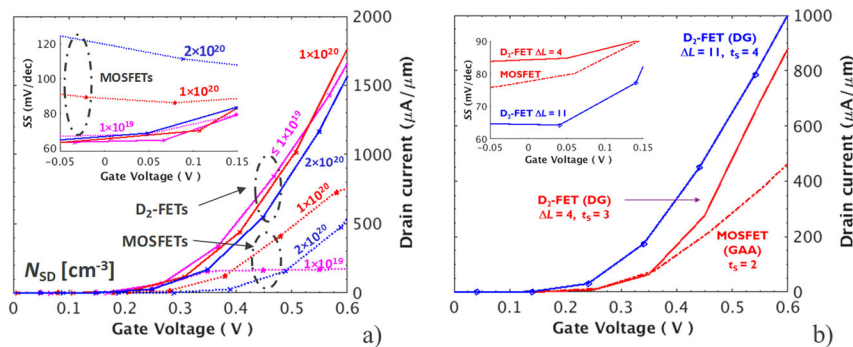

Fig. 11 $L = 3$ nm D$_2$-FETs vs MOSFET transfer characteristics. $I_D(V_G)$ characteristics and $SS(V_G)$ (inset) for **a** HfS$_2$ double gated nMOSFETs and nD$_2$-FETs with $\Delta L = 4$ nm vs N_{SD}, and **b** optimized Si-GAA nMOSFETs ($N_{SD} = 1 \times 10^{20}$ cm^{-3}) and DG nD$_2$-FETs (intrinsic N_{SD}) for $\Delta L = 4$ and 11 nm. For the D$_2$-FETs the optimized film thickness t_S (also indicated in the figure in nm) is larger than for the MOSFET. $V_D = 0.6$ V. $L = 3$ nm. $I_{OFF} = 10$ nA/μm.

charge through the device. This leads to a weak increase and eventually a saturation of the current in the $I_D(V_G)$ characteristics. By increasing the source- and drain-extension doping to ensure good degeneracy at the source side, however, this effect is delayed to higher gate-overdrive values (Figs. 7b and 10). As can also be seen on Fig. 10, increasing N_{SD} has a detrimental effect on SS at such a scaled gate length, so that an optimal value exists. This is related to a reduction of the effective-channel length and an increase of SDT for higher N_{SD} values (Fig. 6). A similar trend is observed for all the n- and p-type devices studied here and an optimal value between $N_{SD} = 2$–4×10^{20} cm^{-3} is observed for $L = 5$ nm and $V_{DD} = 0.6$ V in all cases.

The D$_2$-FET
The trade-off between on- and off-state for the optimal doping concentration becomes more stringent as L is reduced, ultimately degrading transistor performance and preventing further downscaling. Even using 2D materials, scaling below 5-nm gate length becomes very challenging. The case of the monolayer HfS$_2$ transistor with $L = 3$ nm is shown on Fig. 11a. Using $N_{SD} = 2 \times 10^{20}$ cm^{-3} results in strongly degraded SS due to SCEs and SDT. The optimal $N_{SD} = 1 \times 10^{20}$ cm^{-3} value, however, has poor performance. It suffers both from source-starvation-related on-current saturation and degraded slope due to SDT and SCE.

In addition, as transistor dimensions are scaled down, it becomes increasingly difficult to dope, activate, and control the location of the source-and-drain-extension doping atoms, using traditional implantation and annealing techniques[4] (https://irds. ieee.org/editions/2018)[19,45,46]. In thin-film technologies, direct

implantation usually results in a high defect concentration. It is especially the case for 2D materials, where finding a proper way for doping is still an active topic of research (https://irds.ieee.org/editions/2018)[19,47]. Typically, more complex techniques, such as epitaxial regrowth of the source and drain extensions, using in situ doping during the epitaxy, are needed (https://irds.ieee.org/editions/2018)[19,48]. Strict control and positioning of the high-doping concentration to prevent its unwanted diffusion during the high-temperature steps of the fabrication process (e.g., during dopants activation or epitaxial-growth phases) in the channel is also challenging and requires advanced techniques such as Flash and Laser anneal[4] (https://irds.ieee.org/editions/2018)[45], a solution to this particular problem is to use a uniformly doped, or junctionless transistor[4]. In any case, the discrete nature and limited numbers of doping atoms, resulting from the scaling of the device dimensions, leads to a strong and unavoidable statistical variability when using doping impurities at very small dimensions[46].

Owing to the challenge of chemical doping, electrostatic doping is sometime used in today's experimental 2D or carbon-nanotube material devices to dope their source and drain extensions[49]. It consists in using a gate, e.g., the wafer back gate, to electrostatically induce a high carrier concentration and decrease the semiconductor resistivity, which is the desired effect of chemical doping. Using electrostatic doping with a gate is indeed free of all the afore-mentioned problems related to chemical doping. It can dope (i.e., control the carrier concentration in) the entire thickness of the semiconductor film as long as this film is sufficiently thin, typically for $t_S < 10$ nm (the exact value also depends on the residual chemical doping of the film and
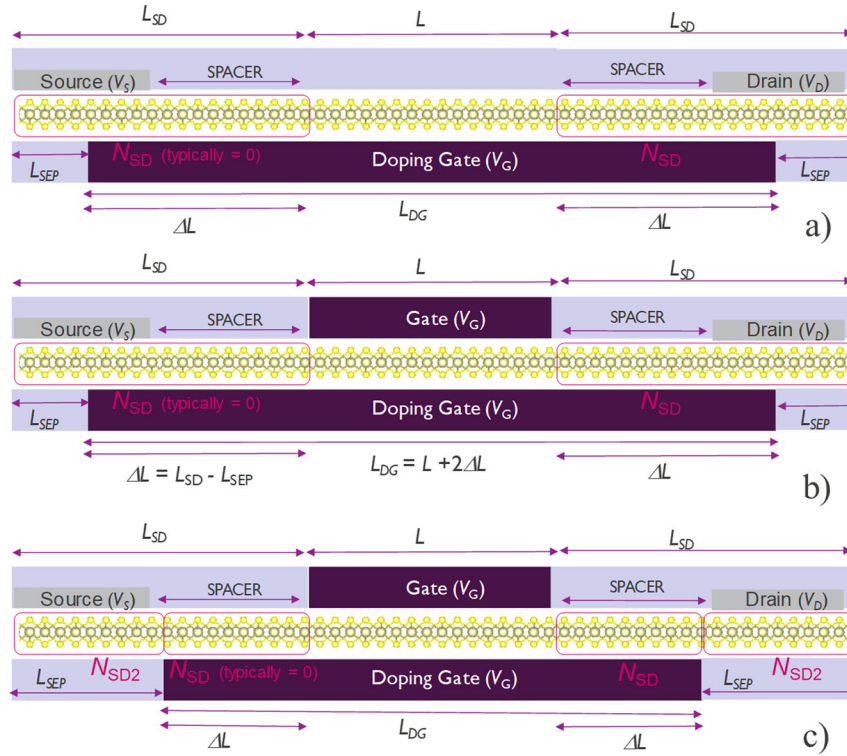
Fig. 12 Schematic view of 1 ML 2D D$_2$-FETs. D$_2$-FET with **a** a single gate, **b** a double gate, and **c** an alternative double-gate design with chemical doping in the source-and-drain-contact regions (N_{SD2}), larger L_{SEP}, and shorter ΔL.

substantially decrease if this residual doping is larger than 1×10^{20} cm^{-3}). Directly, using the wafer back gate is, however, not a manufacturable solution to individually control billions of transistors with different n- and p-type doping on a chip. A local, dedicated gate for each transistor would rather be required for that purpose. Typically, also these techniques are meant to induce a fixed amount of doping in the source and drain extensions of the device, whereas the dynamic-doping implications of this technique have not yet been studied. It would indeed be advantageous to have no or a low carrier concentration in the off-state to minimize leakage and a high carrier concentration in the on-state to maximize drive current, i.e., we want to dynamically control doping with the gate voltage of the transistor to break free of the N_{SD} optimization trade-off.

We propose here a dynamically doped FET, whose purpose is to turn the challenge of scaling into an opportunity (thin-film and multi-gate architecture technologies, which are the by-product of scaling, enable the manufacturability of such a device). It consists of a transistor that is dynamically doped by one of its own gate(s). This doping gate is located opposite (e.g., at the bottom) the source and drain metal contacts (e.g., located on top). Owing to its opposite position, the doping gate, unlike a conventional gate of length L, can now overlap, by a value ΔL on each side, the source and drain extensions to dynamically induces doping without increasing the footprint of the transistor (Fig. 12). This unconventional gate-positioning scheme would alleviate the need for strict alignment control between the doping-gate and the other gates in a multi-gate technology. We insist here that the D$_2$-FET remains a three-terminal device, like a conventional MOSFET. The doping gate is also the gate of the device for a single-gate device and share the same contact-voltage bias that any other conventional gates, if a multi-gate architecture is used (Fig. 12). It, therefore, does not require an additional contact compared with a MOSFET.

It should be reminded here that, scaling L is a way to scale the total contacted gate pitch, CGP, of a transistor, i.e., the minimal distance between the gate of two subsequent transistors. CGP is composed of the sum of L and the length of the highly-doped source-and-drain extensions, L_{SD} (see Figs. 1 and 12). L_{SD} is the sum of the spacer length, that separate the gate contact from the source and drain metal contact of the three-terminal transistor, as well as the metal-contact length (https://irds.ieee.org/editions/2018). The length of the doping gate is $L_{DG} = L + 2^* \ \Delta L$. Technological requirements impose that $\Delta L = L_{SD} - L_{SEP}$, L_{SEP} being a separation distance (typically at least half the spacer length, i.e. a few nm) (https://irds.ieee.org/editions/2018) needed to separate the doping gate from one transistor to that of the next. It is, therefore, longer than L, although it does not require a larger CGP footprint than a traditional top-sided gate of length L. To quantify this, the 2031 IRDS-dimensional targets for the so-called 1-nm-technology node and beyond are $L = 12$ nm, $L_{SD} = 14$ nm, and CGP $= 40$ nm (https://irds.ieee.org/editions/2018). The spacer length is 6 nm so that $L_{SEP} \geq 3$ nm, $\Delta L \leq 11$ nm, and $L_{DG} \leq L + 22$ nm. As can be seen, a comfortable margin is available for L_{DG}. This allows for keeping a good electrostatic control, as well as a relaxed t_S scaling, as can be seen for the Si case in Fig. 12b, even when using a very aggressive pitch scaling ($L = 3$ nm).

In the rest of this section, if not specified otherwise, we have, however, assumed a worst-case scenario for the D$_2$-FET with smaller L_{DG}. We used $\Delta L = L/2$ with a minimum value of $\Delta L = 4$ nm for $L \leq 8$ nm. This either assumes a very aggressive L_{SD} scaling, or the possibility of choosing a smaller ΔL value (i.e., $L_{SEP} \gg 3$ nm) in conjunction with high-doping, N_{SD2}, in the L_{SEP} ungated part of the device (as shown in Fig. 12c) to further reduce the contact resistance or for reducing the intrinsic gate capacitance (see discussion below) for instance. Our simulation results show that both cases of D$_2$-FETs (Fig. 12b, c) achieve similar I_{ON} and SS for same L and ΔL in case of ohmic or low Schottky-barrier contacts. The second scheme (Fig. 12c) could be advantageous in case of a high Schottky-barrier contact.

On Fig. 11a, several $L = 3$ nm DG HFS$_2$ D$_2$-FET characteristics are shown, one with no intentional doping ($N_{SD} \leq 1 \times 10^{19}$ cm^{-3}, which correspond to typical residual doping concentrations in the

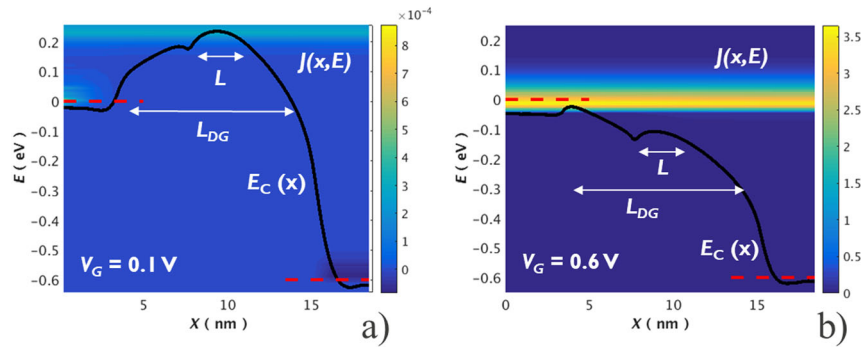**Fig. 13 Current spectrum and top conduction-band edge along the channel direction of a $L = 3$ nm HfS$_2$ D$_2$-FET.** $J(E, x)$ (surface plot) and $E_C(x)$ (—) of the simulated intrinsic ($N_{SD} \leq 1 \times 10^{19}$ cm^{-3}) $L = 3$ nm HfS$_2$ nD$_2$-FET of Fig. 11a (with the design of Fig. 12c) **a** in off-state at $V_G = 0.1$ V, and **b** in on-state at $V_G = 0.6$ V. $\Delta L = 4$ nm. $N_{SD2} = 4 \times 10^{20}$ cm^{-3}.

2D films, i.e., lowly doped or intrinsic extensions) and 2 with highly-doped extensions ($N_{SD} = 1$ and 2×10^{20} cm^{-3}). Our simulations show that for $N_{SD} \leq 1 \times 10^{19}$ cm^{-3}, the presence of a residual doping in the extensions has no impact on the current-voltage characteristics. The carrier concentration in the extensions is mainly determined by the doping-gate bias. In off-state, the conjunction of a low carrier concentration in the extensions and the extended doping-gate geometry allows for a large L_{eff} (typically $\geq 2 \times L$, Fig. 13a) and nearly ideal SS and low off-state current is achieved. In on-state, a high carrier concentration allows for a high-drive current. As can be seen, the intrinsic DG-D$_2$-FET, free from any chemical doping, already strongly outperforms the optimized $N_{SD} = 1 \times 10^{20}$ cm^{-3} DG-MOSFET. On Fig. 13b, however, it can be seen that in on-state, for a large gate overdrive, the current might be limited by the source part of the conduction-band barrier that is mostly controlled by the doping-gate, not by the top-gate.

In case, a large additional N_{SD} doping is used as an attempt to further boost the on-state current, the carrier concentration in the extensions is still dynamically controlled by the doping gate, but the dynamic-doping level at a given V_G can be enhanced vs the intrinsic case. In Fig. 11a, it is observed that the current drive can be slightly increased for $N_{SD} \geq 1 \times 10^{20}$ cm^{-3}, owing to the enhanced carrier concentration in the source, whereas SS is only slightly affected as the doping gate still deplete the extension in the off-state. As in the case of the regular MOSFET, an optimal doping of $N_{SD} = 1 \times 10^{20}$ cm^{-3} is observed for $V_{DD} = 0.6$ V. On the contrary to the MOSFET case, however, I_{ON} and SS sensitivity to doping variations are strongly reduced, and I_{ON} remains high, whereas SS remains low for all the simulated N_{SD} values. Finally, Supplementary Fig. 10 compares the $I_D(V_G)$ characteristics of SG and DG 1ML-HfS$_2$ MOSFETS and D$_2$-FETs for $L = 3$ nm and for $L = 5$ nm. It is shown that for the D$_2$-FET case, a simpler-to-fabricate SG architecture is as good or even better in term of drive current than that of a DG- D$_2$-FET. The SG-D$_2$-FET indeed keeps a similar and good electrostatic control (SS), when compared with that of the DG- D$_2$-FET case, hence similar drive-current (per gate). For $L \leq 3$ nm, the SG-D$_2$-FET I_{ON} typically outperforms that of the DG-D$_2$-FET device as the short top-gate drive only a small amount of additional current compared to the doping gate of the device. In the MOSFET case, a SG architecture is not sufficient to maintain a good electrostatic control for sub-10 nm devices. The SG device SS and I_{ON} is hence degraded compared to the DG-MOSFET case.

Figure 11b compare $L = 3$ nm optimized MOSFETS and D$_2$-FETs for the Si case. For the Si-D$_2$-FETs, the number of gates can be reduced, similarly to what was found for the HfS$_2$ case, and an intrinsic DG device was used instead of a GAA. Furthermore, the t_S scaling was relaxed towards L rather than ½ L, assuming $\Delta L = 4$ nm. This strongly reduces QC and boost I_{ON} of the D$_2$-FET, as for the square cross-section GAA the confinement is both in the width

(y-) and height (z-direction). In the rectangular cross-section DG-case, the width (y direction) is typically large compare to its height t_S that is further relaxed compared to the GAA case. In case $\Delta L = 11$ nm is used, the optimal t_S for the D$_2$-FET is even further relaxed to 4 nm and the performance is further boosted as SS is strongly improved. For Si, we found that the intrinsic case (i.e., unintentionally doped extensions with $N_{SD} \leq 1 \times 10^{19}$ cm^{-3}) is always better than the case with a larger N_{SD}. In any case, even in case chemical doping would be used in the D$_2$-FET, the related challenges (e.g., variability) would be reduced, one reason being the relaxed dimensions (L_{DG}, t_S) at same CGP.

Scaling perspective
Next, we investigate in Fig. 14 the scaling behavior for both optimized MOSFETs and D$_2$-FETs made of the less-explored HfS$_2$, the material showing the highest mobility and performing the best at $L = 5$ nm, and that of WS$_2$, the best performing material of the more studied group-VI TMDs. The optimized device results for the more conventional Si technology are shown as well. The evolution of their on-current vs L at a fixed off-state current of 10 nA/µm and a supply-voltage $V_{DD} = 0.6$ V is shown on the figure.

For the MOSFET case, all materials show degradation of their performance when L is scaled below 7.5 nm. This is related to SCEs and SDT as discussed before. Comparing the Si-GAA and WS$_2$ DG MOSFETs that achieve similar drive current at $L = 10$ nm, it can be seen that the GAA I_{ON} degrades faster when downscaling L. This is related to the better electrostatic control and QC-free characteristics of the 2D material over the GAA. None of these 2 materials, however, are able to meet the stringent high-performance IRDS I_{ON} targeted for year 2031 (https://irds.ieee.org/editions/2018) and a higher-mobility channel material is needed. HfS$_2$ on the other hand, owing to its outstanding transport properties, exceeds the target down to a channel length of ~6 nm.

Using the D$_2$-FET concept, it can be seen that the performance degradation with L is delayed to L below 5 nm in all cases. For such small gate lengths, the drive-current potential is strongly enhanced compared with that of the MOSFET case. At $L = 3$ nm, the Si, WS$_2$ and HfS$_2$ D$_2$-FETs show an I_{ON} gain of ~3× compared with their MOSFET counterparts. It is now possible to reach the IRDS target with a L of ~1 nm HfS$_2$ SG-D$_2$-FET transistor (using $\Delta L = 4$ nm). Our simulations finally show that, using this scheme, it is still possible to have a transistor effect using $L = 0$ nm, and that this device performs as well as a regular chemically doped multi-gate transistor with a 4 nm or longer L values for the case with 2D materials, although using only single-gated intrinsic semiconductor materials. This shows the promise of using the D$_2$-FET concept for sub-10 nm transistors and in particular for ultra-scaled high-mobility material devices that would be required to meet the stringent IRDS HP targets.
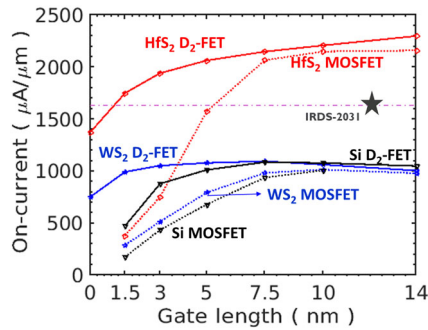
Fig. 14 Maximum achievable I_{ON} vs L for optimized n-type MOSFETs and D$_2$-FETs made of Si, WS$_2$, and HfS$_2$. A DG architecture is assumed for all the 2D MOSFETs. For the Si MOSFETs, a GAA transistor was used. A SG architecture is employed for the 2D D$_2$-FETs, whereas a DG transistor is used for the Si-D$_2$-FETs. EOT = 0.5 nm. $V_{DD} = 0.6$ V. $I_{OFF} = 10$ nA/μm. The current is normalized by the gate perimeter. For the D$_2$-FETs, we used $\Delta L = L/2$ with a minimum value of $\Delta L = 4$ nm for $L \leq 8$ nm.

It is to be noted that the larger L_{DG}, will however increase the intrinsic gate capacitance at same L and a trade-off between ΔL and the I_{ON} gain may exist for the speed performance of the D$_2$-FET. In modern scaled technologies, the total load capacitance of inverters or other digital circuits is often dominated by extrinsic (back-end-of-line) capacitances, such as the metal-line capacitance that is proportional to CGP (not L_{DG}) (https://irds.ieee.org/editions/2018). By enabling further downscaling with strongly improved drive-current, the D$_2$-FET may therefore also enable a power-delay benefit. To investigate the trends, we performed, here, a power-delay performance analysis of the basic building block of a digital circuit, i.e., a CMOS inverter, using scaled D$_2$-FET and MOSFET devices. The details about the process assumptions and layout, which determine the number of stacked-transistors per device and their geometries, as well as their back-end-of-line capacitance load that is considered in this analysis, are detailed in Supplementary Note 5 and Supplementary Figs 11 and 12.

Figure 15 compares the switching energy vs delay (EDP) of back-end-loaded high-performance invertor cells made with HfS$_2$ DG MOSFETs and HfS$_2$ SG-D$_2$-FETs, as well as that made with Si-GAA's and Si SG-D$_2$-FETs (typically the best device architectures per category of materials for MOSFETs and D$_2$-FETs). A more-detailed analysis to identify the best devices per category is available in Supplementary Fig. 13 and surrounding text. The inverters are loaded with a typical 50 contacted gate pitch-long metal line (https://irds.ieee.org/editions/2018). As CGP is reduced for shorter L, more aggressively scaled devices get a net capacitance reduction. The SG-D$_2$-FET devices that only require one spacer length instead of two (see Supplementary Fig. 11) benefit from a further CGP reduction at same L. The extrinsic capacitances of the cell layout are also included in the load capacitance. Again, the SG-D$_2$-FET devices are free from the extrinsic parasitic components C_{GSext} and C_{GDext} (see Supplementary Fig. 11), as the gate metal contact does not directly face the source and drain metal contacts.

In Fig. 15, V_{DD} is varied from 0.4 to 0.7 V. The best EDP performance is achieved by the $L = 0$ nm, HfS$_2$ SG-D$_2$-FET, which uses the simplest (SG) architecture and further yields the largest pitch reduction (CGP = 22 nm). It is closely followed by the $L = 5$ nm, HfS$_2$ DG MOSFET (CGP = 33 nm) and the $L = 3$ nm, HfS$_2$ SG-D$_2$-FET (CGP = 25 nm). The $L = 3$ nm, HfS$_2$ DG MOSFET (CGP = 31 nm) performance is strongly degraded, and typically worst (especially at larger V_{DD} where the speed performance saturates due to on-current saturation) to that of the $L = 5$ nm Si SG-D$_2$-FET (CGP = 27 nm). The latter two devices still comfortably outperform the $L = 12$ nm Si-GAA (CGP = 40 nm).
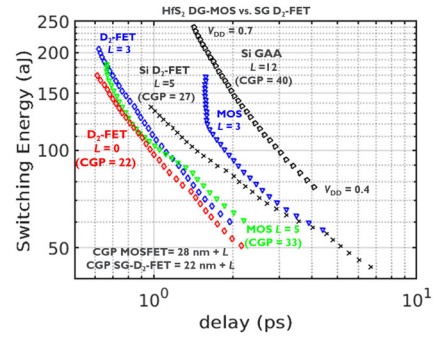

Fig. 15 Switching energy vs delay (EDP) of high-performance MOSFET and D$_2$-FET inverters. EDP of 1ML-HfS$_2$ high-performance inverter cells, at various V_{DD} (0.4 V to 0.7 V), made of $L = 5$ nm and $L = 3$ nm stacked DG MOSFETs (5 ribbons/device) and $L = 0$ nm and $L =$ nm stacked SG-D$_2$-FETs (nine ribbons/device). The EDP performance of Si HP inverter cells made of $L = 12$ nm stacked Si-GAA MOSFETs ($t_S = 5$ nm, 8 wires/device) and $L = 5$ nm Si SG-D$_2$-FETs ($t_S = 3$ nm, 7 ribbons/device) are also shown for comparison. The inverters are loaded with a 50 contacted-gate-pitch-long metal line (https://irds.ieee.org/editions/2018). The extrinsic capacitances of the cell layout are also included in the load capacitance. $I_{OFF} = 10$ nA/μm. $\Delta L = 4$ nm for the D$_2$-FETs.

These results further confirm and highlight the promising potential of the D2-FET device, paving the way towards ultimately scaled devices, with reduced process complexity and variability (e.g., reduced number of gates, larger t_S, doping free or reduced sensitivity to doping fluctuations…) and improved performance.

DISCUSSIONS

In summary, we have presented an in-depth study of the essential physics and performance potential of several 2D materials towards sub-10 nm gate length high-performance CMOS. We have argued that using very-advanced atomistic-simulation techniques including e-ph scattering, such as the DFT-NEGF technique we have used here, is required to achieve reliable and accurate results for 2D-materials. We have extracted from our simulations relevant 2D-material parameters such as m_{DoS} and mobilities for 5 TMD n- and p-type materials.

We have benchmarked n- and p-type MOSFETs made of six different 2D materials, including the most used Mo and W-based TMDs, P$_4$, as well as, the less-explored HfS$_2$ and ZrS$_2$ materials, against Si-GAA. Our results demonstrate the interest and better scaling potentials of HfS$_2$, ZrS$_2$, and WS$_2$ for sub-10 nm HP CMOS providing that a high-doping concentration could be used in the source and drain extensions to mitigate source-starvation effects. We have further shown that a high-drive current, meeting the stringent IRDS-2031 target down to $\sim L = 6$ nm, would be achievable using HfS$_2$. We finally proposed the D$_2$-FET concept that scales better than its MOSFET counterpart and seems very attractive for the sub-10-nm gate-length regime. Used in combination with a high-mobility material such as HfS$_2$, it allows for keeping the stringent ITRS HP on current when scaling down to 1 nm gate length. It further shows very attractive power-delay performance and its EDP performance keeps increasing when ultimately scaling down L to 0 nm using a simpler SG architecture with an ultra-compact design (CGP = 22 nm). The D$_2$-FET further addresses the grand-challenge of doping in ultra-scaled devices and 2D material in particular.

METHODS

Quantum transport solver

Our quantum transport solver, ATOMOS[18], was specifically developed for high-performance computing and the use of computationally heavy DFT

Hamiltonians. It is written in C++ and uses multi-threaded MPI with various levels of parallelism. Ultimately, any heavy vector-matrix or matrix-matrix operations are performed using BLAS and LAPACK.

ATOMOS core transport solver is a Real-Space NEGF solver based on the recursive-Green's function (RGF) algorithm[50]. For completeness, the equations for retarded (G^R), lesser ($G^<$) and greater ($G^>$) Green's functions read[33]:

$$G^R = \left(EI_N - H - \Sigma^R\right)^{-1}, \tag{1}$$

$$G^< = G^R \Sigma^< G^{R\dagger}, \tag{2}$$

$$G^> = G^R - G^{R\dagger} + G^<. \tag{3}$$

E is the scalar energy. I_N the identity matrix, H the device Hamiltonian, and $\Sigma^{R,<}$ the retarded, lesser self-energies that include the interaction terms (e.g. with the semi-infinite leads $\Sigma_C^{R,<}$ and the electron–phonon scattering terms $\Sigma_S^{R,<}$) are matrices of rank N, the total number of atoms in the device × the number of orbitals/atoms. We efficiently store H and other G matrices using our dedicated sparse block-matrix class, which we specifically customized for the RGF method.

The contact self-energies are computed with the Sancho-Rubio method[51]. Electron–phonon scattering is considered using the self-consistent Born approximation[25]. Assuming the phonons stay in equilibrium, the scattering self-energy may be written as[33]:

$$\Sigma_S^{\lessgtr}(\mathbf{r}_i, \mathbf{r}_j, E) = \int \frac{d\mathbf{q}}{(2\pi)^3} e^{i\mathbf{q}.(\mathbf{r}_i - \mathbf{r}_j)} |M_q|^2 \times \left(N_q + \frac{1}{2} \pm \frac{1}{2}\right) G^<(\mathbf{r}_i, \mathbf{r}_j, E \pm \hbar\omega_q) \tag{4}$$

where $\mathbf{q}$ and ω_q are the phonon wave vector and the corresponding angular frequency, $\hbar$ is the reduced Plank's constant, N_q is the phonon-occupation number. M_q is the e-ph coupling matrix, which depends on the exact scattering mechanism. For TMDs we used the DFT-computed e-ph parameters from ref. [15], for P_4 those from ref. [43]. Additional details about the e-ph scattering implementation can also be found in Supplementary Note 3. To ensure efficient load-balancing, a master-slave-approach-based dynamic scheduler is used to distribute the various energy-momentum (e-k) points between the different parallel ranks. For optimally generating the energy points, we rely on a recursive adaptive-grid algorithm[52], using a trapezoidal integration rule and a global-error estimator.

Similarly, a parallel Poisson solver with its own sparse class is used. To expedite the self-consistent Poisson-NEGF convergence, we employ a predictor-corrector method using the Newton scheme[53]. To predict the carrier changes with respect to potential variations, various semi-classical predictor functions have been implemented. For 2D materials, we can well fit the NEGF data using a 2D-DoS model, i.e., using a Fermi-Dirac integral of order 0 (See Charge and DoS Fitting section below and Supplementary Fig. 4). Additional adaptive-damping methods can be used, if the current and charge convergence criteria are not met within a given number of iterations. The anisotropic dielectric permittivity's are taken from ref. [54].

In this work, we have used the Wannierization technique[55] to express the DFT-Hamiltonian in a localized-orbital basis that is compatible with the RGF algorithm. The use of advanced and well-optimized algorithms together with high-performance parallel computing allow for scalable and fast calculations. For the 1 ML 2D devices simulated here, using a real-space DFT-Hamiltonian with longer-range interactions and dissipative transport, a full $I_D(V_G)$ curve is typically achieved within about an hour on one to a few 100 cores, using the latest generation Intel Xeon CPU.

DFT-Hamiltonian computation

The electronic states in the various TMD and BP monolayers are modeled using the DFT-based ab initio tool QUANTUM ESSPRESSO[20]. Both the geometry relaxation and the computation of the electronic structure are performed using the generalized gradient approximation and the optB86b exchange-correlation functional[21]. Spin-orbit coupling was not included. The plane-wave cutoff energy and the Monkhorst-Pack k-point grid for the Brillouin-zone integration, that we used for the relaxation and band structure calculations, were selected so that the total energy was well converged. The convergence criteria are set to <10^{-3} eV/Å for the forces acting on each ion, and a difference smaller than 10^{-3} eV for the total energy variation between two subsequent iterations. A vacuum layer of 25 Å was employed in the DFT simulations to cutoff the spurious interactions of the periodic images along the out-of-plane (z-direction, see inset of Fig. 2).

We then transformed the Bloch wavefunctions into MLWF, typically centered on the ions, with the wannier90 package[22]. Figure 2 and Supplementary Fig. 9 demonstrate the validity of our MLWF representation for the case of WS_2, ZrS_2, HfS_2, and P_4. ATOMOS uses the resulting supercell information, i.e., MLWF and atom positions, lattice vectors, and the localized Hamiltonian-matrix elements, as building blocks to create the full-device atomic structure and Hamiltonian-matrix. We kept in the device Hamiltonian, the required Wannier-Hamiltonian longer-range interactions (typically 12–15 Å). ATOMOS can further rotate the device geometry to a preferential channel orientation within the 2D layer. We assumed periodic boundary conditions in the width (y axis) direction. They were modeled with 24 k_Y points.

Charge and DoS fitting

From our NEGF simulations, we can extract the electron concentration vs $E_F - E_C$. As can be seen in Supplementary Fig. 4, this can be well fitted by a 2D-DoS model (using a Fermi-Dirac integral of order 0), from which the conduction-band DoS, N_{2D}, can be extracted by:

$$n = N_{2D}.\ln\left(1 + e^{\frac{E_F - E_C}{k_B T}}\right) \tag{5}$$

In Eq. (5), E_F is the Fermi level, E_C is the conduction-band edge, k_B is the Boltzmann constant and T is the absolute temperature. The NEGF-simulated carrier concentrations vs $E_F - E_C$ curves are extracted from a $L = 14$ nm device using the averaged values of a cross-section in the middle of the channel under low V_D (typically 1 mV) bias condition, whereas varying V_G. Owing to non-equilibrium transport, the NEGF Fermi level is only known and well defined at the source and drain contacts. At very low V_D, we can, however, safely assume a quasi-linear and close to equilibrium regime and that the Fermi level value in the middle of the channel is halfway between that of the source, E_{FS}, and drain, E_{FD}. Assuming an equivalent single-band parabolic effective-mass model, we can further define an equivalent DoS mass, m_{DoS}, by:

$$N_{2D} = \frac{m_{DoS}.k_B T}{\pi.\hbar^2} \times \frac{1}{t_S} \tag{6}$$

Note that this DoS mass capture the DFT-computed non-parabolicity of the occupied bands close to the conduction-band edge. In Eq. (6), $\hbar$ is the reduced Plank constant and t_S is the 2D-film thickness, ~0.6 nm for a monolayer TMD (the exact value we used for each studied monolayer can be found in Supplementary Table 1). t_S is used to convert the 2D density from per-area to per-volume unit. Similarly, for a p-type device, we can extract the valence-band DoS, N_{2D} from the DFT-NEGF-simulated hole concentration vs $E_V - E_F$ using:

$$p = N_{2D}.\ln\left(1 + e^{\frac{E_V - E_F}{k_B T}}\right) \tag{7}$$

where E_V is the valence-band edge. Using Eq. (6), it is then possible to extract the equivalent hole DoS mass.

REFERENCES

1. Kuhn, K. J. Considerations for ultimate CMOS scaling. *IEEE Trans. Electron Devices* **59**, 1813–1828 (2012).
2. Lu, W.-Y. & Taur, Y. On the scaling limit of ultrathin SOI MOSFETs. *IEEE Trans. Electron Devices* **53**, 1137–1141 (2006).
3. van Dal, M. J. H., et al. Demonstration of scaled Ge p-channel FinFETs integrated on Si. *Int. Electron Devices Meeting (IEDM)*, 521–524 (2012).

4. Colinge, J.-P. et al. Nanowire transistors without junctions. *Nat. Nanotechnol.* **5**, 225–229 (2010).

5. Vasen, T. et al. InAs Nanowire GAA n-MOSFETs with 12–15 nm Diameter. *IEEE 2016 Symposium on VLSI Technology (VLSIT)*, 1–2 (2016). https://doi.org/10.1109/VLSIT.2016.7573417.

6. Loubet, N. et al. Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET. *2017 Symposium on VLSI Technology (VLSIT)*, 230–231 (2017). https://doi.org/10.23919/VLSIT.2017.7998183.

7. Lee, C.-W., Yu, C. G., Park, J. T. & Colinge, J.-P. Device design guidelines for nanoscale MuGFETs. *Solid-State Electron.* **51**, 505–510 (2007).

8. Suk, S. D. et al. Investigation of nanowire size dependency on TSNWFET. *2007 Int. Electron Devices Meeting (IEDM)*, 891–894 (2007).

9. Afzalian, A. et al. Quantum confinement effects in capacitance behavior of multigate silicon nanowire MOSFETs. *IEEE Trans. Nanotech.* **10**, 300–309 (2011).

10. Afzalian, A. Ultimate FDSOI multi-gate MOSFETs and multi-barrier boosted Gate Modulated Resonant tunneling-FETs for a new high-performance low-power paradigm. *Nano-Semiconductors: Devices and Technology*, Kris Iniewski (Ed.) (CRC Press, Boca Raton 2012). https://doi.org/10.1201/9781315217468.

11. Afzalian, A. et al. A new F(ast)-CMS NEGF algorithm for efficient 3d simulations of switching characteristics enhancement in constricted tunnel barrier silicon nanowire MuGFETs. *J. Comput. Electron.* **8**, 287–306 (2009).

12. Novoselov, K. S. et al. Two-dimensional atomic crystals. *Proc. Natl. Acad. Sci. USA* **102**, 10451–10453 (2005).

13. Novoselov, K. S., Mishchenko, A., Carvalho, A. & Castro Neto, A. H. 2D materials and van der Waals heterostructures. *Science* **353**, 9439 (2016).

14. Radisavljevic, B. et al. Single-layer MoS2 transistors. *Nat. Nanotechnol.* **6**, 147–150 (2011).

15. Huang, Z., Zhang, W. & Zhang, W. Computational search for two-dimensional MX2 semiconductors with possible high electron mobility at room temperature. *Materials* **9**, 716 (2016).

16. Li, L. et al. Black phosphorus field-effect transistors. *Nat. Nanotechnol.* **9**, 372–377 (2014).

17. Haastrup, S. et al. The computational 2d materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Mater.* **5**, 042002 (2018).

18. Afzalian, A. & Pourtois, G. ATOMOS: An ATomistic MOdelling Solver for dissipative DFT transport in ultra-scaled HfS2 and Black phosphorus MOSFETs. *2019 International conference on simulation of semiconductor processes and devices (SISPAD)*, 1–4 (2019). https://doi.org/10.1109/SISPAD.2019.8870436.

19. Chuang, H.-J. et al. Low-resistance 2D/2D ohmic contacts: a universal approach to high-performance WSe2, MoS2, and MoSe2 transistors. *Nano Lett.* **16**, 1896–1902 (2016).

20. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 155–178 (2009).

21. Klimeš, J., Bowler, D. R. & Michaelides, A. Chemical accuracy for the van der Waals density functional. *J. Phys. Condens Matter* **22**, 022201 (2010).

22. Mostofi, A. A. et al. An updated version of wannier90: a tool for obtaining maximally-localised Wannier functions. *Comput. Phys. Commun.* **185**, 2309 (2014).

23. Keldysh, P. Diagram technique for nonequilibrium processes. *Sov. Phys. JETP* **20**, 1018 (1965).

24. Kadanoff, P. & Baym, G. *Quantum Statistical Mechanics*. (Benjamin, New York, 1962).

25. Afzalian, A. Computationally efficient self-consistent Born approximation treatments of phonon scattering for coupled-mode space non-equilibrium green's functions. *J. Appl. Phys.* **110**, 094517 (2011).

26. Liu, L., Kumar, S. B., Ouyang, Y. & Guo, J. Performance limits of monolayer transition metal dichalcogenide transistors. *IEEE Trans. Electron. Devices* **58**, 3042–3047 (2011).

27. Cao, W., Kang, J., Sarkar, D., Liu, W. & Banerjee, K. 2D semiconductor FETs-projections and design for sub-10 nm VLSI. *IEEE Trans. Electron. Devices* **62**, 3459–3469 (2015).

28. Agarwal, T. et al. Benchmarking of MoS2 FETs with multigate Si-FET options for 5 nm and beyond. *IEEE Trans. Electron. Devices* **62**, 4051–4056 (2015).

29. Szabó, Á., Rhyner, R. & Luisier, M. Ab initio simulation of single- and few-layer MoS2 transistors: effect of electron-phonon scattering. *Phys. Rev. B* **92**, 035435 (2015).

30. Song, S. et al. Wafer-scale production of patterned transition metal ditelluride layers for two-dimensional metal–semiconductor contacts at the Schottky–Mott limit. *Nat. Electron.* **3**, 207–215 (2020).

31. Suryavanshi, S. V. & Pop, E. S2DS: physics-based compact model for circuit simulation of two-dimensional semiconductor devices including non-idealities. *J. Appl. Phys.* **120**, 224503 (2016).

32. Rim, K., Narasimha, S., Longstreet, M., Mocuta, A. & Cai, J. Low field mobility characteristics of sub-100 nm unstrained and strained Si MOSFETs. *Int. Electron Devices Meeting (IEDM)*, 43–46 (2002). https://doi.org/10.1109/IEDM.2002.1175775.

33. Mahan, G. D. *Many-Particle Physics*. (Plenum Press, New York, 1988).

34. Afzalian, A. et al. Physics and performance of III-V nanowire broken-gap heterojunction TFETs using an efficient tight-binding mode-space NEGF model enabling million-atom nanowire simulations. *J. Phys. Condens. Matter* **30**, 254002 (2018).

35. Cheng, L. & Liu, Y. What limits the intrinsic mobility of electrons and holes in two dimensional metal dichalcogenides? *J. Am. Chem. Soc.* **140**, 17895–17900 (2018).

36. Fischetti, M. V. & Vandenberghe, W. G. Mermin-Wagner theorem, flexural modes, and degraded carrier mobility in two-dimensional crystals with broken horizontal mirror symmetry. *Phys. Rev. B* **93**, 155413 (2016).

37. Zhang, W. et al. Two-dimensional semiconductors with possible high room temperature mobility. *Nano Res.* **7**, 1731–1737 (2014).

38. Jin, Z., Li, X., Mullen, J. T. & Kim, K. W. Intrinsic transport properties of electrons and holes in monolayer transition-metal dichalcogenides. *Phys. Rev. B* **90**, 045422 (2014).

39. Klinkert, C. et al. 2-D materials for ultrascaled field-effect transistors: one hundred candidates under the ab initio microscope. *ACS Nano* **14**, 8605–8615 (2020).

40. Afzalian, A., Passlack, M. & Yeo, Y.-C. Scaling perspective for III-V broken gap nanowire TFETs: an atomistic study using a fast tight-binding 15 mode-space NEGF model. *Int. Electron Devices Meeting (IEDM)*, 30.1.1–30.1.4 (2016). https://doi.org/10.1109/IEDM.2016.7838510.

41. Liu, F., Wang, Y., Liu, X., Wang, J. & Guo, H. Ballistic transport in monolayer black phosphorus transistors. *IEEE Trans. Electron Devices* **61**, 3871–3876 (2014).

42. Lam, K.-T., et al. Effects of interlayer interaction in van der Waals layered black phosphorus for sub-10 nm FET. *IEEE Int. Electron Device Meeting (IEDM)*, 12.2.1–12.2.4 (2015). https://doi.org/10.1109/IEDM.2015.7409681.

43. Gaddemane, G. et al. Theoretical studies of electronic transport in mono- and bilayer phosphorene: a critical overview. *Phys. Rev. B* **98**, 115416 (2018).

44. Fischetti, M. V., et al. Simulation of electron transport in high-mobility MOSFETs: density of states bottleneck and 'source starvation'. *Int. Electron Devices Meeting (IEDM)*, 109–112 (2007).

45. Huet, K., Mazzamuto, F., Tabata, T., Toqué-Tresonne, I. & Morib, Y. Doping of semiconductor devices by laser thermal annealing. *Mater. Sci. Semicond. Process.* **62**, 92–102 (2017).

46. Aldegunde, M., Martinez, A. & Barker, J. R. Study of discrete doping-induced variability in junctionless nanowire MOSFETs using dissipative quantum transport simulations. *IEEE Electron Device Lett.* **33**, 194–196 (2012).

47. Dolui, K., Rungger, I., Das, C. P. & Sanvito, S. Possible doping strategies for MoS2 monolayers: an ab initio study. *Phys. Rev. B* **88**, 075420 (2013).

48. Giddings, A. D. et al. Sn incorporation in ultrathin inas nanowires for next-generation transistors characterized by atom probe tomography. *ACS Appl. Nano Mater.* **2**, 1253–1258 (2019).

49. Moriyama, N., Ohno, Y., Kitamura, T., Kishimoto, S. & Mizutani, T. Change in carrier type in high-k gate carbon nanotube field-effect transistors by fixed charges. *Nanotechnology* **21**, 165201 (2010).

50. Svizhenko, A., Anantram, M., Govindan, T., Biegel, R. & Venugopal, R. Two-dimensional quantum mechanical modeling of nanotransistors. *J. Appl. Phys.* **91**, 2343–2354 (2002).

51. Lopez Sancho, M. P., Lopez Sancho, J. M., Sancho, J. M. L. & Rubio, J. Highly convergent schemes for the calculation of bulk and surface green functions. *J. Phys. F: Met. Phys.* **15**, 851–858 (1985).

52. Afzalian, A., Colinge, J. P. & Flandre, D. Physics of gate modulated resonant tunneling (RT)-FETs: multi-barrier MOSFET for steep slope and high on-current. *Solid-State-Electron.* **59**, 50–61 (2011).

53. Trellakis, A. & Galick, A. T. Iteration scheme for the solution of the two-dimensional Schrödinger-Poisson equations in quantum structures. *J. Appl. Phys.* **81**, 7880 (1997).

54. Laturia, A., Van de Put, M. L. & Vandenberghe, W. G. Dielectric properties of hexagonal boron nitride and transition metal dichalcogenides: from monolayer to bulk. *npj 2D Mater. Appl* **2**, 6 (2018).

55. Calzolari, A., Marzari, N., Souza, I. & Buongiorno Nardelli, M. Ab initio transport properties of nanostructures from maximally localized Wannier functions. *Phys. Rev. B* **69**, 035108 (2004).

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

A.A. developed the theory, simulation code, and concepts (e.g., the D$_2$-FET). He performed the simulations and analyses and wrote the manuscript.

COMPETING INTERESTS

The author declares no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41699-020-00181-1.

Correspondence and requests for materials should be addressed to A.A.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.