Ultra-strength materials

Ting Zhu\textsuperscript{a,*,} Ju Li\textsuperscript{b, *}

\textsuperscript{a} Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30068, USA
\textsuperscript{b} Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA

A R T I C L E   I N F O

Article history:
Received 26 February 2010
Received in revised form 12 April 2010
Accepted 13 April 2010

A B S T R A C T

Recent experiments on nanostructured materials, such as nanoparticles, nanowires, nanotubes, nanopillars, thin films, and nanocrystals have revealed a host of “ultra-strength” phenomena, defined by stresses in a material component generally rising up to a significant fraction ($>\frac{1}{10}$) of its ideal strength – the highest achievable stress of a defect-free crystal at zero temperature. While conventional materials deform or fracture at sample-wide stresses far below the ideal strength, rapid development of nanotechnology has brought about a need to understand ultra-strength phenomena, as nanoscale materials apparently have a larger dynamic range of sustainable stress (“strength”) than conventional materials. Ultra-strength phenomena not only have to do with the shape stability and deformation kinetics of a component, but also the tuning of its physical and chemical properties by stress. Reaching ultra-strength enables “elastic strain engineering”, where by controlling the elastic strain field one achieves desired electronic, magnetic, optical, phononic, catalytic, etc. properties in the component, imparting a new meaning to Feynman’s statement “there’s plenty of room at the bottom”. This article presents an overview of the principal deformation mechanisms of ultra-strength materials. The fundamental defect processes that initiate and sustain plastic flow and fracture are described, and the mechanics and physics of both displacive and diffusive mechanisms are reviewed. The effects of temperature, strain rate and sample size are discussed. Important unresolved issues are identified.

\textcopyright 2010 Elsevier Ltd. All rights reserved.
1. Introduction

This article examines a broad category of materials behavior that can be characterized as ultra-strength phenomena, meaning the material's sample-wide stress has reached a significant fraction of its ideal strength [1–3], and this state persists for a considerable period of time, say seconds or years, without significant stress relaxation. The realm of ultra-strength phenomena is broader than that of high-pressure physics, because large non-hydrostatic stresses due to shear, tension and compression may be sustained. In materials, the applied non-hydrostatic load tends to be relaxed by plastic deformation and fracture. So the amount of stress a conventional material can sustain is often only a tiny fraction (a few thousandth) of the corresponding elastic moduli. In nanostructured materials, however, a much larger dynamic range of non-hydrostatic stresses can be achieved, thereby opening wide avenues for controlling the physical and chemical properties of materials, by for example elastic strain engineering (ESE).

This article reviews how to achieve ultra-strength by defeating various mechanisms of relaxation – the so-called inelastic deformation mechanisms – at finite temperatures and sample sizes, with various initial material imperfections, so that ultra-strength can be obtained and controlled. At the beginning, we will motivate why ultra-strength is highly desirable, especially in the age of nanotechnology with advance of tools capable of resolving, manipulating and building materials at the nanometer scale [4,5]. It will become clear that ultra-strength phenomena not only have to do with shape stability and deformation kinetics of materials under stress and temperature – the traditional concerns of mechanics of materials – but also the functional properties of materials.
2. Ultra strength and elastic strain engineering

It is known that a material’s physical and chemical properties are functions of the lattice parameters of the underlying crystal lattice, or the elastic strain $\varepsilon_{\text{elastic}}$ with respect to the stress-free reference state. This is because fundamentally, electronic structure changes with $\varepsilon_{\text{elastic}}$ [6]. Electronic band gaps open or close with $\varepsilon_{\text{elastic}}$ [7–10,15,16,20,23], magnetic [13,14,24,25], optical or plasmonic [16–19,21], ionic [26,27] or phononic [3,12,28], thermoelectric [11], catalytic [22], etc. Generally speaking, given any material property $A$ that one wants to optimize, say the carrier mobility [29–33,23], its derivative with respect to the elastic strain $\partial A/\partial \varepsilon_{\text{elastic}}$ is non-zero at zero strain (some components of this gradient may be zero due to symmetry, but not all components). Thus, if an increase of elastic strain does not improve the property with respect to the zero-elastic-strain state, a decrease will do. Following the gradient $\partial A/\partial \varepsilon_{\text{elastic}}$ in six-dimensional strain space with a steepest descent kind of algorithm can lead to the optimal property of $A$.

The idea of elastic strain engineering is rather straightforward and has existed in various communities for many years. In at least one instance it has found a major commercial application: the so-called Strained Silicon technology [29–33,23,34–39] is already deployed in almost all the computers we buy today, at a moment in history when the characteristic feature size in integrated circuits has shrunk to 45 nm and below. The working principle of Strained Silicon technology is that by applying tensile elastic strain of the order 1% on a thin channel of silicon, more than 50% enhancement in room-temperature mobility of charge carriers can be obtained. Strained Silicon technology is now advertised as one of the major advancements postponing the eventual breakdown of Moore’s law [35,39].

What are the limits to elastic strain engineering? There is an idealistic answer and a practical answer. The practical answer is more conservative, by taking into account temperature, time and pre-existing defects.

Even idealistically, a crystal lattice is not infinitely strong and cannot be elastically strained beyond its ideal strain [40–60,3] at temperature $T = 0$ K, without losing homogeneity of the lattice or undergoing phase transformation. The ideal strain $\varepsilon_{\text{ideal}}$ is an intrinsic property of an “ideal” starting atomic structure that contains absolutely no “defects”. Here “defects” are in quotation marks because it depends on the perspective: for instance one may wish to define the ideal strain of an atomically flat Pt (1 1 1) surface, the so-called surface ideal strength [61,62], in contrast to the usual bulk ideal strength. According to this view, the atomically flat surface which has two translational symmetries is not a “defect” but rather the ideal starting structure itself. Under very large elastic strain at $T = 0$ K it may spontaneously break translational symmetries to nucleate displacive defects or undergo surface phase transformations.

Because in 3D, strain is a symmetric tensor with 6 degrees of freedom, $\varepsilon_{\text{ideal}}$ is in fact a five-dimensional bounding surface, enclosing the origin in a six-dimensional strain space. Crossing this boundary at $T = 0$ K by elastically straining the ideal structure leads to spontaneous and nearly instantaneous (much faster than most experimental timescales, usually picoseconds) large structural changes, whereas moving along arbitrary strain paths within the bounding surface causes no such discontinuities.

Associated with $\varepsilon_{\text{ideal}}$ is the ideal stress $\sigma_{\text{ideal}}$. The reasons for defining these tensorial ideal limits are (a) they can be easily calculated from first principles, (b) they reveal the basic features of electronic structure and chemical bonding, such as bond directionality [53] and “brittle” versus “ductile” tendencies [58,63], and (c) as to be shown in Section 3, the actual measured mechanical strengths in certain classes of experiments scale roughly with the theoretically calculated $\sigma_{\text{ideal}}$, and sometimes even closely approach them. Although anomalies can happen, in most cases we expect a material’s strength to decrease with increasing temperature, waiting time and addition of defects. Therefore, we may regard
the ideal stress as the upper bound to material strength, and the ideal strain as the upper bound to elastic strain engineering (ESE).

Even though ESE is a rather straightforward concept, it is very powerful and potentially a revolutionary one in the age of nanotechnology. The reason is that elastic strain is a continuous variable. So instead of just “silicon”, we actually have access to 1% strained silicon, 2% strained silicon, 5% strained silicon, 5.1% strained silicon, etc. indeed an uncountably infinite possibilities of elastically strained silicons, with wildly changing physical properties [3,6–39]. By controlling the elastic strain, we add six continuous new dimensions to the conventional materials we know.

However, ESE is still under-appreciated and vastly under-explored today, unlike alloying – the engineering of continuous chemical composition which started in the Bronze Age – as a means to control material properties. The reason becomes more obvious when one realizes that the 1% tensile elastic strain adopted in today’s Strained Silicon technology is applied to silicon features at 45 nm. If one applies tensile strain to a macroscopic block of silicon or even a silicon wafer, our day-to-day experience tells us it would fracture long before 1% strain is reached. In other words, the 45-nm feature size somehow enables elastic strain engineering, that historically was not feasible on say, 100 μm feature-sized silicons.

Similarly, although metallic crystals with absolutely no defects can sustain ideal strain of the order 10%, a macroscopic block of conventional metals with an initial defect population would flow plastically long before 1% applied strain is reached. That is, even though from our day-to-day experience, conventional metals can indeed deform up to hundreds of percent, most of the applied strain is carried by inelastic strain \( \varepsilon_{\text{inelastic}}(\mathbf{x}) \), where \( \mathbf{x} \) denotes the position vector, and only a tiny part (typically less than two tens of one percent) is elastic strain, as represented by, e.g., the additive partition \( \varepsilon(\mathbf{x}) = \varepsilon_{\text{elastic}}(\mathbf{x}) + \varepsilon_{\text{inelastic}}(\mathbf{x}) \), where \( \varepsilon \) represents total strain of the volume element.

For a material component in service to be functional, such as a silicon channel in metal–oxide–semiconductor field-effect transistor (MOSFET), two conditions must hold. First, the component shape must be stable, over long periods of time such as a few years – the expected life of the device. If the silicon channel fractures, or a metallic interconnect deforms too much plastically, the component and subsequently the device may fail prematurely. Second, the component material must have the right properties, such as carrier mobility, band gap, thermoelectric figure-of-merit, etc. It turns out that in the case of elastic-strain engineered components, both conditions depend critically on how \( \varepsilon_{\text{inelastic}} \) behaves under stress. In an ideal defect-free structure at \( T = 0 \, \text{K} \), \( \varepsilon_{\text{inelastic}} \) stays zero unless the ideal strength is reached. This is not the case in realistic materials which have initial defects, and \( T \) is non-zero.

The inelastic strain is physically attributed to dislocation glide [64,65], deformation twinning [66], crack formation and growth [67], phase transformation [68], mass transport by diffusion [69,70], etc. In general, \( \varepsilon_{\text{elastic}} \) and \( \varepsilon_{\text{inelastic}} \) have very distinct roles and different dependencies on stress. \( \varepsilon_{\text{elastic}} \) depends locally on stress in a one-on-one correspondence, whereas the rate of change of \( \varepsilon_{\text{inelastic}} \) depend on stress with a highly non-linear, and temperature and history dependent relation (may also be non-local, but we will not delve into it here). At low to intermediate temperatures, the rate of inelastic strain can be sharply thresholded: that is, unless the stress exceeds a certain threshold \( \sigma_{\text{thresh}} \), this rate is practically negligible. It is then convenient to call \( \sigma_{\text{thresh}} \) the material strength, for if stressed below \( \sigma_{\text{thresh}} \), the material does not appear to deform inelastically. \( \sigma_{\text{thresh}} \) itself is a function of temperature, material microstructure, and the timescale of observation. This last dependence has to do with the precise meaning of the term “practically negligible” inelastic strain rate, for even “Mountains flow before the Lord” [71]. Clearly, “practically negligible” strain rate means different things to an engineer working in the semiconductor industry where the relevant timescale is a few years, as to a geologist studying glacier flow over geological timescales.

The threshold stress \( \sigma_{\text{thresh}} \) is the aforementioned practical limit to ESE. It is smaller than the ideal limit \( \sigma_{\text{ideal}} \). Let us take uniaxial tension as an example. The ideal tensile strength \( \sigma_{\text{ideal}} \) is roughly \( E/10 \) [72,59], as shown from ab initio calculations [58], where \( E \) is Young’s modulus at the stress-free state. Most metals, however, plastically deform at stresses around \( 10^{-4}E \) to \( 10^{-3}E \). The large discrepancy between the ideal and the practical limits of stress also occurs for materials undergoing shear [72]. This actually motivated Orowan [73], Polanyi [74] and Taylor [75] to postulate the existence of dislocations in 1934, after the ground-breaking work of Frenkel on ideal strength in 1926 [40,76].
In recent years, however, a large number of experiments have emerged whereby the measured average stress of components starts to approach a significant fraction of the ideal strength [77]. This means a very large elastic strain energy density can be stored in real material components without significant relaxation, over seconds, hours or even years. These experiments, which typically involve some nanoscale characteristic lengths (but not always) such as grain size, film thickness or sample size, are introduced in detail in Section 3. It is worth noting here that by achieving very high threshold stresses, the dynamic range for modifying a material’s physical and chemical properties by elastic strain engineering becomes larger and larger, thus imparting an additional meaning to Feynman’s statement “there’s plenty of room at the bottom” [78].

The above experimental realities motivate defining a new concept: ultra-strength [1–3], which means the material’s sample-wide stress reaching a significant fraction of its ideal strength, over certain extended period of time at finite temperature. The qualifications of space and time are important here. “Sample-wide” is necessary because even in conventional material components such as a steel bridge truss or a rivet, near dislocation cores, and crack tips the stress can be very high, but those regions are only a tiny fraction of the component volume. “Extended time period” is also necessary because in molecular dynamics simulations [79–81], it is often rather easy to achieve high stress without significant relaxation, over a simulated time period of nanoseconds, but that does not mean relaxations cannot happen over seconds, hours or years in real experiments. Thus achieving pervasive high stresses inside a large enough space–time volume without appreciable relaxation is the true meaning of ultra-strength.

For the sake of concreteness, we take “significant fraction” to be \( \sigma / \sigma_{\text{ideal}} > 1/10 \) irrespective of shear, tensile or compressive stresses, unless otherwise specified. The \( 1/10 \)-rule is convenient because the ideal strength for tension and shear is approximately \( \sigma / \sigma_{\text{ideal}} \approx 1/22 \) of the elastic modulus [59]. Thus, under uniaxial tension, if the sample can sustain tensile stress greater than \( E/100 \), ultra-strength is likely reached. While the stress is high, linear elasticity, in general, holds well [49–51,56,59]. Therefore, as a rule of thumb, if the sample-wide elastic strain exceeds \( 1\% \), ultra-strength is likely reached. According to this definition, the present Strained Silicon technology [34–39] is already in the ultra-strength regime. Similarly in shear, if the resolved shear stress is greater than \( \mu / 100 \), where \( \mu \) is the shear modulus, ultra-strength is likely reached. Correspondingly, the sample-wide elastic shear strain (engineering) can exceed \( 1\% \). The “1% elastic strain rule” is a convenient mnemonic for ultra-strength.

The shear stress tends to drive bond switching (correlated bond breaking and bond reformation with no reduction of total bond count), whereas the tensile stress can drive irreversible bond loss, i.e., net reduction of an atom’s coordination number, especially in brittle materials like ionic and covalent solids. Compressive stresses, including hydrostatic compression, can also drive phase transformations [21,82–84]. With finite elastic strain energy stored in a material, there are often two options for relaxation: plastic deformation which corresponds to bond switching, and fracture which corresponds to net bond loss and eventual surface creation, although there are many subtle connections between the two [80]. Both plasticity and fracture contribute to inelastic strain. Plasticity tends to be the more effective energy dissipation channel in the long run, because bond switching can be repeated many times and is a “renewable” dissipation channel, whereas bond loss tends to be less “renewable”. This fundamentally has to do with how the effective elastic compliance of a system changes with the bond density [63]. The actual bond switching and bond loss processes occur in materials asynchronously and inhomogeneously, near defect cores such as dislocations and cracks. Nonetheless the ideal strengths, which correspond to homogeneous straining of the bonds, reveal much about the brittle versus ductile tendencies of a material. A systematic trend has been discovered by comparing \( \sigma_{\text{ideal}} \) to \( \sigma_{\text{ideal}} \) in 22 simple metals and ceramics [58]. It was found that \( \sigma_{\text{ideal}} / \sigma_{\text{ideal}} \) is relatively small in metals compared to ceramics; and within metals, the more ductile the metal, the smaller is \( \sigma_{\text{ideal}} / \sigma_{\text{ideal}} \), so the shear dissipation channel is triggered frequently long before the stress can rise high enough to trigger decohesion.

Large electronic-structure changes are expected to occur in materials in the ultra-strength regime [7,8,14,9,10,20]. For example, reaching ideal shear strength in semiconductors can be related to vanishing of the band gap and electron de-localization [6,54]. Under high pressure, metals like lithium and sodium can turn semiconducting [83,84], and become optically transparent [21]. These
striking stress effects on physi-chemical properties should be anticipated because bond breaking is fundamentally an electronic phenomena, and ultra-strength means getting somewhat “close” to spontaneous bond breaking. In contrast, if a material is not in the ultra-strength regime, for example in conventional materials where the elastic strain typically stays below 0.2%, then it is unlikely to experience striking changes in its electronic structure and its physi-chemical properties are not likely to be dramatically different from those of the stress-free material; at most, it may experience linear effects like piezoelectricity.

A vast space for material property exploration and tuning thus opens up with elastic strain engineering [3,6–39], if we can elevate the “practical strength” of materials to ultra-strength. Feynman’s statement “there’s plenty of room at the bottom” [78] again proves to be visionary, because introducing nanoscale microstructure is a main method of achieving ultra-strength. In this article, we will review how much we understand about different inelastic relaxation mechanisms, so we can purposefully defeat them and maintain pervasive high stresses in a prescribed space–time domain. Given how much was already gained in human civilization by tuning the continuous chemical composition of materials, it is quite exciting to speculate on how much are yet to be gained by tuning the six elastic strain variables, now that ultra-strength is an experimental reality (Section 3). One may even argue that continuous elastic strain has certain advantages over continuous chemical composition, since elastic strain can be changed very quickly in situ, and does not require implantation, diffusion or annealing.

3. Ultra-strength experiments

In 1949 Bragg and Lomer reported an experimental investigation of in situ deformation of an extended raft of bubbles floating on a soap solution [85,86], and found the bubble crystal can deform elastically up to shear strain about 10%. So the bubble raft qualifies as an ultra-strength system. When they compared the measured elastic limit with theoretical estimate of ideal shear strength using a bubble-bubble interaction pair potential [87], they even noted that the experimental elastic limit “cannot be simply correlated with the critical shear strain for an infinite lattice, because in this experiment the dislocations start from a free edge”, so the “surface ideal strength” [61,62] rather than “bulk ideal strength” should be the more appropriate benchmark. Many years later, by applying indentation load with a perfectly round indenter, the maximum in shear stress was achieved inside the bubble crystal in a controlled manner, and homogeneous dislocation nucleation

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured strength of μm-diameter sized whiskers (adapted from [97]).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Material</th>
<th>Max elastic strain (%)</th>
<th>Method of testing</th>
<th>Method of growth</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fe</td>
<td>4.9</td>
<td>Tension</td>
<td>Halide reduction</td>
<td>[96]</td>
</tr>
<tr>
<td>Cu</td>
<td>2.8</td>
<td>Tension</td>
<td>Halide reduction</td>
<td>[96]</td>
</tr>
<tr>
<td>Ag</td>
<td>4.0</td>
<td>Tension</td>
<td>Halide reduction</td>
<td>[96]</td>
</tr>
<tr>
<td>Au</td>
<td>1.9</td>
<td>Tension</td>
<td></td>
<td>[99]</td>
</tr>
<tr>
<td>Ni</td>
<td>1.8</td>
<td>Tension</td>
<td>Halide reduction</td>
<td>[96]</td>
</tr>
<tr>
<td>Si</td>
<td>2.0</td>
<td>Tension</td>
<td>Halide reduction</td>
<td>[100,101]</td>
</tr>
<tr>
<td>Zn</td>
<td>2.0</td>
<td>Tension</td>
<td></td>
<td>[102]</td>
</tr>
<tr>
<td>NaCl</td>
<td>2.6</td>
<td>Tension</td>
<td>Vapor condensation</td>
<td>[103]</td>
</tr>
<tr>
<td>SiO2</td>
<td>5.2</td>
<td>Tension</td>
<td>Vapor condensation</td>
<td>[97]</td>
</tr>
<tr>
<td>Al2O3</td>
<td>3.0</td>
<td>Tension</td>
<td>Vapor condensation</td>
<td>[97]</td>
</tr>
<tr>
<td>MoO3</td>
<td>1.0</td>
<td>Tension</td>
<td>Vapor condensation</td>
<td>[97]</td>
</tr>
<tr>
<td>C</td>
<td>2.0</td>
<td>Tension</td>
<td>Vapor condensation</td>
<td>[97]</td>
</tr>
<tr>
<td>Sn</td>
<td>2–3</td>
<td>Bending</td>
<td>Growth from solid</td>
<td>[95]</td>
</tr>
<tr>
<td>Ge</td>
<td>1.8</td>
<td>Bending</td>
<td>Halide reduction</td>
<td>[101]</td>
</tr>
<tr>
<td>ZnO</td>
<td>1.5</td>
<td>Bending</td>
<td>Halide reduction</td>
<td>[101]</td>
</tr>
<tr>
<td>ZnS</td>
<td>1.5</td>
<td>Bending</td>
<td>Vapor condensation</td>
<td>[104]</td>
</tr>
<tr>
<td>LiF</td>
<td>3</td>
<td>Bending</td>
<td>Cleavage</td>
<td>[105,97]</td>
</tr>
<tr>
<td>MgSO4·7H2O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydroquinone, etc.</td>
<td>&gt;2</td>
<td>Bending</td>
<td>Precipitation</td>
<td>[106]</td>
</tr>
</tbody>
</table>
was observed [88–93]. The bulk ideal strength was achieved in this case. Note that because the bubbles themselves are macroscopic objects (0.2–2 mm in diameter), the effect of thermal fluctuations is negligible at the bubble crystal level. Thus effectively the bubble crystal mimics a $T = 0$ K crystal.

For real atomic crystals at finite temperatures, in 1924 G.F. Taylor performed mechanical tests on 30 μm diameter antimony wire, and found “it may be bent repeatedly without breaking” [94], although bulk antimony is very brittle. Taylor estimated the tensile strength of his antimony wires to be 180–220 MPa. Since the Young’s modulus of antimony is 55 GPa, the maximum elastic strain achieved in the wire is 0.4%, approaching but still not in the ultra-strength regime.

In 1952 Herring and Galt performed experiments on tin whiskers 1.8 μm in diameter, and found those whiskers can “sustain large strain without creep”. They showed that an elastically bent whisker sustaining maximum tensile strain of 0.6% – after a week in this position – straightened out perfectly when the restraints were removed. Herring and Galt estimated these whiskers can sustain 2–3% elastic strain [95]. This is in contrast to bulk tin, which can only sustain about 0.01% elastic strain before yielding in simple uniaxial tension [95].

Brenner subsequently performed tension tests on a variety of whiskers with low initial defect density, typically μm scale in diameter and mm scale in length [96,97], at finite temperature. Table 1 also presents some other strength measurements on ceramic and metallic whiskers performed in the 1950s. These results demonstrated that ultra-strength is indeed achievable. Although whiskers sounded “small” in the 1950s, dramatic miniaturization of devices has occurred in the last few decades so that now many material components in commercial microchips and MEMS/NEMS devices [98] are dwarfed in size by these whiskers of the 1950s. Elastic strain engineering is now an industrial reality, in the form of Strained Silicon technology [34–39] in microchips, and poised to become realities in other domains of nanotechnology, as surely as chemical metallurgy – the engineering of continuous chemical composition – spread in civilizations.

For widespread adaptation of elastic strain engineering, four major scientific challenges need to be overcome: (a) how to synthesize material systems that can potentially sustain ultra-high elastic strain, (b) how to experimentally control the loads and measure the strain/stress using advanced, miniaturized instrumentation, (c) how to measure, understand and predict the dependencies of physi-chemical properties (electronic, magnetic, optical, phononic, thermoelectric, catalytic, etc.) on elastic strain, and (d) how to understand the different mechanisms by which stored elastic strain energy can relax from the material component, so ultra-strength can be reliably achieved and designed for different temperatures, sizes and expected service lifetimes.

For Challenge (a), in the last two decades a large number of nanoscale objects and nanostructured materials have been synthesized, such as nanotubes, nanowires, nanoparticles, bulk nanocrystals, nano-porous materials, and thin films and multi-layers. Many are potentially ultra-strength systems ready for elastic strain engineering. In Table 1 of Ref. [2] there is a sample list of mechanical strength measured for carbon nanotubes, ceramic and metallic nanowires, nano-spheres and graphene. One can be convinced from these data that ultra-strength can be achieved in a diverse number of systems, and sometimes very close to the theoretical strength. For Challenge (c), physical and chemical measurements [8–10,14,15,17–19,21,24,25,27–30,32,34–39] combined with first-principles and semi-empirical modeling [3,6,7,11–13,16,20,22,23,26,31,33,38,82–84] are needed, which is quite involved and may be the subject of another review. Challenges (b) and (d) fall in the realm of mechanics of materials, and are the main subjects of this review.

In relation to (d), Brenner noted that silver whiskers “exhibited considerable plastic deformation prior to fracture”, while Cu and Fe whiskers “snapped without any observable amount of plastic deformation”. Some ceramic whiskers certainly failed by what appeared to be cleavage fracture [104,103,100,101,107,108]. Thus both plasticity (bond switching) and fracture (bond loss) channels of dissipation may be activated, driven by the nucleation and motion of defects such as dislocations [64,65], grain boundaries [109–113], deformation twins [66,114,77], cracks [67], and point defects [69,70], which can also interact with each other in intricate manners. To experimentally resolve the evolution dynamics of defect populations inside ultra-strength materials, for (d), advanced characterization techniques, using for instance in situ electron microscopy [109,111,115–117], in situ X-ray [118,119] and neutron [120] diffraction, etc. are needed. The confluence of advanced mechanical
testing and advanced characterization techniques [121,122] is often the experimental basis for making further progress in (d).

Here we make note of several significant experimental developments since the 1950s. The first is the invention of atomic force microscopy (AFM) [123] and a related method, nanoindentation, where a sharp conical or pyramidal tip is pushed into a material half-space and both the indentation force and displacement are measured [124]. The tip of the mechanical probe is not perfectly sharp, of course, the radius of curvature of which can be estimated from the elastic part of the indentation response. Nanoindentation means the radius of curvature of the probe tip is of nanoscale dimension, from several hundred to a few nanometers. Table 1 of Ref. [80] shows the estimated peak shear stresses occurring inside the material during nanoindentation experiments, when serrated, inelastic displacement bursts occur. A feature of the typical nanoindentation experiment is the extremely large stress gradient, where very high stresses are induced in local material volumes but the stress falls off rapidly as one leaves the indented spot. In a typical nanoindentation experiment, even if the material itself has no intrinsic microstructural lengthscale, for instance a perfect crystal, a lengthscale is introduced in the measurement which is the mechanical probe size. The so-called strain-gradient plasticity [125–127] theory has been developed to rationalize the dramatic increase in measured strength with decreasing probe size. In this case, the microstructural lengthscale is related to the buildup of plastic strain gradient and geometrically necessary dislocations [128]. Although supposedly of a different origin from the uniformly tensile loaded whiskers in Brenner’s experiments [96,97,99,108,129], both kinds of measurements tend to satisfy the trend: “smaller is stronger”.

Another significant experimental development is that fully dense, bulk nanostructured materials such as nanocrystals [130–132] and nanotwinned materials [133–135] are now routinely synthesized, and their macroscopic flow strengths measured by traditional means are also exceedingly high compared to materials with conventional microstructures. Similarly, thin films [136], nanostructured multilayers [137,138] and coatings [139] can sustain exceptionally high stresses as well. The Hall–Petch power-law relation

\[ \sigma_{\text{thresh}}(d) = \sigma_{\text{thresh}}^0 + kd^{-\alpha} \]  

(1)

where \( d \) is a characteristic lengthscale, and \( \sigma_{\text{thresh}}^0 \), \( k \) and \( \alpha \) are \( d \)-independent positive constants, was originally used to correlate the measured yield strength [140] or cleavage strength [141] of a bulk metal with the grain size. The power-law (1) is however a convenient fitting form for size effect and now used to numerically represent the “smaller is stronger” trend for a variety of measured strengths in materials, be \( d \) the mechanical probe size [124], the exterior sample size [142,77] or thin film thickness [136], or an interior size scale like the grain size [130–132], twin lamella thickness [133–135] or multi-layer spacing [137,138].

Ultra-strength can even be achieved in bulk materials without pre-existing nanostructures. For instance, titanium-based Gum metals with initial grain size \( \sim 10^5 \) \( \mu \)m are demonstrated to have an elastic strain limit of \( \sim 1\% \) under uniaxial tension [143]. This anomaly has been attributed to a singularity in the crystal’s elastic compliance as a function of electron filling in the alloy [60]. Certainly, many bulk materials can sustain exceedingly high pressures indefinitely without phase transformation [144]. The uniform hydrostatic pressure is applied experimentally by using, for instance, a diamond anvil cell and pressure transmitting media such as silicone oil. The effects of high pressure on the physical properties of materials are well recorded [24,26,83,84].

In 2004 Uchic et al. published an experimental investigation in which micro-pillars were sculpted out from a substrate using focused ion beam (FIB) [145,146], and the mechanical properties of the pillars were tested using indenter with a flat punch head [147]. The benefits of this approach are that small-volume samples can be extracted out of almost any material, the sample geometry and stress conditions are relatively simple (smaller strain gradients), and the sample size can be continuously tuned, to study strength–size correlation. A large variety of metallic and ceramic materials have been tested this way. In particular, when the pillar diameter is reduced to below a few hundred nanometers, the pillar becomes electron transparent, which enables in situ transmission electron microscopy (TEM) observations, concurrent with load–displacement measurements [116,117,77]. The direct observation
of internal microstructural evolution takes much guesswork out of interpreting the load–displacement curve. Also, in situ X-ray and electron backscattering diffraction [148] can be used to track lattice elastic strain, lattice rotation and dislocation density evolution inside the pillars. These advances in experimental diagnostics [149,142], along with atomistic and mesoscale modeling of defect evolutions [150–153], are leading to deeper insights into the fundamental characters of stress relaxation in small-volume materials, and spatio-temporal and statistical features such as intermittency [154,155] and localization [156].

Because FIB processing introduces surface damage, there have been attempts recently at obtaining FIB-less samples [157–160] that have low (or zero) initial mobile defect density. The measured strengths of these small-volume components can indeed be quite sensitive to the fabrication method, because the existence of a single Frank–Read type mobile dislocation source [161] may matter a great deal to the subsequent mechanical response of small components. Also, uniaxial compression tests tend to introduce complications due to bending and friction, so instrumented uniaxial tensile test protocols at micro- and nano-scales are currently under active development [159,160,162,163].

Still going down in characteristic size scale, the mechanical testing of metallic and semiconducting carbon nanotubes [164–169], nanowires [170–173], nanoparticles [174,122], etc. have revealed that these nanoscale components can indeed sustain exceedingly large deviatoric or tensile elastic strains indefinitely. See Table 1 of Ref. [2] for a list of measured strengths, normalized by the elastic modulus. It seems that the “ceiling” of strength, the ideal strength, are indeed being approached in many of these material components when the characteristic lengthscale reaches tens of nanometers. The same conclusion that ideal strength is being approached experimentally was drawn by Volkert and coworkers [175] in rationalizing the measured strength of nanoporous Au, when individual metal ligaments were tens of nanometers in diameter.

An extreme limit of nanoscaling is monolayer graphene which is only one atomic layer thin [176], but can extend hundreds of microns in-plane. Using AFM, Lee et al. performed indentation on free-standing graphene suspended on top of a micron-sized hole [177]. From the load–displacement response, they extracted the non-linear elastic response of graphene, and inferred a biaxial tensile elastic strain limit of 25% and stress limit of $\frac{\sigma}{E}\approx 0.25$, where $E$ is Young’s modulus. The measured result is very close to the ideal strength calculated ab initio from quantum mechanics [3].

Similar to stress, temperature is an important controlling factor of inelastic deformation. Although one expects the majority of ESE applications to be near room temperature, such as the present Strained Silicon technology [29–33,23,34–39], there are many possible scenarios where high temperature capability would be desirable. Currently, although exquisite force and displacement controls have been achieved with experimental instrumentation, the control of temperature is not as advanced. For instance, so far there have only been a few systematic attempts at temperature-controlled nanoindentation [178–182]. Surface diffusion and other non-conservative processes may be actively involved in stress relaxation at very small probe sizes [183] or elevated temperatures [184]. By applying Joule heating using a scanning probe inside TEM, Huang et al. have demonstrated large plastic elongations (>100%) of nominally rigid single-wall carbon nanotubes [185], attributed to vacancy diffusion and dislocation climb [186–188] in graphene. High temperatures can cause extensive reconstruction and annealing of the nanostructures [189,190], which upon cooling to room temperature may be suitable for ultra-strength applications.

Like temperature, the environment (chemical environment, irradiation, etc.) may also greatly influence the mechanical properties of material components. It was found, for instance, that ZnO nanobelts stiffen significantly upon exposure to photon illumination [191]. Controlled studies of these environmental effects on strength, as well as their coupling with physical and chemical properties, are important for assessing the service life of ultra-strength components.

4. General theoretical considerations

The upper limit of material’s strength is the ideal strength, defined as the highest stress a perfect crystal can sustain without undergoing immediate structural transformation at temperature 0 K. The
ideal tensile and shear strength can be estimated as 1/10 of the elastic modulus, as discussed in Section 2. The specific value of ideal strength depends on material’s composition and all the components of applied strain. They can be quantified by using the \textit{ab initio} quantum mechanical calculations, which give typical values in the range of 1–100 GPa. So we consider the ultra-strength behavior as a GPa-level phenomenon, for typical hard inorganic materials.

For bulk crystals, the experimentally measured strengths are far below the theoretical values. The low strength arises because there are usually a large number of grown-in defects (e.g., dislocations or cracks) in the sample. In well-annealed crystals, the dislocation density is initially low. But it increases rapidly once deformation starts, due to, for example, double cross slip and other Frank–Read type multiplication processes [192]. The yield strength of crystals is then limited by the resistance to dislocation motion, typically a factor of 100 below the ideal strength or even more in coarse-grained samples.

Recent years have seen rapid growth of experimental studies on ultra-strength material systems and phenomena. This is largely enabled by the refinement of nanoscale mechanical experiments which allows one to study the near-ideal-strength behavior quantitatively in a controlled fashion, as reviewed in Section 3. Generally speaking, the attainment of ultra strength can result from both the lengthscale and timescale effects.

4.1. Lengthscale effect

The strength of materials with pre-existing defects can be increased by reducing systems’ characteristic lengthscales. The strength-controlling lengthscale can be the sample size, grain size, layer thickness or contact loading zone size. In crystals, the strengthening mechanism is typically based on the blockage or elimination of dislocations. Whereas the well-known Hall–Petch strengthening effect [140,141] in polycrystalline metals arises because of dislocation blockage, the ultra-strength in nanoscale pillars is a representative example of strengthening by dislocation elimination [193]. Various detailed mechanisms of size-strengthening will be reviewed in Section 5. Parenthetically, in amorphous materials such as metallic glass, the room for strength improvement is limited, because the yield strength of macroscopic monolithic glasses are already in the ultra-strength range, of around 2% elastic strain at room temperature [194–196]. A more significant effect of reducing the lengthscale is to suppress the shear band instability, thereby increasing the tensile ductility, as shown in metallic glass samples with dimensions of the order of 100 nm [197] and hybrid crystalline–amorphous nanolaminates [138,198].

4.2. Timescale effect

Ultra-strength phenomena are closely related to the observation timescale, as manifested by the dependence of strength on loading rate and temperature. To understand the strength, temperature, and timescale relations, one needs to know the controlling deformation mechanism. Then the kinetic rate theory can be used to link time and strength. Generally, increasing temperature can lower the strength because of thermal fluctuations in assisting the activation of the stress–relaxation mechanism. Since thermal activation requires a waiting period, the rate mechanism is naturally associated with a characteristic time and the strength becomes rate dependent [199]. Fundamental concepts in the kinetic rate theory of strength, with emphasis on the activation volume as a kinetic signature of deformation mechanism, are reviewed next.

A unit process of strength-controlling structural transformation can be characterized by two types of quantitative measures: the athermal (ideal) strength and activation parameters. The former measures the zero-temperature elastic limit, at which the structural transformation occurs instantaneously without the aid of thermal fluctuations. The latter, including the activation energy and activation volume, characterizes the probabilistic nature of transformation when the applied load is below the athermal limit. Specifically, at a given temperature $T$ and applied shear stress $\tau$, the transformation rate is

$$v = N_0 \exp \left( - \frac{Q(\tau, T)}{k_B T} \right).$$

(2)
Here $v_0$ is the physical trial frequency, $N$ is the number of equivalent sites of transformation, $k_B T$ is the thermal energy, and $Q$ is the activation free energy whose magnitude is controlled by the local shear stress $\tau$ along with the temperature. The athermal (ideal) strength corresponds to the critical stress, giving zero activation energy.

To develop a quantitative sense of Eq. (2), it is instructive to consider some numbers. The physical trial frequency $v_0$ is typically on the order of $10^{11}/s$ as dictated by the atomic vibration. For the transformation observable in typical laboratory experiments, the rate $v$ should be of the order of say $10^{-3}/s$, so that the activation energy needs to be around $30 k_B T$ [200]. For example, a process with a barrier of 0.7 eV would be relevant to laboratory experiments at room temperature ($k_B T \approx 1/40$ eV). For competing processes with the same activation energy, one can determine the operating mechanism by finding the activation volume, equivalently, the rate sensitivity. Specifically, the sensitivity of the transformation rate to stress can be measured by the true activation volume $\Omega$ defined as

$$\Omega(\tau, T) \equiv -\frac{\partial Q}{\partial \tau} |_T \approx k_B T \frac{\partial \ln(v)}{\partial \tau}. \quad (3)$$

Physically, the activation volume is proportional to the number of atoms involved in a thermally activated process, such that it measures the individualistic and collective nature of transition. During thermal activation, the stress does work on the activation volume to assist the transition by reducing the effective energy barrier. Different rate processes may have drastically different characteristic activation volumes, e.g., $\Omega \approx 0.1 b^2$ for lattice diffusion versus $\Omega \approx 1000 b^2$ for dislocation Orowan looping in coarse-grained metals. As a result, activation volume can serve as an effective kinetic signature of deformation mechanism. Note that Eq. (3) can be generalized to a definition of the activation volume tensor when all the stress components are considered [200,80]. Albeit its broad implications, the concept of tensorial activation volume will be only discussed briefly in the context of diffusive deformation mechanisms in Section 6. In this review, we focus on the simple scalar form of $\Omega$ for highlighting its physical characteristics and usefulness. Importantly, the activation volume can be determined by both experiment and atomistic modeling, thus providing a unique link in coupling the two approaches for revealing the rate-controlling deformation mechanisms.

In experiments, the activation volume can be determined by measuring the strain rate sensitivity. Consider, as an example, uniaxial tension of a polycrystalline specimen. The empirical power-law relation between stress $\sigma$ and strain rate $\dot{e}$ is

$$\frac{\sigma}{\sigma_0} = \left( \frac{\dot{e}}{\dot{e}_0} \right)^m, \quad (4)$$

where $\sigma_0$ is the reference stress, $\dot{e}_0$ is the reference strain rate, $m$ is the non-dimensional rate sensitivity index and generally in between 0 and 1; $m = 0$ gives the rate-independent limit and $m = 1$ corresponds to the linear Newtonian flow. The apparent activation volume $\Omega^*$ is conventionally defined as [201]

$$\Omega^* = \sqrt{3} k_B T \frac{\partial \ln \dot{e}}{\partial \sigma}. \quad (5)$$

Combining Eqs. (4) and (5), one can readily show that $m$ is related to $\Omega^*$ by

$$m = \sqrt{3} \frac{k_B T}{\sigma_0 \Omega^*}. \quad (6)$$

In Eqs. (5) and (6), the factor of $\sqrt{3}$ arises because the von Mises yield criterion is invoked to convert the normal stress $\sigma$ to the effective shear stress $\tau^*$, i.e., $\tau = \sigma^*/\sqrt{3}$. Since $\tau^*$ is related to the resolved shear stress on a single slip plane $\tau$ by $\tau^* = M/\sqrt{3} \tau$, where $M = 3.1$ is the Taylor factor, it follows that the true activation volume $\Omega$ associated with a unit process and the apparent activation volume $\Omega^*$ measured from a polycrystalline sample are related by

$$\Omega^* = \frac{\sqrt{3}}{M} \Omega. \quad (7)$$
The activation volume and rate sensitivity provide a direct link between experimentally measurable plastic flow characteristics and underlying deformation mechanisms. However, this link can be complicated by such important factors as mobile dislocation density and strain hardening, as discussed in Section 5.6.

In atomistic modeling, the stress dependent activation energy can be calculated by using the reaction pathway sampling technique. In a first approximation of the temperature effect on the activation free energy $Q(s,T)$, one may assume \[ Q(s,T) = Q_0(s) \left(1 - \frac{T}{T_m}\right), \] \[ Q_0(s) = \frac{Q_0(s)}{T_m}, \] \[ Q_0(s) = \frac{\partial Q_0}{\partial s}. \] The prediction of the transformation rate is then reduced to a problem of finding the minimum energy path (MEP) \[203\] of transformation on the zero-$T$ PES. Namely, $Q_0(\tau)$ can be calculated as the energy difference between the saddle point and initial equilibrium state on the MEP, and $v_0$ is vibrational frequency of the normal mode in the MEP reaction coordinate direction at the initial equilibrium state. One can determine the MEP by using, for example, the nudged elastic band (NEB) method \[203,204\] for an efficient exploration of the PES.

In Eq. (8) the ratio of $Q_0(\tau)/T_m$ corresponds to the activation entropy. There is a well-known Meyer–Neldel empirical compensation rule \[205\], which suggests that activation entropy is likely to correlate with $Q_0(\tau)$ by $Q_0(\tau)/T_{MN}$, where $T_{MN}$ denotes the Meyer–Neldel temperature. This rule was obeyed by a extremely wide range of kinetic processes and it has been justified on the basis of multi-phonon combinatorics inside the activation volume \[206\]. The particular choice of $T_m$ as $T_{MN}$ seems to be reasonable in the context of dislocation nucleation. Namely, $Q_0(\tau,T_m) = 0$ when $T = T_m$, meaning that even a small stress should lead to dislocation nucleation immediately at the melting temperature. Eq. (8) has been invoked in the study of surface dislocation nucleation \[152\], as reviewed in Section 5.1.3, where $T_m$ is replaced by a surface disordering temperature. Recently, temperature dependent activation energies was directly evaluated by Warner and Curtin \[207\] in their study of crack-tip dislocation nucleation.

5. Deformation mechanisms: shear and fracture

The line defect of dislocations, through self-similar translations in crystals, offers a highly efficient means of producing plastic shear strain. In Sections 5.1–5.6, the dislocation-based deformation mechanisms in ultra-strength materials are reviewed. Under high stresses, the non-dislocation shear, such as deformation twinning and martensitic (shear) transformation, can also play an important role in plastic deformation. This will be addressed in Section 5.7. As a competing mechanism to shear, fracture can occur in ultra-strength materials, as discussed in Section 5.8. In these sections, we focus on the mechanisms in crystalline materials, while noting that deformation of amorphous solids often operates at very high stresses; the deformation mechanisms in metallic glasses have recently been reviewed by Schuh et al. \[195\].

5.1. Dislocation nucleation

Ultra-strength can be achieved in a small volume of defect-free, perfect crystal. At low temperatures, the strength of a perfect crystal is limited by dislocation nucleation in the bulk or from the surface, depending on the sample geometry and loading mode. Homogeneous dislocation nucleation in a bulk perfect crystal has been observed in experiments of atomic model systems of bulb raft \[88\] and colloidal crystal \[91–93\], as well as in atomistic simulations \[89,124\].

5.1.1. Athermal nucleation and influence of the stress state

Athermal dislocation nucleation in a perfect crystal is caused by the elastic instability of the lattice under large shear stress; equivalently, it occurs when the energy barrier of nucleation vanishes. It has
been common to use the critical resolved shear stress (CRSS) as the athermal threshold of instantaneous nucleation or motion of a dislocation; for example, the well-known Peierls stress refers to the critical load of motion of a pre-existing dislocation [208]. However, caution should be taken for the case of homogeneous dislocation nucleation because the corresponding CRSS depends sensitively on stress components other than shear [209,210]. In general, the single-valued CRSS should be only considered as an order-of-magnitude estimate.

One can invoke the non-linear theory of hyperelastic instability to precisely determine the athermal threshold of homogeneous nucleation [211]. Such kind of nucleation analysis can account for the influences of different stress components, crystal anisotropy, and elastic softening at large deformation. An atomistics-based hyperelastic instability criterion, termed \( \lambda \)-criterion, has been developed that balances the elastically softening tangent modulus and the GPa-level high stress at large shear [89]. Using this criterion, the critical nucleation conditions at zero Kelvin have been predicted, including the nucleation site, critical stress, and activated slip system [90,210,212–214]. Fig. 1 shows, as an example, predictions of dislocation nucleation for single crystal Cu subjected to nanoindentation. The influence of crystallography on nucleation is evidently shown in the figure; different values of CRSS and effects of elastic softening for each indentation orientation are detailed in [213]. Such criterion can be generalized to crystals with multi-lattices [214–216] and applied for predictions based on quantum mechanical calculations [212,217].

It should be emphasized that the \( \lambda \)-criterion depends on the local measures of stress and tangent elastic modulus. As a result, it is only applicable when the variation of deformation is not too large at the atomic scale. Recent studies have questioned the number of atoms that should be included for the local instability criterion [218,219]. These raise a general question on the size of the material element.

---

**Fig. 1.** Homogeneous dislocation nucleation during nanoindentations on the (111), (110), and (100) surfaces of single crystal Cu [210,213]. (a) Schematic illustration of predicted nucleation sites (stars). (b) Simulations of nanoindentation using the interatomic potential finite element method (IPFEM); circles indicate the nucleation sites visible on the surface of the finite element mesh. (c) Molecular dynamics simulations showing nucleation of dislocations; atoms are colored by the coordination number and perfectly coordinated atoms are removed for clarity.
below which such type of criterion is invalid. Clearly, for the case of dislocation nucleation from an atomically sharp crack, atomic bonds near the crack tip are at every stage of elongation and rupture, so that both the atomic-level stress and strain vary drastically at the scale of the lattice. Under these circumstances, the local instability criterion is expected to break down. An accurate analysis should account for the non-local effect of nucleation on system’s energy, e.g., Rice [220] proposed an energy-based nucleation criterion that balances the elastic energy release rate of the system and the energy penalty associated with formation of a dislocation embryo as characterized by unstable stacking energy.

Free surface can play an important role in dislocation nucleation [221–223]. Bei et al. [158] measured different shear strengths in the nanoindentation and uniaxial compression tests of Mo-alloy single crystals, and obtained $\sim 1/8 \mu$ versus $\sim 1/26 \mu$. They attributed this strength difference to the higher critical shear stress required to nucleate a full dislocation loop in the bulk as opposed to a half loop at a surface. Theoretical analysis of surface or edge dislocation nucleation is complicated by the presence of surface stress [222,224] or edge stress in 2D graphene [225]. This issue arises due to the bonding environment of surface atoms different from that of the bulk atoms. The athermal nucleation from the surface can, in principle, be modeled within the hyperelasticity framework by modifying the $\lambda$-criterion with an appropriate description of the surface stress effect [226].

5.1.2. Thermally activated nucleation

Increasing temperature can lower the strength because of thermal fluctuations in assisting dislocation nucleation. The effect of temperature on nucleation can be studied theoretically at three levels: the continuum analysis based on the Volterra model of dislocation line [192,65], the Peierls concept of a periodic stress–displacement relation across the slip plane [227–229], and direct atomistic simulation [204]. The first two approaches, as discussed in this subsection, provide a useful basis, at the scaling law level, for understanding the nucleation stress as a function of geometrical and material parameters. However, the non-linear and atomic-sized nature of the nucleation process warrants the atomistic analysis for quantitative predictions, as detailed in Section 5.1.3.

It is instructive to first consider the analysis by the Volterra dislocation line model for thermally activated dislocation nucleation in a perfect crystal [192,92]. Two competing effects control the nucleation process: an increase in the elastic energy associated with the formation of a dislocation loop, and a decrease in the potential energy due to the work done by the local shear stress, $\tau$, acting on the area of the loop. For the nucleation in a bulk crystal, the elastic energy of a loop of radius $r$ is approximately given by $A(\mu b^2 \tau)/2 \ln(r/r_0)$, where $r_0$ is the radius of the dislocation core, $b$ is the Burgers vector magnitude, and $A$ is a constant of order one. The mechanical work is $-\pi r^2 \tau b$. The activation energy $Q$ corresponds to the maximum of the sum of the two energies, as given by

$$\frac{Q}{\mu b^2} = \frac{A}{4} \frac{r_c}{b} \ln \frac{r_c}{r_0},$$

where the critical radius of the loop is

$$r_c = \frac{A}{4\pi \mu} \frac{\tau b}{\ln \left(\frac{r_c}{r_0}\right)}.$$  \hspace{1cm} (10)

In order for the nucleation observable in laboratory experiments (e.g., at room temperature $T = 300$ K and on the seconds-to-minutes time scale), the activation energy $Q$ needs to be reduced to around $0.7eV \approx 30k_b T$. It follows that the shear stress $\tau$ is estimated as $\mu/10$ from Eqs. (9) and (10). The corresponding $r_c$ is a few $b$s at most, which is on the size scale of dislocation core. Clearly, a quantitative analysis of such an atomic-sized loop, which is most relevant to the ultra-strength phenomena in typical laboratory experiments, requires the non-linear model to explicitly account for the dislocation core effect [230]. This is beyond the scope of the Volterra model.

Xu and Argon [231] studied the homogeneous dislocation nucleation in a perfect crystal under shear based on a non-linear cohesive law that describes the stress–displacement relation across the slip plane. By using a variational boundary integral method, they determined the saddle-point structure and associated activation energy of a dislocation loop. Numerical results showed that when the
applied shear stress, $\tau$, is at 50% of the athermal nucleation stress, $\tau_0$, the activation energies are still at the level of ten to a few tens of eVs for metals such as Au, Cu, Al, and a covalent solid of Si. These large values reinforce the notion that the homogeneous dislocation nucleation requires ultra-high stresses close to the ideal strength. Similar conclusions were drawn for the surface nucleation of a half loop [65,232,233].

5.1.3. Temperature and strain-rate dependence of nucleation stress

Direct molecular dynamics (MD) simulations have been performed to explore the temperature and strain-rate dependence of dislocation nucleation [234,235,223,236–241]. To overcome the limitations of exceedingly high stresses and strain rates in MD, the statistical models have been developed that integrate direct MD simulations, transition-state theoretical analysis, and reaction pathway calculation [242,152]. Consider, as an example, surface nucleation in a Cu nanopillar under a constant applied strain rate [152]. Because of the probabilistic nature of thermally activated nucleation processes, the nucleation stress has a distribution even if identical samples are used. The most probable nucleation stress is defined by the peak of the frequency distribution of nucleation events. Specifically, the statistical distribution of nucleation events is the product of an nucleation rate that increases in time and a likelihood of pillar survival without nucleation that decreases with time. These two competing effects

![Diagram](image-url)

**Fig. 2.** Temperature and strain rate dependence of surface dislocation nucleation [152]. (a) Nucleation of a partial dislocation from the side surface of a single crystal Cu nanowire under uniaxial compression. (b) Under a constant strain rate, a peak of nucleation events arises because of the two competing effects: the increasing nucleation rate with time (stress) and decreasing survival probability without nucleation. (c) Nucleation stress as a function of temperature and strain rate from predictions (solid lines) and direct molecular dynamics simulations (circles). (d) Illustration of the surface effect on the rate-controlling process and the size dependence of yield strength in micro- and nano-pillars of diameter $d$ under compression.
lead to a maximum at a specific time (stress) as illustrated in Fig. 2b. The nucleation stress, therefore, represents the most likely moment of nucleation under a particular loading rate. It is not a constant.

Based on the above nucleation statistics-based definition, Zhu et al. [152] developed a non-linear theory of the most probable nucleation stress as a function of temperature and strain rate. Here, the non-linearity arises primarily because of the non-linear stress dependence of activation energy, which can be numerically determined using the reaction pathway calculations. A key result from their analysis is that the activation volume associated with surface dislocation source is in a characteristic range of $1-10b^3$, much lower than that of bulk dislocation processes, $100-1000b^3$. The physical effect of such small activation volume can be clearly seen from a simplified linear version of the theory, giving an analytic formula of the nucleation stress

$$\sigma = \sigma_a - \frac{k_B T}{\Omega} \ln \frac{k_B T N_0}{E \varepsilon_0 \Omega},$$

(11)

where $\sigma_a$ is the athermal stress of instantaneous nucleation and $E$ is Young's modulus. Notice that the nucleation stress $\sigma$ has a temperature scaling of $T \ln T$, and the activation volume $\Omega$ appears outside the logarithm, such that a small $\Omega$ associated with a surface source should lead to sensitive temperature and strain-rate dependence of nucleation stress, as quantitatively shown in atomistic simulations, see Fig. 2c. In nano-sized volumes, surface dislocation nucleation is expected to dominate, as supported by recent experiment [158]. As schematically shown in Fig. 2d, the strength mediated by surface nucleation should provide an upper bound to the size-strength relation in nanopillar compression experiments. This upper bound is strain-rate sensitive because of the small activation volume of surface nucleation at ultra-high stresses.

5.2. Dislocation exhaustion

What if mobile dislocations pre-exist in small-volume materials? Under such conditions, ultra strength can be achieved by dislocation exhaustion. This has been experimentally measured in crystals of very small dimensions where the few dislocations present cannot multiply sufficiently before they are annihilated at free surfaces. Greer et al. [243,193] measured uniaxial compression strength of Au pillars fabricated using focused ion beam (FIB). As shown in Fig. 3, when the pillar diameter is reduced to 300 nm, the compressive strength is increased to about 800 MPa, a value approximately

![Fig. 3. Experiment of uniaxial compression of (001)-oriented gold pillars, showing the size strengthening effect [193]. (a) Stress-strain curves with different pillar diameters D. (b) SEM image of the compressed pillar after deformation. Slip lines in multiple orientations indicate a homogeneous shape change.](image-url)
50 times higher than for bulk Au. Volkert and Lilleodden [149] have reported similar high strength for single crystal Au pillars in diameter about 300 nm. In addition, Volkert et al. [149] have investigated nanoporous Au by uniaxial compression. Using the scaling laws for foams, they estimated the yield strength of the 15 nm diameter ligaments to be about 1.5 GPa. All of these experimental values are extremely high for Au, close to the theoretical predictions of ideal shear strength from density functional theory (DFT) calculations, i.e., 850 MPa–1.4 GPa, depending on the loading mode [58].

Greer and Nix [193] have explained the ultra-strength in Au nanopillars based on the concept of dislocation starvation [244]. Unlike bulk samples, dislocations in nanopillars can travel only very small distances before annihilating at free surfaces, thereby reducing the overall dislocation multiplication rate. Gliding dislocations leave the crystal more rapidly than they multiply, decreasing the overall dislocation density. Such processes would lead to a dislocation-starved state requiring very high stresses to nucleate new mobile dislocations.

To support the hardening mechanism by dislocation starvation, Greer et al. [243] have compared the stress–strain behavior of Au pillars with that of Cu whiskers in micron diameters [96]. These whiskers initially contained a small number of defects (dislocations). The initial elastic loading leads to a very high yield stresses followed by huge strain softening and flow at very small stresses. Unlike the whiskers, the stress–strain curve of the gold nanopillars is composed of elastic loading segments separated by discrete displacement bursts throughout the deformation without a significant reduction in the flow stress. This major difference between the whisker-like and the nanopillar behavior, that the flow stress does not fall to a low value following the strain burst, strongly suggests that the pillars may become dislocation starved just after each strain burst.

Shan et al. performed in situ nanocompression experiments of submicrometer nickel pillars inside a transmission electron microscope (TEM) [116]. They directly observed the deformation of pillar structures and correlate the measured stress values with discrete plastic events. The experiments show that the pillar structures contain a high density of initial defects after processing but can be made disloca-

![Image](image_url)

**Fig. 4.** Sample-size effects on the yield behavior of nanocrystalline Ni [249]. (a) FIB-machined pillar array. (b) Compression stress–strain curves for the two sets of pillars tested: 24 samples at a diameter of 160 ± 30 nm (red) and 15 samples at 272 ± 30 nm (blue). (c) Yield strength (Y1) versus pillar diameter (D) showing power-law behavior in compression. (d) A median-rank (MR) Weibull statistical analysis for fitting the Weibull modulus m.
tion free by applying purely mechanical stress. This phenomenon, termed mechanical annealing, leads to clear evidence of source limited deformation where the hardening occurs through the progressive exhaustion of dislocation sources, causing dislocation starvation. Recently, Oh et al. reported the in situ TEM tensile tests of a submicrometer aluminum single crystal [117]. They observed that single-ended sources emit dislocations that escape the crystal before being able to multiply. They also found that the deformation is strain-rate sensitive. This was shown by a sudden increase in strain rate that causes a noticeable surge in dislocation density as the nucleation rate outweighs the loss rate.

The dislocation starvation model was developed to interpret the size strengthening effect in nanopillars of face-centered cubic (FCC) crystals. Recent experiments showed that the size effects on the flow stress in nanopillars of the body-centered cubic (BCC) crystal such as Mo is less pronounced [245,246]. This difference has been attributed to the relative low mobility of screw dislocations in BCC metals [153], which is related to the high Peierls barriers and non-planar core structures [230]. Experiments of different BCC nanopillars by Schneider et al. [247] showed that the relationship between yield strength and pillar diameter correlated with the critical temperature $T_c$ when screw and edge dislocations have equal mobility. The results indicated that for temperature close to $T_c$, where the influence of the low mobility of screw dislocations becomes negligible, the behavior of BCC approaches that of FCC metals.

5.3. Sample-size effects on yield strength

Size effects on the yield strength of small crystals can be caused by the statistical nature of dislocation sources [248]. In nanometer-sized volumes the number of potential dislocation sources is severely limited, such that the yield events evolving from a single source become measurable, as manifested by the displacement burst in a load-controlled test or load drop in a displacement-controlled test. The statistical distribution of source strengths can lead to the fluctuation of yield strength for a fixed sample size. More interestingly, it will also cause the sample size dependence of yield strength, defined as the mean stress at the onset of the first “pop-in”. Here we review a study of sample size

![Fig. 5. Strength-controlling dislocation mechanisms in multilayers operative at different lengthscales [137].](image-url)
effects in the yield behavior of nanocrystalline Ni pillars based on the statistical analysis of pop-in events [249].

Consider metallic pillars under compression. If the first dislocation that is nucleated from a weakest-link source moves out of the crystal or into a grain boundary and the geometric configurations of the source remains unaltered, one can expect continued nucleation from this source. Such continued nucleations from the same source manifest in the stress–strain response as a displacement burst. As long as this situation prevails and no other sources are activated, the weakest link scenario should be a good statistical model of yield behavior. According to the Weibull statistics of the weakest link, the yield probability is

$$F(\sigma; L^d) = 1 - \exp \left[-\left(\frac{L}{L_0}\right)^d \left(\frac{\sigma}{\sigma_0}\right)^m\right],$$  \hspace{1cm} (12)

where $L$ is a characteristic sample lengthscale, $d (=2 \text{ or } 3)$ is the dimensionality number, $\sigma_0$ is the reference stress and $m$ is the Weibull modulus. Using Eq. (12), Rinaldi et al. [249] showed that the yield stress $\sigma_Y$ should satisfy the scaling law

$$\sigma_Y \sim L^{-d/m}.$$  \hspace{1cm} (13)

In addition to the sample size effect, a statistical evaluation of $d$ can further reveal whether surface ($d = 2$) or volume ($d = 3$) defects control the observed yield behavior in pillars. Specifically, for a given sample size $L$, one can determine the Weibull modulus $m$ by the statistical analysis of the yield strengths based on Eq. (12). From the yield stresses of pillars with different sizes $d$, one can fit the power-law exponents $\beta \equiv -d/m$ using Eq. (13), so that the dimensionality number $d$ can be determined.

Fig. 4 shows the experimental measurements of plastic yielding in nanocrystalline Ni pillars by Rinaldi et al. [249]. The yield strength versus pillar diameter is plotted in Fig. 4c, and the Weibull statistical analysis of strength variation is given in Fig. 4d. Both the power-law exponent $\beta$ and Weibull modulus $m$ were then determined. The results show that surface defects ($d \approx 2$) are most likely responsible for plastic yielding in nanocrystalline Ni pillars. Evidently, both the sample size effects and nature of underlying dislocation sources are revealed based on the weakest link model and Weibull statistics. These results indicate that while extreme value statistics have been conventionally applied for brittle fracture [67], they could play an important and previously unrecognized role in the stochastic, jerky plastic flow at small scales [242,181,248,152].

![Ultra-high strength with high ductility in nano-twinned Cu (nt-Cu)](image)

**Fig. 6.** Ultra-high strength with high ductility in nano-twinned Cu (nt-Cu) [254]. (a) Uniaxial stress–strain curve of nt-Cu, as compared with nanocrystalline (nc) and coarse-grained counterparts. (b) TEM image of twin lamellae. (c) High resolution TEM image of a deformed sample showing the accumulation of dislocations at twin boundaries that are nearly dislocation free before deformation.
5.4. Dislocation–interface interaction

In thin films and layered materials, the strengthening arises from the constraints of surrounding substrate and layers [250–253]. Achieving ultra-strength requires that the film and layer thickness be in the range of a few to a few tens of nanometers. Under such conditions, the dislocation–interface interaction is expected to play an important role in controlling the strength.

The effect of layer thickness on flow strength has been summarized by Misra et al. [137] for metallic multilayers, see Fig. 5. They showed that a Hall–Petch type model based on dislocation pile-up is applicable at the sub-micrometer length scales. At the few to a few tens of nanometers length scales, confined layer slip of single dislocations is the operative mechanism; the strength is influenced by interface stress and interface dislocation arrays. As layer thicknesses are further reduced to a few nanometers or less, the strength reaches a peak. This peak strength is set by the interface resistance to single dislocation transmission. For example, a strength peak has been clearly shown in a nanolayered system of ultra-fine crystalline Cu with nanoscale growth twins when the average twin thickness is around 15 nm [133].

The dislocation–interface interactions may not only give rise to ultra strength but also high ductility [134]. For example, Lu et al. [254] have reported experiments of an unusual combination of ultra-high yield strength (≈1 GPa) and high ductility (14% elongation to failure) in nano-twinned copper, i.e., ultrafine-grained Cu (with grain size of several hundred nanometers) containing layered growth twins, typically tens of nanometers in thickness, see Fig. 6. While the ultrahigh strength can be attributed to the size-strengthening effect of thin twin lamellae [255], the mechanistic connection between retained high ductility and twinned nanostructures is not fully understood.

In bulk nanocrystalline metals the tensile ductility is known to be limited by the necking instability. The associated strain localization in the neck region induce microscopic damage that rapidly progresses to final failure. The onset of necking has been extensively studied at the continuum level.

![Fig. 7. Atomistic reaction pathway modeling of twin boundary-mediated slip transfer reactions [204,271]. (a) Schematics of dislocation–interface reactions based on double Thompson tetrahedra, showing an example of absorption of a lattice dislocation into a coherent twin boundary. (b) Atomic configurations from the nudged elastic band method calculations; capturing of the 3D cross-slip process of absorption (mid-image) enables the predictions of activation energy, activation volume, and yield stress consistent with experimental measurements.](image-url)
The central result is that the high strain rate sensitivity and high rate of strain hardening can both help delay the onset of necking, thereby promoting the tensile ductility. To connect the phenomenological theories with nanostructure-mediated plastic deformation, one needs to first determine the rate-controlling mechanisms. The experiment by Lu et al. [258] showed that the nano-twinned Cu increases the rate sensitivity ($m \approx 0.02$) by up to an order of magnitude relative to microcrystalline metals with grain size in the micrometer range, and a concomitant decrease in the activation volume by two orders of magnitude (e.g., down to $\Omega \approx 20b^3$ when twin lamellae are approximately 20 nm thick).

Zhu et al. [204] have studied the mechanistic origin of decreased activation volume and increased rate sensitivity in the nano-twinned system. They analyzed slip transfer reactions between lattice dislocation and coherent twin boundary (TB), involving dislocation absorption and desorption into the TB and slip transmission across the TB. The dislocation–interface reactions have been studied by experiments and molecular dynamics simulations [66,135,259–270]. To overcome the well-known timescale limitation of molecular dynamics, Zhu et al. used the free-end nudged elastic band method to determine the minimum energy paths of reactions, such that predictions of yield stress, activation volume and rate sensitivity can be directly compared with experimental measurements. The modeling predictions were consistent with experiments, thereby showing that the slip transfer reactions are the rate-controlling mechanisms in nano-twinned Cu (see Fig. 7).

In slip-transfer reactions, small activation volume arises because only a small group of atoms are involved during the cross-slip process of dislocation-TB reactions. This is in contrast to the large activation volume associated with the controlling mechanism of forest dislocation cutting in coarse-grained crystals. In the latter case, the activation volume associated with dislocation bow-out can be estimated by $lb^2$ [272], where $l$ denotes the length of dislocation line held between two pinning points and the number of atoms involved in the activation process scales as $l/b$. For a forest cutting process with a characteristic dislocation density $\rho$ of order $100/\mu\text{m}^2$, i.e., $l \sim 1/\sqrt{\rho} \sim 0.1 \mu\text{m}$, a large number of atoms are involved, leading to a big activation volume of about $100 \sim 1000b^3$ [273]. In contrast, the small activation volume in nano-twinned Cu can be attributed to the fact that the sample is subjected to the ultra-high stress at the level of GPa, so that the stress work ($\sigma\Omega$) associated with a small activation volume can lower the effective energy barrier to a level sufficient to initiate the dislocation–interface reaction within the timescale (seconds to minutes) of laboratory experiments.

Besides the rate sensitivity, the dislocation–TB interaction can also cause an increase in the strain-hardening rate, which may play a key role in tensile ductility. The modeling study by Zhu et al. [204] indicated that in contrast to the general high angle grain boundaries in nanocrystals, which have a limited capacity of hardening, the coherent TB is quite hardenable as it gradually loses coherence during the initial stage of plastic deformation. This effective hardenability facilitates the accommodation of incompatible deformation between adjacent twins; it raises the resistance of plastic flow, promotes strain hardening, and helps retain tensile ductility. Also, the coherent TB is more resistant to tensile decohesion compared to random boundaries [274–277] and, therefore, can store a larger amount of deformation-induced incompatibility (interfacial dislocation content) before fracture occurs.

5.5. Intragrain versus intergrain mechanisms

Nanocrystalline metals and alloys represent an important class of ultra-strength materials. They are single or multi-phase polycrystals with nanoscale grain size (1–100 nm). Several recent reviews [130–132,278–281] have addressed the deformation mechanisms in nanocrystalline materials from various experimental and modeling perspectives. Of particular note is that because of the increasing volume of grain boundary materials with decreasing grain size, the deformation in nanocrystalline materials can be critically influenced by the grain boundary mediated mechanisms [282–286], which compete with the intragrain mechanisms such as intersection of bulk dislocations. The competition between the intergrain and intragrain mechanisms on the overall mechanical response has been studied for coarse-grained materials [256,287,288], most of which focused on the regime of the low stress and high homologous temperature for understanding the influences of diffusion flow.

Two theoretical models on the transition from the intragrain to intergrain mechanism [289,290] are reviewed next. These models reveal the mechanistic underpinnings of grain size-dependent...
strength and activation volume. They not only can be applied to interpret experiments and atomistic simulations, but also provide a mechanistic basis of developing the constitutive models, so that 3D numerical simulations can be performed towards a quantitative understanding of the mechanical behavior in nanocrystalline materials [132,291–295].

5.5.1. The strongest size

The ultra-strength phenomenon in nanocrystalline materials involves an interesting possibility of the strongest grain size [296,297]. Specifically, the well-known Hall–Petch effect refers to the rise of plastic resistance with decreasing grain size [140,141]. However, if the grain size drops into the nanometer range, a peak flow resistance could result at a grain size of around 20 nm [296], leading to the breaking down of the Hall–Petch effect. Such a nanostructural size effect has been observed by experiments, e.g., in nano-twinned copper where a strength peak occurs at the twin thickness of approximately 15 nm [133], and in computer simulations as well [298,299,135].

Argon and Yip [289] developed an iso-stress model of the strongest size by focusing on the competition between grain boundary shear and near-boundary dislocation mechanism. Parenthetically, based on molecular dynamics simulations at high temperatures, it has been suggested that the plastic deformation may involve the Coble creep type of diffusional flow [300]. Argon and Yip argued that at low temperatures diffusional flow is unlikely to dominate because of the extremely high levels of stress where shear mechanisms almost certainly overtake processes of diffusional transport.

The Argon–Yip model [289] considers the shear of high angle grain boundaries as a flow process in an amorphous metal, constrained into a disordered layer of thickness \( \delta \approx b \). An empirical power-law was assumed for the shear strain rate versus stress

\[
\dot{\gamma}_{GB} = \gamma^T v_{GB} \left( \frac{\sigma}{\tau_{GB}} \right)^{1/m_1}, \tag{14}
\]

where \( \gamma^T \) is the characteristic shear transformation strain (\( \gamma^T = 0.015 \)), \( v_{GB} \) is the frequency of atomic vibration, and \( \tau_{GB} \) is the ideal shear strength of high angle grain boundary materials (\( \tau_{GB} = \mu/30 \)). It is important to realize that the grain boundary shear operates close to its athermal threshold, so that the rate sensitivity should be taken an increased value (\( m_1 = 1/30 \)). The shear strain rate due to dislocation plasticity is given similarly by

\[
\dot{\gamma}_{DL} = \dot{\gamma}_{DL}^0 \left( \frac{\sigma}{\tau_{DL}} \right)^{1/m_2}, \tag{15}
\]

where \( \dot{\gamma}_{DL}^0 \) is the characteristic dislocation velocity, \( \tau_{DL} \) is the shear stress required to move a dislocation, and \( m_2 \) is the rate sensitivity parameter.

---

**Fig. 8.** The strongest-size behavior for polycrystalline Cu from the Argon–Yip model [289] (solid line) and molecular dynamics simulations (symbols) by Schiotz and Jacobsen [299], where \( \sigma \) is the von Mises effective shear stress and \( d \) is the grain size.
\[ \dot{\gamma}_D = \left( \frac{b}{d} \right) v_D \left( \frac{\tau_D}{\sigma} \right)^{1/m_2}, \] (15)

where \( b/d \) is the shear strain increment caused by the released dislocation traversing over the grain of diameter \( d \). The frequency factor \( v_D \) was taken to be in the range of \( 10^{-3} \) \( v_{GB} \), proper for dislocation intersections. The dislocation mechanism was also considered to operate near its athermal threshold, so that the corresponding rate sensitivity was taken an increased value (\( m_2 = m_1 = 1/30 \)). Argon and Yip [289] considered plasticity occurs by thermally assisted release of dislocations from pinning points, namely, the cutting of forest dislocations. The threshold of shear resistance \( \tau_D \) was taken to be grain-size dependent \( 0.25 \mu \sqrt{b/d} \). They estimated \( \tau_D \) by extrapolating the experimental results for coarse-grained polycrystals [301] to nanocrystals, and assumed that strengthening is still caused by the accumulation of geometrically necessary dislocations whose density inversely scales with grain size. In addition, they also analyzed an alternative dislocation mechanism: the rate controlling process is considered as thermally activated emission of dislocation from the triple junction of grain boundaries. This alternative mechanism results in a nearly identical result.

To a first approximation, Argon and Yip assumed that both mechanisms were under the same applied stress but contributed to the overall strain rate \( \dot{\gamma} \) in proportion to their volume fraction, i.e.,

\[ \dot{\gamma} = f \dot{\gamma}_{GB} + (1 - f) \dot{\gamma}_D. \] (16)

where \( f = 6(b/d) \) is the volume fraction of the grain boundary material by taking the boundary thickness \( \delta = b \).

Fig. 8 shows the modeling prediction of the strongest size, which is consistent with molecular dynamics simulations. According to the model, the breakdown of the Hall–Petch effect can be qualitatively understood as follows. At the large grain size, the near boundary dislocation mechanism dominates, and the strength increases with decreasing grain size by following the classic Hall–Petch relation. As the grain size reduces to the nanometer range, grain boundary shear gradually takes over as the dominant mechanism because of the increasing volume fraction of grain boundary materials, i.e., increasing \( f \) with decreasing \( d \). The strength begins to decrease toward the athermal strength of grain boundary shear, which is lower than that of the near-boundary dislocation mechanism. Consequently, the competition between the two rate mechanisms leads to a peak strength. While the Argon–Yip model is only validated by molecular dynamics simulations at high strain rate \( \approx 10^9/s \) [299], it provides a physically transparent framework for understanding the temperature and strain rate effects on the strongest size of nanocrystals. It could also be generalized to study other related phenomena such as plastic strain recovery in nanocrystalline metals that similarly involves the competition between intragrain and intergrain processes [302–304].

![Fig. 9](image-url) Size dependent activation volume measured from the nanotwinned and nanocrystalline metals [271,305]. (a) Activation volume \( \Omega \) versus grain size \( d \) for nanocrystalline Cu (blue symbols) and versus twin thickness \( \kappa \) for nano-twinned Cu (red symbols) [152]; the solid line shows the results from the theoretical model of Eq. (18). (b) TEM images showing the controlled increase of twin densities (from the right to left image) with the same average grain size of 0.5 \( \mu m \).
5.5.2. **Size-dependent activation volume**

In addition to strength, the variation of nanostructural lengthscales can strongly influence the kinetics of plastic flow. This is manifested as the grain/twin size dependent activation volume and strain rate sensitivity, as measured in nanocrystalline and nano-twinned materials [271], see Fig. 9 for example. Such size dependence was interpreted in terms of the transition from the intragrain to intergrain/twin mechanisms with decreasing grain/twin sizes [305].

Conrad [306], Armstrong and Rodriguez [307] developed iso-strain models to explain this size effect. They considered two competing mechanisms that are in parallel to sustain the applied load: one involves the intragrain dislocation mechanism (e.g., intersection of forest dislocations, giving a large, but size-independent activation volume typical of coarse grained polycrystals); the other is the near-boundary mechanism (e.g., punching of a dislocation through a dense bundle of grain boundary dislocations, the emission of dislocations from grain boundaries by impurities or ledges). The size dependence of activation volume is associated with the latter mechanism through a dislocation pile-up model.

In Conrad’s model [306], the total shear stress is taken as the sum of the shear resistances of two mechanisms, \( \tau = \tau_i + \tau_b \), where \( \tau_i \) is the resistance associated with dislocation intersections in the crystal and \( \tau_b \) is the resistance to dislocation motion near the grain boundary. This averaging scheme of the simple stress sum differs from the Argon–Yip model that involves the average of strain rate based on volume fractions. For the near-boundary mechanism, the number of pile-up dislocations can be estimated by

\[
\frac{1}{\Omega} = \frac{1}{\Omega_i} + \frac{M \mu b}{2\pi x \tau_b d} \frac{1}{\Omega_b},
\]

where \( M \approx 3.1 \) is the Taylor factor. Eq. (17) can be recast as

\[
\frac{1}{\Omega} = \frac{1}{\Omega_i} + \frac{M^2 \mu b}{2 \pi k_{HP} d^{1/2}} \frac{1}{\Omega_b},
\]

where the Hall–Petch equation \( \tau_b = k_{HP} d^{-1/2} M^{-1} \) is invoked to relate the grain size to \( \tau_b \). Fig. 9 shows that the experimentally measured activation volumes in both the nanocrystalline and nano-twinned

![Fig. 10. Stress relaxation experiment of nano-twinned Cu [305]. (a) Stress relaxation series for a sample with average twin thickness of 15 nm (nt-15). (b) Mobile dislocation density \( \rho_m \) versus time for samples without twins (ufg-Cu) and with the twin thickness of 35 nm (nt-35) and nt-15, where \( \rho_{m0} \) denotes \( \rho_m \) at time \( t = 0 \).](image-url)
Cu follow the Hall–Petch-type size dependence as given in Eq. (18) [271,305]. This result provides a firmer basis of the notion that the size dependence of activation volume arises from a transition of the rate-controlling deformation mechanism with decreasing nanostructural lengthscale.

The foregoing models assume that grain size is fixed in the course of plastic deformation. In contrast, grain growth has been observed in nanocrystals during plastic deformation at room to liquid-nitrogen temperatures [308,279,110,113]. This grain coarsening at low temperatures is largely driven by mechanical energy, as opposed to the thermally driven grain growth in coarse grained samples at high temperatures. It has been attributed to the high-energy state of grain boundaries [309]. Recent atomistic modeling has revealed the detailed atomic-level mechanisms of shear mediated grain boundary migration [284,81].

5.6. Effects of mobile dislocation density and strain hardening

In the limiting cases of nanopillars and nanotwins, the density of mobile dislocations is negligibly small. Under these conditions, the ultra-strength deformation can be rate-limited by the activation of a single dislocation source, so that the experimentally measured apparent activation parameters is directly related to the true activation parameters associated with unit activation process of a single dislocation. However, care should be taken in making such connections when the microstructure and deformation process are complex [310,311]. This is because the apparent activation volume, which acts as a kinetic signature of deformation mechanism, could be significantly affected by the mobile dislocation density and strain hardening, both of which are dependent on the plastic deformation history. It is, therefore, important to conceptually distinguish the apparent and physical (true) activation volume [201,312,313].

In a stress relaxation experiment of uniaxial tension, a single relaxation transient often exhibits a logarithmic variation of stress with time

\[ \sigma(t) = \sigma(0) - \frac{\sqrt{3} k_B T}{\Omega_a} \ln \left(1 + \frac{t}{t_r}\right), \]  

(19)

where \( \Omega_a \) is the apparent activation volume and \( t_r \) is the relaxation time constant. On the other hand, the so-called physical activation volume \( \Omega \) can be determined by a jump increase of stress (and strain rate) in a stress relaxation experiment

\[ \Omega = \sqrt{3} k_B T \frac{\ln(\dot{\varepsilon}_2/\dot{\varepsilon}_1)}{\Delta \sigma}, \]  

(20)

where \( \dot{\varepsilon}_2 \) and \( \dot{\varepsilon}_1 \) denote the strain rate at the finish of relaxation 1 and at the onset of relaxation 2, respectively, see Fig. 10a. According to the Orowan equation [41], the plastic strain rate is \( \dot{\gamma} = \rho_m v_b \), where \( \rho_m \) is the mobile dislocation density and \( v \) is the dislocation velocity. In a stress relaxation experiment, because of the short time between consecutive relaxation transients, it is reasonable to assume that the variation of mobile dislocation density is negligibly small as the applied stress jumps from the end of relaxation 1 to the beginning of relaxation 2. It follows that the variation of plastic strain rate is dictated by that of dislocation velocity, so that the physical activation volume directly characterizes the sensitivity of dislocation velocity with stress. In contrast, the apparent activation volume, as defined in Eq. (19), additionally incorporates the influence of variation of mobile dislocation density.

The separation between the apparent and physical activation volumes enables the study of dynamic evolution of mobile dislocations during stress relaxation. It can be shown that the mobile dislocation density \( \rho_m \) at time \( t \) and \( \rho_{m0} \) at time \( t = 0 \) are related by [201]

\[ \frac{\rho_m}{\rho_{m0}} = \left(\frac{t_r}{t_r + t}\right)^{\beta/(1+\beta)} \]  

(21)

Here, \( \beta \) is the immobilization parameter defined as \( \beta = (\Omega_a/\Omega)/(1 + K/M) - 1 \), where \( K \) is the strain hardening coefficient and \( M \) the machine stiffness. In a limiting case both the strain hardening and change of mobile dislocation density are negligible, giving \( K = 0 \) and \( \beta = 0 \), respectively. Then the
two activation volumes are exactly the same. In general, the apparent and physical activation volumes are different [305], such that the mobile dislocation density decreases with time. It is interesting to note that the exhaustion rate of mobile dislocations can be strongly influenced by the twin thickness, as shown in Fig. 10b. Such a twin size dependence has been interpreted in terms of the increased strain hardening rate associated with a high density of twin boundaries, as well as the unique role of coherent twin boundaries in preserving mobile dislocations [305].

In summary, we have reviewed in Sections 5.1–5.6 the lengthscale and timescale effects on the dislocation mediated deformation mechanisms in ultra-strength materials. It has been shown that the plastic deformation in small-volume nanomaterials is generally dislocation source limited and nucleation controlled. Large areas of surface and interface can play an important role in controlling the strength and plastic flow kinetics. The ultrahigh stress may significantly reduce the activation volume and increase the rate sensitivity, leading to the sensitive temperature and strain rate dependence of plastic yielding. In small volumes the impact of the statistical nature of dislocation sources becomes increasingly significant, as manifested by the stochastic jerky plastic flow and the sample size effect on the yield stress in the absence of strain gradient. Most of recent research has been focused on FCC metals and preliminary studies have been performed for BCC [158,245–247] and hexagonal close-packed (HCP) materials [77]. Further research is necessary for better understanding the ultra-strength behavior in different material systems, including alloys and ceramics. Besides strength, the ductility, fracture toughness and fatigue resistance need to be more systematically measured and characterized. The competition between dislocation shear with other displacive and diffusive mechanisms also needs to be investigated; some preliminary studies will be reviewed in succeeding sections. Because of the ultrahigh stress, boundaries of the deformation mechanism map [314] in the space of stress, temperature, and strain-rate is expected to shift significantly compared to those of coarse-grained materials, thereby allowing exploration of crossover behavior that can provide new understanding of the relaxation and hardening mechanisms of ultra-strength materials.

5.7. Twinning and martensitic transformation

The non-dislocation shear, such as deformation twinning and martensitic transformation, can play an important role in the plastic deformation of ultra-strength materials. It is important to conceptually distinguish between dislocation slip, deformation twinning, and martensitic transformation. Cottrell [315] has classified the shear processes in crystals according to the relation of the lattice in the sheared region to the coherent surrounding lattice as follows:

1. **Slip.** In this case the lattice generated in the sheared region is identical in all aspects with the surrounding lattice. This is the common mode of dislocation movement.
2. **Deformation twinning.** Here the lattice generated in the sheared region is the same as the parent lattice but is oriented in twin relation to it, i.e., the lattices of the twinned and untwinned parts of the crystal are mirror images of each other by reflection in some simple crystal plane.
3. **Martensitic (shear) transformation.** They are similar to deformation twinning, but differ in that a new phase (lattice) is generated in the sheared region. The best known example of a shear transformation occurs in the decomposition of FCC austenite to body-centered tetragonal (BCT) martensite in steel; the name of martensitic transformation is often used in place of shear transformation. Martensitic transformation is also commonly observed in shape memory alloys [316].

5.7.1. Deformation twinning versus slip

In ultra-strength materials, the formation of deformation twins can dominate over dislocation slip. This has been observed by the TEM observations of plastic deformation in nanocrystalline metals [132,317]. Significant twinning behavior is also observed during straining of Au thin films [318]. In a plastic shear process of an FCC metal, a leading partial is first formed. The subsequent step may involve the formation of a twinning partial parallel to the leading partial on an adjacent slip plane; the repetition of such a process leads to emission of a deformation twin. Alternatively, a
trailing partial can form in the same slip plane as the first leading partial and thus complete the emission of a perfect dislocation.

Two mechanistic models on the competition between dislocation slip and twinning in nanocrystals are reviewed next. One is the thermodynamic model to account for the energy balance associated with emission of a perfect dislocation or a deformation twin from a pre-existing boundary dislocation [114,297]. The other is the kinetic barrier-based model for determining the athermal stress of emission of a perfect dislocation or a deformation twin from a source of stress concentration at the boundary [290]. The two models differ qualitatively by giving different dependences of critical emission stress on grain size $d$, i.e., $1/d$ versus $1/\sqrt{d}$.

Consider the emission of dislocations versus twins from a grain boundary. The competition between the two processes can be understood by comparing the critical resolved shear stress (CRSS) needed to emit a perfect dislocation, $\tau_f$, with that required to nucleate a Shockley partial, i.e., a stacking fault, which is also the first step of deformation twinning, $\tau_p$ [114]. These CRSSs can be estimated by considering the energy balance between the elastic energy increase in expanding a boundary dislocation loop and the potential energy decrease associated with the work done by the shear stress. Assuming the source size is approximately equal to the grain size $d$, the CRSSs for the emission of a perfect dislocation and twinning partials are given, respectively, by

$$\tau_f = \frac{b}{\mu} \frac{d}{a}$$

(22)

and

$$\tau_p = \frac{b_p}{\mu} \frac{d}{a} + \Gamma_{sf}.$$ 

(23)

Here, $\Gamma_{sf} = \gamma_{sf} / (\mu b_p)$ is the normalized stacking fault energy, $b_p$ is the magnitude of the Burgers vector of a partial dislocation. Equating Eqs. (22) and (23) gives

$$d_c = \frac{b - b_p}{\Gamma_{sf}}.$$ 

(24)

At the critical grain size $d_c$, the transition will occur from the perfect dislocation dominated plastic response for large grains to the twinning and stacking faults controlled plastic behavior for small grains. The estimated $d_c$ is about 10 nm for aluminum [114].

It should be emphasized that the above analyses of slip versus twinning are based on the thermodynamic energy balance of expansion of existing boundary dislocations. As a result, the CRSS only involves the material energetic parameter of stack fault energy, $\gamma_{sf}$, which characterizes the metastable stacking state of the lattice. Van Swygenhoven et al. [319] have emphasized the importance of examining the generalized stacking fault energy curve, which additionally provides activation related parameters such as unstable stacking energy $\gamma_{us}$ and unstable twinning energy $\gamma_{ut}$.

Asaro and Suresh [290] analyzed the athermal emission of a perfect dislocation and a deformation twin from a source of stress concentration at the boundary. They envisioned that the sliding of the grain boundary facet acts as a small crack, leading to the emission of dislocations or twins from the grain boundary triple junction. Based on the analysis of Rice [220] for crack-tip dislocation emission, they derived the athermal stress of nucleation of a leading partial from a triple junction

$$\tau_{lp} = \frac{4b_p}{\pi(1-v)d} \Gamma_{us},$$ 

(25)

where $\Gamma_{us} = \gamma_{us} / (\mu b_p)$ is the normalized unstable stacking energy. Comparing the critical shear stresses given by Eqs. (23) and (25), one sees a qualitative difference in the grain size dependence of nucleation stress. This is because Eq. (23) was derived from the thermodynamic energy balance of expansion of existing dislocations, whereas Eq. (25) was obtained by considering the athermal nucleation with vanishing kinetic energy barrier.
The athermal shear stress for the nucleation of a trailing partial was derived along the same line

\[
\frac{\tau_{tO}}{\mu} = \sqrt{\frac{4b_p}{\pi(1-v)d} \left( \Gamma_{sD} + (4 - 3v)(\Gamma_{us} - \Gamma_{gf}) \right)}. \tag{26}
\]

Here, a simplified loading condition was considered, i.e., the resolved shear stress is applied in the
direction of the leading partial. Eq. (26) can be understood by noting that the leading partial dislocation
has the effect of (1) modifying the local stress near the triple junction due to the leading partial
and (2) resetting the energy of the unslipped state from zero to \( \gamma_{gf} \) so that the peak energy to be
surmounted for the athermal instability of nucleation is reduced from \( \gamma_{us} \) \[220\]. The terms in curly
brackets of Eq. (26) represent the above two effects, respectively. The critical nucleation stress for
other loading conditions can be derived along the same line.

In attempting to compare the tendency for twinning partials and trailing partials, a parameter \( T \),
termed the twinning tendency, was introduced \[320\]. It is defined as

\[
T = \lambda_{crit} \sqrt{\frac{\Gamma_{us}}{\Gamma_{ut}}}, \tag{27}
\]

where \( \Gamma_{ut} = \gamma_{ut}(\mu b_p) \) is the normalized unstable twinning energy. When \( T > 1 \), a nanotwin will nucleate,
and a dislocation when \( T < 1 \). In Eq. (27), \( \lambda_{crit} \) characterizes the additional load necessary to nucleate the trailing partial relative to the leading partial, and \( \Gamma_{us}/\Gamma_{ut} \), the ratio of the energy barriers for
dislocation emission and deformation twin.

The above models address the energy, stress, and grain size dependence of deformation twinning. It is
equally important to resolve the atomic-scale kinematics of twinning processes. As commented by
Cottrell and Bilby \[321\] about half a century ago, “the principal difficulty in developing a dislocation
theory of mechanical twinning has been to explain how a twin grows through a finite thickness of
crystal”. Such a difficulty certainly persists in ultra-strength materials. Furthermore, it is not clear
whether the classical pole mechanism of twinning \[321\] or its variants \[66\] will still dominate in
nanomaterials characterized with large areas of free surface or interface. Various atomic-level twinning
mechanisms have been proposed based on high resolution TEM observations and molecular
dynamics simulations \[317,322,323\]. TEM imaging is useful to reveal the twinning product. However,
it is generally difficult to capture the 3D dynamic process of twinning that proceeds very fast. On the
other hand, MD simulations are limited to exceedingly high strain rates and stresses, such that the
twining mechanisms identified in MD may not be relevant to laboratory experiments. It is, therefore,
important to bridge the experiment and modeling for elucidating the twinning mechanisms. For
example, a recent study coupling the in situ high-resolution TEM experiment and atomistic simulations
has successfully revealed the detailed reversible twinning process at the crack tip of pure Al
\[324\]. Finally, It is noteworthy that for twinning in small volumes with limited sources, the thermal
activation could play an important role; this has not yet been studied to understand the temperature
and strain rate dependence of nucleation stresses of deformation twins at small lengthscales.

5.7.2. Martensitic transformation

Martensitic transformation refers to the diffusionless solid-to-solid phase transition of the crystal
lattice in response to changes in temperature and applied stresses \[68\]. Such transformation is
commonly observed in shape memory alloys (SMA) \[316\]. The research of nanoscale SMAs is rapidly
growing in recent years \[325\]. This development provides opportunities of revealing the fundamental
mechanisms of martensitic transformation, and may offer practical solutions of efficient shape memory
\[325–327\], actuation \[328\], and mechanical damping \[329\] in microscale and nanoscale devices.

Ye et al. \[330\] reported in situ TEM observation of martensitic phase transformation in NiTi nano-
pillars. By recording the diffraction patterns in real time, they have obtained evidence that the stress-
induced B2 to B19’ (austenite to martensite) transformation exists even in NiTi pillars with diameter
below 200 nm. In their experiments, the forward transformation starts at a high stress level of about
1 GPa. This value is higher than what is reported for bulk NiTi, but comparable to the measurement of
sub-micron compression pillars \[331\]. The increased transformation stress with decreasing sample
size was also measured in Cu–Al–Ni single crystal pillars with diameter at the submicrometer scale [329]. This sample size effect was attributed to a low number of nucleation sites for martensite in nanoscale volumes, similar to the hardening by dislocation starvation in FCC metallic nanopillars, as reviewed in Section 5.2.

Bulk SMAs, in general, feature the pseudoelastic effect, meaning that the net deformation is effectively zero after a load–unload cycle at constant temperature. In addition, a large stress–strain hysteresis often accompanies the pseudoelastic response. For NiTi sub-micron pillars, Frick et al. [331] showed that decreased diameter leads to a smaller stress–strain hysteresis, with a full loss of recoverable pseudoelastic strain for samples with a diameter smaller than 200 nm. In contrast, San Juan et al. [329] measured a larger hysteresis loop in Cu–Al–Ni single crystal pillars at the submicrometer scale, as opposed to bulk counterparts. It should be emphasized that the substrate deformation could significantly influence the force–displacement measurement of small-sized pillars, as shown by the in situ TEM experiment of compression of NiTi nanopillars [330]. Therefore, care needs to be taken when interpreting the reversibility and hysteresis responses of SMA pillars.

Unique martensitic phases have been observed in nanostructured SMAs [325]. Twins of the martensite as narrow as 1 nm (four atomic layers) are made possible by a variant combination of very low specific twin boundary energy. For example, thin [001] compound twins have been found in nanocrystalline NiTi [332], whereas this type of twins is not typically observed in coarse-grained NiTi. In nanocrystals, twin-related variants of the martensite can span the entire nano-sized crystal, so that the elastic deformation of the lattice can be accommodated by grain boundaries instead of habit planes. In contrast, extended martensite and austenite interfaces in coarse-grained SMAs require deformation compatibility, thus favoring the formation of conventional type-I and type-II twins that can better accommodated by the parent B2 phase of austenite.

The experiments reviewed above call for a reconsideration of the classical theories of martensitic transformation [333–336] for their applications to the SMA nanomaterials. These theories were developed with assumptions such as abundant martensite sources and/or the occurrence of twin-related variants to facilitate compatibility at habit planes. These conditions may not hold in nanomaterials. At the most fundamental level, the formation of various martensitic phases and twin variants can be studied by atomistic modeling based on first-principles quantum mechanical or empirical interatomic potential-based calculations. Recent research along this line has been focused on the atomic geometries, phase energies, and transformation pathways of various phases (B2, R, BCO, B19, B19′) in NiTi at \( T = 0 \) K [332,337–344]. On the other hand, continuum models have been developed for understanding the influence of grain size on the martensitic transformation of NiTi nanograins [345]. Integrated atomistic and continuum modeling, coupled with experiments, is needed to provide a mechanistic understanding of the nanometer length scale effects on martensitic transformation.

Dislocations in SMAs received less attention compared to martensites in the past. However, there are growing efforts in characterizing dislocation plasticity, which could play an important role in small-volume SMAs with limited defect sources. Recent TEM experiments have provided the detailed information on dislocation activities in SMAs [346,347]. It has been shown that dislocations form either as the result of the interaction of different plates of martensite or the plastic deformation of martensite. Further research is needed to understand the competition and cooperation between dislocation shear and martensitic transformation, as well as the impact on fracture [348,349] and fatigue [350,351].

5.8. Fracture

Fracture is an important deformation mode in ultra-strength materials. However, compared to plastic shear, there are much less quantitative fracture experiments reported in the literature. This is largely because of the practical difficulties of precisely controlling the fracture testing at the micrometer and nanometer scales [352,353]. Here we focus on a few theoretical concepts of nanoscale fracture. Unresolved issues are discussed at the end.

5.8.1. Brittle versus ductile fracture

Plastic flow and brittle fracture are competing deformation modes in ultra-strength materials. Bulk silicon is brittle, but considerable plastic deformation has been measured for silicon nanowires in
tension [172] and silicon nanopillars in compression [355] at room temperature. A revealing analysis of brittle-ductile competition has been given by Dumitrica et al. [354] for a perfect single-wall carbon nanotube. By molecular dynamics simulations at various temperatures $T$ and at short time scale (about 0.1 ns), they identified two primary mechanisms as shown in Fig. 11a: the “cool” mechanism of brittle fracture at low $T$ and “hot” mechanism of formation of Stone–Wales (SW) defects (i.e., dislocation dipoles) at high $T$. To explore the competition at long time scales, they have studied the energy landscapes governing the two processes. They found that the brittle mode of bond breaking requires little thermal agitation; the fracture stress is controlled by the ideal tensile strength of C–C bond. In contrast, the SW defect formation by single bond rotation sensitively depends on temperature, implying a small activation volume associated with this process. According to transition state theory, the activation energy under constant stress needs to be reduced to

$$Q = k_B T \ln(Nvt),$$

in order to observe the formation of SW dipole over the test duration $t$ (probability $\approx 1$). Dumitrica et al. [354] have further calculated $Q$ as a function of stress. Then, the stress needed to form the SW dislocation dipole can be back out from the $Q$ given by Eq. (28). Fig. 11b shows the predicted breaking-strain map. Such a mechanism map is expected to provide useful guidance for the experimental study of temperature and strain rate dependence of mechanical responses of carbon nanotubes.

The foregoing analysis has considered an idealized situation of competition between the brittle and ductile response in a perfect system. In an alternative scenario, such competition could occur at the crack tip in a system with a pre-existing crack. In this case, bond breaking at the crack tip may have different temperature and strain rate dependences. The analysis along the line of Rice and others [220, 227, 228, 356, 357] needs to be performed to better understand the brittle to ductile transition in ultra-strength materials.

### 5.8.2. Size and loading effects on nanoscale fracture

In the study of brittle fracture at the nanometer scale, two questions often arise: (a) Is the classical theory of Griffith’s fracture [358] is still applicable? (b) What is the influence of the discreteness of atomic lattice? These questions can be directly addressed by the atomistic calculation of nano-sized cracks [359].

Recall that in the Griffith theory of fracture [358, 67], one considers a large body with a central crack of length $2a$, see Fig. 12a for example. Suppose the system is under an average tensile stress $\sigma$. This load can be imposed by either a fixed displacement or a constant force at the far field. Relative to the uncracked body under the same load (e.g., fixed displacement), the elastic energy decrease due
to the presence of a crack of length $2a$ is $\pi \sigma^2 a^2 / E$ per unit thickness, where $E$ is Young's modulus, and the corresponding increase of surface energy is $4\gamma a$, where $\gamma$ is surface energy density. As a result, the total energy change is

$$U(a) = 4\gamma a - \frac{\pi \sigma^2}{E} a^2. \quad (29)$$

The critical crack length of Griffith’s fracture $2a_{cr}$ is defined in terms of the condition when $U(a)$ reaches the maximum, giving

$$a_{cr} = \frac{2\gamma E}{\pi \sigma^2}. \quad (30)$$

Note that the displacement/strain-controlled and force/stress-controlled loadings give the same formula of $U(a)$ and $a_{cr}$.

Fig. 12b shows the atomistic calculations of $U(a)$ for single crystal Si based on the Stillinger–Weber interatomic potential [360]. In the figure, a circle represents the energy of a metastable state with a
nominal crack length $2a$ given by the number of broken bonds times the lattice spacing. The envelope curves connecting circles give $U(a)$ under stress-controlled (red) and strain-controlled (blue) loadings. The Griffith crack length can be determined by the maximum of $U(a)$, giving $2a_{cr} \approx 2.8$ nm. Alternatively, one can evaluate the Griffith crack length based on Eq. (30) using the material constants of $E$ and $\gamma$ calculated from the Stillinger–Weber potential, giving $2a_{cr} \approx 2.74$ nm as indicated by vertical lines in Fig. 12b–d. The agreement between the two methods of predicting the critical crack length (with a difference less than one atomic spacing of 0.33 nm) suggests that the Griffith formula is applicable to the nanoscale fracture.

Energy barriers of crack extension arise because of the lattice discreteness, leading to the so-called lattice trapping effect [361–364]. In Fig. 12b, each spike-like curve linking adjacent circles gives the minimum energy path (MEP) of breaking a single bond at the crack tip, i.e., unit crack extension by one lattice spacing [365,366]. The maximum of each MEP gives the energy barrier of bond breaking. Such MEPs manifest the corrugation of the atomic-scale energy landscape of the system due to the lattice discreteness. As a result, a crack can be locally “trapped” in a series of metastable states with different crack lengths and crack-tip atomic structures. The time-dependent kinetic crack extension [367] then corresponds to the transition of the system from one state to another via thermal activation.

Fig. 12 also demonstrates the effects of system size and loading method on the nanoscale brittle fractures. It is seen from Fig. 12b that when the system size is about 10 times larger than the crack size, the curves of $U(a)$ are close for stress-controlled (red) and strain-controlled (blue) fracture. Comparison of Fig. 12c and d indicates that $U(a)$ for strain-controlled fracture is much more sensitive to the reduction of system size than the stress-controlled fracture. This is because the average stress in a sample under strain control can significantly change with crack length in small systems (i.e., the sample size is less than 10 times the crack length) [359].

5.8.3. Grain boundary engineering

While Sections 5.8.1 and 5.8.2 focus on single crystals, a current challenge is to understand fracture in nanostructured materials such as nanocrystalline metals. The tensile fracture stress of nanocrystals is high, because of the high plastic yield stress associated with the size strengthening effect of nanostructures; but their tensile fracture strain, i.e., ductility, is low compared to the coarse-grained materials [131,368]. As discussed in Section 5.4, this can result from the early onset of necking owing to the low strain hardening rate. Consequently, strain localization in the neck region induces microscopic damages (e.g., cracking, cavitation) that rapidly progress to final failure. In contrast, nano-twinned metals exhibit the delayed onset of necking and improved tensile ductility [254]. These arise due to the increased strain hardening rate attributable to the fact that the coherent twin boundary is more hardenable than random grain boundaries [134,204]. Furthermore, the twin boundary is more fracture resistant than random boundaries, thus further promoting the tensile ductility [277]. These beneficial effects of twin boundary on fracture resistance can be generally correlated to the notion of grain boundary engineering [274–276], namely, to increase the fraction of so-called special grain boundaries with low sigma number ($\Sigma$) defined in terms of the coincident-site lattice (CSL) model [369]. Grain boundary engineering will also decrease the connectivity of the remaining random boundaries that are susceptible to intergranular degradation such as cracking, cavitation, corrosion, and rapid self-diffusion [370–372]. So introducing nanotwins (i.e., a large number of twin boundaries) into ultra-fine grained crystals can be considered as a successful example of grain boundary engineering at the nanoscale. Further research is needed to exploit the grain boundary engineering of nanostructural materials to yield improvements in strength, ductility, chemical corrosion, stress corrosion cracking, high-temperature creep and fatigue resistance [81,134,276,372–375].

We conclude this section by noting that there is a wide range of unresolved issues related to nanoscale fracture. These fracture problems often feature a strong coupling with the chemical environment and composition. For example, while bulk silicon is not susceptible to environmentally induced cracking (i.e., stress-corrosion cracking) and fatigue, experiments showed that the micron-scale silicon, which is being widely in micro-electro-mechanical systems, can failure prematurely by fatigue in air [376,377]. There is compelling experimental evidence suggesting that the formation and growth of nanoscale cracks in surface reaction layers are responsible to those failures [378,379]. As another
example, silicon is being considered for the next generation anode materials of rechargeable lithium batteries. However, applications of silicon in batteries are limited because silicon’s volume changes drastically (by about 400%) upon insertion and extraction of lithium, which results in pulverization and capacity fading [380]. Recent experiments showed that silicon nanowire battery electrodes can mitigate these issues because the small diameter of nanowires allows for better accommodation and relaxation of large volume change without the initiation of fracture [381,382]. To address these fracture problems, understanding the mechanical-electrochemical coupling at the nanoscale is the key [383]. The research in this direction is not only important for preventing fracture and failure, a conventional realm of corrosion science [384], it may also open new avenues of engineering nanostructured materials for novel applications in, e.g., battery [385–387] and hydrogen storage [388,389] through elastically tuning the thermodynamic and kinetic aspects of phase transition. Finally, it is noteworthy that understanding the nanoscale fracture has implications for elucidating the failure mechanisms in biomaterials [390], irradiation-embrittled nuclear materials [391], quasi-brittle engineering structures [392], etc. and for guiding the hierarchical material design as it has been suggested that materials become insensitive to flaws at the nanoscale [393].

6. Diffusion-controlled deformation mechanisms

In Section 5, we reviewed mostly displacive deformation mechanisms, where atom registries change in a more collective and deterministic manner. For example, before and after dislocation glide, all atoms above the slip plane move by the same distance with respect to the bottom half of the crystal, the displacement pattern taken from a finite possibilities of Burgers vectors.

In contrast, when diffusion happens, the atom registries are disrupted in a more random and chaotic manner. Consider three atoms A, B, C which were originally nearest neighbors before diffusion: it is quite conceivable that after diffusion, atom B sits at (10, –3, 5)a0 with respect to atom A, where a0 is the lattice constant, while atom C sits at (–4, 10, –2)a0 with respect to atom A. Such atomic separation patterns are quite difficult to achieve with displacive processes; furthermore the separations (10, –3, 5)a0 and (–4, 10, –2)a0 are very much random and differ from place to place.

Because collective action (i.e., a displacement pattern) is required in displacive processes, the activation volume of thermally activated displacive events tend to be larger than one single atomic volume, sometimes much larger [152,204]. This is because displacive transformation is a “collective decision” of multiple atoms. In contrast, a diffusion event is very much an individual atom’s decision, when seen at the saddle points of thermal activations, so the activation volume of diffusion is on the order of one atomic volume, and sometimes even smaller.

In thinking about displacive and diffusive processes, it is important to make a distinction between atomic sites and atoms. The same lattice site inside a crystal may be occupied by atoms of different species, different atoms of the same species, or even a vacancy, without changing the topological relationship between the lattice sites. This notion of a relatively robust site lattice gives rise to so-called network constraint in the theory of diffusion [394–396]. Indeed, in deformation one is more concerned with what happens to the site lattice, rather than what happens to individual atoms, since one cannot tell (and does not care) which labeled atom sits where anyway. Macroscopic shape change means remodeling of the (occupied) site lattice. And in X-ray or electron diffraction, one measures the average lattice spacing between sites [397], without knowing the explicit labeling scheme of atoms.

Both displacive and diffusive processes can cause shape change and stress relaxation [398]. Diffusive (also called civilian) processes require nearly uncorrelated diffusions of individual atoms through the material – even though as a result the site lattice is remodeled more or less deterministically. Displacive (also called military) processes involve more collective displacement of atoms from their original positions. It is important to recognize that real deformation can be a complex pathway involving both diffusive and displacive events. Thus taken as a whole, the deformation mechanism can have a hybrid diffusive-displacive character. A well-known example of hybrid diffusive-displacive deformation is dislocation creep (see Fig. 13), operative at intermediate temperatures and stresses on the deformation-mechanism map [314]. It has been proposed that in dislocation creep, most of the plastic strain is still generated by dislocation glide, but the creep rate is governed by the unlocking of...
dislocation locks and dislocation climb, controlled by bulk or dislocation core (pipe) diffusion. Thus the effective activation energy of dislocation creep is the same as that for tracer self-diffusion.

The key questions for the ultra-strength community, when diffusion may be involved, are:

(a) Can ultra-strength (other than hydrostatic pressure) be sustained in systems where diffusion can happen?
(b) If not, what is the size-temperature boundary of stress relaxation? The answer is not only important for assessing the domain of elastic strain engineering, but also for forming components by diffusion at high temperatures [185–190] and then using it as ultra-strength component at lower temperatures.

Currently there is no consensus on (a), and many unknowns exist for (b). A main reason is because systematic temperature-controlled nanomechanical testing is still somewhat rare today. However, some theoretical considerations are offered below.

Regarding (a), it is helpful to recall in Eq. (4), \( \dot{\epsilon} \propto \sigma^{1/m} \). The smaller \( m \) is, the more non-linear and “thresholded” the strain rate – stress relation becomes. For example, \( m = 0.1 \) means a reduction of stress by half would cause a thousand-fold decrease in the strain rate: this means a component that fails in 1 day at 2 GPa, would survive three years at 1 GPa, figuratively speaking. A sharply thresholded \( \dot{\epsilon}(\sigma) \) behavior allows one to easily define the concept of “strength”, such that whenever \( \sigma < \sigma_{\text{thresh}} \), the strain rate is practically negligible, and the strain rate is only significant if \( \sigma \geq \sigma_{\text{thresh}} \). When \( m \) is large, as \( m = 1 \) in the case of the Nabarro–Herring creep [69,70] and Coble creep [399], even the concept of “strength” becomes murky. In these cases, a component that fails in 1 day at 2 GPa would survive only 2 days at 1 GPa, figuratively speaking. This is not good for the practical application of elastic strain
engineering, because a service life of thousands of days is expected. In other words, creep is detrimental for elastic strain engineering and should be suppressed. A useful component should probably have $m < 0.1$ in the regime of practical operations.

We see in Eq. (6) that $1/m$ is proportional to the activation volume. Based on the difference between diffusive and displacive events explained at the beginning of this section, one knows that diffusive events tend to be more “individualistic” at the atomic level, seen at the saddle point of each thermal activation event, and tend to have smaller activation volumes. The scalar activation volume for bulk vacancy migration in pure copper is as small as $0.12\Omega_0$, where $\Omega_0$ is the atomic volume [400] (see also elaborations on this below). Therefore, the stress threshold for diffusive deformation tend to be not as sharp as that for displacive deformation. In other words, displacive mechanisms introduced in Section 5 do tend to have stronger non-linearities in stress and sharper stress thresholding behavior. The atomistically computable activation volume is a quantitative measure separating “yield” from “creep” in mechanical deformation phenomena [80].

It is however important to recall that even displacive processes like surface dislocation nucleation [152] can have small true activation volume (as small as $2b^2$) at high stresses, and near pre-existing defects. Also, the effective $m$ can have extraneous controlling factors not directly related to the activation event itself, such as grain size effect in a dislocation pile-up. A microscopic deformation pathway can in fact posses a continuum of diffusive versus displacive characters, as measured by (a) the activation volume of individual activation event, and (b) geometric correlations between different thermally activated, stress-biased events. As J.W. Christian remarked, “the main categories of transformation are called military and civilian, but rigid classifications is not required since soldiers may sometimes be out of step and civilians may sometimes form paramilitary organizations!” (p. 6 of Ref. [68]). The actual situation can become very complex, for instance when (boundary) diffusion and dislocation slip may couple and occur in tandem [405,181–183,406].

Schuh, Mason and Lund performed high-temperature nanoindentation experiments [181,182], and Cross et al. performed small-contact-area ($\sim 10 \text{ nm}^2$) indentations [183] into metal surfaces. Both studies found deformation kinetics that cannot be easily explained by classic models. At these length-scales and temperatures, surface diffusion can be shown to be more important than bulk vacancy diffusion [184]. It is thus conceivable that at initial contact between the indenter and surface, the indenter displacement can be completely accommodated by surface-diffusion mass transport. Once the contact radius increases above a critical value, the imposed displacement rate begins to exceed the ability of mass removal by surface diffusion. Elastic energy starts to be stored inside the material, which will eventually trigger dislocation nucleation. Such dislocation nucleation could be aided by previously stored mass-action defects such as surface vacancies and steps, which means then the dislocation nucleation would be heterogeneous and the process as a whole would have a hybrid diffusive-displacive character [184].

Syed Asif and Pethica [407] performed nanoindentation creep experiments on single-crystal W and GaAs, and Ngan et al. performed creep experiments on Ni$_3$Al single-crystal pillars [406]. Both found that the creep rate greatly accelerated after a large displacement burst (stress-driven) had happened in the sample. A coupling effect in reverse could operate in these cases: Ngan et al. think that the creep response is due to dislocation core diffusion. Before the large displacement burst, the dislocation population density is small in the pillar, and therefore the creep rate is low and cannot match the applied strain rate. So the stress keeps increasing, until large displacive shear occurs in sample which relaxes the stress somewhat. However even under these reduced stresses, due to the increased dislocation core density, creep accelerated. This would be another scenario of coupled diffusive-displacive process, with displacive events taking the lead. The dislocation population is likely to have an inherent lengthscale dependence, which means the onset of significant diffusion could be shifted to lower temperatures than in coarse-grained materials.

Sun et al. recently observed with in situ TEM the extrusion of Au, Pt, W and Mo nanoparticles from shrinking graphitic nanocages [404], at temperatures estimated to be $\sim 1/2$ of the melting temperature. At such high temperatures and small sizes, one expects surface diffusion may dominate the deformation. Yet, based on the large lattice spacing shrinkage (as much as 5%) which implies a large stress in the nanoparticle (could be hydrostatic, though), Sun et al. proposed the shape change of nanoparticle is mostly due to displacive dislocation slip, instead of diffusive mass action. This
proposition was supported by molecular dynamics simulations [404], which however has a strain-rate problem. Traditional deformation-mechanism maps [314] were developed for coarse-grained materials; how the borders of the different deformation-mechanism regimes shift, at these extremely small scales and high stresses, is currently unknown/unvalidated (Fig. 13). This highlights the urgent need to draft out the entire size-temperature deformation-mechanism maps [1], which include the size-dependent melting [401–403] and boundary pre-melting [81], and Hall-Petch scaling for low-temperature displacive plasticity [152] as special limiting cases. There have been important contributions at constructing such maps from molecular dynamics simulations [323,408], but the efforts are still far from complete and validated.

From a modeling point of view, one should be aware that the problem of diffusion in stressed solids is conceptually quite challenging [394–396]. Even regarding the bulk vacancy flux, one may find in the literature different expressions of the chemical potentials, field equation (partial differential equation, PDE) and boundary conditions (BC) as a function of stress. Often, even wrong PDE and BC formulations can give the correct scaling result from dimensional-analysis argument, but are conceptually problematic. Only a careful treatment which separates vacancy source (surface/GB/climbing dislocation) action as a BC relying on the boundary traction, and vacancy transport in crystal as a PDE relying on the internal stress (and the vacancy relaxation volume), can resolve the conceptual problems and reconcile continuum equations with the atomistic view. Again, it is worth repeating that a solid under any externally applied load other than uniform hydrostatic pressure cannot be in global thermodynamic equilibrium. Therefore, the vacancy concentration can at best be in local equilibrium with the nearest patch of surface/GB source or climbing dislocation.

Most existing diffusional theories, such as those used in computing the boundaries of the Nabarro-Herring creep [69,70] and Coble creep [399] in Fig. 13, as well as the formulation for Asaro-Tiller-Grinfeld instability [409–411], are based on the assumptions of small stress and small stress gradient. Because point defect has a formation volume, and its migration involves an activation volume/strain [80], the point-defect concentration, migration barrier and mobilities are in fact coupled to the stress, which may lead to dramatic non-linear effects at high stresses, unaccounted for in existing small-stress models.

For example, consider a vacancy in FCC Cu, which has cubic symmetry when it is stationary (Fig. 14a and c). However, when the vacancy moves, say in the (011) diagonal direction shown in Fig. 14, it will involve an orthorhombic-symmetry saddle state (Fig. 14b), with an anisotropic activation volume/strain tensor [80] for the migration:

$$\Omega_m \equiv -\frac{\partial Q_m(\sigma)}{\partial \sigma} = \begin{bmatrix} 1.2505 & 0 & 0 \\ 0 & -0.5675 & -0.1519 \\ 0 & -0.1519 & -0.5675 \end{bmatrix} \Omega_0, \quad (31)$$

where $\Omega_0$ is the atomic volume. The above is calculated using the Mishin interatomic potential [400], and the minimum energy path is calculated using the nudged elastic band (NEB) method [203]. Upon

Fig. 14. Vacancy migration in FCC Cu. (a and c) The initial and final states, both having cubic symmetry. (b) The saddle point, which has orthorhombic symmetry. (d) The minimum energy path at zero stress. At large stresses, the migration energy barrier is expected to shift significantly, with coupling to deviatoric stress components.
rotation to a $x' = \langle 100 \rangle$, $y' = \langle 011 \rangle$, $z' = \langle 011 \rangle$ coordinate frame, the orthorhombic activation volume tensor is

$$
\Omega_m' = \begin{bmatrix}
1.2505 & 0 & 0 \\
0 & -0.7194 & 0 \\
0 & 0 & -0.4157
\end{bmatrix}, \Omega_0.
$$

(32)

While the aforementioned scalar activation volume $\text{Tr}(\Omega_m) = 0.12\Omega_0$ is quite small in magnitude due to sign cancellations, the component that couples to the $\langle 100 \rangle$ uniaxial stress is of the order $\Omega_0$. Therefore, to leading order, the vacancy $\langle 011 \rangle$-migration barrier is modified by stress as:

$$
Q_m(\sigma) = Q_m^0 - 1.25\Omega_0\sigma_{x'x'} + 0.72\Omega_0\sigma_{y'y'} + 0.42\Omega_0\sigma_{z'z'} + C(\sigma^2),
$$

(33)

where $Q_m^0$, the stress-free vacancy migration barrier, is 0.68 eV. Because $Q_m(\sigma)$ appears inside the exponential, a change of $-0.3$ eV to $Q_m$ would cause a factor of $10^5$ change in vacancy mobility at $T = 300$ K. Another way to say it is that since $Q_m(\sigma)/k_B T$ always appear together, the effective temperature has changed from 300 K to $300 \times 0.68/0.38 = 537$ K if $Q_m$ changes by $-0.3$ eV. Thus, very significant changes in bulk diffusional behavior will occur when the deviatoric stress component reaches the level of $0.3$ eV/$\Omega_0 \sim 4$ GPa, in the ultra-strength regime. The stress can in fact be used to either enhance or suppress diffusion, so the temperature boundaries of diffusive and hybrid diffusive-displacive regimes in Fig. 13 can shift very significantly under a high stress. This is also true for surface and grain boundary diffusion, and pipe diffusion.

The above estimates should not be surprising, because vacancy migration occurs by thermally activated bond breaking, and some bonds are severely pre-weakened in the ultra-strength regime, by the definitions of ideal strength and strain. As we said before, ultra-strength leads to significant changes in the physi-chemical properties of the material, which should include atomic/ionic diffusivities as well. This may be exploited, for instance, to tune the ionic conductivity of electrolyte material used for electrochemical energy conversion.

Furthermore, originally isotropic atomic/ionic diffusivities would likely become highly anisotropic in the ultra-strength regime, with anisotropy ratio as high as $10^5$, because $\Omega_m$ couples to the shear stress. The effect of large stress gradient, especially in small-probe experiments, on surface diffusion may be significant as well. Because of these considerations, the boundaries of the diffusive and hybrid diffusive-displacive deformation mechanisms are expected to shift in complex ways [184]. So our answer to question (b), posed at the beginning, is rather uncertain at this moment. Careful temperature-controlled experiments and advanced atomistic modeling are expected to improve our understanding in the future.

7. Conclusions

Materials science is about understanding and exerting control over the behavior of materials, made of many atoms and electrons. What are the limits of materials science? Chemically, there are only finite selections on the periodic table. Structurally, ideal strength puts a basic limit on atomic organizations. Within such basic limits lies the richness of materials behavior.

The ultra-strength behavior of materials differs naturally from traditional mechanical responses, because when elastic strains exceeding $\sim 1\%$ permeate in a material component, significant changes in physi-chemical properties with respect to the stress-free state are expected to occur, such as carrier mobility, band gap, and thermoelectric figure-of-merit. Taking advantage of this over an extended period is elastic strain engineering. Understanding ultra-strength is not only necessary for forming material components and maintaining their shape stabilities, which are certainly critical for their functions, but also elastic strain engineering of physi-chemical properties.

In the last 20 years, we have seen tremendous advances in materials processing, mechanical instrumentation, symbolized by the AFM, as well as in materials characterization techniques. We have also witnessed the dramatic miniaturization trend, which started in the 1950s, homing in at tens of nanometers. This convergence has brought about an explosion of ultra-strength observations, that has not only made the ultra-strength concept experimental realities, but also an industrial reality such as...
Strained Silicon technology. The much enlarged dynamic range of sustainable stress on atomic organization gives a new meaning to Feynman's statement “there's plenty of room at the bottom”. The practical consequence of this could be significant, because elastic strain, like chemical composition, is continuously tunable field variable. Therefore, just like one understands that β-brass is not pure copper, strained silicon is not normal silicon. Indeed, elastic strain field can be tuned in real time, easier than one could tune the composition field.

From a thermodynamic point of view, all ultra-strength systems are far from equilibrium, because the large stresses at the ultra-strength level can change the chemical potential by $\sim 10^{-2} - 10^{-1}$ eV/atom compared to the stress-relaxed state \[412]\]. Therefore, to take advantage of elastic strain engineering requires defeating various mechanisms of stress relaxation, including plastic flow and fracture. Our article primarily deals with the temperature–time–lengthscale characteristics of these inelastic relaxation processes. Our goal is to show that despite the disparate appearances of the various material systems, certain common threads can be identified for all ultra-strength systems, such as the activation volume and temperature/strain-rate sensitivity. The deformation-mechanism map approach of Frost and Ashby, after taking account of the size dependence, should provide a common language for the different communities that share the ultra-strength features.

Modeling is a critical component of ultra-strength research, not only for studying the basic defect processes and their space–time–temperature characteristics for the design and manufacture of reliable systems, but also for numerically exploring elastic strain engineering. Just like computer modeling is indispensable nowadays for the design of cars and airplanes, we envision in the future, hyperelastic (i.e., non-linear elastic) modeling will be used in the design of ultra-strength components, coupling stress field with charge-density field, and physi-chemical properties such as ionic conductivity, catalytic activity, and band bending. Quantum mechanical calculations of ideal strength and the impact of stress on physi-chemical properties will be important workhorses in ESE. This review motivates such modeling, but does not give sufficient details. Much future efforts are needed in this direction.

Acknowledgements

TZ acknowledges the support by NSF CMMI-0653769, 0758265, 0758554, 0825435, 0835484, and DOE DE-AC07-05ID14517. JL acknowledges the support by NSF CMMI-0728069 and DMR-0520020, ONR N00014-05-1-0504 and AFOSR FA9550-08-1-0325. We thank Professors Sidney Yip and Subra Suresh for their guidance and encouragement.

References


Volkert CA, Lilloedden ET. Size effects in the deformation of sub-micron Au columns. Philos Mag 2006;86:5567–79.


[280] Ma E. Eight routes to improve the tensile ductility of bulk nanostructured metals and alloys. JOM 2006;58:49–53.


