Direct and indirect transitions in the region of the band gap using electron-energy-loss spectroscopy

B. Rafferty and L. M. Brown
Cavendish Laboratory, Cambridge University, Madingley Road, Cambridge, CB3 0HE, United Kingdom
(Received 17 December 1997; revised manuscript received 17 April 1998)

The momentum and energy dependence of the matrix elements for direct and indirect transitions across the band gap is studied both theoretically and experimentally. The accepted theory for the inelastic scattering cross section of fast electrons by condensed matter is extended to show how the nature of a transition can change the shape of the measured energy-loss spectrum in the region of the onset. For the case of direct transitions the matrix element acts only as a multiplicative factor to the *joint density of states* (JDOS), and so an $(E-E_g)^{1/2}$ term is observed in the spectrum. The matrix elements for indirect transitions are shown to be dependent on the momentum transferred by the incident fast electron to the crystal electrons. The product of the indirect matrix element and the JDOS contributes an $(E-E_g)^{3/2}$ term to the spectrum. To test this theory it is shown that the *matrix-element-weighted joint density of states* should be extracted from the electron-energy-loss spectra via a Kramers-Kronig transformation. An objective method is proposed for plotting the data to determine both the principal band gaps and their direct or indirect nature. The method is tested and succeeds in well-known cases. It is now possible with the electron microscope to measure these fundamental electronic properties of semiconductors and insulators accurately by electron spectroscopy.
[S0163-1829(98)06740-X]

I. INTRODUCTION

Since the discovery of the electron,[1] the energy loss of electrons as they traverse matter has been one of the liveliest research fronts this past century. Bohr,[2,3] Bethe,[4] Fermi,[5] and Bohm and Pines[6–8] were among the first to study theoretically the stopping power of electrons by matter, and to show how this is related to the physical and electronic structure of the material. These theoretical ideas are now used every day in the understanding of electron energy loss spectra.

The low loss region ($<$50 eV) contains much information about the excitations of valence electrons. These excitations are manifest as both collective plasma oscillations[6–8] and single electron transitions that depend upon the position of critical points with the band structure of the sample the most important critical points being those which define the band gap itself. To view the band-gap region of semiconducting and insulating materials is an ideal way to study their electronic structure. Although optical techniques have higher energy resolution ($\sim$2 meV as compared to $\sim$300 meV) they are restricted to a spatial resolution of at best 0.2 μm for transmission measurements and only direct transitions across the band gap can be studied. The electron-energy-loss spectroscopy (EELS) system used here is attached to a dedicated VG HB501 scanning transmission electron microscope (STEM) which can focus an electron probe to a diameter of approximately 4 Å and position it anywhere on the sample. Since the incident electrons in the probe have much larger momentum than equivalent photons, both energy and momentum can be transferred to the crystal electrons during an inelastic scattering event, thus allowing both direct and indirect transitions to be observed.

Previous measurements of the bandgap region using EELS (Refs. 9 and 10) showed that the spectra did not have a sharp onset. Although it was recognized that both direct and indirect transitions were involved the result was not correctly explained using the accepted *joint density of states* (JDOS) picture of the Bethe theory. In this paper other terms that are present in the Bethe theory, which were then neglected, will be taken into consideration. It will be shown that the matrix elements defining the transitions can be decoupled for direct and indirect transitions. The contributions from the indirect transitions across the band gap have the effect of smoothing out the onset to a band edge. These effects must also be considered in optical experiments if phonons are used as a source of momentum transfer.

Band-gap EEL spectra from six crystalline samples have been acquired and a numerical procedure for comparing the spectral data with the extended Bethe theory is outlined. This method reveals the sizes of the band gaps and the nature of the transitions across them that are in excellent agreement with other experimental measurements and theoretical predictions.

II. THEORY

If we are to understand how the features in the EEL spectrum are produced, we must consider both the interaction of the incident electron with the sample and also how the geometry of the microscope contributes to the spectrum. Figure 1 is a schematic of the scattering process in the STEM. The incident electron beam is formed by the coherent sum of plane waves whose range of wave vectors is defined by the size of the objective aperture. These waves form a probe on the sample that has a spatial diameter of approximately 4 Å. The momentum that is transferred to the sample ($\hbar\mathbf{q}$) can be split into components that are parallel and perpendicular to the beam direction. The perpendicular component is the larger of the two for most valence energy losses. The mag-

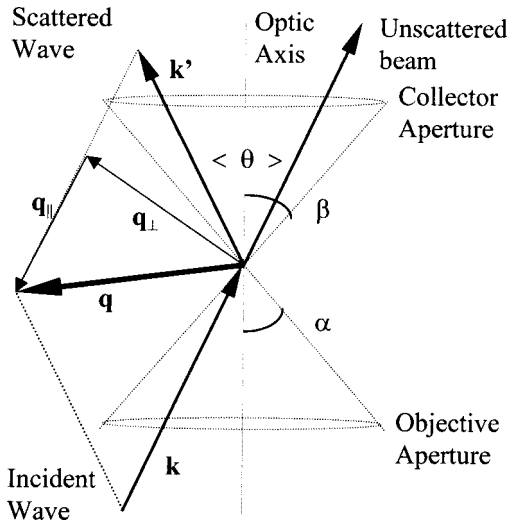


FIG. 1. Schematic diagram for the inelastic scattering of electrons in the STEM. The incident electron beam is produced by the coherent sum of plane waves whose wave vectors are defined by the angular size of the objective aperture. The momentum transferred to the sample, $\hbar\mathbf{q}$, is largely within the plane perpendicular to the direction of the incident beam. The magnitude of the component of $\mathbf{q}$ parallel to the beam depends on both the energy of the incoming beam and also the energy loss associated with that scattering event. The relative magnitude of $\mathbf{q}$ is exaggerated for clarity.

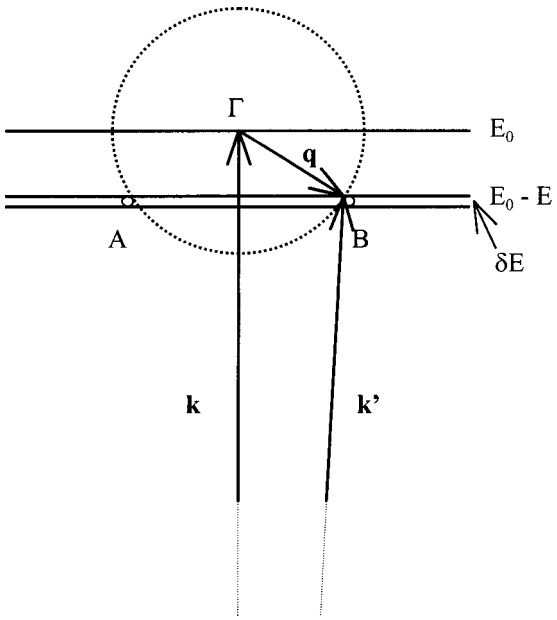


FIG. 2. A schematic diagram showing the ring of intersection (*AB*) between the Ewald sphere of the scattered traversing electron and the constant energy surface of the samples band structure in k space. The Ewald spheres can be considered to be flat on the scale of the scattering vector $\mathbf{q}$ because the wave vectors of the incident electron before and after the collision are much greater than $\mathbf{q}$.

nitude of the parallel component depends on the energy of the incident electron and also the energy loss associated with a particular inelastic scattering event.

There are two main theories that are used to understand and interpret the features observed in an EEL spectrum. The first method is based upon a dielectric formulation where the sample is modeled as a homogeneous dielectric through which the incident electron passes. The second theory, Bethe theory, is quantum mechanical in nature.

In the dielectric formulation, which was first used by Fermi,[5] it is taken that the ensemble of valence electrons in a solid may be characterised by a complex dielectric function $\varepsilon(\mathbf{q},\omega)$, which is a function of both the frequency and the wave vector of the electromagnetic disturbance in the solid. As any electron moves through the solid other ions and electrons are moved by Coulomb repulsion; each fast electron is thus screened. This screening is negligible for insulators, and even in metals the impact parameter for a 1-eV loss is smaller than the dynamic screening length. So the fast electron may be regarded as an effective free particle. This is one of the key assumptions used in the dielectric formulation. The transmitted electron having coordinate $\mathbf{r}$ and velocity $\mathbf{v}$ in the z direction is represented by a point charge $-e\,\delta(\mathbf{r}-\mathbf{v}t)$ which generates within the medium a spatially and time-dependent electrostatic potential satisfying Poisson's equation. Solving this equation leads to the double-differential scattering cross section[11–13]

$$\frac{d^2\sigma}{dE\ d\Omega}=\frac{\mathrm{Im}\left[\dfrac{-1}{\varepsilon(\mathbf{q},\omega)}\right]}{\pi^2 a_0 m_e v^2 n_a}\left(\frac{1}{\theta^2+\theta_E^2}\right), \tag{1}$$

where $\mathrm{Im}[-1/\varepsilon(\mathbf{q},\omega)]$ is known as the *energy-loss function*, a_o is the Bohr radius, m_e is the electron mass, v is the incident electron velocity, n_a is the valence electron density, θ is the scattering angle, and $\theta_E=E/2E_0$, where E is the energy loss and E_0 is the energy of the incident electron.

Bethe theory has been reviewed many times.[14] An electron with velocity $\mathbf{v}$, mass m_e, and charge Ze collides with a stationary mass m_t in an initial state $|i\rangle$, and is deflected into the solid angle element $d\Omega$ along polar coordinates θ and ϕ measured in the center-of-mass frame of reference. Simultaneously, the atom undergoes a transition to a state $|f\rangle$ at excitation energy E, measured from $|i\rangle$. When the fast electron is nonrelativistic, the double-differential cross section can be calculated in the lowest order in the interaction potential V between the particle and the atom (i.e., the Born approximation), to give

$$\frac{d^2\sigma}{d\Omega\ dE}=\frac{1}{\pi\varepsilon_0}\left(\frac{Zm_{\mathrm{red}}e^2}{\hbar^2}\right)^2\left(\frac{k'}{k}\right)\frac{1}{q^4}\times\left|\sum_{j=1}^{N}\langle f|\exp(i\mathbf{q}\cdot\mathbf{r}_j)|i\rangle\right|^2\rho(E), \tag{2}$$

where $Z=-1$ for incident electrons, $\mathbf{r}_j$ is the coordinate of the jth electron, N is the number of electrons in the target atom and m_{red} is the reduced mass of the incident electron and the target atom. The term $\rho(E)$ is the JDOS, which is the convolution of the valence- and conduction-band densities of states. The factor (k'/k) arises from the need to keep the incident current of particles to be equal to the current of scattered particles. Here k and k' refer to the incident electron before and after the collision respectively.[15]

All semiconductors and insulators have a fundamental bandgap that can be probed and studied by a traversing electron. Since nearly all band gaps are of only a few eV the band-gap region of an EEL spectrum covers at most the first

~10 eV of the spectrum. The intensity that is observed within this region is produced by the excitation of electrons from the valence to the conduction band of the material, or possibly to localized states which we do not consider here. The excitations have no collective character and so the nature of the transitions is best described by Bethe theory. Equation (2) should be the most accurate description of the contributions to the band-gap region of the spectrum. To extract the most information from a band-gap EEL spectrum, we must first fully understand the terms in Eq. (2) with reference to the electronic properties of a material. There are three terms to consider. First, there is the joint density of states, which has much direct information about the electronic structure of the sample being studied. Second, we must see whether the q^{-4} factor makes any significant changes to the shape of the spectrum. Finally, we will see how the matrix elements change the detailed shape of the features in the band-gap spectrum.

It has been shown by Bruley and Brown[9] that for parabolic bands the JDOS probed by the electrons is

$$\rho(E)=\frac{V}{(2\pi)^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\alpha\sqrt{E-E_g}, \tag{3}$$

where m^* is the sum of the masses of the holes in the valence band and the electrons in the conduction band, E_g is the value of the band gap, and α is the convergence angle of the incident electron beam. Figure 2 shows a schematic diagram of the scattering process for a direct band-gap material and a cross section of the band structure centered on the Ewald sphere of the unscattered traversing electron in k space. The volume of the ring of intersection (AB) that is produced by the overlap of the Ewald sphere and the sphere of the excited crystal electron measures the number of final states available to the excited electron. This is a direct measure of the density of states. The number of states within this volume is equal to the product of the density of states and an infinitesimal energy element corresponding to some energy resolution. The curvature of the valence band can be incorporated into this simple description by the use of the summed masses of the valence-band holes and conduction-band electrons. This result does not change if the transitions from the valence to conduction band are direct or indirect. Thus the JDOS has a sharp onset at an energy loss of E_g.

When the electron probe passes down the [001] direction in a material such as diamond, transitions across the direct band gap, and also to four pockets along the ΓX directions to the indirect band-gap contributions (Fig. 3), are allowed. In relation to Fig. 2, the indirect contributions are four equally spaced pockets displaced a distance $\mathbf{q}_0$ ($\mathbf{q}_0$ is the momentum separation between the top of the valence band and the bottom of the indirect conduction band) from the center of the Ewald sphere of the incident electron. Due to the symmetry of the crystal, we need to consider only one of these contributions from the indirect bandgap. Rather than just considering $d^3\mathbf{k}$, we now need to examine $d^3\mathbf{k}/q^4$ integrated over θ and ϕ. For indirect transitions, q^2 is given by

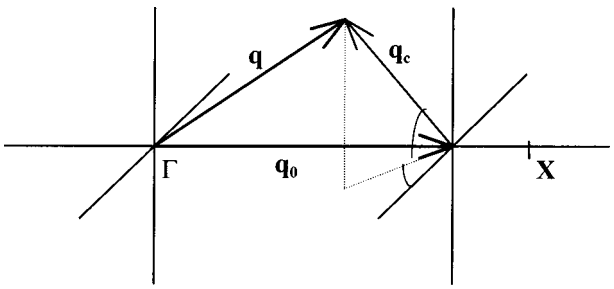


FIG. 3. A schematic diagram showing the geometry of the q^{-4} term for an indirect band-gap excitation.

$$q^2=A+B\cos(\theta)\sin(\phi),$$

$$A=q_c^2+q_0^2,\quad B=-2q_0q_c.$$

We require[15]

$$\int\frac{d^3\mathbf{k}}{q^4}=\int\int\frac{\sin(\theta)q_c^2dq_cd\theta\,d\phi}{[A+B\cos(\theta)\sin(\phi)]^2}.$$

The integral over θ is trivial leaving an integral of the form

$$\int\frac{d^3\mathbf{k}}{q^4}=\int_0^{2\pi}\frac{2q_c^2dq_c^2d\phi}{A^2+B^2\sin^2(\phi)}.$$

This is a third-order elliptical integral $\Pi(\phi,n,k)$, with $k=0$, which can be solved analytically,[16]

$$\Pi(\phi,n,k)=\int\frac{d\phi}{(1+n\sin^2\phi)\sqrt{1-k^2\sin^2\phi}},$$

$$\Pi(\phi.n.0)=\frac{2\pi}{1+n}-\frac{\pi n[\Lambda_0(\beta,0)-1]}{2\sqrt{n^2(1+n)}},$$

$$\Lambda_0(\beta,0)=\frac{2}{\pi}[EF(\beta,1)+KE(\beta,1)-KF(\beta,1)],$$

$$\beta=\sin^{-1}\left[\frac{1}{\sqrt{1+n}}\right].$$

On substituting these expressions into the required integral, we obtain

$$\int\frac{d^3\mathbf{k}}{q^4}=\pi q_cdq_c\left[\frac{5(q_0^2+q_c^2)^2+(q_0^2+q_c^2)\sqrt{(q_0^2+q_c^2)^2+(2q_0q_c)^2}}{(q_0^2+q_c^2)^2+(2q_0q_c)^2}\right],$$

$$q_c=\left(\frac{2m_c}{\hbar^2}\right)^{1/2}\sqrt{E-E_g},$$

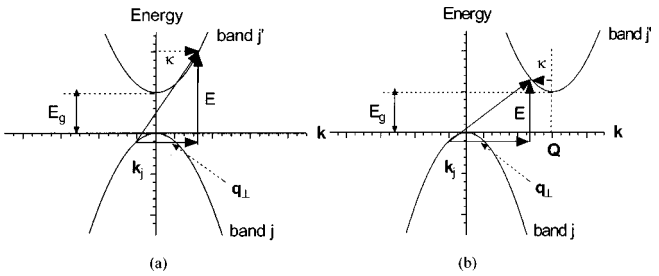


FIG. 4. Schematic diagrams showing the transitions across a (a) direct band gap and (b) an indirect band gap.

where E_g is the indirect band gap and m_c is the summed masses of the valence- and conduction-band hole and electron masses. For transitions just above the onset of the band gap, this expression still has an abrupt onset as before, and so the geometric contributions of indirect band gaps cannot explain the detailed shape of the band-gap EELs spectrum. The term in the square brackets affects the density of states only when q_c becomes itself comparable to q_0 in size, at which point the parabolic band approximation has broken down.

So far we have not asked anything about the energy or momentum dependence of the initial and final states of the atomic electron at position $\mathbf{r}_j$. It is in considering exactly this dependence of the atomic wave functions that produces a more complete understanding of the EELS scattering process. The electrons are described as Bloch electrons, $\phi_{\mathbf{k}_j\sigma}$, where $\mathbf{k}_j$ is the wave number of the Bloch electron at position $\mathbf{r}_j$, $u(\mathbf{r}_j)$ is a function which has the same periodicity of the lattice, and χ_σ is the spin part of the wave function which can be ignored for this discussion. The crystal electrons can also be described as plane waves within this framework by setting the function $u(\mathbf{r}_j)$ to unity. We thus have

$$\phi_{k_j\sigma}=\frac{1}{\sqrt{V}}\exp(i\mathbf{k}_j\cdot\mathbf{r}_j)u_k(\mathbf{r}_j)\chi_\sigma .$$

We can also express the momentum transferred to the jth Bloch electron, which is equal to that lost by the incident electron, $\hbar q$, in operator form:

$$\mathbf{q}\equiv\frac{\hbar}{i}\nabla .$$

Also, since we will only be considering small scattering angles in the electron microscope, we can expand the exponent for small $\mathbf{q}\cdot\mathbf{r}_j$. Substituting these expressions in Eq. (2), and using the orthogonality of the Bloch electrons we find that the matrix element describing the transition from state $|i\rangle$ to state $|f\rangle$ becomes

$$\langle f|\exp(i\mathbf{q}\cdot\mathbf{r}_j)|i\rangle=\frac{\hbar}{i}\langle f|\nabla\mathbf{r}_j|i\rangle,$$

$$\langle f|\nabla\mathbf{r}_j|i\rangle=\int\{u^*_{\mathbf{k}_j+\mathbf{q}}(\mathbf{r}_j)\exp[-i(\mathbf{k}_j+\mathbf{q})\mathbf{r}_j]\nabla\mathbf{r}_j[u_{\mathbf{k}_j}(\mathbf{r}_j)\exp(-i\mathbf{k}_j\cdot\mathbf{r}_j)]\}d\mathbf{r}_j^3$$

$$=\int u^*_{\mathbf{k}_j+\mathbf{q}}(\mathbf{r}_j)\exp(-i\mathbf{q}\cdot\mathbf{r}_j)\nabla[\mathbf{r}_j u_{\mathbf{k}_j}(\mathbf{r}_j)]d\mathbf{r}_j^3+\int u^*_{\mathbf{k}_j+\mathbf{q}}(\mathbf{r}_j)\exp(-i\mathbf{q}\cdot\mathbf{r}_j)i\mathbf{k}_j\cdot\mathbf{r}_j u_{\mathbf{k}_j}(\mathbf{r}_j)d\mathbf{r}_j^3=M_1+i\mathbf{k}_j\cdot\mathbf{M}_2,$$

where

$$M_1=\int u^*_{\mathbf{k}_j+\mathbf{q}}(\mathbf{r}_j)\exp(-i\mathbf{q}\cdot\mathbf{r}_j)\nabla[\mathbf{r}_j u_{\mathbf{k}_j}(\mathbf{r}_j)]d\mathbf{r}_j^3,$$

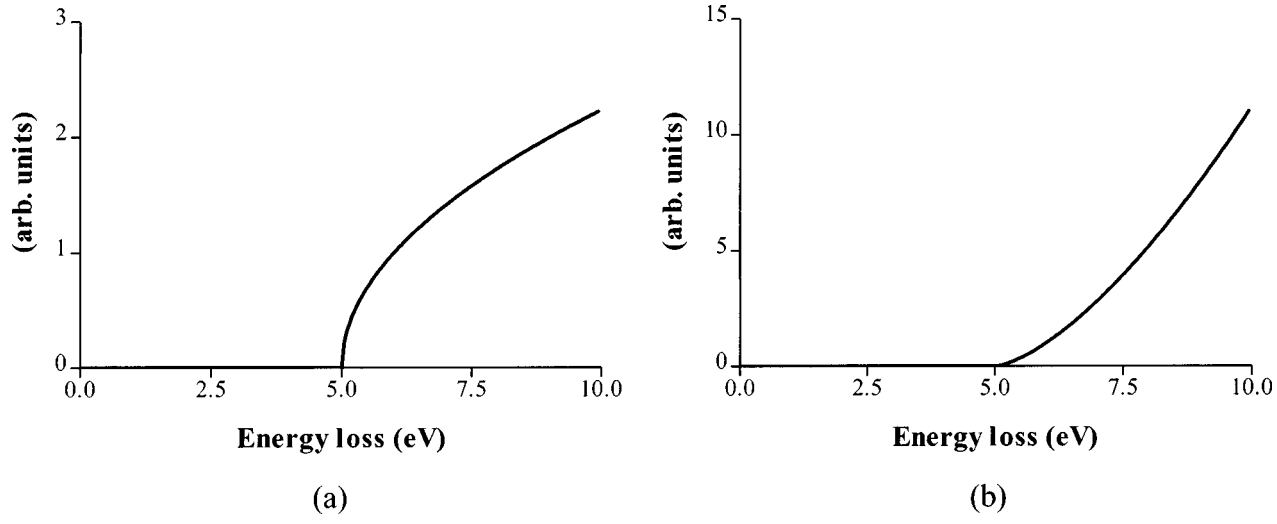


FIG. 5. Schematic diagrams showing the contributions from the JDOS and matrix elements for (a) direct and (b) indirect transitions.

$$\mathbf{M}_2 = \int u^*_{\mathbf{k}_j+\mathbf{q}}(\mathbf{r}_j)\exp(-i\mathbf{q}\cdot\mathbf{r}_j)\mathbf{r}_j u_{\mathbf{k}_j}(\mathbf{r}_j)d\mathbf{r}_j^3. \quad (4)$$

How do we sum these two terms together? In the process of summing and squaring the two terms of the matrix element describing the transition, the mixing term $M_1 \times ik_j \cdot \mathbf{M}_2$ vanishes, since the integrand of M_1 is asymmetric and that of $\mathbf{M}_2$ is symmetric. Therefore, the two terms do not *interfere*; they are summed *incoherently*. Thus the double-differential cross section becomes

$$\frac{d^2\sigma}{dE\ d\Omega} = \frac{\hbar}{\pi\varepsilon_0}\left(\frac{zMe^2}{\hbar^2}\right)^2\left(\frac{k'}{k}\right)\frac{1}{q^4}[M_1^2 + k_j^2 M_2^2]\rho(E). \quad (5)$$

When the incident electron is scattered by the crystal, the momentum which it transfers to the crystal can be split into its components parallel and perpendicular to the incident electrons trajectory. For relatively flat bands, the parallel component of this momentum is used to take the crystal electron from its initial energy band, j (effective mass m) to its final energy band j', (effective mass m'), while the perpendicular component dictates the position of the crystal electron in the final energy band. Thus, using the conservation of energy one can write [Fig. 4(a)]

$$E - E_g = E_n + E_{n'},$$

thus

$$\frac{2(E-E_g)}{\hbar^2} = \frac{\kappa^2}{m'} + \frac{k_j^2}{m}, \quad (6)$$

where $\kappa = \mathbf{k}_j + \mathbf{q}_\perp$ and

$$k_j^2 = \frac{m_{\rm red}(E-E_g)}{\hbar^2} - q_\perp^2\left(\frac{m_{\rm red}}{m'}\right)^2\left(\frac{m'}{m_{\rm red}} - 2\right) \pm \frac{q_\perp m_{\rm red}}{2m'}\left[\frac{2m_{\rm red}(E-E_g)}{\hbar^2} - q_\perp^2\left(\frac{m_{\rm red}}{m'}\right)^2\left(\frac{m'}{m_{\rm red}} - 1\right)\right]^{1/2},$$

where $(1/m_{\rm red}) = (1/m) + (1/m')$ is the reduced mass. The second term in Eq. (6) is zero if the effective masses of bands j and j', are equal and negligibly small for deviations from this.

If we consider a valence and conduction band centered on the same point in the Brillouin zone, then for direct transitions, $\mathbf{q}_\perp = 0$, the main contribution comes from electrons excited from the top of the valence band, $\mathbf{k}_j = 0$, and so the main contribution in the spectral intensity has the energy dependence of the joint density of states, $(E-E_g)^{1/2}$. The contributions from direct transitions with $\mathbf{k}_j \neq 0$ will be negligible since the joint density of states includes appreciable contributions only for excitations between parallel bands. For the contributions from indirect transitions between parallel bands there will also be a small extra term with a spectral intensity proportional to $(E-E_g)^{3/2}$; the second term in the above expression for $\mathbf{k}_j$ will be negligible for these transitions. A similar argument can be used for the case when the valence and conduction bands are not centered on the same point in the Brillouin zone, the only difference being the neglect of the first term in the expression for $\mathbf{k}_j$. We now have [Fig. 4(b)]

$$\kappa = \mathbf{k}_j + \mathbf{q}_\perp - \mathbf{Q},$$

where $\mathbf{Q}$ is the separation of the valence and conduction bands in reciprocal space. Therefore $\mathbf{q}_\perp$ is replaced by $q_\perp - \mathbf{Q}$, which is close to zero for transitions just above the onset of allowed transitions. Thus for the two contributions (Fig. 5) we have leading terms with different energy dependencies corresponding to direct and indirect transitions:

$$I_{\rm direct} \propto (E-E_g)^{1/2}, \quad I_{\rm indirect} \propto (E-E_g)^{3/2}. \quad (7)$$

This result is similar to that of Elliot[17] and Batson.[10] Elliot considered bound excitons produced by optical excitations of semiconductors, and came to the same energy dependence of the optical spectra. Batson interpreted Elliot's result as being an envelope function of the spectrum. The way in which we have presented this theory shows that the second term is a consequence of indirect transitions that also need to be considered in optical spectroscopy when phonon collisions are included as a source of momentum transfer.

III. RESULTS

All of the band-gap EEL spectra in this section have had their complex dielectric function extracted using the Kramers-Kronig transformation as described by Egerton.[13] We emphasize the imaginary part of the dielectric function that is directly related to the single electron scattering and so to the results of the extended Bethe theory and the JDOS. The zero loss peak has been removed using a Fourier-ratio deconvolution with a vacuum zero loss peak. This method was chosen over a scaled subtraction of a vacuum zero loss peak, since large artifacts are left in the spectral data and we find the subtraction method limits the size of observable band gaps to >2 eV (Fig. 6). All of the spectra were acquired using the same objective and collector apertures that had semiangles of 9 and 7 mrad, respectively. These apertures ensure that all excitations to the conduction band in the first Brillouin zone would be collected. The energy drift present in the electronics has been minimized by alignment of the

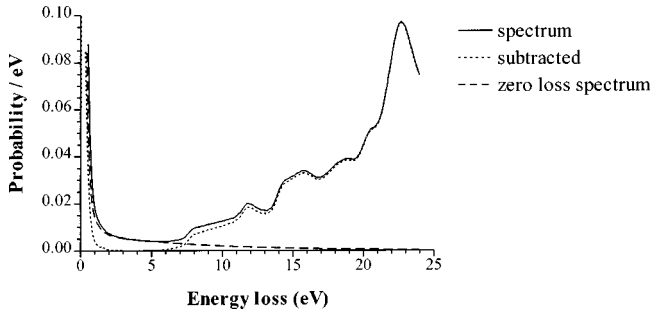


FIG. 6. A scaled subtraction of the zero loss peak works reasonably well down to energies of approximately 2 eV for large-band-gap materials, as shown for MgO, but does not reveal an onset to the intensity which corresponds to the size of the band gap. The subtracted spectrum is the zero loss spectrum scaled to fit the observed spectrum from the specimen over a region below the band gap.

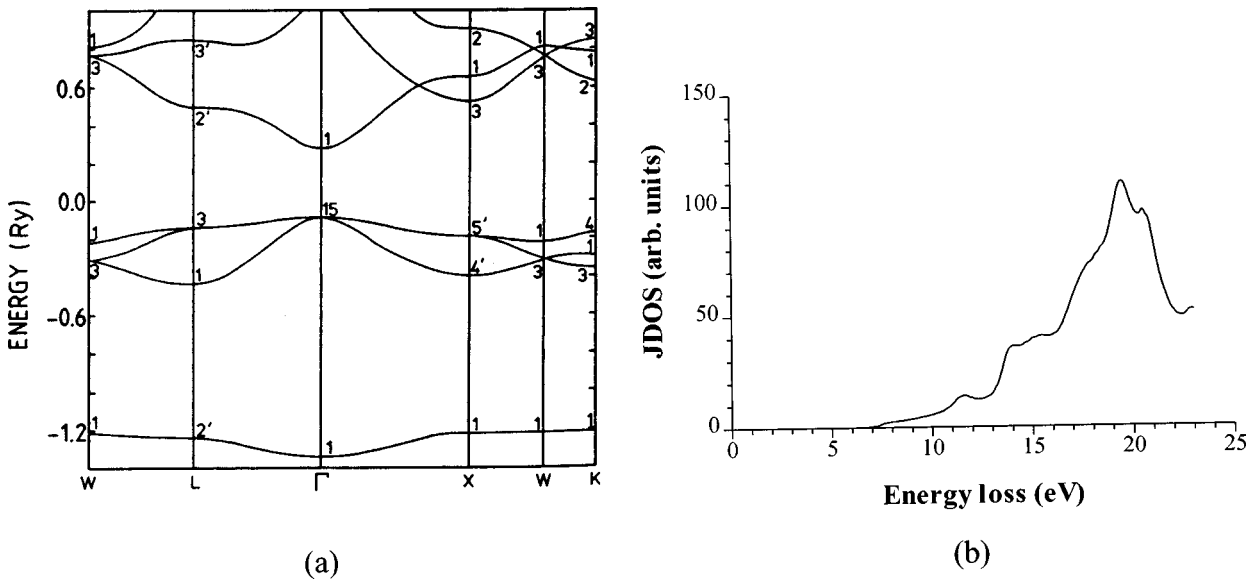


FIG. 7. (a) The electronic band structure of MgO (Ref. 17): the top of the valence band is marked at zero energy; and (b) its JDOS, showing an onset at ~7.2 eV.

EEL spectra on the zero loss of approximately 1000 individual spectra with exposure times of 50 ms and dispersion of 0.05 eV/channel.

Magnesium oxide

The MgO cube that was used to acquire the spectrum was tilted so that four faces of the cube are parallel to the electron beam. This means that the electron probe was centered on a [100] direction of the cube. The probe was also positioned so that it was passing through the centers of the faces that were perpendicular to the beam. This reduced the contribution of any additional surface effects in the spectrum.

The Brillouin zone of MgO cubes has body-centered-cubic symmetry; Fig. 7(a) shows a calculated band structure.[18] It has a direct band gap of 7.8 eV (Ref. 19) centered on the Γ point of the band structure. The bands between −6 and O eV have predominantly oxygen $2p$, character and form the highest occupied states of the valence band. The conduction-band edge at 7.8 eV is formed by the unoccupied magnesium $3s$ states.

Figure 7(b) shows the experimental JDOS spectrum from bulk MgO. The intensity starts to increase at an energy of ~7.22 eV, which is the onset of allowed transitions. This is the onset of the band gap. There is some residual intensity within the band gap that is produced by surface states.

CVD diamond

The chemical vapor deposition (CVD) diamond was tilted so that the electron beam was centered on the [001] pole direction, and positioned so that no defects or twin bound-

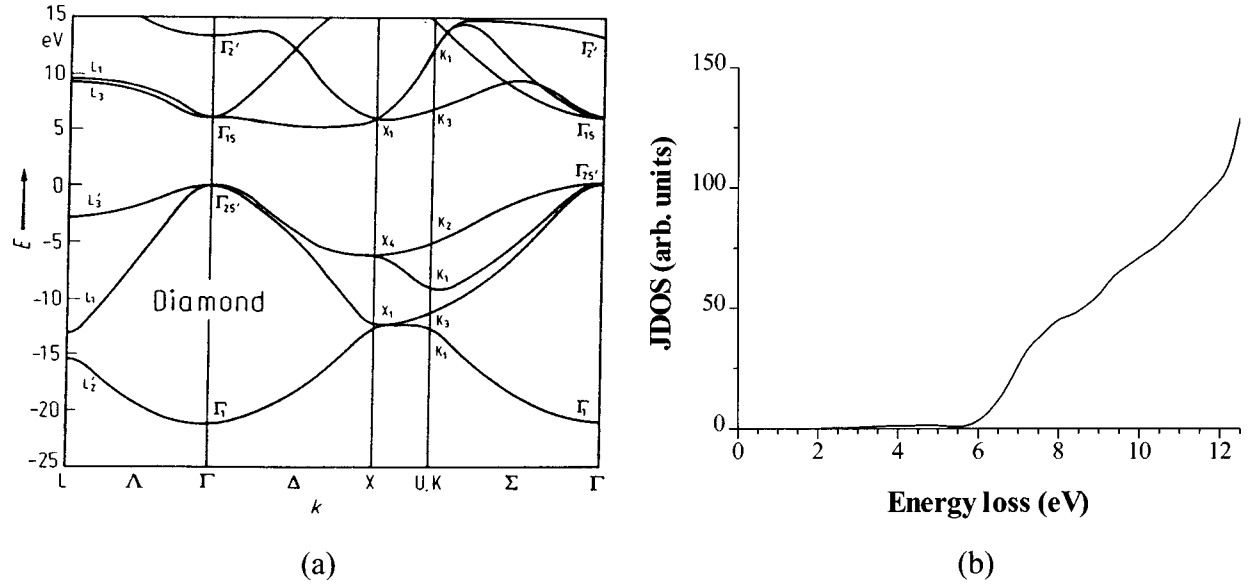


FIG. 8. (a) The electronic band structure of diamond (Ref. 20): the top of the valence band is marked at zero energy; and (b) its JDOS, showing a slow onset at ~5.5 eV.

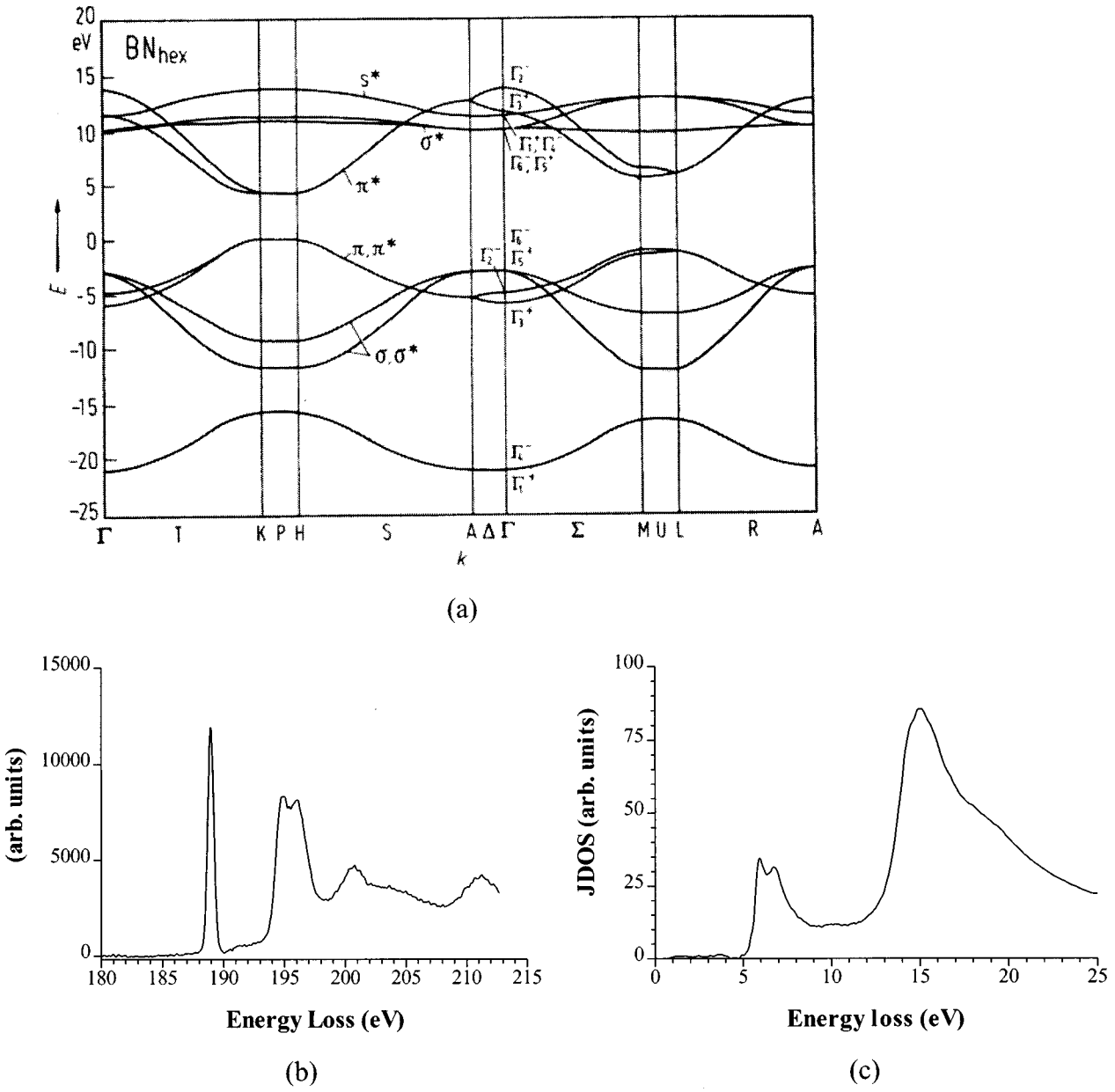


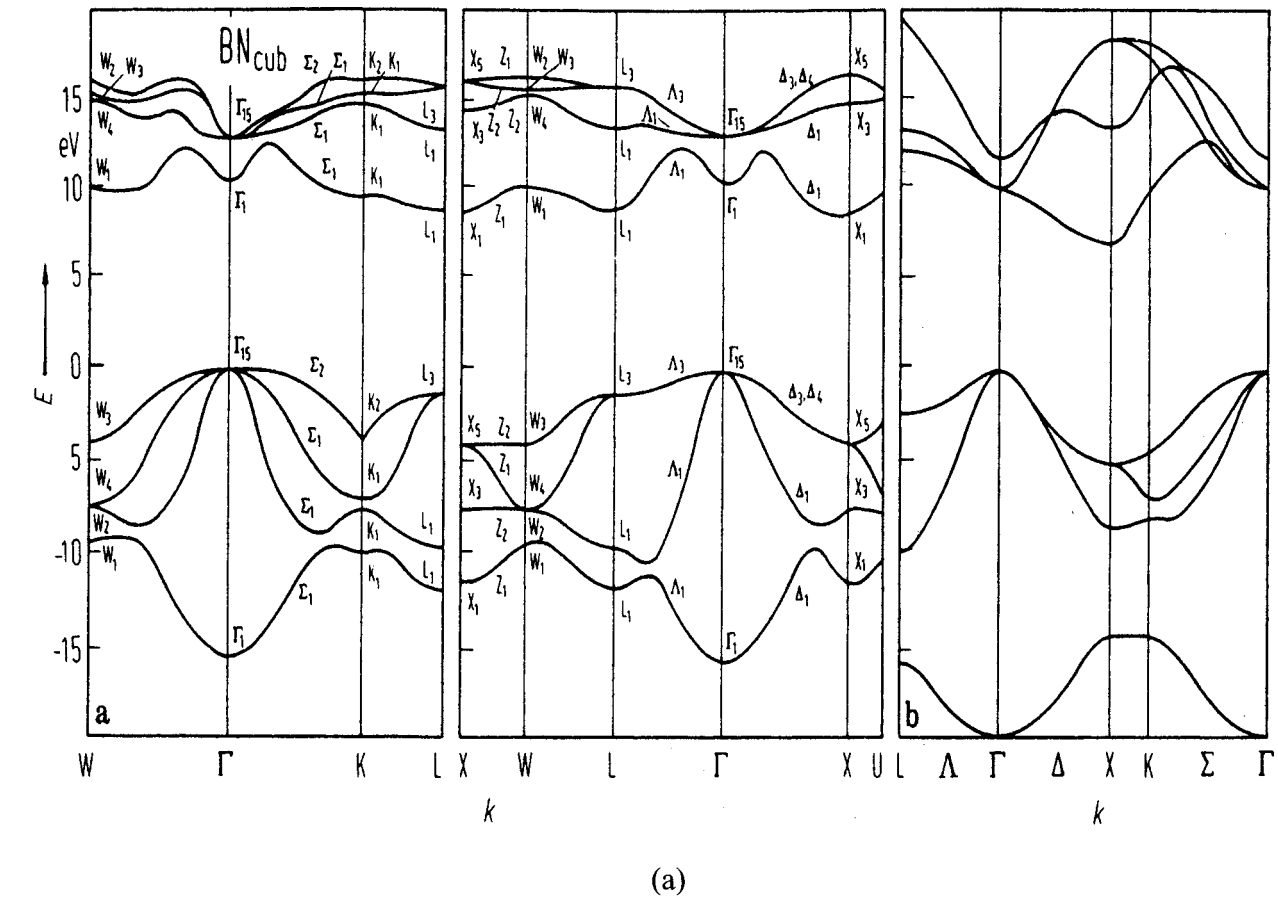
FIG. 9. (a) The electronic structure of hexagonal BN (Ref. 20); the top of the valence band is marked at zero energy. (b) The corresponding B *K* edge showing a sharp peak at ~187 eV due to the $1s-\pi^*$ states, and the doublet at ~193 eV due to the $1s-\sigma^*$ states. (c) The joint density of states showing a doublet with an onset at ~5 eV with an energy dependence of $(E-E_g)^{3/2}$ corresponding to the indirect band gap to the spin-split π^* states centered on the *M* point of the Brillouin zone.

aries where illuminated during acquisition. Diamond has a body-centered-cubic Brillouin zone, and has an indirect band gap of 5.5 eV for transitions from the top of the valence band centered on Γ to the bottom of the conduction band about 0.8 of the way along the Γ*X* direction [Fig. 8(a)]. It also has a direct band gap of 6.2 eV centered on the Γ point.[20]

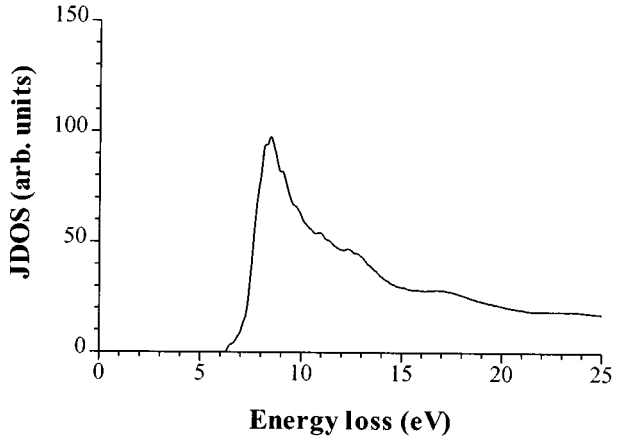
Figure 8(b) is the experimental joint density of states plot of diamond in the region of the band gap. Below 5.5 eV there is little or no intensity, showing the size of the band gap to be 5.5 eV, in good agreement with the optical data. Even though the two band gaps are of a similar energy and most electrons are scattered to small angles, we see the indirect contribution quite clearly because the conduction band in the vicinity of the indirect band gap is very flat and has a large JDOS.

Boron nitride: hexagonal and cubic forms

The hexagonal boron nitride has to be tilted to orientate the electron beam onto the *c* axis of a crystallite [refractive index $n_{c\ \mathrm{axis}}=2.13$ (Ref. 20) used in Kramers-Kronig analysis]. Crystalline cubic boron nitride has a zinc-blende structure. Both hexagonal and cubic boron nitrides have indirect band gaps. The band gap for hexagonal boron nitride is 5.2 eV, and is between a valence-band maximum at *H* and a conduction band minimum at *M* [Fig. 9(a)]. This is the gap between the hybridized π and π^* states. The boron *K*-edge excitation will show the *p*-projected density of empty states for the structure [Fig. 9(b)]. The peak at ~189 eV is the π^* peak and the doublet at ~196 eV is produced by the σ^* levels and s^* levels that can be seen centered on the ΓΔ*A*



(a)



(b)

FIG. 10. (a) The electronic structure of cubic BN (Ref. 20): the top of the valence band is marked by zero energy; and (b) its JDOS, showing the band gap to be approximately 6.2 eV.

direction of the band structure [Fig. 9(a)]. The states between the π^* and σ^* peaks are interlayer states which have a characteristic form which is shown in the K edge.[21] The joint density of states [Fig. 9(c)] can be seen to have a structure that mirrors the general structure in the boron K edge. It can be seen that the *splitting* of the peaks has been reversed; the boron K edge has the σ^* peak split, while the π-π^* peak in the JDOS is seen as a doublet. For the excitation of the boron K edge, this transition is almost perfectly direct. Thus we see the $1s-\pi^*$ peak as a singlet, produced by the flat region of the electronic structure in the region of the Γ point, and the $1s-\sigma^*$ peak mirrors the separation of the σ^* and s^* bands centered on the $\Gamma\Delta A$ region of the Brillouin zone, a separation varying between 2 and 3 eV. For the excitations that make up the JDOS, the splitting of the $\pi-\pi^*$ peak mirrors the splitting of the π^* band at the M point of the Brillouin zone, a much smaller splitting of 0.5 eV. The second peak is produced by both the direct transitions from the σ (centered

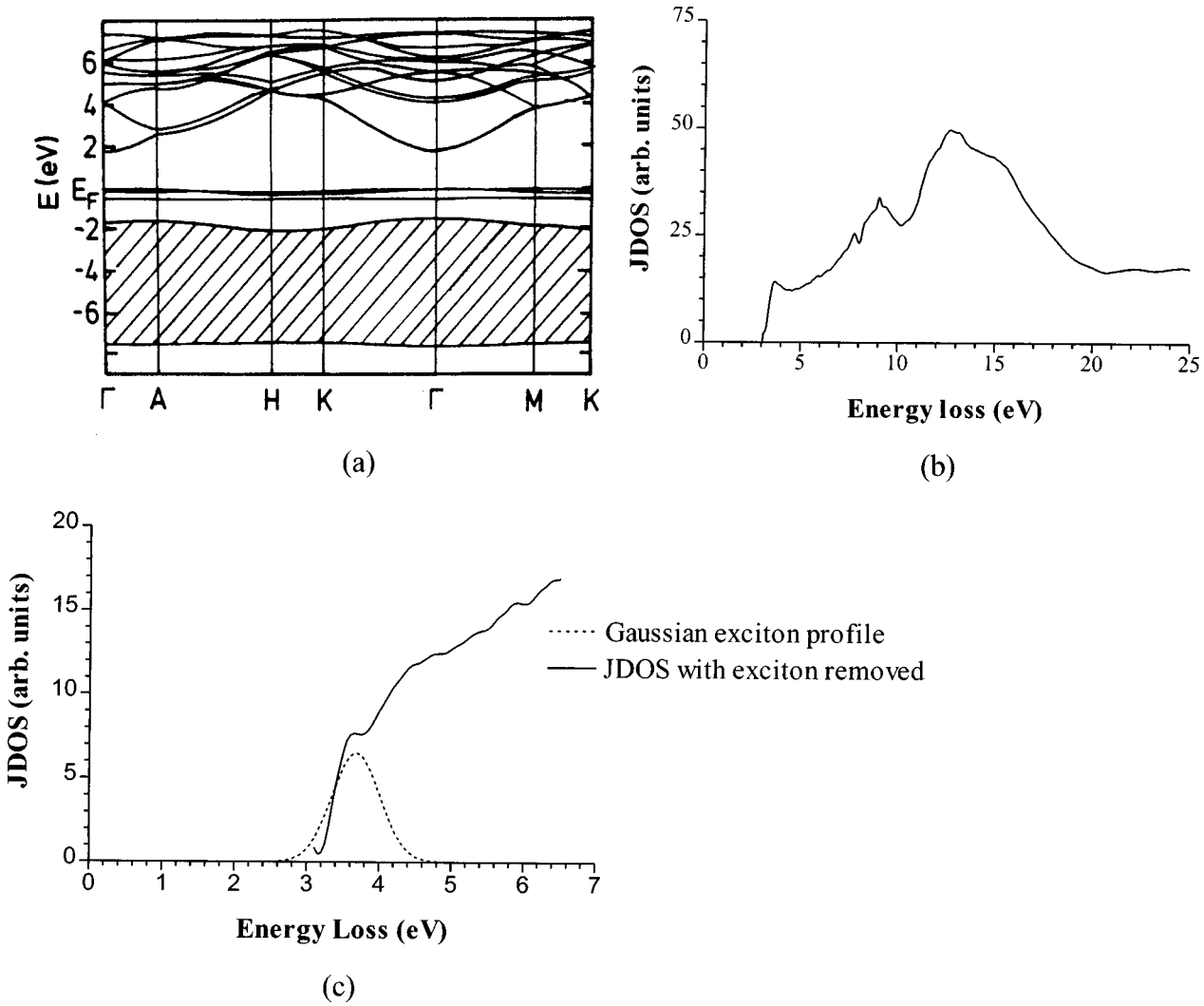


FIG. 11. (a) The electronic band structure of zinc oxide which was calculated with a Cu-atom defect which added a narrow band at the Fermi level (Ref. 21). (b) The JDOS of ZnO showing an excitionic peak at ~3.2 eV. (c) The JDOS with a Gaussian exciton profile removed.

on the Γ point) states to the σ^* states which will be dominated by indirect transitions because of the extremely low dispersion of the σ^* states.

The band gap can be clearly seen at 5.13 eV in Fig. 9(c). Close inspection of this figure shows that the intensity for the second main peak has an onset at approximately 11 eV. This is probably due to the indirect σ-σ^* transitions, although there will be some contribution from the direct $\pi-\sigma^*$ transitions centered on the *K*, *P*, and *H* symmetry points of the Brillouin zone. Even though there is the possibility for direct transitions at energies of ~5 eV, there is no sign of them in the spectra. This is because these transitions would be direct transitions between π and π^* states. The matrix elements for these transitions are just the dipole matrix elements, $\langle\pi^*|\mathbf{r}|\pi\rangle$; the angular term in this expression is independent of **r**, and so we are left with the selection rule $\Delta l = \pm 1$ for nonzero matrix elements. This argument also holds for the absence of direct σ-σ^* transitions.

Cubic boron nitride has an indirect band gap of 6.2 eV (Ref. 20) for transitions from the top of the valence band at Γ to the bottom of the conduction band at *X* [Fig. 10(a)]. There is also a direct band gap but this is at a higher energy, about 10 eV above the top of the valence band. Figure 10(b) shows the JDOS for cubic boron nitride with an onset of allowed transitions at approximately 6.22 eV. Just above the initial peak at ~11 eV there is some additional intensity which is probably the contribution of the direct transitions. Since it is sitting on a large contribution, it is difficult make any comments about the detailed shape of this contribution.

Zinc oxide

Zinc oxide is an ionic crystal with a wurtzite structure. The orientation of the crystallite that was used to acquire the data was difficult to determine because of the proximity of other crystallites and the low intensity of the diffracted discs in the microdiffraction pattern. There is a direct band gap of 3.4 eV at the center of the Brillouin zone, and a wide range of possible transitions from the occupied $3d$-like zinc bands to the unoccupied $2p$-like oxygen bands at energies of about ~10 eV [Fig. 11(a)].[22] The JDOS for zinc oxide can be seen in Fig. 11(b). Although this should have a direct band gap, the shape of the band edge looks more like that of the two forms of boron nitride rather than, say, magnesium oxide. There seems to be a peak at the onset that is obscuring the true shape of the band edge. This is probably due to a bound

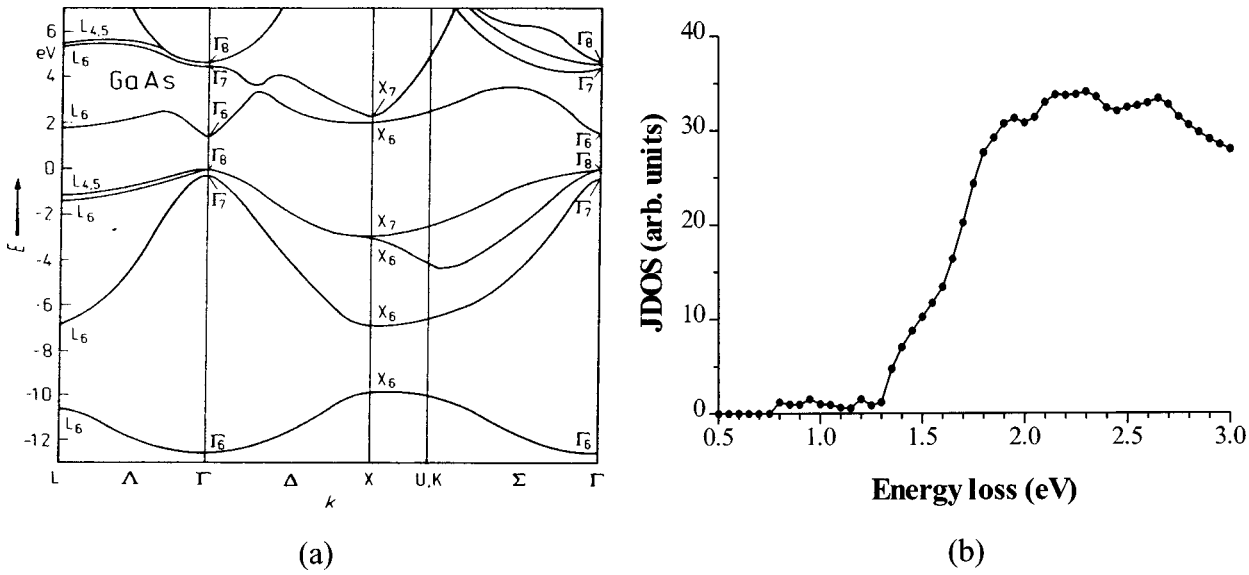


FIG. 12. The electronic band structure of gallium arsenide showing the characteristic camel's back's structure to the conduction-band edge centered on the Γ point. Zero energy corresponds to the top of the valence band (Ref. 19). (b) The JDOS of GaAs reflecting the detailed shape of the valence and conduction bands.

exciton that is common in these luminescent materials. Figure 11(c) shows the JDOS data after subtracting a Gaussian profile to account for the excitonic peak. It can be seen that there is still a residual peak at 3.7 eV due to the crude method employed to remove the excitonic peak.

Gallium arsenide

The gallium arsenide, which has a cubic zinc-blende structure, was orientated onto a [110] pole direction from which EELs spectra were then acquired, gallium arsenide has a direct band gap centered on the Γ point of the Brillouin zone of 1.5 eV (Ref. 20) [Fig. 12(a)], which has the characteristic *Camel's back structure* of the conduction-band edge.

Figure 12(b) shows the JDOS for gallium arsenide, it shows an onset at an energy of ~1.3 eV which is in very good agreement with the optical data. This spectrum is also in close agreement to that which was acquired by Batson *et al.*[23] The JDOS also contains structure at energies of 2.0 and 2.6 eV. The first of these energies corresponds to the energy separation between the top of the valence band for light holes [labeled Γ_7 in Fig. 12(a)] and the bottom of the conduction band [labeled Γ_6 in Fig. 12(a)]. The second energy corresponds to the energy separation between the top of the valence band for heavy holes (Γ_8) and the camel's hump on the ΓL side of the conduction-band minima (Γ_6).

The intensity in the JDOS just before the onset is ascribed to noise from the deconvolution routine. This puts a limit on the smallest band-gap material we can look at as ~0.8–1 eV. This limit is intrinsic to the system, and is set by the energy distribution of the electrons tunneling from the tip: the tail of the zero loss peak necessarily produces noise which obscures the onset of losses in the spectrum.

Comparison between theory and experiment

The question to be answered now is if the onsets of the JDOS plots follow a power law of the form $(E-E_g)^n$, and if the value of n is $\frac{1}{2}$ or $\frac{3}{2}$ for direct and indirect transitions, respectively. From the original JDOS data it is easy to define an energy range which includes the critical point that we wish to examine; this range comprises of about 30–40 data points. Once this is done we choose the first datum point in this range, and assign its energy value to be that of the critical point. The data are offset so that this critical point is the origin of the energy axis, and a log-log plot is then constructed. To this data a general straight line ($Y=A+B\times X$) is fitted, using a linear regression routine, to the spectral data. The value of B will be the best-fit value of the index n for this critical point. From this fitted line we can also extract the correlation factor R^2. This procedure is repeated for each energy point in the initial energy range around the critical point. We can now plot the index n and the correlation factor R^2 as a function of the position of the critical point.

Figures 13(a)–13(f) show the results of this analysis for MgO cubes and CVD diamond, both hexagonal and cubic BN, GaAs and ZnO, respectively. The error bars on the data points for the index n are $\pm\sigma$ the standard deviation derived from the fitting procedure. It can be seen that the error in the value of the index n is minimized when the correlation factor is maximized. All of these plots show the correlation factor to be maximized at energies that are in good agreement with the optical values for the fundamental band gaps (Table I). When the correlation factor is maximized, the value of the index n is also seen to be at either $\frac{1}{2}$ or $\frac{3}{2}$, showing that the band gap is either direct or indirect, respectively. The values of the index n that are observed in the plots indicate the correct nature of the type of transitions to the critical points for each of the materials. In particular the correlation factor in diamond has two peaks corresponding to both indirect and direct band gaps. The plot for GaAs shows the correlation function to be at a slightly lower energy than the accepted value for the band gap. This is probably due to the presence of an excitonic peak near the onset that is masking the true position of the onset of interband transitions.[23] The JDOS for ZnO can be analyzed once the excitonic peak has been removed. Although this has been done in a rather crude fash-

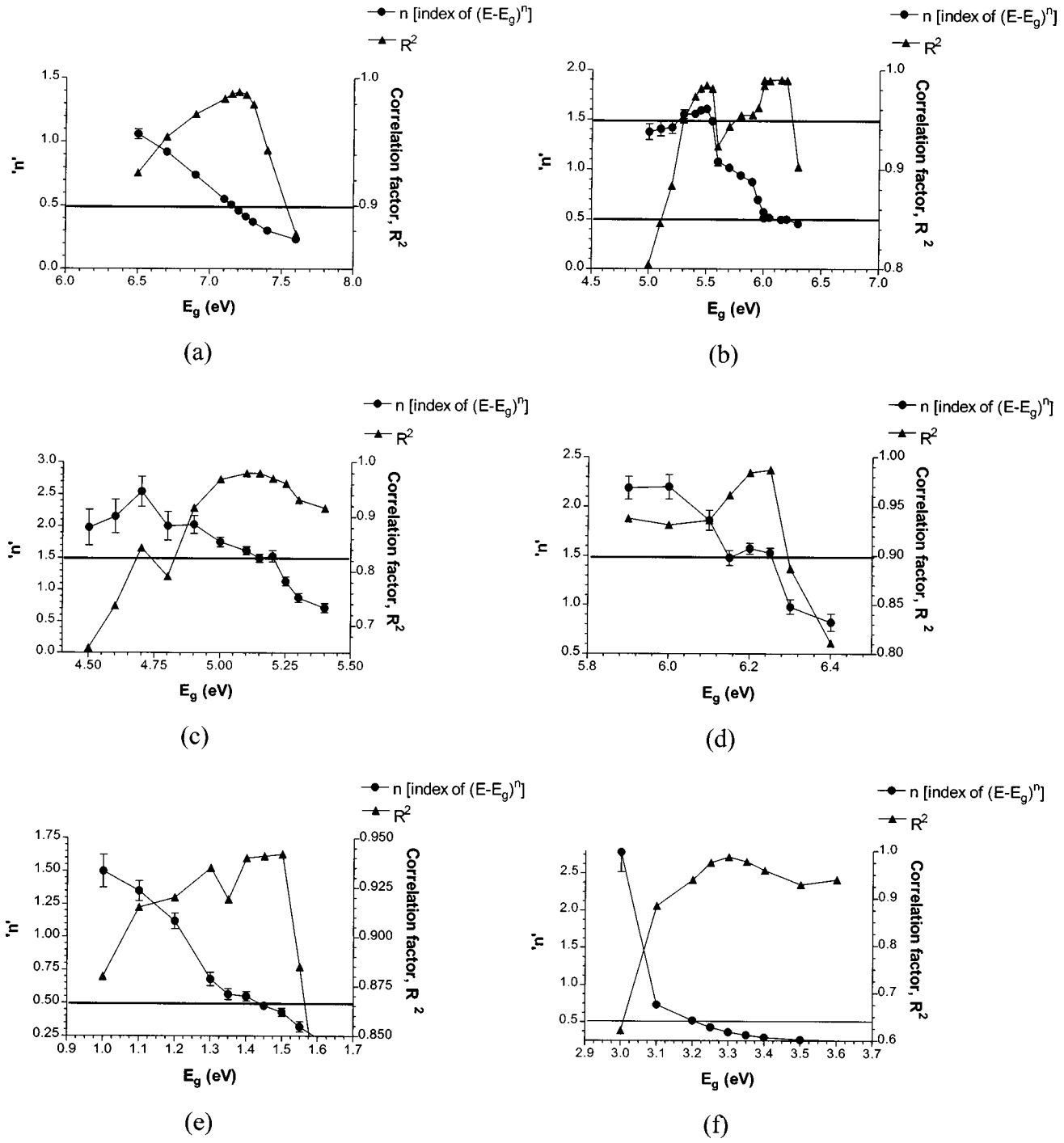


FIG. 13. Plots showing the index n of the $(E-E_g)^n$ term and the correlation factor for the fitting procedure, described in the text, as a function of the position of the critical point for (a) MgO cubes, (b) CVD diamond, (c) hexagonal BN, (d) cubic BN, (e) GaAs, and (f) ZnO. The correlation curves all peak at the correct position of the critical point and the correct index n is observed at these critical points which define the nature of the bandgap transitions: $\frac{1}{2}$ for direct transitions, and $\frac{3}{2}$ for indirect transitions.

ion, the above procedure has still revealed a band gap of 3.3 ± 0.05 eV which is excellent agreement with the true optical band gap.[24] The errors in the values for the band gaps measured in this way from the EELS data can be estimated from the error in the peak position of the correlation curves. Using this criterion the errors in the values of E_g are all approximately ±0.05 eV.

IV. CONCLUSIONS

A theory has been developed to understand the detailed shape of bandgap EELs spectra. This theory shows that the matrix elements have significant effects on the shape of the spectra, especially when considering direct or indirect transitions across the fundamental band gap of a semiconductor or insulator. The theory shows how the matrix elements for

TABLE I. A comparison of the size and types of band gaps between values in the literature and those extracted from the correlation plots of Fig. 13. These errors indicated by * are estimates from the uncertainty in the position of the maximum in the correlation coefficients shown in Fig. 13.

	Literature		Correlation plots	
Material	E_g (eV)	Type of transitions	E_g (eV)	Type of transitions
MgO	7.8 (Ref. 18)	direct	7.22±0.03*	direct
CVD diamond	5.5 (Ref. 20)	indirect	5.50±0.04*	indirect
	6.2 (Ref. 20)	direct	6.12±0.07*	direct
BN (hexagonal)	5.2 (Ref. 20)	indirect	5.13±0.08*	indirect
(cubic)	6.2 (Ref. 20)	indirect	6.22±0.03*	indirect
ZnO	3.1–3.3 (Ref. 22)	direct	3.3±0.05*	direct
GaAs	1.5 (Ref. 20)	direct	1.45±0.05*	direct

direct and indirect transitions can be decoupled from one another. For the case of direct transitions the matrix elements still act as a scaling factor to the JDOS, and so an $(E-E_g)^{1/2}$ term is observed in the spectrum. The matrix elements for indirect transitions are shown to be dependent on the momentum transferred by the incident fast electron to the crystal electrons, resulting in the product of the indirect transition matrix elements and the JDOS contributing as an $(E-E_g)^{3/2}$ term to the EEL spectrum.

This theory has been tested on six crystalline materials. It is shown that before any information can be extracted from the spectra the zero loss peak must be removed. By deconvolving a vacuum zero loss peak from the acquired spectrum, it is possible to study the band-gap region of the EELs spectrum for materials with band gaps as small as ~0.9 eV at an experimental energy resolution (the full width at half maximum of the zero loss peak) of at most 0.22 eV. To test the theory, it is shown that the *matrix element weighted joint density of states* should be extracted from the EEL spectra. These data contain all of the information needed to study the electronic structure. To derive the nature of a fundamental band gap, a numerical routine has been outlined which plots the index n of the $(E-E_g)^n$ term and the fitting correlation factor as a function of the assumed size of the band gap. This routine reveals the nature and size of a band gap that is in excellent agreement with the data in the literature for the materials that were studied. The routine provides an objective method to assess band gaps by electron spectroscopy to an accuracy of ±0.05 eV. It should be stressed that the accuracy of the method far surpasses the energy resolution of the spectrometer because it is merely limited by the location of the critical points in the data that is entirely a question of noise.

Although this theory has been developed with inelastic electron scattering in mind, it must be noted that it is equally valid for inelastic photon scattering when phonons are used as a source of momentum transfer.

ACKNOWLEDGMENTS

One of us (B.R.) would like to thank De Beers Industrial Diamond Division for its financial support throughout this work, and D. A. Ritchie of the Semiconductor Physics Group, Cavendish Laboratory, for providing the GaAs sample.

[1] J. J. Thompson, Philos. Mag. **XLIV**, 301 (1897).
[2] N. Bohr, Philos. Mag. **25**, 10 (1913).
[3] N. Bohr, Philos. Mag. **30**, 581 (1915).
[4] H. A. Bethe, Z. Phys. **76**, 293 (1932).
[5] E. Fermi, Phys. Rev. **57**, 485 (1940).
[6] D. Bohm and D. Pines, Phys. Rev. **82**, 625 (1950).
[7] D. Pines and D. Bohm, Phys. Rev. **85**, 338 (1952).
[8] D. Bohm and D. Pines, Phys. Rev. **92**, 609 (1953).
[9] J. Bruley and L. M. Brown, *Analytical Electron Microscopy Workshop*, 1987 Proceedings, edited by G. W. Lorimer (The Institute of Metals, London, 1988).
[10] P. E. Batson, in *Transmission Electron Energy Loss Spectroscopy in Materials Science*, edited by M. M. Disko, C. C. Ahn, and B. Fultz, Electronic, Magnetic and Photonic Materials Divison Monograph Series Vol. 2 (The Minerals, Metals and Materials Society, Pennsylvania, 1991), p. 217.
[11] J. Hubbard, Proc. Phys. Soc. London **68**, 976 (1955).
[12] R. H. Ritchie, Phys. Rev. **85**, 338 (1957).
[13] R. F. Egerton, *Electron Energy Loss Spectroscopy in the Electron Microscope*, 2nd ed. (Plenum, New York, 1989).
[14] M. Inokuti, Rev. Mod. Phys. **43**, 297 (1971).
[15] J. Bruley, Ph.D. thesis, Cambridge University, 1987.
[16] P. F. Bryd and M. D. Friedman, *Handbook of Elliptical Integrals for Engineers and Scientists*. 2nd ed. (Springer-Verlag, Heidelberg, 1971).
[17] R. J. Elliot, Phys. Rev. **108**, 1384 (1957).
[18] G. Timmer and G. Borstel, Phys. Rev. B **43**, 5098 (1991).
[19] D. M. Roessler and W. C. Walker, Phys. Rev. **159**, 733 (1967).
[20] O. Madelung, *Semiconductors—Basic Data*. 2nd ed. (Springer-Verlag, Berlin, 1996).
[21] C. Pickard (private communication).
[22] K. C. Mishra *et al.*, Phys. Rev. B **42**, 1423 (1990).
[23] P. E. Batson, *et al.*, Phys. Rev. Lett. **57**, 2729 (1986).
[24] V. Srikant and D. R. Clarke, J. Appl. Phys. **83**, 5447 (1998).