Theory of the scanning tunneling microscope

J. Tersoff* and D. R. Hamann
AT&T Bell Laboratories, Murray Hill, New Jersey 07974
(Received 25 June 1984)

We present a theory for tunneling between a real surface and a model probe tip, applicable to the recently developed "scanning tunneling microscope." The tunneling current is found to be proportional to the local density of states of the surface, at the position of the tip. The effective lateral resolution is related to the tip radius R and the vacuum gap distance d approximately as $[(2\ \text{Å})(R+d)]^{1/2}$. The theory is applied to the 2×1 and 3×1 reconstructions of Au(110); results for the respective corrugation amplitudes and for the gap distance are all in excellent agreement with experimental results of Binnig *et al.* if a 9-Å tip radius is assumed. In addition, a convenient approximate calculational method based on atom superposition is tested; it gives reasonable agreement with the self-consistent calculation and with experiment for Au(110). This method is used to test the structure sensitivity of the microscope. We conclude that for the Au(110) measurements the experimental "image" is relatively insensitive to the positions of atoms beyond the first atomic layer. Finally, tunneling to semiconductor surfaces is considered. Calculations for GaAs(110) illustrate interesting qualitative differences from tunneling to metal surfaces.

I. INTRODUCTION

One of the most fundamental problems in surface physics is the determination of surface structure. Recently a new and uniquely promising technique, the "scanning tunneling microscope" (STM), was introduced.[1–4] This method offers, for the first time, the possibility of *direct, real-space* determination of surface structure in three dimensions, including nonperiodic structures. A small metal tip is brought near enough to the surface that the vacuum tunneling resistance between surface and tip is finite and measurable. The tip scans the surface in two dimensions, while its height is adjusted to maintain a constant tunneling resistance. The result is essentially a contour map of the surface.

For electronic states at the Fermi level, the surface represents a potential barrier whose height is equal to the work function ϕ. As expected by analogy with planar tunneling, the current varies exponentially with the vacuum gap distance, with decay length $\hbar(8m\phi)^{-1/2}$. For typical metallic work functions, this length is about 0.4 Å. Thus, aside from issues of *lateral* resolution, in the constant-current scanning mode the tip may be expected to follow the surface height to 0.1 Å or better. It can be seen from the data that the new microscope designs have sufficient mechanical stability to achieve this in practice.[1–4]

The one-dimensional tunneling problem (i.e., through two-dimensionally uniform barriers) has been treated extensively,[5] and field emission from a tip is well understood. The usefulness of STM stems from the fact that it is neither one dimensional nor operating as a field emitter, but is instead sensitive to the full three-dimensional structure of the surface. Little was known quantitatively about tunneling in this case, until the recent development of STM motivated the present work (parts of which were reported briefly elsewhere[6]), and other approaches,[7,8] which are discussed briefly below.

Here we develop a theory of STM which is at once sufficiently realistic to permit quantitative comparison with experimental "images," and sufficiently simple that the implementation is straightforward. The surface is treated "exactly," while the tip is modeled as a locally spherical potential well where it approaches nearest the surface. This treatment is intuitively reasonable and is consistent with the current poor understanding of the actual microscopic geometry of the tip, which is prepared in an uncontrolled and nonreproducible manner.[9]

In Sec. II we present the formal development of the theory. The tunneling current is found to be proportional to the (bare) surface local density of states (LDOS) at the Fermi level (E_F) at the position of the tip. The effective lateral resolution is roughly $[(2\ \text{Å})(R+d)]^{1/2}$, where R is the tip radius of curvature and d is the vacuum gap. General features of the surface LDOS are discussed, as are the various approximations. Some other recent approaches[7,8] to the problem are also considered.

Section III describes a calculation for the 2×1 and 3×1 reconstructions of the Au(110) surface. The results are in quantitative agreement with recent measurements of Binnig *et al.*[4] if a 9-Å tip radius is assumed. General features and limitations of the numerical implementation are discussed. In particular, self-consistent electronic structure calculations of vacuum charge far from the surface are at present only feasible for systems with small unit cells.

We therefore introduce in Sec. IV a crude approximation for the surface LDOS, which permits convenient calculation of the STM image even for large unit cells or nonperiodic systems. Comparison with results of Sec. III shows that the approximation works rather well, at least for Au(110). Using this approximation, we compare the images expected for different possible structures of Au(110). We conclude that STM is rather insensitive to

the position of the surface layer relative to the underlying layers. For the Au(110) 3×1 surface, even the presence or absence of a missing row in the second layer cannot be reliably distinguished.

Finally, in Sec. V, we consider the case of a semiconducting surface. The theory is expected still to apply, though with some modifications. In particular, the image may be qualitatively different for different tunneling polarity or sample doping. This effect is illustrated with calculations for GaAs(110).

II. THEORY OF STM

While it is easy to write down a formal expression for the tunneling current, many approximations are needed to derive an expression which permits practical computation. Some of the approximations made below are sufficiently drastic that they can be justified only because of the relative insensitivity of any conclusions to the resulting errors. It is therefore not convenient to justify the various approximations as they are introduced. Instead, we first present the theory in Sec. II A. Then in Sec. II B we consider general features of the surface local density of states and, hence, of the tunneling current as a function of tip position. These results determine the intrinsic resolution and sensitivity of STM. Finally, in Sec. II C we consider the various approximations and their possible effect.

A. Tunneling current

The tunneling current is given to first order in Bardeen's[10] formalism by

$$I=\frac{2\pi e}{\hbar}\sum_{\mu,\nu}f(E_\mu)[1-f(E_\nu+eV)]\,|M_{\mu\nu}|^2\delta(E_\mu-E_\nu)\,, \tag{1}$$

where $f(E)$ is the Fermi function, V is the applied voltage, $M_{\mu\nu}$ is the tunneling matrix element between states ψ_μ of the probe and ψ_ν of the surface, and E_μ is the energy of state ψ_μ in the absence of tunneling. Note that while (1) resembles ordinary first-order perturbation theory, it is formally different in that ψ_μ and ψ_ν are nonorthogonal eigenstates of different Hamiltonians. For high temperatures there is a corresponding term for reverse tunneling. Since the experiments are performed at room temperature or below and at small voltage ($\sim$10 meV for metal-metal tunneling), we take the limits of small voltage and temperature,

$$I=\frac{2\pi}{\hbar}e^2V\sum_{\mu,\nu}|M_{\mu\nu}|^2\delta(E_\nu-E_F)\delta(E_\mu-E_F)\,. \tag{2}$$

Before attempting a realistic treatment, it is worthwhile to consider the limit where the tip is replaced with a point probe. This case represents the ideal of a nonintrusive measurement of the surface, with the maximum possible resolution. If the tip wave functions are arbitrarily localized, then the matrix element is simply proportional to the amplitude of ψ_ν at the position $\vec{r}_0$ of the probe, and (2) reduces to

$$I\propto\sum_\nu|\psi_\nu(\vec{r}_0)|^2\delta(E_\nu-E_F)\,.$$

The quantity on the right is simply the surface local density of states (LDOS) at E_F, i.e., the charge density from states at E_F. Thus the tunneling current is proportional to the surface LDOS at the position of the point probe, and the microscope image represents a contour map of constant surface LDOS. This almost trivial result anticipates major features of the more complete treatment below.

In handling (2) in general, the essential problem is to calculate $M_{\mu\nu}$. Bardeen[10] has shown that

$$M_{\mu\nu}=\frac{\hbar^2}{2m}\int d\vec{S}\cdot(\psi_\mu^*\vec{\nabla}\psi_\nu-\psi_\nu\vec{\nabla}\psi_\mu^*)\,, \tag{3}$$

where the integral is over any surface lying entirely within the vacuum (barrier) region separating the two sides. The quantity in parentheses is simply the current operator.

To evaluate $M_{\mu\nu}$, we expand the surface wave function in the form

$$\psi_\nu=\Omega_s^{-1/2}\sum_G a_G\exp[(\kappa^2+|\vec{\kappa}_G|^2)^{1/2}z]\exp(i\vec{\kappa}_G\cdot\vec{x})\,, \tag{4}$$

which is a completely general expression for ψ in the region of negligible potential. Here Ω_s is sample volume, $\kappa=\hbar^{-1}(2m\phi)^{1/2}$ is the minimum inverse decay length for the wave functions in vacuum, ϕ is the work function, and $\vec{\kappa}_G=\vec{k}_{||}+\vec{G}$, where $\vec{k}_{||}$ is the surface Bloch wave vector of the state, and $\vec{G}$ is a surface reciprocal-lattice vector. The first few a_G are typically of order unity. For a nonperiodic surface the sum over G becomes an integral.

Since the microscopic structure of the tip is not yet known, we model it as a locally spherical potential well where it approaches nearest to the surface, as illustrated in Fig. 1. R is the local radius of curvature about the center located at $\vec{r}_0$, and d is the distance of nearest approach to the surface. In the region of interest, the wave functions of the tip are taken to have the asymptotic spherical form

$$\psi_\mu=\Omega_t^{-1/2}c_t\kappa Re^{\kappa R}(\kappa|\vec{r}-\vec{r}_0|)^{-1}e^{-\kappa|\vec{r}-\vec{r}_0|}\,, \tag{5}$$

where Ω_t is the probe volume and κ is defined as above.

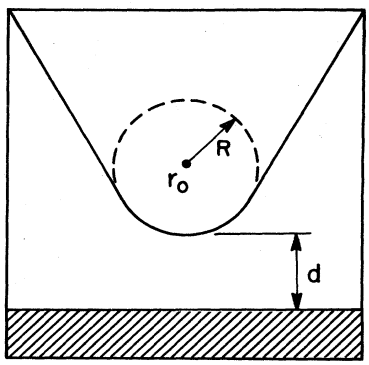

FIG. 1. Schematic picture of tunneling geometry. Probe tip has arbitrary shape but is assumed locally spherical with radius of curvature R, where it approaches nearest the surface (shaded). Distance of nearest approach is d. Center of curvature of tip is labeled $\vec{r}_0$.

(We assume for simplicity that the work function ϕ for the tip is equal to that of the surface.) The form is chosen to be correctly normalized when the parameter c_t (which is determined by the tip geometry, detailed electronic structure, and tip-vacuum boundary condition) is of order 1. We have neglected the possible angular dependence of ψ_μ, which introduces some quantitative modifications discussed below.

We expand the tip wave function (5) in the same form as the surface (4) using the fact that

$$(\kappa\vec{r})^{-1}e^{-\kappa\vec{r}}=\int d^2q b(\vec{q})\exp[-(\kappa^2+q^2)^{1/2}|z|] \times\exp(i\vec{q}\cdot\vec{x}) , \tag{6}$$

$$b(\vec{q})=(2\pi)^{-1}\kappa^{-2}(1+q^2/\kappa^2)^{-1/2} . \tag{7}$$

The matrix element is then almost trivial to evaluate. Substituting the surface and the tip wave functions in (3) and evaluating the expansion term by term in G, one finds

$$M_{\mu\nu}=\frac{\hbar^2}{2m}4\pi\kappa^{-1}\Omega_t^{-1/2}\kappa R e^{\kappa R}\psi_\nu(\vec{r}_0) , \tag{8}$$

where $\vec{r}_0$ is the position of the *center of curvature* of the tip. Substituting into (2), the desired result is

$$I=32\pi^3\hbar^{-1}e^2V\phi^2D_t(E_F)R^2\kappa^{-4}e^{2\kappa R} \times\sum_\nu|\psi_\nu(\vec{r}_0)|^2\delta(E_\nu-E_F) , \tag{9}$$

where D_t is the density of states per unit volume of the probe tip. Note that (8) does *not* imply that the value of the surface wave function ψ_ν at $\vec{r}_0$ is *physically* relevant. The matrix element is determined by an integral entirely within the gap region. However, because of the analytic properties of (4) and (5), the formal evaluation of ψ_ν at distance $R+d$ correctly describes the lateral averaging due to finite tip size.

The spherical-tip approximation entered only the normalization of (5). The crucial approximation was evaluating the matrix element only for an s-wave tip wave function. The $\vec{q}$ dependence of $b(\vec{q})$ in (7) then canceled that of the z derivative in the matrix element (3), so that (9) involved only undistorted wave functions of the surface. For tip wave functions with angular dependence ($l\neq0$), it is sufficient to include the $m=0$ term (other m give a node towards the surface). In that case, the terms in the Fourier expansion of ψ_ν are weighted by a factor $\sim(1+q^2/\kappa^2)^{l/2}$ in the matrix element, which for relevant values of q can be approximated by unity for small l. (In the example below the relevant $q^2/\kappa^2\approx0.1$.) The tip model therefore becomes less accurate for large R, where higher l values become more important. A more exact treatment would probably be far less useful, since it would require more specific information about the tip wave functions, and would not reduce to an explicit equation such as (9) or (10) below.

Substituting typical metallic values into (9), one obtains for the tunneling conductance

$$\sigma\approx0.1R^2e^{2\kappa R}\rho(\vec{r}_0,E_F) , \qquad \rho(\vec{r}_0,E)\equiv\sum_\nu|\psi_\nu(\vec{r}_0)|^2\delta(E_\nu-E) , \tag{10}$$

where σ is in ohms^{-1}, distances are in a.u., and energy in eV. Since $|\psi_\nu(\vec{r}_0)|^2\propto e^{-2\kappa(R+d)}$, we see from (10) that $\sigma\propto e^{-2\kappa d}$ as expected. Because of the exponential dependence on distance, it is not essential that the coefficient in (10) be very accurate.

We considered above the limit of a point probe. Realistically, the sharpest tip imaginable is a single atom, supported on a cluster or small plateau. The form (5) is not really appropriate for determining the normalization of ψ_μ in that case. However, because of the insensitivity of results to the coefficients, an adequate estimate for the single-atom case may be obtained simply by taking $R=2\kappa^{-1}$ (roughly the metallic radius for most metals) in (10).

Note that $\rho(\vec{r},E_F)$ is simply the surface local density of states (at E_F) at the point $\vec{r}$ or, equivalently, the charge per unit energy from states of the surface at E_F. According to (10), at constant current the tip follows a contour of constant $\rho(\vec{r},E_F)$. We therefore consider the behavior of $\rho(\vec{r},E_F)$ in some detail.

B. General features of $\rho(\vec{r},E_F)$

Within the approximations above, the microscope image is simply a contour of constant $\rho(\vec{r},E_F)$ of the surface. The behavior of $\rho(\vec{r},E_F)$, along with the tip radius, therefore determines the resolution and sensitivity of STM. Moreover, a detailed picture of $\rho(\vec{r},E_F)$ is essential in assessing the approximate method described in Sec. IV.

The starting point here is Eq. (4) for the surface wave function. A given wave function ψ_ν contributes a charge density

$$|\psi_\nu|^2=\Omega_s^{-1}\sum_{G,G'}a_Ga_{G'}^*\exp[-(\kappa^2+\kappa_G^2)^{1/2}z-(\kappa^2+\kappa_{G'}^2)^{1/2}z+i(\vec{\kappa}_G-\vec{\kappa}_{G'})\cdot\vec{r}] . \tag{11}$$

Since $\vec{\kappa}_G-\vec{\kappa}_{G'}=\vec{G}-\vec{G}'$, $|\psi_\nu|^2$ has the periodicity of the lattice and can be Fourier expanded,

$$|\psi_\nu|^2=\sum_G u_{\nu G}(z)e^{i\vec{G}\cdot\vec{r}} . \tag{12}$$

The total $\rho(\vec{r},E)$ may similarly be written

$$\rho(\vec{r},E)=\sum_\nu|\psi_\nu|^2\delta(E_\nu-E) =\sum_G\rho_G(z,E)e^{i\vec{G}\cdot\vec{r}} , \tag{13}$$

$$\rho_G(z,E)=\sum_\nu u_{\nu G}(z)\delta(E_\nu-E). \tag{14}$$

At sufficiently large distances $\rho(\vec{r},E)$ becomes rather smooth, and only the lowest nonzero Fourier component need be retained for this discussion. This is so in the example of Au(110) in Sec. III, where the STM image is practically sinusoidal. (The case where the image is highly structured is considered below.) Then

$$\rho(\vec{r},E_F)=\rho_0(z,E_F)+2\rho_{G_1}(z,E_F)\cos(\vec{G}_1\cdot\vec{x}) , \tag{15}$$

where we have assumed a reflection symmetry and where G_1 is the smallest G. Far from the surface $\rho_0(z,E_F)$ is dominated by states near the center ($\bar{\Gamma}$) of the surface Brillouin zone, since $\kappa_G=0$ gives the longest decay length (decay constant $=2\kappa$), in (4) and (11). It can be seen from minimizing the exponents in (11) that the longest decay length for ρ_{G_1} occurs at $\vec{\kappa}_{||}=\frac{1}{2}\vec{G}_1$. Then $|\kappa_G|=\frac{1}{2}G_1$ for $G=0$ or $-G_1$. The corresponding asymptotic decay constant for ρ_{G_1} is

$$\frac{d}{dz}\ln[\rho_{G_1}(z,E_F)]=2(\kappa^2+\tfrac{1}{4}G_1^2)^{1/2}$$
$$\approx 2\kappa+\tfrac{1}{4}\kappa^{-1}G_1^2 \tag{16}$$

using $\frac{1}{2}G_1\ll\kappa$. [For Au(110), $G_1/2\kappa\approx 0.1$.]

The extremal values of z for constant current [constant $\rho(\vec{r},E_F)$] occur at $\cos(\vec{G}_1\cdot\vec{x})=\pm 1$, and these are denoted $z_\pm$ here. Then from (15),

$$\rho(\vec{r})=\rho_0(z_\pm)\pm 2\rho_G(z_\pm) , \tag{17}$$

where the argument E_F is omitted for simplicity. Defining the corrugation amplitude $\Delta=z_+-z_-$ and using $\rho_0(z_+)\approx e^{-2\kappa\Delta}\rho_0(z_-)$, (17) gives

$$e^{-2\kappa\Delta}\approx\frac{\rho_0(z)-2\rho_G(z)}{\rho_0(z)+2\rho_G(z)} , \tag{18}$$

where z is some average value between z_+ and z_-. At distances were the image is sufficiently smooth ($\kappa\Delta\ll 1$), using (16), (18) becomes

$$\Delta\approx 2\kappa^{-1}\rho_G(z)/\rho_0(z)\propto e^{-\beta z} ,$$
$$\beta\equiv 2(\kappa^2+\tfrac{1}{4}G_1^2)^{1/2}-2\kappa\approx\tfrac{1}{4}\kappa^{-1}G_1^2 . \tag{19}$$

Thus, the corrugation decreases exponentially with distance from the surface, the decay length β being very sensitive to the surface lattice constant. This corrugation decay length is in agreement with numerical calculations described below. Of course, the result applies only far from the surface ($\beta z>1$), and is not strictly correct (though it still works well in practice) if there is a gap in the projected one-dimensional density of states at E_F for $\vec{k}_{||}=0$ or $\frac{1}{2}\vec{G}_1$.

If the surface unit cell is large, then the features of interest in the image may be well localized within the unit cell. This is the case, for example, in images of the Si(111) 7×7 surface.[3] Then the Bloch wave vector may be neglected, and it is easy to show that for any G such that $G_1\ll G<\kappa$, the most slowly decaying term in (11) con tributing to ρ_G has the same falloff as given in (16), with G_1 replaced by G, to lowest order in $(G^2/4\kappa^2)$. Thus the decay constant (16) seems to have rather general validity.

These results may be used to define an effective real-space resolution for STM. The suppression of the Fourier term for $G\neq 0$ by a factor $\exp(-\frac{1}{4}\kappa^{-1}G^2z)$ is precisely the effect of averaging over a Gaussian resolution function of rms width $(0.5\kappa^{-1}z)^{1/2}$, i.e., full width at half maximum $1.66(\kappa^{-1}z)^{1/2}$. Recall that for the relevant contours, $z=R+d$; if $R\gg d$, the resolution is determined by the tip radius but is nonetheless much smaller than R, since $\kappa^{-1}<1$ Å. For $R\ll d$, as in the case of a single-atom effective tip, the resolution is limited by d and, therefore, by how small a tunneling resistance is experimentally feasible. Note, however, that reducing d from 6 to 4 Å requires a decrease in tunneling voltage, or an increase in current, by roughly a factor of 200, and yet gives only a 20% increase in resolution.

In the plane of the surface atoms, $z=0$, $\rho(r)$ is rather localized within the unit cell. It is reasonable, therefore, to assume $\rho_0(z=0)\approx\rho_G(z=0)$, as long as there is only one atom per surface cell. Then at large z the corrugation (19) becomes

$$\Delta\approx 2\kappa^{-1}e^{-\beta z} . \tag{20}$$

This crude approximation, in fact, gives a good semiquantitative description of the results of the self-consistent calculations for Au(110) described below. A more systematic (though similarly crude) prescription for approximating $\rho(r,E_F)$ is suggested below.

C. Assessment of approximations

We now return to the theory developed in Sec. II A and consider the accuracy and generality of the many approximations made there.

The most crucial point is that rather large errors can be tolerated in the *coefficient* in (9) and (10). A factor of $e^2\approx 7$ error in the coefficient shifts the inferred value of $\vec{r}_0$ for a given current by only $\kappa^{-1}<1$ Å. The corresponding change in the corrugation Δ is, using (19), roughly a factor of $\exp(-\frac{1}{4}G^2/\kappa^2)\approx 0.92$ for Au(110) 2×1 ($G\approx 0.8$ Å^{-1}), an error under 10%. The substitution of typical metallic values in (10) is thus quite adequate.

As mentioned above, the use of an s-wave tip wave function is adequate if the real wave functions are restricted to small angular momentum l. For a sufficiently large effective tip the approximation is expected to lose its detailed validity. In any case, the s-wave treatment here is not intended as an accurate description of a real tip, but rather as a useful way of parametrizing the effect of finite tip size, which is otherwise relatively intractible.

In Sec. II A we implicitly assumed that the potential goes to zero in a region between the surface and tip and that the integral (3) is taken in that region. Actually, the electron is never more than about 3–6 Å from the surface, so the magnitude of the potential is never less than ~ 1 eV. Locally the effective value of κ is

$$\kappa(\vec{r})=\hbar^{-1}(2m)^{1/2}[\phi+V(\vec{r})]^{1/2} .$$

The resulting modest ($\sim 10\%$) change in κ is unimportant

except as it affects the wave function [as opposed to the coefficient in (9)]. The surface wave functions are calculated using the full potential, so we anticipate no problems *except* that we neglect the contribution of the tip to the potential. Without a more precise description of the real tip, we see no way to incorporate the tip potential in a consistent fashion.

(The local-density potential and thin slab geometries used in Sec. III below may be a source of inaccuracy in the implementation. However, such problems of *implementation* are a separate issue from the *intrinsic* limitations of the model presented here.)

When the tip and surface have different work functions, the resulting potential gradient causes a smooth variation in the effective $\kappa(\vec{r})$ between surface and tip. Again, as long as the difference is not great (<1 eV), we expect no crucial changes.

While the nominal current density may be quite large, the nanoamp current used corresponds to one electron per 1.6×10^{-10} sec. This is long compared to relevant transit times[11] as well as phonon vibration and relaxation times. The electrons may therefore be viewed as tunneling one at a time, and effects such as space charge and sample heating should be negligible.

We conclude that the approximations of Sec. II A are adequate for a quantitative understanding of STM, within the constraint of nearly complete ignorance of the microscopic structure of the tip. We hope that in the future improved characterization of the tip will permit a better evaluation of the model presented here.

Some other theoretical treatments of STM have been reported recently.[7,8] Garcia *et al.*[7] have applied methods developed for atom diffraction to calculate the current between two periodic metal surfaces. One surface is taken as strongly corrugated, and represents the tip. The potential is taken as flat throughout, with abrupt discontinuities at the two surfaces. A major drawback of this approach is that it is strictly numerical, and gives no direct insight or intuition into what is measured in STM. Moreover the quantitative results must be viewed with some caution. The model form for the potential is somewhat arbitrary; and, more important, the tip is treated in a peculiar fashion. The tip is apparently taken as infinitely extended in one dimension; moreover it is periodically repeated, which could give rise to irrelevant and unphysical interference effects. The model of Garcia *et al.* is appropriate for studying qualitative aspects of vacuum tunneling, but it is not clear how it could be usefully applied to aid in making structural inferences from experimental images. We prefer the approach taken here because it is at once more quantitative and more conceptually transparent.

Feuchtwang *et al.*[8] have pointed out that, instead of assuming a specific form for the tip wave function, one may represent the current as a convolution of spectral functions of the surface and tip. This suggests a possible avenue of investigation, but may not be applicable to imaging in the constant current mode (the only mode of STM imaging now considered practical). In any case the spectral function of the tip is not known, so implementing this approach would probably require approximations similar in spirit to those here. We prefer to retain the explicit asymmetry between surface and tip, reflecting the asymmetry both in our interest and in our understanding of the two.

III. AN EXAMPLE: Au(110)

In this section we describe a self-consistent calculation for the Au(110) surface, and compare our results with recent measurements of Binnig *et al.*[4] Agreement is excellent if a tip radius of 9 Å is assumed. We also discuss factors limiting the accuracy of the calculation and conclude that such calculations are feasible for relatively few systems of interest for STM. In the next section we consider an alternative approach which is less reliable but is feasible even for extremely complex systems.

The Au(110) surface normally exhibits 2×1 reconstruction with a missing-row geometry.[12] A 3×1 reconstruction has also been observed.[13] Recently Binnig *et al.*[4] reported high-resolution STM measurements for an Au(110) surface with regions of both 2×1 and 3×1 structure and concluded that the 3×1 structure consisted of (111) microfacets analogous to the 2×1. Measured STM corrugations were 0.45 and 1.4 Å for 2×1 and 3×1, respectively. (The two phases occurred together and were measured in the same scan with the same tip, permitting direct comparison.)

Since Au(110) is the only surface with a tractible unit-cell size for which high-resolution STM images are available, we have chosen it for detailed study in this section and in the next. We have calculated $\rho(\vec{r},E_F)$ for both 2×1 and 3×1 surfaces using a recently developed linearized augmented-plane-wave (LAPW) method described elsewhere.[14] For the 2×1 surface we used a slab geometry of three complete layers with a half layer [alternate $(1\bar{1}0)$ rows missing] on either side. The 3×1 geometry suggested by Binnig *et al.*[4] was employed; an asymmetric slab was constructed of two complete layers, a third layer with one missing row, and a fourth layer with two missing rows (see Fig. 2). The calculation is similar to that in Ref. 13, with $\rho(\vec{r},E_F)$ approximated by the charge in states within 0.5 eV of E_F, divided by the finite interval width of 1 eV.

Figure 2 shows the calculated $\rho(\vec{r},E_F)$ for Au(110). Since the actual tip geometry is not known, we choose a tip radius $R=9$ Å, so that (10) gives a (2×1) corrugation of 0.45 Å at tunneling resistance[9] 10^7 Ω to fit experiment. Then d is found to be 6 Å, measured from the surface Au nuclei to the edge of the tip potential well (i.e., the shell at which the tip wave function becomes decaying in character). This value is consistent[9] with experimental estimates of d based on resonant tunneling oscillations.[4] Given R, (10) yields a corrugation of 1.4 Å for the (3×1) surface, in excellent agreement with experiment.

The agreement here is gratifying; with one parameter, R, we obtain good agreement with two experimental corrugations, as well as with the gap distance. Nevertheless, it is worth briefly considering the numerical aspects of the calculation, which limit its accuracy.

In the surface LAPW method the wave functions are expanded in a Laue basis beyond the last plane of atoms, so the exponential decay poses no problems for simple surfaces. However, for the very "open" Au(110) surface,

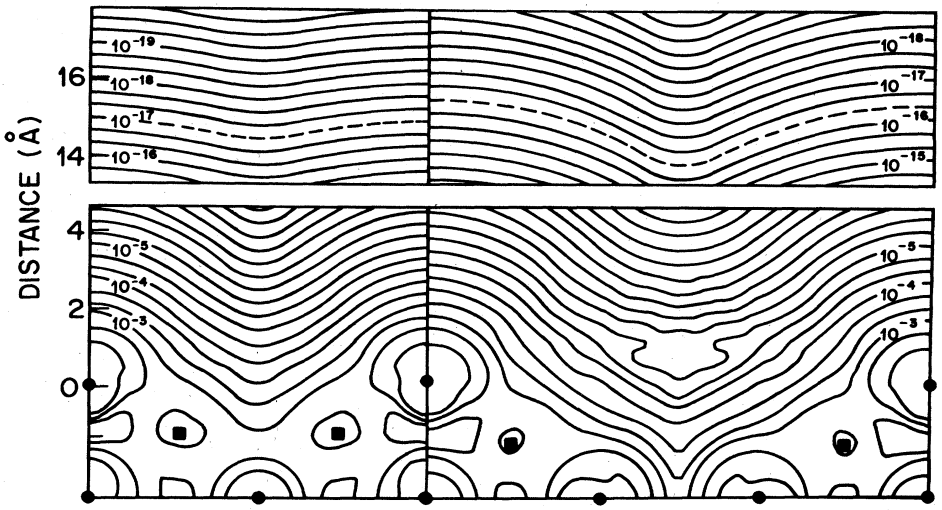

FIG. 2. Calculated $\rho(\vec{r},E_F)$ for Au(110) (2×1) (left) and (3×1) (right) surfaces. Figure shows $(1\bar{1}0)$ plane through outermost atoms. Positions of nuclei are indicated by solid circles (in plane) and squares (out of plane). Contours of constant $\rho(\vec{r},E_F)$ are labeled in units of a.u.$^{-3}$eV^{-1}. Note break in distance scale. Peculiar structure around contour 10^{-5} of (3×1) is due to limitations of the plane-wave part of the basis in describing the exponential decay inside the deep troughs. Center of curvature of probe tip follows dashed line.

we are obliged to expand the wave functions in a plane-wave basis in the "trough" region where surface atoms are missing. Since the wave functions decay exponentially there, the expansion converges slowly. The persistence of Gibbs's oscillations in the charge (Fig. 2) suggests that the convergence is still imperfect, but the 400-plane-wave expansion is the maximum possible with a CRAY-1 computer and our current code. (Some improvement could be be obtained by taking advantage of inversion and reflection symmetry for a suitable slab geometry.)

The other major source of inaccuracy is the very thin "slab" geometry employed. This might result in an inaccurate work function, which would certainly affect the results to some extent. The calculated work functions are 5.7 and 5.2 eV for 2×1 and 3×1 surfaces, respectively. Also, the thin slab gives only a few discrete states for a given wave vector. This sparse sampling of the bulk continuum leads to a numerical noise in energy-projected quantities. For this reason we included states from a rather large (1-eV) interval to approximate the charge $\rho(\vec{r},E_F)$ from states at E_F.

The local-density approximation used here does not reproduce the correct image form of the correlation potential at large distances from the surface; presumably it also gives incorrect lateral structure in the correlation potential in this region. Neither of these shortcomings has a significant effect on the results, however. The "crossover" from the high-density regime to the image-potential regime occurs well outside the classical turning point for electrons at the Fermi level, where the potential is small compared to the (negative) kinetic energy. The structure in the wave functions is determined by the strong potential near the atom cores, and the evolution of the wave functions at large distances from the surface is determined primarily by kinetic energy, as discussed above.

None of these sources of inaccuracy can be expected to greatly alter the results obtained; the calculation certainly gives a good overall representation of the true $\rho(\vec{r},E_F)$. Nevertheless, the accumulation of numerical uncertainties dictates some care in drawing conclusions. In the analysis above, d was determined rather directly by (10), since the dependence of current upon R largely cancels as noted above. However, R was inferred by fitting the experimental corrugation, which depends on $R+d$, and subtracting d. The corrugation is more susceptible to errors, both experimental and theoretical, than is the current. Moderate errors ($\sim20\%$) in either the calculated or measured corrugation amplitude have little effect on our conclusions. Nevertheless, since this is the first such calculation for STM, we believe it would be premature to rule out a tip consisting in effect of one or two atoms. For a sufficiently small cluster of atoms, the effective value of R depends on the precise geometry.

IV. APPROXIMATE METHODS FOR STM

The unique strength of STM is that it is a truly local real-space probe of surface geometry. As such it can resolve isolated steps, defects, and impurities. The direct computation of electronic structure for such nonperiodic structures is not, in general, feasible. Conversely, STM provides little information for relatively smooth low-Miller-index surfaces, the only kind which are presently susceptible to accurate calculation of the vacuum charge. It is therefore imperative that methods for treating more complex structures be developed, if the theoretical analysis of STM results is to progress. Such methods need not be highly accurate to be useful.

A. Atom superposition for $\rho(\vec{r},E_F)$

The calculated $\rho(\vec{r},E_F)$ in Fig. 2 bears a strong resemblance to total charge densities, which have been em-

ployed in understanding helium scattering.[14,15] It is known[15,16] that the charge is sometimes well approximated by the superposition of atom charge densities,

$$\rho(r)=\sum_R \phi(\vec{r}-\vec{R}) , \tag{21}$$

where $\phi(r)$ is the charge density of the free atom, and R are atom positions, which need not form a periodic lattice. While this approximation has never been tested at the large distances relevant for STM, we show below that it is well worth trying.

A natural next assumption is

$$\rho(\vec{r},E_F)\approx\rho(\vec{r})/E_0 , \tag{22}$$

where $\rho(\vec{r})$ is the total charge. To estimate E_0, we write

$$\rho(\vec{r},E)\sim A\exp[-\hbar^{-1}(2mE)^{1/2}z] , \tag{23}$$

where the variation of A with E is assumed small over the range contributing to $\rho(\vec{r})$. Then using

$$\rho(\vec{r})=\int_{-\infty}^{E_F}\rho(\vec{r},E)dE$$

and evaluating the integral, we find $E_0\approx E_F/\kappa z$. (The derivation assumes $\kappa z\gg 1$, so $E^{1/2}$ can be expanded about E_F.) For example, if $\kappa z\approx 10$, then $E_0\approx\frac{1}{2}$ eV. The precise value is unimportant, as discussed in Sec. II C. Note that the most drastic assumption is not (22), but rather the use of (21) at distances so great that only states near E_F contribute.

The use of (21) and (22), however crude, is not totally without justification. If the atom wave function is $\phi(r)\sim e^{-\kappa r}$, it is not hard to show[16] that the asymptotic decay length of the corrugation is identical to (16) and (19), to first order in $(G/2\kappa)^2$. For Au the atom eigenvalue is close to the work function in the local density approximation, and so κ for the atom and for the surface are almost the same. Were this not the case, one could replace the true atom charge with a model charge having the decay length appropriate for the surface. The fractional error in the atom superposition estimate of the corrugation therefore approaches an asymptotic value rather than growing without bound for large z. The decay length for the charge is correct by construction, so the method gives an excellent estimate of the gap distance d.

Intuitively one expects the greatest success for noble metals such as Au, where directional bonding is minimal. In other cases, $\rho(\vec{r},E_F)$ may show marked electronic structure effects, even when the total charge does not. For interesting examples, see Ref. 17 and Sec. V below.

B. Comparison with self-consistent results

We have repeated the calculation of Sec. III using the atom superposition approximation described above. Assuming as before a 9-Å tip radius, the calculated 2×1 and 3×1 corrugations are 0.30 and 0.93 Å, both about 30% less than in the self-consistent calculation. (While the accuracy of that calculation could not be calibrated quantitatively, better agreement would probably be fortuitous in any case.) This level of accuracy is enough to permit semiquantitative estimates of the STM image to be expected for a given atomic geometry.

We can also compare these atomic results directly with experiment, as we did in Sec. III. Using R as a fitting parameter as before, we obtain excellent agreement with both of the two measured corrugation amplitudes, and a gap of $d=6$ Å as before, by assuming $R=4$ Å. Thus the atom superposition calculation is entirely consistent with the experimental data but leads to a different (and presumably less reliable) conclusion regarding the magnitude of R.

C. Structure sensitivity of STM

We are now in a position to calculate conveniently (albeit crudely) the STM image for an arbitrary geometry. By comparing the images expected from different geometries, we can judge what structural conclusion can (or cannot) be drawn from experimental data.

As the first example, we consider the Au(110) 3×1 surface. Binnig *et al.*[4] inferred a geometry with two rows missing in the first layer and one in the second layer (see Fig. 2), to account for the deep observed corrugation of the 3×1 surface (1.4 Å versus 0.45 Å for the 2×1). While this inference is quite reasonable, we consider now a more quantitative test. We have calculated $\rho(\vec{r},E_F)$ with approximations (21) and (22) for two 3×1 geometries, one with and one without a row missing in the second layer. The results are shown in Fig. 3. While the

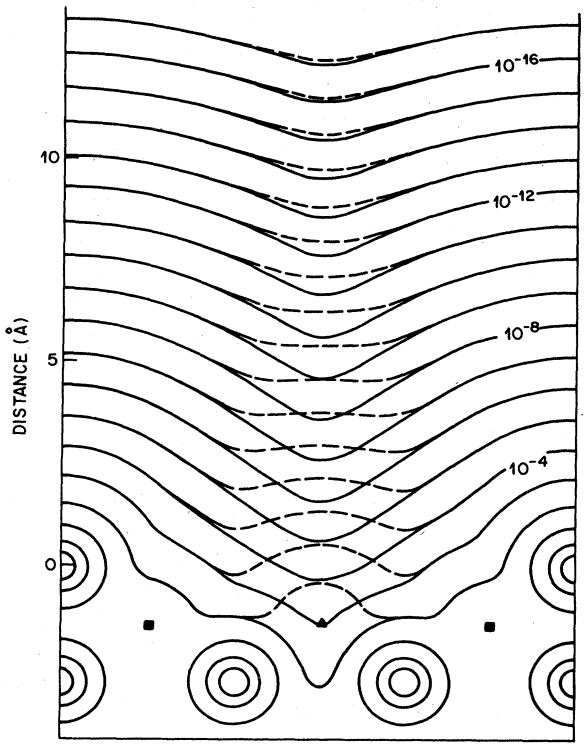

FIG. 3. Atom superposition charge density (a.u.$^{-3}$) for two possible geometries of Au(110) 3×1. Solid lines are for same geometry as in Fig. 2. Dashed lines are for geometry with no atom missing in second layer. Triangle shows site of atom (out of plane of figure) present only in latter case; squares show position of other out-of-plane atoms.

charge densities look radically different close to the surface, at a distance of 10 Å (where the corrugation is 1.4 Å) the corrugation amplitudes differ by only 15%; at 15 Å (appropriate for a 9-Å tip) the corrugation amplitudes differ by less than 5%. Realistically, even the 15% difference is far too little to reliably distinguish the two geometries. The greater corrugation in the 3×1 case (as compared to the 2×1) is attributable entirely to the greater surface lattice constant, which permits clearer resolution of the peaks and troughs. According to (19), a smaller surface lattice constant (as for the 2×1) results in an exponentially smaller corrugation at large distances.

As another sensitivity test, we compared the charge for the 2×1 surface to that for the (half-filled) first layer alone. At distances greater than 8 or 9 Å, the removal of the second and all subsequent layers has no noticeable effect. While atom superposition neglects the electronic changes for such a monolayer, the result at least tells us that the STM data carry no useful information whatever on the position of the first layer relative to the underlying substrate. A more methodical study of the relationship between geometry and charge density (and hence the STM image), also within the atom superposition approximation, is presented by Tersoff *et al.*[16]

V. SEMICONDUCTOR SURFACES

As noted above, for small voltages the tunneling occurs between states at E_F. At semiconducting surfaces, E_F lies near the conduction- or valence-band edge, depending on whether the doping is n or p type. The character of the states at E_F, and hence the form of $\rho(\vec{r},E_F)$, may be drastically different for these two cases, giving correspondingly different STM images.

For low doping or high voltages, the voltage polarity rather than the doping may determine whether tunneling involves valence or conduction states. In the one reported example, Binnig *et al.*[3] found that a measurable tunneling current for the Si(111) surface required a large voltage, over 2.5 V. These measurements were repeated with heavily doped Si samples, however, and comparable STM images were obtained with voltages in the 10-meV range used for Au.[9] The high voltage in the first instance was probably developed across a non-Ohmic contact, a surface barrier due to band-bending, or both. We now have no reason to believe that the tunneling conductance in the STM imaging regime is significantly different for semiconductors and metals.

One of the simplest semiconductor surfaces is the cleaved GaAs(110) surface. The geometry of the 1×1 reconstruction is reasonably well established, and there are known to be no surface states in the band gap. We therefore use this surface to illustrate the difference between expected STM images for tunneling involving valence and conduction states. Figure 4 shows $\rho(\vec{r},E_F)$ calculated for the GaAs(110) surface, based on the charge in states within 1 eV of the respective band edges. The total charge density is also shown. Far from the surface, the valence-edge charge looks quite similar to the total charge density. The charge is concentrated on the As atoms, which are raised above the Ga by the reconstruction. As a result, the image is well approximated by a superposition of As atom charge densities.

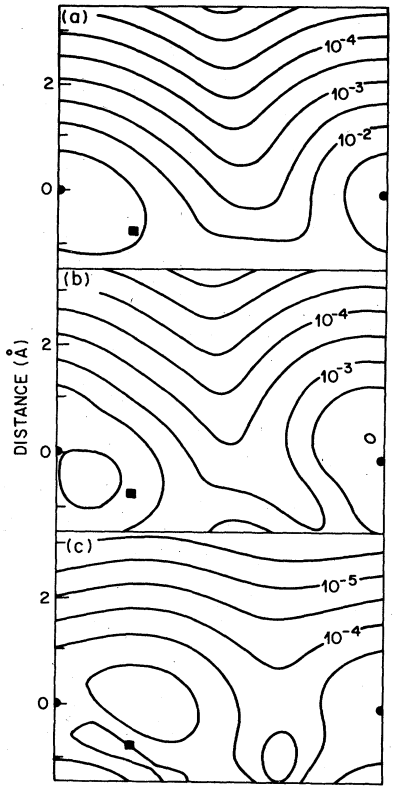

FIG. 4. Projected charge densities at the GaAs(110) surface, in a $(1\bar{1}0)$ plane midway between the Ga and As atoms, in units of bohr^{-3}. The vacuum charge density is much smoother in the direction perpendicular to the figure. The three panels show (a) total charge; (b) charge in states within 1 eV of the valence-band edge; (c) charge in states within 1 eV of the conduction-band edge. Positions of the surface atoms, projected into the plane of the figure, are given by circles (As) and squares (Ga). Horizontal direction is $\langle 001 \rangle$, vertical is $\langle 110 \rangle$.

The conduction-band charge, however, looks quite different. Charge is concentrated on the Ga atoms; but these are lower than the As, with the net effect that both contribute comparably to the vacuum charge. The total corrugation is thus much smaller than for the valence charge, with the charge density peaking weakly above the Ga sites.

The surface lattice constant of GaAs(110) is less than 6 Å, which may be beyond the power of STM to resolve. However, the qualitative difference predicted for valence and conduction-band tunneling here should be observable in a wide variety of semiconducting surfaces.

VI. CONCLUSION

We have presented a simple theory for STM, which includes fully the detailed electronic structure of the surface and yet is computationally tractable. The tunneling current is found to be proportional to the surface LDOS at the position of the tip. The approximations made appear to introduce relatively little inaccuracy, *except* that

the tip is treated in a model way; even this approximation probably cannot be significantly improved until the microscopic structure of the tip is better understood.

The theory provides explicit expressions for the intrinsic spatial resolution and for the dependence of the tunneling current on tip size and position. When applied to the Au(110) surface, the theory agreed well with experiment. Moreover, the accuracy appeared to be limited by the computational implementation rather than by intrinsic factors.

Motivated both by the difficulty in carrying out accurate electronic structure calculations for STM, and by the need for a local technique for treating nonperiodic structures now observable with STM, we have proposed a simple approximate technique based on atom superposition, analogous to methods often used for the helium-diffraction problem. Despite its crudeness, this method has some analytic justification in terms of its asymptotic behavior, and gives good results for Au(110). It also provides a convenient way to test the sensitivity of STM to details of the surface structure, such as the presence or absence of a missing row in the second layer of Au(110) 3×1. We conclude that STM is [at least for the Au(110) surface] quite insensitive to positions of atoms beyond the first layer.

Finally, we have considered in a qualitative way the novel effects which may be observable in tunneling to semiconductor surfaces, where valence and conduction states have very different charge distributions. For the Au(110) surface, the LDOS (and hence the image) reflected surface topography in a relatively straightforward way. For GaAs(110), on the other hand, this was true only if tunneling involved the valence band (as for tunneling out of *p*-type GaAs). For conduction-band tunneling (into *n*-type GaAs), the image bore no simple relation to the topography, since the (lower) Ga atoms were emphasized by electronic structure effects.

We have concentrated on the use of STM to determine surface atomic structure, since this has been the main application to date. However, the fact that STM really measures the surface LDOS may be exploited in the future to give novel information about not only the topography of surfaces, but their electronic structure as well.

ACKNOWLEDGMENT

We are grateful to H. Rohrer and A. Baratoff for stimulating discussions.

---

*Present address: IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

[1]G. Binnig, H. Rohrer, Ch. Gerber, and E. Weibel, Appl. Phys. Lett. **40**, 178 (1982).

[2]G. Binnig, H. Rohrer, Ch. Gerber, and E. Weibel, Phys. Rev. Lett. **49**, 57 (1982); G. Binnig and H. Rohrer, Surf. Sci. **126**, 236 (1983).

[3]G. Binnig, H. Rohrer, Ch. Gerber, and E. Weibel, Phys. Rev. Lett. **50**, 120 (1983).

[4]G. Binnig, H. Rohrer, Ch. Gerber, and E. Weibel, Surf. Sci. **131**, L379 (1983).

[5]C. B. Duke, *Tunneling in Solids,* Suppl. 10 of *Solid State Physics,* edited by F. Seitz and D. Turnbull (Academic, New York, 1969), p. 1.

[6]J. Tersoff and D. R. Hamann, Phys. Rev. Lett. **50**, 1998 (1983).

[7]N. Garcia, C. Ocal, and F. Flores, Phys. Rev. Lett. **50**, 2002 (1983).

[8]T. E. Feuchtwang, P. H. Cutler, and N. M. Miskovsky, Phys. Lett. **99A**, 167 (1983).

[9]H. Rohrer (private communication).

[10]J. Bardeen, Phys. Rev. Lett. **6**, 57 (1961).

[11]T. E. Hartman, J. Appl. Phys. **33**, 3427 (1962).

[12]I. K. Robinson, Phys. Rev. Lett. **50**, 1145 (1983), and references therein; L. D. Marks and D. J. Smith, Nature (London) **303**, 316 (1983).

[13]W. Moritz and D. Wolf, Surf. Sci. **88**, L29 (1979); Y. Kuk, Bull. Am. Phys. Soc. **28**, 260 (1983), and unpublished.

[14]D. R. Hamann, Phys. Rev. Lett. **46**, 1227 (1981).

[15]R. B. Laughlin, Phys. Rev. B **25**, 2222 (1982); M. Manninen, J. F. Nørskov, and C. Umrigar, Surf. Sci. **119**, L393 (1982).

[16]J. Tersoff, M. J. Cardillo, and D. R. Hamann (unpublished).

[17]P. J. Feibelman and D. R. Hamann, Phys. Rev. Lett. **52**, 61 (1984).