Appendix A

Time–Independent Perturbation Theory

References

- Davydov - *Quantum Mechanics*, Ch. 7.
- Morse and Feshbach, *Methods of Theoretical Physics*, Ch. 9.
- Shankar, *Principles of Quantum Mechanics*, Ch. 17.
- Cohen-Tannoudji, Diu and Laloë, *Quantum Mechanics*, vol. 2, Ch. 11.
- T-Y. Wu, Quantum Mechanics, Ch. 6.

A.1 Introduction

Another review topic that we discuss here is time–independent perturbation theory because of its importance in experimental solid state physics in general and transport properties in particular.

There are many mathematical problems that occur in nature that cannot be solved exactly. It also happens frequently that a *related* problem can be solved *exactly*. Perturbation theory gives us a method for relating the problem that can be solved exactly to the one that cannot. This occurrence is more general than quantum mechanics –many problems in electromagnetic theory are handled by the techniques of perturbation theory. In this course however, we will think mostly about quantum mechanical systems, as occur typically in solid state physics.

Suppose that the Hamiltonian for our system can be written as

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}' \tag{A.1}$$

where $\mathcal{H}_0$ is the part that we can solve exactly and $\mathcal{H}'$ is the part that we cannot solve. Provided that $\mathcal{H}' \ll \mathcal{H}_0$ we can use perturbation theory; that is, we consider the solution of the unperturbed Hamiltonian $\mathcal{H}_0$ and then calculate the effect of the perturbation Hamiltonian $\mathcal{H}'$. For example, we can solve the hydrogen atom energy levels exactly, but when we apply an electric or a magnetic field, we can no longer solve the problem exactly. For

this reason, we treat the effect of the external fields as a perturbation, provided that the energy associated with these fields is small:

$$\mathcal{H} = \frac{p^2}{2m} - \frac{e^2}{r} - e\vec{r}\cdot\vec{E} = \mathcal{H}_0 + \mathcal{H}' \tag{A.2}$$

where

$$\mathcal{H}_0 = \frac{p^2}{2m} - \frac{e^2}{r} \tag{A.3}$$

and

$$\mathcal{H}' = -e\vec{r}\cdot\vec{E}. \tag{A.4}$$

As another illustration of an application of perturbation theory, consider a weak periodic potential in a solid. We can calculate the free electron energy levels (empty lattice) exactly. We would like to relate the weak potential situation to the empty lattice problem, and this can be done by considering the weak periodic potential as a perturbation.

A.1.1 Non-degenerate Perturbation Theory

In non-degenerate perturbation theory we want to solve Schrödinger's equation

$$\mathcal{H}\psi_n = E_n\psi_n \tag{A.5}$$

where

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}' \tag{A.6}$$

and

$$\mathcal{H}' \ll \mathcal{H}_0. \tag{A.7}$$

It is then assumed that the solutions to the unperturbed problem

$$\mathcal{H}_0\psi_n^0 = E_n^0\psi_n^0 \tag{A.8}$$

are known, in which we have labeled the unperturbed energy by E_n^0 and the unperturbed wave function by ψ_n^0. By *non-degenerate* we mean that there is only one eigenfunction ψ_n^0 associated with each eigenvalue E_n^0.

The wave functions ψ_n^0 form a complete orthonormal set

$$\int \psi_n^{*0}\psi_m^0 d^3r = \langle\psi_n^0|\psi_m^0\rangle = \delta_{nm}. \tag{A.9}$$

Since $\mathcal{H}'$ is small, the wave functions for the total problem ψ_n do not differ greatly from the wave functions ψ_n^0 for the unperturbed problem. So we expand $\psi_{n'}$ in terms of the complete set of ψ_n^0 functions

$$\psi_{n'} = \sum_n a_n\psi_n^0. \tag{A.10}$$

Such an expansion can always be made; that is no approximation. We then substitute the expansion of Eq. A.10 into Schrödinger's equation (Eq. A.5) to obtain

$$\mathcal{H}\psi_{n'} = \sum_n a_n(\mathcal{H}_0 + \mathcal{H}')\psi_n^0 = \sum_n a_n(E_n^0 + \mathcal{H}')\psi_n^0 = E_{n'}\sum_n a_n\psi_n^0 \tag{A.11}$$

and therefore we can write

$$\sum_n a_n(E_{n'} - E_n^0)\psi_n^0 = \sum_n a_n \mathcal{H}'\psi_n^0. \tag{A.12}$$

If we are looking for the perturbation to the level m, then we multiply Eq. A.12 from the left by ψ_m^{0*} and integrate over all space. On the left hand side of Eq. A.12 we get $\langle\psi_m^0|\psi_n^0\rangle = \delta_{mn}$ while on the right hand side we have the matrix element of the perturbation Hamiltonian taken between the unperturbed states:

$$a_m(E_{n'} - E_m^0) = \sum_n a_n\langle\psi_m^0|\mathcal{H}'|\psi_n^0\rangle \equiv \sum_n a_n\mathcal{H}'_{mn} \tag{A.13}$$

where we have written the indicated matrix element as $\mathcal{H}'_{mn}$. Equation A.13 is an iterative equation on the a_n coefficients, where each a_m coefficient is related to a complete set of a_n coefficients by the relation

$$a_m = \frac{1}{E_{n'} - E_m^0}\sum_n a_n\langle\psi_m^0|\mathcal{H}'|\psi_n^0\rangle = \frac{1}{E_{n'} - E_m^0}\sum_n a_n\mathcal{H}'_{mn} \tag{A.14}$$

in which the summation includes the $n = n'$ and m terms. We can rewrite Eq. A.14 to involve terms in the sum $n \neq m$

$$a_m(E_{n'} - E_m^0) = a_m\mathcal{H}'_{mm} + \sum_{n\neq m} a_n\mathcal{H}'_{mn} \tag{A.15}$$

so that the coefficient a_m is related to all the other a_n coefficients by:

$$a_m = \frac{1}{E_{n'} - E_m^0 - \mathcal{H}'_{mm}}\sum_{n\neq m} a_n\mathcal{H}'_{mn} \tag{A.16}$$

where n' is an index denoting the energy of the state we are seeking. The equation (A.16) written as

$$a_m(E_{n'} - E_m^0 - \mathcal{H}'_{mm}) = \sum_{n\neq m} a_n\mathcal{H}'_{mn} \tag{A.17}$$

is an identity in the a_n coefficients. If the perturbation is small then $E_{n'}$ is very close to E_m^0 and the first order corrections are found by setting the coefficient on the right hand side equal to zero and $n' = m$. The next order of approximation is found by substituting for a_n on the right hand side of Eq. A.17 and substituting for a_n the expression

$$a_n = \frac{1}{E_{n'} - E_n^0 - \mathcal{H}'_{nn}}\sum_{n''\neq n} a_{n''}\mathcal{H}'_{nn''} \tag{A.18}$$

which is obtained from Eq. A.16 by the transcription $m \to n$ and $n \to n''$. In the above, the energy level $E_{n'} = E_m$ is the level for which we are calculating the perturbation. We now look for the a_m term in the sum $\sum_{n''\neq n} a_{n''}\mathcal{H}'_{nn''}$ of Eq. A.18 and bring it to the left hand side of Eq. A.17. If we are satisfied with our solutions, we end the perturbation calculation at this point. If we are not satisfied, we substitute for the $a_{n''}$ coefficients in Eq. A.18 using the same basic equation as Eq. A.18 to obtain a triple sum. We then select out the a_m term, bring it to the left hand side of Eq. A.17, etc. This procedure gives us an easy recipe to find the energy in Eq. A.11 to any order of perturbation theory. We now write these iterations down more explicitly for first and second order perturbation theory.

1^{st} Order Perturbation Theory

In this case, no iterations of Eq. A.17 are needed and the sum $\sum_{n\neq m} a_n \mathcal{H}'_{mn}$ on the right hand side of Eq. A.17 is neglected, for the reason that if the perturbation is small, $\psi_{n'} \sim \psi^0_n$. Hence only a_m in Eq. A.10 contributes significantly. We merely write $E_{n'} = E_m$ to obtain:

$$a_m(E_m - E^0_m - \mathcal{H}'_{mm}) = 0. \tag{A.19}$$

Since the a_m coefficients are arbitrary coefficients, this relation must hold for all a_m so that

$$(E_m - E^0_m - \mathcal{H}'_{mm}) = 0 \tag{A.20}$$

or

$$E_m = E^0_m + \mathcal{H}'_{mm}. \tag{A.21}$$

We write Eq. A.21 even more explicitly so that the energy for state m for the perturbed problem E_m is related to the unperturbed energy E^0_m by

$$E_m = E^0_m + \langle \psi^0_m | \mathcal{H}' | \psi^0_m \rangle \tag{A.22}$$

where the indicated diagonal matrix element of $\mathcal{H}'$ can be integrated as the average of the perturbation in the state ψ^0_m. The wave functions to lowest order are not changed

$$\psi_m = \psi^0_m. \tag{A.23}$$

2^{nd} order perturbation theory

If we carry out the perturbation theory to the next order of approximation, one further iteration of Eq. A.17 is required:

$$a_m(E_m - E^0_m - \mathcal{H}'_{mm}) = \sum_{n\neq m} \frac{1}{E_m - E^0_n - \mathcal{H}'_{nn}} \sum_{n''\neq n} a_{n''} \mathcal{H}'_{nn''} \mathcal{H}'_{mn} \tag{A.24}$$

in which we have substituted for the a_n coefficient in Eq. A.17 using the iteration relation given by Eq. A.18. We now pick out the term on the right hand side of Eq. A.24 for which $n'' = m$ and bring that term to the left hand side of Eq. A.24. If no further iteration is to be done, we throw away what is left on the right hand side of Eq. A.24 and get an expression for the arbitrary a_m coefficients

$$a_m \left[(E_m - E^0_m - \mathcal{H}'_{mm}) - \sum_{n\neq m} \frac{\mathcal{H}'_{nm}\mathcal{H}'_{mn}}{E_m - E^0_n - \mathcal{H}'_{nn}} \right] = 0. \tag{A.25}$$

Since a_m is arbitrary, the term in square brackets in Eq. A.25 vanishes and the second order correction to the energy results:

$$E_m = E^0_m + \mathcal{H}'_{mm} + \sum_{n\neq m} \frac{|\mathcal{H}'_{mn}|^2}{E_m - E^0_n - \mathcal{H}'_{nn}} \tag{A.26}$$

in which the sum on states $n \neq m$ represents the 2nd order correction.

To this order in perturbation theory we must also consider corrections to the wave function

$$\psi_m = \sum_n a_n \psi_n^0 = \psi_m^0 + \sum_{n \neq m} a_n \psi_n^0 \tag{A.27}$$

in which ψ_m^0 is the large term and the correction terms appear as a sum over all the other states $n \neq m$. In handling the correction term, we look for the a_n coefficients, which from Eq. A.18 are given by

$$a_n = \frac{1}{E_n' - E_n^0 - \mathcal{H}_{nn}'} \sum_{n'' \neq n} a_{n''} \mathcal{H}_{nn''}'. \tag{A.28}$$

If we only wish to include the lowest order correction terms, we will take only the most important term, i.e., $n'' = m$, and we will also use the relation $a_m = 1$ in this order of approximation. Again using the identification $n' = m$, we obtain

$$a_n = \frac{\mathcal{H}_{nm}'}{E_m - E_n^0 - \mathcal{H}_{nn}'} \tag{A.29}$$

and

$$\psi_m = \psi_m^0 + \sum_{n \neq m} \frac{\mathcal{H}_{nm}' \psi_n^0}{E_m - E_n^0 - \mathcal{H}_{nn}'}. \tag{A.30}$$

For homework, you should do the next iteration to get 3rd order perturbation theory, in order to see if you really have mastered the technique (this will be an optional homework problem).

Now look at the results for the energy E_m (Eq. A.26) and the wave function ψ_m (Eq. A.30) for the 2nd order perturbation theory and observe that these solutions are implicit solutions. That is, the correction terms are themselves dependent on E_m. To obtain an explicit solution, we can do one of two things at this point.

1. We can ignore the fact that the energies differ from their unperturbed values in calculating the correction terms. This is known as Raleigh-Schrödinger perturbation theory. This is the usual perturbation theory given in Quantum Mechanics texts and for homework you may review the proof as given in these texts.

2. We can take account of the fact that E_m differs from E_m^0 by calculating the correction terms by an iteration procedure; the first time around, you substitute for E_m the value that comes out of 1st order perturbation theory. We then calculate the second order correction to get E_m. We next take this E_m value to compute the new second order correction term etc. until a convergent value for E_m is reached. This iterative procedure is what is used in *Brillouin–Wigner* perturbation theory and is a better approximation than Raleigh-Schrödinger perturbation theory to both the wave function and the energy eigenvalue for the same order in perturbation theory.

The Brillouin–Wigner method is often used for practical problems in solids. For example, if you have a 2-level system, the Brillouin–Wigner perturbation theory to second order gives an exact result, whereas Rayleigh–Schrödinger perturbation theory must be carried out to infinite order.

Let us summarize these ideas. If you have to compute only a small correction by perturbation theory, then it is advantageous to use Rayleigh-Schrödinger perturbation theory

because it is much easier to use, since no iteration is needed. If one wants to do a more convergent perturbation theory (i.e., obtain a better answer to the same order in perturbation theory), then it is advantageous to use Brillouin–Wigner perturbation theory. There are other types of perturbation theory that are even more convergent and harder to use than Brillouin–Wigner perturbation theory (see Morse and Feshbach vol. 2). But these two types are the most important methods used in solid state physics today.

For your convenience we summarize here the results of the second–order non–degenerate Rayleigh-Schrödinger perturbation theory:

$$E_m = E_m^0 + \mathcal{H}'_{mm} + \sum_n{}' \frac{|\mathcal{H}'_{nm}|^2}{E_m^0 - E_n^0} + ... \tag{A.31}$$

$$\psi_m = \psi_m^0 + \sum_n{}' \frac{\mathcal{H}'_{nm}\psi_n^0}{E_m^0 - E_n^0} + \tag{A.32}$$

where the sums in Eqs. A.31 and A.32 denoted by primes exclude the $m = n$ term. Thus, Brillouin–Wigner perturbation theory (Eqs.A.26 and A.30) contains contributions in second order which occur in higher order in the Rayleigh-Schrödinger form. In practice, Brillouin–Wigner perturbation theory is useful when the perturbation term is too large to be handled conveniently by Rayleigh–Schrödinger perturbation theory, but still small enough for perturbation theory to work insofar as the perturbation expansion forms a convergent series.

A.1.2 Degenerate Perturbation Theory

It often happens that a number of quantum mechanical levels have the same or nearly the same energy. If they have exactly the same energy, we know that we can make any linear combination of these states that we like and get a new eigenstate also with the same energy. In the case of degenerate states, we have to do perturbation theory a little differently, as described in the following section.

Suppose that we have an f-fold degeneracy (or near-degeneracy) of energy levels

$$\underbrace{\psi_1^0, \psi_2^0, ...\psi_f^0}_{\text{states with the same or nearly the same energy}} \qquad \underbrace{\psi_{f+1}^0, \psi_{f+2}^0, ...}_{\text{states with quite different energies}}$$

We will call the set of states with the same (or approximately the same) energy a "nearly degenerate set" (NDS). In the case of degenerate sets, the iterative Eq. A.17 still holds. The only difference is that for the degenerate case we solve for the perturbed energies by a different technique, as described below.

Starting with Eq. A.17, we now bring to the left-hand side of the iterative equation all terms involving the f energy levels that are in the NDS. If we wish to calculate an energy within the NDS in the presence of a perturbation, we consider all the a_n's within the NDS as large, and those outside the set as small. To first order in perturbation theory, we ignore the coupling to terms outside the NDS and we get f linear homogeneous equations in the a_n's where $n = 1, 2, ...f$. We thus obtain the following equations from Eq. A.17:

$$\begin{array}{ccccc}
a_1(E_1^0 + \mathcal{H}'_{11} - E) & +a_2\mathcal{H}'_{12} & +... & +a_f\mathcal{H}'_{1f} & = 0 \\
a_1\mathcal{H}'_{21} & +a_2(E_2^0 + \mathcal{H}'_{22} - E) & +... & +a_f\mathcal{H}'_{2f} & = 0 \\
\vdots & \vdots & \ddots & \vdots & \\
a_1\mathcal{H}'_{f1} & +a_2\mathcal{H}'_{f2} & +\ldots & +a_f(E_f^0 + \mathcal{H}'_{ff} - E) & = 0.
\end{array} \tag{A.33}$$

In order to have a solution of these f linear equations, we demand that the coefficient determinant vanish:

$$\begin{vmatrix} (E_1^0 + \mathcal{H}'_{11} - E) & \mathcal{H}'_{12} & \mathcal{H}'_{13} & \dots & \mathcal{H}'_{1f} \\ \mathcal{H}'_{21} & (E_2^0 + \mathcal{H}'_{22} - E) & \mathcal{H}'_{23} & \dots & \mathcal{H}'_{2f} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{H}'_{f1} & \mathcal{H}'_{f2} & \dots & \dots & (E_f^0 + \mathcal{H}'_{ff} - E) \end{vmatrix} = 0 \tag{A.34}$$

The f eigenvalues that we are looking for are the eigenvalues of the matrix in Eq. A.34 and the set of orthogonal states are the corresponding eigenvectors. Remember that the matrix elements $\mathcal{H}'_{ij}$ that occur in the above determinant are taken between the unperturbed states in the NDS.

The generalization to second order degenerate perturbation theory is immediate. In this case, Eqs. A.33 and A.34 have additional terms. For example, the first relation in Eq. A.33 would then become

$$a_1(E_1^0 + \mathcal{H}'_{11} - E) + a_2\mathcal{H}'_{12} + a_3\mathcal{H}'_{13} + \ldots + a_f\mathcal{H}'_{1f} = -\sum_{n \neq NDS} a_n\mathcal{H}'_{1n} \tag{A.35}$$

and for the a_n in the sum in Eq. A.35, which are now small (because they are outside the NDS), we would use our iterative form

$$a_n = \frac{1}{E - E_n^0 - \mathcal{H}'_{nn}} \sum_{m \neq n} a_m \mathcal{H}'_{nm}. \tag{A.36}$$

But we must only consider the terms in the above sum which are large; these terms are all in the NDS. This argument shows that every term on the left side of Eq. A.35 will have a correction term. For example the correction term to a general coefficient a_i will look as follows:

$$a_i\mathcal{H}'_{1i} + a_i \sum_{n \neq NDS} \frac{\mathcal{H}'_{1n}\mathcal{H}'_{ni}}{E - E_n^0 - \mathcal{H}'_{nn}} \tag{A.37}$$

where the first term is the original term from 1st order degenerate perturbation theory and the term from states outside the NDS gives the 2nd order correction terms. So, if we are doing higher order degenerate perturbation theory, we write for each entry in the secular equation the appropriate correction terms (Eq. A.37) that are obtained from these iterations. For example, in 2nd order degenerate perturbation theory, the (1,1) entry to the matrix in Eq. A.34 would be

$$E_1^0 + \mathcal{H}'_{11} + \sum_{n \neq NDS} \frac{|\mathcal{H}'_{1n}|^2}{E - E_n^0 - \mathcal{H}'_{nn}} - E. \tag{A.38}$$

As a further illustration let us write down the (1,2) entry:

$$\mathcal{H}'_{12} + \sum_{n \neq NDS} \frac{\mathcal{H}'_{1n}\mathcal{H}'_{n2}}{E - E_n^0 - \mathcal{H}'_{nn}}. \tag{A.39}$$

Again we have an implicit dependence of the 2nd order term in Eqs. A.38 and A.39 on the energy eigenvalue that we are looking for. To do 2nd order degenerate perturbation we again

have two options. If we take the energy E in Eqs. A.38 and A.39 as the unperturbed energy in computing the correction terms, we have 2nd order degenerate Rayleigh-Schrödinger perturbation theory. On the other hand, if we iterate to get the best correction term, then we call it Brillouin–Wigner perturbation theory.

How do we know in an actual problem when to use degenerate 1st or degenerate 2nd order perturbation theory? If the matrix elements $\mathcal{H}'_{ij}$ coupling members of the NDS vanish, then we must go to 2nd order. Generally speaking, the first order terms will be much larger than the 2nd order terms, provided that there is no symmetry reason for the first order terms to vanish.

Let us explain this further. By the matrix element $\mathcal{H}'_{12}$ we mean $(\psi_1^0|\mathcal{H}'|\psi_2^0)$. Suppose the perturbation Hamiltonian $\mathcal{H}'$ under consideration is due to an electric field $\vec{E}$

$$\mathcal{H}' = -e\vec{r}\cdot\vec{E} \tag{A.40}$$

where $e\vec{r}$ is the dipole moment of our system. If now we consider the effect of inversion on $\mathcal{H}'$, we see that $\vec{r}$ changes sign under inversion $(x, y, z) \rightarrow -(x, y, z)$, i.e., $\vec{r}$ is an odd function. Suppose that we are considering the energy levels of the hydrogen atom in the presence of an electric field. We have s states (even), p states (odd), d states (even), etc. The electric dipole moment will only couple an even state to an odd state because of the oddness of the dipole moment under inversion. Hence there is no effect in 1^{st} order non–degenerate perturbation theory for situations where the first order matrix element vanishes. For the $n = 1$ level, there is, however, an effect due to the electric field in second order so that the correction to the energy level goes as the square of the electric field, i.e., $|\vec{E}|^2$. For the n =2 levels, we treat them in degenerate perturbation theory because the $2s$ and $2p$ states are degenerate in the simple treatment of the hydrogen atom. Here, first order terms only appear in entries coupling s and p states. To get corrections which split the p levels among themselves, we must go to 2nd order degenerate perturbation theory.

Appendix B

1D Graphite: Carbon Nanotubes

In this appendix we show how the tight binding approximation (§B.1.1) can be used to obtain an excellent approximation for the electronic structure of carbon nanotubes which are a one dimensional form of graphite obtained by rolling up a single sheet of graphite into a seamless cylinder. In this appendix the structure and the electronic properties of a single atomic sheet of 2D graphite and then discuss how this is rolled up into a cylinder, then describing the structure and properties of the nanotube using the tight binding approximation.

B.1 Structure of 2D graphite

Graphite is a three-dimensional (3D) layered hexagonal lattice of carbon atoms. A single layer of graphite, forms a two-dimensional (2D) material, called 2D graphite or a graphene layer. Even in 3D graphite, the interaction between two adjacent layers is very small compared with intra-layer interactions, and the electronic structure of 2D graphite is a first approximation of that for 3D graphite.

In Fig. B.1 we show (a) the unit cell and (b) the Brillouin zone of two-dimensional graphite as a dotted rhombus and shaded hexagon, respectively, where $\vec{a}_1$ and $\vec{a}_2$ are unit vectors in real space, and $\vec{b}_1$ and $\vec{b}_2$ are reciprocal lattice vectors. In the x, y coordinates shown in Fig. B.1, the real space unit vectors $\vec{a}_1$ and $\vec{a}_2$ of the hexagonal lattice are expressed as

$$\vec{a}_1 = \left(\frac{\sqrt{3}}{2}a, \frac{a}{2}\right), \quad \vec{a}_2 = \left(\frac{\sqrt{3}}{2}a, -\frac{a}{2}\right), \tag{B.1}$$

where $a = |\vec{a}_1| = |\vec{a}_2| = 1.42 \times \sqrt{3} = 2.46$Å is the lattice constant of two-dimensional graphite. Correspondingly the unit vectors $\vec{b}_1$ and $\vec{b}_2$ of the reciprocal lattice are given by:

$$\vec{b}_1 = \left(\frac{2\pi}{\sqrt{3}a}, \frac{2\pi}{a}\right), \quad \vec{b}_2 = \left(\frac{2\pi}{\sqrt{3}a}, -\frac{2\pi}{a}\right) \tag{B.2}$$

corresponding to a lattice constant of $4\pi/\sqrt{3}a$ in reciprocal space.

Three σ bonds for 2D graphite hybridize in a sp^2 configuration, while, and the other $2p_z$ orbital, which is perpendicular to the graphene plane, makes π covalent bonds. In Sect. B.1.1 we consider only the π energy bands for 2D graphite, because we know that the π energy bands are covalent and are the most important for determining the solid state properties of 2D graphite.

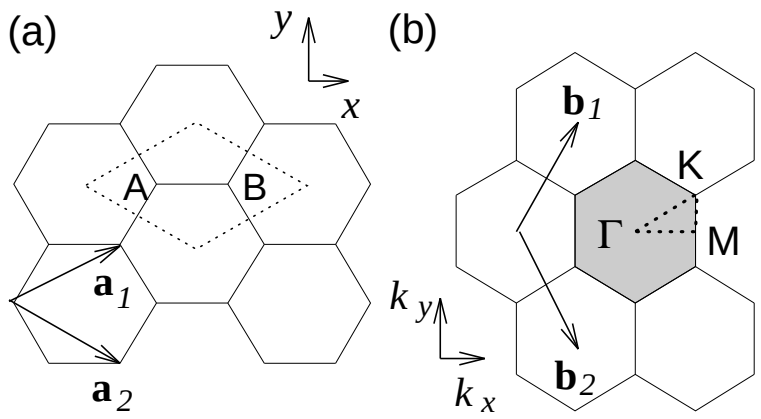


Figure B.1: (a) The unit cell and (b) Brillouin zone of two-dimensional graphite are shown as the dotted rhombus and shaded hexagon, respectively. $\vec{a}_i$, and $\vec{b}_i$, $(i = 1, 2)$ are unit vectors and reciprocal lattice vectors, respectively. Energy dispersion relations are obtained along the perimeter of the dotted triangle connecting the high symmetry points, Γ, K and M.

B.1.1 Tight Binding approximation for the π Bands of Two-Dimensional Graphite

Two Bloch functions, constructed from atomic orbitals for the two inequivalent carbon atoms at A and B in Fig. B.1, provide the basis functions for 2D graphite. When we consider only nearest-neighbor interactions, then there is only an integration over a single atom in the diagonal matix elements $\mathcal{H}_{AA}$ and $\mathcal{H}_{BB}$, as is shown in Eq. 1.81 and thus $\mathcal{H}_{AA} = \mathcal{H}_{BB} = \epsilon_{2p}$. For the off-diagonal matrix element $\mathcal{H}_{AB}$, we must consider the three nearest-neighbor B atoms relative to an A atom, which are denoted by the vectors $\vec{R}_1, \vec{R}_2$, and $\vec{R}_3$. We then consider the contribution to Eq. 1.82 from $\vec{R}_1, \vec{R}_2$, and $\vec{R}_3$ as follows:

$$\begin{aligned} \mathcal{H}_{AB} &= t(e^{i\vec{k}\cdot\vec{R}_1} + e^{i\vec{k}\cdot\vec{R}_2} + e^{i\vec{k}\cdot\vec{R}_3}) \\ &= tf(k) \end{aligned} \tag{B.3}$$

where t is given by Eq. 1.83[1] and $f(k)$ is a function of the sum of the phase factors of $e^{i\vec{k}\cdot\vec{R}_j}$ $(j = 1, \cdots, 3)$. Using the x, y coordinates of Fig. B.1(a), $f(k)$ is given by:

$$f(k) = e^{ik_x a/\sqrt{3}} + 2e^{-ik_x a/2\sqrt{3}} \cos\left(\frac{k_y a}{2}\right). \tag{B.4}$$

Since $f(k)$ is a complex function, and the Hamiltonian forms a Hermitian matrix, we write $\mathcal{H}_{BA} = \mathcal{H}^*_{AB}$ in which * denotes the complex conjugate. Using Eq. (B.4), the overlap integral matrix is given by $\mathcal{S}_{AA}$= $\mathcal{S}_{BB} = 1$, and $\mathcal{S}_{AB} = sf(k) = \mathcal{S}^*_{BA}$. Here s has the same definition

[1] We often use the symbol $\gamma_0 = |t|$ for the nearest neighbor transfer integral.

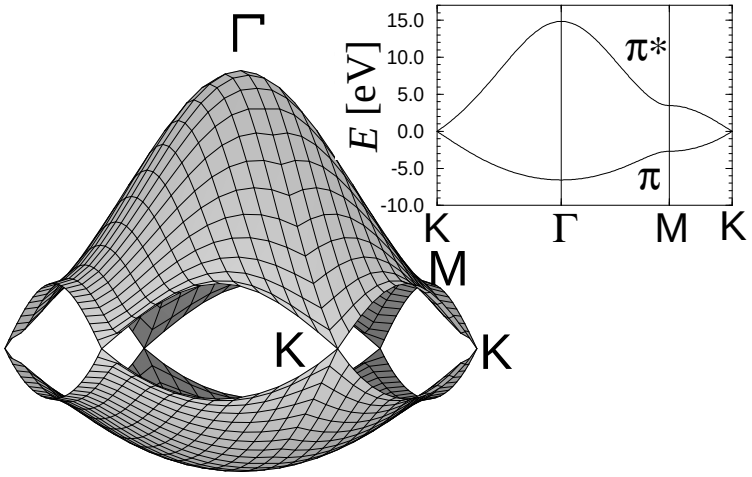


Figure B.2: The energy dispersion relations for 2D graphite are shown throughout the whole region of the Brillouin zone. Here we use the parameters $\epsilon_{2p} = 0$, $t = -3.033$eV and $s = 0.129$. The inset shows the electronic energy dispersion along the high symmetry directions of the triangle ΓMK shown in Fig. B.1(b) (see text).

as in Eq. 1.84, so that the explicit forms for $\mathcal{H}$ and $\mathcal{S}$ can be written as:

$$\mathcal{H} = \begin{pmatrix} \epsilon_{2p} & tf(k) \\ tf(k)^* & \epsilon_{2p} \end{pmatrix}, \quad \mathcal{S} = \begin{pmatrix} 1 & sf(k) \\ sf(k)^* & 1 \end{pmatrix}. \tag{B.5}$$

By solving the secular equation $\det(\mathcal{H} - E\mathcal{S}) = 0$ and using $\mathcal{H}$ and $\mathcal{S}$ as given in Eq. (B.5), the eigenvalues $E(\vec{k})$ are obtained as a function $w(\vec{k})$, k_x and k_y:

$$E_{g2D}(\vec{k}) = \frac{\epsilon_{2p} \pm tw(\vec{k})}{1 \pm sw(\vec{k})}, \tag{B.6}$$

where the + signs in the numerator and denominator go together giving the bonding π energy band, and likewise for the − signs, which give the anti-bonding π^* band, while the function $w(\vec{k})$ is given by:

$$w(\vec{k}) = \sqrt{|f(\vec{k})|^2} = \sqrt{1 + 4\cos\frac{\sqrt{3}k_x a}{2}\cos\frac{k_y a}{2} + 4\cos^2\frac{k_y a}{2}}. \tag{B.7}$$

In Fig. B.2, the energy dispersion relations of two-dimensional graphite are shown throughout the 2D Brillouin zone and the inset shows the energy dispersion relations along the high symmetry axes along the perimeter of the triangle shown in Fig. B.1(b). The upper half of the energy dispersion curves describes the π^*-energy anti-bonding band, and the lower half is the π-energy bonding band. The upper π^* band and the lower π band are degenerate at the K points through which the Fermi energy passes. Since there are two π electrons per unit cell, these two π electrons fully occupy the lower π band. Since a detailed calculation of the density of states shows that the density of states at the Fermi level is zero, two-dimensional graphite is a zero-gap semiconductor.

When the overlap integral s becomes zero, the π and π^* bands become symmetrical around $E = \epsilon_{2p}$ which can be understood from Eq. (B.6). The energy dispersion relations in the case of $s = 0$ are commonly used as a simple approximation for the electronic structure of a graphene layer:

$$E_{g2D}(k_x, k_y) = \pm t \left\{ 1 + 4\cos\left(\frac{\sqrt{3}k_x a}{2}\right)\cos\left(\frac{k_y a}{2}\right) + 4\cos^2\left(\frac{k_y a}{2}\right)\right\}^{1/2}. \tag{B.8}$$

The simple approximation given by Eq. (B.8) is used next to obtain a simple approximation for the electronic dispersion relations for carbon nanotubes, and provides an excellent first approximation for the analysis of presently available experiments on carbon nanotubes.

B.2 Single Wall Carbon Nanotubes

In §B.2 we briefly review the structure of single wall carbon nanotubes and relate this structure to the 2D graphene sheet discussed in §B.1, while §B.2.1 gives the electronic structure of the single wall carbon nanotube, as obtained from the tight binding approximation and from $E(k)$ for the graphene sheet, given by Eq. B.8.

B.2.1 Structure

A single-wall carbon nanotube can be described as a graphene sheet rolled into a cylindrical shape so that the structure is one-dimensional with axial symmetry, and in general exhibits a spiral conformation, called *chirality*. The chirality, as defined in this appendix, is given by a single vector called the chiral vector. To specify the structure of carbon nanotubes, we define several important vectors, which are derived from the chiral vector.

Chiral Vector: $\mathbf{C}_h$

The structure of a single-wall carbon nanotube (see Fig. B.3) is specified by the vector ($\overrightarrow{OA}$ in Fig. B.4) which corresponds to a section of the nanotube perpendicular to the nanotube axis (hereafter we call this section the equator of the nanotube). In Fig. B.4, the unrolled honeycomb lattice of the nanotube is shown, in which $\overrightarrow{OB}$ is the direction of the nanotube axis, and the direction of $\overrightarrow{OA}$ corresponds to the equator. By considering the crystallographically equivalent sites O, A, B, and B', and by rolling the honeycomb sheet so that points O and A coincide (and points B and B' coincide), a paper model of a carbon nanotube can be constructed. The vectors $\overrightarrow{OA}$ and $\overrightarrow{OB}$ define the chiral vector $\mathbf{C}_h$ and the translational vector $\mathbf{T}$ of a carbon nanotube, respectively, as further explained below.

The chiral vector $\mathbf{C}_h$ can be expressed by the real space unit vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ (see Fig. B.4) of the hexagonal lattice defined in Eq. (B.1):

$$\mathbf{C}_h = n\mathbf{a}_1 + m\mathbf{a}_2 \equiv (n, m), \quad (n, m \text{ are integers}, 0 \le |m| \le n). \tag{B.9}$$

The specific chiral vectors $\mathbf{C}_h$ shown in Fig. B.3 are, respectively, (a) $(5, 5)$, (b) $(9, 0)$ and (c) $(10, 5)$, and the chiral vector shown in Fig. B.4 is $(4, 2)$. An armchair nanotube corresponds to the case of $n = m$, that is $\mathbf{C}_h = (n, n)$ [see Fig. B.3(a)], and a zigzag nanotube corresponds

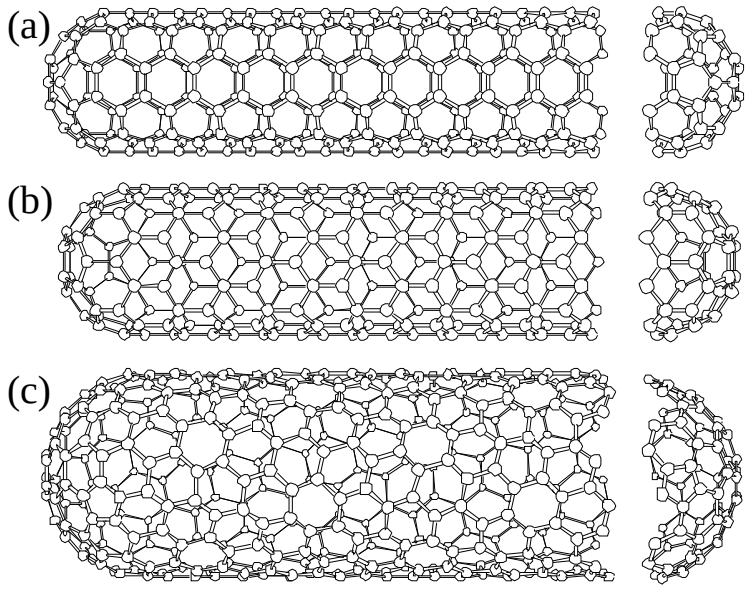


Figure B.3: Classification of carbon nanotubes: (a) armchair, (b) zigzag, and (c) chiral nanotubes, showingcross-sections and caps for the 3 basic kinds of nanotubes.

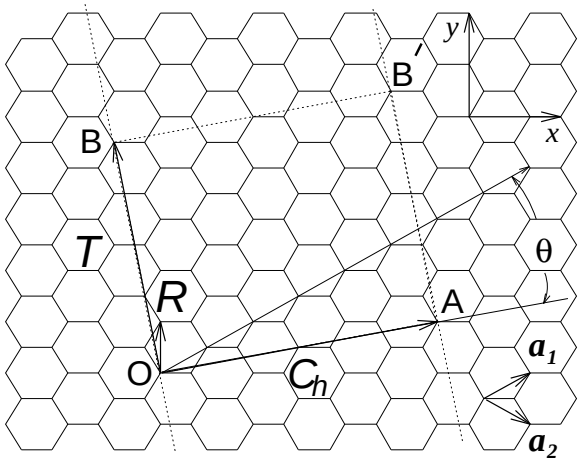


Figure B.4: The unrolled honeycomb lattice of a nanotube, showing the unit vectors $\vec{a}_1$ and $\vec{a}_2$ for the graphene sheet. When we connect sites O and A, and B and B', a nanotube can be constructed. $\overrightarrow{OA}$ and $\overrightarrow{OB}$ define the chiral vector $\mathbf{C}_h$ and the translational vector $\mathbf{T}$ of the nanotube, respectively. The rectangle $OAB'B$ defines the unit cell for the nanotube. The figure corresponds to $\mathbf{C}_h = (4, 2)$, $d = d_R = 2$, $\mathbf{T} = (4, -5)$, $N = 28$, $\mathbf{R} = (1, -1)$.

to the case of $m = 0$, or $\mathbf{C}_h = (n, 0)$ [see Fig. B.3(b)]. All other (n, m) chiral vectors correspond to chiral nanotubes [see Fig. B.3(c)]. Because of the hexagonal symmetry of the honeycomb lattice, we need to consider only $0 < |m| < n$ in $\mathbf{C}_h = (n, m)$ for chiral nanotubes.

The diameter of the carbon nanotube, d_t, is given by L/π, in which L is the circumferential length of the carbon nanotube:

$$d_t = L/\pi, \;\; L = |\mathbf{C}_h| = \sqrt{\mathbf{C}_h \cdot \mathbf{C}_h} = a\sqrt{n^2 + m^2 + nm}. \tag{B.10}$$

It is noted here that $\mathbf{a}_1$ and $\mathbf{a}_2$ are not orthogonal to each other and that the inner products between $\mathbf{a}_1$ and $\mathbf{a}_2$ yield:

$$\mathbf{a}_1 \cdot \mathbf{a}_1 = \mathbf{a}_2 \cdot \mathbf{a}_2 = a^2, \quad \mathbf{a}_1 \cdot \mathbf{a}_2 = \frac{a^2}{2}, \tag{B.11}$$

where the lattice constant $a = 1.42\,\text{Å} \times\sqrt{3} = 2.46\,\text{Å}$ of the honeycomb lattice is given in Eq. (B.1).

The chiral angle θ (see Fig. B.4) is defined as the angle between the vectors $\mathbf{C}_h$ and $\mathbf{a}_1$, with values of θ in the range $0 \le |\theta| \le 30^\circ$, because of the hexagonal symmetry of the honeycomb lattice. The chiral angle θ denotes the tilt angle of the hexagons with respect to the direction of the nanotube axis, and the angle θ specifies the spiral symmetry. The chiral angle θ is defined by taking the inner product of $\mathbf{C}_h$ and $\mathbf{a}_1$, to yield an expression for $\cos\theta$:

$$\cos\theta = \frac{\mathbf{C}_h \cdot \mathbf{a}_1}{|\mathbf{C}_h||\mathbf{a}_1|} = \frac{2n + m}{2\sqrt{n^2 + m^2 + nm}}, \tag{B.12}$$

thus relating θ to the integers (n, m) defined in Eq. (B.9). In particular, zigzag and armchair nanotubes correspond to $\theta = 0^\circ$ and $\theta = 30^\circ$, respectively.

B.2.2 Translational Vector: T

The translation vector $\mathbf{T}$ is defined to be the unit vector of a 1D carbon nanotube. The vector $\mathbf{T}$ is parallel to the nanotube axis and is normal to the chiral vector $\mathbf{C}_h$ in the unrolled honeycomb lattice in Fig. B.4. The lattice vector $\mathbf{T}$ shown as $\overrightarrow{OB}$ in Fig. B.4 can be expressed in terms of the basis vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ as:

$$\mathbf{T} = t_1\mathbf{a}_1 + t_2\mathbf{a}_2 \equiv (t_1, t_2), \;\; \text{(where } t_1, t_2 \text{ are integers)}. \tag{B.13}$$

The translation vector $\mathbf{T}$ corresponds to the first lattice point of the 2D graphene sheet through which the vector $\overrightarrow{OB}$ (normal to the chiral vector $\mathbf{C}_h$) passes. From this fact, it is clear that t_1 and t_2 do not have a common divisor except for unity. Using $\mathbf{C}_h \cdot \mathbf{T} = 0$ and Eqs. (B.9), (B.11), and (B.13), we obtain expressions for t_1 and t_2 given by:

$$t_1 = \frac{2m + n}{d_R}, \;\; t_2 = -\frac{2n + m}{d_R} \tag{B.14}$$

where d_R is the greatest common divisor (gcd) of $(2m+n)$ and $(2n+m)$. Also, by introducing d as the greatest common divisor of n and m, then d_R can be related to d by[2]

$$d_R = \begin{cases} d & \text{if } n - m \text{ is not a multiple of } 3d \\ 3d & \text{if } n - m \text{ is a multiple of } 3d. \end{cases} \tag{B.15}$$

[2]This relation is obtained by repeated use of the fact that when two integers, α and β ($\alpha > \beta$), have a common divisor, γ, then γ is also the common divisor of $(\alpha - \beta)$ and β (Euclid's law). When we denote the

The length of the translation vector, T, is given by:

$$T = |\mathbf{T}| = \sqrt{3}L/d_R, \tag{B.16}$$

where the circumferential nanotube length L is given by Eq. (F.18). We note that the length T is greatly reduced when (n, m) have a common divisor or when $(n - m)$ is a multiple of $3d$. In fact, for the $\mathbf{C}_h = (5, 5)$ armchair nanotube, we have $d_R = 3d = 15$, $\mathbf{T} = (1, -1)$ [Fig. B.3(a)], while for the $\mathbf{C}_h = (9, 0)$ zigzag nanotube we have $d_R = d = 9$, and $\mathbf{T} = (1, -2)$ [Fig. B.3(b)].

The unit cell of the 1D carbon nanotube is the rectangle $OAB'B$ defined by the vectors $\mathbf{C}_h$ and $\mathbf{T}$ (see Fig. B.4), while the unit vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ define the area of the unit cell of 2D graphite. When the area of the nanotube unit cell $|\mathbf{C}_h \times \mathbf{T}|$ (where the symbol $\times$ denotes the vector product operator) is divided by the area of a hexagon ($|\mathbf{a}_1 \times \mathbf{a}_2|$), the number of hexagons per unit cell N is obtained as a function of n and m in Eq. (B.9) as:

$$N = \frac{|\mathbf{C}_h \times \mathbf{T}|}{|\mathbf{a}_1 \times \mathbf{a}_2|} = \frac{2(m^2 + n^2 + nm)}{d_R} = \frac{2L^2}{a^2 d_R}, \tag{B.17}$$

where L and d_R are given by Eqs. (F.18) and (B.15), respectively, and we note that each hexagon contains two carbon atoms. Thus there are $2N$ carbon atoms (or $2p_z$ orbitals) in each unit cell of the carbon nanotube.

Unit Cells and Brillouin Zones

The unit cell for a carbon nanotube in real space is given by the rectangle generated by the chiral vector $\mathbf{C}_h$ and the translational vector $\mathbf{T}$, as is shown in $OAB'B$ in Fig. B.4. Since there are $2N$ carbon atoms in this unit cell, we will have N pairs of bonding π and anti-bonding π^* electronic energy bands. Similarly the phonon dispersion relations will consist of $6N$ branches resulting from a vector displacement of each carbon atom in the unit cell.

Expressions for the reciprocal lattice vectors $\mathbf{K}_2$ along the nanotube axis and $\mathbf{K}_1$ in the circumferential direction[3] are obtained from the relation $\mathbf{R}_i \cdot \mathbf{K}_j = 2\pi\delta_{ij}$, where $\mathbf{R}_i$ and $\mathbf{K}_j$ are, respectively, the lattice vectors in real and reciprocal space. Then, using Eqs. (B.14), (B.17), and the relations

$$\begin{array}{ll} \mathbf{C}_h \cdot \mathbf{K}_1 = 2\pi, & \mathbf{T} \cdot \mathbf{K}_1 = 0, \\ \mathbf{C}_h \cdot \mathbf{K}_2 = 0, & \mathbf{T} \cdot \mathbf{K}_2 = 2\pi, \end{array} \tag{B.18}$$

we get expressions for $\mathbf{K}_1$ and $\mathbf{K}_2$:

$$\mathbf{K}_1 = \frac{1}{N}(-t_2\mathbf{b}_1 + t_1\mathbf{b}_2), \quad \mathbf{K}_2 = \frac{1}{N}(m\mathbf{b}_1 - n\mathbf{b}_2), \tag{B.19}$$

where $\mathbf{b}_1$ and $\mathbf{b}_2$ are the reciprocal lattice vectors of two-dimensional graphite given by Eq. (B.2). In Fig. B.5, we show the reciprocal lattice vectors, $\mathbf{K}_1$ and $\mathbf{K}_2$, for a $\mathbf{C}_h =$

greatest common divisor as $\gamma = gcd(\alpha, \beta)$, we get

$$d_R = gcd(2m + n, 2n + m) = gcd(2m + n, n - m) = gcd(3m, n - m) = gcd(3d, n - m),$$

which gives Eq. (B.15).

[3]Since nanotubes are one-dimensional materials, only $\mathbf{K}_2$ is a reciprocal lattice vector. $\mathbf{K}_1$ gives discrete k values in the direction of $\mathbf{C}_h$.

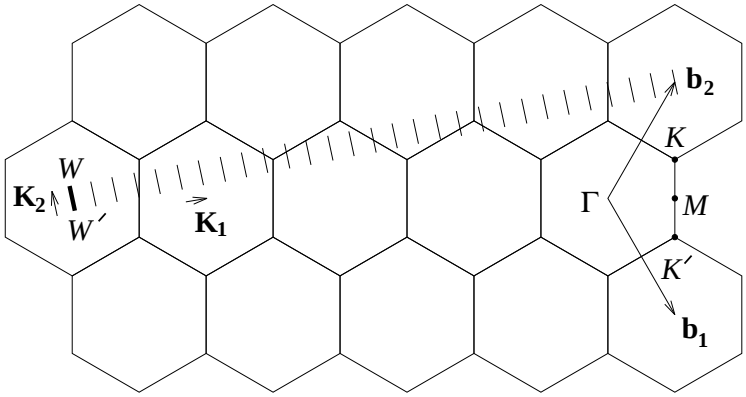


Figure B.5: The Brillouin zone of a carbon nanotube is represented by the line segment WW' which is parallel to $\mathbf{K}_2$. The vectors $\mathbf{K}_1$ and $\mathbf{K}_2$ are reciprocal lattice vectors corresponding to $\mathbf{C}_h$ and $\mathbf{T}$, respectively. The figure corresponds to $\mathbf{C}_h = (4, 2)$, $\mathbf{T} = (4, -5)$, $N = 28$, $\mathbf{K}_1 = (5\mathbf{b}_1 + 4\mathbf{b}_2)/28$, $\mathbf{K}_2 = (4\mathbf{b}_1 - 2\mathbf{b}_2)/28$ (see text).

$(4, 2)$ chiral nanotube. The first Brillouin zone of this one-dimensional material is the line segment WW'. Since $N\mathbf{K}_1 = -t_2\mathbf{b}_1 + t_1\mathbf{b}_2$ corresponds to a reciprocal lattice vector of two-dimensional graphite, two wave vectors which differ by $N\mathbf{K}_1$ are equivalent. Since t_1 and t_2 do not have a common divisor except for unity (see Sect. B.2.2), none of the $N - 1$ vectors $\mu\mathbf{K}_1$ (where $\mu = 1, \cdots, N - 1$) are reciprocal lattice vectors of two-dimensional graphite. Thus the N wave vectors $\mu\mathbf{K}_1$ ($\mu = 0, \cdots, N - 1$) give rise to N discrete k vectors, as indicated by the $N = 28$ parallel line segments in Fig. B.5, which arise from the quantized wave vectors associated with the periodic boundary conditions on $\mathbf{C}_h$. The length of all the parallel lines in Fig. B.5 is $2\pi/\mathbf{T}$ which is the length of the one-dimensional first Brillouin zone. For the N discrete values of the k vectors, N one-dimensional energy bands will appear. Because of the translational symmetry of $\mathbf{T}$, we have continuous wave vectors in the direction of $\mathbf{K}_2$ for a carbon nanotube of infinite length. However, for a nanotube of finite length L_t, the spacing between wave vectors is $2\pi/L_t$.

B.3 Electronic Structure of Single-Wall Nanotubes

B.3.1 Zone-Folding of Energy Dispersion Relations

The electronic structure of a single-wall nanotube can be obtained simply from that of two-dimensional graphite. By using periodic boundary conditions in the circumferential direction denoted by the chiral vector $\mathbf{C}_h$, the wave vector associated with the $\mathbf{C}_h$ direction becomes quantized, while the wave vector associated with the direction of the translational vector $\mathbf{T}$ (or along the nanotube axis) remains continuous for a nanotube of infinite length. Thus the energy bands consist of a set of one-dimensional energy dispersion relations which are cross sections of those for two-dimensional graphite (see Fig. B.2).

When the energy dispersion relations of two-dimensional graphite, $E_{g2D}(\mathbf{k})$ [see Eqs. (B.6) and/or (B.8)] at line segments shifted from WW' by $\mu\mathbf{K}_1$ ($\mu = 0, \cdots, N - 1$) are folded so that the wave vectors parallel to $\mathbf{K}_2$ coincide with WW' as shown in Fig. B.5, N pairs of

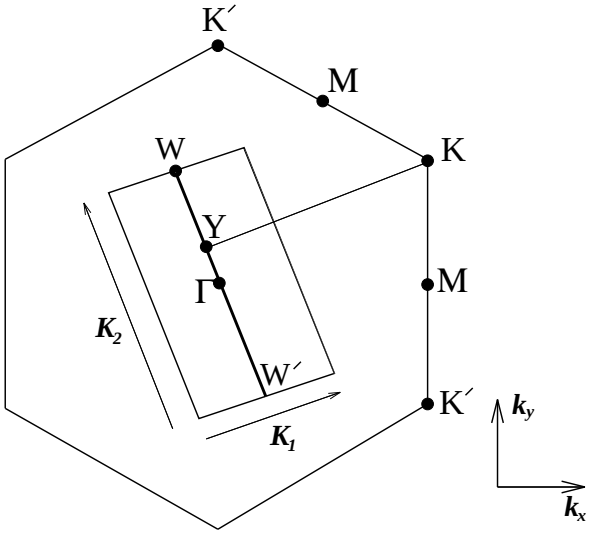


Figure B.6: The condition for metallic energy bands: if the ratio of the length of the vector $\overrightarrow{YK}$ to that of $\mathbf{K}_1$ is an integer, metallic energy bands are obtained.

1D energy dispersion relations $E_\mu(k)$ are obtained, where N is given by Eq. (B.17). These 1D energy dispersion relations are given by

$$E_\mu(k) = E_{g2D}\left(k\frac{\mathbf{K}_2}{|\mathbf{K}_2|} + \mu\mathbf{K}_1\right), \quad (\mu = 0, \cdots, N-1, \text{ and } -\frac{\pi}{T} < k < \frac{\pi}{T}), \tag{B.20}$$

corresponding to the energy dispersion relations of a single-wall carbon nanotube. The N pairs of energy dispersion curves given by Eq. (B.20) correspond to the cross sections of the two-dimensional energy dispersion surface shown in Fig. B.2, where cuts are made on the lines of $k\mathbf{K}_2/|\mathbf{K}_2| + \mu\mathbf{K}_1$. If for a particular (n, m) nanotube, the cutting line passes through a K point of the 2D Brillouin zone (Fig. B.1), where the π and π^* energy bands of two-dimensional graphite are degenerate by symmetry, the one-dimensional energy bands have a zero energy gap. In this case, the density of states at the Fermi level has a finite value for these carbon nanotubes, and they therefore are metallic. If, however, the cutting line does not pass through a K point, then the carbon nanotube is expected to show semiconducting behavior, with a finite energy gap between the valence and conduction bands.

The condition for obtaining a metallic energy band is that the ratio of the length of the vector $\overrightarrow{YK}$ to that of $\mathbf{K}_1$ in Fig. B.6 is an integer.[4] Since the vector $\overrightarrow{YK}$ is given by

$$\overrightarrow{YK} = \frac{2n+m}{3}\mathbf{K}_1, \tag{B.21}$$

[4]There are two inequivalent K and K' points in the Brillouin zone of 2D graphite as is shown in Fig. B.6 and thus the metallic condition can also be obtained in terms of K'. However, the results in that case are identical to the case specified by YK.

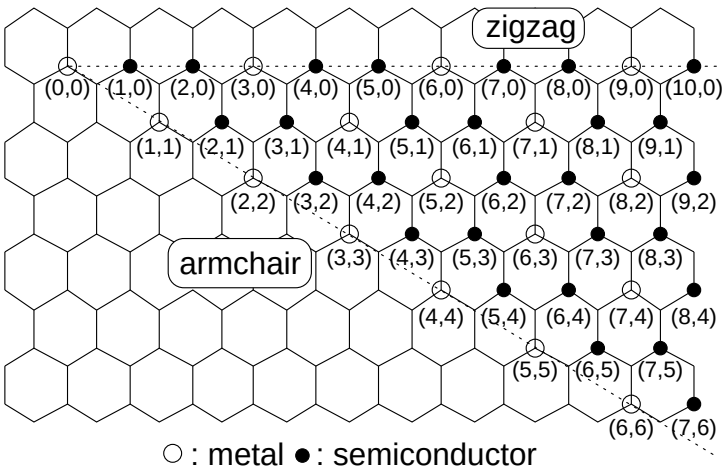


Figure B.7: The carbon nanotubes (n, m) that are metallic and semiconducting, respectively, are denoted by open and solid circles on the map of chiral vectors (n, m). For very small diameter nanotubes (e.g., $d_t < 0.7\,\mathrm{nm}$), the tight binding approximation is not sufficiently accurate, and more detailed approaches are needed. For example, small diameter nanotubes, such as the (4,2) nanotube is predicted to be semiconducting by tight binding approximation, though more detailed calculations show (4,2) to be metallic and experiments indicate that it may be superconducting.

the condition for metallic nanotubes is that $(2n + m)$ or equivalently $(n - m)$ is a multiple of 3.[5] In particular, the armchair nanotubes denoted by (n, n) are always metallic, and the zigzag nanotubes $(n, 0)$ are only metallic when n is a multiple of 3.

In Fig. B.7, we show which carbon nanotubes are metallic and which are semiconducting, denoted by open and solid circles, respectively. From Fig. B.7, it follows that approximately one third of the carbon nanotubes are metallic and the other two thirds are semiconducting.

B.3.2 Energy Dispersion of Armchair and Zigzag Nanotubes

To obtain explicit expressions for the dispersion relations, the simplest cases to consider are the nanotubes having the highest symmetry, i.e. the achiral armchair and zigzag nanotubes. The appropriate periodic boundary conditions used to obtain the energy eigenvalues for the (n, n) armchair nanotube define the small number of allowed wave vectors $k_{x,q}$ in the circumferential direction

$$n\sqrt{3}k_{x,q}a = 2\pi q, \quad (q = 1, \ldots, 2n). \tag{B.22}$$

[5]Since $3n$ is a multiple of 3, the remainders of $(2n + m)/3$ and $(n - m)/3$ are identical.

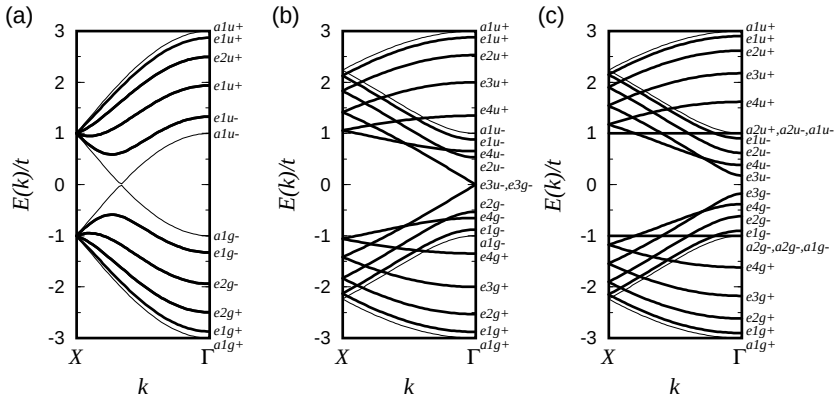


Figure B.8: One-dimensional energy dispersion relations for (a) armchair $(5,5)$, (b) zigzag $(9,0)$, and (c) zigzag $(10,0)$ carbon nanotubes labeled by the irreducible representations of the point group D_{nd} or D_{nh} (which describe the symmetry of these nanotubes), depending on whether there are even or odd numbers of bands n at the Γ point ($k = 0$). The a-bands are nondegenerate and the e-bands are doubly degenerate at a general k-point. The X points for armchair and zigzag nanotubes correspond to $k = \pm\pi/a$ and $k = \pm\pi/\sqrt{3}a$, respectively. (See Eqs. B.23–B.25.)

Substitution of the discrete allowed values for $k_{x,q}$ given by Eq. (B.22) into Eq. (B.8) yields the energy dispersion relations $E_q^a(k)$ for the armchair nanotube, $\mathbf{C}_h = (n,n)$,

$$E_q^a(k) = \pm t \left\{ 1 \pm 4\cos\left(\frac{q\pi}{n}\right)\cos\left(\frac{ka}{2}\right) + 4\cos^2\left(\frac{ka}{2}\right) \right\}^{1/2}, \qquad \begin{matrix} (-\pi < ka < \pi), \quad (q = 1, \ldots, 2n) \end{matrix} \tag{B.23}$$

in which the superscript a refers to armchair and k is a one-dimensional vector in the direction of the vector $\mathbf{K}_2 = (\mathbf{b}_1 - \mathbf{b}_2)/2$. This direction corresponds to the vector from the Γ point to the K point in the two-dimensional Brillouin zone of graphite[6] [see Fig. B.1(b)]. The resulting calculated 1D dispersion relations $E_q^a(k)$ for the $(5,5)$ armchair nanotube are shown in Fig. B.8(a), where we see six dispersion relations for the conduction bands[7] and an equal number for the valence bands.

Because of the degeneracy point between the valence and conduction bands at the band crossing which occurs at the Fermi energy, the $(5,5)$ armchair nanotube is thus a zero-gap semiconductor which will exhibit metallic conduction at finite temperatures, because only infinitesimal excitations are needed to excite carriers into the conduction band. All (n,n) armchair nanotubes have a band degeneracy between the highest valence band and the lowest conduction band at $k = \pm 2\pi/(3a)$, where the bands cross the Fermi level. Thus, all armchair nanotubes are expected to exhibit metallic conduction, similar to the behavior of 2D graphene sheets.

[6]Note that $\mathbf{K}_2$ vector is *not* a reciprocal lattice vector of the 2D graphite.

[7]The Fermi energy E_F corresponds to $E/t = 0$. The upper half of Fig. B.8 corresponds to the unoccupied conduction bands.

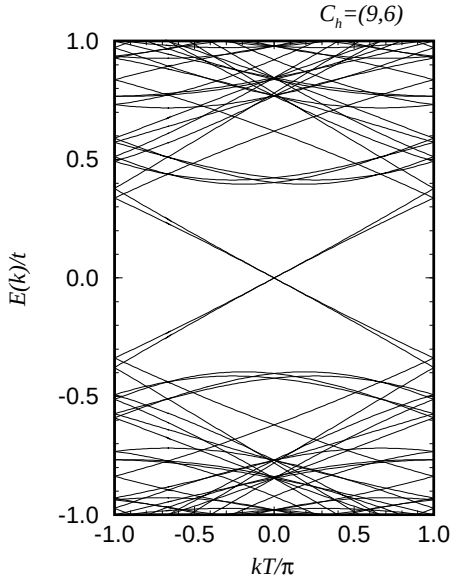


Figure B.9: Plot of the energy bands $E(k)$ for the metallic 1D nanotube $(n, m) = (9, 6)$ for values of the energy between $-t$ and t, in dimensionless units $E(k)/|t|$. The Fermi level is at $E = 0$. The largest common divisor of (9,6) is $d = 3$, and the value of d_R is $d_R = 3$. The general behavior of the four energy bands intersecting at $k = 0$ is typical of the case where $d_R = d$.

The energy bands for the $\mathbf{C}_h = (n, 0)$ zigzag nanotube $E_q^z(k)$ can be obtained likewise from Eq. (B.8) by writing the periodic boundary condition on k_y as:

$$nk_{y,q}a = 2\pi q, \qquad (q = 1, \ldots, 2n), \tag{B.24}$$

to yield the 1D dispersion relations for the $4n$ states for the $(n, 0)$ zigzag nanotube (denoted by the superscript z)

$$\begin{gathered} E_q^z(k) = \pm t \left\{ 1 \pm 4\cos\left(\frac{\sqrt{3}ka}{2}\right)\cos\left(\frac{q\pi}{n}\right) + 4\cos^2\left(\frac{q\pi}{n}\right) \right\}^{1/2}, \\ \left(-\frac{\pi}{\sqrt{3}} < ka < \frac{\pi}{\sqrt{3}}\right), \qquad (q = 1, \ldots, 2n). \end{gathered} \tag{B.25}$$

The resulting calculated 1D dispersion relations $E_q^z(k)$ for the $(9, 0)$ and $(10, 0)$ zigzag nanotubes are shown in Figs. B.8(b) and (c), respectively. There is no energy gap for the metallic $(9, 0)$ nanotube at $k = 0$, whereas the $(10, 0)$ nanotube indeed shows an energy gap. For a general $(n, 0)$ zigzag nanotube, when n is a multiple of 3, the energy gap at $k = 0$ becomes zero; however, when n is not a multiple of 3, an energy gap opens at $k = 0$, as seen in Fig. B.8(c).

B.3.3 Dispersion of Chiral Nanotubes

Chiral nanotubes have usually much larger unit cells and, therefore a large number of branches in their dispersion relation. In Fig. B.9, we show dispersion relations for the $(9, 6)$

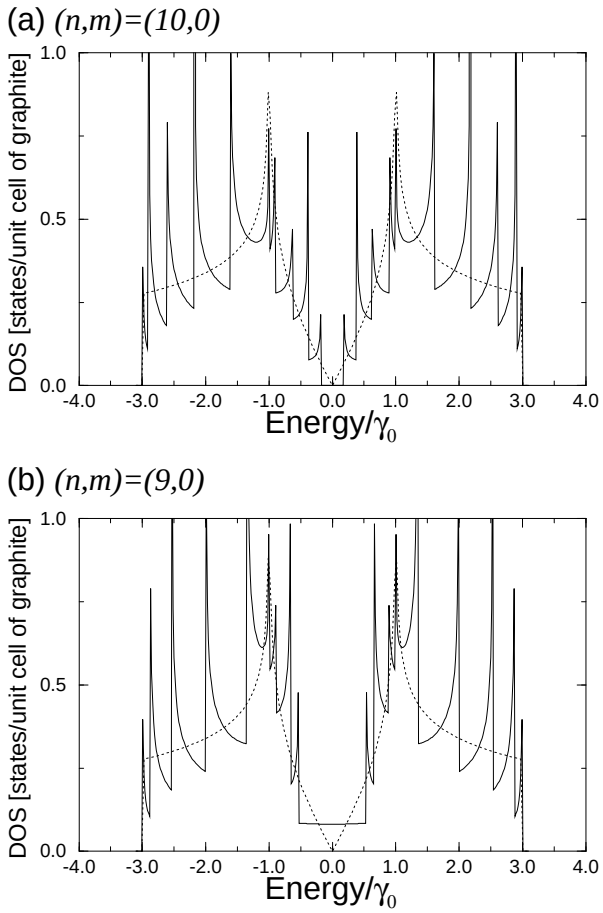


Figure B.10: Electronic 1D density of states per unit cell of a 2D graphene sheet for two $(n, 0)$ zigzag nanotubes: (a) the $(9, 0)$ nanotube which has metallic behavior, (b) the $(10, 0)$ nanotube which has semiconducting behavior. Also shown as a dashed line in the figure is the density of states for the 2D graphene sheet.

chiral nanotube. Since $n - m$ is a multiple of 3, this chiral nanotube is metallic.

B.4 Density of States, Energy Gap

Of particular interest has been the energy dependence of the nanotube density of states, as shown in Fig. B.10 which compares the density of states for metallic (9,0) and semiconducting (10,0) zigzag nanotubes. In this figure, we see that the density of states near the Fermi level E_F (located at $E = 0$) is different for metallic and semiconducting nanotubes. The density of states at E_F has a value of zero for semiconducting nanotubes, but is non-zero (and small) for metallic nanotubes. Also of great interest are the singularities in the 1D density of states, corresponding to extrema in the $E(k)$ relations. The comparison between the 1D density of states for the nanotubes and the 2D density of states for a graphene layer is included in the figure. Another important result, pertaining to semiconducting nanotubes,

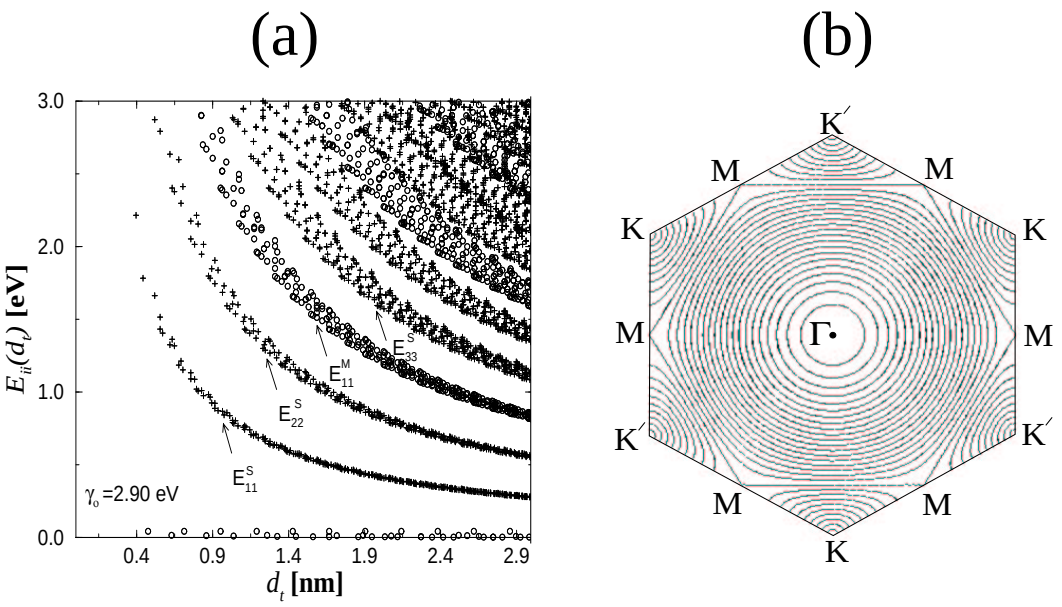


Figure B.11: (a) Calculated energy separations $E_{ii}(d_t)$ between van Hove singularities i in the 1D electronic density of states of the conduction and valence bands for all (n, m) values vs nanotube diameter $0.4 < d_t < 3.0\,\text{nm}$, using a value for the carbon-carbon energy overlap integral of $\gamma_0 = 2.9\,\text{eV}$ and a nearest neighbor carbon-carbon distance $a_{\text{C}-\text{C}} = 1.42\,\text{Å}$. Semiconducting (S) and metallic (M) nanotubes are indicated by crosses and open circles, respectively. The index i in the interband transitions E_{ii} denotes the transition between the van Hove singularities, with $i = 1$ being closest to the Fermi level. (b) Plot of the 2D equi-energy contours of graphite, showing trigonal warping effects in the contours, as we move from the K point in the $K - \Gamma$ or $K - M$ directions. The equi-energy contours are circles near the K point and near the center of the Brillouin zone. But near the zone boundary, the contours are straight lines which connect the nearest M points.

shows that their energy gap depends upon the reciprocal nanotube diameter d_t, according to the relation $E_g = (|t|a_{\mathrm{C-C}})/d_t$, independent of the chiral angle of the semiconducting nanotube, where $a_{\mathrm{C-C}} = a/\sqrt{3}$.

It is significant that every (n, m) nanotube has a different and unique set of energies where the singularities in the 1D electronic density of states occur. Figure B.11(a) shows a plot of the energy differences E_{ii} between singularities i in the conduction and valence bands for every possible nanotube as a function of nanotube diameter, showing the uniqueness of the energies of these singularities in the density of states. This uniqueness arises from the trigonal warping effect. Figure B.11(b) shows that the constant energy surfaces around the origin (Γ point where $k = 0$) and around the K and K' points in the 2D Brillouin zone are circular only near the Γ, K, and K' high symmetry points. Away from these symmetry points, trigonal warping effects become important, giving rise to a different set of singularities in the density of states, depending on the nanotube diameter and chirality. We can measure the E_{ii} singularities in the density of states at the single nanotube level by the Raman effect, which shows a strong resonance with an individual (n, m) carbon nanotube when the laser excitation energy is equal to one of these singularities. Therefore, the resonance Raman effect can be used to identify the (n, m) values for individual carbon nanotubes. Because of the unique properties of these particular low dimensional systems, spectroscopy can be used to obtain structural information about individual carbon nanotubes.

Appendix C

Harmonic Oscillators, Phonons, and Electron-Phonon Interaction

C.1 Harmonic Oscillators

In this section we review the solution of the harmonic oscillator problem in quantum mechanics using raising and lowering operators. This is aimed at providing a quick review as background for the lecture on phonon scattering processes and other topics in this course.

The Hamiltonian for the harmonic oscillator in one-dimension is written as:

$$\mathcal{H} = \frac{p^2}{2m} + \frac{1}{2}\kappa x^2. \tag{C.1}$$

We know classically that the frequency of oscillation is given by $\omega = \sqrt{\kappa/m}$ so that

$$\mathcal{H} = \frac{p^2}{2m} + \frac{1}{2}m\omega^2 x^2 \tag{C.2}$$

Define the lowering and raising operators a and $a^\dagger$ respectively by

$$a = \frac{p - im\omega x}{\sqrt{2\hbar m\omega}} \tag{C.3}$$

$$a^\dagger = \frac{p + im\omega x}{\sqrt{2\hbar m\omega}} \tag{C.4}$$

Since $[p, x] = \hbar/i$, then $[a, a^\dagger]$ =1 so that

$$\mathcal{H} = \frac{1}{2m}\Big[(p + i\omega mx)(p - i\omega mx) + m\hbar\omega\Big] \tag{C.5}$$

$$= \hbar\omega[a^\dagger a + 1/2]. \tag{C.6}$$

Let $N = a^\dagger a$ denote the number operator and its eigenstates $|n\rangle$ so that $N|n\rangle = n|n\rangle$ where n is any real number. However

$$\langle n|N|n\rangle = \langle n|a^\dagger a|n\rangle = \langle y|y\rangle = n \geq 0 \tag{C.7}$$

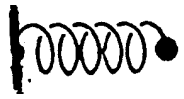


Figure C.1: Simple harmonic oscillator with single spring.

where $|y\rangle = a|n\rangle$ and the absolute value square of the eigenvector cannot be negative. Hence n is a positive number or zero.

$$Na|n\rangle = a^\dagger aa|n\rangle = (aa^\dagger - 1)a|n\rangle = (n-1)a|n\rangle \tag{C.8}$$

Hence $a|n\rangle = c|n-1\rangle$ and $\langle n|a^\dagger a|n\rangle = |c|^2$. However from Eq. C.7 $\langle n|a^\dagger a|n\rangle = n$ so that $c = \sqrt{n}$ and $a|n\rangle = \sqrt{n}|n-1\rangle$. Since the operator a lowers the quantum number of the state by unity, a is called the annihilation operator. Therefore n also has to be an integer, so that the null state is eventually reached by applying operator a for a sufficient number of times.

$$Na^\dagger|n\rangle = a^\dagger aa^\dagger|n\rangle = a^\dagger(1 + a^\dagger a)|n\rangle = (n+1)a^\dagger|n\rangle \tag{C.9}$$

Hence $a^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle$ so that $a^\dagger$ is called a raising operator or a creation operator. Finally,

$$\mathcal{H}|n\rangle = \hbar\omega[N + 1/2]|n\rangle = \hbar\omega(n + 1/2)|n\rangle \tag{C.10}$$

so the eigenvalues become

$$E = \hbar\omega(n + 1/2), \qquad n = 0, 1, 2, \ldots. \tag{C.11}$$

C.2 Phonons

In this section we relate the lattice vibrations to harmonic oscillators and identify the quanta of the lattice vibrations with phonons. Consider the 1-D model of atoms connected by springs (see Fig. C.1). The Hamiltonian for this case is written as:

$$\mathcal{H} = \sum_{s=1}^{\mathcal{N}} \left(\frac{p_s^2}{2m} + \frac{1}{2}\kappa(x_{s+1} - x_s)^2 \right) \tag{C.12}$$

This equation doesn't look like a set of independent harmonic oscillators since x_s and x_{s+1} are coupled. Let

$$\begin{aligned} x_s &= 1/\sqrt{\mathcal{N}} \textstyle\sum_k Q_k e^{iksa} \\ p_s &= 1/\sqrt{\mathcal{N}} \textstyle\sum_k P_k e^{iksa}. \end{aligned} \tag{C.13}$$

These Q_k, P_k's are called phonon coordinates. It can be verified that the commutation relation for momentum and coordinate implies a commutation relation between P_k and $Q_{k'}$

$$[p_s, x_{s'}] = \frac{\hbar}{i}\delta_{ss'} \Longrightarrow [P_k, Q_{k'}] = \frac{\hbar}{i}\delta_{kk'}. \tag{C.14}$$

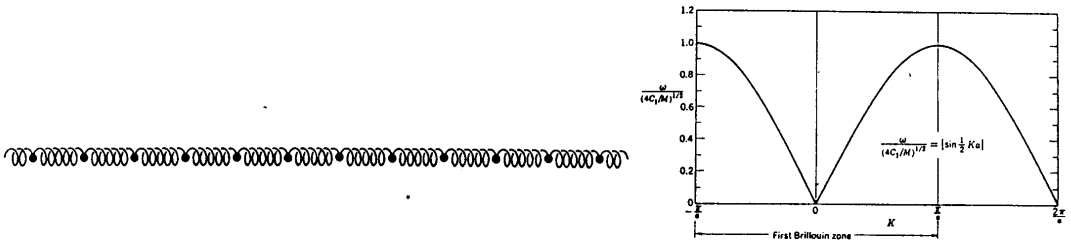


Figure C.2: Schematic for a one dimensional phonon model and the corresponding dispersion relation.

The Hamiltonian in phonon coordinates is:

$$\mathcal{H} = \sum_k \left(\frac{1}{2m} P_k^\dagger P_k + \frac{1}{2} m\omega_k^2 Q_k^\dagger Q_k \right) \tag{C.15}$$

with the dispersion relation given by

$$\omega_k = \sqrt{2\kappa}(1 - \cos ka) \tag{C.16}$$

This is all in Kittel ISSP, pp. 611-613. (see Fig. C.2) Again let

$$a_k = \frac{iP_k^\dagger + m\omega_k Q_k}{\sqrt{2\hbar m\omega_k}}, \tag{C.17}$$

$$a_k^\dagger = \frac{-iP_k + m\omega_k Q_k^\dagger}{\sqrt{2\hbar m\omega_k}} \tag{C.18}$$

so that the Hamiltonian is written as:

$$\mathcal{H} = \sum_k \hbar\omega_k (a_k^\dagger a_k + 1/2) \Rightarrow E = \sum_k (n_k + 1/2)\hbar\omega_k \tag{C.19}$$

The quantum of energy $\hbar\omega_k$ is called a phonon. The state vector of a system of phonons is written as $|n_1, n_2, \ldots, n_k, \ldots\rangle$, upon which the raising and lowering operator can act:

$$a_k|n_1, n_2, \ldots, n_k, \ldots\rangle = \sqrt{n_k}\, |n_1, n_2, \ldots, n_k - 1, \ldots\rangle \tag{C.20}$$

$$a_k^\dagger|n_1, n_2, \ldots, n_k, \ldots\rangle = \sqrt{n_k + 1}\, |n_1, n_2, \ldots, n_k + 1, \ldots\rangle \tag{C.21}$$

From Eq. C.21 it follows that the probability of annihilating a phonon of mode k is the absolute value squared of the diagonal matrix element or n_k.

C.3 Electron-Phonon Interaction

The basic Hamiltonian for the electron-lattice system is

$$\mathcal{H} = \sum_k \frac{p_k^2}{2m} + \frac{1}{2}{\sum_{kk'}}' \frac{e^2}{|\vec{r}_k - \vec{r}_{k'}|} + \sum_i \frac{P_i^2}{2M} + \frac{1}{2}{\sum_{ii'}}' V_{\text{ion}}(\vec{R}_i - \vec{R}_{i'}) + \sum_{k,i} V_{\text{el-ion}}(\vec{r}_k - \vec{R}_i) \tag{C.22}$$

where the first two terms constitute $\mathcal{H}_{\rm electron}$, the third and fourth terms are denoted by $\mathcal{H}_{\rm ion}$ and the last term is $\mathcal{H}_{\rm electron-ion}$. The electron-ion interaction term can be separated into two parts: the interaction of electrons with ions in their equilibrium positions, and an additional term due to lattice vibrations:

$$\begin{aligned} \mathcal{H}_{\rm el-ion} &= \mathcal{H}^0_{\rm el-ion} + \mathcal{H}_{\rm el-ph} && \text{(C.23)} \\ \sum_{k,i} V_{\rm el-ion}(\vec{r}_k - \vec{R}_i) &= \sum_{k,i} V_{\rm el-ion}[\vec{r}_k - (\vec{R}^0_i + \vec{s}_i)] && \text{(C.24)} \end{aligned}$$

where $\vec{R}^0_i$ is the equilibrium lattice site position and $\vec{s}_i$ is the displacement of the atoms from their equilibrium positions in a lattice vibration so that

$$\mathcal{H}^0_{\rm el-ion} = \sum_{k,i} V_{\rm el-ion}(\vec{r}_k - \vec{R}^0_i) \tag{C.25}$$

and

$$\mathcal{H}_{\rm el-ph} = -\sum_{k,i} \vec{s}_i \cdot \vec{\nabla} V_{\rm el-ion}(\vec{r}_k - \vec{R}^0_i). \tag{C.26}$$

In solving the Hamiltonian we use an adiabatic approximation, which solves the electronic part of the Hamiltonian by

$$(\mathcal{H}_{\rm electron} + \mathcal{H}^0_{\rm el-ion})\psi = E_{\rm el}\psi \tag{C.27}$$

and seeks a solution of the total problem as

$$\Psi = \psi(\vec{r}_1, \vec{r}_2, \cdots \vec{R}_1, \vec{R}_2, \cdots)\varphi(\vec{R}_1, \vec{R}_2, \cdots) \tag{C.28}$$

such that $\mathcal{H}\Psi = E\Psi$. Here Ψ is the wave function for the electron-lattice system. Plugging this into the Eq. C.22, we find

$$E\Psi = \mathcal{H}\Psi = \psi(\mathcal{H}_{\rm ion} + E_{\rm el})\varphi - \sum_i \frac{\hbar^2}{2M_i}\left(\varphi\nabla^2_i\psi + 2\vec{\nabla}_i\varphi\cdot\vec{\nabla}_i\psi\right) \tag{C.29}$$

Neglecting the last term, which is small, we have

$$\mathcal{H}_{\rm ion}\varphi = (E - E_{\rm el})\varphi \tag{C.30}$$

Hence we have decoupled the electron-lattice system.

$$(\mathcal{H}_{\rm electron} + \mathcal{H}^0_{\rm el-ion})\psi = E_{\rm el}\psi \tag{C.31}$$

which gives us the energy band structure and ψ satisfies Bloch's theorem while φ is the wave function for the ions

$$\mathcal{H}_{\rm ion}\varphi = E_{\rm ion}\varphi \tag{C.32}$$

which gives us phonon spectra and harmonic oscillator like wave functions, as we have already seen in §C.2.

The discussion has thus far left out the electron-phonon interaction $\mathcal{H}_{\rm el-ph}$

$$\mathcal{H}_{\rm el-ph} = -\sum_{k,i} \vec{s}_i \cdot \vec{\nabla} V_{\rm el-ion}(\vec{r}_k - \vec{R}^0_i) \tag{C.33}$$

which is then treated as a perturbation. Since the displacement vector can be written in terms of the normal coordinates $Q_{\vec{q},j}$

$$\vec{s}_i = \frac{1}{\sqrt{\mathcal{N}M}} \sum_{\vec{q},j} Q_{\vec{q},j} e^{i\vec{q}\cdot\vec{R}_i^0}\ \hat{e}_j \tag{C.34}$$

where j denotes the polarization index, $\mathcal{N}$ is the total number of ions and M is the ion mass. Hence

$$\mathcal{H}_{\text{el-ph}} = -\sum_{k,i} \frac{1}{\sqrt{\mathcal{N}M}} \sum_{\vec{q},j} Q_{\vec{q},j} e^{i\vec{q}\cdot\vec{R}_i^0}\ \hat{e}_j \cdot \vec{\nabla} V_{\text{el-ion}}(\vec{r}_k - \vec{R}_i^0) \tag{C.35}$$

where the normal coordinate can be expressed in terms of the lowering and raising operators

$$Q_{\vec{q},j} = \left(\frac{\hbar}{2\omega_{\vec{q},j}}\right)^{\frac{1}{2}} (a_{\vec{q},j} + a^{\dagger}_{-\vec{q},j}). \tag{C.36}$$

Writing out the time dependence explicitly,

$$a_{\vec{q},j}(t) = a_{\vec{q},j} e^{-i\omega_{\vec{q},j}t} \tag{C.37}$$

$$a^{\dagger}_{\vec{q},j}(t) = a^{\dagger}_{\vec{q},j} e^{i\omega_{\vec{q},j}t} \tag{C.38}$$

we obtain

$$\begin{aligned} \mathcal{H}_{\text{el-ph}} &= -\sum_{\vec{q},j} \left(\frac{\hbar}{2\mathcal{N}M\omega_{\vec{q},j}}\right)^{\frac{1}{2}} (a_{\vec{q},j} e^{-i\omega_{\vec{q},j}t} + a^{\dagger}_{\vec{q},j} e^{i\omega_{\vec{q},j}t}) \\ &\times \sum_{k,i} (e^{i\vec{q}\cdot\vec{R}_i^0} + e^{-i\vec{q}\cdot\vec{R}_i^0}) \hat{e}_j \vec{\nabla} V_{\text{el-ion}}(\vec{r}_k - \vec{R}_i^0) \qquad \text{(C.39)} \\ &= -\sum_{\vec{q},j} \left(\frac{\hbar}{2\mathcal{N}M\omega_{\vec{q},j}}\right)^{\frac{1}{2}} \left(a_{\vec{q},j} \sum_{k,i} \hat{e}_j e^{i(\vec{q}\vec{R}_i^0 - \omega_{\vec{q},j}t)} \cdot \vec{\nabla} V_{\text{el-ion}}(\vec{r}_k - \vec{R}_i^0) \right. \\ &+ \left. c.c. \right) \qquad \text{(C.40)} \end{aligned}$$

If we are only interested in the interaction between one electron and a phonon on a particular branch, say the longitudinal acoustic (LA) branch, then we drop the summation over j and k

$$\mathcal{H}_{\text{el-ph}} = -\left(\frac{\hbar}{2\mathcal{N}M\omega_{\vec{q}}}\right)^{\frac{1}{2}} \left(a_{\vec{q}} \sum_i \hat{e} e^{i(\vec{q}\cdot\vec{R}_i^0 - \omega_{\vec{q}}t)} \cdot \vec{\nabla} V_{\text{el-ion}}(\vec{r} - \vec{R}_i^0) + c.c \right) \tag{C.41}$$

where the first term in the bracket corresponds to the phonon absorption and the c.c. term corresponds to the phonon emission.

With $\mathcal{H}_{\text{el-ph}}$ at hand, we can solve transport problems (e.g., τ due to phonon scattering) and optical problems (e.g., indirect transitions) exactly since all of these problems involve the matrix element $\langle f|\mathcal{H}_{\text{el-ph}}|i\rangle$ of $\mathcal{H}_{\text{el-ph}}$ linking states $|i\rangle$ and $|j\rangle$.

Appendix D

Artificial Atoms

PHYSICS TODAY JANUARY 1993
Marc A. Kastner
Marc Kastner is the Donner Professor of Science in the department of physics at the Massachusetts Institute of Technology, in Cambridge.

The charge and energy of a sufficiently small particle of metal or semiconductor are quantized just like those of an atom. The current through such a quantum dot or one-electron transistor reveals atom-like features in a spectacular way.

The wizardry of modern semiconductor technology makes it possible to fabricate particles of metal or "pools" of electrons in a semiconductor that are only a few hundred angstroms in size. Electrons in these structures can display astounding behavior. Such structures, coupled to electrical leads through tunnel junctions, have been given various names: single electron transistors, quantum dots, zero-dimensional electron gases and Coulomb islands. In my own mind, however, I regard all of these as artificial atoms-atoms whose effective nuclear charge is controlled by metallic electrodes. Like natural atoms, these small electronic systems contain a discrete number of electrons and have a discrete spectrum of energy levels. Artificial atoms, however, have a unique and spectacular property: The current through such an atom or the capacitance between its leads can vary by many orders of magnitude when its charge is changed by a single electron. Why this is so, and how we can use this property to measure the level spectrum of an artificial atom, is the subject of this article.

To understand artificial atoms it is helpful to know how to make them. One way to confine electrons in a small region is by employing material boundaries by surrounding a metal particle with insulator, for example. Alternatively, one can use electric fields to confine electrons to a small region within a semiconductor. Either method requires fabricating very small structures. This is accomplished by the techniques of electron and x-ray lithography. Instead of explaining in detail how artificial atoms are actually fabricated, I will describe the various types of atoms schematically.

Figures D.1a and D.1b show two kinds of what is sometimes called, for reasons that will soon become clear, a single-electron transistor. In the first type (figure D.1a), which I call the all-metal artificial atom,[1] electrons are confined to a metal particle with typical dimensions of a few thousand angstroms or less. The particle is separated from the leads by thin insulators, through which electrons must tunnel to get from one side to the other. The leads are labeled "source" and "drain" because the electrons enter through the former

and leave through the latter the same way the leads are labeled for conventional field effect transistors, such as those in the memory of your personal computer. The entire structure sits near a large, well-insulated metal electrode, called the gate.

Figure D.1b shows a structures[2] that is conceptually similar to the all-metal atom but in which the confinement is accomplished with electric fields in gallium arsenide. Like the all-metal atom, it has a metal gate on the bottom with an insulator above it; in this type of atom the insulator is AlGaAs. When a positive voltage V_g is applied to the gate, electrons accumulate in the layer of GaAs above the AlGaAs. Because of the strong electric field at the AlGaAs-GaAs interface, the electrons' energy for motion perpendicular to the interface is quantized, and at low temperatures the electrons move only in the two dimensions parallel to the -interface. The special feature that makes this an artificial atom is the pair of electrodes on the top surface of the GaAs. When a negative voltage is applied between these and the source or drain, the electrons are repelled and cannot accumulate underneath them. Consequently the electrons are confined in a narrow channel between the two electrodes. Constrictions sticking but into the channel repel the electrons and create potential barriers at either end of the channel. A plot of a potential similar to the one seen by the electrons is shown in the inset in figure D.1. For an electron to travel from the source to the drain it must tunnel through the barriers. The "pool" of electrons that accumulates between the two constrictions plays the same role that the small particle plays in the all-metal atom, and the potential barriers from the constrictions play the role of the thin insulators. Because one can control the height of these barriers by varying the voltage on the electrodes, I call this type of artificial atom the controlled-barrier atom. Controlled- barrier atoms in which the heights of the two potential barriers can be varied independently have also been fabricated.[2] (The constrictions in these devices are similar to those used for measurements of quantized conductance in narrow channels as reported in PHYSICS TODAY, November 1988, page 21.) In addition, there are structures that behave like controlled-barrier atoms but in which the barriers are caused by charged impurities or grain boundaries.[2,4]

Figure D.1c shows another, much simpler type of artificial atom. The electrons in a layer of GaAs are sandwiched between two layers of insulating AlGaAs. One or both of these insulators acts as a tunnel barrier. If both barriers are thin, electrons can tunnel through them, and the structure is analogous to the single-electron transistor without the gate. Such structures, usually called quantum dots, have been studied extensively.[5,6] To create the structure, one starts with two-dimensional layers like those in figure D.1b. The cylinder can be made by etching away unwanted regions of the layer structure, or a metal electrode on the surface, like those in figure D.1b, can be used to repel electrons everywhere except in a small circular section of GaAs. Although a gate electrode can be added to this kind of structure, most of the experiments have bee done without one, so I call this the two-probe atom.

D.1 Charge quantization

One way to learn about natural atoms is to measure the energy required to add or remove electrons. This usually done by photoelectron spectroscopy. For example the minimum photon energy needed to remove an electron is the ionization potential, and the maximum energy (photons emitted when an atom captures an electron is the electron affinity. To learn about artificial atoms we also measure the energy needed to add or subtract electron.

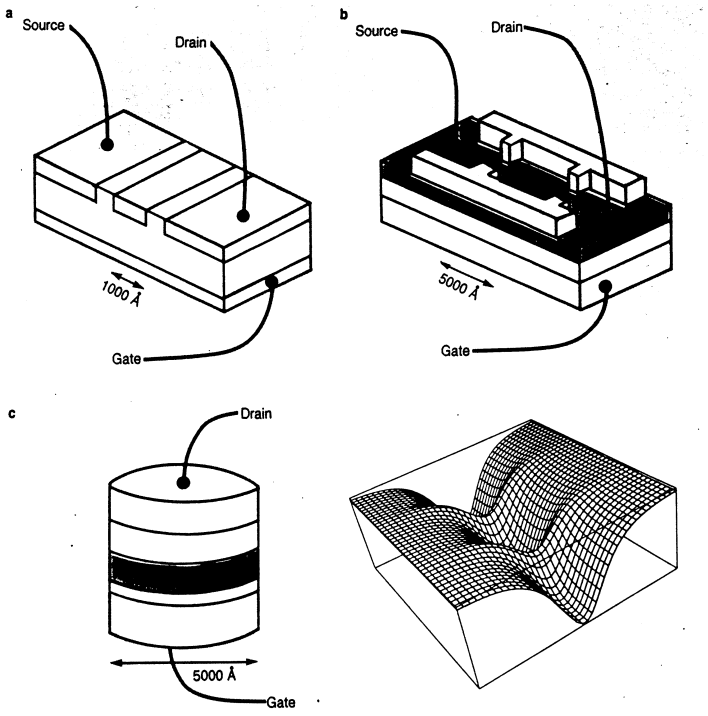


Figure D.1: The many forms of artificial atoms include the all-metal atom (a), the controlled-barrier atom (b) and the two-probe atom, or "quantum dot" (c). Areas shown in blue are metallic, white areas are insulating, and red areas are semiconducting. The dimensions indicated are approximate. The inset shows a potential similar to the one in the controlled-barrier atom, plotted as a function of position at the semiconductor-insulator interface. The electrons must tunnel through potential barriers caused by the two constrictions. For capacitance measurements with a two-probe atom, only the source barrier is made thin enough for tunneling, but for current measurements both source and drain barriers are thin.

However, we do it by measuring the current through the artificial atom.

Figure D.2 shows the current through a controller barrier atom[7] as a function of the voltage V_g between the gate and the atom. One obtains this plot by applying very small voltage between the source and drain, just large enough to measure the tunneling conductance between them. The results are astounding. The conductance(displays sharp resonances that are almost periodic in V_g By calculating the capacitance between the artificial atom and the gate we can show[2,8] that the period is the voltage necessary to add one electron to the confined pool of electrons. That is why we sometimes call the controller barrier atom a single-electron transistor: Whereas the transistors in your personal computer turn on only on(when many electrons are added to them, the artificial atom turns on and off again every time a single electron added to it.

A simple theory, the Coulomb blockade model, explains the periodic conductance resonances.[9] (See PHYSICS TODAY, May 1988, page 19.) This model is quantitatively correct for the all-metal atom and qualitatively correct for the controlled-barrier atom.[10] To understand the model, think about how an electron in the all-metal atom tunnels from one lead onto the metal particle and then onto the other lead. Suppose the particle is neutral to begin with. To add a charge Q to the particle requires energy $Q^2/2C$, where C is the total capacitance between the particle and the rest of the system; since you cannot add less than one electron the flow of current requires a Coulomb energy $e^2/2C$. This energy barrier is called the Coulomb blockade. A fancier way to say this is that charge quantization leads to an energy gap in the spectrum of states for tunneling: For an electron to tunnel onto the particle, its energy must exceed the Fermi energy of the contact by $e^2/2C$, and for a hole to tunnel, its energy must be below the Fermi energy by the same amount. Consequently the energy gap has width e^2/C. If the temperature is low enough that $kT < e^2/2C$, neither electrons nor holes can flow from one lead to the other.

The gap in the tunneling spectrum is the difference between the "ionization potential" and the "electron affinity" of the artificial atom. For a hydrogen atom the ionization potential is 13.6 eV, but the electron affinity, the binding energy of H^-, is only 0.75 eV. This large difference arises from the strong repulsive interaction between the two electrons bound to the same proton. Just as for natural atoms like hydrogen, the difference between the ionization potential and electron affinity for artificial atoms arises from the electron-electron interactions; the difference, however, is much smaller for artificial atoms because they are much bigger than natural ones.

By changing the gate voltage V_g one can alter the energy required to add charge to the particle. V_g is applied between the gate and the source, but if the drain-source voltage is very small, the source, drain and particle will all be at almost the same potential. With V_g applied, the electrostatic energy of a charge Q on the particle is

$$E = QV_g + Q^2/2C \tag{D.1}$$

For negative charge Q, the first term is the attractive interaction between Q and the positively charged gate electrode, and the second term is the repulsive interaction among the bits of charge on the particle. Equation D.1 shows that the energy as a function of Q is a parabola with its minimum at $Q = -CV_g$. For simplicity I have assumed that the gate is the only electrode that contributes to C; in reality, there are other contributions.[7]

By varying V_g we can choose any value of Q_0, the charge that would minimize the energy in equation D.1 if charge were not quantized. However, because the real charge

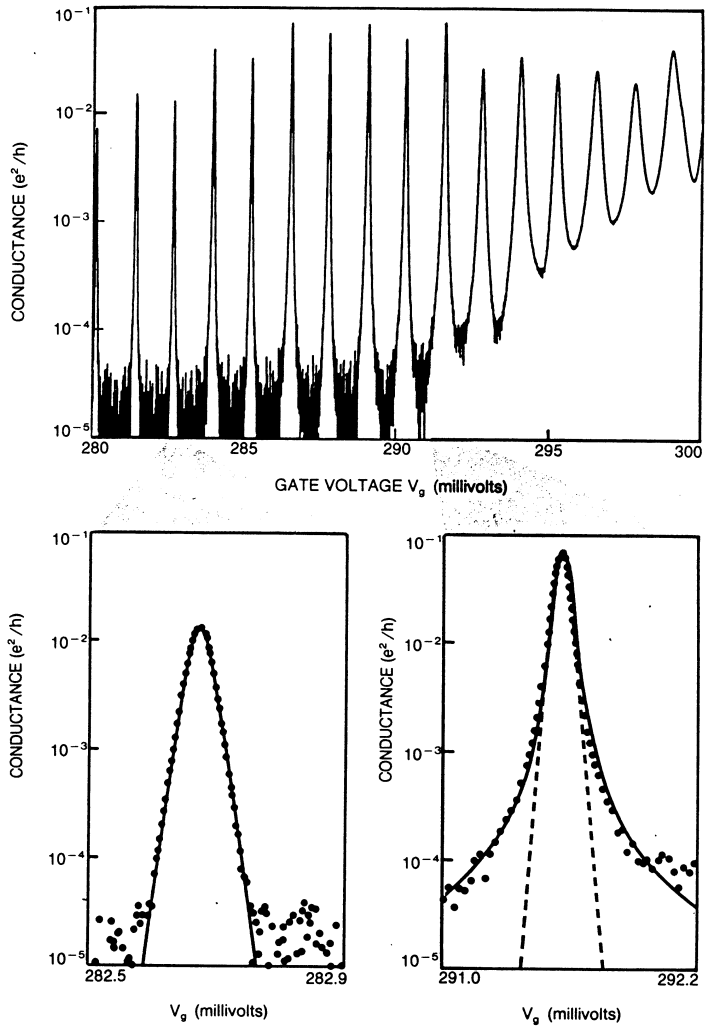


Figure D.2: Conductance of a controlled-barrier atom as a function of the voltage V_g on the gate at a temperature of 60 mK. At low V_g (solid blue curve) the shape of the resonance is given by the thermal distribution of electrons in the source that are tunneling onto the atom, but at high V_g a thermally broadened Lorentzian (red curve) is a better description than the thermal distribution alone (dashed blue curve). (Adapted from ref. 7.)

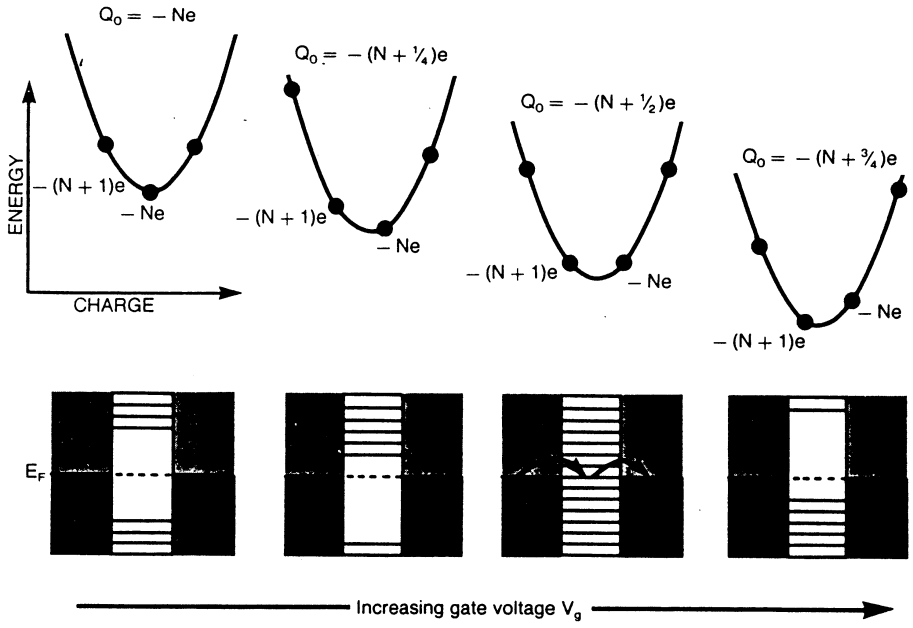


Figure D.3: Total energy (top) and tunneling energies (bottom) for an artificial atom. As voltage is increased the charge Q_0 for which the energy is minimized changes from $-Ne$ to $-(N + 1/4)e$. Only the points corresponding to discrete numbers of electrons on the atom are allowed (dots on upper curves). Lines in the lower diagram indicate energies needed for electrons or holes to tunnel onto the atom. When $Q_0 = -(N + 1/2)e$ the gap in tunneling energies vanishes and current can flow.

is quantized, only discrete values of the energy E are possible. (See figure D.3.) When $Q_0 = -Ne$, an integral number N of electrons minimizes E, and the Coulomb interaction results in the same energy difference $e^2/2C$ for increasing or decreasing N by 1. For all other values of Q_0 except $Q_0 = -(N + 1/2)e$ there is a smaller, but nonzero, energy for either adding or subtracting an electron. Under such circumstances no current can flow at low temperature. However, if $Q_0 = -(N + 1/2)e$ the state with $Q = -Ne$ and that with $Q = (N + 1)e$ are degenerate, and the charge fluctuates between the two values even at zero temperature. Consequently the energy gap in the tunneling spectrum disappears, and current can flow. The peaks in conductance are therefore periodic, occurring whenever $CV_g = Q_0 = -(N + 1/2)e$, spaced in gate voltage by e/C.

As shown in figure D.3, there is a gap in the tunneling spectrum for all values of V_g except the charge-degeneracy points. The more closely spaced discrete levels shown outside this gap are due to excited states of the electrons present on the artificial atom and will be discussed more in the next section. As V_g is increased continuously, the gap is pulled down relative to the Fermi energy until a charge degeneracy point is reached. On moving through this point there is a discontinuous change in the tunneling spectrum: The gap collapses and then reappears shifted up by e^2/C. Simultaneously the charge on the artificial atom increases by 1 and the process starts over again. A charge-degeneracy point and a conductance peak are reached every time the voltage is increased by e/C, the amount necessary to add one electron to the artificial atom. Increasing the gate voltage of an artificial atom is therefore analogous to moving through the periodic table for natural atoms by increasing the nuclear charge.

The quantization of charge on a natural atom is something we take for granted. However, if atoms were larger, the energy needed to add or remove electrons would be smaller, and the number of electrons on them would fluctuate except at very low temperature. The quantization of charge is just one of the properties that artificial atoms have in common with natural ones.

D.2 Energy quantization

The Coulomb blockade model accounts for charge quantization but ignores the quantization of energy resulting from the small size of the artificial atom. This confinement of the electrons makes the energy spacing of levels in the atom relatively large at low energies. If one thinks of the atom as a box, at the lowest energies the level spacings are of the order $\hbar^2/ma^2$, where a is the size of the box. At higher energies the spacings decrease for a three-dimensional atom because of the large number of standing electron waves possible for a given energy. If there are many electrons in the atom, they fill up many levels, and the level spacing at the Fermi energy becomes small. The all-metal atom has so many electrons (about 10^7) that the level spectrum is effectively continuous. Because of this, many experts do not regard such devices as "atoms," but I think it is helpful to think of them as being atoms in the limit in which the number of electrons is large. In the controlled-barrier atom, however, there are only about 30–60 electrons, similar to the number in natural atoms like krypton through xenon. Two-probe atoms sometimes have only one or two electrons. (There are actually many more electrons that are tightly bound to the ion cores of the semiconductor, but those are unimportant because they cannot move.) For most cases, therefore, the spectrum of energies for adding an extra electron to the atom is discrete, just

as it is for natural atoms. That is why a discrete set of levels is shown in figure D.3.

One can measure the energy level spectrum directly by observing the tunneling current at fixed V_g as a function of the voltage $V_{\rm ds}$ between drain and source. Suppose we adjust V_g so that, for example, $Q_0 = -(N + 1/4)e$ and then begin to increase $V_{\rm ds}$. The Fermi level in the source rises in proportion to $V_{\rm ds}$ relative to the drain, so it also rises relative to the energy levels of the artificial atom. (See the inset to figure D.4a.) Current begins to flow when the Fermi energy of the source is raised just above the first quantized energy level of the atom. As the Fermi energy is raised further, higher energy levels in the atom fall below it, and more current flows because there are additional channels for electrons to use for tunneling onto the artificial atom. We measure an energy level by measuring the voltage at which the current increases or, equivalently, the voltage at which there is a peak in the derivative of the current, $dI/dV_{\rm ds}$. (We need to correct for the increase in the energy of the atom with $V_{\rm ds}$, but this is a small effect.) Many beautiful tunneling spectra of this kind have been measured[5] for two-terminal atoms. Figure D.4a shows one for a controlled barrier atom.[7]

Increasing the gate voltage lowers all the energy levels in the atom by eV_g, so that the entire tunneling spectrum shifts with V_g, as sketched in figure D.3. One can observe this effect by plotting the values of $V_{\rm ds}$ at which peaks appear in $dI/dV_{\rm ds}$. (See figure D.4b.) As V_g increases you can see the gap in the tunneling spectrum shift lower and then disappear at the charge-degeneracy point, just as the Coulomb blockade model predicts. You can also see the discrete energy levels of the artificial atom. For the range of V_s shown in figure D.4 the voltage is only large enough to add or remove one electron from the atom; the discrete levels above the gap are the excited states of the atom with one extra electron, and those below the gap are the excited states of the atom with one electron missing (one hole). At still higher voltages (not shown in figure D.4) one observes levels for two extra electrons or holes and so forth. The charge-degeneracy points are the values of V_g for which one of the energy levels of the artificial atom is degenerate with the Fermi energy in the leads when $V_{\rm ds} = 0$, because only then can the charge of the atom fluctuate.

In a natural atom one has little control over the spectrum of energies for adding or removing electrons. There the electrons interact with the fixed potential of the nucleus and with each other, and these two kinds of interaction determine the spectrum. In an artificial atom, however, one can change this spectrum completely by altering the atom's geometry and composition. For the all-metal atom, which has a high density of electrons, the energy spacing between the discrete levels is so small that it can be ignored. The high density of electrons also results in a short screening length for external electric fields, so electrons added to the atom reside on its surface. Because of this, the electron-electron interaction is always e^2/C (where C is the classical geometrical capacitance), independent of the number of electrons added. This is exactly the case for which the Coulomb blockade model was invented, and it works well: The conductance peaks are perfectly periodic in the gate voltage. The difference between the "ionization potential" and the "electron affinity" is e^2/C, independent of the number of electrons on the atom.

In the controlled-barrier atom, as you can see from figure D.4, the level spacing is one or two tenths of the energy gap. The conductance peaks are not perfectly periodic in gate voltage, and the difference between ionization potential and electron affinity has a quantum mechanical contribution. I will discuss this contribution a little later in more detail.

In the two-probe atom the electron-electron interaction can be made very small, so that

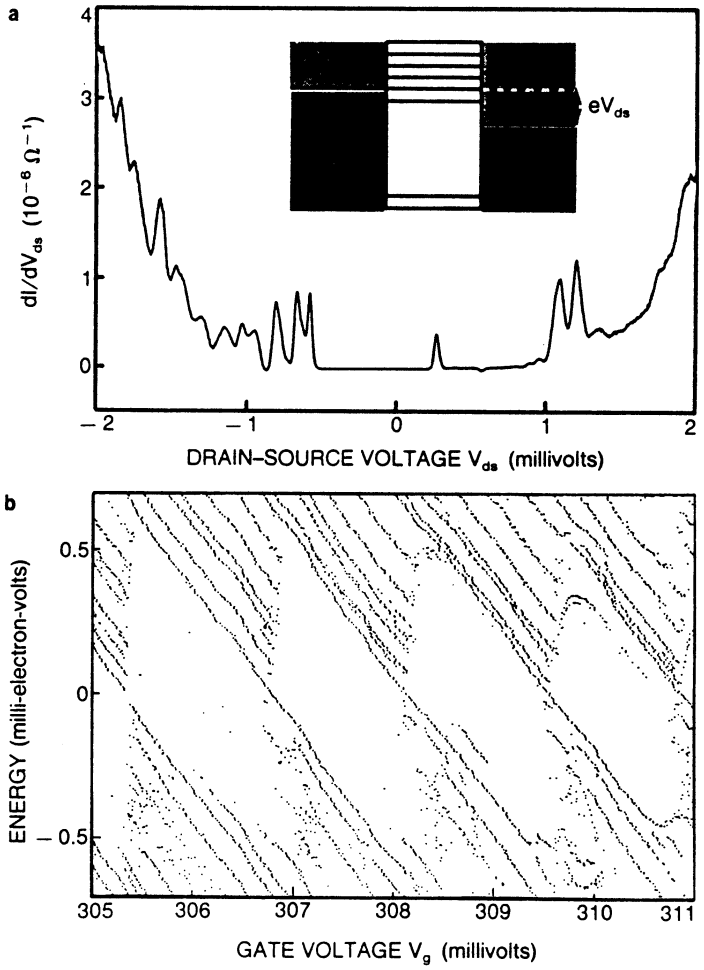


Figure D.4: Discrete energy levels of an artificial atom can be detected by varying the drain-source voltage. When a large enough $V_{\rm ds}$ is applied, electrons overcome the energy gap and tunnel from the source to the artificial atom. (See inset of a.) a: Every time a new discrete state is accessible the tunneling current increases, giving a peak in $dI/dV_{\rm ds}$. The Coulomb blockade gap is the region between about –0.5 mV and + 0.3 mV where there are no peaks. b: Plotting the positions of these peaks at various gate voltages gives the level spectrum. Note how the levels and the gap move downward as V_g increases, just as sketched in the lower part of figure D.3. (Adapted from ref. 7.)

one can in principle reach the limit opposite to that of the all-metal atom. One can find the energy levels of a two-probe atom by measuring the capacitance between its two leads as a function of the voltage between them.[6] When no tunneling occurs, this capacitance is the series combination of the source-atom and atom-drain capacitances. For capacitance measurements, two-probe atoms are made with the insulating layer between the drain and atom so thick that current cannot flow under any circumstances. Whenever the Fermi level in the source lines up with one of the energy levels of the atom, however, electrons can tunnel freely back and forth between the atom and the source. This causes the total capacitance to increase, because the source-atom capacitor is effectively shorted by the tunneling current. The amazing thing about this experiment is that a peak occurs in the capacitance every time a single electron is added to the atom. (See figure D.5a.) The voltages at which the peaks occur give the energies for adding electrons to the atom, just as the voltages for peaks in $dI/dV_{\rm ds}$ do for the controlled-barrier atom or for a two probe atom in which both the source-atom barrier and the atom-drain barrier are thin enough for tunneling. The first peak in figure D.5a corresponds to the one-electron artificial atom.

Figure D.5b shows how the energies for adding electrons to a two-probe atom vary with a magnetic field perpendicular to the GaAs layer. In an all-metal atom the levels would be equally spaced, by e^2/C, and would be independent of magnetic field because the electron-electron interaction completely determines the energy. By contrast, the levels of the two-probe atom are irregularly spaced and depend on the magnetic field in a systematic way. For the two-probe atom the fixed potential determines the energies at zero field. The level spacings are irregular because the potential is not highly symmetric and varies at random inside the atom because of charged impurities in the GaAs and AlGaAs. It is clear that the electron-electron interactions that are the source of the Coulomb blockade are not always so important in the two-probe atom as in the all-metal and controlled-barrier atoms. Their relative importance depends in detail on the geometry.[5]

D.3 Artificial atoms in a magnetic field

Level spectra for natural atoms can be calculated theoretically with great accuracy, and it would be nice to be able to do the same for artificial atoms. No one has yet calculated. an entire spectrum, like that in figure D.4a. However, for a simple geometry we can now predict the charge-degeneracy points, the values of V_g corresponding to conductance peaks like those in figure D.2. From the earlier discussion it should be clear that in such a calculation one must take into account the electron's interactions with both the fixed potential and the other electrons.

The simplest way to do this is with an extension of the Coulomb blockade model.[11–13] It is assumed, as before, that the contribution to the gap in the tunneling spectrum from the Coulomb interaction is e^2/C no matter how many electrons are added to the atom. To account for the discrete levels one pretends that once on the atom, each electron interacts independently with the fixed potential. All one has to do is solve for the energy levels of a single electron in the fixed potential that creates the artificial atom and then fill those levels in accordance with the Pauli exclusion principle. Because the electron-electron interaction is assumed always to be e^2/C, this is called the constant-interaction model.

Now think about what happens when one adds electrons to a controlled-barrier atom by increasing the gate voltage while keeping $V_{\rm ds}$ just large enough so one can measure the

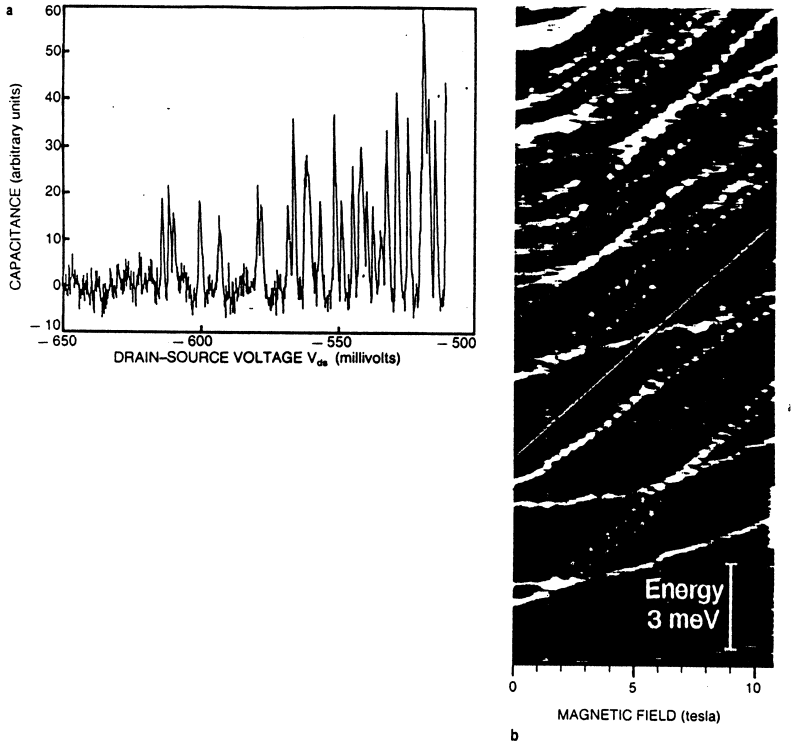


Figure D.5: Capacitance of a two-probe atom that has only one barrier thin enough to allow tunneling. a: The capacitance has a peak every time a single electron is added to the atom. The positions of the peaks give the energy spectrum of the atom. b: Peaks in capacitance plotted versus applied magnetic field. The green line indicates the rate of change of the energy expected when the magnetic field dominates. (Adapted from ref. 6.)

conductance. When there are $N-1$ electrons on the atom the $N-1$ lowest energy levels are filled. The next conductance peak occurs when the gate voltage pulls the energy of the atom down enough that the Fermi level in the source and drain becomes degenerate with the Nth level. Only when an energy level is degenerate with the Fermi energy can current flow; this is the condition for a conductance peak. When V_g is increased further and the next conductance peak is reached, there are N electrons on the atom, and the Fermi level is degenerate with the $(N+1)$-th level. Therefore to get from one peak to the next the Fermi energy must be raised by $e^2/C + (E_{N+1} - E_N)$, where E_N, is the energy of the Nth level of the atom. If the energy levels are closely spaced the Coulomb blockade result is recovered, but in general the level spacing contributes to the energy between successive conductance peaks.

It turns out that we can test the results of this kind of calculation best if a magnetic field is applied perpendicular to the GaAs layer. For free electrons in two dimensions, applying the magnetic field results in the spectrum of Landau levels with energies $(n+1/2)\hbar\omega_c$ where the cyclotron frequency is $\omega_c = eB/m^*c$, and m^* is the effective mass of the electrons. In the controlled barrier atom and the two-probe atom, we expect levels that behave like Landau levels at high fields, with energies that increase linearly in B. This behavior occurs because when the field is large enough the cyclotron radius is much smaller than the size of the electrostatic potential well that confines the electrons, and the electrons act as if they were free. Levels shifting proportionally to B, as expected, are seen experimentally. (See figure D.5b.)

To calculate the level spectrum we need to model the fixed potential, the analog of the potential from the nucleus of a natural atom. The simplest choice is a harmonic oscillator potential, and this turns out to be a good approximation for the controlled-barrier atom. Figure D.6a shows the calculated level spectrum as a function of magnetic field for non-interacting electrons in a two dimensional harmonic oscillator potential. At low fields the energy levels dance around wildly with magnetic field. This occurs because some states have large angular momentum and the resulting magnetic moment causes their energies to shift up or down strongly with magnetic field. As the field is increased, however, things settle down. For most of the field range shown there are four families of levels, two moving up, the other two down. At the highest fields there are only two families, corresponding to the two possible spin states of the electron.

Suppose we measure, in an experiment like the one whose results are shown in figure D.2, the gate voltage at which a specific peak occurs as a function of magnetic field. This value of V_g is the voltage at which the Nth energy level is degenerate with the Fermi energy in the source and drain. A shift in the energy of the level will cause a shift in the peak position. The blue line in figure D.6a is the calculated energy of the 39th level (chosen fairly arbitrarily for illustration purposes), so it gives the prediction of the constant-interaction model for the position of the 39th conductance peak. As the magnetic field increases, levels moving up in energy cross those moving down, but the number of electrons is fixed, so electrons jump from upward-moving filled levels to downward-moving empty ones. The peak always follows the 39th level, so it moves up and down in gate voltage.

Figure D.6b shows a measurement[14] of V_g for one conductance maximum, like one of those in figure D.2, as a function of B. The behavior is qualitatively similar to that predicted by the constant-interaction model: The peak mover up and down with increasing B, and the frequency of level crossings changes at the field where only the last two families of levels

remain. However, at high B the frequency is predicted to be much lower than what is observed experimentally. While the constant-interaction model is in qualitative agreement with experiment, it is not quantitatively correct.

To anyone who has studied atomic physics, the constant-interaction model seems quite crude. Even the simplest models used to calculate energies of many electron atoms determine the charge density and potential self- consistently. One begins by calculating the charge density that would result from noninteracting electrons in the fixed potential, and then one calculates the effective potential an electron sees because of the fixed potential and the potential resulting from this charge density. Then one calculates the charge density again. One does this repeatedly until the charge density and potential are self-consistent. The constant-interaction model fails because it is not self- consistent. Figure D.6c shows the results of a self-consistent calculation for the controlled-barrier atom.[14] It is in good agreement with experiment-much better agreement than the constant-interaction model gives.

D.4 Conductance line shapes

In atomic physics, the next step after predicting energy levels is to explore how an atom interacts with the electromagnetic field, because the absorption and emission of photons teaches us the most about atoms. For artificial atoms, absorption and emission of electrons plays this role, so we had better understand how this process works. Think about what happens when the gate voltage in the controlled-barrier atom is set at a conductance peak, and an electron is tunneling back and forth between the atom and the leads. Since the electron spends only a finite time τ on the atom, the uncertainty principle tells us that the energy level of the electron has a width $\hbar/\tau$. Furthermore, since the probability of finding the electron on the atom decays as $e^{t/\tau}$, the level will have a Lorentzian line shape.

This line shape can be measured from the transmission probability spectrum $T(E)$ of electrons with energy E incident on the artificial atom from the source. The spectrum is given by

$$T(E) = \frac{\Gamma^2}{\Gamma^2 + (E - E_N)^2} \tag{D.2}$$

where Γ is approximately $\hbar/\tau$ and E_N is the energy of the Nth level. The probability that electrons are transmitted from the source to the drain is approximately,proportional[15] to the conductance G. In fact, $G \simeq (e^2/h)T$, where e^2/h is the quantum of conductance. It is easy to show that one must have $G < e^2/h$ for each of the barriers separately to observe conductance resonances. (An equivalent argument is used to show that electrons in a disordered conductor are localized for $G < e^2/h$. See, for example, the article by Boris L. Al'tshuler and Patrick A. Lee in PHYSICS TODAY, December 1988, page 36.) This condition is equivalent to requiring that the separation of the levels is greater than their width Γ.

Like any spectroscopy, our electron spectroscopy of artificial atoms has a finite resolution. The resolution is determined by the energy spread of the electrons in the source, which are trying to tunnel into the artificial atom. These electrons are distributed according to the Fermi-Dirac function,

$$f(E) = \frac{1}{\exp[(E - E_F)/kT] + 1} \tag{D.3}$$

where E_F is the Fermi energy. The tunneling current is given by

$$I = \int \frac{e}{h} T(E)[f(E) - f(E - eV_{\rm ds})]dE. \tag{D.4}$$

Equation D.4 says that the net current is proportional to the probability $f(E)T(E)$ that there is an electron in the source with energy E and that the electron can tunnel between the source and drain minus the equivalent probability for electrons going from drain to source. The best resolution is achieved by making $V_{\rm ds} \ll kT$. Then $[f(E) - f(E - eV_{\rm ds})] \simeq eV_{\rm ds}(df/dE)$, and I is proportional to $V_{\rm ds}$, so the conductance is $I/V_{\rm ds}$.

Figure D.2 shows that equations D.2–D.4 describe the experiments well: At low V_g, where Γ is much less than kT, the shape of the conductance resonance is given by the resolution function df/dE. But at higher V_g one sees the Lorentzian tails of the natural line shape quite clearly. The width Γ depends exponentially on the height and width of the potential barrier, as is usual for tunneling. The height of the tunnel barrier decreases with V_g, which is why the peaks become broader with increasing V_g. Just as we have control over the level spacing in artificial atoms, we also can control the coupling to the leads and therefore the level widths. It is clear why the present generation of artificial atoms show unusual behavior only at low temperatures: When kT becomes comparable to the energy separation between resonances, the peaks overlap and the features disappear.

D.5 Applications

The behavior of artificial atoms is so unusual that it is natural to ask whether they will be useful for applications to electronics. Some clever things can be done., Because of the electron-electron interaction, only one electron at a time can pass through the atom. With devices like the "turnstile" device[16,17] shown on the cover of this issue the two tunnel barriers can be raised and lowered independently. Suppose the two barriers are raised and lowered sequentially at a radio or microwave frequency ν. Then, with a small source-drain voltage applied, an electron will tunnel onto the atom when the source-atom barrier is low and off it when the atom-drain barrier is low. One electron will pass in each time interval ν^{-1}, producing a current eV. Other applications, such as sensitive electrometers, can be imagined.[9,18] However, the most interesting applications may involve devices in which several artificial atoms are coupled together to form artificial molecules[16,17,19] or in which many are coupled to form artificial solids. Because the coupling between the artificial atoms can be controlled, new physics as well as new applications may emerge. The age of artificial atoms has only just begun.

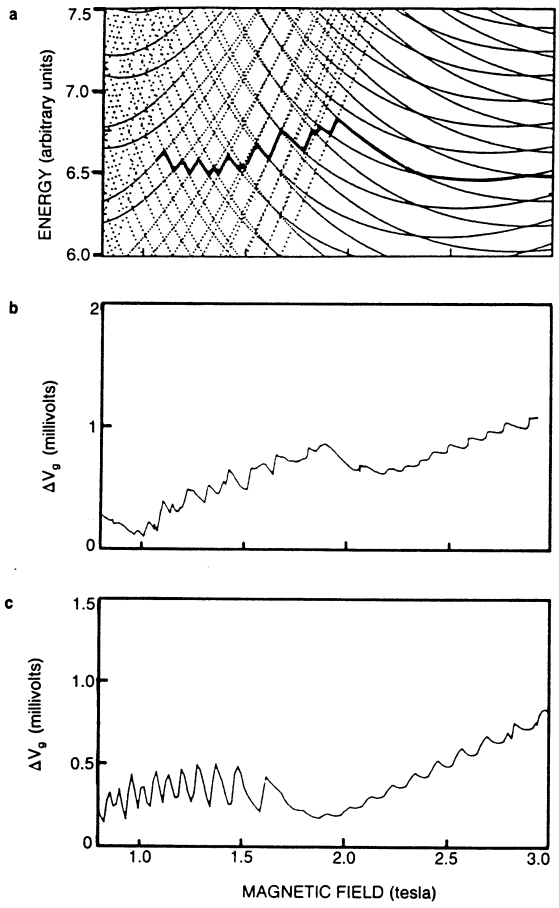


Figure D.6: Effect of magnetic field on energy level spectrum and conductance peaks. a: Calculated level spectrum for noninteracting electrons in a harmonic oscillator electrostatic potential as a function of magnetic field. The blue line is the prediction that the constant interaction model gives for the gate voltage for the 39th conductance peak. b: Measured position of a conductance peak in a controlled-barrier atom as a function of field. c: Position of the 39th conductance peak versus field, calculated self-consistently. The scale in c does not match that in b because parameters in the calculation were not precisely matched to the experimental conditions. (Adapted from Ref. 14.)

References

1. T. A. Fulton, G. J. Dolan, Phys. Rev. Lett. 59, 109 (1987).
2. U. Meirav, M. A. Kastner, S. J. Wind, Phys. Rev. Lett. 65, 771 (1990). M. A. Kastner, Rev. Mod. Phys. 64, 849 (1992).
3. L. P. Kouwenhoven, N. C. van der Vaart, A. T. Johnson, W. Kool, C. J. P. M. Harmans, J. G. Williamson, A. A. M. Staring, C. T. Foxon, Z. Phys. B 85, 367 (1991), and refs. therein.
4. V. Chandrasekhar, Z. Ovadyahu, R. A. Webb, Phys. Rev. Lett. 67, 2862 (1991). R. J. Brown, M. Pepper, H. Ahmed, D. G. Hasko, D. A. Ritchie, J. E. F. Frost, D. C. Peacock, G. A. C. Jones, J. Phys.: Condensed Matter 2, 2105 (1990).
5. B. Su, V. J. Goldman, J. E. Cunningham, Science 255, 313 (1992). M. A. Reed, J. N. Randall, R. J. Aggarwal, R. J. Matyi, T. M. Moore, A. E. Wetsel, Phys. Rev. Lett. 60, 535 (1988). M. Tewordt, L. Martin-Moreno, J. T. Nicholls, M. Pepper, M. J. Kelly, V. J. Law, D. A. Ritchie, J. E. F. Frost, G. A. C. Jones, Phys. Rev. B 45, 14407 (1992).
6. R. C. Ashoori, H. L. Stormer, J. S. Weiner, L. N. Pfeiffer, S. J. Pearton, K. Baldwin, K. W. West, Phys. Rev. Lett. 68, 3088 (1992).
7. E. B. Foxman, P. L. McEuen, U. Meirav, N. S. Wingreen, Y. Meir, P. A. Belk, N. R. Belk, M. A. Kastner, S. J. Wind, "The Effects of Quantum Levels on Transport Through a Coulomb Island," MIT preprint (July 1992). See also A. T. Johnson, L. P. Kouwenhoven, W. de Jong, N.C. van der Vaart, C. J. P. M. Harmans, C. T. Faxon, Phys. Rev. Lett. 69, 1592 (1992).
8. A. Kumar, Surf. Sci. 263, 335 (1992). A. Kumar, S. E. Laux, F. Stern, Appl. Phys. Lett. 54, 1270 (1989).
9. D. V. Averin, K. K. Likharev, in Mesmcopic Phenomena in Solids, B. L. Al'tshuler, P. A. Lee, R. A. Webb, eds., Elsevier, Amsterdam (1991), p. 173.
10. H. van Houton, C. W. J. Beenakker, Phys. Rev. Lett. 63,1893 (1989).
11. D. V. Averin, A. N. Korotkov, Zh. Eksp. Teor. Fiz. 97, 1661 (1990) [Sov. Phys. JETP 70, 937 (1990)].
12. Y. Meir, N. S. Wingreen, P. A. Lee, Phys. Rev. Lett. 66, 3048 (1991).
13. C. J. Beenakker, Phys. Rev. B 44, 1646 (1991).
14 P.L. McEuen, E. B. Foxman, J. Kinaret, U. Meirav, M. A. Kastner, N. S. Wingreen, S. J. Wind, Phys. Rev. B 45, 11419 (1992).
15. R. Landauer, IBM J. Res. Dev. 1, 223 (1957).
16. L. P. Kouwenhoven, A. T. Johnson, N. C. van der Vaart, W. Kool, C. J. P. M. Harmans, C. T. Foxon, Phys. Rev. Lett. 67, 1626 (1991).
17. L. J. Geerligs, V. F. Anderegg, P A. M. Holweg, J. E. Mooij, H. Pothier, D. Esteve, C. Urbina, M. H. Devoret, Phys. Rev. Lett. 64, 2691 (1990).
18. H. Grabert, M. H. Devoret, eds., Single Charge Tunneling, Plenum, New York (1992).
19. R.J. Haug, J.M. Hong, K.Y.Lee, Surf. Sci. 263, 415 (1991).

Appendix E

Transport in 1D materials

E.1 Carbon Nanotubes

The nanometer dimensions of the carbon nanotubes together with the unique electronic structure of a graphene sheet make the electronic properties of these one-dimensional structures highly unusual. This Chapter reviews some theoretical work on the relation between the atomic structure and the electronic and transport properties of single-walled carbon nanotubes. In addition to the ideal tubes, results on the quantum conductance of nanotube junctions and tubes with defects will be discussed. On-tube metal-semiconductor, semiconductor-semiconductor, and metal-metal junctions have been studied. Other defects such as substitutional impurities and pentagon-heptagon defect pairs on tube walls are shown to produce interesting effects on the conductance. The effects of static external perturbations on the transport properties of metallic nanotubes and doped semiconducting nanotubes are examined, with the metallic tubes being much less affected by long-range disorder. The structure and properties of crossed nanotube junctions and ropes of nanotubes have also been studied. The rich interplay between the structural and the electronic properties of carbon nanotubes gives rise to new phenomena and the possibility of nanoscale device applications.

E.2 Introduction

Carbon nanotubes are tubular structures that are typically several nanometers in diameter and many microns in length. This fascinating new class of materials was first discovered by S. Iijima [1] in the soot produced in the arc-discharge synthesis of fullerenes. Because of their nanometer dimensions, there are many interesting and often unexpected properties associated with these structures, and hence there is the possibility of using them to study new phenomena and employing them in applications [2, 3, 4]. In addition to the multi-walled tubes, single-walled nanotubes [5, 6, 7], and ropes of close-packed single-walled tubes have been synthesized. [8] Also, carbon nanotubes may be filled with foreign materials [9, 10] or collapsed into flat, flexible nanoribbons [11]. Carbon nanotubes are highly unusual electrical conductors, the strongest known fibers, and excellent thermal conductors. Many potentially important applications have been explored, including the use of nanotubes as nanoprobe tips [12], field emitters [13, 14], storage or filtering media [15], and nanoscale electronic

devices [16, 17, 18, 19, 20, 21, 22, 23, 24]. Further, it has been found that nanotubes may also be formed with other layered materials [25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36]. In particular, BN, BC_3, and other $B_xC_yN_z$ nanotubes have been theoretically predicted [25, 26, 27, 28, 29] and experimentally synthesized [30, 31, 32, 33, 34, 35].

Many different aspects of carbon nanotubes are treated in the various sections of this Volume. In this Chapter, we focus on a review of some selected theoretical studies on the electronic and transport properties of carbon nanotube structures, in particular, those of junctions, impurities, and other defects. Structures such as ropes of nanotubes and crossed nanotubes are also discussed.

The organization of the Chapter is as follows. Section 2 contains an introduction to the geometric and electronic structure of ideal single-walled carbon nanotubes. Section 3 gives a discussion of the electronic and transport properties of various on-tube structures. Topics presented include on-tube junctions, impurities, and local defects. On-tube metal-semiconductor, semiconductor-semiconductor, and metal-metal junctions may be formed by introducing topological structural defects. These junctions have been shown to behave like nanoscale device elements. Other defects such as substitutional impurities and Stone–Wales defects on tube walls also are shown to produce interesting effects on the conductance. Crossed nanotubes provide another means to obtain junction behavior. The crossed-tube junctions, nanotube ropes, and effects of long-range disorder are the subjects of Section 4. Intertube interactions strongly modify the electronic properties of a rope. The effects of long-range disorder on metallic nanotubes are quite different from those on doped semiconducting tubes. Finally, a summary and some conclusions are given in Section 5.

E.3 Geometric and Electronic Structure of Carbon Nanotubes

In this Section, we give an introduction to the structure and electronic properties of the single-walled carbon nanotubes (SWNTs). Shortly after the discovery of the carbon nanotubes in the soot of fullerene synthesis, single-walled carbon nanotubes were synthesized in abundance using arc discharge methods with transition metal catalysts [5, 6, 7]. These tubes have quite small and uniform diameter, on the order of one nanometer. Crystalline ropes of single-walled nanotubes with each rope containing tens to hundreds of tubes of similar diameter closely packed have also been synthesized using a laser vaporization method [8] and other techniques, such as arc-discharge and CVD techniques. These developments have provided ample amounts of sufficiently characterized samples for the study of the fundamental properties of the SWNTs. As illustrated in Fig. 1, a single-walled carbon nanotube is geometrically just a rolled up graphene strip. Its structure can be specified or indexed by its circumferential periodicity [37]. In this way, a SWNT's geometry is completely specified by a pair of integers (n, m) denoting the relative position $\vec{c} = n\vec{a}_1 + m\vec{a}_2$ of the pair of atoms on a graphene strip which, when rolled onto each other, form a tube.

Theoretical calculations [**?**, 39, 40, 41] have shown early on that the electronic properties of the carbon nanotubes are very sensitive to their geometric structure. Although graphene is a zero-gap semiconductor, theory has predicted that the carbon nanotubes can be metals or semiconductors with different size energy gaps, depending very sensitively on the diameter and helicity of the tubes, i.e., on the indices (n, m). As seen below, the intimate connection between the electronic and geometric structure of the carbon nanotubes gives rise to many of the fascinating properties of various nanotube structures, in particular nanotube junctions.

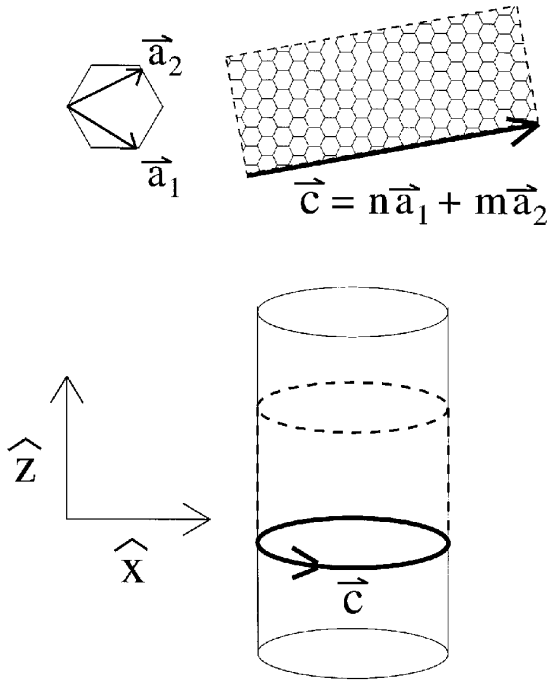


Figure E.1: Geometric structure of an (n, m) single-walled carbon nanotube

The physics behind this sensitivity of the electronic properties of carbon nanotubes to their structure can be understood within a band-folding picture. It is due to the unique band structure of a graphene sheet, which has states crossing the Fermi level at only 2 inequivalent points in k-space, and to the quantization of the electron wavevector along the circumferential direction. An isolated sheet of graphite is a zero-gap semiconductor whose electronic structure near the Fermi energy is given by an occupied π band and an empty π^* band. These two bands have linear dispersion and, as shown in Fig. 2, meet at the Fermi level at the K point in the Brillouin zone. The Fermi surface of an ideal graphite sheet consists of the six corner K points. When forming a tube, owing to the periodic boundary conditions imposed in the circumferential direction, only a certain set of $\vec{k}$ states of the planar graphite sheet is allowed. The allowed set of k's, indicated by the lines in Fig. 2, depends on the diameter and helicity of the tube. Whenever the allowed k's include the point K, the system is a metal with a nonzero density of states at the Fermi level, resulting in a one-dimensional metal with 2 linear dispersing bands. When the point K is not included, the system is a semiconductor with different size energy gaps. It is important to note that the states near the Fermi energy in both the metallic and the semiconducting tubes are all from states near the K point, and hence their transport and other properties are related to the properties of the states on the allowed lines. For example, the conduction band and valence bands of a semiconducting tube come from states along the line closest to the K point.

The general rules for the metallicity of the single-walled carbon nanotubes are as follows:

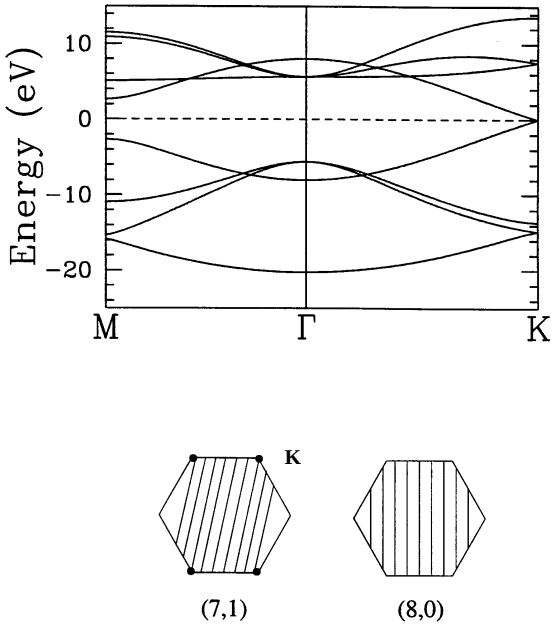


Figure E.2: (Top) Tight-binding band structure of graphene (a single basal plane of graphite). (Bottom) Allowed $\vec{k}$-vectors of the (7,1) and (8,0) tubes (solid lines) mapped onto the graphite Brillouin zone.

(n, n) tubes are metals; (n, m) tubes with $n - m = 3j$, where j is a nonzero integer, are very tiny-gap semiconductors; and all others are large-gap semiconductors. Strictly within the band-folding scheme, the $n - m = 3j$ tubes would all be metals, but because of tube curvature effects, a tiny gap opens for the case that j is nonzero. Hence, carbon nanotubes come in three varieties: large-gap, tiny-gap, and zero gap. The (n, n) tubes, also known as armchair tubes, are always metallic within the single-electron picture, independent of curvature because of their symmetry. As the tube radius R increases, the band gaps of the large-gap and tiny-gap varieties decreases with a $1/R$ and $1/R^2$ dependence, respectively. Thus, for most experimentally observed carbon nanotube sizes, the gap in the tiny-gap variety which arises from curvature effects would be so small that, for most practical purposes, all the $n - m = 3j$ tubes can be considered as metallic at room temperature. Thus, in Fig. 2, a (7,1) tube would be metallic, whereas a (8,0) tube would be semiconducting.

This band-folding picture, which was first verified by tight-binding calculations [38, 39, 40], is expected to be valid for larger diameter tubes. However, for a small radius tube, because of its curvature, strong rehybridization among the σ and π states can modify the electronic structure. Experimentally, nanotubes with a radius as small as 3.5 Å have been produced. *Ab initio* pseudopotential local density functional (LDA) calculations [41] indeed revealed that sufficiently strong hybridization effects can occur in small radius nanotubes which significantly alter their electronic structure. Strongly modified low-lying conduction band states are introduced into the band gap of insulating tubes because of hybridization of the σ^* and π^* states. As a result, the energy gaps of some small radius tubes are decreased by more than 50%. For example, the (6,0) tube which is predicted to be semiconducting in the band-folding scheme is shown to be metallic. For nanotubes with diameters greater

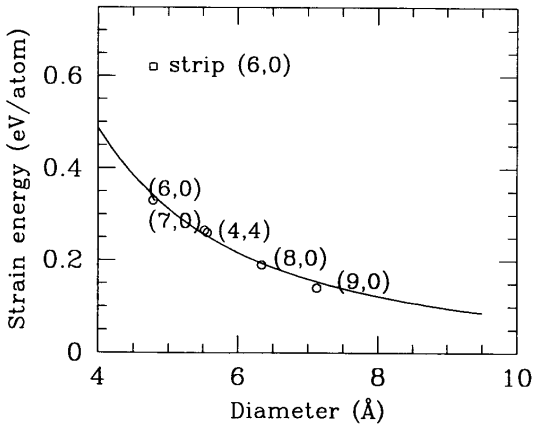


Figure E.3: Strain energy/atom for carbon nanotubes from *ab initio* total energy calculations [44]

than 1 nm, these rehybridization effects are unimportant. Strong σ-π rehybridization can also be induced by bending a nanotube [42]

Energetically, *ab initio* total energy calculations have shown that carbon nanotubes are stable down to very small diameters. Figure 3 depicts the calculated strain energy per atom for different carbon nanotubes of various diameters [41]. The strain energy scales nearly perfectly as d^{-2} where d is the tube diameter (solid curve in Fig. 3), as would be the case for rolling a classical elastic sheet. Thus, for the structural energy of the carbon nanotubes, the elasticity picture holds down to a subnanometer scale. The elastic constant may be determined from the total energy calculations. This result has been used to analyze collapsed tubes [11] and other structural properties of nanotubes. Also shown in Fig. 3 is the energy/atom for a (6,0) carbon strip. It has an energy which is well above that of a (6,0) tube because of the dangling bonds on the strip edges. Because in general the energy per atom of a strip scales as d^{-1}, the calculation predicts that carbon nanotubes will be stable with respect to the formation of strips down to below 4 Å in diameter, in agreement with classical force-field calculations [43].

There have been many experimental studies on carbon nanotubes in an attempt to understand their electronic properties. The transport experiments [19, 20, 45, 46, 47] involved both two- and four-probe measurements on a number of different tubes, including multiwalled tubes, bundles of single-walled tubes, and individual single-walled tubes. Measurements showed that there are a variety of resistivity behaviors for the different tubes, consistent with the above theoretical picture of having both semiconducting and metallic tubes. In particular, at low temperature, individual metallic tubes or small ropes of metallic tubes act like quantum wires [19, 20]. That is, the conduction appears to occur through well-separated discrete electron states that are quantum-mechanically coherent over distances exceeding many hundreds of nanometers. At sufficiently low temperature, the system behaves like an elongated quantum dot.

Figure 4 depicts the experimental set up for such a low temperature transport measurement on a single-walled nanotube rope from Ref. [20]. At a few degrees Kelvin, the low-bias conductance of the system is suppressed for voltages less than a few millivolts, and there

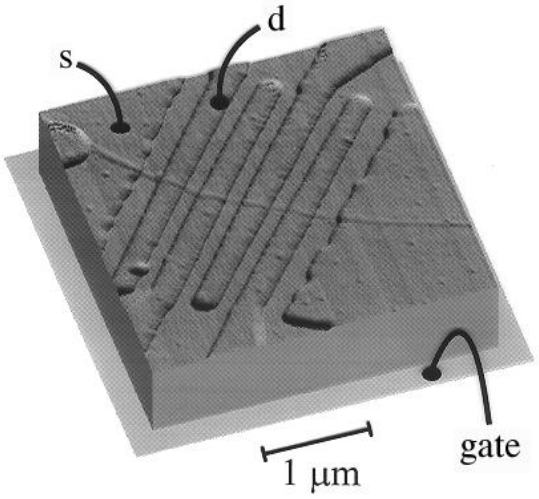


Figure E.4: Experimental set-up for the electrical measurement of a single-walled nanotube rope, visible as the diagonal curved line [20]

are dramatic peaks in the conductance as a function of gate voltage that modulates the number of electrons in the rope. (See Fig. 5.) These results have been interpreted in terms of single-electron charging and resonant tunneling through the quantized energy levels of the nanotubes. The data are explained quite well using the band structure of the conducting electrons of a metallic tube, but these electrons are confined to a small region defined either by the contacts or by the sample length, thus leading to the observed quantum confinement effects of Coulomb blockade and resonant tunneling.

There have also been high resolution low temperature scanning tunneling microscopy (STM) studies, which directly probe the relationship between the structural and electronic properties of the carbon nanotubes [48, 49]. Figure 6 is a STM image for a single carbon nanotube at 77 K on the surface of a rope. In these measurements, the resolution of the measurements allowed for the identification of the individual carbon rings. From the orientation of the carbon rings and the diameter of the tube, the geometric structure of the tube depicted in Fig. 6 was deduced to be that of a (11,2) tube. Measurement of the normalized conductance in the scanning tunneling spectroscopy (STS) mode was then used to obtain the local density of states (LDOS). Data on the (11,2) and the (12,3) nanotubes gave a constant density of states at the Fermi level, showing that they are metals as predicted by theory. On another sample, a $(14, -3)$ tube was studied. Since 14+3 is not equal to 3 times an integer, it ought to be a semiconductor. Indeed, the STS measurement gives a band gap of 0.75 eV, in very good agreement with calculations.

The electronic states of the carbon nanotubes, being band-folded states of graphene, lead to other interesting consequences, including a striking geometry dependence of the electric polarizability. Figure 7 presents some results from a tight-binding calculation for the static polarizabilities of carbon nanotubes in a uniform applied electric field [50]. Results for 17 single-walled tubes of varying size and chirality, and hence varying band gaps, are given. The unscreened polarizability α_0 is calculated within the random phase approximation. The cylindrical symmetry of the tubes allows the polarizability tensor to be divided into

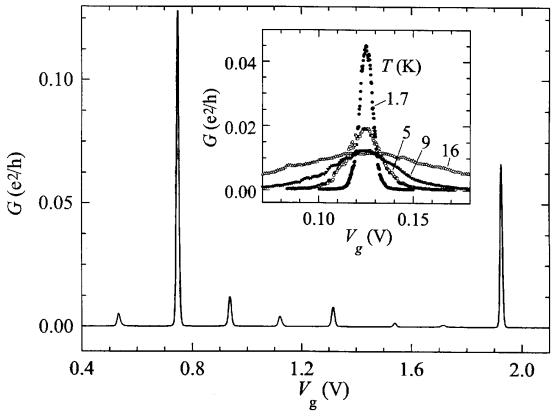


Figure E.5: Measured conductance of a single-walled carbon nanotube rope as a function of gate voltage [20]

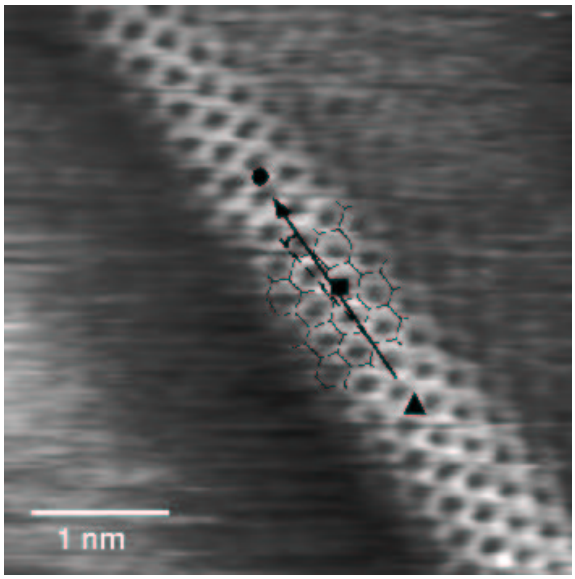


Figure E.6: STM images at 77 K of a single-walled carbon nanotube at the surface of a rope [49]

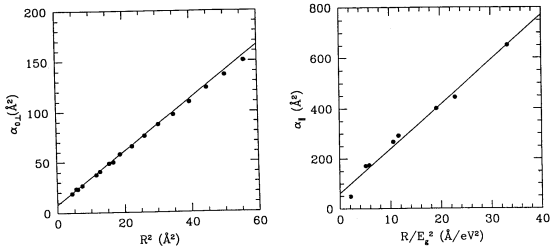


Figure E.7: Calculated static polarizability of single wall carbon nanotubes, showing results both for $\alpha_{0\perp}$ vs R^2 on the left and $\alpha_{\|}$ vs R/E_g^2 on the right [50]

components perpendicular to the tube axis, $\alpha_{0\perp}$, and a component parallel to the tube axis, $\alpha_{0\|}$. Values for $\alpha_{0\perp}$ predicted within this model are found to be totally independent of the band gap E_g and to scale linearly as R^2, where R is the tube radius. The latter dependence may be understood from classical arguments, but the former is rather unexpected. The insensitivity of $\alpha_{0\perp}$ to E_g results from selection rules in the dipole matrix elements between the highest occupied and the lowest unoccupied states of these tubes. On the other hand, Fig. 7 shows that $\alpha_{0\|}$ is proportional to R/E_g^2, which is consistent with the static dielectric response of standard insulators. Also, using arguments analogous to those for C_{60} [51, 52], local field effects relevant to the screened polarizability tensor α may be included classically, resulting in a saturation of $\alpha_\perp$ for large $\alpha_{0\perp}$, but leaving $\alpha_{\|}$ unaffected. Thus, in general, the polarizability tensor of a carbon nanotube is expected to be highly anisotropic with $\alpha_{\|} \gg \alpha_\perp$. And the polarizability of small gap tubes is expected to be greatly enhanced among tubes of similar radius.

Just as for the electronic states, the phonon states in carbon nanotubes are also quantized into phonon subbands. This has led to a number of interesting phenomena [2, 3] which are discussed elsewhere in this Volume [53]. Here we mention several of them. It has been shown that twisting motions of a tube can lead to the opening up of a minuscule gap at the Fermi level, leading to the possibility of strong coupling between the electronic states and the twisting modes or twistons [54]. The heat capacity of the nanotubes is also expected to show a dimensionality dependence. Analysis [55] shows that the phonon contributions dominate the heat capacity, with single-walled carbon nanotubes having a $C_{\rm ph} \sim T$ dependence at low temperature. The temperature below which this should be observable decreases with increasing nanotube radius R, but the linear T dependence should be accessible to experimental investigations with presently available samples. In particular, a tube with a 100 Å radius should have $C_{\rm ph} \sim T$ for $T < 7\,{\rm K}$. Since bulk graphite has $C_{\rm ph} \sim T^{2-3}$, a sample of sufficiently small radius tubes should show a deviation from graphitic behavior. Multi-walled tubes, on the other hand, are expected to show a range of behavior intermediate between $C_{\rm ph} \sim T$ and $C_{\rm ph} \sim T^{2-3}$, depending in detail on the tube radii and the number of concentric walls.

In addition to their fascinating electronic properties, carbon nanotubes are found to have exceptional mechanical properties [56]. Both theoretical [57, 58, 59, 60, 61, 62, 63] and experimental [64, 65] studies have demonstrated that they are the strongest known fibers. Carbon nanotubes are expected to be extremely strong along their axes because of the strength of the carbon-carbon bonds. Indeed, the Young's modulus of carbon nanotubes

has been predicted and measured to be more than an order of magnitude higher than that of steel and several times that of common commercial carbon fibers. Similarly, BN nanotubes are shown [66] to be the world's strongest large-gap insulating fiber.

E.4 Electronic and Transport Properties of On-tube Structures

In this section, we discuss the electronic properties and quantum conductance of nanotube structures that are more complex than infinitely long, perfect nanotubes. Many of these systems exhibit novel properties and some of them are potentially useful as nanoscale devices.

E.4.1 Nanotube junctions

Since carbon nanotubes are metals or semiconductors depending sensitively on their structures, they can be used to form metal-semiconductor, semiconductor-semiconductor, or metal-metal junctions. These junctions have great potential for applications since they are of nanoscale dimensions and made entirely of a single element. In constructing this kind of on-tube junction, the key is to join two half-tubes of different helicity seamlessly with each other, without too much cost in energy or disruption in structure. It has been shown that the introduction of pentagon-heptagon pair defects into the hexagonal network of a single carbon nanotube can change the helicity of the carbon nanotube and fundamentally alter its electronic structure [16, 17, 18, 67, 68, 69, 70, 71]. This led to the prediction that these defective nanotubes behave as the desired nanoscale metal-semiconductor Schottky barriers, semiconductor heterojunctions, or metal-metal junctions with novel properties, and that they could be the building blocks of nanoscale electronic devices.

In the case of nanotubes, being one-dimensional structures, a local topological defect can change the properties of the tube at an infinitely long distance away from the defect. In particular, the chirality or helicity of a carbon nanotube can be changed by creating topological defects into the hexagonal network. The defects, however, must induce zero net curvature to prevent the tube from flaring or closing. The smallest topological defect with minimal local curvature (hence less energy cost) and zero net curvature is a pentagon-heptagon pair [16, 17, 18, 67, 68, 69, 70, 71]. Such a pentagon-heptagon defect pair with its symmetry axis nonparallel to the tube axis changes the chirality of a (n, m) tube by transferring one unit from n to m or vice versa. If the pentagon-heptagon defect pair is along the (n, m) tube axis, then one unit is added or subtracted from m. Figure 8 depicts a (8,0) carbon tube joined to a (7,1) tube via a 5-7 defect pair. This system forms a quasi-1D semiconductor/metal junction, since within the band-folding picture the (7,1) half tube is metallic and the (8,0) half tube is semiconducting.

Figures 9 and 10 show the calculated local density of states (LDOS) near the (8,0)/(7,1) junction. These results are from a tight-binding calculation for the π electrons [16]. In both figures, the bottom panel depicts the density of states of the perfect tube, with the sharp features corresponding to the van Hove singularities of a quasi-1D system. The other panels show the calculated LDOS at different distances away from the interface, with cell 1 being the closest to the interface in the semiconductor or side and ring 1 the closest to the interface in the metal side. Here, cell refers to one unit cell of the tube and "ring" refers

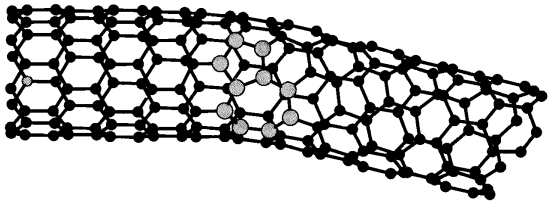


Figure E.8: Atomic structure of an (8,0)/(7,1) carbon nanotube junction. The large light-gray balls denote the atoms forming the heptagon-pentagon pair [16]

to a ring of atoms around the circumference. These results illustrate the spatial behavior of the density of states as it transforms from that of a metal to that of a semiconductor across the junction. The LDOS very quickly changes from that of the metal to that of the semiconductor within a few rings of atoms as one goes from the metal side to the semiconductor side. As the interface is approached, the sharp van Hove singularities of the metal are diluted. Immediately on the semiconductor side of the interface, a different set of singular features, corresponding to those of the semiconductor tube, emerges. There is, however, still a finite density of states in the otherwise bandgap region on the semiconductor side. These are metal induced gap states [72], which decay to zero in about a few Å into the semiconductor. Thus, the electronic structure of this junction is very similar to that of a bulk metal-semiconductor junction, such as Al/Si, except it has a nanometer cross-section and is made out of entirely the element carbon.

Similarly, semiconductor-semiconductor and metal-metal junctions may be constructed with the proper choices of tube diameters and pentagon-heptagon defect pairs. For example, by inserting a 5-7 pair defect, a (10,0) carbon nanotube can be matched to a (9,1) carbon nanotube [16]. Both of these tubes are semiconductors, but they have different bandgaps. The (10,0)/(9,1) junction thus has the electronic structure of a semiconductor heterojunction. In this case, owing to the rather large structural distortion at the interface, there are interesting localized interface states at the junction. Theoretical studies have also been carried out for junctions of B-C-N nanotubes [73], showing very similar behaviors as the carbon case, and for other geometric arrangements, such as carbon nanotube T-junctions, where one tube joins to the side of another tube perpendicularly to form a "T" structure [74].

Calculations have been carried out to study the quantum conductance of the carbon nanotube junctions. Typically these calculations are done within the Landauer formalism [75, 76]. In this approach, the conductance is given in terms of the transmission matrix of the propagating electron waves at a given energy. In particular, the conductance of metal-metal nanotube junctions is shown to exhibit a quite interesting new effect which does not have an analog in bulk metal junctions [67]. It is found that certain configurations of pentagon-heptagon pair defects in forming the junction completely stop the flow of electrons, while other arrangements permit the transmission of current through the junction. Such metal-metal junctions thus have the potential for use as nanoscale electrical switches. This phenomenon is seen in the calculated conductance of a (12,0)/(6,6) carbon nanotube junction in Fig. 11. Both the (12,0) and (6,6) tubes are metallic within the tight-binding model, and they can be matched perfectly to form a straight junction. However, the con-

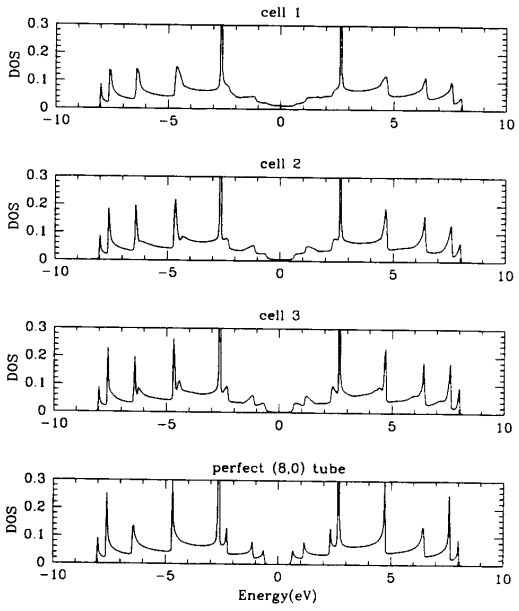


Figure E.9: Calculated LDOS of the (8,0)/(7,1) metal-semiconductor junction at the semiconductor side. From top to bottom, LDOS at cells 1, 2, and 3 of the (8,0) side. Cell 1 is at the interface [16]

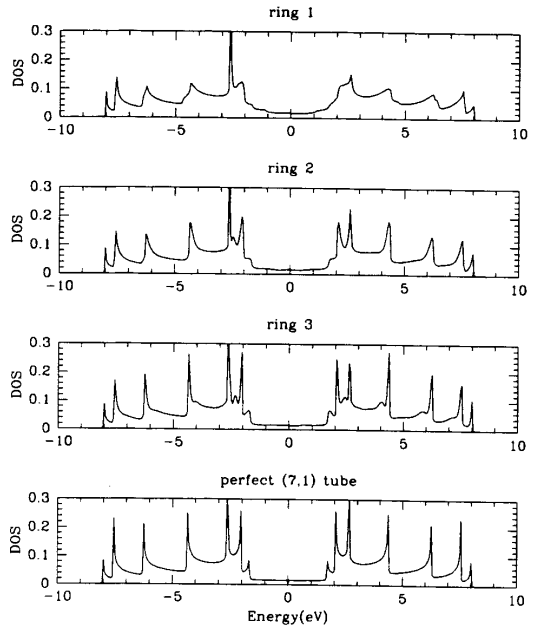


Figure E.10: Calculated LDOS of the (8,0)/(7,1) metal-semiconductor junction at the metal side. From top to bottom, the LDOS at rings 1, 2, and 3 of the (7,1) side. Ring 1 is at the interface [16]

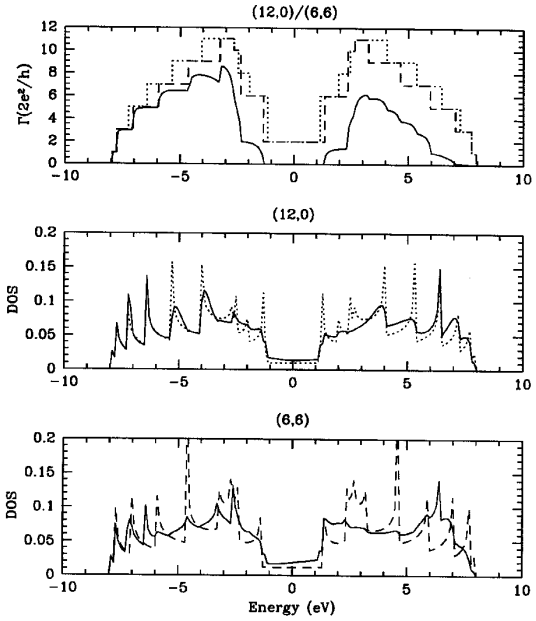


Figure E.11: Calculated results for the (12,0)/(6,6) metal-metal junction. Top: conductance of a matched tube (solid line), a perfect (12,0) tube (dashed line), and a perfect (6,6) tube (dotted line). Center: LDOS at the interface on the (12,0) side (full line) and of the perfect (12,0) tube (dotted line). Bottom: LDOS at the interface on the (6,6) side (full line) and of the perfect (6,6) tube (dashed line) [67]

ductance is zero for electrons at the Fermi level, E_F. This peculiar effect is not due to a lack of density of states at E_F. As shown in Fig. 11, there is finite density of states at E_F everywhere along the whole length of the total system for this junction. The absence of conductance arises from the fact that there is discrete rotational symmetry along the axis of the combined tube. But, for electrons near E_F, the states in one of the half tubes are of a different rotational symmetry from those in the other half tube. As an electron propagates from one side to the other, the electron encounters a symmetry gap and is completely reflected at the junction.

The same phenomenon occurs in the calculated conductance of a (9,0)/(6,3) metal-metal carbon nanotube junction. However, in forming this junction, there are two distinct ways to match the two halves, either symmetrically or asymmetrically. In the symmetric matched geometry, the conductance is zero at E_F for the same symmetry reason as discussed above. (See Fig. 12.) But, in the asymmetric matched geometry, the discrete rotational symmetry of the total system is broken and the electrons no longer have to preserve their rotational quantum number as they travel across the junction. The conductance for this case is now nonzero. Consequently, in some situations, bent junctions can conduct better than straight junctions for the nanotubes. This leads to the possibility of using these metal-metal or other similar junctions as nanoswitches or strain gauges, i.e., one can imagine using some symmetry breaking mechanisms such as electron-photon, electron-phonon or mechanical deformation to switch a junction from a non-conducting state to a conducting state [67].

Junctions of the kind discussed above may be formed during growth, but they can also

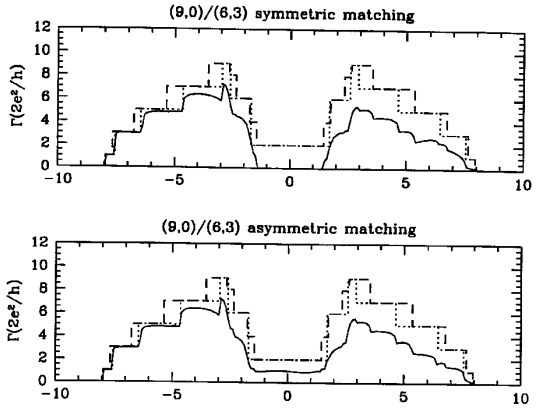


Figure E.12: Calculated conductance of the (9,0)/(6,3) junction – matched system (solid line), perfect (6,3) tube (dotted line), and perfect (9,0) tube (dashed line) [67]

be generated by mechanical stress [77]. There is now considerable experimental evidence of this kind of on-tube junction and device behavior predicted by theory. An experimental signature of a single pentagon-heptagon pair defect would be an abrupt bend between two straight sections of a nanotube. Calculations indicate that a single pentagon-heptagon pair would induce bend angles of roughly 0-15 degrees, with the exact value depending on the particular tubes involved. Several experiments have reported sightings of localized bends of this magnitude for multiwalled carbon nanotubes [23, 78, 79]. Having several 5-7 defect pairs at a junction would allow the joining of tubes of different diameters and add complexity to the geometry. The first observation of nonlinear junction-like transport behavior was made on a rope of SWNTs [22], where the current-voltage properties were measured along a rope of single-walled carbon nanotubes using a scanning tunneling microscopy tip and the behavior shown in Fig. 13 was found in some samples. At one end of the tube, the system behaves like a semimetal showing a typical I-V curve of metallic tunneling, but after some distance at the other end it becomes a rectifier, presumably because a defect of the above type has been introduced at some point on the tube. A more direct measurement was carried out recently [23]. A kinked single-walled nanotube lying on several electrodes was identified and its electrical properties in the different segments were measured. The kink was indicative of two half tubes of different chiralities joined by a pentagon-heptagon defect pair. Figure 14 shows the measured I-V characteristics of a kinked nanotube. The inset is the I-V curve for the upper segment showing that this part of the tube is a metal; but the I-V curve across the kink shows a rectifying behavior indicative of a metal-semiconductor junction.

E.4.2 Impurities, Stone–Wales defects, and structural deformations in metallic nanotubes

An unanswered question in the field has been why do the metallic carbon nanotubes have such long mean free paths. This has led to consideration of the effects of impurities and defects on the conductance of the metallic nanotubes. We focus here on the (10,10) tubes; however, the basic physics is the same for all (n, n) tubes. In addition to tight-binding

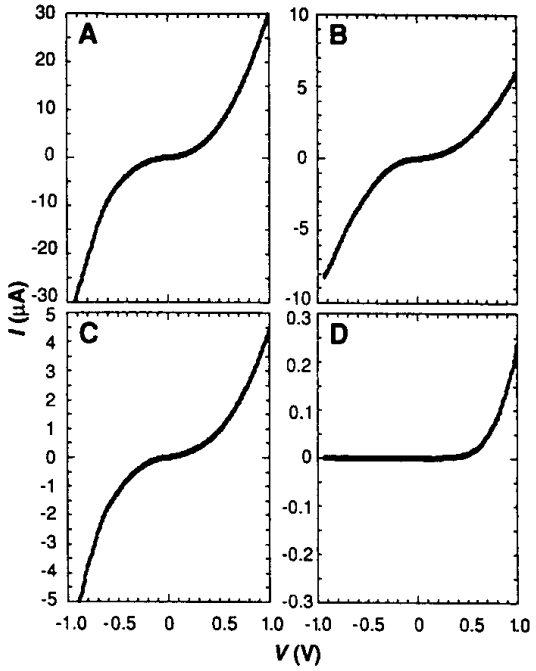


Figure E.13: Current-voltage characteristic measured along a rope of single-walled carbon nanotubes. Panels A, B, C, and D correspond to successive different locations on the rope [22]

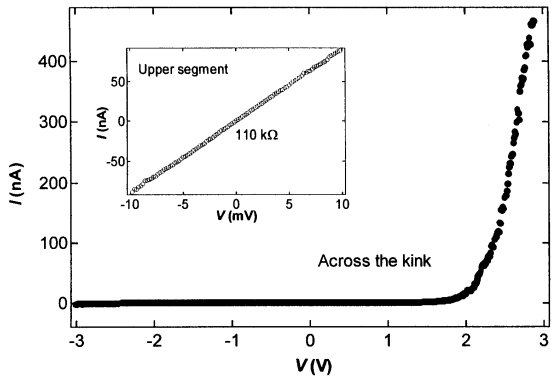


Figure E.14: Measured current-voltage characteristic of a kinked single-walled carbon nanotube [23]

studies, there are now first-principles calculations on the quantum conductance of nanotube structures based on an *ab initio* pseudopotential density functional method with a wavefunction matching technique [80, 81]. The advantages of the *ab initio* approach are that one can obtain the self-consistent electronic and geometric structure in the presence of the defects and, in addition to the conductance, obtain detailed information on the electronic wavefunction and current density distribution near the defect.

Several rather surprising results have been found concerning the effects of local defects on the quantum conductance of the (n, n) metallic carbon nanotubes [81]. For example, the maximum reduction in the conductance due to a local defect is itself often quantized, and this can be explained in terms of resonant backscattering by quasi-bound states of the defect. Here we discuss results for three simple defects: boron and nitrogen substitutional impurities and the bond rotation or Stone–Wales defect. A Stone–Wales defect corresponds to the rotation of one of the bonds in the hexagonal network by 90 degrees, resulting in the creation of a quite low energy double 5-7 defect pair, without changing the overall helicity of the tube.

Figure 15 depicts several results for a (10,10) carbon nanotube with a single boron substitutional impurity. The top panel is the calculated conductance as a function of the energy of the electron. For a perfect tube, the conductance (indicated here by the dashed line) is 2 in units of the quantum of conductance, $2e^2/h$, since there are two conductance channels available for the electrons near the Fermi energy. For the result with the boron impurity, a striking feature is that the conductance is virtually unchanged at the Fermi level of the neutral nanotube. That is, the impurity potential does not scatter incoming electrons of this energy. On the other hand, there are two dips in the conductance below E_F. The amount of the reduction at the upper dip is one quantum unit of conductance and its shape is approximately Lorentzian. In fact, the overall structure of the conductance is well described by the superposition of two Lorentzian dips, each with a depth of 1 conductance quantum. These two dips can be understood in terms of a reduction in conductance due to resonant backscattering from quasi-bound impurity states derived from the boron impurity.

The calculated results thus show that boron behaves like an acceptor with respect to the first lower subband (i.e., the first subband with energy below the conduction states) and forms two impurity levels that are split off from the top of the first lower subband. These impurity states become resonance states or quasi-bound states due to interaction with the conduction states. The impurity states can be clearly seen in the calculated LDOS near the boron impurity (middle panel of Fig. 15). The two extra peaks correspond to the two quasi-bound states. The LDOS would be a constant for a perfect tube in the region between the van Hove singularity of the first lower subband and that of the first upper subband. Because a (n, n) tube with a substitutional impurity still has a mirror plane perpendicular to the tube axis, the defect states have definite parity with respect to this plane. The upper energy state (broader peak) in Fig. 15 has even parity and the lower energy state (narrower peak) has odd parity, corresponding to s-like and p-like impurity states, respectively.

The conductance behavior in Fig. 15 may be understood by examining how electrons in the two eigen-channels interact with the impurity. At the upper dip, an electron in one of the two eigen-channels is reflected completely (99.9%) by the boron impurity, but an electron in the other channel passes by the impurity with negligible reflection (0.1%). The same happens at the lower dip but with the behavior of the two eigen-channels switched. The bottom panel shows the calculated scattering phase shifts. The phase shift of the odd

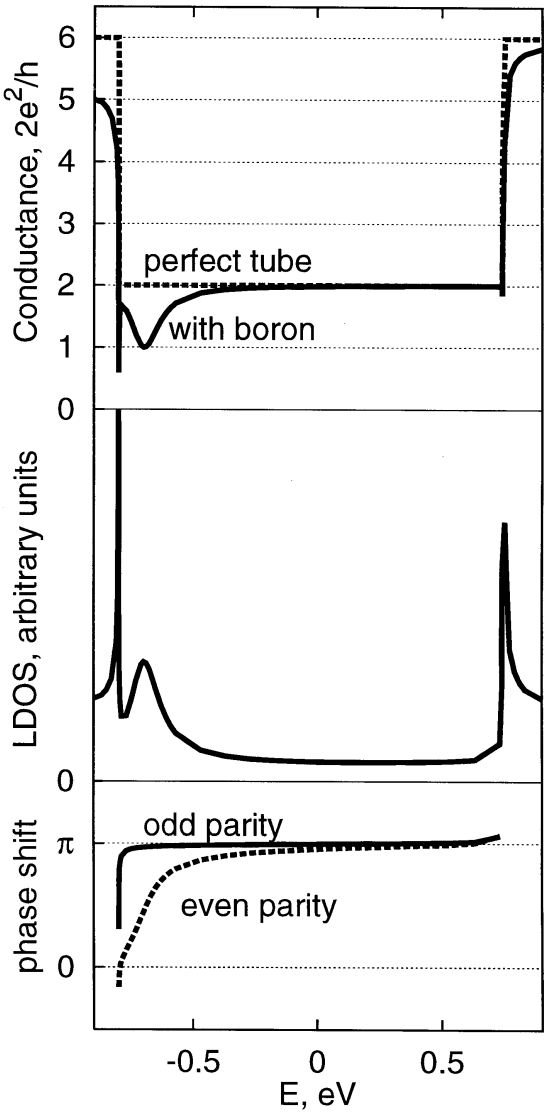


Figure E.15: Energy dependence of the calculated conductance, local density of states, and phase shifts of a (10,10) carbon nanotube with a substitutional boron impurity [81]

parity state changes rapidly as the energy sweeps past the lower quasi-bound state level, with its value passing through $\pi/2$ at the peak position of the quasi-bound state. The same change occurs to the phase shift of the even parity state at the upper impurity-state energy. The total phase shift across a quasi-bound level is π in each case, in agreement with the Friedel sum rule. The picture is that an incoming electron with energy exactly in resonance with the impurity state is being scattered back totally in one of the channels but not the other. This explains the exact reduction of one quantum of conductance at the dip. The upper-energy impurity state has a large binding energy (over 0.1 eV) with respect to the first lower subband and hence is quite localized. It has an approximate extent of $\sim$10 Å, whereas the lower impurity state has an extent of $\sim$250 Å.

The results for a nitrogen substitutional impurity on the (10,10) tube are presented in Fig. 16. Nitrogen has similar effects on the conductance as boron, but with opposite energy structures. Again, the conductance at the Fermi level is virtually unaffected, but there are two conductance dips above the Fermi level just below the first upper subband. Thus, the nitrogen impurity behaves like a donor with respect to the first upper subband, forming an s-like quasi-bound state with stronger binding energy and a p-like state with weaker binding energy. As in the case of boron, the reduction of one quantum unit of conductance at the dips is caused by the fact that, at resonance, the electron in one of the eigen channels is reflected almost completely by the nitrogen impurity but the electron in the other channel passes by the impurity with negligible reflection. The LDOS near the nitrogen impurity shows two peaks corresponding to the two quasi-bound states. The phase shifts of the two eigen channels show similar behavior as in the boron case.

For a (10,10) tube with a Stone–Wales or double 5-7 pair defect, the calculations also find that the conductance is virtually unchanged for the states at the Fermi energy. Thus these results show that the transport properties of the neutral (n, n) metallic carbon nanotube are very robust with respect to these kinds of intra-tube local defects. As in the impurity case, there are two dips in the quantum conductance in the conduction band energy range, one above and one below the Fermi level. These are again due to the existence of defect levels, and the reduction at the two dips is very close to one quantum of conductance for the same reason, as discussed above. The symmetry of the Stone–Wales defect in this case does not cause mixing between the π and π^* bands, and these two bands remain as eigen channels in the defective system. The lower dip is due to a complete reflection of the π^* band and the upper dip is due to complete reflection of the π band. This implies that the conductance of the nanotube, when there are more than one double 5-7 pair defect, would not sensitively depend on their relative positions, but only on their total numbers, as long as the distance between defects is far enough to be able to neglect inter-defect interactions. The analysis of the phase shifts show that the lower quasi-bound state is even with respect to a mirror plane perpendicular to the tube axis, while the upper quasi-bound state is odd with respect to the same plane.

The conductance of nanotubes can also be affected by structural deformations. Two types of deformations involving bending or twisting the nanotube structure have been considered in the literature. It was found that a smooth bending of the nanotube does not lead to scattering [54], but formation of a local kink induces strong σ-π mixing and backscattering similar to that discussed earlier for boron impurity [82]. Twisting has a much stronger effect [54]. A metallic armchair (n, n) nanotube upon twisting develops a band-gap which scales linearly with the twisting angle up to the critical angle at which the tube collapses

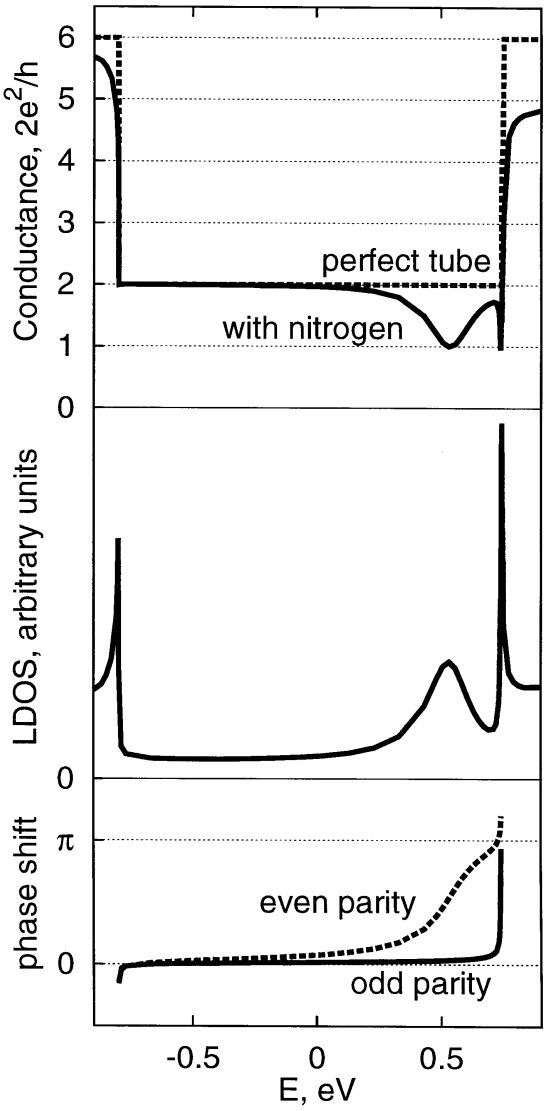


Figure E.16: Calculated conductance, local density of states and phase shifts of a (10,10) carbon nanotube with a substitutional nitrogen impurity [81]

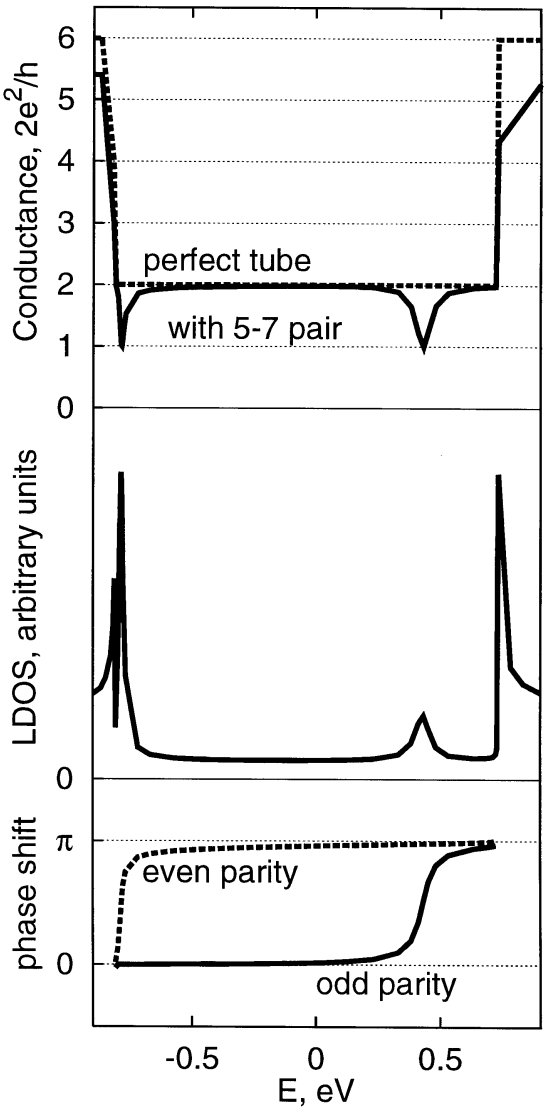


Figure E.17: Energy dependence of the calculated conductance, local density of states, and phase shifts of a (10,10) carbon nanotube with a Stone–Wales defect [81]

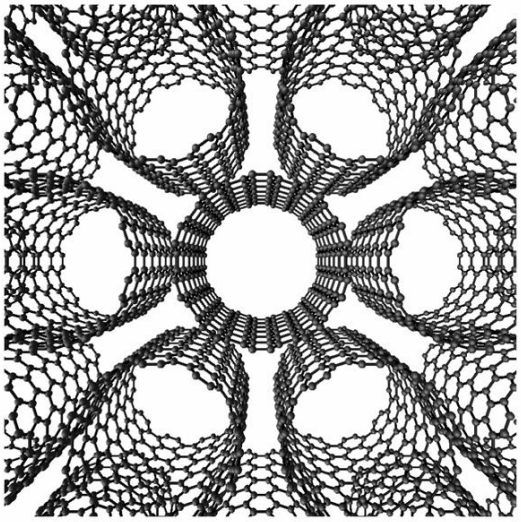


Figure E.18: Perspective view of a model of a rope of (10,10) carbon nanotubes

into a ribbon [82].

E.5 Nanotube Ropes, Crossed-Tube Junctions, and Effects of Long-Range Perturbations

E.5.1 Ropes of nanotubes

Another interesting carbon nanotube system is that of ropes of single-walled carbon nanotubes which have been synthesized in high yield [8]. These ropes, containing up to tens to hundreds of single-walled nanotubes in a close-packed triangular lattice, are made up of tubes of nearly uniform diameter, close to that of the (10,10) tubes (see Fig. 18.) Because of the rather weak interaction between these tubes, a naive picture would be that the packing of individual metallic nanotubes into ropes would not change their electronic properties significantly. Theoretical studies [83, 84, 85] however showed that this is not the case for a rope of (10,10) carbon nanotubes. A broken symmetry of the (10,10) nanotube caused by interactions between tubes in a rope induces formation of a pseudogap in the density of states of about 0.1 eV. The existence of this pseudogap alters many of the fundamental electronic properties of the rope.

As discussed above, an isolated (n, n) carbon nanotube has two linearly dispersing conduction bands which cross at the Fermi level forming two "Dirac" points, as schematically presented in Fig. 19(a). This linear band dispersion in a one-dimensional system gives rise to a finite and constant density of electronic states at the Fermi energy. Thus, an (n, n) tube is a metal within the one-electron picture. The question of interest is: How does the electronic structure change when the metallic tubes are bundled up to form a closely packed two-dimensional crystal, as in the case of the (10,10) ropes. In the calculation, a large (10,10) rope is modeled by a triangular lattice of (10,10) tubes infinitely extended in

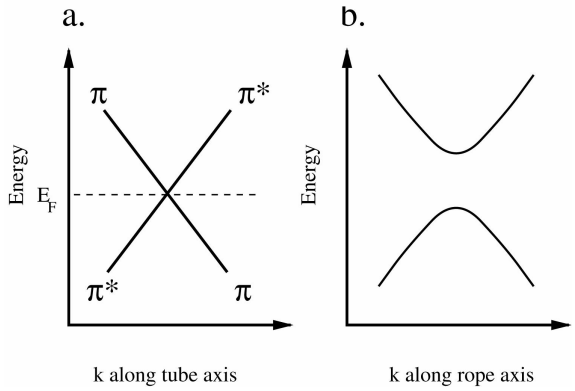


Figure E.19: Band crossing and band repulsion. (a) Schematic band structure of an isolated (n, n) carbon nanotube near the Fermi energy. (b) Repulsion of bands due to the breaking of mirror symmetry.

the lateral directions. For such a system, the electronic states, instead of being contained in the 1-D Brillouin zone of a single tube, are now extended to a three-dimensional irreducible Brillouin zone wedge. If tube-tube interactions are negligibly small, the electronic energy band structure along any line in the wedge parallel to the rope axis would be exactly the same as the band dispersion of an isolated tube. In particular, at the k-wavevector corresponding to the band crossing point, there will be a two-fold degenerate state at the Fermi energy. This allowed band crossing is due to the mirror symmetry of the (10,10) tube. For a tube in a rope, this symmetry is however broken because of intertube interactions. The broken symmetry causes a quantum level repulsion and opens up a gap almost everywhere in the Brillouin zone, as schematically shown in Fig. 19(b).

The band repulsion resulting from the broken-symmetry strongly modifies the density of states (DOS) of the rope near the Fermi energy compared to that of an isolated (10,10) tube. The calculated DOS is presented in Fig. 20(a). Shown are the results for two cases: aligned and misaligned tubes in the rope. In both cases, there is a pseudogap of the order of 0.1 eV in the density of states. Examination of the electronic structure reveals that the system is a semimetal with both electron and hole carriers. The existence of the pseudogap in the rope makes the conductivity and other transport properties of the metallic rope significantly different from those of isolated tubes, even without considering the effect of local disorder in low dimensions. Since the DOS increases rapidly away from the Fermi level, the carrier density of the rope is sensitive to temperature and doping. The existence of both electron and hole carriers leads to qualitatively different thermopower and Hall-effect behaviors from those expected for a normal metal. The optical properties of the rope are also affected by the pseudogap. As illustrated by the calculated joint density of states (JDOS) in Fig. 20(b), there would be a finite onset in the infrared absorption spectrum for a large perfectly ordered (10,10) rope, where one can assume k-conserving optical transitions. In the case of high disorder, an infrared experiment would more closely reflect the DOS rather than the JDOS. For most actual samples, the fraction of (10,10) carbon nanotubes (compared with other nanotubes of the same diameter) in the experimentally synthesized ropes appears to be small. However, the conclusion that broken symmetry induces a gap in the (n, n) tubes is a general result which is of relevance for tubes under any significant

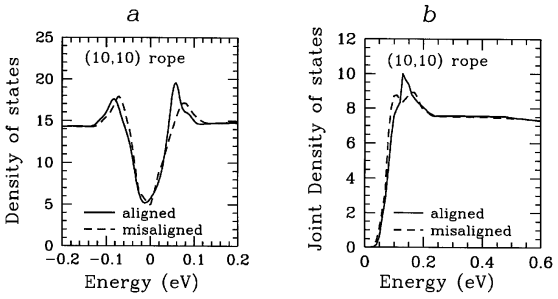


Figure E.20: (a) Calculated density of states for a rope of misaligned (10,10) carbon nanotubes (broken line) and aligned tubes (solid line). The Fermi energy is at zero. (b) Calculated joint density of states for a rope of misaligned (broken line) and aligned (solid line) (10,10) tubes. Results are in units of states per meV per atom [83, 84]

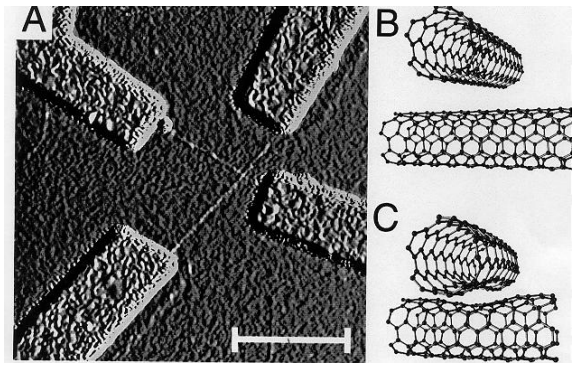


Figure E.21: AFM image of a crossed SWNT device (A). Calculated structure of a crossed (5,5) SWNT junction with a force of 0 nN (B) and 15 nN (C) [86]

asymmetric perturbations, such as those due to structural deformations or external fields.

E.5.2 Crossed-tube junctions

The discussion of nanotube junctions in Sec. 3 is focused on the on-tube junctions, i.e., forming a junction by joining two half tubes together. These systems are extremely interesting, but difficult to synthesize in a controlled manner at this time. Another way to form junctions is to have two tubes crossing each other in contact [86]. (See Fig. 21.) This kind of crossed-tube junction is much easier to fabricate and control with present experimental techniques. When two nanotubes cross in free space, one expects that the tubes at their closest contact point will be at a van der Waals distance away from each other and that there will not be much intertube or junction conductance. However, as shown by Avouris and coworkers [87], for two crossed tubes lying on a substrate, there is a substantial force pressing one tube against the other due to the substrate attraction. For a crossed-tube junction composed of SWNTs with the experimental diameter of 1.4 nm, this contact force has been estimated to be about 5 nN [87]. This substrate force would then be sufficient to deform the crossed-tube junction and lead to better junction conductance.

In Fig. 21, panel A is an AFM image of a crossed-tube junction fabricated from two

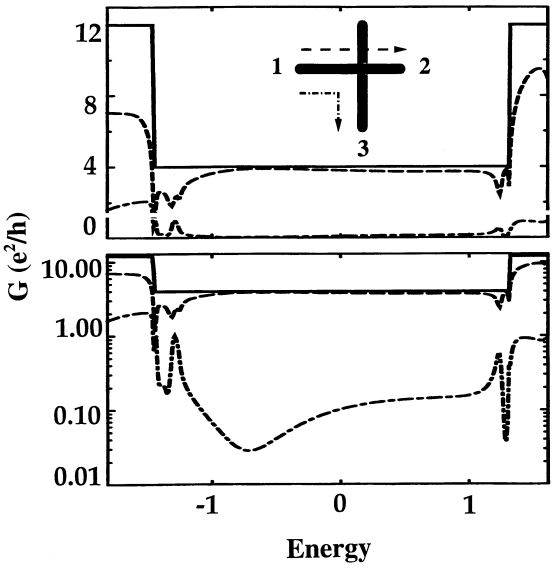


Figure E.22: Calculated conductance (expressed in units of e^2/h) of a crossed (5,5) carbon nanotube junction with a contact force of 15 nN on a linear (top) and log (bottom) scale. The dashed (dotted-dashed) curve corresponds to the intra-tube (intertube) conductance [86]

single-walled carbon nanotubes of 1.4 nm in diameter with electrical contacts at each end [86]. Panels B and C show the calculated structure corresponding to a (5,5) carbon nanotube pressed against another one with zero and 15 nN force, respectively. Because of the smaller diameter of the (5,5) tube, a larger contact force is required to produce a deformation similar to that of the experimental crossed-tube junction. The calculation was done using the *ab initio* pseudopotential density functional method with a localized basis [86]. As seen in panel C, there is considerable deformation, and the atoms on the different tubes are much closer to each other. At this distance, the closest atomic separation between the two tubes is 0.25 nm, significantly smaller than the van der Waals distance of 0.34 nm.

For the case of zero contact force (panel B in Fig. 21), the calculated intra-tube conductance is virtually unchanged from that of an ideal, isolated metallic tube, and the intertube conductance is negligibly small. However, when the tubes are under a force of 15 nN, there is a sizable intertube or junction conductance. As shown in Fig. 22, the junction conductance at the Fermi energy is about 5% of a quantum unit of conductance $G_0 = 2e^2/h$. The junction conductance is thus very sensitive to the force or distance between the tubes.

Experimentally, the conductance of various types of crossed carbon nanotube junctions has been measured, including metal-metal, semiconductor-semiconductor, and metal-semiconductor crossed-tube junctions. The experimental results are presented in Fig. 23. For the metal-metal crossed-tube junctions, a conductance of 2 to 6% of G_0 is found, in good agreement with the theoretical results. Of particular interest is the metal-semiconductor case in which experiments demonstrated Schottky diode behavior with a Schottky barrier in the range of 200–300 meV, which is very close to the value of 250 meV expected from theory for nanotubes with diameters of 1.4 nm [86].

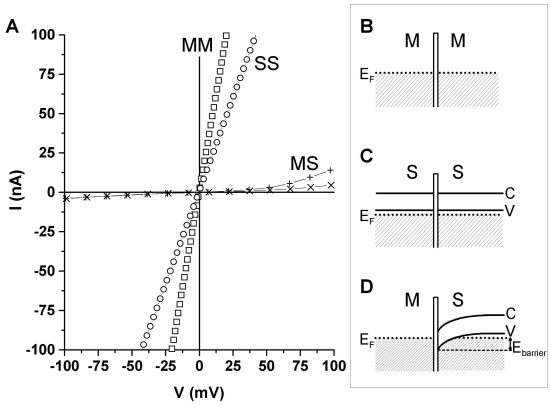


Figure E.23: Current-voltage characteristics of several crossed SWNT junctions [86] (see text)

E.5.3 Effects of Long-Range Disorder and External Perturbations

The effects of disorder on the conducting properties of metal and semiconducting carbon nanotubes are quite different. Experimentally, the mean free path is found to be much longer in metallic tubes than in doped semiconducting tubes [19, 20, 21, 24, 88]. This result can be understood theoretically if the disorder potential is long range. As discussed below, the internal structure of the wavefunction of the states connected to the sublattice structure of graphite lead to a suppression of scattering in metallic tubes, but not in semiconducting tubes. Figure 24 shows the measured conductance for a semiconducting nanotube device as a function of gate voltage at different temperatures. [88]. The diameter of the tube as measured by AFM is 1.5 nm, consistent with a single-walled tube. The complex structure in the Coulomb blockade oscillations in Fig. 24 is consistent with transport through a number of quantum dots in series. The temperature dependence and typical charging energy indicates that the tube is broken up into segments of length of about 100 nm. Similar measurements on intrinsic metal tubes, on the other hand, yield lengths that are typically a couple of orders of magnitude longer [19, 20, 21, 24, 88].

Theoretical calculations have been carried out to examine the effects of long-range external perturbations [88]. In the calculation, to model the perturbation, a 3-dimensional Gaussian potential of a certain width is centered on one of the atoms on the carbon nanotube wall. The conductance with the perturbation is computed for different Gaussian widths, but keeping the integrated strength of the potential the same. Some typical tight-binding results are presented in Fig. 25. The solid lines show the results for the conductance of a disorder-free tube, while the dashed and the dot-dashed lines are, respectively, for a single long-range (σ=0.348 nm, ΔV= 0.5 eV) and a short-range (σ=0.116 nm, ΔV= 10 eV) scatterer. Here ΔV is the shift in the on-site energy at the potential center. The conduction bands (i.e., bands crossing the Fermi level) of the *metallic* tube are unaffected by the long-range scatterer, unlike the lower and upper subbands of both the *metallic* and *semiconducting* tubes, which are affected by boh long- and short-range scatterers. All subbands are influenced by the short-range scatterer. The inset shows an expanded view of the onset of conduction in the semiconducting tube at positive E, with each division corresponding

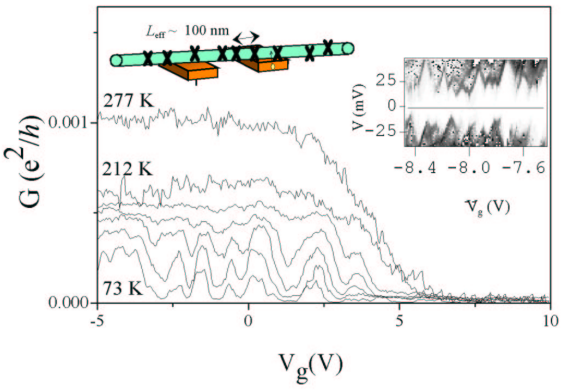


Figure E.24: Conductance vs. gate voltage V_g for a semiconducting single-walled carbon nanotube at various temperatures. The upper insert schematically illustraates the sample geometry and the lower insert shows dI/dV vs. V and V_g plotted as a gray scale [88]

to 1 meV. Also, the sharp step edges in the calculated conductance of the perfect tubes are rounded off by both types of perturbations.

Both the experimental and theoretical findings strongly suggest that long-range scattering is suppressed in the metallic tubes. One can actually understand this qualitatively from the electronic structure of a graphene sheet [89, 90]. The graphene structure has two atoms per unit cell. The properties of electrons near the Fermi energy are given by those states near the corner of the Brillouin zone. (See Fig. 26.) If we look at the states near this point and consider them in terms of a k-vector away from the corner K point, then they can be described by a Dirac Hamiltonian. For these states, the wavefunctions can be written in terms of a product of a plane wave component (with a vector k) and a pseudo-spin which describes the bonding character between the two atoms in the unit cell. The interesting result is that this pseudo-spin points along k. For example, if the state at k is bonding, then the state at $-k$ is antibonding in character. Within this framework, one can work out the scattering between the allowed states in a carbon nanotube due to long-range disorder, i.e., disorder with Fourier components $V(q)$ such that $q \ll K$. This, for example, will be the case for scattering by charged trap states in the substrate (oxide traps). In this case, the disorder does not couple to the pseudo-spin portion of the wavefunction, since the disorder potential is approximately constant on the scale of the interatomic distance. The resulting matrix element between states is then [89, 90]: $|\langle k'|V(r)|k\rangle|^2 = |V(k-k')|^2 \cos^2[(1/2)\theta_{k,k'}]$, where $\theta_{k,k'}$ is the angle between the initial and final states. The first term in $V(k-k')$ is the Fourier component at the difference in k values of the initial and final envelope wavefunctions. The cosine term is the overlap of the initial and final spinor states.

For a metallic tube [Fig. 26(b)], backscattering in the conduction band corresponds to scattering between k and $-k$. Such scattering is forbidden, because the molecular orbitals of these two states are orthogonal. In semiconducting tubes, however, the situation is quite different [Fig. 26(c)]. The angle between the initial and final states is less than π, and scattering is thus only partially suppressed by the spinor overlap. As a result, semiconducting tubes should be sensitive to long-range disorder, while metallic tubes should not. However, short-range disorder which has Fourier components $q \sim K$ will couple the molecular orbitals together and lead to scattering in all of the subbands. These theoretical considerations agree

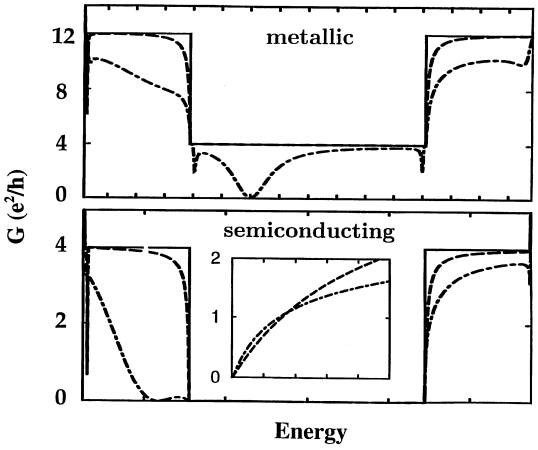


Figure E.25: Tight-binding calculation of the conductance of a (a) metallic (10,10) tube and (b) semiconducting (17,0) tube in the presence of a Gaussian scatterer. The energy scale on the abscissa is 0.2 eV per division in each graph [88]

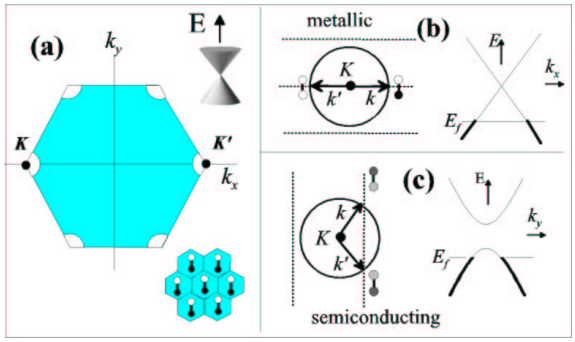


Figure E.26: (a) Filled states (shaded) in the first Brillouin zone of a p-type graphene sheet. There are two carbon atoms per unit cell (lower right inset). The dispersions of the states near E_F are cones whose vertices are located at the corner points of the Brillouin zone. The Fermi circle, defining the allowed k vectors, and the band dispersions are shown in (b) and (c) for a metallic and a semiconducting tube, respectively [88]

well with experiment and with the detailed calculations discussed above. Long-range disorder due to, e.g., localized charges near the tube, breaks the semiconducting tube into a series of quantum dots with large barriers, resulting in a dramatically reduced conductance and a short mean free path. On the other hand, metallic tubes are insensitive to this disorder and remain near-perfect 1D conductors.

E.6 Summary

This Chapter gives a short review of some of our theoretical understanding of the structural and electronic properties of single-walled carbon nanotubes and of various structures formed from these nanotubes. Because of their nanometer dimensions, the nanotube structures can have novel properties and yield unusual scientific phenomena. In addition to the multi-walled carbon nanotubes, single-walled nanotubes, nanotube ropes, nanotube junctions, and non-carbon nanotubes have been synthesized.

These quasi-one-dimensional objects have highly unusual electronic properties. For the perfect tubes, theoretical studies have shown that the electronic properties of the carbon nanotubes are intimately connected to their structure. They can be metallic or semiconducting, depending sensitively on tube diameter and chirality. Experimental studies using transport, scanning tunneling, and other techniques have basically confirmed the theoretical predictions. The dielectric responses of the carbon nanotubes are found to be highly anisotropic in general. The heat capacity of single-wall nanotubes is predicted to have a characteristic linear T dependence at low temperature.

On-tube metal-semiconductor, semiconductor-semiconductor, and metal-metal junctions may be formed by introducing topological structural defects, and these junctions have been shown to behave like nanoscale device elements. For example, different half-tubes may be joined with 5-member ring/7-member ring pair defects to form a metal-semiconductor Schottky barrier. The calculated electronic structure of these junctions is very similar to that of standard metal-semiconductor interfaces, and in this sense, they are molecular level devices composed of the single element, carbon. Recent experimental measurements have confirmed the existence of such Schottky barrier behavior in nanotube ropes and across kinked nanotube junctions. Similarly, 5-7 defect pairs in different carbon and non-carbon nanotubes can produce semiconductor-semiconductor and metal-metal junctions. The existence of metal-metal nanotube junctions in which the conductance is suppressed for symmetry reasons has also been predicted. Thus, the carbon nanotube junctions may be used as nanoscale electronic elements.

The influence of impurities and local structural defects on the conductance of carbon nanotubes has also been examined. It is found that local defects in general form well defined quasi-bound states even in metallic nanotubes. These defect states give rise to peaks in the LDOS and reduce the conductance at the energy of the defect levels by a quantum unit of conductance via resonant backscattering. The theoretical studies show that, owing to the unique electronic structure of the graphene sheet, the transport properties of (n, n) metallic tubes appear to be very robust against defects and long-range perturbations near E_F. Doped semiconducting tubes are much more susceptible to long-range disorder. These results explain the experimental findings of the long coherence length in metallic tubes and the large difference in mean free path between the metallic and doped semiconducting tubes. For nanotube ropes, intertube interactions are shown to alter the

electronic structure of (n, n) metallic tubes because of broken symmetry effects, leading to a pseudogap in the density of states and to semimetallic behavior. Crossed-tube junctions have also been fabricated experimentally and studied theoretically. These systems show significant intertube conductance for metal-metal junctions and exhibit Schottky behavior for metal-semiconductor junctions when the tubes are subjected to contact force from the substrate.

The carbon nanotubes are hence a fascinating new class of materials with many unique and desirable properties. The rich interplay between the geometric and electronic structure of the nanotubes has given rise to many interesting, new physical phenomena. At the practical level, these systems have the potential for many possible applications.

Bibliography

[1] S. Iijima, Nature (London) **354**, 56 (1991).

[2] M. S. Dresselhaus, G. Dresselhaus, and P. C. Eklund, *Science of Fullerenes and Carbon Nanotubes* (Academic Press, New York, NY, 1996).

[3] P. M. Ajayan and T. W. Ebbesen, Rep. Prog. Phys. **60**, 1025 (1997).

[4] C. Dekker, Phys. Today **52**, 22 (1999).

[5] S. Iijima and T. Ichihashi, Nature (London) **363**, 603 (1993).

[6] D. S. Bethune, C. H. Kiang, M. S. de Vries, G. Gorman, R. Savoy, J. Vazquez, and R. Beyers, Nature (London) **363**, 605 (1993).

[7] P. M. Ajayan, J. M. Lambert, P. Bernier, L. Barbedette, C. Colliex, and J. M. Planeix, Chem. Phys. Lett. **215**, 509 (1993).

[8] T. Guo, C.-M. Jin, and R. E. Smalley, Chem. Phys. Lett. **243**, 49–54 (1995).

[9] M. R. Pederson and J. Q. Broughton, Phys. Rev. Lett. **69**, 2689 (1992).

[10] P. M. Ajayan and S. Iijima, Nature (London) **361**, 333 (1993).

[11] N. G. Chopra, L. X. Benedict, V. H. Crespi, M. L. Cohen, S. G. Louie, and A. Zettl, Nature (London) **377**, 135 (1995).

[12] H. Dai, E. W. Wong, and C. M. Lieber, Nature (London) **384**, 147 (1996).

[13] W. A. de Heer, A. Châtelain, and D. Ugarte, Science **270**, 1179 (1995). see also ibid page 1119.

[14] A. G. Rinzler, J. H. Hafner, P. Nikolaev, L. Lou, S. G. Kim, D. Tománek, P. Nordlander, D. T. Colbert, and R. E. Smalley, Science **269**, 1550 (1995).

[15] A. C. Dillon, K. M. Jones, T. A. Bekkedahl, C. H. Kiang, D. S. Bethune, and M. J. Heben, Nature (London) **386**, 377–379 (1997).

[16] L. Chico, V. H. Crespi, L. X. Benedict, S. G. Louie, and M. L. Cohen, Phys. Rev. Lett. **76**, 971–974 (1996).

[17] Ph. Lambin, A. Fonseca, J. P. Vigneron, J. B. Nagy, and A. A. Lucas, Chem. Phys. Lett. **245**, 85–89 (1995).

[18] R. Saito, G. Dresselhaus, and M. S. Dresselhaus, Phys. Rev. B **53**, 2044–2050 (1996).

[19] S. J. Tans, M. H. Devoret, H. Dai, A. Thess, R. E. Smalley, L. J. Geerligs, and C. Dekker, Nature (London) **386**, 474–477 (1997).

[20] M. Bockrath, D. H. Cobden, P. L. McEuen, N. G. Chopra, A. Zettl, A. Thess, and R. E. Smalley, Science **275**, 1922–1924 (1997).

[21] R. Martel, T. Schmidt, H. R. Shea, T. Hertel, and Ph. Avouris, Appl. Phys. Lett. **73**, 2447 (1998).

[22] P. G. Collins, A. Zettl, H. Bando, A. Thess, and R. E. Smalley, Science **278**, 5335 (1997).

[23] Zhen Yao, H. W. C. Postma, L. Balents, and C. Dekker, Nature (London) **402**, 273 (1999).

[24] S. J. Tans, R. M. Verschueren, and C. Dekker, Nature **393**, 49–52 (1998).

[25] X. Blase, A. Rubio, S. G. Louie, and M. L. Cohen, Europhys. Lett. **28**, 335 (1994).

[26] Y. Miyamoto, A. Rubio, M. L. Cohen, and S. G. Louie, Phys. Rev. B **50**, 18360 (1994).

[27] Y. Miyamoto, A. Rubio, S. G. Louie, and M. L. Cohen, Phys. Rev. B **50**, 4976 (1994).

[28] X. Blase, A. Rubio, S. G. Louie, and M. L. Cohen, Phys. Rev. B **51**, 6868 (1995).

[29] Y. Miyamoto, M. L. Cohen, and S. G. Louie, Solid State Comm. **102**, 605 (1997).

[30] Z. Weng-Sieh, K. Cherrey, N. G. Chopra, X. Blase, Y. Miyamoto, A. Rubio, M. L. Cohen, S. G. Louie, A. Zettl, and R. Gronsky, Phys. Rev. B **51**, 11229 (1995).

[31] O. Stephan, P. M. Ajayan, C. Colliex, Ph. Redlich, J. M. Lambert, P. Bernier, and P. Lefin, Science **266**, 1683 (1994).

[32] N. G. Chopra, J. Luyken, K. Cherry, V. H. Crespi, M. L. Cohen, S. G. Louie, and A. Zettl, Science **269**, 966 (1995).

[33] A. Loiseau, F. Willaime, N. Demoncy, G. Hug, and H. Pascard, Phys. Rev. Lett. **76**, 4737 (1996).

[34] K. Suenaga, C. Colliex, N. Demoncy, A. Loiseau, H. Pascard, and F. Willaime, Science **278**, 653 (1997).

[35] P. Gleize, S. Herreyre, P. Gadelle, M. Mermoux, M. C. Cheynet, and L. Abello, J. Materials Science Letters **13**, 1413 (1994).

[36] See also the chapter by R. Tenne and A. Zettl in this Volume.

[37] R. Saito and H. Kataura. In *Carbon Nanotubes*, edited by M. S Dresselhaus and P. Avouris, Springer-Verlag, Berlin, 2000. to be published.

[38] N. Hamada, S. Sawada, and A. Oshiyama, Phys. Rev. Lett. **68**, 1579–1581 (1992).

[39] R. Saito, M. Fujita, G. Dresselhaus, and M. S. Dresselhaus, Appl. Phys. Lett. **60**, 2204–2206 (1992).

[40] J. W. Mintmire, B. I. Dunlap, and C. T. White, Phys. Rev. Lett. **68**, 631–634 (1992).

[41] X. Blase, L. X. Benedict, E. L. Shirley, and S. G. Louie, Phys. Rev. Lett. **72**, 1878 (1994).

[42] A. Rochefort, D. S. Salahub and Ph. Avouris, Chem. Phys. Lett. **297**, 45 (1998).

[43] S. I. Sawada and N. Hamada, Solid State Commun. **83**, 917–919 (1992).

[44] X. Blase and S. G. Louie, unpublished.

[45] L. Langer, V. Bayot, E. Grivei, J. P. Issi, J. P. Heremans, C. H. Olk, L. Stockman, C. Van Haesendonck, and Y. Bruynseraede, Phys. Rev. Lett. **76**, 479–482 (1996).

[46] T. W. Ebbesen, H. J. Lezec, H. Hiura, J. W. Bennett, H. F. Ghaemi, and T. Thio, Nature (London) **382**, 54–56 (1996).

[47] H. Dai, E. W. Wong, and C. M. Lieber, Science **272**, 523–526 (1994).

[48] J. W. G. Wildöer, L. C. Venema, A. G. Rinzler, R. E. Smalley, and C. Dekker, Nature (London) **391**, 59–62 (1998).

[49] T. W. Odom, J. L. Huang, P. Kim, and C. M. Lieber, Nature (London) **391**, 62–64 (1998).

[50] L. X. Benedict, S. G. Louie, and M. L. Cohen, Phys. Rev. B **52**, 8541 (1995).

[51] G. F. Bertsch, A. Bulgac, D. Tománek, and Y. Wang, Phys. Rev. Lett. **67**, 2690 (1991).

[52] B. Koopmans, PhD Thesis, University of Groningen, 1993.

[53] See also the chapter by J. Hone in this Volume.

[54] C. L. Kane and E. J. Mele, Phys. Rev. Lett. **78**, 1932 (1997).

[55] L. X. Benedict, S. G. Louie, and M. L. Cohen, Solid State Commun. **100**, 177–180 (1996).

[56] See also the chapter by B. Yakobson in this Volume.

[57] D. H. Robertson, D. W. Brenner, and J. W. Mintmire, Phys. Rev. B **45**, 12592 (1992).

[58] R. S. Ruoff and D. C. Lorents, Carbon **33**, 925 (1995).

[59] J. M. Molina, S. S. Savinsky, and N. V. Khokhriakov, J. Chem. Phys. **104**, 4652 (1996).

[60] B. I. Yakobson, C. J. Brabec, and J. Bernholc, Phys. Rev. Lett. **76**, 2411 (1996).

[61] C. F. Cornwell and L. T. Wille, Solid State Commun. **101**, 555 (1997).

[62] S. Iijima, C. J. Brabec, A. Maiti, and J. Bernholc, J. Chem. Phys. **104**, 2089 (1996).

[63] J. P. Lu, Phys. Rev. Lett. **79**, 1297 (1997).

[64] M. M. J. Treacy, T. W. Ebbesen, and J. M. Gibson, Nature (London) **381**, 678 (1996).

[65] E. W. Wong, P. E. Sheehan, and C. M. Lieber, Science **277**, 1971 (1997).

[66] N. G. Chopra and A. Zettl, Solid State Comm. **105**, 297 (1998).

[67] L. Chico, L. X. Benedict, S. G. Louie, and M. L. Cohen, Phys. Rev. B **54**, 2600 (1996).

[68] B. I. Dunlap, Phys. Rev. B **49**, 5643 (1994).

[69] J.-C. Charlier, T. W. Ebbesen, and Ph. Lambin, Phys. Rev. B **53**, 11108 (1996).

[70] T. W. Ebbesen and T. Takada, Carbon **33**, 973 (1995).

[71] Ph. Lambin, L. Philippe, J.-C. Charlier, and J. P. Michenaud, Synthetic Metals **2**, 350–356 (1996).

[72] S. G. Louie and M. L. Cohen, Phys. Rev. B **13**, 2461 (1976).

[73] X. Blase, J. C. Charlier, A. de Vila, and R. Car, Appl. Phys. Lett. **70**, 197 (1997).

[74] M. Menon and D. Srivastava, Phys. Rev. Lett. **79**, 4453–4456 (1997).

[75] R. Landauer, Phil. Mag. **21**, 863 (1970).

[76] D. S. Fisher and P. A. Lee, Phys. Rev. B **23**, 6851 (1981).

[77] M. Nardelli, B. I. Yakobson, and J. Bernholc, Phys. Rev. Lett. **81**, 4656 (1998).

[78] N. Koprinarov, M. Marinov, G. Pchelarov, M. Konstantinove, and R. Stefanov, Phys. Rev. Lett. **99**, 2042 (1996).

[79] A. Zettl, private communications.

[80] H. J. Choi and J. Ihm, Phys. Rev. B **59**, 2267 (1999).

[81] H. J. Choi, J. Ihm, S. G. Louie, and M. L. Cohen, Phys. Rev. Lett. **84**, 2917 (2000).

[82] A. Rochefort, Ph. Avouris, F. Lesage, and R. R. Salahub, Phys. Rev. B **60**, 13824 (1999).

[83] P. Delaney, H. J. Choi, J. Ihm, S. G. Louie, and M. L. Cohen, Nature (London) **391**, 466 (1998).

[84] P. Delaney, H. J. Choi, J. Ihm, S. G. Louie, and M. L. Cohen, Phys. Rev. B **60**, 7899 (1999).

[85] Y. K. Kwon, S. Saito, and D. Tománek, Phys. Rev. B **58**, R13314 (1998).

[86] M. S. Fuhrer, J. Nygard, L. Shih, M. Forero, Y. G. Yoon, M. S. C. Mazzone, H. J. Choi, J. Ihm, S. G. Louie, A. Zettl, and P. L. McEuen, Science **288**, 494 (2000).

[87] I. V. Hertel, R. E. Walkup, and P. Avouris, Phys. Rev. B **58**, 13870 (1998).

[88] P. L. McEuen, M. Bockrath, D. H. Cobden, Y. G. Yoon, and S. G. Louie, Phys. Rev. Lett. **83**, 5098 (1999).

[89] T. Ando, T. Nakkanishi, and R. Saito, J. Phys. Soc. Jpn. **67**, 2857 (1998).

[90] T. Ando and T. Nakkanishi, J. Phys. Soc. Jpn. **67**, 1704 (1998).

Appendix F

Low dimensional systems as promising thermoelectric materials: a study of one-dimensional Bismuth Nanowires

References:

- T. M. Tritt, *Recent trends in Thermoelectric Materials III*, Vol. 71 of Series of Semiconductors and Semimetals, Academic Press, Chapter 1.
- Z. Zhang, X. Sun, M. S. Dresselhaus, J. Y. Ying, and J. Heremans, *Appl. Phys. Lett.*, 73, 1589 (1998)
- Y. -M. Lin and X. Sun and M. S. Dresselhaus, *Phys. Rev. B*, 62, 4610 (2000)
- Z. Zhang, X. Sun, M. S. Dresselhaus, J. Y. Ying and J. Heremans, *Phys. Rev. B*, 61, 4850 (2000).

F.0.1 Introduction to thermoelectricity

Evaluation of new materials including low dimensional materials for thermoelectric applications is usually made in terms of the dimensionless thermoelectric figure of merit ZT where T is the temperature (in degrees Kelvin) and Z is given by

$$Z = \frac{S^2\sigma}{\kappa}, \tag{F.1}$$

where S is the thermoelectric power or Seebeck coefficient, σ is the electrical conductivity and

$$\kappa = \kappa_e + \kappa_{\rm ph}, \tag{F.2}$$

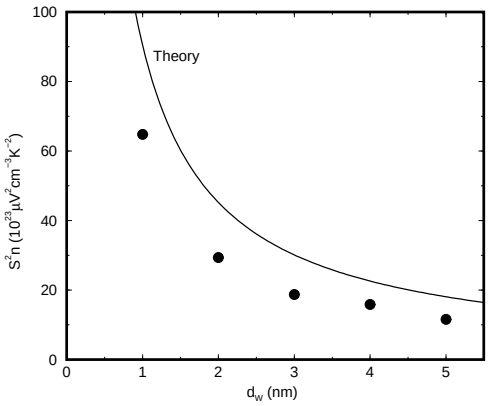


Figure F.1: The comparison between experimental data for S^2n vs quantum well width d_W and the theoretical curve at optimal doping level to maximize $Z_{2D}T$ for the optimum thermoelectric figure of merit for a strain relaxed $Si/Si_{0.7}Ge_{0.3}$ quantum well superlattice of 15 periods at room temperature.

is the thermal conductivity, which includes contributions from carriers (κ_e) and from the lattice (κ_{ph}). Equation (F.1) emphasizes the importance of a large S for high thermoelectric performance (or high ZT), where S denotes the voltage generated by a thermal gradient. Large values of ZT require high S, high σ, and low κ. Since an increase in S normally implies a decrease in σ because of carrier density considerations, and since an increase in σ implies an increase in the electronic contribution to κ as given by the Wiedemann–Franz law, it is very difficult to increase Z in typical thermoelectric materials. The best commercial 3D thermoelectric material is $Bi_{0.5}Sb_{1.5}Te_3$ in the $Bi_{2(1-x)}Sb_{2x}Te_{3(1-y)}Se_{3y}$ family with a room temperature $ZT \approx 1$. It is believed that if materials with $ZT \approx 3$ could be developed, many more practical applications for thermoelectric devices would follow.

F.1 Proof-of-Principle Studies

Early studies of low dimensional thermoelectricity focused on the demonstration of proof-of-principle of enhanced ZT within a quantum well structure using simple theoretical models, comparisons between theory and experiment, and comparisons between the low dimensional (2D) and 3D Seebeck coefficient, all comparisons being carried out under optimum doping conditions.

The demonstration of proof-of-principle in the strain relaxed $Si/Si_{0.7}Ge_{0.3}$ superlattice system is shown in Fig. F.1. For the proof-of-principle studies, superlattices were grown with 15 superlattice periods, having quantum well widths between 10–50 Å alternating with 300 Å of $Si_{0.7}Ge_{0.3}$ of barrier layer, and the measured S^2n at 300 K were compared to model calculations based on the well-established band structures of Si and $Si_{1-x}Ge_x$ alloys, using only literature values and no adjustable parameters. This comparison between the model calculation and measurements for S^2n provides clear confirmation that the reduction of the size of the quantum well results in an increase in S^2n and that the model calculation has a similar dependence on quantum well width as the experimental points, when taking account

Table F.1: Bismuth parameters

Property	Bulk	Trigonal	Binary	Bisectrix
Mass Density (g/cm^3) (300 K)	9.8			
Melting Point (K) (1 atm)	544.4			
Velocity of sound (10^5 cm/s) (4.2 K)		2.02	2.62	2.7
Velocity of sound (10^5 cm/s)(300 K)		1.972	2.540	2.571
Phonon mean free path[a] (nm) (77 K)		14.7	15.2	14.8
Thermal conductivity (300 K) (W/mK)		6.0	9.8	9.8
Lattice constant[b] (Å)		$c = 11.862$	$a = 4.5460$	$a = 4.5460$
Compressibility (Mbar^{-1})		1.82	0.62	0.62
Bulk modulus (Mbar^{-1})	0.326×10^{-3}			
Young's modulus (dyn/cm)		2.12×10^{11}	3.10×10^{11}	3.10×10^{11}
Volume coeff of thermal expansion (K^{-1})	3.965×10^{-5}			
L-point band gap (meV) (0 K)		13.6	13.6	13.6
Plasma frequency (2 K) (cm^{-1})		158 ± 3		
Work function (eV)		4.22		
Debye temperature (K)		112		
Static dielectric constant		84	105	105
Carrier density (77 K)[c] (10^{17} cm^{-3})	4.4			
Carrier density (300 K) (10^{17} cm^{-3})	2.4			

[a]From the relation $\kappa = C_v v_s \ell/3$. Along $[01\bar{1}2]$ and $[10\bar{1}1]$, $\ell = 15.3$ nm and 15.7 nm, respectively.
[b]The rhombohedral angle for Bi is $\alpha = 57°14.2'$ and sublattice parameter $u = 0.237$ as compared to $\alpha = 60°$ and $u = 0.25$ for a cubic system. For Sb the lattice constants are $a = 4.308$ Å, $c = 11.274$ Å, $\alpha = 57°6.5'$ and $u = 0.233$.
[c]The carrier density at 4 K for Bi is $2.7 \times 10^{17}/\text{cm}^3$, for Sb is $3.7 \times 10^{19}/\text{cm}^3$, and for As is $2 \times 10^{20}/\text{cm}^3$.

of experimental uncertainties in the data.

F.2 Basic properties of Bi

Bismuth is a very attractive material for low-dimensional thermoelectricity because of the very large anisotropy of the three ellipsoidal constant energy surfaces for electrons at the L-point in the rhombohedral Brillouin zone (see Fig. F.2), and the high mobility of the carriers for the light mass electrons. In addition, bulk bismuth has carriers with very long mean free paths for electronic transport and heavy mass ions which are highly effective for scattering phonons (see Table F.1). Bismuth can also be alloyed isoelectronically with antimony to yield a high mobility alloy with highly desirable thermoelectric properties.

As a bulk material, semimetallic Bi has a low Seebeck coefficient S because of the approximate cancellation between the contributions to S from the electron and hole carriers,

Table F.2: The band structure parameters of bulk Bi at $T \leq 77\,\mathrm{K}$

Parameters	Notation	Value
Band Overlap[a]	Δ_0	–38 meV
Band Gap at L-point	E_{gL}	13.8 meV
Electron Effective Mass Tensor Elements at the Band Edge for $L(A)$ pocket[b]	m_{e1}	$0.00119\,m_0$
	m_{e2}	$0.263\,m_0$
	m_{e3}	$0.00516\,m_0$
	m_{e4}	$0.0274\,m_0$
T-point Hole Effective Mass Tensor Elements at the Band Edge	m_{h1}	$0.059\,m_0$
	m_{h2}	$0.059\,m_0$
	m_{h3}	$0.634\,m_0$
Electron Mobility[c] at 77 K in units $\times 10^4\,\mathrm{cm}^2/\mathrm{Vs}$	μ_1	68.0
	μ_2	1.6
	μ_3	38.0
	μ_4	–4.3
Hole Mobility[c] at 77 K in units $\times 10^4\,\mathrm{cm}^2/\mathrm{Vs}$	$\mu_{h1} = \mu_{h2}$	12.0
	μ_{h3}	2.1

[a]The band overlap for Sb is 17.75 meV and for As is 356 meV.
[b]The tilt angle of the L-point electron ellipsoids are 6.0°, −4°, −4° for Bi, Sb, and As, respectively. The tilt angle of the H-point hole "ellipsoids" are 53° for Sb and 37.5° for the major As hole ellipsoid.
[c]The form of the effective mass tensor and the mobility tensor are assumed to be the same for a given carrier pocket.

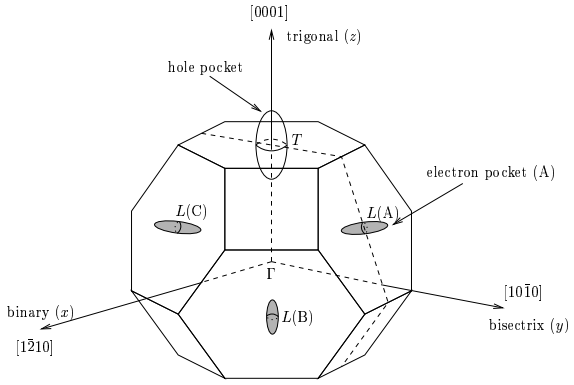


Figure F.2: The Fermi surfaces of Bi, showing the Brillouin zone with the fifth-band hole pocket about the T-point and the three sixth-band L-point electron pockets labeled A, B, and C. For quantum wells with their confinement direction, or for nanowires having their wire axes in the bisectrix-trigonal plane, the mirror plane symmetry of the bulk bismuth structure results in the crystallographic equivalence of the L-point carrier pockets B and C. However, the L-point carrier pocket A is not equivalent crystallographically to carrier pockets B or C.

since for a two carrier system, the Seebeck coefficient can be written as

$$S = \frac{\sigma_e S_e + \sigma_h S_h}{\sigma_e + \sigma_h} \tag{F.3}$$

where σ_e, σ_h, S_e and S_h respectively, denote the electrical conductivity and Seebeck coefficient for the electrons and holes. For the case of bismuth, the mobility of the electrons is much larger that of the holes, so that S tends to be weakly negative. It was early recognized that Bi could be a good thermoelectric material if the hole carriers could be removed; however, no reliable mechanism was proposed to remove the hole carriers in pure bismuth. On the other hand, it was recognized long ago that $Bi_{1-x}Sb_x$ alloys when properly doped and oriented could be among the best presently available thermoelectric materials, especially in the liquid nitrogen temperature range (near 77 K).

Low dimensionality, however, offers an opportunity to overcome the problem of the low Seebeck coefficient in pure bismuth. As the quantum well (or wire) width decreases, the band edge for the lowest subband in the conduction band rises above that for the highest subband in the valence band, thereby inducing a semimetal-semiconductor transition. If the 2D (or 1D) bismuth system is then doped to the optimum doping level, a large enhancement in $Z_{2D}T$ (and even more enhancement in $Z_{1D}T$) should be possible as the quantum well (or wire) width is decreased.

Table F.2 summarizes the band structure parameters of bulk Bi, and their temperature dependence is given in Table F.3. The modeling of Bi nanowires and the prediction of their thermoelectric properties will be based on these band structure parameters.

Table F.3: Temperature dependence of selected bulk Bi band structure parameters.

Parameters	Temperature Dependence
Band Overlap (meV)	$\Delta_0 = \begin{cases} -38 \text{ (meV)} & (T < 80\text{K}) \\ -38 - 0.044(T-80) & \\ +4.58 \times 10^{-4}(T-80)^2 & (T > 80\text{K}) \\ -7.39 \times 10^{-6}(T-80)^3 & \end{cases}$ (F.4)
Direct Band Gap (meV)	$E_{gL} = 13.6 + 2.1 \times 10^{-3}T + 2.5 \times 10^{-4}T^2$ (F.5)
L-point Electron Effective Mass Components	$[\mathbf{m}_e(T)]_{ij} = \dfrac{[\mathbf{m}_e(0)]_{ij}}{1 - 2.94 \times 10^{-3}T + 5.56 \times 10^{-7}T^2}$ (F.6)

F.3 Nanowires

F.3.1 Introduction to Nanowires

Thus far, bismuth is the dominant thermoelectric quantum wire material that has been fabricated and studied for thermoelectric applications, though some success has been demonstrated with the fabrication of quantum wires from antimony, Bi_2Te_3, and Si. However, except for the case of Bi and Sb, little attention has been given to the thermoelectric properties of these wires. Ballistic transport in thin metallic wires has been studied more generally for many years. In this section, the structure, characterization and thermoelectric-related properties of bismuth nanowires is reviewed.

F.3.2 Structure and Synthesis of Bismuth Nanowires

Arrays of hexagonally-packed parallel bismuth nanowires, 7–110 nm in diameter and 25–65 μm in length, have been prepared. These nanowires are embedded in a dielectric matrix of anodic alumina, which, because of its array of parallel nano-channels, is used as a template for preparing the Bi nanowires (see Fig. F.3). The bismuth is confined to these nanochannels and the bismuth does not diffuse into the anodic alumina matrix. The Bi nanowires are highly oriented with a common crystallographic direction along the wire axis. Since the band structure of Bi is highly anisotropic, the transport properties of Bi nanowires are expected to be dependent on the crystallographic orientation along the wire axes. In addition, structural analysis shows that the crystal structure of bulk Bi is maintained in the nanowires, indicating that many properties of bulk Bi may be utilized in modeling the behavior of Bi nanowires.

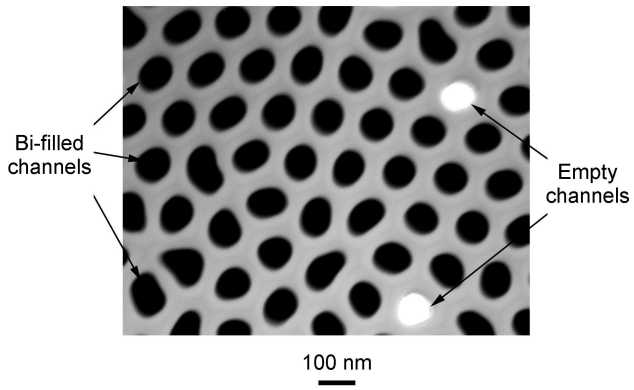


Figure F.3: Cross-sectional view of the cylindrical channels of 65 nm average diameter of an anodic alumina template, shown as a transmission electron microscope (TEM) image. The template has been mostly filled with bismuth, and the TEM image was taken after the top and bottom sides of the sample had been ion milled with 6 KV Ar ions.

F.3.3 Electronic Structure of Nanowires

To model the electronic structure, we assume, as a first approximation, the simplest possible model for an ideal 1D quantum wire, where the carriers are confined inside a cylindrical potential well bounded by a barrier of infinite potential height. An extension of this simple approach provides a reasonable approximation for a Bi nanowire embedded in an alumina template, in view of the large band gap of the anodic alumina template (3.2 eV), which provides excellent carrier confinement for the embedded quantum wires. Due to the small electron effective mass components of Bi, the quantum confinement effects in Bi nanowires are more prominent than for other wires with the same diameter.

Since the electron motion in the quantum wires is restricted in directions normal to the wire axis, the confinement causes the energies associated with the in-plane motion to be quantized, and the energy of the lowest energy subband will be raised by approximately

$$\Delta E \sim \frac{\hbar^2}{m_p^* d_W^2} \tag{F.7}$$

where m_p^* is the in-plane effective mass of the electrons and d_W is the wire diameter. Since electron motion is only allowed along the wire axis, the electrons are expected to behave like a 1D electron system, with a dispersion relation that has the form

$$E_{nm}(k_l) = \varepsilon_{nm} + \frac{\hbar^2 {k_l}^2}{2m_l^*} \tag{F.8}$$

where ε_{nm} represents a quantized energy level labeled by two quantum numbers (n, m), k_l is the wavenumber of the electron wavefunctions traveling along the wire axis, and m_l^* is the

dynamical effective mass for electrons moving along the wire. For materials with a highly anisotropic electronic energy band structure such as Bi, the mass m_p^* which determines the subband energies ε_{nm} can be very different from the mass m_l^* which characterizes the motion along the wire.

In the nanowire system, the quantized subband energy ε_{nm} and the transport effective mass m_l^* along the wire axis are the two most important band parameters that determine almost every electronic property of this unique 1D system. However, due to the highly anisotropic electron and hole pockets, the calculation of the band structure in Bi nanowires has been very challenging.

Recently, calculations of the band structure of 1D Bi quantum wires have been carried out, which explicitly take into account the cylindrical wire boundary conditions and the anisotropic carrier effective masses in Bi. In addition, the non-parabolic features of the L-point conduction band and the temperature dependence of the various band parameters are also included to provide a more accurate model for the electronic structure of the Bi nanowires.

For an infinitely long circular wire with a diameter d_W, the z' axis is taken to be parallel to the wire axis with the x' and y' axes lying on the cross-sectional plane of the wire. Since the wires are allowed to be oriented along an arbitrary direction with respect to the crystallographic directions, the inverse effective mass tensor of one of the carrier pockets in the wire coordinates (x', y', z') has the general form

$$\boldsymbol{\alpha} \equiv \mathbf{M^{-1}} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix} \tag{F.9}$$

where $\alpha_{ij} = \alpha_{ji}$. Explicit values for the components of the effective mass tensor $\mathbf{M}$ for the T-point holes and L-point electrons are given in Table F.2. Since there is only one carrier pocket for holes at the T point, Eq. (F.9) is sufficient for describing the T-point holes. However, for the L-point electrons, Eq. (F.9) describes only carrier pocket A (see Fig. F.2), and rotations of Eq. (F.9) by $\pm 2\pi/3$ around the trigonal axis are needed to obtain the effective mass tensors for the B and C electron carrier pockets.

For the T-point holes, parabolic energy bands are assumed, and the Schödinger equation is simplified and given by

$$-\frac{\hbar^2}{2}\left(\alpha_{11}\frac{\partial^2}{\partial x'^2} + \alpha_{22}\frac{\partial^2}{\partial y'^2}\right) u = \left(E - \frac{\hbar^2 k_{z'}^2}{2m_{33}}\right) u. \tag{F.10}$$

Equation (F.10) has solutions $u(x', y')$ that satisfy the boundary condition $u(r' = d_W/2) = 0$, yielding the eigenvalues of $u(x', y')$ in Eq. (F.10) that are quantized

$$E_{nm}(k_{z'}) = \varepsilon_{nm} + \frac{\hbar^2 k_{z'}^2}{2m_{33}}, \tag{F.11}$$

where ε_{nm} is the eigenvalue of Eq. (F.10) corresponding to the band edge eigenstate at $k_{z'} = 0$ labeled by the quantum numbers (n, m). Here

$$m_{33} = \hat{z}' \cdot \mathbf{M} \cdot \hat{z}' \tag{F.12}$$

Table F.4: Calculated effective mass components of each carrier pocket for determining the band structure of Bi nanowires at 77 K along the indicated crystallographic directions, based on the effective mass parameters of bulk bismuth given in Table F.2. The z' direction is chosen along the wire axis. Calculations are done for Bi nanowires oriented along the three principal axes (trigonal, binary, and bisectrix) and the $[01\bar{1}2]$ and $[10\bar{1}1]$ directions (which are preferential growth directions for Bi nanowires) and correspond to (0,0.8339,0.5519) and (0,0.9503,0.3112) in Cartesian coordinates, respectively. All mass values in this table are in units of the free electron mass, m_0.

Mass Component		Trigonal	Binary	Bisectrix	$[01\bar{1}2]$	$[10\bar{1}1]$
e^- pocket A	$m_{x'}$	0.1175	0.0023	0.0023	0.0029	0.0024
	$m_{y'}$	0.0012	0.2659	0.0012	0.0012	0.0012
	$m_{z'}$	0.0052	0.0012	0.2630	0.2094	0.2542
e^- pocket B	$m_{x'}$	0.1175	0.0023	0.0023	0.0016	0.0019
	$m_{y'}$	0.0012	0.0016	0.0048	0.0125	0.0071
	$m_{z'}$	0.0052	0.1975	0.0666	0.0352	0.0526
e^- pocket C	$m_{x'}$	0.1175	0.0023	0.0023	0.0016	0.0019
	$m_{y'}$	0.0012	0.0016	0.0048	0.0125	0.0071
	$m_{z'}$	0.0052	0.1975	0.0666	0.0352	0.0526
hole pocket	$m_{x'}$	0.0590	0.6340	0.6340	0.1593	0.3261
	$m_{y'}$	0.0590	0.0590	0.0590	0.0590	0.0590
	$m_{z'}$	0.6340	0.0590	0.0590	0.2349	0.1147

is the effective mass component along the wire axis, and the in-plane effective masses are

$$\begin{aligned} m_{x'} &\equiv \alpha_{11}^{-1} = (\hat{x'} \cdot \mathbf{M^{-1}} \cdot \hat{x'})^{-1} \\ m_{y'} &\equiv \alpha_{22}^{-1} = (\hat{y'} \cdot \mathbf{M^{-1}} \cdot \hat{y'})^{-1} \end{aligned} \tag{F.13}$$

in the x' and y' directions, respectively. The eigenvalue ε_{nm} has an analytic expression only when $\alpha_{11} = \alpha_{22}$, but in general, the values of ε_{nm} must be solved numerically. This is true for both the T-point holes and for the L-point electrons.

From the band structure parameters for bulk Bi (Table F.2), values for $m_{x'}$, $m_{y'}$, and $m_{z'}$ (or $\tilde{m}_z$) at 77 K for Bi nanowires for the three principal crystallographic axes (trigonal, binary, and bisectrix directions), and for the preferential $[01\bar{1}2]$ and $[10\bar{1}1]$ growth directions are given in Table F.4. It should be noted that $m_{x'}$ and $m_{y'}$ can be interchanged without affecting any physical results.

The calculated subband structure of Bi quantum wires oriented along the $[01\bar{1}2]$ directions are shown in Fig. F.4 at 77 K. For the $[01\bar{1}2]$ wires, the degeneracy at the L point is lifted, resulting in two inequivalent groups of carrier pockets: a single electron pocket A and two electron pockets B, C with the same symmetry and band parameters as each other but different from pocket A. The L-point electron pocket A has smaller mass components ($m_{x'}$, $m_{y'}$) in the quantum confined direction than the electron pockets B and C (see Table F.4). Therefore, the electron pocket A forms a higher energy conduction subband, while the electron pockets B and C form a two-fold degenerate subband at a lower energy (see Fig. F.4). It should be pointed out that since electron pocket A has a larger mass component ($m_{z'}$) along the wire axis than pockets B and C, the dispersion relation of the $L(A)$ subband has a smaller curvature (see Fig. F.4). The band edge of the lowest subband of the $L(B,C)$ electrons increases with decreasing wire diameter d_W, while the highest subband edges of the T-point and L-point holes move downwards in energy. At $d_W < 49.0$ nm, the energy of the lowest L-point conduction subband edge exceeds that of the highest T-point valence subband edge, indicating that these nanowires become semiconducting.

F.3.4 Doping of Bi nanowires

For intrinsic Bi nanowires, equal numbers of electrons and holes are expected, whether in the semimetallic or semiconducting state. However, in thermoelectric applications of bismuth, it is necessary to control the Fermi level so that: (1) the transport phenomena are dominated by a single type of carrier only, i.e., electrons or holes, (2) the electrochemical potential is placed to achieve the optimum ZT. In addition, to optimize the efficiency of thermoelectric devices, it is essential to obtain a high Seebeck coefficient S. However, since the contributions from holes and from electrons have opposite signs with regard to the Seebeck coefficient, the magnitude of S in pure Bi is usually very small, although individual contributions from electrons or holes can be quite significant. Therefore, it is expected that Bi nanowires can be a very promising thermoelectric material if the Fermi level can be adjusted properly so that only electrons or only holes contribute to S and the electrochemical potential is set to maximize ZT (see §F.2).

Since Bi is a group V element, the Fermi level can be increased by introducing a small amount of group VI element, such as Te, which acts as an electron donor in Bi. Group IV elements such as Sn or Pb, on the other hand, act as electron acceptors in Bi, and can be used to synthesize p-type Bi.

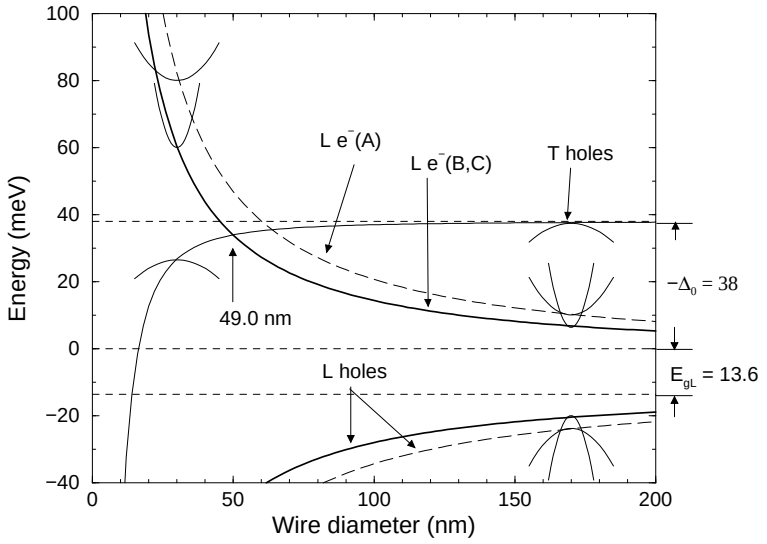


Figure F.4: The subband structure at 77 K of Bi quantum wires oriented along the $[01\bar{1}2]$ growth direction, showing the energies of the highest subbands for the T-point hole carrier pocket, the L-point electron pockets (A, B and C) as well as the L-point holes. The zero energy refers to the conduction band edge in bulk Bi. As the wire diameter d_W decreases, the conduction subbands move up in energy, while the valence subbands move down. At $d_c = 49.0\,\mathrm{nm}$, the lowest conduction subband edge formed by the $L(B, C)$ electrons crosses the highest T-point valence subband edge, and a semimetal-semiconductor transition occurs.

We assume that the electronic band structure of Te-doped Bi is the same as pure Bi in the spirit of a rigid band approximation, except for the presence of a Te donor energy level located below the conduction band edge by E_d.

The ionization energy E_d required for releasing donor electrons from Te atoms can be estimated using the Bohr hydrogen-like model,

$$E_d \simeq 13.6 \times \frac{m^*/m_0}{\epsilon^2} \text{ (eV)} \tag{F.14}$$

where m^* is the effective mass at the conduction band edge, m_0 is the free electron mass, and ϵ is the dielectric constant ($\epsilon \simeq 100$) of Bi in the low frequency limit (see Table F.1). The value for the effective mass m^* is taken as the average value of the electron effective mass tensor, and at low temperatures, $m^*/m_0 \simeq 0.002$. With this large value for ϵ and small value for m^*, we obtain a very small value for E_d

$$E_d \simeq 3.20 \times 10^{-3} \text{meV} \tag{F.15}$$

with a very large effective Bohr radius

$$a_0^* = 0.5 \text{ (Å)} \times \frac{\epsilon}{m^*} \simeq 2.5 \mu\text{m}, \tag{F.16}$$

so that the energy required to ionize the Te atom in Bi is very small. However, it should be noted that the results obtained in Eqs. (F.15) and (F.16) are only valid for bulk materials. In Te-doped Bi nanowires, where the wire diameter d_W is smaller than the effective Bohr radius a_0^*, the ionization energy will be increased due to the confinement of donor electrons to be close to the impurity atom. Since the Coulomb potential energy is inversely proportional to the distance between two charged particles, a rough estimation of the ionization energy $E_{d(1\text{D})}$ in nanowires can be made by multiplying the ionization energy $E_{d(3\text{D})}$ in bulk materials [Eq. (F.15)] by the ratio between the effective Bohr radius and the wire diameter. Thus, for 40 nm Te-doped Bi nanowires, an estimate for the ionization energy would be

$$E_{d(1\text{D})} \simeq E_{d(3\text{D})} \times \frac{a_0^*}{d_W} \simeq 0.2 \,\text{meV}. \tag{F.17}$$

At very low temperatures where the thermal energy $k_B T \leq E_{d(1\text{D})}$, the donor electrons will freeze out, and the freeze-out temperature is $T_{fo} \simeq 2\,\text{K}$ for 40 nm Te-doped Bi nanowires. For the temperatures of interest (ranging from 4 K to 300 K), we can assume that all the donor atoms are ionized in 40 nm Te-doped Bi nanowires and that each Te atom donates one electron to the conduction band.

F.3.5 Semi-Classical Transport Model for Bi Nanowires

The thermoelectric-related transport coefficients of Te-doped Bi nanowires can be derived from the simple semi-classical model, which is based on the Boltzmann transport equation.

We define the transport-related quantities $\mathcal{L}_{1\text{D}}^{(\alpha)}$ as

$$\mathcal{L}_{1\text{D}}^{(\alpha)} = e^2 \int \frac{8dk}{\pi^2 d_W^2} \left(-\frac{df}{dE}\right) \tau(k) v(k) v(k) [E(k) - E_f]^\alpha, \tag{F.18}$$

in which $E(k)$ is the electronic dispersion relation, $\tau(k)$ is the relaxation time, E_f is the Fermi energy, and $f(E)$ is the Fermi–Dirac distribution function

$$f(E) = \frac{1}{1 + e^{(E-E_f)/k_BT}}. \tag{F.19}$$

Since the numerical calculation of Eq. (F.18) requires knowledge of the k dependence of the relaxation time $\tau(k)$, and since the calculation of $\tau(k)$ from fundamental principles of scattering mechanisms is usually very complicated, we use a simple first approximation, known as the constant relaxation time approximation, to simplify the calculations of the thermoelectric properties of materials. In this formalism, $\tau(k) = \tau$ is taken to be constant in k, and in energy, and τ can be related to the carrier mobility μ along the wire by

$$\mu = \frac{e\tau}{m^*} \tag{F.20}$$

where m^* is the transport effective mass along the wire, and μ can, in principle, be obtained from experimental measurements. Thus, the integration of Eq. (F.18) can be carried out readily as long as the dispersion relation $E(k)$ is known. For a one-band system described by a parabolic dispersion relation, Eq. (F.18) becomes

$$\mathcal{L}_{1\mathrm{D}}^{(0)} = D\left[\frac{1}{2}F_{-\frac{1}{2}}\right] \tag{F.21}$$

$$\mathcal{L}_{1\mathrm{D}}^{(1)} = \begin{cases} (k_BT)D\left[\frac{3}{2}F_{\frac{1}{2}} - \frac{1}{2}\zeta^* F_{-\frac{1}{2}}\right] & \text{(for electrons)} \\ -(k_BT)D\left[\frac{3}{2}F_{\frac{1}{2}} - \frac{1}{2}\zeta^* F_{-\frac{1}{2}}\right] & \text{(for holes)} \end{cases} \tag{F.22}$$

$$\mathcal{L}_{1\mathrm{D}}^{(2)} = (k_BT)^2 D\left[\frac{5}{2}F_{\frac{3}{2}} - 3\zeta^* F_{\frac{1}{2}} + \frac{1}{2}\zeta^{*2} F_{-\frac{1}{2}}\right] \tag{F.23}$$

where D is given by

$$D = \frac{16e}{\pi d_W^2}\left(\frac{2m^* k_BT}{\hbar^2}\right)^{\frac{1}{2}}\mu, \tag{F.24}$$

and F_i denotes the Fermi–Dirac related functions which is given by

$$F_i = \int_0^\infty \frac{x^i dx}{\exp(x - \zeta^*) + 1}, \tag{F.25}$$

with fractional indices $i = -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \ldots$. The reduced chemical potential ζ^* is defined as:

$$\zeta^* = \begin{cases} (E_f - \varepsilon_e^{(0)})/k_BT & \text{(for electrons)} \\ (\varepsilon_h^{(0)} - E_f)/k_BT & \text{(for holes).} \end{cases} \tag{F.26}$$

where $\varepsilon_e^{(0)}$ and $\varepsilon_h^{(0)}$ are the band edges for electrons and holes, respectively.

For Bi quantum wire systems, there are many 1D subbands due to the multiple carrier pockets at the L points and the T point, and the quantum confinement-induced band splitting also forms a set of 1D subbands from each single band in bulk materials. Therefore, when considering the transport properties of real 1D nanowire systems, contributions from all of the subbands near the Fermi energy should be included. In a multi-band system, the

$\mathcal{L}^{(\alpha)}$'s, should be replaced by the sum $L^{(\alpha)}_{\rm total} = \sum_i \mathcal{L}^{(\alpha)}_i$ of contributions from each subband (label by i), and the transport coefficients σ, S, and κ_e then become

$$\sigma_{\rm total} = \sum_i \sigma_i \tag{F.27}$$

$$S_{\rm total} = \frac{\sum_i \sigma_i S_i}{\sum_i \sigma_i} \tag{F.28}$$

$$\kappa_{e,{\rm total}} = \frac{1}{e^2 T}\left[L^{(2)}_{\rm total} - \frac{(L^{(1)}_{\rm total})^2}{L^{(0)}_{\rm total}}\right] \tag{F.29}$$

where σ_i and S_i are the electrical conductivity and the thermopower corresponding to each subband, respectively. In Bi nanowires, the sums in Eqs. (F.27)–(F.29) include subbands associated with electron pockets A and (B,C) as well as contributions from T-point holes and L-point holes (see §F.3.3).

Another physical quantity of interest in thermoelectric applications is the lattice thermal conductivity κ_L which, together with electronic thermal conductivity κ_e, determines the total thermal conductivity of the system. From kinetic theory, the thermal conductivity of phonons is given by

$$\kappa_L = \frac{1}{3} C_v v \ell \tag{F.30}$$

where C_v is the heat capacity per unit volume, v is the sound velocity, and ℓ is the mean free path for phonons. We note that, for an ideal quantum wire system embedded in a host material with a large band gap, the electron wavefunctions are well confined within the quantum wire, and they can only travel along the wire axis. However, the host material that confines electrons cannot confine the phonon paths, and thus, because of acoustic mismatch, phonons will be scattered when they move across the wire boundary. This increased boundary scattering of phonons in the quantum wire system will decrease the phonon mean free path ℓ as well as the lattice thermal conductivity along the wire. The simplest approximation to model the lattice thermal conductivity in the quantum wire system is to replace the phonon mean free path ℓ in Eq. (F.30) by the wire diameter d_W if $d_W < \ell$ in the bulk material. It should be noted that for $d_W \ll \ell$, the lattice thermal conductivity is expected to decrease dramatically, more so than the decrease in the electrical conductivity. This reduction in the lattice thermal conductivity is one of the reasons for the expected enhanced thermoelectric performance in low-dimensional systems.

F.3.6 Optimization of $Z_{\rm 1D}T$ through doping at 77 K

In applying the general results of §F.3.5 to calculate $Z_{\rm 1D}T$ for Bi nanowires, the anisotropy of the electronic structure results in anisotropic effective mass tensors, and other related quantities have to be considered. For example, the mobility tensor for each carrier pocket for Bi is also highly anisotropic. For the L-point electron pocket A (see Fig. F.2), the mobility tensor has the form

$$\boldsymbol{\mu}_{e(A)} = \begin{pmatrix} \mu_{e1} & 0 & 0 \\ 0 & \mu_{e2} & \mu_{e4} \\ 0 & \mu_{e4} & \mu_{e3} \end{pmatrix}, \tag{F.31}$$

Table F.5: Values of the mobility tensor elements for electron and hole pockets of Bi at 77 K. The mobility values are given in units of $\mathrm{cm^2V^{-1}s^{-1}}$.

μ_{e1}	μ_{e2}	μ_{e3}	μ_{e4}	μ_{h1}	μ_{h3}
6.8×10^3	1.6×10^4	3.8×10^3	-4.3×10^4	1.20×10^5	2.1×10^4

and the mobility tensors for electron pockets B and C can be derived by a rotation of $\boldsymbol{\mu}_{e(A)}$ by $\pm120^\circ$ about the trigonal axis. For the T-point holes, the mobility tensor has the form

$$\boldsymbol{\mu}_h = \begin{pmatrix} \mu_{h1} & 0 & 0 \\ 0 & \mu_{h1} & 0 \\ 0 & 0 & \mu_{h3} \end{pmatrix}. \tag{F.32}$$

The values for these mobility tensor elements at 77 K are listed in Table F.5. Since the mobility tensors are anisotropic, the mobility $\mu_{\hat{l}}$ for carriers traveling along the wire will depend on the wire orientation, and $\mu_{\hat{l}}$ is given by

$$\mu_{\hat{l}} = (\hat{l} \cdot \boldsymbol{\mu}^{-1} \cdot \hat{l})^{-1} \tag{F.33}$$

where $\hat{l}$ is the unit vector along the wire axis, which follows from the general definition of the carrier mobility in terms of $\mu = e\tau/m^*$ and from Matthiessen's rule summing $1/\tau_i$ for each scattering process i.

The lattice thermal conductivity in bulk Bi is also anisotropic, and has the form

$$\boldsymbol{\kappa}_L = \begin{pmatrix} \kappa_{L,\perp} & 0 & 0 \\ 0 & \kappa_{L,\perp} & 0 \\ 0 & 0 & \kappa_{L,\parallel} \end{pmatrix} \tag{F.34}$$

where $\kappa_{L,\parallel}$ and $\kappa_{L,\perp}$ are the thermal conductivities parallel and perpendicular to the trigonal axis, respectively. By extrapolating the experimental data for $\boldsymbol{\kappa}_L$ measured between 100 K and 300 K, the lattice thermal conductivity tensor elements at 77 K are estimated as $\kappa_{L,\perp} = 13.2\,(\mathrm{W/mK})$ and $\kappa_{L,\parallel} = 9.9\,(\mathrm{W/mK})$, respectively. For Bi nanowires oriented in directions other than the three principal axes, the lattice thermal conductivity along the wire is then given by

$$\begin{aligned} \kappa_{L,\hat{l}} &= \hat{l} \cdot \boldsymbol{\kappa}_L \cdot \hat{l} \\ &= \cos^2\theta\kappa_{L,\perp} + \sin^2\theta\kappa_{L,\parallel} \end{aligned} \tag{F.35}$$

where θ is the angle between the wire axis and the trigonal axis.

The phonon mean free paths ℓ in Bi nanowires are estimated by the heat capacity C_v, sound velocities $\boldsymbol{v}$ and the thermal conductivity $\kappa_{L,\hat{l}}$ via Eq. (F.30). At 77 K, the value for the heat capacity of Bi is measured as $C_v \simeq 1.003\,(\mathrm{JK^{-1}cm^{-3}})$. The measured sound velocities $\boldsymbol{v}$ of Bi at 1.6 K and 300 K along selected directions are listed in Table F.3.6, in which the interpolated values for $\boldsymbol{v}$ at 77 K are also given. The calculated phonon mean free path ℓ of bulk Bi at 77 K is listed in Table F.1 for Bi crystals oriented along the three

Table F.6: The sound velocities v of Bi along the three principal axes. The values of 77 K are interpolated from the experimentally measured results at 1.6 K and 300 K.

	v (10^5 cm/s)					
T (K)	Trigonal	Binary	Bisectrix	$(0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$	$[01\bar{1}2]$	$[10\bar{1}1]$
1.6	2.02	2.62	2.70	2.15	2.26	2.58
77	2.01	2.60	2.67	2.13	2.24	2.45
300	1.972	2.540	2.571	2.082	2.18	2.375

principal axes, and also along the $[10\bar{1}1]$, and the $[01\bar{1}2]$ directions. As the wire diameters become smaller than the phonon mean free path calculated in bulk Bi, the scattering at the wire boundary becomes the dominant scattering process for phonons, and the phonon mean free path in these nanowires is approximately limited by the wire diameter, or $l \simeq d_W$. Thus, the lattice thermal conductivity κ_L in Te-doped Bi nanowires will decrease significantly as the wire diameter decreases below $d_W \leq 15$ nm (see Table F.1).

Using the general formalism presented in §F.3.5 for S, σ, κ_e, and the above discussion on κ_L and procedures to account for the multiple carrier pockets and their anisotropy, the thermoelectric figure of merit $Z_{1\mathrm{D}}T$ has been calculated. Figure F.5 shows the calculated $Z_{1\mathrm{D}}T$ for n-type Bi nanowires oriented along the trigonal axis at 77 K as a function of the dopant concentrations for three different wire diameters. We note that the value of $Z_{1\mathrm{D}}T$ for a given donor concentration increases drastically with decreasing wire diameter d_W, and the maximum $Z_{1\mathrm{D}}T$ for each wire diameter occurs at an optimized donor concentration $N_{d(\mathrm{opt})}$ which increases somewhat as the wire diameter decreases. For 5 nm Bi nanowires oriented along the trigonal axis at 77 K, the maximum $Z_{1\mathrm{D}}T$ at 77 K is about 6, with an optimized electron concentration $N_{d(\mathrm{opt})} \simeq 10^{18}\,\mathrm{cm}^{-3}$. The value of $Z_{1\mathrm{D}}T$ also strongly depends on the wire orientation due to the anisotropic nature of the Bi band structure and of the thermal properties of Bi. Figure F.6 shows the calculated figure of merit $Z_{1\mathrm{D}}T$ at 77 K as a function of donor concentration N_d for 10 nm Bi nanowires oriented in different directions. For 10 nm Bi nanowires at 77 K, the trigonal nanowires have the highest optimal $Z_{1\mathrm{D}}T$ which is about 2.0, while bisectrix wires have the lowest optimal $Z_{1\mathrm{D}}T \simeq 0.4$. The optimum carrier concentrations $N_{d(\mathrm{opt})}$ and the corresponding $Z_{1\mathrm{D}}T$ of n-type Bi nanowires at 77 K are listed in Table F.7 for various wire diameters and orientations. Figure F.7 shows the calculated optimal $Z_{1\mathrm{D}}T$ at 77 K as a function of wire diameter for n-type Bi nanowires oriented along the three principal axes, and the $[10\bar{1}1]$, and the $[01\bar{1}2]$ growth directions.

As a comparison, the optimum acceptor concentration $N_{a(\mathrm{opt})}$ and the corresponding $Z_{1D}T$ for p-type Bi nanowires are calculated and listed in Table F.7 for various wire diameters and orientations. Compared with the results in Table F.7 for n-type Bi nanowires, we note that p-type Bi nanowires in general have a much lower $Z_{1D}T$.

F.3.7 Temperature-Dependent Resistivity of Bi Nanowires

Measurements of the temperature dependence of the resistance $R(T)$ of Bi nanowire arrays have been carried out on samples prepared both from the liquid phase by pressure injection and by vapor phase deposition, yielding results consistent with each other. Due

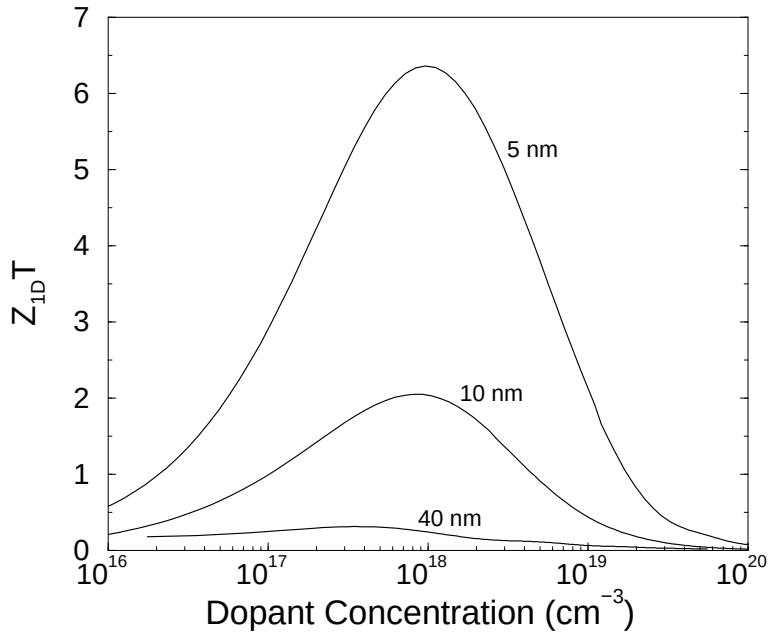


Figure F.5: Calculated $Z_{1D}T$ for Te-doped Bi nanowires oriented along the trigonal axis at 77 K as a function of Te dopant concentration for three different wire diameters.

Table F.7: The optimum dopant concentrations $N_{d(\mathrm{opt})}$ (in $10^{18}\,\mathrm{cm}^{-3}$) and the corresponding $Z_{1D}T$ of n-type and p-type Bi nanowires at 77 K for various wire diameters and orientations.

Wire Orientation		5 nm		10 nm		40 nm	
		$N_{d(\mathrm{opt})}$	$Z_{1D}T$	$N_{d(\mathrm{opt})}$	$Z_{1D}T$	$N_{d(\mathrm{opt})}$	$Z_{1D}T$
Trigonal	(n-type)	0.96	6.36	0.81	2.0	0.38	0.31
	(p-type)	0.96	6.36	12.9	0.72	6.2	0.17
Binary	(n-type)	0.35	3.68	0.28	1.14	0.56	0.13
	(p-type)	0.79	1.78	10.3	0.16	7.9	0.05
Bisectrix	(n-type)	4.1	2.21	1.78	0.40	4.97	0.03
	(p-type)	0.74	0.32	0.19	0.40	0.50	0.07
$[10\bar{1}1]$	(n-type)	3.21	2.69	1.57	0.51	2.57	0.04
	(p-type)	1.04	1.16	0.43	0.19	0.63	0.05
$[01\bar{1}2]$	(n-type)	2.07	3.41	1.33	0.70	2.73	0.06
	(p-type)	2.59	2.46	0.58	0.18	0.75	0.03

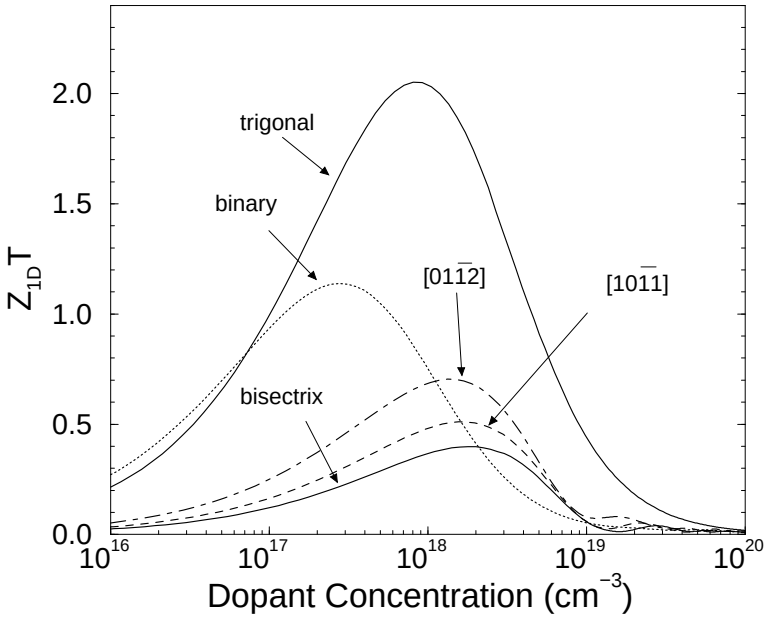


Figure F.6: Calculated $Z_{1\mathrm{D}}T$ at 77 K as a function of Te dopant concentration N_d for 10 nm Te-doped Bi nanowires oriented along different directions: trigonal, binary, bisectrix, [10$\bar{1}$1], and [01$\bar{1}$2]

.

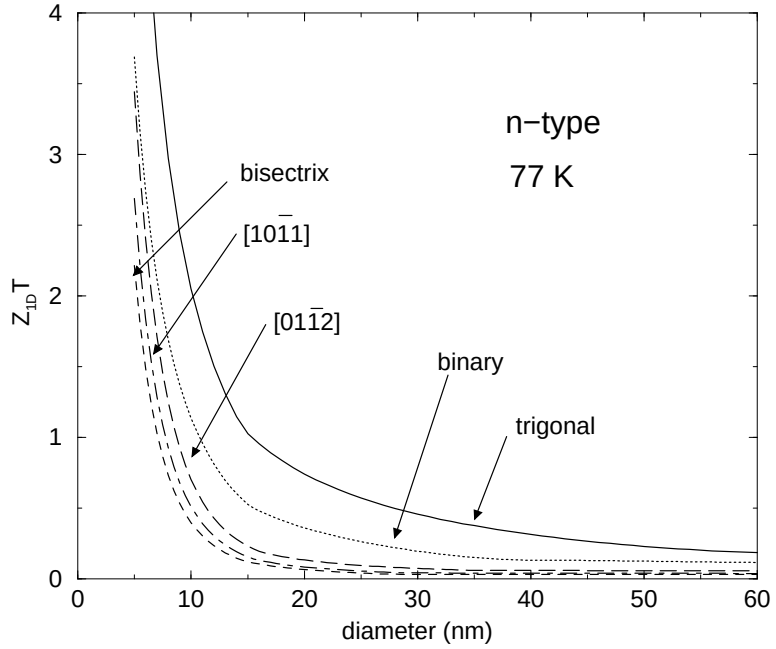


Figure F.7: Calculated $Z_{1\mathrm{D}}T$ at 77 K as a function of Te dopant concentrations N_d for 10 nm Te-doped Bi nanowires oriented along different directions: trigonal, binary, bisectrix, [10$\bar{1}$1], and [01$\bar{1}$2].

to the geometric limitations, the first attempt to study the temperature dependent properties of Bi nanowires was made with a two-probe measurement. Although an absolute value of the resistivity cannot be derived through this two-probe method, the temperature dependence can be examined by normalizing the resistance to that at a common temperature, e.g., $R(300\,\mathrm{K})$. Figure F.8 shows the temperature dependence of the resistivity $R(T)/R(300\,\mathrm{K})$ of Bi nanowire arrays of various wire diameters prepared by the vapor deposition process.

As shown in Fig. F.8, the temperature dependence of the resistance of Bi nanowires is very different from that of bulk Bi, and is very sensitive to the wire diameter. At high temperatures ($T > 70\,\mathrm{K}$), the resistance of all nanowire arrays shown in this figure increases with decreasing temperature. When $T < 70\,\mathrm{K}$, the resistance of the smaller diameter Bi nanowires (7 nm – 48 nm) in the semiconducting regime continues to increase with decreasing temperature, while the resistance decreases with decreasing temperature for the nanowire samples of larger diameters (70 nm and 200 nm) in the semimetallic regime.

The striking difference in the temperature dependence of the resistance between Bi nanowires arrays and bulk Bi can readily be explained qualitatively, and a simple physical argument is given here in terms of the temperature dependence of the carrier density $n(T)$ and mobility $\mu(T)$. For both Bi nanowires and bulk Bi, $n(T)$ increases with increasing temperature, while $\mu(T)$ decreases. However, for Bi nanowires with larger wire diameters which are semimetallic (e.g., 70 nm and 200 nm), the increase in the carrier density with increasing temperature is much slower than that in the semiconducting wires with smaller diameters (e.g., $d_W \leq 48\,\mathrm{nm}$), especially at low temperatures. For smaller diameter nanowires ($< 48\,\mathrm{nm}$ in Fig. F.8), the increase in $n(T)$ outweighs the decrease in $\mu(T)$ when the temperature increases, and therefore the resistance drops. On the other hand, for bulk Bi or Bi nanowire arrays with larger wire diameters (70 nm and 200 nm in Fig. F.8), the temperature dependence of $\mu(T)$ becomes more important due to the weaker T dependence of $n(T)$ in the semimetallic regime, and therefore, the resistance increases with increasing temperature for $T < 100\,\mathrm{K}$. The carrier density of the 7 nm semiconducting Bi nanowires increases by many orders of magnitude ($\sim 10^8$) when the temperature increases from 100 K to 300 K, while that of the 70 nm nanowires also increases by about 16 times. However, since the mobility of bulk Bi decreases by a factor of 13 from 100 K to 300 K, the increase in $n(T)$ overwhelms that of $\mu(T)$, and the resistance of Bi nanowires (7–70 nm) decreases with temperature for $T > 100\,\mathrm{K}$.

Based on the model for the electronic structure for Bi nanowires (see §F.3.3) and the transport model (developed in §F.3.5 for the more general case of a doped Bi nanowire system), the normalized temperature dependent resistance $R(T)/R(300\,\mathrm{K})$ for 70 nm and 36 nm Bi nanowires has been calculated, and the results are shown by the solid curves in Fig. F.9, exhibiting trends consistent with the experimental results in Fig. F.8. Calculations for the wire diameters of 70 nm and 36 nm are particularly interesting, because these wires represent two different types of Bi nanowires: semimetallic and semiconducting, respectively. In the modeling of Fig. F.9, some assumptions were made to take into account the discrepancies between an ideal 1D Bi quantum wire and a real Bi nanowire. First, in a perfect single crystalline Bi quantum wire, there is no scattering at the wire boundary because the electron (or hole) wave-function and the local carrier density vanish at the wire boundary as a result of the assumed ideal infinite-potential interface. Within the Bi nanowire, possible scattering mechanisms are electron-phonon and electron-electron interactions. However, in

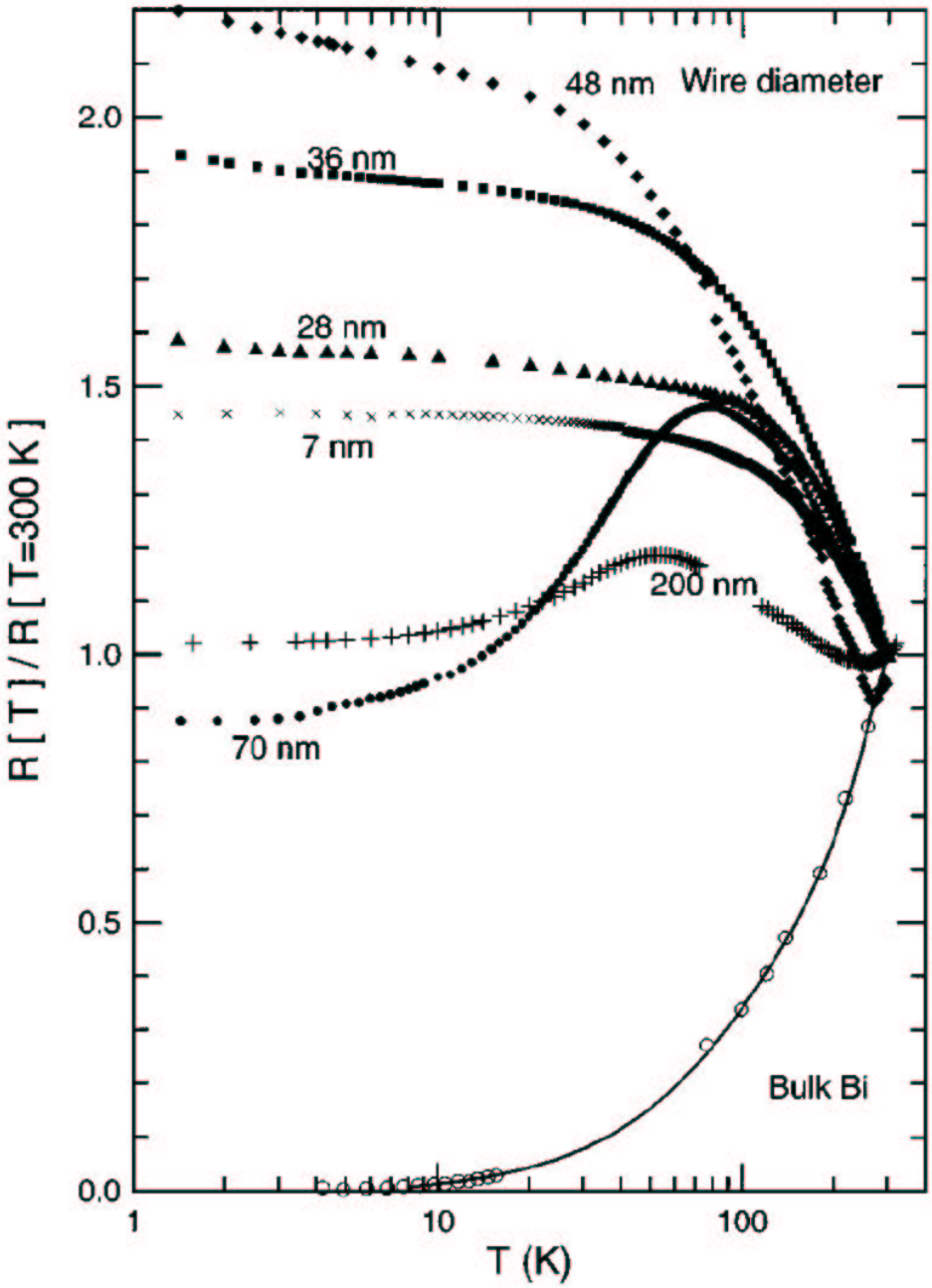


Figure F.8: Temperature dependence of the normalized resistance for Bi nanowire arrays of various wire diameters prepared by the vapor deposition method, in comparison with the corresponding data for bulk Bi. The measurement of the resistance was made while the Bi nanowires were in their alumina templates using a two-probe measurement technique.

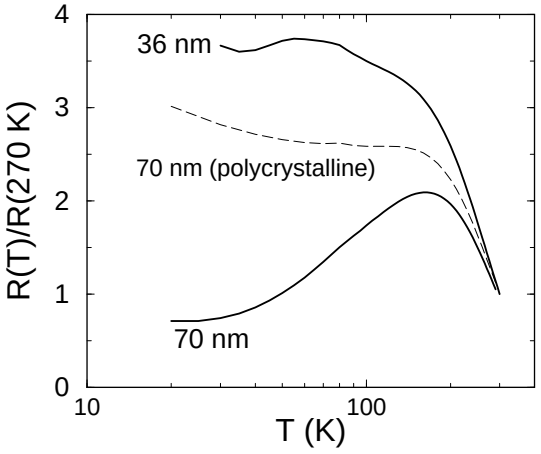


Figure F.9: The calculated temperature dependence of the resistance for Bi nanowires of 36 nm and 70 nm, using a semiclassical transport model.

a real Bi nanowire sample, the boundary conditions are far from ideal, and the energy barrier at the boundary is finite instead of infinite. Furthermore, real Bi nanowires may have a higher defect level in a thin layer at the boundary than in the interior of the nanowires due to the various surface conditions at the Bi/Al_2O_3 interface. Therefore, the electrons will experience substantial boundary scattering, due to the finite amplitude of the electron wave functions at the boundary. This boundary scattering effect has been observed in magnetoresistance measurements (see §F.3.8). In addition to the scattering at the boundary, the electrons can also be scattered at grain boundaries within real Bi nanowire samples, for which a domain size on the order of the wire diameter has been observed. However, since the domains in Bi nanowires possess the same crystal orientation along the wire axis, the small-angle scattering at the grain boundary would be expected to be a minor scattering mechanism in determining the transport properties of real Bi nanowires. Another striking difference between ideal Bi nanowires and real Bi nanowire samples is the uncontrolled impurities which act as dopants in Bi nanowires. For ideal semiconducting Bi quantum wires, the carrier density should decay exponentially with decreasing T at low temperatures, and the resistance should correspondingly increase dramatically. Instead, even for the 7 nm Bi nanowires, the measured resistance increases slowly and steadily with decreasing temperature (see Fig. F.8). This effect can be attributed to the uncontrolled impurities in the Bi nanowires, which give rise to a finite carrier density at low temperatures. The uncontrolled impurities will not only alter the carrier density, but will also decrease the carrier mobility by ionized impurity scattering.

The effect of each scattering mechanism mentioned above can be characterized by a scattering time τ, and the total scattering time $\tau_{\rm tot}$ in a real Bi nanowire can be approximated by adding the scattering rates in accordance with Matthiessen's rule

$$\frac{1}{\tau_{\rm tot}(T)} = \frac{1}{\tau_{\rm bulk}(T)} + \frac{1}{\tau_{\rm boundary}} + \frac{1}{\tau_{\rm imp}(T)} \tag{F.36}$$

in which $\tau_{\rm bulk}$ is the total relaxation time in bulk Bi, and $\tau_{\rm boundary}$ and $\tau_{\rm imp}$ are the relaxation

times for boundary scattering (including wire and grain boundary scattering) and ionized impurity scattering, respectively. It should be noted that boundary scattering and ionized impurity scattering are only important at low temperatures (<100 K), and that phonon scattering becomes the dominant scattering mechanism at higher temperatures (> 100 K). The relaxation time for ionized impurity scattering is approximately proportional to $T^{3/2}$, while boundary scattering is much less temperature dependent, and for simplicity τ_{boundary} was assumed to be a constant, independent of temperature. Since the mobility μ is proportional to the relaxation time τ, the approximation used to find the mobility, considering all of these scattering effects, was:

$$\frac{1}{\mu_{\text{tot}}(T)} = \frac{1}{\mu_{\text{bulk}}(T)} + \frac{1}{\mu_{\text{boundary}}} + \frac{1}{\mu_{\text{imp}}(T)} \tag{F.37}$$

in which $\mu_{\text{bulk}}(T)$ can be found in the literature, μ_{boundary} was assumed to be constant in T, and $\mu_{\text{imp}} \sim T^{3/2}$. In the curve for the 70 nm Bi nanowires in Fig. F.9, the two mobilities μ_{boundary} and μ_{imp} were fitted to 50 $\text{m}^2\text{V}^{-1}\text{s}^{-1}$ and $1.0 \times T^{3/2}\,\text{m}^2\text{V}^{-1}\text{s}^{-1}$, respectively. As for the 36 nm Bi nanowires, the mobility terms were fit to $\mu_{\text{boundary}} \simeq 33\ \text{m}^2\text{V}^{-1}\text{s}^{-1}$ and $\mu_{\text{imp}} \simeq 0.2 \times T^{3/2}\,\text{m}^2\text{V}^{-1}\text{s}^{-1}$, and the carrier density, due to the presence of uncontrolled impurities, was fitted to $N_{\text{imp}} \simeq 5 \times 10^{16}\,\text{cm}^{-3}$, which amounts to less than 100 impurity atoms in the Bi nanowires per 1 μm in length. However, since the 70 nm Bi nanowires are semimetallic at low temperatures, the carrier density contribution due to the small amount of uncontrolled impurities per se has an insignificant effect, and this effect is neglected in the modeling, taking account only of the effect of these uncontrolled impurities on scattering carriers. We also note that μ_{boundary} is smaller for the 36 nm Bi nanowires than for the 70 nm nanowires, due to their smaller wire diameter. As the wire diameter decreases, the contribution of the μ_{imp} term decreases more rapidly that the μ_{boundary} term. Regarding the main contribution of the last two terms to Eq. (F.37), the term $\mu_{\text{boundary}}^{-1}$ dominates over μ_{imp}^{-1} for 77 K and above, and the μ_{imp}^{-1} term is relatively larger for the small diameter nanowires. The normalized resistance $R(T)/R(300\,\text{K})$ curves in Fig. F.9 show general trends for the temperature dependence of the normalized resistance of the 36 nm and 70 nm Bi nanowires, consistent with the experimental results in Fig. F.8 for the actual nanowire arrays, showing strong evidence that the different temperature dependences of $R(T)/R(270\,\text{K})$ for Bi nanowires with different wire diameters are predominantly due to the quantum confinement-induced semimetal-semiconductor transition, which occurs when the wire diameter in Fig. F.8 decreases below 50 nm.

The effect of crystal quality can be accounted for in the same transport model by the value of μ_{boundary} in Eq. (F.37). Instead of a non-monotonic behavior for semimetallic Bi nanowires as shown in Fig. F.8, $R(T)$ is predicted to show a *monotonic* T dependence at a higher defect level. The dashed curve in Fig. F.9 shows the calculated $R(T)/R(300\,\text{K})$ for 70-nm wires with increased boundary scattering ($\mu_{\text{boundary}} \simeq 6\,\text{m}^2\text{V}^{-1}\text{s}^{-1}$), exhibiting a monotonic T dependence, similar to that of Bi nanowires prepared by electrochemical deposition which is likely to produce polycrystalline nanowires. Generally speaking, for samples with many grain boundaries the $1/\tau_{\text{boundary}}$ term is large, leading to a qualitatively different temperature dependence and a lower overall mobility.

The differences in the slopes of the temperature dependence of the low temperature resistance ($T < 10\,\text{K}$) also provide experimental evidence for the semimetal [large nanowire diameter and $(\partial R/\partial T) > 0$] to semiconductor [small nanowire diameter and $(\partial R/\partial T) < 0$]

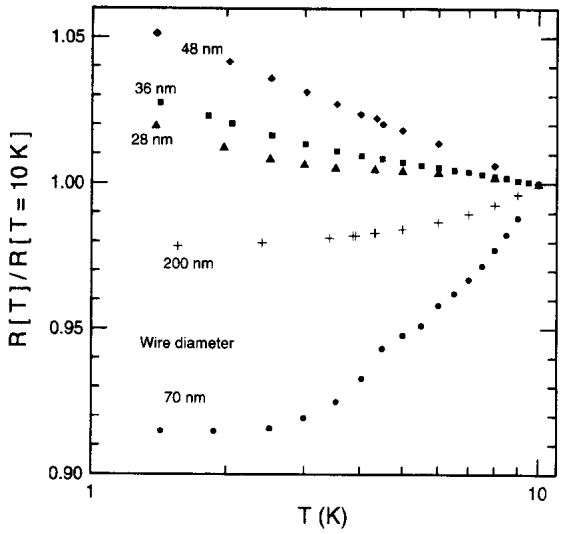


Figure F.10: The temperature dependence of the zero-field resistance for Bi nanowires of different diameters, normalized to the resistance at 10 K.

transition in Bi nanowires, as shown in Fig. F.10.

F.3.8 Magnetoresistance of Bi Nanowires

Because of the inherent one-dimensional geometry of nanowires, certain conventional measurements, such as the Hall effect, which are traditionally carried out to determine the carrier density cannot be performed. Magneto-oscillatory effects cannot be used in many cases to determine the Fermi energy because of wire boundary scattering (which makes it difficult to satisfy $\omega_c\tau \gg 1$), and optical measurements on the Bi/anodic alumina samples to determine the plasma frequency are largely dominated by contributions from the host alumina template, and even single nanowire measurements of the absolute resistivity are quite challenging. Therefore, determining the effects of doping and annealing Bi nanowires often cannot be assessed by conventional means.

Magnetoresistance (MR) measurements provide an informative technique for characterizing Bi nanowires because these measurements yield a great deal of information about electron scattering from wire boundaries, the effects of doping and annealing on scattering, and localization effects in the nanowires.

Figure F.11 shows the longitudinal magnetoresistance (**B** parallel to the wire axis) for 65 nm and 109 nm diameter Bi nanowire samples at 2 K. In the low field regime, the MR increases with B (positive MR), up to some peak value, B_m, beyond which the MR becomes a decreasing function of B (negative MR). This behavior is typical of the longitudinal MR of Bi nanowires in the diameter range 45 nm to 200 nm, and can be understood on the basis of the classical size effect of the nanowire. The MR of wires with diameters smaller than 40 nm shows a strong dependence on B. The peak position B_m moves to lower B field values as the wire diameter increases, as shown in Fig. F.11(b,c) where B_m is seen to vary linearly with $1/d_W$. The application of a longitudinal magnetic field produces helical motion of the electrons along the wire, and above some critical field, approximately B_m,

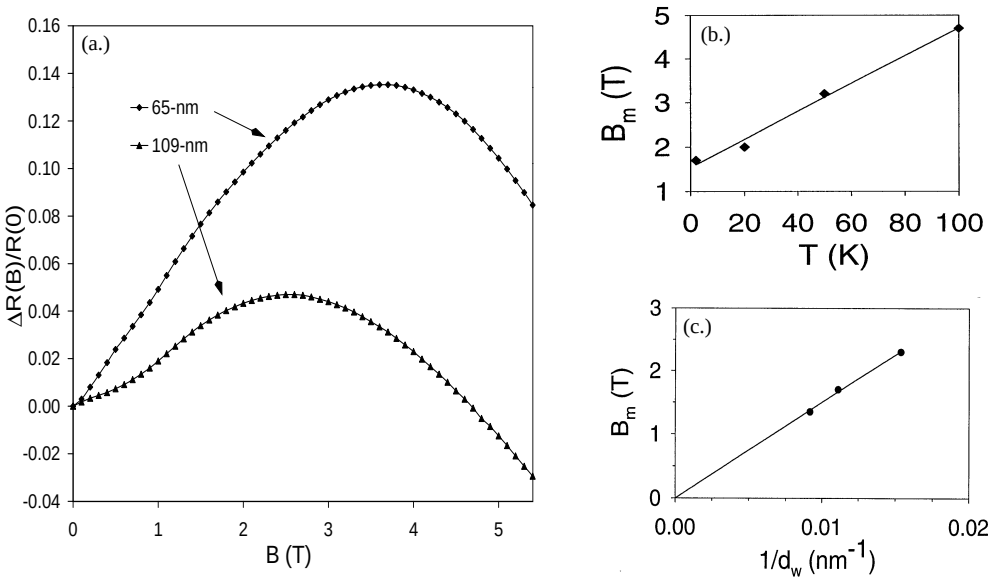


Figure F.11: (a) Longitudinal magnetoresistance, $\Delta R(B)/R(0)$, at 2 K as a function of B for Bi nanowire arrays with diameters 65 and 109 nm before thermal annealing. (b) The peak position B_m as a function of temperature for the 109 nm diameter Bi nanowire array. (c) The peak position B_m of the longitudinal MR at 2 K as a function of $1/d_W$, the reciprocal of the nanowire diameter.

the radius of the helical motion will become smaller than the radius of the wire, causing a decrease in the wire boundary scattering, and giving rise to a negative magnetoresistance $\partial R/\partial B < 0$. However, for low fields $B \leq B_m$, the magnetic field deflects the electrons causing increased scattering with the wire boundary, thereby giving rise to an increase in resistance or a positive magnetoresistance, which is common to most crystalline solids. The condition for B_m is given by $B_m \sim 2c\hbar k_F/ed_W$ where k_F is the wave vector at the Fermi energy. In summary, for $B \leq B_m$ the cyclotron radius is larger than the wire radius and we have a positive MR, while for $B \geq B_m$, the cyclotron radius is smaller than the wire radius, and we have a negative MR. This phenomenon, called the classical size effect for the magnetoresistance, provides much insight into the scattering of electrons in Bi nanowires.

The peak position, B_m, is found to increase linearly with temperature in the range 2 to 100 K, as shown in Fig. F.11(b,c). As T is increased, phonon scattering becomes important and therefore a higher magnetic field is required to reduce the resistivity associated with boundary scattering sufficiently to change the sign of the MR. Likewise increasing the grain boundary scattering also increases the value of B_m at a given T and wire diameter. Application of a transverse magnetic field does not show significant reduction in wire boundary scattering, and therefore the transverse MR is always positive.

Thermal annealing of undoped Bi nanowire samples causes a significant decrease in the magnitude of the magnetoresistance as well as a decrease in the peak position, B_m. This behavior indicates that prior to annealing, the scattering at defects and impurities is dominant over scattering at the wire boundary, even at low temperature (2 K). The observed decrease in MR upon annealing indicates that the Bi nanowires become purer after thermal

treatment, as one would expect.

Bi nanowires doped with Te have been fabricated and characterized, as discussed in §F.3.4. The longitudinal MR of Te-doped samples show no peak in the MR (as can be seen in Fig. F.11 for undoped samples), and instead, the longitudinal MR of Te-doped samples is found to be a monotonically increasing function of magnetic field (positive MR) in the magnetic field range $0 \leq B \leq 5.4\,\mathrm{T}$ at 2 K. The disappearance of the negative MR is attributed to a change in the dominant scattering mechanism from wire boundary scattering (which can be reduced by applying a B field) to magnetic field-independent ionized impurity scattering from the Te dopant ions.

Annealing the Te-doped samples yields MR behavior that is in striking contrast to that of the undoped samples described above. Upon annealing, an *increase* in the MR of the Te-doped samples is observed. This indicates that the dopants are being pushed out of the nanowire to the wire boundary, thereby increasing the role of boundary scattering and decreasing the role of charged impurity scattering.

For Te doped samples with $d_W < 40\,\mathrm{nm}$, the longitudinal MR monotonically increases as B^2 and shows no peak, indicating that B, in the measured range (up to 5.4 T), is too low to reduce the cyclotron radius below the wire radius. Consequently, increasing the magnetic field in this field range always leads to increased boundary scattering.

In addition to the longitudinal magnetoresistance measurements, transverse magnetoresistance measurements (**B** perpendicular to the wire axis) have also been performed on Bi nanowire array samples, where a monotonically increasing B^2 dependence over the entire range $0 \leq B \leq 5.5\,\mathrm{T}$ is found for all Bi nanowires studied thus far. This is as expected, since the wire boundary scattering cannot be reduced by a magnetic field perpendicular to the wire axis. The negative MR observed for the Bi nanowire arrays above B_m shows that wire boundary scattering is a dominant scattering process for the longitudinal magnetoresistance, thereby establishing that the mean free path is larger than the wire diameter and that the Bi nanowires have high crystal quality.

Also encouraging for thermoelectric applications are the results on the high-field classical size effect in the longitudinal magnetoresistance, showing that the defect and impurity levels in the nanowires are sufficiently low so that the wire diameter is comparable to or smaller than the carrier mean free path, and ballistic transport can occur in the nanowires in a high longitudinal magnetic field. The ability of the electronic structure and transport models for Bi nanowires to account for the dependence of the classical size effect in the magnetoresistance on temperature, magnetic field, nanowire diameter and annealing conditions is important for predicting the behavior of Bi nanowires in the smaller diameter range, well below 10 nm, where enhancement in $Z_{1\mathrm{D}}T$ is expected.

By applying a magnetic field, a transition from a 1D localized system, which is characteristic of low magnetic fields, to a 3D localized system can be induced as the magnetic field is increased. The effect of this transition can be seen in Fig. F.12, where the longitudinal magnetoresistance is plotted for Bi nanowire arrays of various nanowire diameters in the range 28 to 70 nm for $T < 5\,\mathrm{K}$. In these curves, a subtle step-like feature is seen at low magnetic fields, and this feature is independent of temperature and of the orientation of the magnetic field, and depends only on wire diameter. The corresponding transverse magnetoresistance curves, also show a step at the same magnetic field strengths. The lack of dependence of the magnetic field of the step on temperature and magnetic field orientation indicates that the phenomenon is not related to the effective masses, which are highly anisotropic in Bi,

but rather is related to the magnetic field length, $L_H = (\hbar/eB)^{1/2}$, which is the spatial extent of the wave function of electrons in the lowest Landau level, and L_H is independent of the effective mass. Setting $L_H(B_c)$ equal to the diameter d_W of the nanowire, defines a critical magnetic field strength, B_c, below which the carrier wavefunction is confined by the nanowire boundary (the 1D localization regime), and above which the wavefunction is confined by the magnetic field (the 3D localization regime). This calculated field strength, B_c, is indicated in Fig. F.12 by vertical lines for the appropriate nanowire diameters, and these calculated B_c values provide a good fit to the step-like like features in the MR curves shown in Fig. F.12. The physical basis for this phenomenon is associated with localization of a single magnetic flux quantum within the nanowire diameter.

The last magnetic field characterization technique discussed in this section is the Shubnikov–de Haas (SdH) quantum oscillatory effect. SdH oscillations, in principle, provide the most direct measurement of the Fermi energy and carrier density for the Bi nanowire system. However, in order to observe SdH oscillations, the magnetic field is applied parallel to the nanowire axis, and the electrons must complete at least one cyclotron orbit without being scattered. Thus the cyclotron radius must be smaller than the wire radius and the mean free path to observe the SdH effect. SdH oscillations occur when the quantized Landau levels pass through the Fermi energy as the magnetic field is increased. By determining the period of the SdH oscillation (periodic in $1/B$), the position of the Fermi energy can be determined by the relation

$$\frac{1}{\Delta(1/B)} = \frac{m_c E_F^e}{\hbar q}\left[1 + E_F^e/E_g\right] \tag{F.38}$$

where $\Delta(1/B)$ denotes the SdH period, m_c is the cyclotron mass, E_F^e is the electron Fermi level and E_g is the L-point gap, where non-parabolic effects are explicitly considered for electrons, while for the holes, the non-parabolic term E_F^h/E_g can be neglected. Figure F.13 shows SdH oscillations reported for an undoped Bi nanowire sample with a 200 nm wire diameter, which was found to be slightly n-type, due to uncontrolled impurities. Also measurements of SdH oscillations were made on Te-doped 200 nm diameter Bi nanowires, also showing two different periods at $14.6\,\mathrm{T}^{-1}$ (identified with light mass electron orbits) and at $19.5\,\mathrm{T}^{-1}$ (identified with heavy mass electron orbits).

F.3.9 Seebeck Coefficient of Bi Nanowires

Thus far, there have been very few measurements on the Seebeck coefficient of Bi nanowires, though the recent achievement of reliable measurements on 200 nm Bi nanowire arrays is encouraging, and corroborates extensive prior studies of the thermoelectric properties of bulk single crystals. Improvement in the measurement technique for the Seebeck coefficient of Bi nanowires is still needed to extend the measurements to the smaller nanowire diameters of interest for possible thermoelectric applications.

The Seebeck coefficient, unlike the resistivity, is intrinsically independent of sample size and the number of nanowires contributing to the signal, because S depends on the ratio of $\mathcal{L}^{(1)}/\mathcal{L}^{(0)}$, on temperature and is expected to depend on wire diameter. The alumina template containing an array of Bi nanowires therefore provides a convenient package for measuring the Seebeck coefficient of Bi nanowires.

Two techniques for measuring the Seebeck coefficient of Bi nanowire arrays are described in the literature, one using a differential thermocouple arrangement, and another, in which

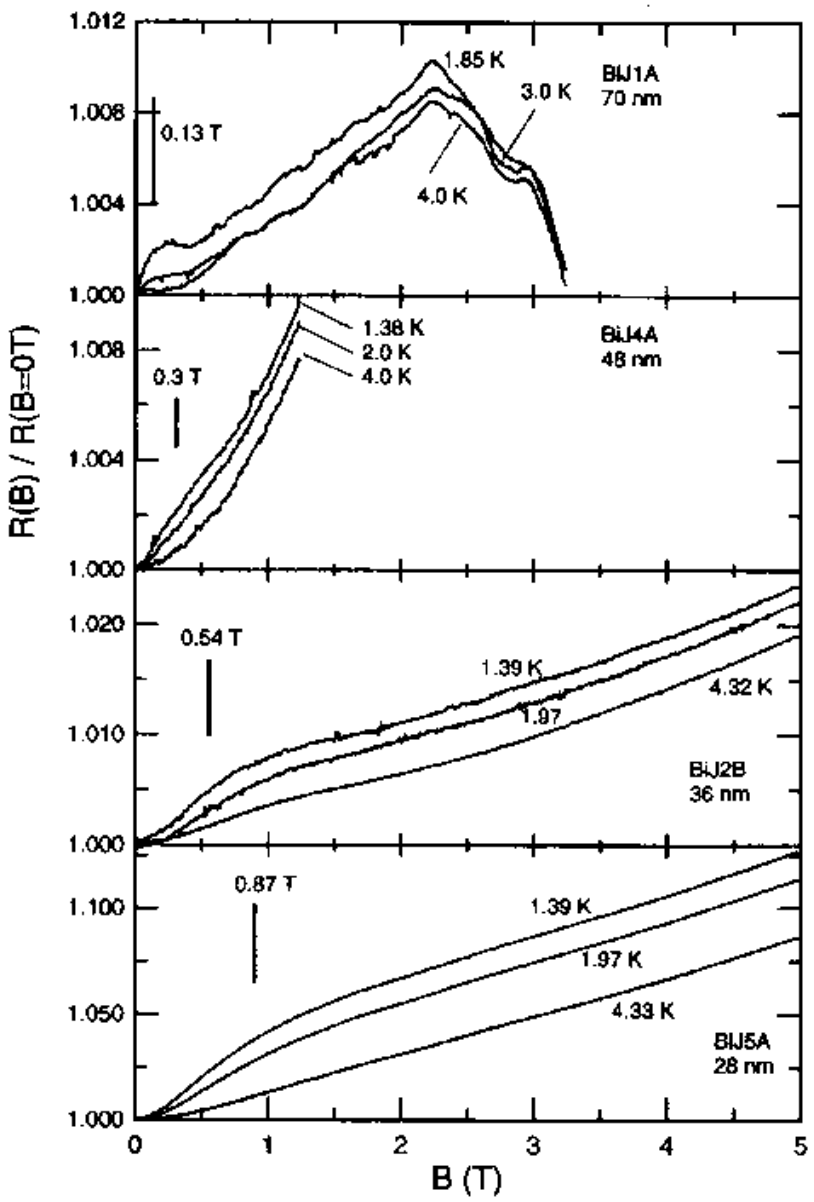


Figure F.12: Longitudinal magnetoresistance as a function of magnetic field for Bi nanowires with 28, 36, 48 and 70 nm diameters. The vertical bars indicate the critical magnetic field B_c at which the magnetic length $L_H = (\hbar/eB)^{1/2}$ equals the nanowire diameter.

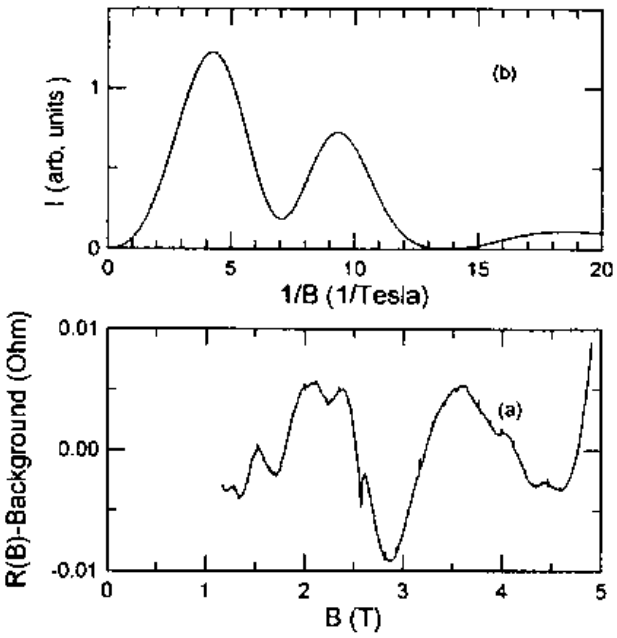


Figure F.13: (a) The oscillatory magnetoresistance for the magnetic field parallel to the nanowire axis of an array of parallel undoped Bi nanowires 200 nm in diameter embedded in an anodic alumina template after the background MR has been subtracted. (b) Fourier transform of the oscillatory part of the magnetoresistance, showing two well-defined SdH periods, the $4.2\,\mathrm{T}^{-1}$ period being identified with the heavy electron cyclotron orbit and another period at $9.25\,\mathrm{T}^{-1}$ identified with T-point holes.

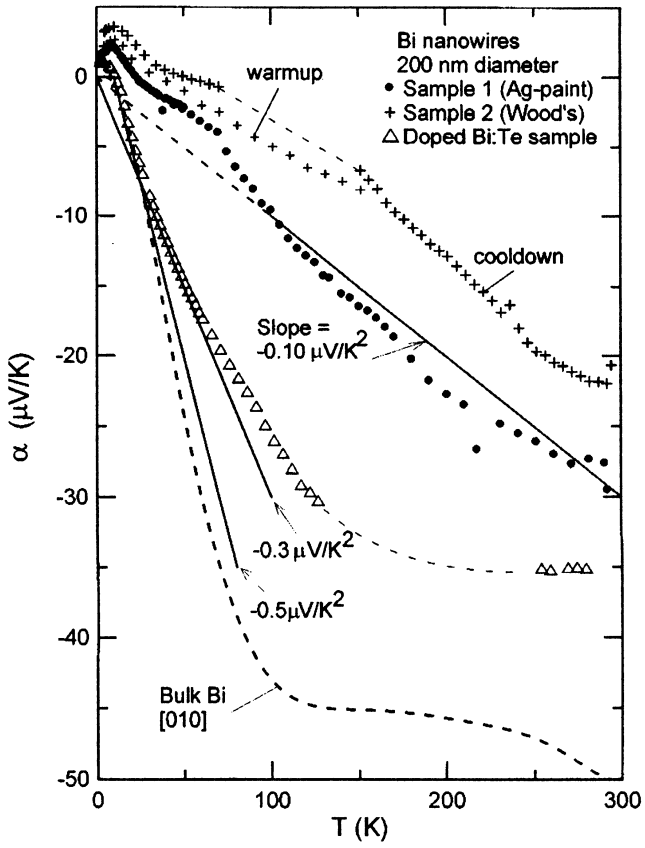


Figure F.14: The temperature dependence of the Seebeck coefficient of two undoped Bi nanowire arrays and one Te-doped Bi nanowire array. All three samples have Bi nanowire diameters of 200 nm. The results for the Seebeck coefficient for the Bi nanowires are compared to that of bulk Bi along the bisectrix direction indicated by the dashed curve.

the thermocouples are mounted in direct electrical contact with the sample to measure the temperature difference, ΔT, across the Bi nanowire sample. The thickness of the thermocouples used in both measurements was 12.5 μm, which is comparable to the 50 μm sample thickness. Because of the relatively large thickness of the thermocouples, the measured ΔT is actually the temperature difference across both the sample and the thermocouples, and therefore, the measured ΔT overestimates the true sample ΔT. Thus, the measured Seebeck coefficient will be a lower limit of the true Seebeck coefficient of the sample. Despite the small thickness of the samples, large ΔT's (in excess of 10 K) are achievable because of the low thermal conductivity of the alumina template ($\sim$1.7 W/mK).

The Seebeck coefficient of bulk Bi is low (-50 to $-100\,\mu$V/K) because of the presence of both electrons and holes, whose contributions to S tend to cancel each other. Figures F.14 and F.15 show the measured Seebeck coefficients of 200 nm Bi nanowires and 40 nm Bi nanowires, respectively. One immediately notices the low magnitudes for these Seebeck coefficients. As mentioned above, the measurement technique underestimates the Seebeck coefficient. For the 200 nm nanowires, the band overlap is not expected to deviate from the

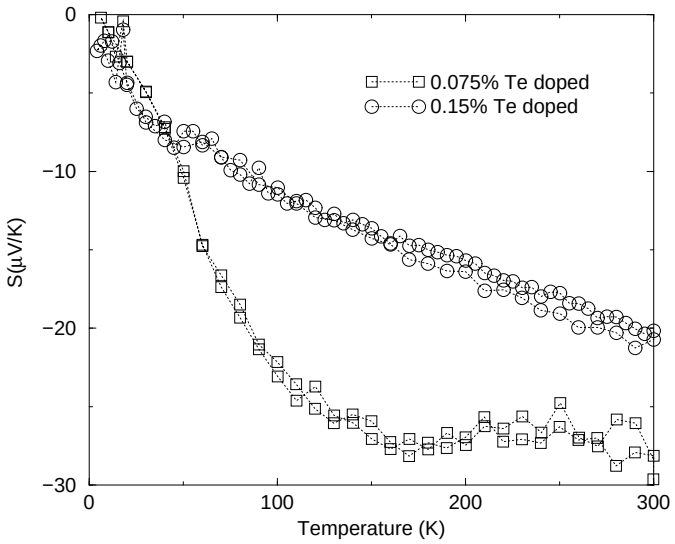


Figure F.15: The temperature dependence of the Seebeck coefficient of two Te-doped Bi nanowire 40 nm diameter arrays with different doping concentrations.

bulk value and therefore 200 nm Bi nanowires contain approximately the same number of holes and electrons as bulk Bi. We thus attribute the low values of S to the same cancellation between electron and hole contributions that occurs in bulk Bi and the underestimate of the measurement technique for measuring S. However for the 40 nm diameter wires, the band structure is expected to be significantly different; in particular, we expect the band overlap between the valence and conduction bands found in bulk Bi to be absent in these nanowires which instead have a small bandgap, appropriate to the semiconducting state, thus potentially providing a higher Seebeck coefficient. However, as shown by the calculations presented in Fig. F.16, S is large when the chemical potential is near the band edge, but S is small when the chemical potential is far from the band edge. The chemical potential can be varied by adding dopants which increase the electron (or hole) concentration. The presence of Te dopants in Bi nanowires moves the chemical potential to lie within the conduction band, thus explaining the low measured values of the Seebeck coefficient. The theoretical optimum $Z_{1D}T$ is predicted to lie close to the conduction band edge (for n-type) as indicated in Fig. F.16. The challenge that now presents itself is how to control the chemical potential ζ sensitively enough to optimize the thermoelectric properties, and how to measure ζ precisely enough for the optimization of $Z_{1D}T$.

¿From Figs. F.14 and F.15 we notice a striking difference in the effect of doping for nanowires of different diameters (40 and 200 nm). Upon the addition of Te dopant, the magnitude of S is *increased* for the 200 nm diameter samples, as shown in Fig. F.14. However, the magnitude of the Seebeck coefficient is *decreased* upon doping of the nanowires with diameter 40 nm, as shown in Fig. F.15. These results appear to be in contrast with each other. However, as we look at the difference in the band structure of these two samples with different diameters, the reasons for this contrasting behavior become clear. For the 200 nm diameter nanowires, there is essentially no shift in the band edge energy relative to bulk bismuth. Therefore, at $T = 77\,\mathrm{K}$ (for example), the nanowires are semimetallic with

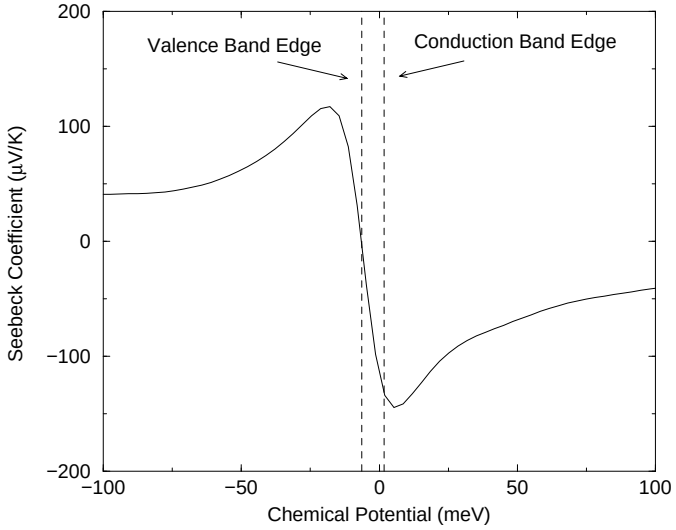


Figure F.16: The calculated Seebeck coefficient for a 40 nm diameter Bi nanowire with transport along the $[01\bar{1}2]$ direction at 77 K, considering non-parabolic dispersion relations for the conduction band and a circular wire cross-section.

a band overlap of 38 meV. S is calculated as a weighted sum of contributions from the hole and electron bands (S_e and S_h), weighted by the hole and electron conductivities (σ_e and σ_h). The addition of Te dopant causes an increase in the electron conductivity and hence an increase in the magnitude of the negative Seebeck coefficient, since electrons dominate over holes in the transport phenomena of Bi nanowires. For the 40 nm diameter nanowires, however, the band edges are shifted appreciably due to quantum confinement effects. The 40 nm nanowire is predicted to be semiconducting at low temperatures, with a calculated bandgap of about 8 meV at 77 K. Figure F.16 shows the calculated Seebeck coefficient for 40 nm diameter Bi nanowires oriented along the $[01\bar{1}2]$ direction (the same orientation as the wires in the sample of Fig. F.15), plotted as a function of the chemical potential. Also indicated on the figure are the conduction and valence band edges. It should be clear from this graph that the S of a lightly doped sample (where the chemical potential is expected to lie just inside the conduction band) is larger in magnitude than the S of a heavily doped sample (where the chemical potential lies further in the conduction band), thus explaining the contrasting behavior that is observed for the 200 nm and 40 nm diameter nanowire samples.

F.4 Summary

In this appendix the predictions of an enhancement in the thermoelectric figure of merit of Bi quantum wires relative to their corresponding bulk counterparts, as well as the present state of experimental confirmation of these predictions are reviewed. Bismuth nanowires, offer significant promise for practical applications, because they can be self assembled and are predicted to have desirable thermoelectric properties when they have wire diameters in the 5–10 nm range. Though temperature-dependent resistance measurements have been

carried out for Bi nanowires in this diameter range, reliable thermoelectric measurements have not yet been reported.

The introduction of low dimensional concepts into the field of thermoelectricity has stimulated new approaches to thinking about better thermoelectric materials, new strategies for achieving higher $Z_{3D}T$, and new applications areas for thermoelectrics, such as thermal management of integrated circuits. It would be fair to say that the introduction of low-dimensional concepts into the thermoelectrics field has injected a large increase in interest and attention to thermoelectric materials and phenomena by the more general scientific community. It is, however, too soon to assess the eventual impact of these low-dimensional concepts on the eventual use of thermoelectricity for practical applications.

Appendix G

Ion Implantation and Rutherford Backscattering Spectroscopy

References:

- S.T. Picraux, *Physics Today*, November (1984), p. 38.
- S.T. Picraux and P.S. Peercy, *Scientific American*, March (1985), p. 102.
- J.W. Mayer, L. Eriksson and J.A. Davies, *Ion Implantation of Semiconductors*, Stanford University Press (1970).
- G. Carter and W.A. Grant, *Ion Implantation of Semiconductors*, Edward Arnold Publishers (1976).
- J.K. Hirvonen, *Ion Implantation, Treatise on Materials Science and Technology*, Vol. 18, Academic Press (1980).
- W.K. Chu, J.W. Mayer and M.A. Nicolet, *Backscattering Spectrometry* (Academic Press, 1978)

G.1 Introduction to the Technique

Ions of all energies incident on a solid influence its materials properties. Ions are used in many different ways in research and technology. We here review some of the physics of the interaction of ion beams with solids and some of the uses of ion beams in semiconductors. Some useful reviews are listed above.

We start by noting that the interaction of ion beams with solids depends on the energy of the incident ion. A directed low energy ion ($\sim$ 10–100 eV) comes to rest at or near the surface of a solid, possibly growing into a registered epitaxial layer upon annealing (Fig. G.1). A 1–keV heavy ion beam is the essential component in the sputtering of surfaces. For this application, a large fraction of the incident energy is transferred to the atoms of the solid resulting in the ejection of surface atoms into the vacuum. The surface is left in a disordered state. Sputtering is used for removal of material from a sample surface on an almost layer–by–layer basis. It is used both in semiconductor device fabrication, ion

etching or ion milling, and more generally in materials analysis, through depth profiling. The sputtered atoms can also be used as a source for the sputter deposition technique discussed in connection with the growth of superlattices.

At higher energies, $\sim$ 100–300 keV, energetic ions are used as a source of atoms to modify the properties of materials. Low concentrations, $\ll 0.1$ atomic percent, of implanted atoms are used to change and control the electrical properties of semiconductors. The implanted atom comes to rest $\sim$ 1000Å below the surface in a region of disorder created by the passage of the implanted ion. The electrical properties of the implanted layer depend on the species and concentration of impurities, the lattice position of the impurities and the amount of lattice disorder that is created.

Lattice site and lattice disorder as well as epitaxial layer formation can readily be analyzed by the channeling of high energy light ions (such as H^+ and He^+), using the Rutherford backscattering technique. In this case MeV ions are used because they penetrate deeply into the crystal (microns) without substantially perturbing the lattice. This is an attractive ion energy regime because scattering cross sections, flux distributions in the crystal, and the rate of energy loss are quantitatively established. Particle–solid interactions in the range from about 0.1 MeV to 5 MeV are understood and one can use this well–characterized tool for investigations of solid state phenomena. Outside of this range many of the concepts discussed below remain valid: however, the use of MeV ions is favored in solid state characterization applications because of both experimental convenience and the ability to probe both surface and bulk properties.

In § 6.2 and § 6.3, we will review the two last energy regimes, namely ion implantation in the 100 keV region and ion beam analysis (Rutherford backscattering and channeling) by MeV light mass particles (H^+, He^+). We start by describing the most important features of ion implantation. Then we will review a two atom collision in order to introduce the basic atomic scattering concepts needed to describe the slowing down of ions in a solid. Then we will state the results of the LSS theory (J. Lindhard, M. Scharff and H. Schiøtt, *Mat. Fys. Medd. Dan. Vid. Selsk.* 33, No. 14 (1963)) which is the most successful basic theory for describing the distribution of ion positions and the radiation–induced disorder in the implanted sample. A number of improvements to this model have been made in the last two decades and are used for current applications. We will conclude this introduction to ion implantation with a discussion of the lattice damage caused by the slowing down of the energetic ions in the solid.

We then go on and describe the use of a beam of energetic (1–2 MeV) light mass particles (H^+, He^+) to study material properties in the near surface region ($< \mu$m) of a solid. We will then see how to use Rutherford backscattering spectrometry (RBS) to determine the stoichiometry of a sample composed of multiple chemical species and the depth distribution of implanted ions. Furthermore, we see that with single–crystal targets, the effect of channeling also allows investigation of the crystalline perfection of the sample as well as the lattice location of the implanted atomic species. Finally, we will review some examples of the modification of material properties by ion implantation, including the use of ion beam analysis in the study of these materials modifications.

G.2 Ion Implantation

Ion implantation is an important technique for introducing impurity atoms in a controlled way, thus leading to the synthesis of new classes of materials, including metastable materials. The technique is important in the semiconductor industry for making p–n junctions by, for example, implanting n–type impurities into a p–type host material. From a materials science point of view, ion implantation allows essentially any element of the periodic table to be introduced into the near surface region of essentially any host material, with quantitative control over the depth and composition profile by proper choice of ion energy and fluence. More generally, through ion implantation, materials with increased strength and corrosion resistance or other desirable properties can be synthesized.

A schematic diagram of an ion implanter is shown in Fig. G.2. In this diagram the "target" is the sample that is being implanted. In the implantation process, ions of energy E and beam current i_b are incident on a sample surface and come to rest at some characteristic distance R_p with a Gaussian distribution of half width at half maximum ΔR_p. Typical values of the implantation parameters are: ion energies $E \sim 100$ keV, beam currents $i_b \sim 50$ μA, penetration depths $R_p \sim 1000$ Å and half-widths $\Delta R_p \sim 300$ Å (see Fig. G.3). In the implanted regions, implant concentrations of 10^{-3} to 10^{-5} relative to the host materials are typical. In special cases, local concentrations as high as 20 at.% (atomic percent) of implants have been achieved.

Some characteristic features of ion implantation are the following:

1. The ions characteristically only penetrate the host material to a depth $R_p \leq 1\mu$m. Thus ion implantation is a near surface phenomena. To achieve a large percentage of impurity ions in the host, the host material must be thin (comparable to R_p), and high fluences of implants must be used ($\phi > 10^{16}/\text{cm}^2$).

2. The depth profile of the implanted ions (R_p) is controlled by the ion energy. The impurity content is controlled by the ion fluence ϕ.

3. Implantation is a non–equilibrium process. Therefore there are no solubility limits on the introduction of dopants. With ion implantation one can thus introduce high concentrations of dopants, exceeding the normal solubility limits. For this reason ion implantation permits the synthesis of metastable materials.

4. The implantation process is highly directional with little lateral spread. Thus it is possible to implant materials according to prescribed patterns using masks. Implantation proceeds in the regions where the masks are not present. An application of this technology is to the ion implantation of polymers to make photoresists with sharp boundaries. Both positive and negative photoresists can be prepared using ion implantation, depending on the choice of the polymer. These masks are widely used in the semiconductor industry.

5. The diffusion process is commonly used for the introduction of impurities into semiconductors. Efficient diffusion occurs at high temperatures. With ion implantation, impurities can be introduced at much lower temperatures, as for example room temperature, which is a major convenience to the semiconductor industry.

6. The versatility of ion implantation is another important characteristic. With the same implanter, a large number of different implants can be introduced by merely changing the ion source. The technique is readily automated, and thus is amenable for use by technicians in the semiconductor industry. The implanted atoms are introduced in an atomically dispersed fashion, which is also desirable. Furthermore, no oxide or interfacial barriers are formed in the implantation process.

7. The maximum concentration of ions that can be introduced is limited. As the implantation process proceeds, the incident ions participate in both implantation into the bulk and the sputtering of atoms off the surface. Sputtering occurs because the surface atoms receive sufficient energy to escape from the surface during the collision process. The dynamic equilibrium between the sputtering and implantation processes limits the maximum concentration of the implanted species that can be achieved. Sputtering causes the surface to recede slowly during implantation.

8. Implantation causes radiation damage. For many applications, this radiation damage is undesirable. To reduce the radiation damage, the implantation can be carried out at elevated temperatures or the materials can be annealed after implantation. In practice, the elevated temperatures used for implantation or for post–implantation annealing are much lower than typical temperatures used for the diffusion of impurities into semiconductors.

A variety of techniques are used to characterize the implanted alloy. Ion backscattering of light ions at higher energies (e.g., 2 MeV He^+) is used to determine the composition versus depth with $\sim$ 10nm depth resolution. Depth profiling by sputtering in combination with Auger or secondary ion mass spectroscopy is also used. Lateral resolution is provided by analytical transmission electron microscopy. Electron microscopy, glancing angle x–ray analysis and ion channeling in single crystals provide detailed information on the local atomic structure of the alloys formed. Ion backscattering and channeling are discussed in Section 6.3.

G.2.1 Basic Scattering Equations

An ion penetrating into a solid will lose its energy through the Coulomb interaction with the atoms in the target. This energy loss will determine the final penetration of the projectile into the solid and the amount of disorder created in the lattice of the sample. When we look more closely into one of these collisions (Fig. G.4) we see that it is a very complicated event in which :

- The two nuclei with masses M_1 and M_2 and charges Z_1 and Z_2, respectively, repel each other by a Coulomb interaction, screened by the respective electron clouds.

- Each electron is attracted by the two nuclei (again with the corresponding screening) and is repelled by all other electrons.

In addition, the target atom is bonded to its neighbors through bonds which usually involve its valence electrons. The collision is thus a many–body event described by a complicated Hamiltonian. However experience has shown that very accurate solutions for the trajectory of the nuclei can be obtained by making some simplifying assumptions. The

most important assumption for us, is that the interaction between the two atoms can be separated into two components:

- ion (projectile)–nucleus (target) interaction
- ion (projectile)–electron (target) interaction.

Let us now use the very simple example of a collision between two masses M_1 and M_2 to determine the relative importance of these two processes and to introduce the basic atomic scattering concepts required to describe the stopping of ions in a solid. Figure G.5 shows the classical collision between the incident mass M_1 and the target mass M_2 which can be a target atom or a nearly free target electron.

By applying conservation of energy and momentum, the following relations can be derived (you will do it as homework).

- The energy transferred (T) (Ref. H. Goldstein, *Classical Mechanics*, Academic Press (1950)) in the collision from the incident projectile M_1 to the target particle M_2 is given by

$$T = T_{max} \quad \sin^2\left(\frac{\Theta}{2}\right) \tag{G.1}$$

  where

$$T_{max} = 4E_0 \quad \frac{M_1 M_2}{(M_1 + M_2)^2} \tag{G.2}$$

  is the maximum possible energy transfer from M_1 to M_2.

- The scattering angle of the projectile in the laboratory system of coordinates is given by

$$\cos\theta = \frac{1 - (1 + M_2/M_1)(T/2E)}{\sqrt{(1 - T/E)}} \tag{G.3}$$

- The energies of the projectile before (E_0) and after (E_1) scattering are related by

$$E_1 = k^2 E_0 \tag{G.4}$$

  where the kinematic factor k is given by

$$k = \left(\frac{M_1 \cos\theta \pm ({M_2}^2 - {M_1}^2 \sin^2\theta)^{1/2}}{M_1 + M_2} \right) \tag{G.5}$$

These relations are absolutely general no matter how complex the force between the two particles, so long as the force acts along the line joining the particles and the electron is nearly free so that the collision can be taken to be elastic. In reality, the collisions between the projectile and the target electrons are inelastic because of the binding energy of the electrons. The case of inelastic collisions will be considered in what follows.

Using Eqs. (G.2) and (G.3) and assuming either that M_2 is of the same atomic species as the projectile ($M_2 = M_1$), or that M_2 is a nearly free electron ($M_2 = m_e$) we construct Table G.1.

With the help of the previous example, we can formulate a qualitative picture of the slowing down process of the incident energetic ions. As the incident ions penetrate into the solid, they lose energy. There are two dominant mechanisms for this energy loss:

Table G.1: Maximum energy transfer T_{max} and scattering angle θ for nuclear ($M_2 = M_1$) and electronic ($M_2 = m_e$) collisions.

	"nuclear" collision	"electronic" collision
T_{max}	$T_{max} \simeq E_0$	$T_{max} \simeq (4m/M_1)\,E_0$
θ	$0 < \theta \leq \pi/2$	$\theta = 0^\circ$

1. The interaction between the incident ion and the electrons of the host material. This inelastic scattering process gives rise to electronic energy loss.

2. The interaction between the incident ions and the nuclei of the host material. This is an elastic scattering process which gives rise to nuclear energy loss.

The ion–electron interaction (see Table G.1) induces small losses in the energy of the incoming ion as the electrons in the atom are excited to higher bound states or are ionized. These interactions do not produce significant deviations in the projectile trajectory. In contrast, the ion–nucleus interaction results in both energy loss and significant deviation in the projectile trajectory. In the ion–nucleus interaction, the atoms of the host are also significantly dislodged from their original positions giving rise to lattice defects, and the deviations in the projectile trajectory will give rise to the lateral spread of the distribution of implanted species.

Let us now further develop our example of a "nuclear" collision. For a given interaction potential $V(r)$, each ion coming into the annular ring of area $2\pi pdp$ with energy E, will be deflected through an angle θ where p is the impact parameter (see Fig. G.6). We define $T = E_0 - E_1$ as the energy transfer from the incoming ion to the host and we define $2\pi pdp = d\sigma$ as the differential cross section. When the ion moves a distance Δx in the host material, it will interact with $N\Delta x 2\pi pdp$ atoms where N is the atom density of the host.

The energy ΔE lost by an ion traversing a distance Δx will be

$$\Delta E = N\Delta x \int T\, 2\pi pdp \tag{G.6}$$

so that as $\Delta x \to 0$, we have for the stopping power

$$\frac{dE}{dx} = N \int T d\sigma \tag{G.7}$$

where σ denotes the cross sectional area. We thus obtain for the stopping cross section E

$$E = \frac{1}{N}\frac{dE}{dx} = \int T d\sigma. \tag{G.8}$$

The total stopping power is due to both electronic and nuclear processes

$$\frac{dE}{dx} = \left(\frac{dE}{dx}\right)_e + \left(\frac{dE}{dx}\right)_n = N(E_e + E_n) \tag{G.9}$$

where N is the target density and E_e and E_n are the electronic and nuclear stopping cross sections, respectively. Likewise for the stopping cross section E we can write

$$E = E_e + E_n. \tag{G.10}$$

Table G.2: Typical Values of E_1, E_2, E_3 for silicon.[a]

Ion	E_1(keV)	E_2(keV)	E_3(keV)
B	3	17	3000
P	17	140	$\sim 3 \times 10^4$
As	73	800	$> 10^5$
Sb	180	2000	$> 10^5$

[a]See Fig. G.7 for the definition of the notation.

From the energy loss we can obtain the ion range or penetration depth

$$R = \int \frac{dE}{dE/dx}. \tag{G.11}$$

Since we know the energy transferred to the lattice (including both phonon generation and displacements of the host ions), we can calculate the energy of the incoming ions as a function of distance into the medium $E(x)$.

At low energies of the projectile ion, nuclear stopping is dominant, while electron stopping dominates at high energies as shown in the characteristic stopping power curves of Figs. G.7 and G.8. Note the three important energy parameters on the curves shown in Fig. G.7: E_1 is the energy where the nuclear stopping power is a maximum, E_3 where the electronic stopping power is a maximum, and E_2 where the electronic and nuclear stopping powers are equal. As the atomic number of the ion increases for a fixed target, the scale of E_1, E_2 and E_3 increases. Also indicated on the diagram is the functional form of the energy dependence of the stopping power in several of the regimes of interest. Typical values of the parameters E_1, E_2 and E_3 for various ions in silicon are given in Table G.2. Ion implantation in semiconductors is usually done in the regime where nuclear energy loss is dominant. The region in Fig. G.7 where $(dE/dx) \sim 1/E$ corresponds to the regime where light ions like H^+ and He^+ have incident energies of 1–2 MeV and is therefore the region of interest for Rutherford backscattering and channeling phenomena.

G.2.2 Radiation Damage

The energy transferred from the projectile ion to the target atom is usually sufficient to result in the breaking of a chemical bond and the permanent displacement of the target atom from its original site (see Fig. G.9). The condition for this process is that the energy transfer per collision T is greater than the binding energy E_d.

Because of the high incident energy of the projectile ions, each incident ion can dislodge multiple host ions. The damage profile for low dose implantation gives rise to isolated regions of damage as shown in Fig. G.10.

As the fluence is increased, these damaged regions coalesce as shown in Fig. G.10. The damage profile also depends on the mass of the projectile ion, with heavy mass ions of a given energy causing more local lattice damage as the ions come to rest. Since $(dE/dx)_n$ increases as the energy decreases, more damage is caused as the ions are slowed down and come to rest. The damage pattern is shown in Fig. G.11 schematically for light ions (such as boron in silicon) and for heavy ions (such as antimony in silicon). Damage is caused both by the incident ions and by the displaced energetic (knock-on) ions.

A schematic diagram of the types of defects caused by ion implantation is shown in Fig. G.12. Here we see the formation of vacancies and interstitials, Frenkel pairs (the pair formed by the Coulomb attraction of a vacancy and an interstitial). The formation of multiple vacancies leads to a depleted zone while multiple interstitials lead to ion crowding.

G.2.3 Applications of Ion Implantation

For the case of semiconductors, ion implantation is dominantly used for doping purposes, to create sharp p-n junctions in the near-surface region. To reduce radiation damage, implantation is sometimes done at elevated temperatures. Post implantation annealing is also used to reduce radiation damage, with elevated temperatures provided by furnaces, lasers or flash lamps. The ion implanted samples are characterized by a variety of experimental techniques for the implant depth profile, the lattice location of the implant, the residual lattice disorder subsequent to implantation and annealing, the electrical properties (Hall effect and conductivity) and the device performance.

A major limitation of ion implantation for modifying metal surfaces has been the shallow depth of implantation. In addition, the sputtering of atoms from the surface sets a maximum concentration of elements which can be added to a solid, typically $\sim$ 20 to 40 at.%. To form thicker layers and higher concentrations, combined processes involving ion implantation and film deposition are being investigated. Intense ion beams are directed at the solid to bring about alloying while other elements are simultaneously brought to the surface, for example by sputter deposition, vapor deposition or the introduction of reactive gases. One process of interest is ion beam mixing, where thin films are deposited onto the surface first and then bombarded with ions. The dense collision cascades of the ions induce atomic–scale mixing between elements. Ion beam mixing is also a valuable tool to study metastable phase formation.

With regard to polymers, ion implantation can enhance the electrical conductivity by many orders of magnitude, as is for example observed (see Fig. G.13) for ion implanted polyacrylonitrile (PAN, a graphite fiber precursor). Some of the attendant property changes of polymers due to ion implantation include cross-linking and scission of polymer chains, gas evolution as volatile species are released from polymer chains and free radical formation when vacancies or interstitials are formed. Implantation produces solubility changes in polymers and therefore can be utilized for the patterning of resists for semiconductor mask applications. For the positive resists, implantation enhances the solubility, while for negative resists, the solubility is reduced. The high spatial resolution of the ion beams makes ion beam lithography a promising technique for sub–micron patterning applications. For selected polymers (such as PAN), implantation can result in transforming a good insulator into a conducting material with an increase in conductivity by more than 10 orders of magnitude upon irradiation. Thermoelectric power measurements on various implanted polymers show that implantation can yield either p–type or n–type conductors and in fact a p–n junction has recently been made in a polymer through ion implantation (T. Wada, A. Takeno, M. Iwake, H. Sasabe, and Y. Kobayashi, *J. Chem. Soc. Chem. Commun.*, **17**, 1194 (1985)). The temperature dependence of the conductivity for many implanted polymers is of the form $\sigma = \sigma_o \exp(T_0/T)^{1/2}$ which is also the relation characteristic of the one–dimensional hopping conductivity model for disordered materials. Also of interest is

the long term chemical stability of implanted polymers.

Due to recent developments of high brightness ion sources, focused ion beams to submicron dimensions can now be routinely produced, using ions from a liquid metal source. Potential applications of this technology are to ion beam lithography, including the possibility of maskless implantation doping of semiconductors. Instruments based on these ideas may be developed in the future. The applications of ion implantation represent a rapidly growing field.

Instruments based on these ideas may be developed in the future. The applications of ion implantation represent a rapidly growing field.

Further discussion of the application of ion implantation to the preparation of metastable materials is presented after the following sections on the characterization of ion implanted materials by ion backscattering and channeling.

G.3 Ion Backscattering

In Rutherford backscattering spectrometry (RBS), a beam of mono-energetic (1–2 MeV), collimated light mass ions (H^+, He^+) impinges (usually at near normal incidence) on a target and the number and energy of the particles that are scattered backwards at a certain angle θ are monitored (as shown in Fig. G.14) to obtain information about the composition of the target (host species and impurities) as a function of depth. With the help of Fig. G.15 we will review the fundamentals of the RBS analysis.

Particles scattered at the surface of the target will have the highest energy E upon detection. Here the energy of the backscattered ions E is given by the relation

$$E = k^2 E_0 \tag{G.12}$$

where

$$k = \left(\frac{M_1 \cos\theta \pm \left({M_2}^2 - {M_1}^2 \sin^2\theta \right)^{1/2}}{M_1 + M_2} \right) \tag{G.13}$$

as discussed in Section G.2. For a given mass species, the energy E_s of particles scattered from the surface corresponds to the edge of the spectrum (see Fig. G.15). In addition, the scattered energy depends through k on the mass of the scattering atom. Thus different species will appear displaced on the energy scale of Fig. G.15, thereby allowing for their chemical identification. We next show that the displacement along the energy scale from the surface contribution gives information about the depth where the backscattering took place. Thus the energy scale is effectively a depth scale.

The height H of the RBS spectrum corresponds to the number of detected particles in each energy channel ΔE.

G.4 Channeling

If the probing beam is aligned nearly parallel to a close-packed row of atoms in a single crystal target, the particles in the beam will be steered by the potential field of the rows of atoms, resulting in an undulatory motion in which the "channeled" ions will not approach the atoms in the row to closer than 0.1–0.2 Å. This is called the channeling effect (D.V.

Morgan, *Channeling* (Wiley, 1973)). Under this channeling condition, the probability of large angle scattering is greatly reduced. As a consequence, there will be a drastic reduction in the scattering yield from a channeled probing beam relative to the yield from a beam incident in a random direction (see Fig. G.16). Two characteristic parameters for channeling are the normalized minimum yield $\chi_{min} = H_A/H$ which is a measure of the crystallinity of the target, and the critical angle for channeling $\psi_{1/2}$ (the halfwidth at half maximum intensity of the channeling resonance) which determines the degree of alignment required to observed the channeling effect.

The RBS–channeling technique is frequently used to study radiation–induced lattice disorder by measuring the fraction of atom sites where the channel is blocked.

In general, the channeled ions are steered by the rows of atoms in the crystal. However if some portion of the crystal is disordered and lattice atoms are displaced so that they partially block the channels, the ions directed along nominal channeling directions can now have a close collision with these displaced atoms, so that the resulting scattered yield will be increased above that for an undisturbed channel. Furthermore, since the displaced atoms are of equal mass to those of the surrounding lattice, the increase in the yield occurs at a position in the yield *vs.* energy spectrum corresponding to the depth at which the displaced atoms are located. The increase in the backscattering yield from a given depth will depend upon the number of displaced atoms, so the depth (or equivalently, the backscattering energy E) dependence of the yield, reflects the depth dependence of displaced atoms, and integrations over the whole spectrum will give a measure of the total number of displaced atoms. This effect is shown schematically in Fig. G.17.

Another very useful application of the RBS–channeling technique is in the determination of the location of foreign atoms in a host lattice. Since channeled ions cannot approach the rows of atoms which form the channel closer than ~ 0.1Å, we can think of a "forbidden region", as a cylindrical region along each row of atoms with radius ~ 0.1Å, such that there are no collisions between the channeled particles and atoms located within the forbidden zone. In particular, if an impurity is located in a forbidden region it will not be detected by the channeled probing beam. On the other hand, any target particle can be detected by (i.e., will scatter off) probing particles from a beam which impinges in a random direction. Thus, by comparing the impurity peak observed for channeling and random alignments, the fraction of impurities sitting in the forbidden region of a particular channel (high symmetry crystallographic axis) can be determined. Repeating the procedure for other crystallographic directions allows the identification of the lattice location of the impurity atom in many cases.

Rutherford backscattering will not always reveal impurities embedded in a host matrix, in particular if the mass of the impurities is smaller than the mass of the host atoms. In such cases, ion induced x–rays and ion induced nuclear reactions are used as signatures for the presence of the impurities inside the crystal, and the lattice location is derived from the changes in yield of these processes for random and channeled impingement of the probing beam.

Ion markers consist of a very thin layer of a guest atomic species are embedded in an otherwise uniform host material of a different species to establish reference distances. Backscattering spectra are taken before and after introduction of the marker. The RBS spectrum taken after insertion of the marker can be used as a reference for various applications. Some examples where marker references are useful include:

- Estimation of surface sputtering by ion implantation. In this case, recession of the surface from the reference position set by the marker (see Fig. G.18) can be measured by RBS and can be analyzed to yield the implantation–induced surface sputtering.
- Estimation of surface material vaporized through laser annealing, rapid thermal annealing or laser melting of a surface.
- Estimation of the extent of ion beam mixing.

Figure G.1: Schematic illustrating the interactions of ion beams with a single–crystal solid. Directed beams of $\sim$ 10 eV are used for film deposition and epitaxial formation. Ion beams of energy $\sim$ 1 keV are employed in sputtering applications; $\sim$ 100 keV ions are used in ion implantation. Both the sputtering and implantation processes damage and disorder the crystal. Higher energy light ions are used for ion beam analysis.

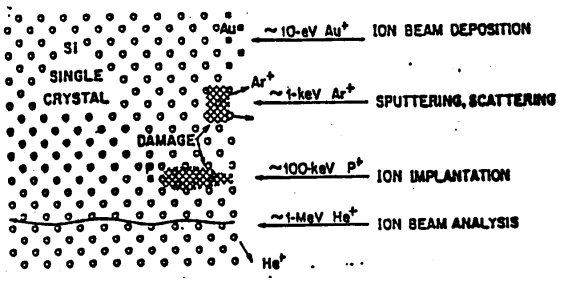
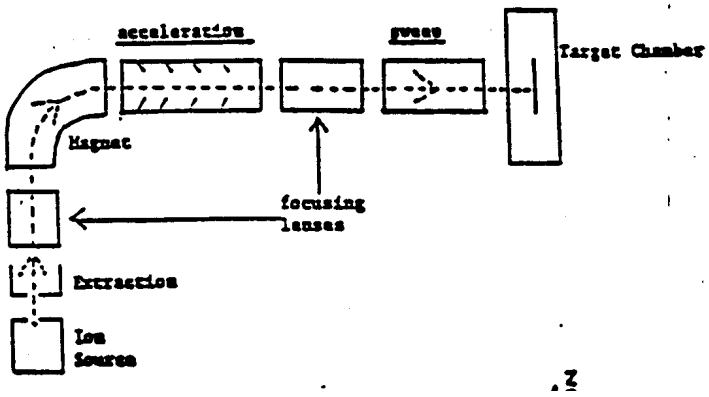


Figure G.2: Schematic Diagram of Ion Implanter.



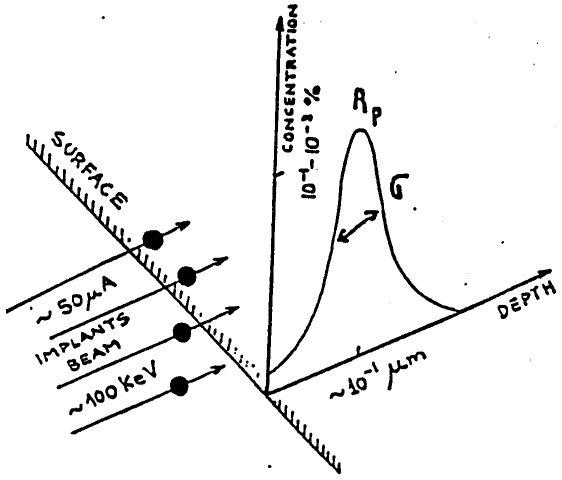


Figure G.3: Typical ion implantation parameters.

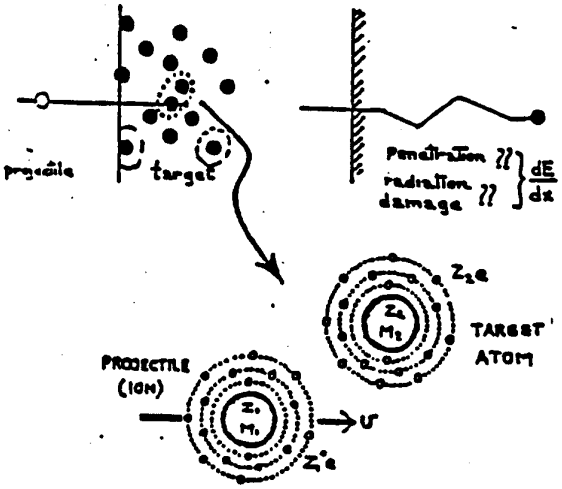


Figure G.4: Penetration of ions into solids.

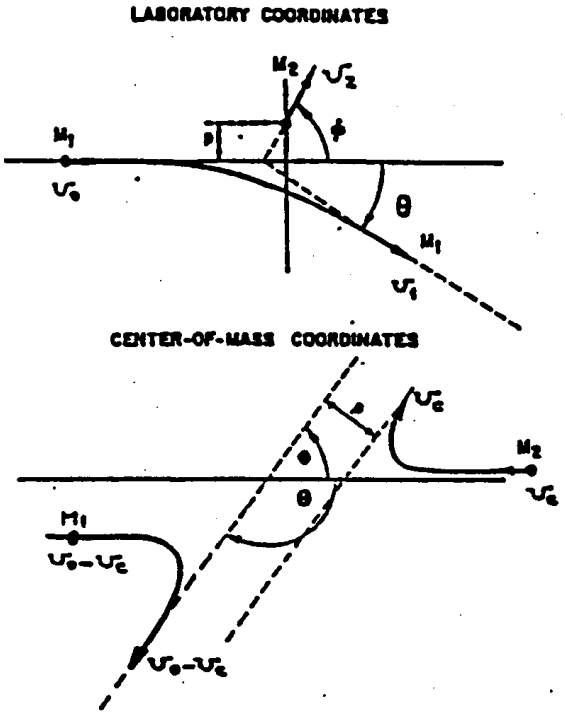


Figure G.5: The upper figure defines the scattering variables in a two–body collision. The projectile has mass M_1 and an initial velocity v_0, and an impact parameter, p, with the target particle. The projectile's final angle of deflection is θ and its final velocity is v_1. The target particle with mass, M_2, recoils at an angle ϕ with velocity v_2. The lower figure is the same scattering event in the center–of–mass (CM) coordinates in which the *total momentum of the system is zero.* The coordinate system moves with velocity v_c relative to the laboratory coordinates, and the angles of scatter and recoil are Θ and Φ in the center of mass system.

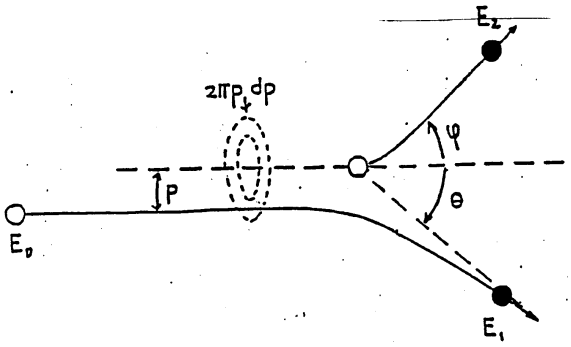


Figure G.6: Classical model for the collision of a projectile of energy E with a target at rest. The open circles denote the initial state of the projectile and target atoms, and the full circles denote the two atoms after the collision.

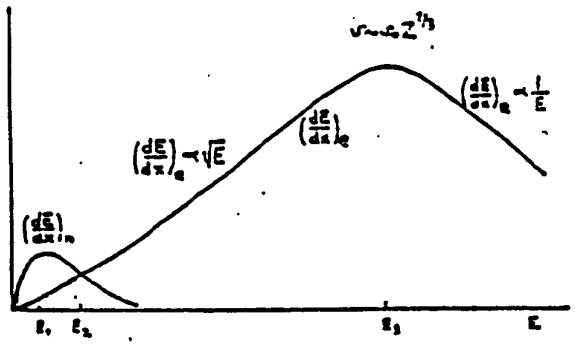


Figure G.7: Nuclear and electronic energy loss (stopping power) *vs.* Energy.

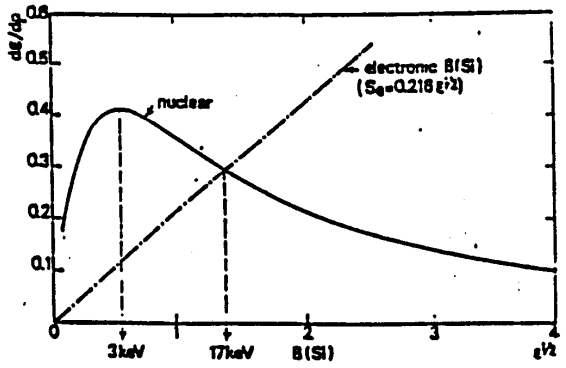


Figure G.8: Nuclear and electronic energy loss (stopping power) *vs.* $\epsilon^{1/2}$ (reduced units of LSS theory).

Figure G.9: Energy transfer from projectile ion to target atoms for a single scattering event for the condition $T > E_d$ where E_d is the binding energy and T is the energy transfer per collision.

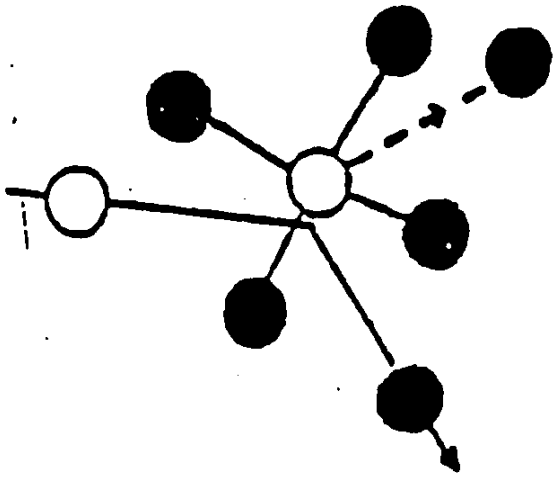
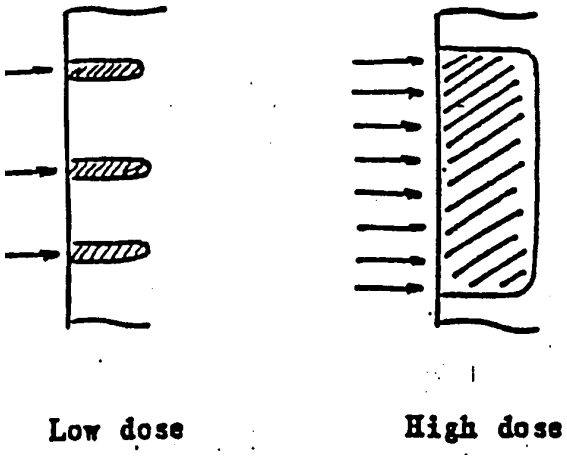


Figure G.10: Schematic diagram showing the range of lattice damage for low dose and high dose implants.



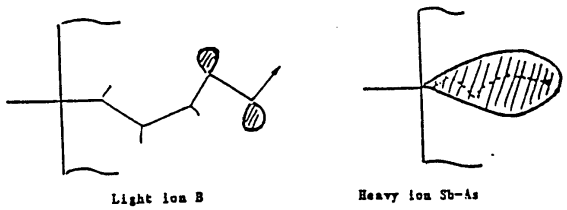


Figure G.11: Schematic diagram of the damage pattern for light ions and heavy ions in the same target (silicon).

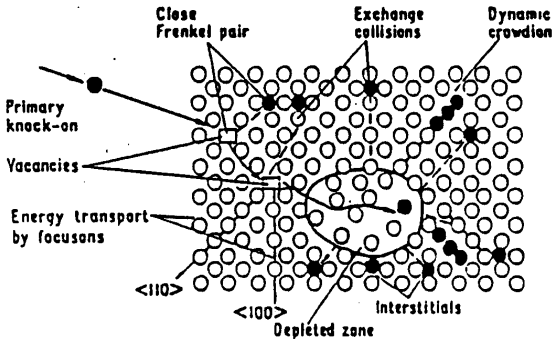
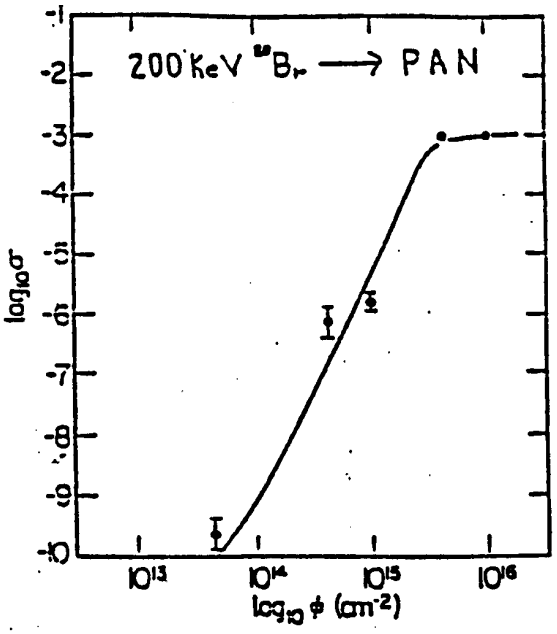


Figure G.12: Example of typical defects induced by ion implantation.

Figure G.13: Implantation induced conductivity of a normally insulating polymer.



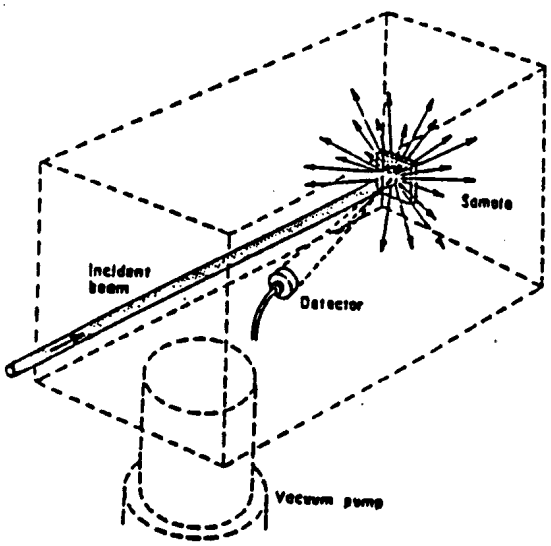


Figure G.14: In the RBS experiment, the scattering chamber where the analysis/experiment is actually performed contains the essential elements: the sample, the beam, the detector, and the vacuum pump.

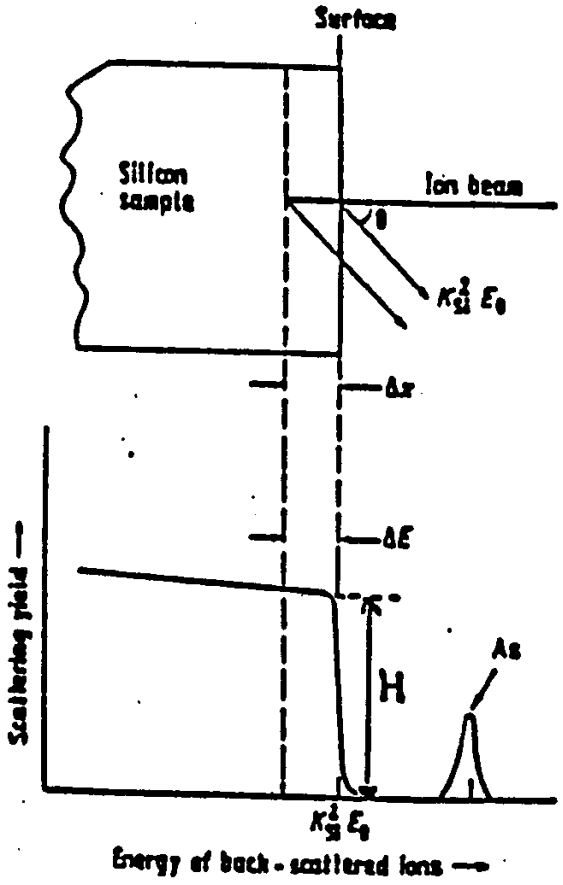


Figure G.15: Schematic diagram showing the energy distribution of ions back-scattered from a Si sample (not aligned) which was implanted with As atoms.

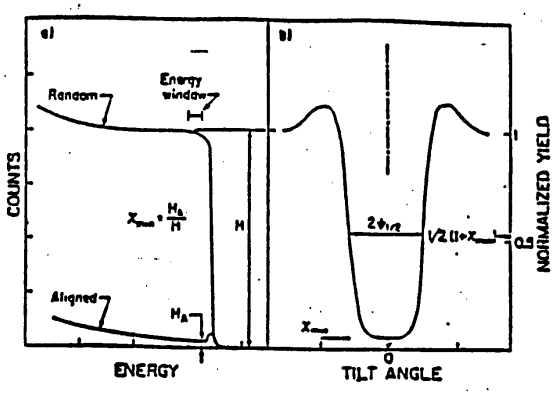
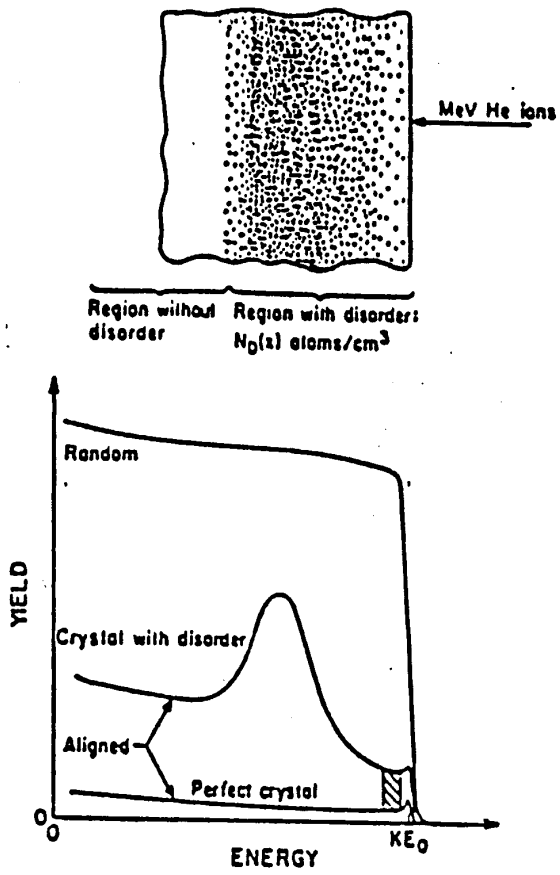


Figure G.16: Schematic backscattering spectrum and angular yield profile.

Figure G.17: Schematic random and aligned spectra for MeV ^{4}He ions incident on a crystal containing disorder. The aligned spectrum for a perfect crystal without disorder is shown for comparison. The difference (shaded portion) in the aligned spectra between disordered and perfect crystals can be used to determine the concentration $N_D(0)$ of displaced atoms at the surface.



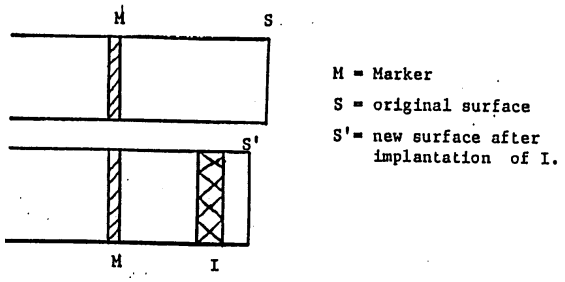


Figure G.18: Schematic of the marker experiment which demonstrates surface recession through surface sputtering.