Part One
Novel Theoretical, Computational, and Experimental Methods and Techniques

Quantum Biochemistry. Edited by Chérif F. Matta

ISBN: 978-3-527-32322-7

1
Quantum Kernels and Quantum Crystallography: Applications in Biochemistry

Lulu Huang, Lou Massa, and Jerome Karle

1.1
Introduction

Professors Bernard and Alberte Pullman were among the first and most important researchers to apply the notions of quantum mechanics to a great number of molecules of biological importance. It has been often noted that their early work was the beginning of quantum biochemistry, pioneering as they did the application of quantum mechanics to carcinogenic properties of aromatic hydrocarbons. Their quantum computations included the electronic structure of nucleic acids and their mechanisms interacting with various drugs, carcinogens and antitumor compounds. They had success in the interpretation of the role of enzyme constituents important in redox reactions, in calculating stability to ultraviolet radiation, in evaluating the role of functional molecular portions (as opposed to whole molecules) in carcinogen action, and in the evaluation of hydrogen bonding through the amino acid residues as potential pathways for electron transfer. Their landmark book entitled *Quantum Biochemistry* [B. Pullman and A. Pullman, Interscience Publishers (John Wiley & Sons), New York, 1963] has been an inspiration for workers in the research field of the same name as the book title. Their success in quantum biology is all the more impressive today in consideration of the computational difficulty of solving the Schrödinger equation in their time. In this chapter we discuss, the origin of our work in the topic title of this chapter, and certain numerical results of quantum biochemistry made possible since the time of the Pullman's by the enormous increase in computing power that has occurred. Remarkable advances in computing have facilitated the treatment of ever increasing molecular size in both crystallography and quantum mechanics.

Quantum Biochemistry. Edited by Chérif F. Matta

ISBN: 978-3-527-32322-7

1.2
Origins of Quantum Crystallography (QCr)

1.2.1
General Problem of *N*-Representability

The origins of our work in the field we named quantum crystallography go back to the ideas that originated in the laboratory of Professor William Clinton of the Physics Department at Georgetown University. In a series of papers the Clinton school introduced into crystallography the concept of *N*-representability.

Over the past many years a voluminous literature concerning the problem of *N*-representability (Figure 1.1) has arisen [1–11]. Because of the physical indistinguishability of particles, every valid approximation to a solution of the Schrödinger equation must be antisymmetric in the coordinate permutation of fermion pairs. Given such antisymmetric functions, Ψ, one may define reduced density matrices:

$$\varrho_p(1 \cdots p, 1' \cdots p') = N(N-1) \cdots (N-p+1) \int \Psi^* \Psi \, d(p+1) \cdots dN \qquad (1.1)$$

The problem of *N*-representability is that of finding conditions by which to recognize these ϱ_p, which are assured to be related to an *N*-body wavefunction according to the rule of Equation 1.1.

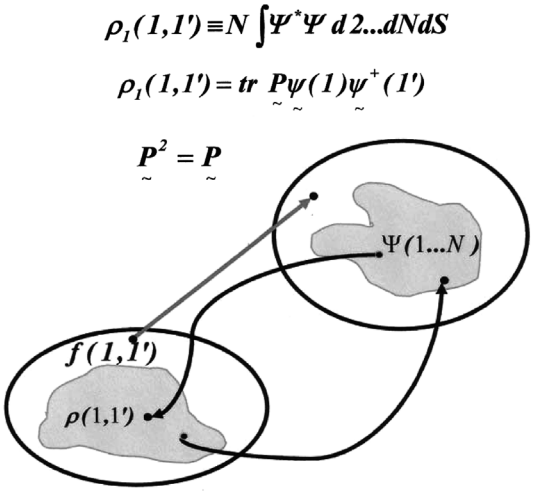


Figure 1.1 Sketch indicating the mapping problem associated with wavefunction representability of density matrices.

Particularly important for the calculation of almost all interesting physical properties are the cases for $p = 2$, and $p = 1$, viz.:

$$\varrho_2(1, 2, 1', 2') = N(N-1) \int \Psi^* \Psi \mathrm{d}3 \cdots \mathrm{d}N \tag{1.2}$$

$$\varrho_1(1, 1') = N \int \Psi^* \Psi \, \mathrm{d}2 \cdots \mathrm{d}N \tag{1.3}$$

In the case of spinless density matrices, integration occurs over all spins.

For the usual case of Hamiltonians containing at most two-body interactions, the second-order reduced density matrix determines completely the energy of the system.

The problem of finding the conditions that allow the mapping of the objects of Equation 1.1, viz., ϱ_p and Ψ into one another is important mathematically. Moreover, there are important physical and computational aspects to the problem. One sees immediately that, for example, ϱ_2 is, inherently a simpler object than is $\Psi(1 \ldots N)$, since it depends only upon the coordinates of two particles, no matter how great is N. Knowledge of N-representable ϱ_2 would allow direct minimization of the energy with respect to the parameters of ϱ_2, thus eliminating the need for handling an N-body wavefunction. The variation principle, which supplies an upper bound for every approximate ϱ_2, will hold so long as N-representability of ϱ_2 is satisfied. A practical quantum mechanics might, in such fashion, be framed entirely within the context of density matrices without any explicit computational role played by N-body wavefunctions.

The problem of N-representability is still a subject of current interest. Although very much has been learned the complete problem of N-representability of ϱ_2 has not been solved. Interestingly, the case of N-representability by a single determinant of orbitals is well understood. Idempotency of the one-body density matrix ϱ_1 completely characterizes this case, for which moreover all higher order density matrices are known functionals of ϱ_1. Of course independent particle models, including the Hartree–Fock and density functional theory cases, are all encompassed within single determinant wavefunctions. The N-representability problem is solved, as far as single Slater determinants are concerned, [6]. Another case for which N-representability is no difficulty occurs for the density itself, that is, $\varrho(1) = \varrho_1(1,1')|_{1' \to 1}$, the diagonal elements of the one-body density matrix. It occurs by a theorem of Gilbert [12] that any normalized, well behaved density is N-representable by a single Slater determinant of orbitals. We have shown by calculations with select examples that an exact density is N-representable by a Slater determinant of physically meaningful orbitals [13].

1.2.2
Single Determinant *N*-Representability

In one case, that characterized by a Slater determinant wavefunction, N-representability of reduced density matrices presents no problem. Such density matrices have

been studied exhaustively and their properties are well understood. We review points of interest.

We take a set of orthonormal molecular spin orbitals $\{\varphi_i(i=1 \cdots N\}$ and with them construct a Slater determinant (an antisymmetric function carrying the physical implications of the Pauli principle):

$$\Psi_{\text{det}}(1 \cdots N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi(1) & \cdots & \phi_1(N) \\ \vdots & & \vdots \\ \phi_N(1) & \cdots & \phi_N(N) \end{vmatrix} \tag{1.4}$$

Such a determinant satisfies the normalization condition:

$$\int \Psi^* \Psi \, \mathrm{d}1 \cdots \mathrm{d}N = 1 \tag{1.5}$$

By direct integration over the product of the Slater determinant with itself, the reduced density matrices of every order may be constructed. For example:

$$\varrho_{N\text{det}} = N!\Psi^*_{\text{det}}\Psi_{\text{det}} = \begin{vmatrix} \varrho(1,1') & \cdots & \varrho(1,N') \\ \vdots & & \vdots \\ \varrho(N,1') & \cdots & \varrho(N,N') \end{vmatrix} \tag{1.6}$$

$$\varrho_{2\,\text{det}} = N(N-1)\int \Psi^*_{\text{det}}\Psi_{\text{det}} \, \mathrm{d}3 \cdots \mathrm{d}N = \begin{vmatrix} \varrho(1,1') & \varrho(1,2') \\ \varrho(2,1') & \varrho(2,2') \end{vmatrix} \tag{1.7}$$

$$\varrho_{1\,\text{det}} = N\int \Psi^*_{\text{det}}\Psi_{\text{det}} \, \mathrm{d}2 \cdots \mathrm{d}N = \varrho(1,1') \tag{1.8}$$

The necessary and sufficient conditions for this one-body density matrix to be N-representable by a single Slater determinant are:

$$\varrho_1^2 = \varrho_1, \quad \int \varrho_1 \, \mathrm{d}1 = N, \quad \varrho_1^\dagger = \varrho_1 \tag{1.9}$$

The density matrix must be idempotent, normalized and hermetian, conditions both simple and of practical utility. McWeeney [8] has shown that a density matrix may be purified to idempotency via an iterative expression and also that an idempotent density matrix can always be factored into a sum of squares of orbitals. The orbitals are not unique in the sense that the one-body density matrix is invariant to a unitary transformation among them. Knowledge of $\varrho_{1\text{-det}}$ fixes $\varrho_{2\text{-det}}$ and every higher reduced density matrix up to and including $\varrho_{N\text{-det,}}$ and Ψ_{det} itself. For a two-body Hamiltonian of the usual type:

$$\hat{H} = \sum \hat{h}_i + \sum \hat{h}_{ij} \tag{1.10}$$

the energy:

$$E = \int \hat{h}_1 \, \varrho_{1\text{det}}(1,1')|_{1' \to 1} \, \mathrm{d}1 + \int \hat{h}_{12} \, \varrho_{2\,\text{det}}(1,2) \, \mathrm{d}1 \, \mathrm{d}2 \geq E_0 \tag{1.11}$$

satisfies the variational theorem. We mention in passing, for the above expression of the energy, that the off-diagonal elements of ϱ_1 are required, but only the diagonal

elements of ϱ_2. E is, of course, invariant to a unitary transformation among the orbitals. Direct minimization of E, expressed by Equation 1.11, produces the approximate Hartree–Fock energy appropriate to the basis used for expansion of the density matrix.

According to the theorem of Gilbert [12] *every* well-behaved electron density (positive and normalized) is N-representable by a single Slater determinant of orbitals. Of course this is obvious for any Hartree–Fock density, but interestingly the theorem is totally general, and holds equally well for the exact density corresponding to the full Hamiltonian. Every $\varrho(1)$ is N-representable by some $\Psi_{\text{det}}(1\ldots N)$.

McWeeney's purification to idempotency [8] may be modified to include conditions of constraint as in Clinton's equations [14]:

$$\underset{\sim}{\mathbf{P}}_{n+1} = 3\underset{\sim}{\mathbf{P}}_n^2 - 2\underset{\sim}{\mathbf{P}}_n^3 + \sum\nolimits_k \lambda_k \underset{\sim}{\mathbf{Q}}_k + \lambda_N \underset{\sim}{\mathbf{1}} \tag{1.12}$$

In Equation 1.12 the λ's are Lagrangian multipliers determined from equations of constraint, for example:

$$O_k = \operatorname{tr} \underset{\sim}{\mathbf{P}} \underset{\sim}{\mathbf{Q}}_k \tag{1.13}$$

$$1 = \operatorname{tr} \underset{\sim}{\mathbf{P}} \underset{\sim}{\mathbf{1}} \tag{1.14}$$

where $\underset{\sim}{\mathbf{Q}}_k$ is the matrix representative of an arbitrary quantum operator $\mathbf{Q}_k$ and $\underset{\sim}{\mathbf{1}}$ is the matrix representative of the normalization operator $\mathbf{1}$. $\underset{\sim}{\mathbf{P}}$ is the Löwdin population matrix or density matrix in an orthonormal basis.

Clinton's equations have the physical significance of delivering a one-body density matrix, N-representable by a single Slater determinant, and satisfying chosen quantum conditions of constraint. Applied in context of the X-ray coherent diffraction experiment [15] these equations can deliver the exact experimental electron density. For such a case, the experimental Bragg structure factors $F(K)$ provide conditions of constraint via the Fourier transform relation:

$$F(\mathbf{K}) = \int \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}} \varrho(\mathbf{r})\, \mathrm{d}^3\mathbf{r} \tag{1.15}$$

where the electron density is:

$$\varrho(\mathbf{r}) = \varrho(\mathbf{r}, \mathbf{r}')|_{\mathbf{r}' \to \mathbf{r}} \tag{1.16}$$

the diagonal elements of the density matrix. Clinton's equations, applied with an appropriate basis, are capable, consistent with Gilbert's theorem, of delivering physically meaningful orbitals that satisfy the experimental (and therefore exact) density. Within quantum crystallography, this has proven to be one of their important uses.

1.2.3 Example Applications of Clinton's Equations

1.2.3.1 Beryllium

We applied the Clinton equations to a beryllium crystal using the very accurate X-ray scattering factor data of Larsen and Hansen [15]. As may be seen in Figure 1.2 the

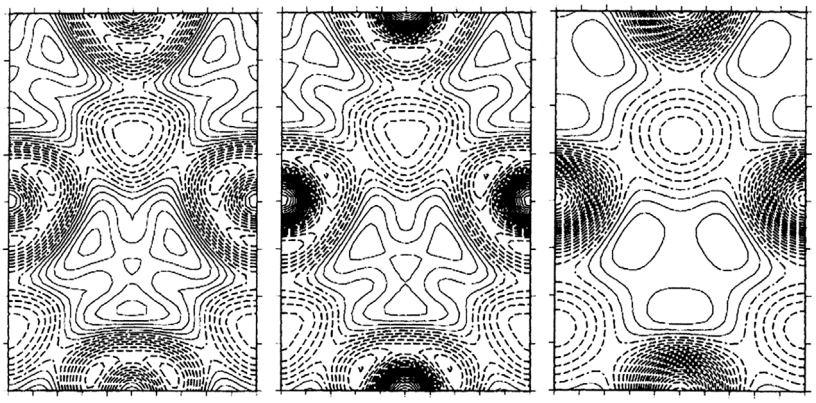


Figure 1.2 Valence density from (a) Dovesi *et al.* [16], (b) this work [15] and (c) Chou, Lam and Cohen [17]. Projections of the tetrahedral and octahedral holes are indicated..

experimental density obtained was obtained was quite accurate, as measured by reference to the best theoretical densities that were available for that crystal. The experimental density contours are very similar to the theoretical contours [15–17]. In Figure 1.3 the errors in scattering factor F are plotted as a function of scattering angle [15]. The errors are randomly distributed out to high angles of scattering. At the time of this result, $R = 0.0018$, achieved with an N-representable density matrix, was perhaps the smallest R factor in the literature of crystallography. This established that N-representable density descriptions of actual X-ray scattering data were practicable and of high accuracy.

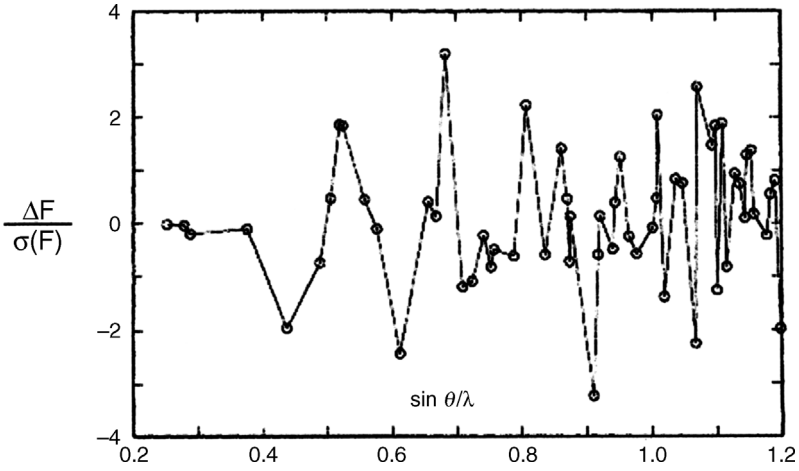


Figure 1.3 Distribution of errors [15]; $R_{wf} = 0.0018$ and G.O.F. = 1.33.

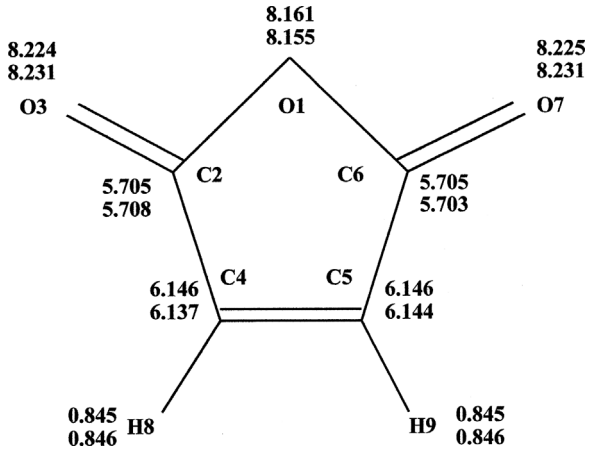


Figure 1.4 Electrons per atom in maleic anhydride. Upper numbers obtained from optimized theoretical calculation with B3LYP/cc-pVTZ. Lower numbers obtained from experimental coordinates and a single point calculation with B3LYP/cc-pVTZ.

1.2.3.2 **Maleic Anhydride**

This section concerns the application of Clinton's equations to a crystal of maleic anhydride [18], a small, flat molecule, having only nine atoms (Figure 1.4). Data collection and crystallographic refinement for this study were carried out by Louis Todaro. The authors refined the elements of the projector matrix by use of the Clinton iterative equations and the structure factor magnitudes obtained from an X-ray diffraction investigation. The final *R*-factor between the experimental structure factor magnitudes and the theoretical ones from the projector matrix for 6-31G** was less than 1.5%.

A total of 507 independent data were used. The experimental data were collected with CuKα radiation at 110(1) K. A calculation of the resolution of these data yielded a value of about 0.80 Å, and the number of independent elements in the projector matrix was 2250. The total number of data available for the refinement of the elements in the projector matrix was $8 \times 507 = 4056$, and so the ratio of data to independent unknowns was 1.80. After the independent data were corrected for vibrational effects and expanded to include all equivalent reflections for space group $P2_12_12_1$, the following results were obtained.

Tables 1.1 and 1.2 display calculated energies and atomic charges, respectively. Clinton's equations yielded both an experimental density matrix and experimental atomic coordinates. There was no significant difference in the coordinates obtained using Clintons equations and those obtained from an ordinary crystallographic least-squares determination, except for the hydrogen atoms, which are placed differently in X-ray diffraction experiments and in quantum mechanical modeling. The implications for maleic anhydride were that perhaps an accurate and efficient way to combine diffraction data with quantum mechanics is to use the heavy atom coordinates obtained crystallographically, holding them fixed, and then carry out the *ab initio* quantum mechanical calculations for the system. The burden for obtaining

Table 1.1 Energies for maleic anhydride.

	Energies (au)[a]					
	E_{total}	T	NE	EE	NN	$-V/T$
Exp. Coord[b].	−379.432	377.065	−1439.600	407.672	275.431	2.0063
OPT[c]	−379.435	377.126	−1439.658	407.659	275.438	2.0061

a) Total energy (E_{total}); total electronic kinetic energy (T); total nuclei–electrons attractive potential energy (NE); total electron–electron repulsion energy (EE); total nuclear–nuclear repulsion energy (NN); and the negative of the virial ratio ($-V/T$) the ratio of the potential energy ($V = NE + EE + NN$) to the kinetic energy (T), a ratio that should ideally be exactly 2 according to the "virial theorem" (in a calculation of infinite precision).
b) The single point calculation was performed with the use of the experimental coordinates and B3LYP/cc-pVTZ.
c) OPT refers to geometry optimization with B3LYP/cc-pVTZ.

Table 1.2 Electrons per atom for maleic anhydride.

	Atoms								
	H_8	C_4	O_3	C_2	O_1	H_9	C_5	O_7	C_6
Exp. Coord.[a]	0.846	6.137	8.231	5.708	8.155	0.846	6.144	8.231	5.703
OPT[b]	0.845	6.146	8.224	5.705	8.161	0.845	6.146	8.225	5.705

a) Single point calculations were performed with the use of the experimental coordinates and B3LYP/cc-pVTZ.
b) OPT refers to geometry optimization with B3LYP/cc-pVTZ.

quantum mechanical information is then placed upon use of a sufficiently accurate chemical model, and the problem of atomic coordinates is simply taken from the normal crystallography. This observation had an influence in the creation of the kernel energy method discussed below. The experimental density matrix obtained from Clinton's equations delivered energies and atomic charges similar to those obtained directly from the density functional theory calculations at the experimental coordinates, the latter of which are shown in Figure 1.4, and are compared to the analogous atomic charges at the DFT optimized coordinates. The overall result is a close correspondence between the N-representable experimental and theoretically calculated charge distribution and energies for the maleic anhydride molecule.

1.3 Beginnings of Quantum Kernels

1.3.1 Computational Difficulty of Large Molecules

Large molecules are a special problem. For example, the computational difficulty of solving the Schrödinger equation increases with a high power of the number of atoms

(or basis functions) in the molecule. In addition, when fixing the elements of the density matrix by a least-squares fit to the X-ray scattering data, it is desirable that the number of data should exceed in good measure the number of matrix elements. But, as the size of a molecule increases the ratio of number of data to number of matrix elements tends to become too small for a reliable determination of the density matrix. The desire to represent increasingly large molecules forced us to consider how to surmount the computational difficulties associated with size. This led to a simple idea, variations of which had occurred to tens of different research groups. That idea was to take a large molecule, break it into smaller pieces, represent the smaller and more tractable pieces, and then put them back together in such fashion as to reconstitute a representation of the original large molecule. One particular method in which this idea is carried out occurs within quantum crystallography.

1.3.2
Quantum Kernel Formalism

The basic formalism that introduces the important idea of the essential molecular pieces, called kernels, [19, 20] is presented in the following paragraphs.

The kernel calculations to be presented here are based on structural data, that is, atomic positions. X-Ray scattering data are used routinely to determine molecular structure, that is, equilibrium atomic arrangements and thermal (disorder) parameters. The same data, when sufficiently accurate, can also be used to obtain the electron density distribution of the unit cell of a crystal [21]. The electron density distribution for a crystal, ϱ, can also be expressed in terms of the trace of a suitable matrix product [13–15, 22–27] according to:

$$\varrho = 2\mathrm{tr}\,\underset{\sim}{\boldsymbol{\varphi}}\underset{\sim}{\boldsymbol{\varphi}}^{\dagger} \tag{1.17}$$

The column matrix $\underset{\sim}{\boldsymbol{\varphi}}$ is composed of doubly occupied orthonormal molecular orbitals, giving rise to the factor of 2. Most molecular ground states have doubly occupied orbitals. In other cases, the formalism may be appropriately generalized.

If we write:

$$\underset{\sim}{\boldsymbol{\varphi}} = \underset{\sim}{\mathbf{C}}\boldsymbol{\Psi} \tag{1.18}$$

Equation 1.17 becomes:

$$\varrho = 2\mathrm{tr}\,\underset{\sim}{\mathbf{C}}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\dagger}\underset{\sim}{\mathbf{C}}^{\dagger} = 2\mathrm{tr}\,\underset{\sim}{\mathbf{C}}^{\dagger}\underset{\sim}{\mathbf{C}}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\dagger} \tag{1.19}$$

The value of the trace is insensitive to the cyclic interchange of the position of $\underset{\sim}{\mathbf{C}}^{\dagger}$. The following definitions are made:

$$\underset{\sim}{\mathbf{S}} = \int \boldsymbol{\Psi}\boldsymbol{\Psi}^{\dagger}\,\mathrm{d}\mathbf{r} \tag{1.20}$$

where the integration is performed over the individual elements of the product matrix $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\dagger}$:

$$\underset{\sim}{\mathbf{R}} = \underset{\sim}{\mathbf{C}}^{\dagger}\underset{\sim}{\mathbf{C}} \tag{1.21}$$

and:

$$\underset{\sim}{\mathbf{R}}\underset{\sim}{\mathbf{S}} = \underset{\sim}{\mathbf{P}}_{\mathbf{a}} \tag{1.22}$$

where $\underset{\sim}{\mathbf{P}}_{\mathbf{a}}$ is a projector. The subscript **a** indicates that unless special steps are taken in forming the matrix $\underset{\sim}{\boldsymbol{\Psi}}$ the projector matrix will not be symmetric. It can be shown that, as a consequence of the fact that the $\underset{\sim}{\boldsymbol{\varphi}}$ are composed of elements that are orthonormal:

$$\underset{\sim}{\mathbf{P}}_{\mathbf{a}}^{2} = \underset{\sim}{\mathbf{P}}_{\mathbf{a}} \tag{1.23}$$

and:

$$\mathrm{tr}\,\underset{\sim}{\mathbf{P}}_{\mathbf{a}} = N \tag{1.24}$$

where N is the number of doubly occupied orbitals in the molecule of interest (N should not to be confused with the symbol number of electrons as the context indicates elsewhere in this chapter). Equation 1.23 is the projector property.

It is convenient to have a projector $\underset{\sim}{\mathbf{P}}_{\mathbf{s}}$ that is symmetric since it reduces the number of elements in the projector that must be evaluated. From Equation 1.22, it follows that:

$$\underset{\sim}{\mathbf{S}}^{1/2}\underset{\sim}{\mathbf{R}}\underset{\sim}{\mathbf{S}}\underset{\sim}{\mathbf{S}}^{-1/2} = \underset{\sim}{\mathbf{S}}^{1/2}\underset{\sim}{\mathbf{P}}_{\mathbf{a}}\underset{\sim}{\mathbf{S}}^{-1/2} \tag{1.25}$$

This matrix product is a symmetric projector $\underset{\sim}{\mathbf{P}}_{\mathbf{s}}$ and may be written:

$$\underset{\sim}{\mathbf{P}}_{\mathbf{s}} = \underset{\sim}{\mathbf{S}}^{1/2}\underset{\sim}{\mathbf{R}}\underset{\sim}{\mathbf{S}}^{1/2} \tag{1.26}$$

It follows from Equations 1.20–1.22 and 1.25 that the electron density can be written:

$$\varrho = 2\mathrm{tr}\,\underset{\sim}{\mathbf{R}}\underset{\sim}{\boldsymbol{\Psi}}\underset{\sim}{\boldsymbol{\Psi}}^{\dagger} = 2\mathrm{tr}\,\underset{\sim}{\mathbf{P}}_{\mathbf{a}}\underset{\sim}{\mathbf{S}}^{-1}\underset{\sim}{\boldsymbol{\Psi}}\underset{\sim}{\boldsymbol{\Psi}}^{\dagger} \tag{1.27}$$

and:

$$\varrho = 2\mathrm{tr}\,\underset{\sim}{\mathbf{P}}_{\mathbf{s}}\underset{\sim}{\mathbf{S}}^{-1/2}\underset{\sim}{\boldsymbol{\Psi}}\underset{\sim}{\boldsymbol{\Psi}}^{\dagger}\underset{\sim}{\mathbf{S}}^{\dagger-1/2} \tag{1.28}$$

It will be seen that in the application of the calculation of fragment densities to obtain kernel densities it is convenient to compute the projector $\underset{\sim}{\mathbf{P}}_{\mathbf{a}}$. In the further application of quantum crystallography, to adjust the values of the projector with the use of diffraction data from a crystal, it is more suitable to use $\underset{\sim}{\mathbf{P}}_{\mathbf{s}}$.

There is a third type of projector, $\underset{\sim}{\mathbf{P}}_{\mathbf{s}\Gamma}$, that is useful because it is a symmetric projector that has fewer elements than $\underset{\sim}{\mathbf{P}}_{\mathbf{s}}$. It arises from the use of point group symmetry to form symmetry orbitals as a basis for the molecular orbitals. Matrices $\underset{\sim}{\mathbf{T}}_{\mathbf{s}\Gamma}$ associated with the irreducible representations of the point group of a molecule [28] can be formed that transform atomic orbitals into symmetry orbitals by the operation $\underset{\sim}{\mathbf{T}}_{\mathbf{s}\Gamma}\underset{\sim}{\boldsymbol{\Psi}}_{\mathbf{m}}$, where the subscript **s** associates $\underset{\sim}{\mathbf{T}}$ with symmetry orbitals and the subscript $\boldsymbol{\Gamma}$ associates $\underset{\sim}{\mathbf{T}}$ with the irreducible representations. The subscript **m** denotes the fact that $\underset{\sim}{\boldsymbol{\Psi}}_{\mathbf{m}}$ is composed of orbitals for a molecule (not the entire unit cell). The coefficients associated with $\underset{\sim}{\mathbf{T}}_{\mathbf{s}\boldsymbol{\Gamma}}\underset{\sim}{\boldsymbol{\Psi}}_{\mathbf{m}}$ are denoted by $\underset{\sim}{\mathbf{C}}_{\boldsymbol{\Gamma}}$,

giving:

$$\varrho = \sum_{\Gamma} 2\mathrm{tr}\, \underset{\sim}{\mathbf{C}}^{\dagger}_{\mathbf{\Gamma}} \underset{\sim}{\mathbf{C}}_{\mathbf{\Gamma}} \underset{\sim}{\mathbf{T}}_{\mathbf{s\Gamma}} \left[\sum_{\hat{R}} \hat{R} \underset{\sim}{\mathbf{\Psi}}_{\mathbf{m}} \underset{\sim}{\mathbf{\Psi}}^{\dagger}_{\mathbf{m}} \right] \underset{\sim}{\mathbf{T}}^{\dagger}_{\mathbf{s\Gamma}}, \tag{1.29}$$

or:

$$\varrho = \sum_{\Gamma} 2\mathrm{tr}\, \underset{\sim}{\mathbf{R}}_{\mathbf{\Gamma}} \underset{\sim}{\mathbf{T}}_{\mathbf{s\Gamma}} \left[\sum_{\hat{R}} \hat{R} \underset{\sim}{\mathbf{\Psi}}_{\mathbf{m}} \underset{\sim}{\mathbf{\Psi}}^{\dagger}_{\mathbf{m}} \right] \underset{\sim}{\mathbf{T}}^{\dagger}_{\mathbf{s\Gamma}} \tag{1.30}$$

where $\hat{R}$ represents the symmetry operations of the crystallographic space group of interest and $\underset{\sim}{\mathbf{\Psi}}_{\mathbf{m}}$ are composed of the atomic orbitals for a molecule. To change $\underset{\sim}{\mathbf{R}}_{\Gamma}$ into a symmetric projector, we write an expression equivalent to Equation 1.30:

$$\varrho = \sum_{\Gamma} 2\mathrm{tr}\, \underset{\sim}{\mathbf{S}}^{1/2}_{\mathbf{\Gamma}} \underset{\sim}{\mathbf{R}}_{\mathbf{\Gamma}} \underset{\sim}{\mathbf{S}}^{1/2}_{\mathbf{\Gamma}} \underset{\sim}{\mathbf{S}}^{-1/2}_{\mathbf{\Gamma}} \underset{\sim}{\mathbf{T}}_{\mathbf{s\Gamma}} \times \left[\sum_{\hat{R}} \hat{R} \underset{\sim}{\mathbf{\Psi}}_{\mathbf{m}} \underset{\sim}{\mathbf{\Psi}}^{\dagger}_{\mathbf{m}} \right] \underset{\sim}{\mathbf{T}}^{\dagger}_{\mathbf{s\Gamma}} \underset{\sim}{\mathbf{S}}^{\dagger\,-1/2}_{\mathbf{\Gamma}} \tag{1.31}$$

or:

$$\varrho = \sum_{\Gamma} 2\mathrm{tr}\, \underset{\sim}{\mathbf{P}}_{\mathbf{s\Gamma}} \underset{\sim}{\mathbf{S}}^{-1/2}_{\mathbf{\Gamma}} \underset{\sim}{\mathbf{T}}_{\mathbf{s\Gamma}} \times \left[\sum_{\hat{R}} \hat{R} \underset{\sim}{\mathbf{\Psi}}_{\mathbf{m}} \underset{\sim}{\mathbf{\Psi}}^{\dagger}_{\mathbf{m}} \right] \underset{\sim}{\mathbf{T}}^{\dagger}_{\mathbf{s\Gamma}} \underset{\sim}{\mathbf{S}}^{\dagger\,-1/2}_{\mathbf{\Gamma}} \tag{1.32}$$

$\underset{\sim}{\mathbf{P}}_{\mathbf{s}\Gamma}$ is symmetric, is associated with symmetry orbitals and:

$$\mathbf{S}_{\Gamma} = \int \underset{\sim}{\mathbf{T}}_{\mathbf{s}\Gamma} \underset{\sim}{\mathbf{\Psi}}_{\mathbf{m}} \underset{\sim}{\mathbf{\Psi}}^{\dagger}_{\mathbf{m}} \underset{\sim}{\mathbf{T}}^{\dagger}_{\mathbf{s}\Gamma}\, \mathrm{d}\mathbf{r} \tag{1.33}$$

where the integration over all space is performed for all individual elements of the product matrix, $\underset{\sim}{\mathbf{\Psi}}_{\mathbf{m}} \underset{\sim}{\mathbf{\Psi}}^{\dagger}_{\mathbf{m}}$.

In the single-determinant approach taken here, the Fourier transforms of Equations 1.28 or 1.32 may be considered to be the basic equations of quantum crystallography. Their Fourier transforms yield the structure factors of crystallographic theory, whose magnitudes are definable in terms of the measured diffraction intensities. The mathematical objective of quantum crystallography is to optimize the fit of the elements of the projector matrix to the experimental structure factor magnitudes and also the fit of some other parameters that occur in the Fourier transform of the right-hand side of Equations 1.28 or 1.32. In addition to the positional coordinates of the atoms, adjustments are made to three scaling factors, which set the average value of the calculated structure factor magnitudes equal to the average of the observed one. Provision may also be made, in quantum crystallography, to adjust the value of thermal parameters attached to the atomic basis orbitals, which have the effect of simulating a smearing of density due to atomic motions.

The fragment calculations that will now be described deliver parts of the $\underset{\sim}{\mathbf{R}}$ matrix with good accuracy. They may then be assembled into the complete $\underset{\sim}{\mathbf{R}}$ matrix, and by use of Equation 1.26 the symmetric projector $\underset{\sim}{\mathbf{P}}_{\mathbf{s}}$ may be formed for use in the quantum crystallography calculations.

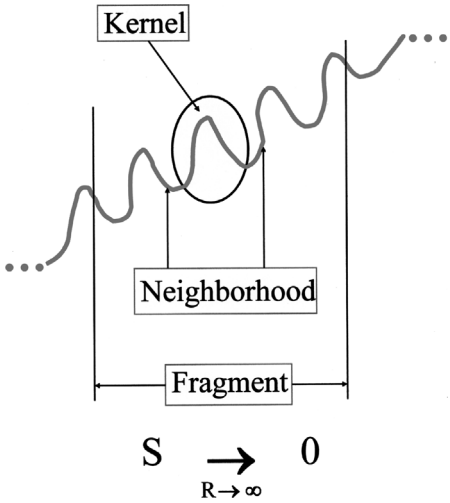


Figure 1.5 Kernel, neighborhood and fragment.

1.3.3
Kernel Matrices: Example and Results

The purpose of kernel calculations is to obtain an accurate $\underset{\sim}{\mathbf{R}}$ matrix when *ab initio* calculations of an entire molecule are either not feasible or considered to be too time-consuming. As illustrated in Figure 1.5, a fragment consists of an inner core or "kernel" and several neighboring atoms called a "neighborhood." The molecule is divided into a suitable number of kernels, which, when recombined, form the complete molecule.

Since the coordinates of the structure of interest are available, it is readily possible to calculate which atoms would occur within a certain chosen distance from all the atoms in a kernel. Such atoms would form the neighborhood. To maintain an electron balance, it may be necessary to attach hydrogen atoms to some of the neighborhood atoms in a fragment. There are various schemes conceivable for choosing the ways in which a molecule may be broken up into kernels and neighborhoods. One of general applicability, which still allows some arbitrary choice, can be based on the rule that all atoms present must be a member of some kernel once and only once. With atomic positions held fixed, the electron density distribution in a fragment is computed. Such a calculation can deliver contributions to an $\underset{\sim}{\mathbf{R}}$ matrix from which the portion that concerns the kernel is saved. Those contributions to the $\underset{\sim}{\mathbf{R}}$ matrix involving orbitals from a neighborhood atom and an atom in the kernel are saved at the fractional value of one-half, in accordance with the above rule. If all neighborhood atoms occur only as part of a kernel, another one-half value would be added to those contributions already saved at one-half values, when the values associated with the adjoining kernels are calculated. Contributions from pairs of atoms, both in the same kernel, are saved with a coefficient of one. The final $\underset{\sim}{\mathbf{R}}$ matrix will be multiplied by the $\underset{\sim}{\mathbf{S}}$ matrix to give $\underset{\sim}{\mathbf{P}}_a$, and since $\underset{\sim}{\mathbf{S}}$ is an overlap matrix, values close to zero will be

obtained for pairs of atoms that are separated by large distances. The pattern of zeros in $\underset{\sim}{\mathbf{S}}$ is used to generate zeros in $\underset{\sim}{\mathbf{R}}$, justified by the symmetry of $\underset{\sim}{\mathbf{S}}$, namely, $\underset{\sim}{\mathbf{S}} = \underset{\sim}{\mathbf{S}}^{\dagger}$, and the invariance of Tr $\underset{\sim}{\mathbf{P}}\underset{\sim}{\mathbf{S}}$ to the insertion into $\underset{\sim}{\mathbf{R}}$ of the pattern of zeros in $\underset{\sim}{\mathbf{S}}$. The behavior of $\underset{\sim}{\mathbf{S}}$ is the reason why the fragment calculations can give accurate values for the molecule as a whole.

The kernel calculations for a hydrated hexapeptide [29] were performed by defining the kernels as the six peptide residues in the ring with each of the three water molecules associated with the appropriate residues as determined by proximity. The neighborhoods in the fragments were formed by the amino acid residues and associated water molecules, if any, adjoining the one considered as the kernel, for example, residues 3 and 5 were the neighborhood for residue 4 acting as a kernel.

We may write the **R**-matrix for the full hexapeptide molecule as:

$$\mathbf{R} = \begin{vmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \cdots & \mathbf{R}_{16} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \cdots & \mathbf{R}_{26} \\ \vdots & & & \vdots \\ \mathbf{R}_{61} & \mathbf{R}_{62} & \cdots & \mathbf{R}_{66} \end{vmatrix} \tag{1.34}$$

where the subscripts refer to each of the six kernels composed of the six amino acid residues (some associated with water molecules) in the hexapeptide. Each element of the matrix Equation 1.34 is itself a matrix whose dimensions are those of the bases associated with each of the kernels labeled by the subscripts.

A matrix associated with each of the six kernels, $\mathbf{R}_j$ ($j = 1, \ldots 6$), may be formed consistent with the rules of Mulliken population analysis, giving:

$$\mathbf{R} = \sum_{j=1}^{6} \mathbf{R}_j \tag{1.35}$$

where $\mathbf{R}_j$ is composed of the sum of two matrices, one whose only nonzero components are $0.5\mathbf{R}_{jk}$ ($k = 1, 2, \ldots 6$) and one whose only nonzero components are $0.5\mathbf{R}_{kj}$ ($k = 1, 2, \ldots 6$). For example, when $j = 4$:

$$\mathbf{R}_4 = \begin{vmatrix} & & & \mathbf{R}_{14}/2 & & \\ & 0 & & \mathbf{R}_{24}/2 & & 0 \\ & & & \mathbf{R}_{34}/2 & & \\ \mathbf{R}_{41}/2 & \mathbf{R}_{42}/2 & \mathbf{R}_{43}/2 & \mathbf{R}_{44} & \mathbf{R}_{45}/2 & \mathbf{R}_{46}/2 \\ & 0 & & \mathbf{R}_{54}/2 & & 0 \\ & & & \mathbf{R}_{64}/2 & & \end{vmatrix} \tag{1.36}$$

The correspondence of Equations 1.34 and 1.35 may be readily verified. The approximation was made that each kernel "overlaps" with a neighborhood that includes only one kernel on either side of the given kernel. This limits the range of k in the $\mathbf{R}_j$ of Equation 1.35 to $k = j - 1, j, j + 1$ instead of $k = 1, 2, \ldots 6$, with $k = 0$

equivalent to $k=6$ and $k=7$ equivalent to $k=1$. Equation 1.36 becomes:

$$\mathbf{R}_4(0) = \begin{vmatrix} & & & 0 & & \\ & 0 & & 0 & & 0 \\ & & & \mathbf{R}_{34}/2 & & \\ 0 & 0 & \mathbf{R}_{43}/2 & \mathbf{R}_{44} & \mathbf{R}_{45}/2 & 0 \\ & 0 & & \mathbf{R}_{54}/2 & & 0 \end{vmatrix} \tag{1.37}$$

where $\underset{\sim}{\mathbf{R}}_4(0)$ indicates that $\underset{\sim}{\mathbf{R}}_4$ is modified by the introduction of truncated neighborhoods, which introduces a pattern of zeros.

The full molecule R-matrix is approximated by summing over the kernel matrices associated with truncated neighborhoods, that is:

$$\mathbf{R}(0) = \sum_{j=1}^{6} \mathbf{R}_j(0) \tag{1.38}$$

which, when written out, is:

$$\underset{\sim}{\mathbf{R}}(0) = \begin{vmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & 0 & 0 & 0 & \mathbf{R}_{16} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \mathbf{R}_{23} & 0 & 0 & 0 \\ 0 & \mathbf{R}_{32} & \mathbf{R}_{33} & \mathbf{R}_{34} & 0 & 0 \\ 0 & 0 & \mathbf{R}_{43} & \mathbf{R}_{44} & \mathbf{R}_{45} & 0 \\ 0 & 0 & 0 & \mathbf{R}_{54} & \mathbf{R}_{55} & \mathbf{R}_{56} \\ \mathbf{R}_{61} & 0 & 0 & 0 & \mathbf{R}_{65} & \mathbf{R}_{66} \end{vmatrix} \tag{1.39}$$

Thus, the electron density distribution for the full molecule is approximately:

$$\varrho(0) = 2\mathrm{tr}\,\underset{\sim}{\mathbf{R}}(0)\left[\underset{\sim}{\boldsymbol{\Psi}}\underset{\sim}{\boldsymbol{\Psi}}^{\dagger}\right](0) \tag{1.40}$$

where the last (0) indicates a pattern of zeros in the matrix $\left[\underset{\sim}{\boldsymbol{\Psi}}\underset{\sim}{\boldsymbol{\Psi}}^{\dagger}\right]$ analogous to the pattern of zeros in $\underset{\sim}{\mathbf{R}}(0)$. The pattern of zeros in $\left[\underset{\sim}{\boldsymbol{\Psi}}\underset{\sim}{\boldsymbol{\Psi}}^{\dagger}\right](0)$ is the same as in $\underset{\sim}{\mathbf{S}}(0)$ and $\underset{\sim}{\mathbf{R}}(0)$. Not only are the overlap integrals very small for the product of those elements that are set equal to zero, the product before integration is also very small. We see how a density function for a complete molecule may be obtained approximately from matrices of smaller kernels.

Approximate matrices based on kernels may produce electron densities whose suitability may be further enhanced by ensuring their *N*-representability [6–8]. This is achieved by requiring the matrices to be normalized projectors. These properties may be imposed on a matrix $\underset{\sim}{\mathbf{R}}$ [and also $\underset{\sim}{\mathbf{R}}(0)$] by use of Clinton's iterative equations [14] in the form:

$$\underset{\sim}{\mathbf{R}}_{n+1} = 3\underset{\sim}{\mathbf{R}}_n\underset{\sim}{\mathbf{S}}\underset{\sim}{\mathbf{R}}_n - 2\underset{\sim}{\mathbf{R}}_n\underset{\sim}{\mathbf{S}}\underset{\sim}{\mathbf{R}}_n\underset{\sim}{\mathbf{S}}\underset{\sim}{\mathbf{R}}_n + \lambda\underset{\sim}{\mathbf{1}} \tag{1.41}$$

subject to the normalization condition given by:

$$\mathrm{tr}\,\underset{\sim}{\mathbf{R}}\underset{\sim}{\mathbf{S}} = N \tag{1.42}$$

where $N = 113$ is the number of doubly occupied molecular orbitals for the hydrated hexapeptide. Condition 1.42 requires that:

$$\lambda = \left[N - \mathrm{tr}\,(3\underset{\sim}{\mathbf{R}}_n\underset{\sim}{\mathbf{S}}\underset{\sim}{\mathbf{R}}_n\underset{\sim}{\mathbf{S}} - 2\underset{\sim}{\mathbf{R}}_n\underset{\sim}{\mathbf{S}}\underset{\sim}{\mathbf{R}}_n\underset{\sim}{\mathbf{S}}\underset{\sim}{\mathbf{R}}_n\underset{\sim}{\mathbf{S}})\right]/M \tag{1.43}$$

where $M = 173$ is the dimension of the Gaussian basis: $\phi_i = \sum_{j=1}^{173} C_{ij}\psi_j$

1.3.4 Applications of the Idea of Kernels

1.3.4.1 Hydrated Hexapeptide Molecule

Isodensity surfaces have been calculated for a hydrated hexapeptide molecule [29], c[Gly-Gly-D-Ala-D-Ala-Gly-Gly]·$3H_2O$, by use of Equation 1.40 from the Hartree–Fock orbitals for the fully hydrated hexapeptide molecule, the orbitals associated with the $\underset{\sim}{\mathbf{R}}(0)$ matrix obtained from the sum over the $\underset{\sim}{\mathbf{R}}(0)$ for the six kernels, and the orbitals associated with the $\underset{\sim}{\mathbf{R}}(0)$ matrix obtained from the Clinton iterative equations. The three types of density appear to be quite similar. Therefore, only that for the Hartree–Fock orbitals at an isodensity surface of 0.23 e Å^{-3} is shown in Figure 1.6a. To obtain a more quantitative insight into the similarity among the three densities, a series of difference isodensity surfaces were calculated in which differences that did not exceed increasingly larger values were omitted.

Figure 1.6b and c shows the difference isodensity surfaces.

Figure 1.6b and c were obtained from $\underset{\sim}{\mathbf{R}}_{\mathrm{HF}} - \underset{\sim}{\mathbf{R}}_{\mathrm{K}}(0)$ and $\underset{\sim}{\mathbf{R}}_{\mathrm{HF}} - \underset{\sim}{\mathbf{R}}_{\mathrm{P}}(0)$, respectively, by use of Equation 1.40 where the subscripts imply Hartree–Fock (HF), a sum over kernels (K), and a projector ($\underline{\mathbf{P}}$) with a more accurate projector property obtained by

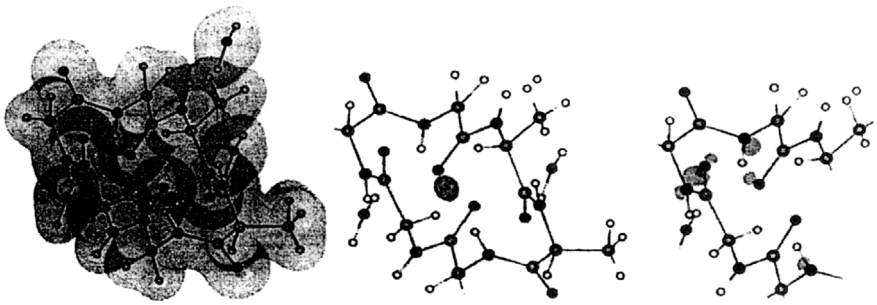


Figure 1.6 (a) Isodensity surface of 0.023 e Å^{-3} for the cyclic hexapeptide trihydrate. (b) Difference isodensity surface of 5×10^{-4} e Å^{-3} for the cyclic hexapeptide trihydrate. The difference isodensity was obtained from $R_{\mathrm{HF}} - R_{\mathrm{K}}(0)$. The small fuzzy region near the center of the diagram, representing the remaining difference isodensity surface, encloses a very small fraction of the molecular volume. (c) Difference isodensity surface of 3×10^{-3} e Å^{-3} for the cyclic hexapeptide trihydrate. The difference isodensity was obtained from $R_{\mathrm{HF}} - R_{\mathrm{p}}(0)$. The small fuzzy regions near the ring represent the difference isodensity surface. They are close to disappearing, enclosing a very small fraction of the molecular volume.

use of Equations 1.41–1.43. For the case of the sum over kernels, a difference isodensity surface of 5×10^{-4} e Å^{-3} is indicated in Figure 1.6b by the small fuzzy region in the center. It encloses a very small fraction of the molecular volume in the center of the molecular framework. The isodensity surface disappears entirely somewhere between 10^{-4} and 10^{-3} e Å^{-3}. For the case of $\underset{\sim}{\mathbf{R}}(0)$ enhanced to form a more accurate projector matrix, the differences are somewhat larger before they disappear, that is, somewhere between 10^{-3} and 10^{-2} e Å^{-3}. The isodensity surface of 3×10^{-3} e Å^{-3} encloses a very small fraction of the molecular volume, as indicated in Figure 1.6c by the tiny fuzzy regions near the ring.

These difference studies indicate that the Hartree–Fock density is closely approximated by the density obtained from the sum over kernels and by the density obtained from enhancing the projector property. This indicates that it is possible to find a projector, $\underset{\sim}{\mathbf{P}}(0) = \underset{\sim}{\mathbf{R}}(0)\,\underset{\sim}{\mathbf{S}}(0)$, and thus an N-representable matrix of the same simplified form as that of the sum over kernels that gives a good approximation to the Hartree–Fock matrix.

1.3.4.2 Hydrated Leu^1-Zervamicin

Fragment Calculations The fragment calculations for the Leu^1-zervamicin [30] (Figure 1.7a) were performed by defining 19 kernels as the 16 peptide residues,

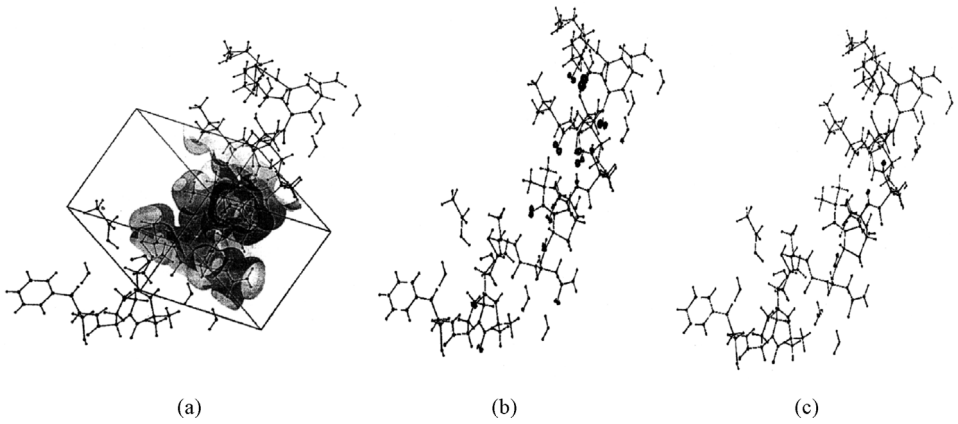


Figure 1.7 (a) Isodensity surface of 0.005 e Å^{-3} within a selected volume for hydrated Leu^1-zervamicin. This isodensity surface was obtained from a Hartree–Fock calculation performed on the entire hydrated molecule and the further use of $\phi_i = \sum_{j=1}^{835} C_{ij}\psi_j$ and $\varrho = 2\sum_{i=1}^{524} \phi_i\phi_i$ applied to the resulting wavefunctions. A ball-and-stick model of the molecule is superimposed on the isodensity surface. (b) Difference isodensity surface of 1.0×10^{-3} e Å^{-3} for the hydrated Leu^1-zervamicin molecule. The difference isodensity surface was obtained from $P_F - P_K$. Small fuzzy regions represent the remaining difference. They involve a small molecular volume and are evidently small in magnitude. (c) Difference isodensity surface of -1.2×10^{-3} e Å^{-3} for the hydrated Leu^1-zervamicin molecule. The difference isodensity surface was obtained from $P_F - P_K$. Small fuzzy regions represent the remaining difference. They involve a small molecular volume and are evidently small in magnitude.

two clusters of water molecules and a cluster of a water and an ethanol molecule. In this application, the neighborhoods were formed with atoms within 5 Å of the kernels plus some few additions to assure that all electrons were paired and the number of electron pairs was even. Table 1.3 lists the kernels and their neighborhoods. The numbers refer to the peptide residues in the sequence Ac-Leu-Ile-Gln-Iva-Ile-Thr-Aib-Leu-Aib-Hyp-Gln-Aib-Hyp-Aib-Pro-Phol (Aib: α-aminoisobutyric acid; Iva: isovaline; Hyp: 4-hydroxyproline; Phol: phenylalininol) description of the chemical content of the Leu^1-zervamicin molecule. The symbols in Table 1.3 correspond to those found in the crystal structure analysis [30]. The crystal structure analysis provided the atomic coordinates used in the calculations reported and also afforded the information from which the selection of the associated solvent molecules was based. The last four columns of Table 1.3 show the number of atoms and the number of basis functions for each kernel and for each neighborhood. Each row of Table 1.3 can be considered to symbolize one individual kernel-neighborhood-fragment calculation. All the calculations of all the rows can be run in parallel on modern supercomputers. The natural parallelization of the calculations is one of the computational advantages of the KEM.

With atomic positions held fixed, the electron density distribution in a fragment is computed. Such a calculation delivers contributions to an $\underset{\sim}{\mathbf{R}}$ matrix and an $\boldsymbol{S}$ matrix from which the portion that concerns the kernel is saved in the form of $\underset{\sim}{\mathbf{P}}_k = \underset{\sim}{\mathbf{S}}_k^{1/2}\underset{\sim}{\mathbf{R}}_k\underset{\sim}{\mathbf{S}}_k^{1/2}$ where the subscript k refers to a kernel matrix. The elements that are saved in a kernel projector matrix are described as follows. Those contributions to the $\underset{\sim}{\mathbf{P}}$ matrix involving orbitals from a neighborhood atom and an atom in the kernel are saved at the fractional value of one-half. If all neighborhood atoms occur only once as part of a kernel, another one-half value would be added to those contributions to the $\underset{\sim}{\mathbf{P}}$ matrix already saved at one-half values, when the values associated with the adjoining kernels are calculated. Contributions from pairs of atoms, both in the same kernel, are saved with a coefficient of 1.

In our previous example (the cyclic hexapeptide trihydrate) we saved the $\underset{\sim}{\mathbf{R}}_k(0)$ instead of the $\underset{\sim}{\mathbf{P}}_k(0)$ and obtained an $\underset{\sim}{\mathbf{R}}(0)$ matrix for the full molecule by combining all the $\underset{\sim}{\mathbf{R}}_k(0)$ for the various kernels. The $\underset{\sim}{\mathbf{P}}_{\mathbf{a}}(0)$ matrix was then obtained by multiplying the $\underset{\sim}{\mathbf{R}}(0)$ matrix by $\underset{\sim}{\mathbf{S}}(0)$ according to Equation 1.22. The $\underset{\sim}{\mathbf{P}}_k$ are saved here instead and in symmetric form. The $\underset{\sim}{\mathbf{P}}_k$ are very good kernel representations and lead to a full $\underset{\sim}{\mathbf{P}}_{\mathbf{s}}$ matrix that is a very good projector, an improvement on $\underset{\sim}{\mathbf{P}}_{\mathbf{a}}(0)$. For the hexapeptide in Section 1.3.4.1, it was possible to obtain good results by defining the single adjacent peptide residue on both sides of a kernel residue as a neighborhood. As a consequence of the denser packing of residues in Leu^1-zervamicin, more residues were required to form the neighborhoods of each kernel.

The matrix $\underset{\sim}{\mathbf{S}}$, defined in Equation 1.20, is a matrix representing the overlap integrals of pairs of orbitals. For pairs of orbitals belonging to atoms that are separated by large distances, the values of the overlap integrals will be close to zero. This behavior of $\underset{\sim}{\mathbf{S}}$ is the reason why the fragment calculations can give accurate values for the elements of $\underset{\sim}{\mathbf{P}}$ for the molecule as a whole.

Table 1.3 Composition of the 19 fragments, that is, kernels and their corresponding neighborhoods, used in the calculation of the **P**-matrix for the hydrated hexadecapeptide, Leu[1]-zervamicin.[a)]

		Kernel		Neighbors	
Kernel	**Neighborhood**	**No. of atoms**	**No. of basis functions**	**No. of atoms**	**No. of basis functions**
1 (Ac-Leu)	2,3,4,5, H6a	27	69	99	275
2 (Ile)	1,3,4,5,6, H7a, Wb3	19	51	124	348
3 (Gln)	1,2,4,5,6,7, N8, H8a Wb2,Wa3, W4	17	53	147	403
4 (Iva)	1,2,3,5,6,7,8, H9a, Wb2	16	44	148	412
5 (Ile)	1,2,3,4,6,7,8,9, H1Of	19	51	159	447
6 (Thr)	O1,2,3,4,5,7,8,9,10, Wb2, Wa3, Wb3	14	42	158	446
7 (Aib)	O2,3,4,5,6,8,9,10, H11e, H11 h, Wb2, Wa3	13	37	153	441
9 (Leu)	O3,4,5,7,9,10,11,12	19	51	143	411
9 (Aib)	O4,5,6,7,8,10,11,12,13	13	37	164	465
10 (Hyp)	O5,6,7,8,9,11,12,13,14, Wa1, Wa2	15	47	142	414
11 (Gln)	O7,8,9,10,12,13,14,15, Wa1, Wa2	17	53	164	480
12 (Aib)	8,9,10,11,13,14,15, H16a, H16c, EtOH, Wa1	13	37	141	401
13 (Hyp)	9,10,11,12,14,15,16, EtOH, Wa1, Wa2, W8	15	47	142	410
14 (Aib)	10,11,12,13,15,16, Wa1, Wa2	13	37	117	345
15 (Pro)	11,12,13,14,16	14	42	143	411
16 (Phol)	O12,13,14,15, EtOH, W8	23	67	112	311
EtOH, W8	12,13,16	12	28	116	324
Wa1,Wa2	10,11,12,13,14	6	14	85	253
Wb2,Wa3,Wb3,W4	2,3,4,6,7	12	28	145	411

a) The individual numbers in the first column, associated with the 16 sequential peptide residues, imply the same corresponding residues in column 2. Other numbers in column 2 have letters with them, for example, H for hydrogen, O for oxygen, N for nitrogen, and W for water. EtOH symbolizes ethanol. The structural aspects of these symbols are to be found in Reference [30].

The Hartree–Fock calculations for the entire hydrated Leu^{l}-zervamicin molecule and for the separate fragment calculations were made by use of the Gaussian 94 program [31] employing the STO-3G basis.

Comparison of Electron Densities Isodensity surfaces have been calculated by use of $\phi_i = \sum_{j=1}^{835} C_{ij}\psi_j$ and $\varrho = 2\sum_{i=1}^{524} \phi_i\phi$ from the Hartree–Fock orbitals for the hydrated Leu^{l}-zervamicin molecule. Use was made of the $\underset{\sim}{\mathbf{P}}$ matrix obtained at once for the full molecule and that obtained from the fragment calculations. The two sources gave electron densities that appeared to be quite similar. Therefore, a portion of the one obtained from the full molecule calculation, as a representative of both types of calculation, is illustrated in Figure 1.7a, at an isodensity surface of 0.005 e Å^{-3}. Confinement of the computed volume to a region of interest, as illustrated, saves time and memory when desirable or necessary. A ball-and-stick model of the structure is superimposed. To obtain a more quantitative insight into the similarity of both types of density calculation, a series of difference isodensity surfaces were calculated in which differences that did not exceed increasingly larger values were omitted. Evidently, this is a calculation that can determine and locate the largest differences between the electron densities.

The difference isodensity surfaces shown in Figure 1.7b and c were obtained from $\underset{\sim}{\mathbf{P}}_{\mathrm{F}} - \underset{\sim}{\mathbf{P}}_{\mathbf{K}}$, where the subscripts imply full molecule (F) and sum over kernel (K) matrices. A difference isodensity surface is shown at 1.0×10^{-3} e Å^{-3} in Figure 1.7b, and -1.2×10^{-3} e Å^{-3} in Figure 1.7c. Some small fuzzy regions are visible at which there are differences as large as, or larger than, the values of the difference isodensities shown. The fuzzy regions should all disappear at slightly larger difference isodensities. Evidently, the fuzzy regions in Figure 1.7b and c are quite small and highly localized, indicating that the electron density is well represented by the $\underset{\sim}{\mathbf{P}}$ matrix obtained from the fragment calculations.

Comments Regarding Kernels and Quantum Crystallography We have presented the basic ideas of quantum crystallography. This entails the treatment of the X-ray scattering experiment in a manner consistent with the requirements of quantum mechanics. In particular, the electron density must be *N*-representable, that is, obtainable from an antisymmetric wavefunction. We indicate how the projector matrix is ensured to be single-determinant *N*-representable by imposition of the condition that it be a hermitian, normalized projector. By adopting the approximation that a full molecule can be "broken" into smaller "fragments," consisting of a "kernel" of atoms and its "neighborhood" of atoms, a simplified representation is obtained that reduces the number of parameters required. The kernels are each "extracted" from their fragments by rules patterned upon those of Mulliken population analysis. An approximate matrix for the full molecule is reconstructed by summing over the kernel matrices and imposing the projection property.

The virtue of introducing the concept of kernel matrices is that their use could allow very large molecules to be studied within the context of quantum crystallography. The fundamental feature that explains the applicability of the kernel approximation, as it is applied here, is the vanishing of orbital overlap as the distance

between orbital centers increases. This has the consequence that elements of the matrix **R** that weight the relative importance of such vanishing overlap contributions to the density may be neglected without affecting the density. Thus, a pattern of zeros is introduced into the matrix $\underline{\mathbf{R}}(0)$, which defines the size of fragments that, in general, will be smaller than the full molecule. The fragments of reasonable size contain the essential information for determining the matrices for kernels. Reconstruction of the full matrix in an approximation that can deliver a good density follows and the projection property maintains the structure of quantum mechanics. The formalism is flexible enough that all the electronic and atomic structural variables may be refined by least-squares methods.

The hexapeptide molecule of this chapter was treated within the context of the *ab initio* Hartree–Fock approximation. However, we point out that the concept of extracting kernel matrices from fragments smaller than a full molecule would be applicable within the context of any method based upon a molecular orbital representation, including extended Hückel, empirical Hartree–Fock, configuration interaction and density functional methods. Our initial exploration of other MO methods bears this out.

1.4
Kernel Density Matrices Led to Kernel Energies

Although our initial interest in the kernel neighborhood fragment approximation of the density matrix concerned its applications within quantum crystallography, we also indicate that it should be useful in the purely quantum mechanical problem of solving the Schrödinger equation. These concepts led us to calculations of kernel energies. Following that the kernel energy method (KEM) evolved, which we now discuss.

Given that the problem of large molecule interactions would be interesting to study by use of the techniques of quantum mechanics, the problem they present is often the considerable size of targets composed of, for example, proteins, DNA, RNA, and so on. That problem is addressed here by using the KEM approximation, whose main features are now reviewed.

In the KEM, the results of X-ray crystallographic coordinates are combined with those of quantum mechanics. This leads to a reduction of computational effort and an extraction of quantum information from the crystallography. Central to the KEM is the concept of the kernel. These are the quantum pieces into which the full molecule is mathematically broken. All quantum calculations are carried out on kernels and double kernels. Because the kernels are chosen to be smaller than a full biological molecule, the calculations are accomplished efficiently, and the computational time is much reduced. Subsequently, the properties of the full molecule are reconstructed from those of the kernels and double kernels. Thus a quantum realization of the aphorism that the whole is the sum of its parts is obtained.

It is assumed that the crystal structure is known for a molecule under study. With known atomic coordinates, the molecule is mathematically broken into tractable

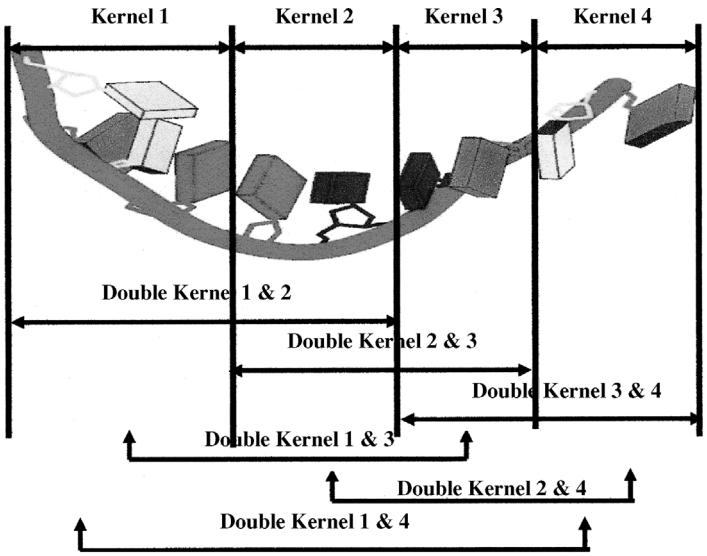


Figure 1.8 Abstract sketch of RNA showing the definitions of the single and double kernels.

pieces called kernels. The kernels are chosen such that each atom occurs in only one kernel. Figure 1.8 shows schematically defined kernels and double kernels, and only these objects are used for all quantum calculations.

The total molecular energy is then reconstructed by summation over the contributions of the double-kernels reduced by those of any single kernels that have been over counted. Two approximations have been found to be useful. In the simpler case, only the chemically bonded double kernels are considered, and the total energy E in this approximation is:

$$E_{\text{total}} = \sum_{i=1, j=i+1}^{n-1} E_{ij} - \sum_{i=2}^{n-1} E_i \tag{1.44}$$

E_{ij} = energy of a chemically bonded double kernel of name ij

E_i = energy of a single kernel of name i

i, j = running indices

n = number of kernels.

In the more accurate case, all double kernels are included, and the total energy is:

$$E_{\text{total}} = \sum_{m=1}^{n-1} \left(\sum_{\substack{i=1 \\ j=i+m}}^{n-m} E_{ij} \right) - (n-2) \sum_{i=1}^{n} E_i \tag{1.45}$$

E_{ij} = energy of a double kernel of name ij

E_i = energy of a single kernel of name i

$i, j, m =$ running indices

$n =$ number of single kernels.

The purpose of the calculations is to obtain kernel contributions to the energy when it is not computationally feasible to treat the entire molecule as a whole. When a structure of interest has known crystallographic coordinates one may easily define kernels, which altogether represent the entire composite molecule. The use of the single kernels and double kernels indicated above is an approximation that is made to obtain a simplification in the quantum calculation. The validity of this approximation, in the case of various peptides, proteins, DNA and RNA structures, is shown in various works discussed below.

1.4.1 KEM Applied to Peptides

Molecules of biological importance have been chosen for the calculation of molecular energy using the concepts of single kernel and double kernel in the KEM [32]. The examples chosen were sufficiently large to provide significant demonstrations of *ab initio* energy calculations using the kernel energy method, but not so large as to prevent energy calculations of whole molecules using supercomputers. The latter cases were required to provide a standard of excellence against which the approximations using kernels could be judged.

The group of peptides that were selected is shown with their crystal structure geometry in Figure 1.9. Peptides, of course, are of vast biological importance, having the capacity to control many crucial functions of an organism, including cell reproduction, immune response, appetite, and so on. The human organism makes a great many peptides that act as neurotransmitters, hormones and antibiotics. Synthetic peptides are studied as possibly effective drugs. The fundamental biological activity of peptides depends upon their conformation, which is, in turn, determined by the energy of the conformation. Thus, the ability to calculate the molecular energy associated with peptide structure is basic to the study of peptides and their function. In this section, we show how the concept of kernels allows for accurate calculation of peptide energy. All of their crystal structures are known [30, 33–42] and have been used in the energy calculations presented here. Figure 1.9 illustrates various natural and synthetic peptides that vary in size, shape and function.

Table 1.4 shows the energies obtained with Equation 1.44 for 16 different peptides. For one of these, Leu^1-zervamicin [30], we calculated the energy for two different conformations, labeled closed and open. The number of atoms and amino acids in the table range from a minimum of 80 atoms contained in six amino acids to a maximum of 327 atoms contained in a 19 amino acid chain. All energy calculations correspond to the Hartree–Fock approximation using a minimal STO-3G basis, and the effects of solvent were not considered.

The results of Table 1.4 all correspond to a kernel size defined as one amino acid. The KEM requires much less calculation time than would be the case for the full molecule Hartree–Fock calculation in the same basis set without approximation.

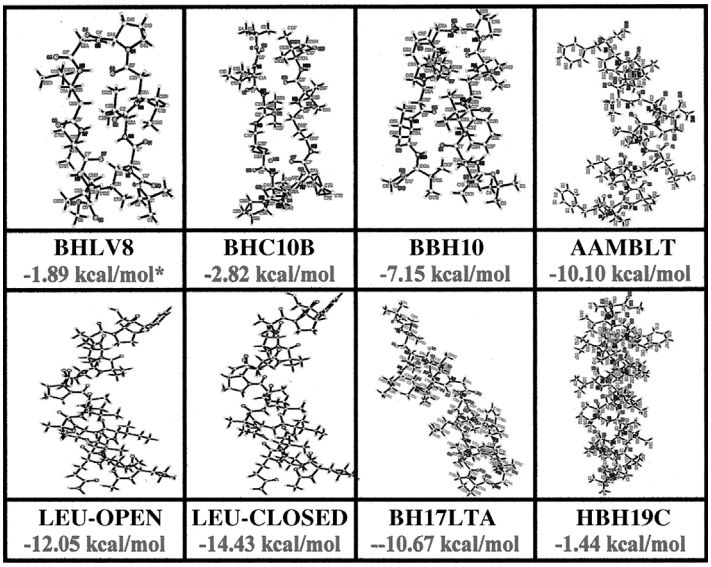


Figure 1.9 Peptide structures from X-ray crystallography. *The energy differences $E_{HF} - E_{KEM}$ are from Equation 1.45, which includes all the double kernels.

A distinction has been made between two approximations. In the first, with the use of Equation 1.44, energy contributions are considered only from those double kernels composed of chemically bonded pairs of single kernels, as in the results of Table 1.4. In the second, using Equation 1.45, energy contributions are considered from all double kernels, whether or not they are composed of chemically bonded single kernels. Our anticipation was that including all double kernels would increase the accuracy of the KEM results. Also, with the use of Equation 1.45, if the size of the kernels were increased, it was presumed that it would also increase the accuracy of the KEM approximation. In most cases the use of the Equation 1.44 approximation, as seen in Table 1.4, is fairly accurate. The worst case occurs for HBH19C [34] in 20 kernels for which the difference is 223 kcal mol^{-1} (1 kcal = 4.184 kJ) out of a total exact value of 6748 au (4 234 370 kcal mol^{-1}) representing about a 0.0053% difference (au stands for "atomic units"). Apparently the approximation based upon the kernels of small size (one amino acid), and including only the chemically bonded double kernels, is a reasonable one.

If the approximation including all double kernels is applied, an increased accuracy is obtained [32]. Also, as is physically reasonable, as the kernel size increases and all double kernels are considered in the calculation, the errors should decrease, as does occur. Thus, judged by the results of peptides represented in Table 1.5, the energy approximations of Equations 1.44 and 1.45 have good accuracy.

The computational results indicate that the kernel energy method is worthwhile. It has yielded results that have small differences. Sixteen calculations have been tabulated for various peptides that have a range of geometries from 4 to 19 residues

Table 1.4 Energy calculation for peptides[a], using Equation 1.44.

Energy	BMA4	ISARAM	ISARIAX	ALAC7ALT	ADPGV7b	BHF4LT	BDPGV7A	BH2L2
Atoms (kernels)	80 (6)	104 (6)	107 (6)	125 (7)	126 (7)	134 (7)	142 (8)	144 (9)
E_{HF} (au)	−1781.44	−2274.28	−2312.95	−2522.80	−2539.50	−3006.70	−2805.72	−2970.77
E_{KEM} (au)[b]	−1781.43	−2274.27	−2312.94	−2522.79	−2539.50	−3006.69	−2805.71	−2970.73
$E_{HF} - E_{KEM}$ (kcal mol^{-1})	−1.82	−8.16	−6.90	−5.02	−3.39	−8.28	−5.02	−21.46
Energy	**BHLV8**	**BHC10B**	**BBH10**	**AAMBLT**	**Leu-open**	**Leu-closed**	**BH17LTA**	**HBH19C**
Atoms (Kernels)	150 (9)	164 (11)	190 (11)	246 (16)	265 (16)	265 (16)	269 (17)	327 (20)
EHF (a.u.)	−3047.10	−3528.00	−3986.31	−5529.32	−5849.50	−5851.57	−5800.36	−-6748.41
$^{*}E_{KEM}$(a.u.)	−3047.07	−3527.96	−3986.28	−5529.26	−5849.44	−5851.50	−5800.32	−6748.05
E_{HF} -E_{KEM} (kcal mol^{-1})	−15.81	−21.90	−18.39	−33.89	−39.28	−43.55	−24.85	−222.64

a) KEM applied to peptides using Equation 1.44, with 1 kernel ≡ 1 amino acid. (Calculations were performed without solvent at the HF/STO-3G level of theory.)
b) Including only double kernels composed of single kernel pairs chemically bonded to one another, Equation 1.44.

and atoms numbering from 80 to 327. The total energy range is 1781–6748 au. The differences for the energies are quite small as a percentage of the total energy. The total energy was calculated by summing the energies of double kernels. In so doing, the contribution of some single kernels are counted twice and thus the contribution of over-counted single kernels must be subtracted from the total.

The basic assumption is that the energy of any given kernel is most affected by its own atoms and those of the neighboring kernels with which it interacts. A pair of interacting kernels forms a double kernel. Perhaps the most important double kernels are those formed of chemically bonded single kernels. Thus kernels and double kernels are used to define the energy of the full molecule as in Equations 1.44 and 1.45. As a molecule grows in size there are more double kernels and more single kernels, but the basic formula is the same. The total energy is a sum of contributions of double kernels reduced by single kernels that have been over-counted. Tables 1.4 and 1.5 show the energy is well represented by the above kernel energy method. The fragment calculations are carried out on double kernels and single kernels whose ruptured bonds have been mended by attachment of H atoms. A satisfactory occurrence in the summation of energies is that the total contribution of hydrogen atoms introduced to saturate the broken bonds tends to zero. The effect on the energy of the hydrogen atoms added to the double kernels effectively cancels that of the hydrogen atoms added to the pure single kernels that enter with opposite sign.

There are, of course, limitations to the accuracy of the KEM. The basic assumption is that the total energy can be built up so long as the atoms of one kernel are mainly affected by themselves and those of neighboring kernels. The tabulated calculations show that the most important double kernels are those composed of pairs of single kernels that are chemically bonded to one another. For best accuracy, however, all double kernels are calculated.

The effect of kernel size on the accuracy of the energy has been considered. In our calculations, increasing kernel size improves the accuracy of energy results. Based upon the peptides calculated thus far we conclude that increasing kernel size reduces the already small difference that occurs when the size of a kernel is specified to be the size of one amino acid. Including all double kernels gives the smallest difference.

The times for entire molecular calculations have been compared to those based upon Equation 1.44. In Figure 1.10 the full molecule Hartree–Fock case has been fit to a fourth power polynomial, and the results based upon Equation 1.44 have been fit to a linear expression. Clearly, the approximation of Equation 1.44 saves computing time. When the two curves are extrapolated beyond the computational data points represented by Table 1.6, the discrepancy between fourth and first power grows. The main diagram in Figure 1.10 plots the projected times shown in Table 1.7. With 1000 atoms, the computing time for an entire molecule is about 13 hours, and the computing time for the KEM is about 18 minutes. At 10 000 atoms the computing time for an entire molecule is about 145 days, and the computing time for the KEM is about 3.5 hours. The use of the KEM with Equation 1.44 applied to peptides gives good accuracy at a significant saving of computing time. This augers well for application of the same method to even larger molecules.

Table 1.5 Energy calculation for peptides, using Equations 1.44 and 1.45.[a)]

Energy	BMA4	ISARAM	ISARIAX	ALAC7ALT	ADPGV7b	BHF4LT	BDPGV7A	BH2L2
Atoms (kernels)	80 (4)	104 (3)	107 (3)	125 (4)	126 (4)	134 (4)	142 (4)	144 (4)
E_{HF} (au)	−1781.44	−2274.28	−2312.95	−2522.80	−2539.50	−3006.70	−2805.72	−2970.76
E_{KEM} (au)[b)]	−1781.43	−2274.27	−2312.94	−2522.79	−2539.49	−3006.69	−2805.71	−2970.75
Difference (kcal mol^{-1})	−4.83	−3.51	−1.26	−5.08	−7.84	−4.96	−7.03	−7.47
E_{KEM} (au)[c)]	−1781.43	−2274.27	−2312.94	−2522.79	−2539.50	−3006.70	−2805.71	−2970.76
$E_{HF} - E_{KEM}$ (kcal mol^{-1})	−2.82	−3.51	−1.26	−3.14	−3.20	−2.70	−3.51	−3.20
Energy	**BHLV8**	**BHC10B**	**BBH10**	**AAMBLT**	**Leu-open**	**Leu-closed**	**BH17LTA**	**HBH19C**
Atoms (kernels)	150 (4)	164 (5)	190 (6)	246 (6)	265 (7)	265 (7)	269 (7)	327 (3s)
E_{HF} (au)	−3047.10	−3528.00	−3986.31	−5529.32	−5849.50	−5851.57	−5800.36	−6748.41
E_{KEM} (au)[b)]	−3047.08	−3527.98	−3986.30	−5529.29	−5849.46	−5851.52	−5800.34	−6748.41
Difference (kcal mol^{-1})	−10.42	−10.35	−12.11	−16.50	−28.30	−29.74	−14.37	−2.63
E_{KEM} (au)[c)]	−3047.09	−3527.99	−3986.30	−5529.30	−5849.48	−5851.55	−5800.34	−6748.41
$E_{HF} - E_{KEM}$ (kcal mol^{-1})	−1.89	−2.82	−7.15	−10.10	−12.05	−14.43	−10.67	−1.44

KEM applied to peptides using Equations 1.44 and 1.45, and with kernel sizes larger than one amino acid. (Calculations were performed without solvent at the HF/STO-3G level of theory.)
The only double kernels included are those made of single kernel pairs that are chemically bonded to one another, Equation 1.44.
All double kernels are included, Equation 1.45.

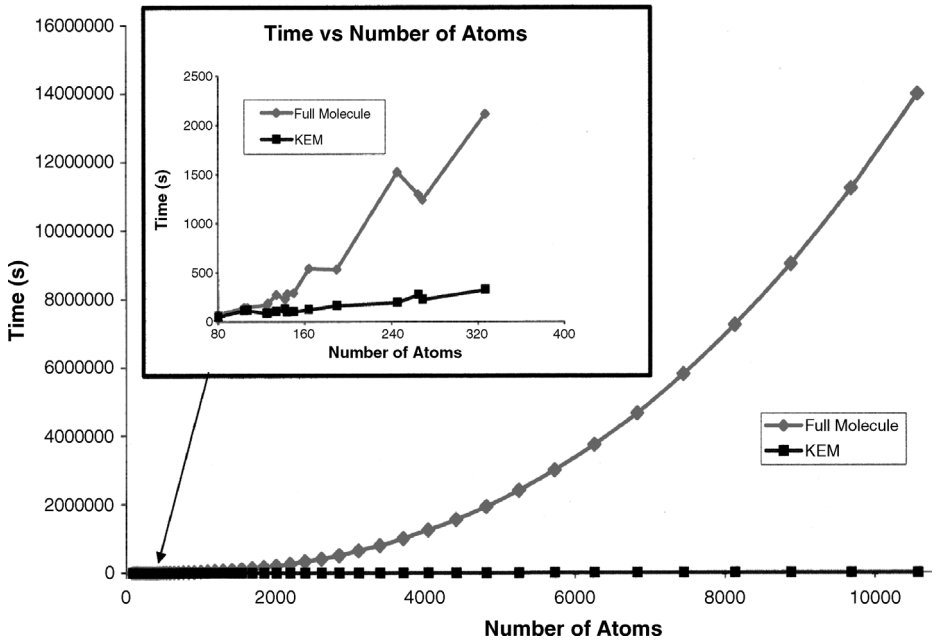


Figure 1.10 Calculation time comparison of full molecule versus KEM. Inset: actual calculation time data for the molecules of Table 1.6. Main figure: projected times obtained from a fourth-order polynomial fit to the HF calculation times for full molecules, and a linear function fit to the KEM calculation times for the same molecules (Table 1.7).

With the use of the known structures of peptides, from crystal structure analysis, it has been shown that it is feasible to make *ab initio* quantum mechanical calculations to good approximation for very large molecules, employing the notion that the whole may be obtained from its parts. In our procedure the parts are the quantum mechanical kernels. The key to such computations is the fragment calculation wherein a molecule is divided into kernels and *ab initio* calculations are performed on each of the kernel fragments and double kernel fragments. The results of our calculations suggest that the larger the kernels the greater the relative accuracy.

1.4.2 Quantum Models within KEM

A model chemistry specifies a quantum method of calculation and a set of basis functions. Given the computational advantages alluded to above, the question arises: What is the effect of the choice of basis functions and quantum methods on the KEM approximation [43] All the previous calculations used to test the approximation were, in the first instance, for reasons of simplicity, based on the use of STO-3G basis functions and HF calculations. It is therefore reasonable to wonder whether the approximation will work equally well with another choice of model chemistry.

Table 1.6 Calculation time (in seconds) for peptides.[a)]

Time (s)	BMA4	ISARAM	ISARIAX	ALAC7ALT	ADPGV7b	BHF4LT	BDPGV7A	BH2L2
Atoms (kernels)	80 (6)	104 (6)	107 (6)	125 (7)	126 (7)	134 (7)	142 (8)	144 (9)
$t_{\text{full-molecule}}$	67	143	148	171	193	276	236	284
t_{KEM}[b)]	51	112	118	85	91	106	127	96
Time (s)	**BHLV8**	**BHC10B**	**BBH10**	**AAMBLT**	**Leu-open**	**Leu-closed**	**BH17LTA**	**HBH19C**
Atoms (kernels)	150 (9)	164 (11)	190 (11)	246 (16)	265 (16)	265 (16)	269 (17)	327 (20)
$t_{\text{full-molecule}}$	296	546	534	1529	1300[c)]	1300[c)]	1241	2122
t_{KEM}[b)]	102	126	165	196	274[c)]	274[c)]	226	327

a) The same supercomputer and the same number of parallel nodes were employed for all calculation times shown here. All energy calculations are in the approximation HF/STO-3G.
b) Only chemically bonded double kernels are included, Equation 1.44; 1 kernel ≡ 1 amino acid.
c) Average time of Leu-open and Leu-closed calculations.

Table 1.7 Comparison between the estimated calculation times for the full molecule and for the KEM.[a)]

No. atoms	Full molecule in hours	KEM in hours
86	0.022	0.019
94	0.028	0.021
103	0.035	0.023
112	0.043	0.025
122	0.054	0.027
133	0.067	0.030
145	0.083	0.033
159	0.104	0.036
173	0.129	0.040
189	0.161	0.044
206	0.201	0.048
225	0.250	0.053
246	0.311	0.059
268	0.387	0.064
293	0.482	0.071
320	0.601	0.078
349	0.748	0.086
381	0.932	0.094
416	1.160	0.104
454	1.445	0.114
495	1.799	0.126
540	2.240	0.138
590	2.790	0.152
644	3.474	0.167
702	4.327	0.184
767	5.388	0.202

Table 1.7 (*Continued*)

No. atoms	Full molecule in hours	KEM in hours
837	6.709	0.222
913	8.355	0.245
997	10.405	0.269
1088	12.957	0.296
1187	16.135	0.326
1296	20.093	0.358
1414	25.022	0.394
1544	31.160	0.433
1685	38.803	0.477
1839	48.322	0.524
2007	60.175	0.577
2190	74.935	0.634
2390	93.316	0.698
2609	116.206	0.767
2847	144.711	0.844
3107	180.208	0.929
3392	224.413	1.022
3702	279.460	1.124
4040	348.010	1.236
4409	433.376	1.360
4812	539.681	1.496
5252	672.062	1.645
5732	836.915	1.810
6256	1042.207	1.991
6828	1297.855	2.190
7452	1616.213	2.409
8133	2012.663	2.649
8877	2506.359	2.914
9688	3121.158	3.206
10 574	3886.763	3.526

a) Comparison of times obtained by fitting polynomials to the actual computing time data for the molecules of Table 1.6, for the full molecule calculation to a fourth-order polynomial and for KEM to a linear function.

Because the previous investigation examined such a wide variety of different peptides, in terms of size, shape, and structure, all with positive results, it seems unlikely that the KEM would depend sensitively on a particular choice. However, to preclude that possibility, we examine here the effect of the choice of model chemistry on the applicability of KEM. The issue is whether KEM is more or less independent of a choice of model chemistry. This question is pursued within the context of both (i) the various choices of basis set and (ii) the use of different quantum chemical methods of calculation. For (i) tests of KEM sensitivity to basis functions have been carried out by applying the KEM approximation repeatedly to the same molecule, ADPGV7b (Figure 1.11) [42], which contains 126 atoms, using various basis functions. For each basis, the energy of the full molecule has been calculated and is labeled

$E_{\text{full-molecule}}$. The difference between the full molecule result and that obtained by KEM in the same basis has been examined. It is of interest to know whether the difference depends in a sensitive way on the choice of basis functions. For example, does that energy difference change systematically with the size and quality of the basis set employed? Alternatively, do the errors fluctuate within limits, not correlated to the size and quality of the basis functions used for the calculations? These questions of basis set dependence are examined in the numerical experiments discussed in Section 1.4.2.1.

Inasmuch as the first KEM paper [32] was restricted to calculations within a HF model chemistry, a question arises as to whether applications of KEM will prove to be valid across a whole spectrum of commonly used quantum methods, characterized by differing levels of accuracy. This is answered by choosing a particular peptide as a test case, namely, Zaib4 (which contains 74 atoms), and by calculating its energy with several different quantum chemical methods. These include HF and DFT calculations, but range widely from there. In the direction of more approximate calculations, semiempirical models are used. In the opposite direction of accuracy, to the same test molecule, several higher-level quantum mechanical chemistry models are applied. It is found that KEM is widely applicable across the spectrum of models tested. Thus, in the Zaib4 study, the above formulas were applied in calculating the molecular energy, to test the accuracy of KEM for various basis functions, as well as chemistry models characterized by different levels of accuracy.

1.4.2.1 Calculations and Results Using Different Basis Functions for the ADPGV7b Molecule

It may be shown that the accuracy of KEM does not depend on a particular choice of basis functions. This is done by calculating the ground-state energy of a representative peptide, ADPGV7b, containing seven amino acid residues, using seven different commonly employed basis function sets, ranging in size from small to medium to large. The study of sensitivity of the KEM approximation to choice of basis functions employed the following basis sets: STO-3G [44, 45], 3-21G [46–51], SV [52, 53], 6-31G [54–63], D95 [64], 6-31G* [65, 66] and cc-pVDZ [67–71]. The accuracy of the KEM does not vary in any systematic way with the size or mathematical completeness of the basis set used, and good accuracy is maintained over the entire variety of basis sets tested. We conclude that the accuracy inherent in the KEM is not dependent on a particular choice of basis functions. The first application, to different peptides mentioned above, employed only HF calculations.

The peptide ADPGV7b of known crystal structure [42] is pictured in Figure 1.11 and broken into four single kernels. The amino acid sequence defining the peptide is as follows: Ac-Val-Ala-Leu-Dpg-Val-Ala-Leu-OMe (Dpg = α, α – di-*n*-propyl glycine). Equations 1.44 and 1.45 were applied repeatedly to the calculation of the energy of the peptide ADPGV7b using each of seven different sets of basis functions. This was done in both the HF approximation and the density functional theory (DFT) approximation, using the standard potential B3LYP. The purpose in both cases was to assess whether the accuracy of the KEM was critically dependent on the choice of basis functions.

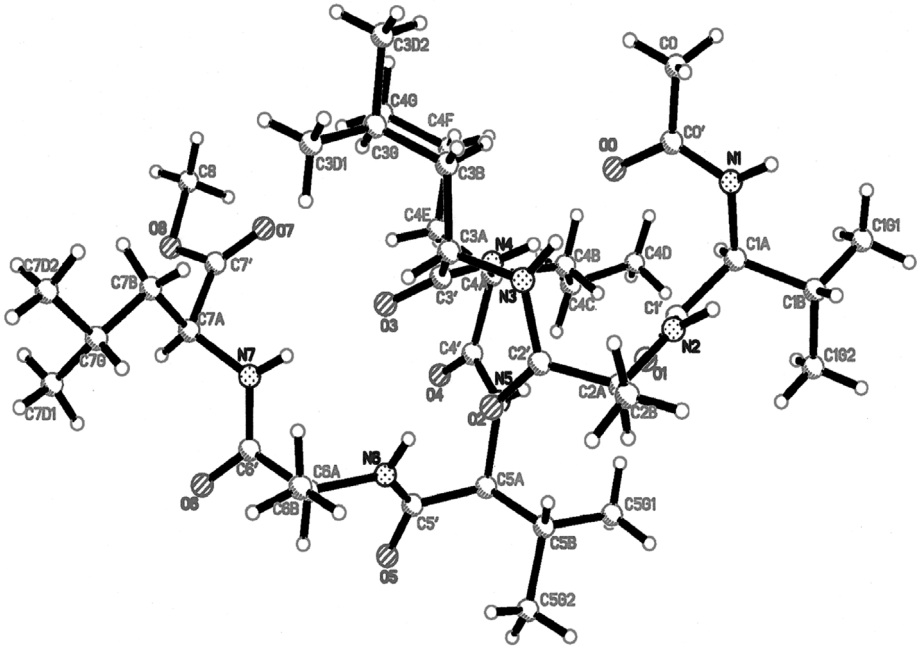


Figure 1.11 ADPGV7b X-ray crystal structure.

Table 1.8 presents the energies obtained with Equation 1.44 for the seven different basis sets. The HF results are in Table 1.8, while the DFT results, which are qualitatively similar, are not shown. The effects of solvent were not considered in this study. All results shown in Table 1.8 correspond to the peptide ADPGV7b, composed of a total of 126 atoms, broken into four kernels. The results for the full molecule calculations are labeled as $E_{\text{full-molecule}}$. A main conclusion that can be drawn from Table 1.8 is that the energy obtained from the KEM is quite accurate for all the basis sets used, and moreover that the accuracy does not correlate in any obvious way with the choice of basis. It may be seen, for example, that the energy differences associated with application of KEM do not correlate with the increasing

Table 1.8 KEM calculation for ADPGV7b, using different basis functions (126 atoms, 4 kernels).

HF/basis	$E_{\text{full-molecule}}$ (au)	E_{KEM} (Equation 1.44) (au)	E_{diff} (kcal mol^{-1})
STO-3G	−2539.5022	−2539.4971	−3.20
3-21G	−2557.8857	−2557.8809	−3.01
SV	−2568.7546	−2568.7475	−4.46
6-31G	−2570.9939	−2570.9872	−4.20
D95	−2571.3472	−2571.3383	−5.58
6-31G*	−2572.1191	−2572.1125	−4.19
cc-pVDZ	−2572.2345	−2572.2285	−3.78

mathematical completeness of the basis used for the energy calculation. The same is equally true for the DFT results and the HF results.

Now a distinction is made between two approximations. In the first, with the use of Equation 1.44, energy contributions are considered only from those double kernels composed of chemically bonded pairs of single kernels. In the second, using Equation 1.45, energy contributions are considered from all double kernels, whether or not they are composed of chemically bonded single kernels. As expected, results for the peptide ADPGV7b indicate a general trend in which accuracy is increased when all double kernels are included in the calculation as specified by Equation 1.45. The result is that the already small differences associated with Equation 1.44 are even smaller with the use of Equation 1.45. It is physically reasonable that when all double kernels are considered in the calculation the difference should decrease, as occurs in the tables. However, and this is a main point of interest, the differences associated with the results of Equation 1.45, just as with Equation 1.44, are relatively small and fluctuate rather randomly with the choice of basis set employed in the calculations. This occurs in both the HF and DFT approximations.

1.4.2.2 Calculations and Results Using Different Quantum Methods for the Zaib4 Molecule

The second question (ii) that arises is whether the results obtained with the use of KEM will be accurate only within the HF approximation. Therefore, we also studied whether KEM is applicable across various quantum computational methods, characterized by differing levels of accuracy. The peptide, Zaib4, containing 74 atoms, was used to calculate its energy at seven different levels of accuracy. These include the semi-empirical methods, AM1 and PM5, a DFT B3LYP model, and *ab initio* HF, MP2, CID and CCSD calculations. KEM was found to be widely applicable across the spectrum of quantum methods tested.

The calculations below, which test the sensitivity of the KEM approximation to choice of model accuracy, employ seven different quantum methods as follows: AM1 [72], PM5 [73], HF [74], DFT [75], CID [76], MP2 [77] and CCSD [78]. For this study we have adopted as a test molecule a 74-atom peptide called Zaib4. Figure 1.12 shows a picture of the molecule arising from the X-ray crystal structure. The amino acid sequence defining the Zaib4 peptide is as follows: Z-Aib-Aib-Aib-Aib-OMe. Table 1.9 gives the calculated molecular energy results for the chemistry models tested. All calculations correspond to the crystal structure geometry. The same STO-3G basis functions were used for all *ab initio* quantum mechanical methods listed. $E_{\text{full-molecule}}$ is listed for each chemistry model along the table, and represents the calculated energy of the full molecule taken as a whole without being broken into kernels.

This is the standard of excellence against which KEM results are to be judged. Table 1.9 lists the calculated energies that derive from KEM using the approximations given by Equation 1.45. Also given are the corresponding differences between $E_{\text{full-molecule}}$ and the values calculated with Equation 1.45. Equations 1.44 (not shown) and 1.45 were applied repeatedly to the calculation of the energy of the peptide Zaib4 using each of seven different methods of quantum chemical calculation indicated in

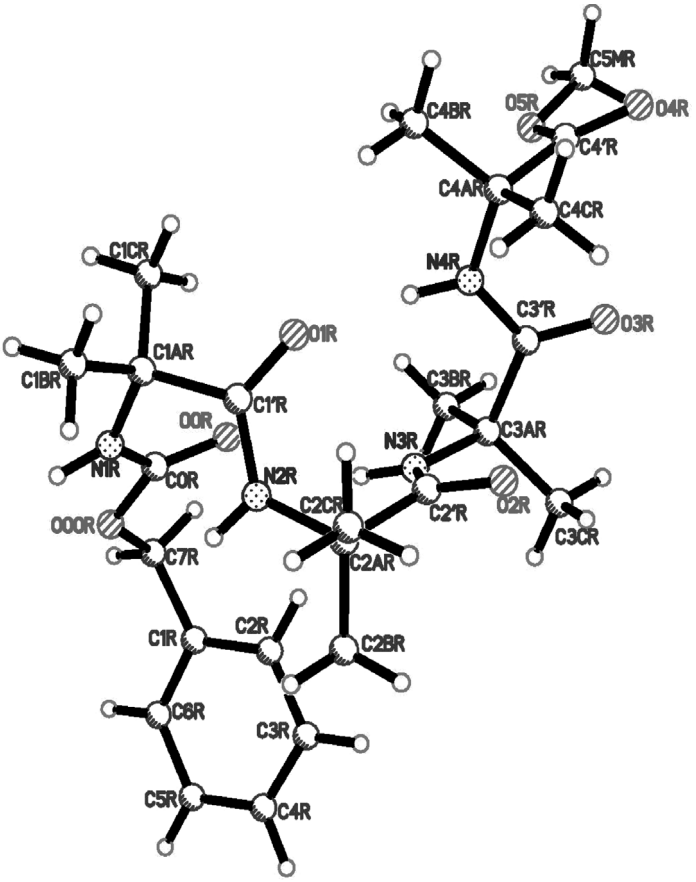


Figure 1.12 Zaib4 X-ray crystal structure.

Table 1.9. The purpose of these calculations was to assess whether the accuracy of KEM was critically dependent on the choice of quantum chemical calculation method employed. Table 1.9 shows the energies obtained with Equation 1.45 for the seven different quantum chemical calculation methods. The main point associated with the

Table 1.9 KEM calculation for Zaib4, using different quantum methods (74 atoms, 3 kernels).

Methods	$E_{\text{full-molecule}}$ (au)	E_{KEM} (Equation 1.45) (au)	E_{diff} (kcal mol^{-1})
AM1[a)]	−248.9642	−248.9619	−1.41
PM5[a)]	−228.0289	−228.0259	−1.88
HF	−1688.4786	−1688.4755	−1.97
B3LYP	−1698.2907	−1698.2870	−2.31
MP2	−1690.2155	−1690.2125	−1.88
CID	−1690.5196	−1690.5094	−6.39
CCSD	−1690.5589	−1690.5564	−1.60

a) Semiempirical methods that consider only the valence electrons.

results of Table 1.9 is that it appears all types of quantum calculations tested, within the limits discussed, are compatible with KEM. The quantum methods displayed in Table 1.9 represent a broad sample of the methodologies commonly used in computational chemistry. Thus, they present a good test of how widely applicable KEM may be for obtaining molecular energies. The numerical values of Table 1.9 indicate that the KEM results are uniformly applicable, for all the model chemistries that have been tested. The errors associated with basing the molecular energy on the approximation related to summing over the kernels, in accordance with Equations 1.44 and 1.45, is generally quite small.

1.4.2.3 Comments Regarding KEM

In judging the accuracy of KEM, the differences of interest are those between the $E_{\text{full-molecule}}$ energy and that predicted by the KEM, both in the same basis set and using the same equations of motion. At least for as the seven basis sets used thus far, it seems that the validity of the KEM approximation does not depend on a particular choice of basis. Therefore, in future applications of KEM, the choice of basis may be made freely, in accordance with those considerations usually apropos of a particular molecular problem, including the absolute accuracy to be achieved, given the computational power, and computational time available, for the task at hand. Turning our attention to the numerical comparisons afforded between the $E_{\text{full-molecule}}$ energies for the various quantum methods and the corresponding energies obtained from KEM approximations, we have seen that they are quite close. It is a favorable result for KEM that it has proved to be applicable with all the quantum methods tested. At least with respect to the limited number of tests that we have been able to carry out, it seems that the validity of KEM will not depend in a sensitive way on either the basis sets or the calculation level of quantum methods used.

1.4.3 KEM Applied to Insulin

1.4.3.1 KEM Calculation Results

An application has been made with the protein insulin [79–81], which is composed of 51 amino acids. Accurate KEM Hartree–Fock energies were obtained for the separate A and B chains of insulin and for their composite structure in the full insulin molecule. A limited basis is used to make possible calculation of the full insulin molecule, which can be used as a standard of accuracy for the KEM calculation.

Insulin is composed of two peptide chains named A and B. The chains are linked by two disulfide bonds, and an additional disulfide is formed within the A chain. The A chain contains 21 amino acids, composed of 309 atoms, including hydrogen, and the B chain contains 30 amino acids, composed of 478 atoms, including hydrogen. Figure 1.13 shows a ribbon diagram of the insulin molecule that gives an impression of the three-dimensional structure of the molecule. The quantum mechanical method chosen for testing the KEM in the case of insulin is that of the Hartree–Fock (HF) equations using atomic orbital basis functions of type STO-3G.

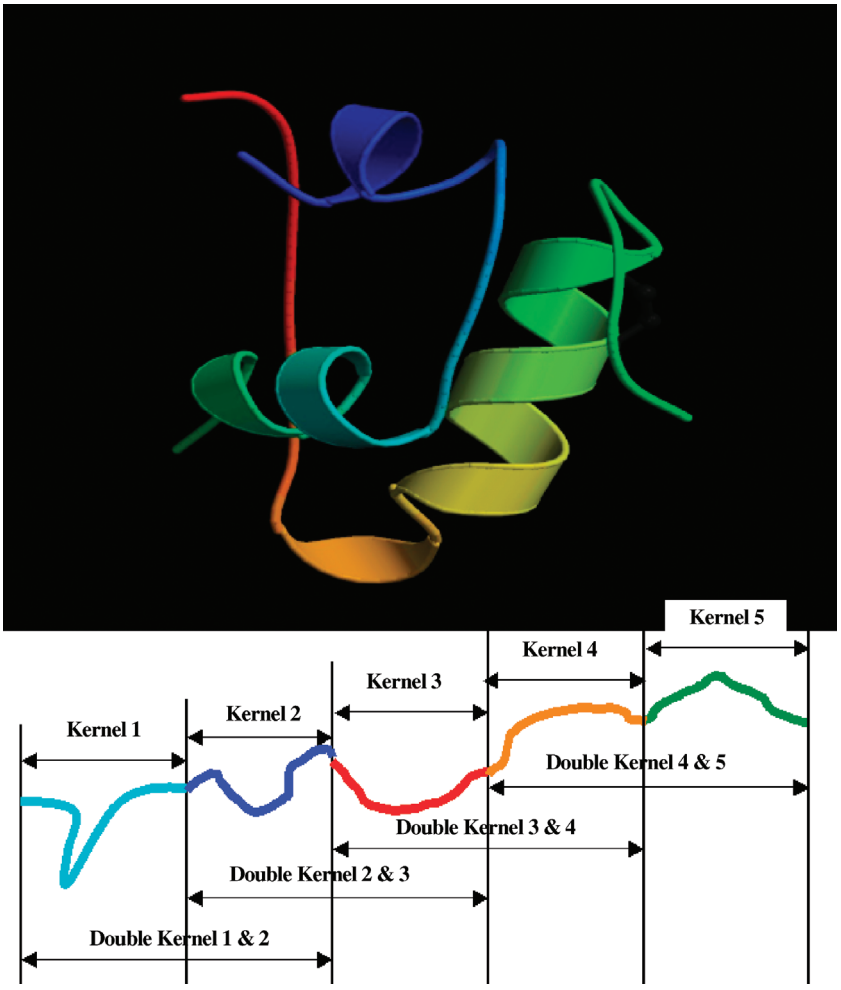


Figure 1.13 The insulin molecule is composed of two chains, A in blue (shown as two shorter helices) and B in green-red (shown as one longer helix). The whole molecule is divided into five kernels as shown. The insulin figure was generated by KING Viewer in the PDB web site.

The full insulin molecule (chains A and B) yields a calculated total energy of $E_{HF} = -21\,104.7660$ au. The KEM result, $E_{KEM} = -21\,104.7656$ au (Equation 1.45), differs from this by as little as 0.0004 au. For all three calculations, that is, chain A, chain B and the complete solvated insulin molecule, the energy differences were calculated corresponding to the full molecule result and its approximation by the KEM. The energy differences are relatively small. In all three cases, the Equation 1.45 differences are less than those of Equation 1.44, and are of magnitude 1 kcal mol^{-1}.

Table 1.10 Energy calculation for solvated insulin.

No. of atoms	No. of kernels	E_{HF} (au)	E_{KEM}[a)] (au)	$E_{HF\text{-}EKEM}$ (kcal mol^{-1})
959	6	−26275.4187	−26275.4127	−3.79

a) KEM calculation with HF/STO-3G, using all the double kernels (Equation 1.45).

Table 1.10 considers the case of the full insulin molecule in the presence of solvent molecules. In the crystal, the solvent molecules are present as 56 H_2O and a single 1,2-dichloroethane. The fully solvated insulin contains a total of 959 atoms. All of the atomic positions of the solvent molecules together with those of the full insulin have been determined crystallographically (Protein Data Bank, PDB ID code 1APH), except for hydrogen atoms, which we have added. In the KEM calculations that have included solvent, all atoms of the solvent together have been used to define one additional kernel, over and above the five kernels chosen to represent chain A and chain B of the full insulin molecule. The KEM results are $E_{KEM} = -26\,275.4013$ au (Equation 1.44) and −26 275.4127 au (Equation 1.45). The results using the KEM are compared with those obtained for the fully solvated molecule, having a total energy of $E_{HF} = -26\,275.4187$ au. The KEM energies differ from this by 0.0174 au (Equation 1.44) and 0.0060 au (Equation 1.45).

1.4.3.2 Comments Regarding the Insulin Calculations

The electronic structure of protein molecules is still not routinely accessible for study by quantum mechanical methods. Here, it has been shown to be possible using the KEM in the case of the protein insulin. Thus, a quantum mechanical explanation, so useful in application to molecules of moderate size, will prove useful too with protein molecules. Here the KEM, which represents a combination of crystallography and quantum mechanics, while simplifying calculations, has achieved near *ab initio* accuracy in the energy for insulin. This has been demonstrated with the components of insulin called chains A and B, the full insulin molecule, and the fully solvated crystalline insulin molecule. The demonstration was carried out by using the HF approximation in a limited Gaussian basis. The numerical results indicate the validity of the KEM in its application to the various aspects of insulin structure studied in this work.

Table 1.10, which gives the results for the explicit treatment of the solvent molecules that have been crystallized together with insulin, shows that the solvent molecules may be collected into one solvent kernel with results whose accuracy is good. The differences are only of magnitude 10.9428 and 3.7921 kcal mol^{-1}, respectively, using Equations 1.44 and 1.45. The corresponding percentage differences are 0.000 066% and 0.000 023%, respectively. Thus, it is shown here that solvent molecules of crystallization may also be included in the KEM calculations with good accuracy.

The KEM has proven to be applicable to all aspects of the insulin molecule that we have tested [82]. The magnitude of all energy differences obtained between E_{HF} and E_{KEM} are relatively small. Moreover, the energy differences are of the same order of

magnitude as would be expected from the previous work in the case of peptides. We conclude that the KEM calculations are applicable to the energy and electronic structure of proteins.

1.4.4 KEM Applied to DNA

1.4.4.1 KEM Calculation Results

The results for structures from X-ray crystallography and energy differences ($E_{HF} - E_{KEM}$ for all of the double kernels) calculated for each of a dozen different DNA systems are displayed in Figure 1.14 [83–93] and Table 1.11. For these DNA systems the number of atoms and the number of kernels involved range from 198

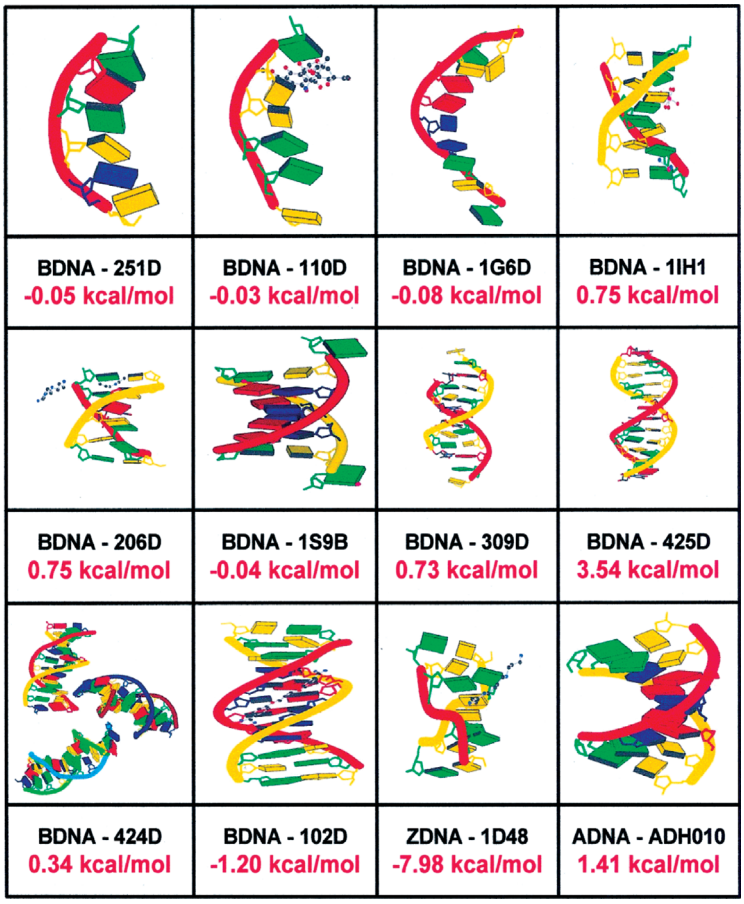


Figure 1.14 DNA structures from X-ray crystallography; range of molecule size: 197 to 2418 atoms. (DNA diagrams are from the Nucleic Acid Database http://ndbserver.rutgers.edu/atlas/xray/index.html.)

Table 1.11 Energy calculation for DNA without solvent.[a)]

	B-DNA 251D	B-DNA 110D	B-DNA 1G6D	B-DNA 1IH1	B-DNA 206D	B-DNA 1S9B
Atoms (kernels)	198 (3)	197 (3)	330 (5)	394 (6)	395 (6)	466 (7)
E_{HF} (au)	−8127.68	−8143.38	−13614.26	−16287.63	−16270.50	−19082.30
E_{KEM} (au)[b)]	−8127.68	−8143.38	−13614.26	−16287.65	−16270.50	−19082.30
$E_{HF} - E_{KEM}$ (kcal mol^{-1})[b)]	0.08	−0.08	0.46	13.02	−5.08	−0.61
E_{KEM} (au)[c)]	−8127.68	−8143.38	−13164.26	−16287.63	−16270.50	−19082.30
$E_{HF} - E_{KEM}$ (kcal mol^{-1})[c)]	−0.05	−0.03	−0.08	0.75	0.75	−0.04
	B-DNA 309D	**B-DNA 425D**	**B-DNA 424D[d)]**	**B-DNA 102D**	**Z-DNA 1D48**	**A-DNA ADH010**
Atoms (kernels)	658 (6)	788 (6)	2364 (18)	790 (6)	394 (6)	528 (8)
E_{HF} (au)	−27079.08	−32509.97	−97529.42	−32476.74	−16286.64	−21652.54
E_{KEM} (au)[b)]	−27079.09	−32509.96	—	−32476.75	−16286.64	−21652.54
$E_{HF} - E_{KEM}$ (kcal mol^{-1})[b)]	6.85	−5.00	—	1.50	2.07	3.24
E_{KEM} (au)[c)]	−27079.08	−32509.98	−97529.42	−32476.75	−16286.63	−21652.54
$E_{HF} - E_{KEM}$ (kcal mol^{-1})[c)]	0.74	3.55	0.34	−1.20	−7.98	1.41

a) The KEM applied to DNA using Equations 1.44 and 1.45, and with HF/STO-3G.
b) The only double kernels included are those made of single kernel pairs that are chemically bonded to one another.
c) All double kernels are included.
d) 424D has three double helix chains, $E_{HF} = E_{ab} + E_{cd} + E_{ef}$.

atoms and 3 kernels for the smallest molecule (B-DNA-251D) up to 2364 atoms and 18 kernels for the largest molecule considered (B-DNA-424D). For each DNA molecular system the full molecule Hartree–Fock energy E_{HF} was calculated. This number is the standard against which the accuracy of KEM results is judged. The energies listed as E_{KEM} represent the results obtained by dividing the DNA molecular systems into kernels and then calculating the total energy in the approximations formalized within Equations 1.44 and 1.45 above. The results of Equations 1.44 and 1.45 were calculated separately. Table 1.11 lists for each molecular system the energy differences $E_{HF} - E_{KEM}$, for Equation 1.45. (Note that the full molecular energies are usually listed in units au, but the energy differences are listed in the smaller units kcal mol^{-1}.)

The results of Table 1.11 show that the KEM is quite accurate, as one may observe from the energy differences $E_{HF} - E_{KEM}$. For Equation 1.44 the absolute magnitude of the energy differences range from a minimum of 0.0795 to a maximum of 13.0105 kcal mol^{-1}. These differences are relatively small, and thus the accuracy of the KEM as implemented in Equation 1.44 is good. The results of Equation 1.45 are even more accurate. For Equation 1.45 the absolute magnitude of the energy differences range from a minimum of 0.0328 to a maximum of 7.9827 kcal mol^{-1}. The Equation 1.45 results are generally expected to be more accurate than the case for Equation 1.44.

1.4.4.2 Comments Regarding the DNA Calculations

The DNA molecular systems of this chapter were treated within the context of the *ab initio* Hartree–Fock approximation. The basis set used for all cases was a limited basis, of Gaussian STO-3G type. A limited basis was chosen to make the energy calculations on full molecular systems (i.e., E_{HF}) as convenient as possible. The numerical values of E_{HF} provided the standard of comparison for the energy values obtained by the KEM. Comparisons between E_{HF} and E_{KEM} have shown that the KEM can be applied to a wide variety of DNA molecular systems with good accuracy. In particular, such calculation accuracy holds true for A-, B- and Z-DNA, the three main types of DNA configuration. The most common configuration of DNA, that is, B-DNA, was examined in ten different molecular systems of variable geometry, and magnitude, as judged by the number of atoms in the system, and was in each case found to be described with good accuracy by the KEM [94].

1.4.5 KEM Applied to tRNA

The quantum mechanical molecular energy of a particular tRNA, of known crystal structure [95], has been calculated with the use of the KEM [96]. The molecule chosen is the yeast initiator tRNA ($ytRNA_i^{Met}$), designated in the Protein Data Bank as 1YFG and in the Nucleic Acid Database as ID TRNA12 (Figure 1.15).

The structure of this molecule is stabilized by a complicated network of hydrogen bonds that have been identified through crystallography. The numerical results obtained in this work use the Hartree–Fock equations, and a limited basis. Table 1.12

Figure 1.15 Crystal structure of tRNA; 1YFG picture is from the Protein Data Bank (PDB).

lists the results that follow from application of Equations 1.44 and 1.45 to the initiator tRNA molecule 1YFG.

The molecule consists of 2565 atoms, which have been broken into 19 kernels. Thus, the average number of atoms per kernel is about 135, which is of such a size as to be readily calculable, whereas the original number of atoms, 2565, is very much less convenient to treat as a whole. We emphasize that Table 1.12 shows that Equations 1.44 and 1.45 results are quite close. They differ by only −0.0073 au, or -1.79×10^{-3} (kcal mol^{-1} atom^{-1}).

Table 1.12 Energy calculation for 1YFG (tRNA) by HF/STO-3G.

No. of atoms	No. of kernels	E_{KEM} [a] (au)	E_{KEM} [b] (au)	$\Delta E = E_{KEM}$ [b] $- E_{KEM}$ [a] (au)	ΔE per atom (kcal mol^{-1})
2565	19	−108995.17	−108995.17	−0.0073	-1.79×10^{-3}

a) The double kernels included are only those made of single kernel pairs chemically bonded to one another, and hydrogen bond interaction energies are added to the results of Equation 1.44.
b) All double kernels are included, Equation 1.45.

We turn now to the matter of the hydrogen bonding network for the 1YFG initiator tRNA that has been established by crystallography (see Nucleic Acid Database, NDB ID TRNA12, in Derivative Data: Hydrogen Bonding Classifications, http://ndbserver.rutgers.edu/atlas/xray/structures/T/trna12/TRNA12-hbc.html), based upon the experimental distances between putative hydrogen bonding donor and acceptor atoms. The interaction energy between a pair of kernels should be negative if that pair is stabilized by the presence of hydrogen bonds. Moreover, the magnitude of the interaction energy would be a measure of the hydrogen bonding stabilization. The interaction energies between pairs of kernels are data that are automatically generated in application of the KEM. The interaction energy, I, between kernels is defined as:

$$I_{ij} = E_{ij} - E_i - E_j, \tag{1.46}$$

where the symbols on the right-hand side of the equation retain their prior meaning. We found that in every instance, corresponding to the hydrogen bonding network established by crystallography, the interaction energy is negative, which is consistent with a stabilizing hydrogen bonding interaction between the relevant kernels. Thus the energetics available from the KEM provide independent confirmation of the hydrogen bonding network obtained experimentally from crystallography.

1.4.6 KEM Applied to Rational Design of Drugs

1.4.6.1 Importance of the Interaction Energy for Rational Drug Design

The importance of the interaction energy for rational drug design may be envisioned by consideration of Figure 1.16. The efficacy of drugs is based upon a geometrical "lock and key" fit of the drug to the target, complemented by an electronic interaction between the two. As indicated in Figure 1.16 by dashed lines, there will be several interactions between the drug and the kernels that constitute its target. The KEM delivers the *ab initio* quantum mechanical interaction energy between the drug and its target. This is computationally practical for molecular targets containing even tens of thousands of atoms. That is the great advantage of using the KEM for rational drug design. Moreover, not only is the total interaction energy obtained, so too as a natural consequence of the KEM approximation are the individual kernel components of the interaction energy. That is to say, the interaction energy of the drug with each individual kernel in the target is obtained. Thus the contribution from each kernel to the efficacy of binding to the drug, which may be large or small, and attractive or repulsive, may be obtained. In this way the most important interactions between the drug and the kernels of the target become evident.

Here we describe our calculations of the energy of various drug–RNA interactions. All calculations here employ a limited basis and the Hartree–Fock approximation. The definition of the interaction energy between any pair of kernels is Equation 1.46 in the previous section. In this section, we use it to calculate the interaction energies between the drug and RNA.

Knowledge of the list of the double kernel interaction energies is critical to rational drug design. That list determines the total drug–target interaction energy as well as

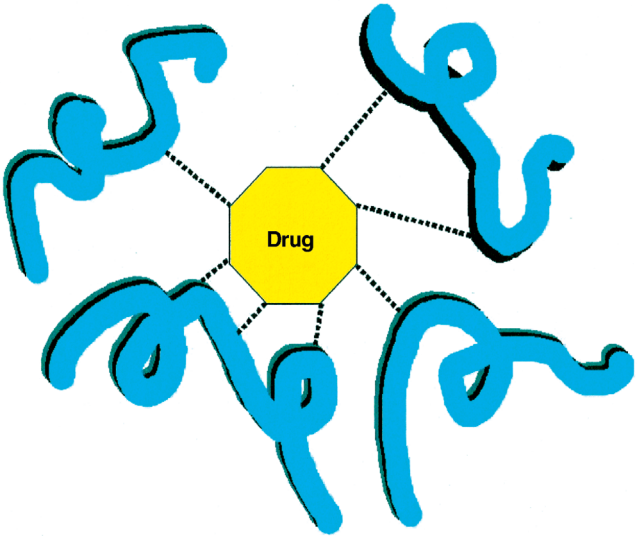


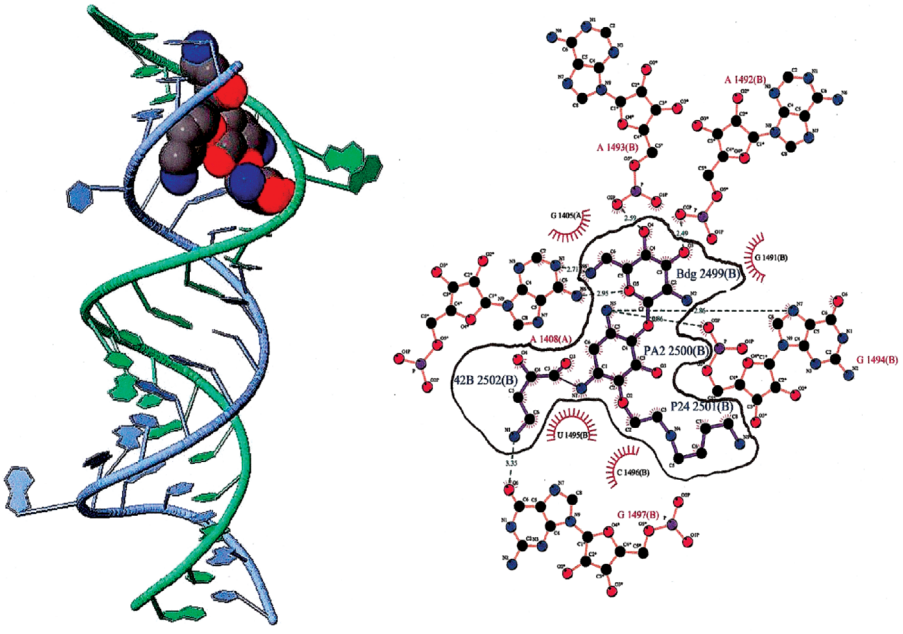
Figure 1.16 Sketch indicating the interaction of a putative drug molecule with its target, a very large medicinal molecular structure. The drug fits geometrically within a reactive "pocket" of the target. The dashed lines indicate interactions with the various kernels that compose the target. The interaction may be either positive or negative; both types of interaction (attractive and repulsive) are expected to occur.

the analysis of exactly which kernels contribute most importantly. Such knowledge may be obtained for the hundreds, or even thousands, of different chemical substitutions at various sites around the drug periphery, and the effect upon the interaction between the drug and the target computed. Such computational information can effectively replace the perhaps thousands of laboratory synthesis experiments needed to obtain related information. Moreover, it would be extremely difficult to obtain, by experimental methods, the double kernel interaction energies that flow naturally from implementation of the KEM to the problem.

1.4.6.2 Sample Calculation: Antibiotic Drug in Complex (1O9M) with a Model Aminoacyl Site of the 30s Ribosomal Subunit

The ribosome is a well-known target for antibiotic drugs. The crystal structure of one such drug, when attached to an A site RNA, is a complex named 1O9M, which has been solved [97] (Figure 1.17). Solvent water molecules included in the crystal structure are not shown in the figure. Utilizing the crystal structure we have calculated by the KEM the relevant energy quantities. These include the total energy of the complex made up of RNA, solvent and drug, and the separate RNA, solvent and drug molecules. We have obtained interaction energies descriptive of the drug–RNA target interaction, and of the hydrogen bonding network within the RNA molecule.

Table 1.13 displays the calculated energy results for the 1O9M drug–RNA target and solvent complex. The total complex, consisting of 1673 atoms, has been "broken"



(a) (b)

Figure 1.17 (a) Crystal structure of the drug–RNA complex 1O9M (molecule picture generated by Jmol Viewer); (b) drug–RNA interactions in the crystal. (Modified from PDBSum web site, LIGPLOT of interactions involving ligand.)

Table 1.13 Drug–target interaction energies (au) for rational design of drugs (see text for details).

Double kernels *ij* (RNA & drug)	Single kernel *i* (RNA)	Single kernel *j* (drug)	I_{ij} (au)	Kernel *i*–kernel *j* (RNA–drug)
−6219.279785	−4402.131144	−1817.149679	0.001038	1–15
−5984.204590	−4167.047454	−1817.149679	−0.007456	2–15
−5964.670898	−4147.520166	−1817.149679	−0.001053	3–15
−6183.414063	−4366.264309	−1817.149679	−0.000074	4–15
−6129.126465	−4311.976759	−1817.149679	−0.000026	5–15
−6129.086914	−4311.937498	−1817.149679	0.000263	6–15
−6038.689453	−4221.539976	−1817.149679	0.000203	7–15
−6219.246582	−4402.096702	−1817.149679	−0.000201	8–15
−5984.210449	−4167.060881	−1817.149679	0.000111	9–15
−5964.679199	−4147.529518	−1817.149679	−0.000002	10–15
−6183.388184	−4366.238058	−1817.149679	−0.000446	11–15
−6129.133301	−4311.980653	−1817.149679	−0.002968	12–15
−6129.113770	−4311.964279	−1817.149679	0.000189	13–15
−6038.665039	−4221.514542	−1817.149679	−0.000818	14–15
−6539.791016	−1817.149679	−4722.624492	−0.016844	16–15

into 16 kernels. Of these kernels, 1–14 represent the RNA target, kernel 15 represents the drug and kernel 16 represents the crystalline water of solvation.

Table 1.13 lists the interaction energies between the drug kernel and the kernels of RNA. The hydrogen atom positions have been energy optimized. The first three columns of the table list the calculated KEM energies for each double kernel and each of its two single kernel components, respectively. The fourth column lists each double kernel interaction energy. The fifth column names the double kernels. The single kernels that make up the RNA target are numbered 1–14. The antibiotic drug is kernel number 15 and the water is kernel 16. The interaction energy of RNA and drug, obtained from the sum of all the 14 RNA kernels and drug kernel interaction energies, is −0.01 124 au, and the interaction energy of RNA in water and drug is −0.02 809 au. We have shown how to begin with a crystal structure, and obtain therefrom quantum mechanical information not otherwise known from the structure alone. Such information includes the energy of the structure, the interaction energy between a drug and its target, and the analysis of such interaction energy in terms of the contribution of each contributing kernel pair. Thus the relative importance of individual kernels to the drug interaction efficacy can be assessed. This forms the basis of a rational drug design improvement from use of a lead drug structure.

1.4.6.3 Comments Regarding the Drug–Target Interaction Calculations

Assume the knowledge of a lead compound that displays the usual list of necessary properties, including adsorption, distribution, metabolism, excretion, and toxicity (ADMET). The critical factor that computational chemistry can contribute is the interaction energy between a putative drug and its target. If the target is a molecular structure containing thousands, or even tens of thousands of atoms, and if an *ab initio* quantum mechanical description of the interaction is to be obtained, then clearly an approximation such as that of the KEM is indicated. Thus, targets composed of peptides, proteins, DNA, RNA and various of their molecular composites can contain enormous numbers of atoms. Because the straightforward computational difficulty of a fully quantum mechanical calculation rises in proportion to a high power of the number of atoms in the molecular system, such calculations have typically been computationally impractical. The use of the KEM alleviates such computational difficulty by means of a formalism that divides a large molecular system into kernels, which are much smaller than the molecular system considered as a whole. Computations with each of the kernels are thus a relatively smaller problem, and can be assigned individually to separate nodes of a parallel processor. Thus a kind of twofold advantage accrues to the KEM, since individual calculations are smaller piecewise than otherwise, and they may be computed in parallel with modern computers designed for that purpose. The entire molecular system is reconstituted from a sum over kernels. What has been shown by the calculations of this chapter is that the KEM may be applied for purposes of rational design of drugs to the large molecules of medicinal chemistry. *Ab initio* results of expected high accuracy, within computational times of reasonable practicality, are obtained. Therefore, in general the KEM will be well suited for obtaining the interaction energy between drug molecules and their target medicinal chemical molecules of large size.

The point that has been made here is that the KEM can be useful for the rational design of drug molecules [98]. The key ideas that result and are useful for drug design are the interaction energy between a drug and its large molecular target, and all the component interaction energies for the individual double kernels.

1.4.7 KEM Applied to Collagen

This discussion combines a collagen molecule of given structure with quantum-mechanical KEM calculations to obtain the energies and interaction energies of a triple helix protein. Knowledge of such energetics allows one to understand the stability of known structures, and the rational design of new protein interacting chains. It is shown that the kernel energy method accurately represents the energies and interaction energies of each of the chains separately and in combinations with one another. This is a challenging problem for the case of large molecular protein chains. However, here the computational chemistry calculations are simplified, and the information derived from the atomic coordinates of the structure is enhanced by quantum mechanical information extracted therefrom.

1.4.7.1 Interaction Energies

The interaction energy among a triplet of protein chains is generalized to:

$$I_{abc} = E_{abc} - (E_a + E_b + E_c) \tag{1.47}$$

where the subscript indices name the triplet of protein chains in question, I_{abc} is the triplet chain interaction energy, E_{abc} is the energy of a triplet of chains, and E_a, E_b and E_c are each the energies of a single protein chain. Again, importantly, the sign of the interaction energy, I_{abc}, indicates whether the triplet of protein chains a, b and c altogether attract (negative I) or repel (positive I). It would be difficult to obtain from atomic coordinates alone the magnitude of the interaction energies that flow naturally from implementation of the KEM. The KEM delivers the *ab initio* quantum mechanical interaction energy between and among protein chains. This may be envisioned to be computationally practical for molecular structures containing thousands, or even tens of thousands of atoms.

1.4.7.2 Collagen 1A89

Collagen is a protein that is essential to the physical structure of the animal body. The molecule is made of three peptide chains that form a triple helix. These are incorporated in a vast number of ways to create structure. Collagen molecular cables provide strength in tendons, resilience to skin, support to internal organs, and a lattice structure to the minerals of bones and teeth. A repeated sequence of three amino acids forms the chains out of which the collagen triple helix is composed. Every third amino acid is glycine. Remaining positions in the chain often contain proline and hydroxyproline.

We selected for study a particular collagen molecule whose molecular structure is known, namely, 1A89 [99], and whose atomic coordinates are readily available in the

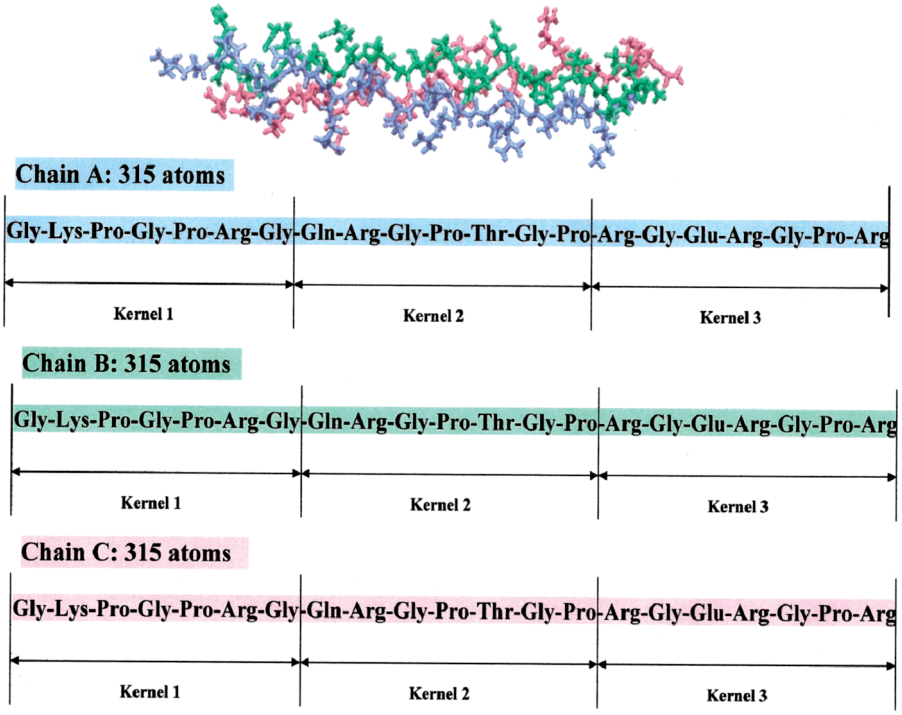


Figure 1.18 Picture of the collagen triple helix 1A89 and the primary structure of each of its individual protein chains broken into kernels.

Protein Data Bank. The atomic coordinates are the starting information from which the KEM proceeds. Clearly, from the structural role it plays in the animal body, collagen must be a stable molecule, with the chains of the triple helix structure adhering to one another. We applied the KEM to the molecular structure 1A89 to establish whether the approximation is sufficiently accurate to reveal the expected adhesion of the collagen triple chains.

Figure 1.18 shows a triple helix of protein chains that make up the collagen molecule that we have studied. Also shown is the amino acid primary structure of the three identical protein chains that make up the helix. Each chain is broken into three kernels, as shown in the figure. The total triplex contains 945 atoms, each chain contains 315 atoms, with kernels 1, 2 and 3 containing 96, 98 and 121 atoms, respectively. The atomic coordinates used in all of the calculations are obtained from the known molecular structure.

Table 1.14 contains the KEM calculations for each of the protein chains considered as a single entity. All calculations of this chapter are of quantum mechanical Hartree–Fock type, using an STO-3G limited basis of atomic orbitals. An "exact" result refers to the Hartree–Fock calculation of an entire molecule, including all of its atoms together, without use of the kernel approximation. The KEM calculated energies are meant to approximate the "exact" results. The difference between the

Table 1.14 Energy calculations for collagen triple helix (1A89) at the HF/STO-3G level of theory.

Chain	Atoms	Kernels	E_{HF} (au)	E_{KEM} (au)	$E_{HF} - E_{KEM}$ (au)	$E_{HF} - E_{KEM}$ (kcal mol^{-1})
A	315	3	−7381.86	−7381.86	0.0000	0.0047
B	315	3	−7382.16	−7382.16	0.0000	0.0260
C	315	3	−7382.83	−7382.83	−0.0002	−0.1027
Triple helix	945	9	−22146.92	−22146.91	−0.0059	−3.7332

two types of calculation is listed in both au and kcal mol^{-1}. One may conclude that the KEM calculation represents well the "exact" result. The percentage difference between the two types of calculation is small. For the single chains A, B and C the percentage differences are 1.0×10^{-7}%, 5.6×10^{-7}% and 2.2×10^{-6}%, respectively. Notice also that the percentage difference for the entire triple helix is only 2.7×10^{-5}%. This level of accuracy is in accord with our previous experiences [32, 43, 82, 94, 96, 98].

Table 1.15 lists the calculation results for the triplex protein chains considered in pairs. The rows and columns are arranged as in Table 1.14, except that a new quantity, the interaction energy between the chains of the pairs, is also listed. As previously, the accuracy of the KEM energies is as expected, with differences for pairs AB, AC and BC of approximately 2.6×10^{-5}%, 2.2×10^{-5}% and 2.8×10^{-5}%, respectively. Notably, not only do we obtain the chain pair interaction energies but, as expected, the interaction is attractive.

Table 1.16 contains the calculation results for the full triple helix of the collagen structure. As indicated above, the KEM result for the total energy is accurate. The HF and KEM interaction energies of the triple helix are also listed.

Table 1.15 Interaction energy calculations[a)] of chain pairs at the HF/STO-3G level of theory.

Chains	Atoms/ kernels	E_{HF} (au)	E_{KEM} (au)	I_{HF} (kcal mol^{-1})	I_{KEM} (kcal mol^{-1})	$I_{HF} - I_{KEM}$ (kcal mol^{-1})
AB	630/6	−14764.05	−14764.05	−23.1488	−20.7075	−2.4413
AC	630/6	−14764.71	−14764.71	−13.2896	−11.2950	−1.9946
BC	630/6	−14765.01	−14765.01	−8.6151	−6.2123	−2.4028

a) Interaction energies are calculated from: $I_{ab} = E_{ab} - E_a - E_b$.

Table 1.16 Interaction energy calculations[a)] of collagen triple helix at the HF/STO-3G level of theory.

$E_{HF(abc)}$ (au)	$E_{HF(a+b+c)}$ (au)	$E_{KEM(abc)}$ (au)	$E_{KEM(a+b+c)}$ (au)	I_{HF} (kcal mol^{-1})	I_{KEM} (kcal mol^{-1})	I_{HF} -I_{KEM} (kcal mol^{-1})
−22146.92	−22146.85	−22146.91	−22146.85	−41.48	−37.90	−3.58

a) Interaction energies calculated from: $I_{abc} = E_{abc} - E_{a+b+c}$, where $E_{a+b+c} = E_a + E_b + E_c$.

1.4.7.3 Comments Regarding the Collagen Calculations

The protein molecule chains, and their pair and triplex aggregates, taken from the molecular structure 1A89, in this chapter were treated within the context of the *ab initio* Hartree–Fock approximation. The basis set used was a limited basis, of Gaussian STO-3G type. A limited basis was chosen simply to make the energy calculations as convenient as possible, for a protein structure of this size. Previous numerical experience has shown that the KEM can be applied to a wide variety of molecules with good accuracy, and such expectations were realized in this instance.

We have shown how to begin with a known molecular structure and obtain therefrom quantum mechanical information not otherwise known from the structure alone. With collagen, such information includes the energy of the individual protein chains and their combinations in pairs and as a triplex. Importantly, the interaction energy between chains of a pair, or among those of a triplex are well represented by the KEM. Notably, the KEM approximation is sufficiently accurate to reveal the expected adhesion that must prevail among the collagen triple chains. This forms the basis of an understanding of the structure of collagen in particular, but more generally of a rational design of protein chain interactions [100].

What has been shown by the calculations here is that the KEM may be applied for purposes of obtaining the interaction energy between protein chains for an understanding of known molecular structures and for the rational design of proposed structures of considerable size.

1.4.8
KEM Fourth-Order Calculation of Accuracy

Remarkable accuracy has been achieved in the calculation of the energy of the ground state of the important biological molecule Leu^1-zervamicin [30], whose crystal structure is known and used in the calculations.

Figure 1.19 shows schematically defined kernels, double, triple and quadruple kernels; only these objects are used for all quantum calculations. The total molecular energy is reconstructed therefrom by summation over the contributions of the kernels and multiple-kernels up to the highest order of interaction to be imposed. In this description we extend the KEM to a fourth order of approximation. The aim, of course, is to increase the accuracy of the KEM calculations. Remarkable accuracy, as we indicate below, can be achieved.

1.4.8.1 Molecular Energy as a Sum over Kernel Energies

The formulas for invoking the KEM up to orders of approximation including double, triple and quadruple energies are displayed as Equations 1.48,1.49 and 1.50, respectively [101]:

$$E_n^{\text{total}} = \sum_{\substack{i=1 \\ i<j}}^{n-1} E_{ij} - (n-2)\sum_{i=1}^{n} E_i \tag{1.48}$$

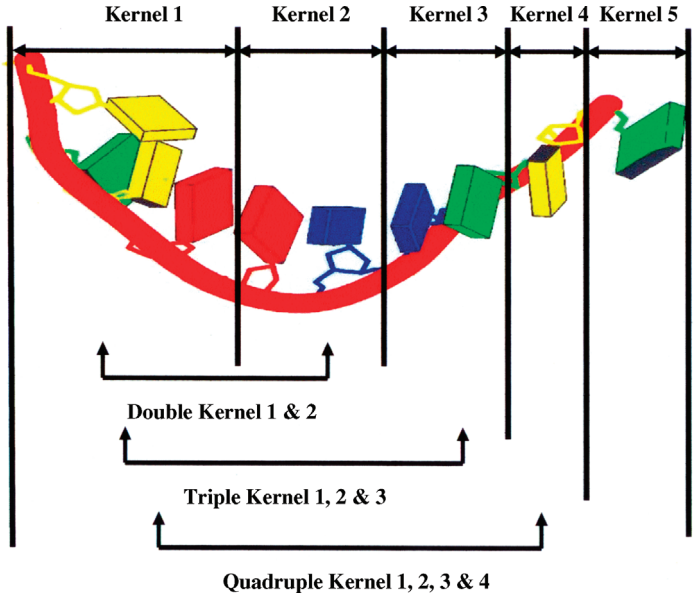


Figure 1.19 Abstract sketch of a molecule showing the definitions of the single, double, triple and quadruple kernels.

$$E_n^{\text{total}} = \sum_{\substack{i=1 \\ i<j<k}}^{n-2} E_{ijk} - (n-3) \sum_{\substack{i=1 \\ i<j}}^{n-1} E_{ij} + \left(\sum_{i=1}^{n-3} i\right)\left(\sum_{i=1}^{n} E_i\right) \tag{1.49}$$

$$E_n^{\text{total}} = \left(\sum_{\substack{i=1 \\ i<j<k<l}}^{n-3} E_{ijkl}\right) - (n-4)\left(\sum_{\substack{i=1 \\ i<j<k}}^{n-2} E_{ijk}\right) + \left(\sum_{i=1}^{n-4} i\right)\left(\sum_{\substack{i=1 \\ i<j}}^{n-1} E_{ij}\right) - \left(\sum_{j=1}^{n-4}\sum_{i=1}^{j} i\right)\left(\sum_{i=1}^{n} E_i\right) \tag{1.50}$$

1.4.8.2 Application to Leu^1-zervamicin of the Fourth-Order Approximation of KEM

We tested the accuracy achievable with the above formulas by application to the important biological molecule Leu^1-zervamicin (Figure 1.20). It is an antibiotic that transports potassium ions across cell membranes. Groups of zervamicin molecules assemble to form channels that serve to allow ion passage. The molecule has hydrophobic side chains extending from the peptide residues on one side and polar side chains extending from other peptide residues on the other side. A side chain of particular interest is a side chain of (residue 11) glutamine. This side chain is attached

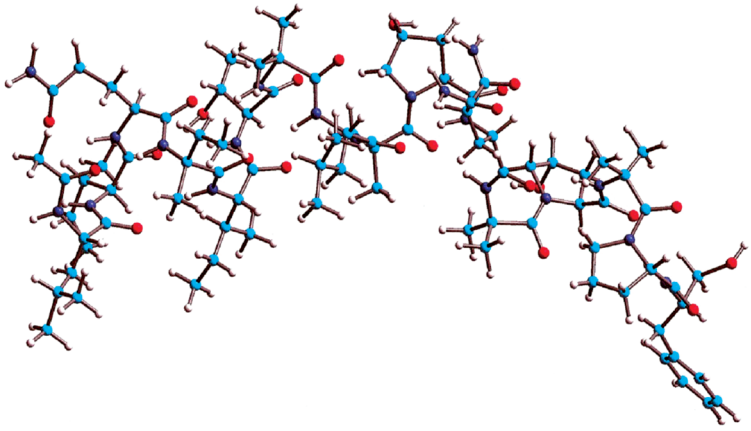


Figure 1.20 Leu[1]-zervamicin (closed form) X-ray crystal structure.

on the hydrophobic side of the peptide, as determined from a crystal-structure investigation by Karle *et al.* [30]. The side chain acts as a gate that allows K^+ ions to proceed through a channel membrane. Glutamine 11, by a swinging motion, acts to open and close the channel each time an ion goes through it. If the structural arrangement of the zervamicin molecules in a cell membrane is similar to that in the crystal, it is possible that the zervamicin crystal structures offer a model of a gating mechanism on the level of atomic resolution. Table 1.17 lists the results of calculation related to Leu[1]-zervamicin.

The program Gaussian 03 [102] was used to carry out the calculations, related to each of the formulas 1, 2 and 3 applied to the Leu[1]-zervamicin molecule. Table 1.17 lists the results in order of increasing accuracy, reading across the table, for the double, triple and quadruple interaction approximations. For each of these approximations, the result listed as the exact energy is the brute force Hartree–Fock calculation on the full molecule using the limited basis of functions STO-3G. The table also lists the corresponding energy calculated by means of the KEM equations. Equations 1.48–1.50 are used to obtain the results listed for the double, triple and quadruple kernel approximations, respectively. The difference between the brute force and the KEM calculations are listed in au. Finally the results are listed on the basis of energy difference per atom in units of kcal mol^{-1}. As expected, the difference between the exact and the KEM results decrease as the order of the KEM approxima-

Table 1.17 HF STO-3G energy calculated up to fourth-order approximation of KEM.

	Single	**Double**	**Triple**	**Quadruple**
E_{KEM} (au)	−5851.8663	−5851.5469	−5851.5686	−5851.5703
E_{exact} (au)	−5851.5703	−5851.5703	−5851.5703	−5851.5703
ΔE (au)	0.296	−0.0234	−0.0017	0.0000
ΔE per atom (kcal mol^{-1})	0.70	−0.06	0.00	0.00

tions taken into account increase from double to triple to quadruple interactions. The numerical differences are −0.0234, −0.0017 and 0.0000 au for the double, triple and quadruple interactions, respectively.

What are the limits of accuracy based upon the idea of quantum kernels? To answer this question, we calculated the KEM energy to an order of approximation including terms up to a fourth order of interaction among the kernels.

The standard of accuracy for these calculations was the brute force Hartree–Fock calculation of the energy for the full molecule in the same basis as for the KEM calculations. As the results of Table 1.17 show, the accuracy of the KEM increases with each order of approximation. For example, using Table 1.17, consider the differences on the basis of difference per atom. At the level of double kernel interactions (Equation 1.48) the magnitude of difference is of the order of 10^{-2} kcal mol^{-1} $atom^{-1}$, an already small error. If triple kernel interactions (Equation 1.49) are invoked the magnitude of difference is reduced by an order of magnitude to 10^{-3} kcal mol^{-1} $atom^{-1}$. Finally, using quadruple interactions (Equation 1.50) induces a further reduction in difference by an additional four orders of magnitude to 10^{-7} kcal mol^{-1} $atom^{-1}$. For this molecule, at least, and those similar to it one need not contemplate going beyond the quadruple level of accuracy in the KEM approximation. The results here allow one to conclude that *ab initio* accuracy is obtainable for biological molecules within the KEM, carrying the approximation up to quadruple interactions. For large enough molecules, for which brute force calculations are not feasible, the KEM calculations will still be practicable, because the kernels and multiple kernels can be chosen to be very much smaller than the full molecule.

The KEM suggested here for especially high accuracy might find application in many problems in which the object of calculation might concern the quantum mechanics of large molecules. These include the rational design of drugs, peptide folding and the study of weak interactions among biological molecules.

1.4.9
KEM Applied to Vesicular Stomatitis Virus Nucleoprotein, 33 000 Atom Molecule

1.4.9.1 Vesicular Stomatitis Virus Nucleoprotein (2QVJ) Molecule

The vesicular stomatitis virus nucleoprotein (2QVJ) molecule (Figure 1.21) [103] is composed of five chains (A–E). Each chain has 421 residues, 6635 atoms and carries a charge of + 3 au. The entire molecule contains 33 175 atoms. Each of the five chains is divided into 66 kernels, so we have altogether $66 \times 5 = 330$ kernels, contained in the entire molecule composed of five chains.

To calculate the energies of the molecule and each of its five sub-chains we have used the atomic coordinates of the crystal structure solved for the 2QVJ molecule. However, the crystal structure does not deliver the hydrogen atom coordinates. These have been added automatically to the heavy atoms of the crystal structure using the procedures of the computer program HyperChem [104]. The same amino acid sequence defines each of five chains that make up the full molecule. However, the relative positions occupied by each of the chains in the full molecule will differ, and

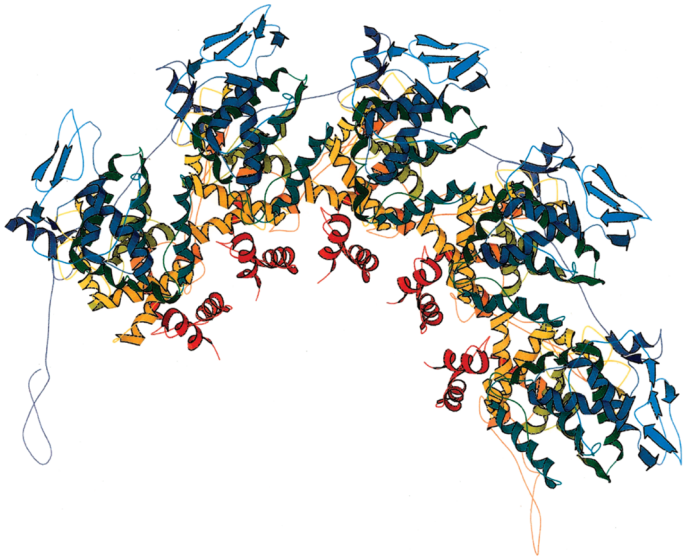


Figure 1.21 The crystal structure of 2QVJ is composed of five similar chains. The picture was generated with KiNG Viewer on the PDB web site.

this in turn will affect slightly the automatic placement of hydrogen atoms from chain to chain. As a result the energy of the chains will differ slightly. In Table 1.18 we list the limited basis (STO-3G) Hartree–Fock energies, calculated using Equation 1.45, for each of the five chains (A–E) that make up the full molecule.

Equation 1.45 was applied to calculate the total energy of each chain and the full molecule 2QVJ; the total energy of the full molecule is −825 954.57 au.

1.4.9.2 Hydrogen Bond Calculations

We determined all hydrogen bond interactions between the chains and calculated their magnitudes using Equation 1.46. All the hydrogen bond calculations reported used 6-31G** basis functions in both the Hartree–Fock and MP2 approximations [105].

1.4.9.3 Comments regarding the 2QVJ Calculations

Within RNA viruses the viral genome RNA is completely enwrapped by a nucleoprotein. Vesicular stomatitis virus is such a case. The nucleoprotein in VSV is a ten-member "cylindrical" oligomer, half of which is the five-member oligomer 2QVJ, which retains a "half cylinder shape," has been crystallized and is the subject of study

Table 1.18 Energies for individual chains of 2QVJ.

Chain	A	B	C	D	E
Energy (au)	−165191.04	−165192.19	−165189.29	−165193.27	−165189.20

here. As suggested by the authors of the crystal structure [103] the intermolecular interactions among the chains that make up the nucleoprotein play a critical role in providing the structural stability it acquires before encapsulation of the viral RNA. This conclusion follows from the crystal structure study. Knowledge of the crystal structure alone does not dictate the actual magnitude of the inter-chain interaction energies. However, given the crystal structure it becomes possible to extract from it the hydrogen bond donors and acceptors and with that information to calculate the inter-chain hydrogen bond interaction energies. That has been accomplished in this chapter.

To begin, we calculated the total energy of the entire 2QVJ molecule using the basic ideas of the KEM. The coordinates of the atoms used were obtained from the crystal structure at 2.8 Å resolution, except for the hydrogen atoms that were added using HyperChem [104]. For the fairly large number of atoms in the molecule as a whole (33 175 atoms) we calculated the energy in the Hartree–Fock approximation, using a limited basis of Gaussian orbitals. We considered each of the molecule's five chains separately, breaking each chain into 66 kernels. A total of 330 kernels make up the whole molecule. Each kernel was chosen to contain approximately 100 atoms, which is of practical size. In this way, using Equation 1.45 the energy of each of the five molecular chains was obtained. Equation 1.45 also delivers a total energy for the full molecule of −825 954.57 au.

It is likely that the inter-chain hydrogen bonds are among the most important contributors to the stability of the inter-chain structure of the whole molecule. All of the many possible inter-chain hydrogen bonds have been considered, their geometries displayed and their corresponding energies calculated [105]. The calculated hydrogen bond energies indicate the importance of correlation energy in representing the hydrogen bond interactions. Not only are the MP2 interaction energies quite a bit lower than the Hartree–Fock values, in some cases the Hartree–Fock results indicate repulsion (positive sign) instead of attraction (negative sign).

In summary the quantum calculations of VSV complement the crystal structure determination of the molecule by delivering the energetics that follow from knowledge of the atomic coordinates. One obtains by the KEM an approximation to the total energy of the whole molecule, and the individual chains that make it up. Principal contributors to the chain interactions are the hydrogen bonds between them. All of these hydrogen bond interactions have been calculated in both the Hartree–Fock and MP2 approximations.

1.5
Summary and Conclusions

Quantum mechanics and crystallography are borderline fields of great importance to the infrastructure of biochemistry. They are unified by sharing the electron density as a cornerstone object. X-rays scatter off the electron density. Density functional theory (DFT) obtains quantum properties as a function of the electron density. For a complete description of a molecular system to evolve from the electron density by

means of either experiment, as in X-ray scattering, or theory, as in density functional theory, *N*-representability is a practical mathematical necessity. For this reason, in our analysis of the X-ray scattering experiment, *N*-representability was introduced into the literature of crystallography. Clinton's equations are a practical method for finding a normalized single determinant *N*-representable electron density. Examples of this are given in the applications to crystals of beryllium and maleic anhydride. It may be mentioned in passing that the general importance of *N*-representability is now well recognized. It has been indicated by the US National Research Council to be one of the ten most prominent research challenges in quantum chemistry [106].

Among the earliest studies to recognize the importance of the fact that X-ray densities derive from a wavefunction is that of W. N. Lipscomb and coworkers concerning Hartree–Fock calculations of various models of the molecule of diborane. These were transformed into structure factors that were compared to the experimental ones from X-ray diffraction data. A "best-fit" provided a choice among quantum mechanical models. This work was reviewed in an article by Lipscomb [107].

To obtain *N*-representable densities from the X-ray data of ever larger biological molecules, one finds the number of experimental density matrix elements tends toward an inconveniently large number. Introducing the concept of quantum kernels reduces the size of the density matrices and makes their accurate determination practicable. This was demonstrated by the examples of a cyclic hexapeptide trihydrate and Leu^1-zervamicin.

The case of maleic anhydride suggests that nuclear positions are not much affected by variations of the density matrix. That is to say, the sum-of-spherical-atoms model used to solve the crystal structure gives good nuclear positions. If one simply adopts the atomic coordinates given by the crystal structure determination, and holds them fixed, that leads to the subfield of quantum crystallography called the kernel energy method (KEM). In the KEM a large biological molecule is mathematically broken into smaller pieces called kernels. Practicable calculations are carried out only on kernels (and multiple kernels). Subsequently, a direct summation over the kernels delivers the energy of the full molecule. Happily the kernel representation of the *ab initio* quantum problem is accurate. This point is made by application to peptides, proteins, DNA and tRNA examples. Moreover, the KEM has been tested with various basis functions and quantum methods. The KEM works for the whole variety of chemical models that have been tested. In addition, the computational time of the calculations is reduced by adoption of quantum kernels. At the level of double kernels approximation for calculating the total energy of large molecules, the KEM has a lower limit in both computing time and accuracy as represented by equation 1.44. The KEM calculation is flexible enough to accommodate the addition to equation 1.44 of any particular interaction energies between kernels. Of course increased interaction energies added to equation 1.44 will increase both accuracy and computational time associated with the results. The upper limit of accuracy and computational time is achieved in the representation of equation 1.45. So KEM makes possible a choice, between the lower & upper limits of accuracy and computing time, dependent on needs.

A natural consequence of the KEM is that it represents well the effect of the hydrogen bonding so important in the structure of biological molecules. The case of a

collagen triplex represents this fact. Also, KEM obtains the energetics that underlie a rational design of drugs. An example of this is shown in an antibiotic drug in complex (1O9M) with a model aminoacyl site of the 30S ribosomal subunit.

To push the limits of KEM accuracy and molecular size, two calculations have been put forward. Expressions for the quantum kernel expansion of the energy have been carried to a fourth order of accuracy. Applied to the ground state of the important biological molecule Leu^1-zervamicin, the fourth-order expansion achieves remarkable accuracy. In a quantum mechanical representation of a truly large molecule, the KEM has been applied to the 2QVJ molecule, composed of five chains, each of 421 residues. The entire molecule contains 33 175 atoms.

The density matrices for the large biological molecules that may be built up from the density matrices of kernels are rendered *N*-representable by means of Clinton's equations. The KEM discussed here applies with advantage to a host of problems in which the object of calculation concerns the true quantum mechanics of large molecules. These include the rational design of proteins, the study of protein folding and molecular self-assembly. Furthermore, perhaps it is not too speculative to insist that the quantum mechanics of large biological molecules can with advantage be brought to bear upon important medical problems more frequently than presently occurs. As a medical oncologist, with an interest in a multidisciplinary approach to cancer research, Provenzano has recently observed [108] that the application of quantum mechanics to medical problems is much to be wished for and encouraged. With this remark, surely the Pullmans would agree.

Acknowledgments

We thank the Office of Naval Research for supporting the work at the Naval Research Laboratory (NRL). One of us (L.M.) wishes to thank the U.S. Navy Summer Faculty Research Program administered by the American Society of Engineering Education for the opportunity to spend summers at NRL. Also, L.M. thanks NIH for grant RR-03037 the National Center for Research Resources, and PSC CUNY for grant 69701-00 38.

References

1 Von Neumann, J. (1927) *Nachr. Ahad. Wiss. Gottigen, Math. Pysik. K1. II a. Math Physik. Chem. Abt.*, 245–272.
2 Dirac, P.A.M. (1931) *Proc. Cambridge Phil. Soc.*, **27**, 240–253.
3 Huisimi, K. (1940) *Proc. Phys. – Math. Soc. Japan*, **22**, 264–314.
4 Mayer, J.E. (1955) *Phys. Rev.*, **100**, 1579–1586.
5 Tredgold, R.H. (1957) *Phys. Rev.*, **105**, 1421–1423.
6 Lowdin, P.O. (1955) *Phys. Rev.*, **97**, 1474–1489.
7 Coleman, J. (1963) *Rev. Mod. Phys.*, **35**, 668–686.
8 McWeeney, R. (1960) *Rev. Mod. Phys.*, **32**, 335–369.
9 Fano, K. (1957) *Rev. Mod. Phys.*, **29**, 74–93.
10 Terhaar, D. (1976) *Reduced Density Matrices in Quantum Chemistry*, Academic Press, New York.
11 Davidson, E.R. (1976) *Reduced Density Matrices in Quantum Chemistry*, Theoretical Chemistry; A Series Of Monographs, vol. 6, Academic Press Inc., London.

12 Gilbert, T.L. (1975) *Phys. Rev. B*, **12**, 2111–2120.
13 Frishberg, C.A. and Massa, L. (1981) *Phys. Rev. B*, **24**, 7018–7024.
14 Clinton, W.L., Galli, A., and Massa, L. (1969) *Phys. Rev.*, **177**, 7–13.
15 Massa, L., Goldberg, M., Frishberg, C.A., Boehme, R., and La Placa, S. (1985) *Phys. Rev. Lett.*, **55**, 622–625.
16 Dovesi, R., Pisani, C., and Ricca, F. (1982) *Phys. Rev. B*, **25**, 3731–3739.
17 Chou, M.Y., Lam, P.K., and Cohen, M.L. (1983) *Phys. Revs. B*, **28**, 4179–4185.
18 Huang, L., Massa, L., and Karle, J. (1999) *Int. J. Quant. Chem.*, **73**, 439–450.
19 Massa, L., Huang, L., and Karle, J. (1995) *Int. J. Quant. Chem. Quant. Chem. Symp.*, **29**, 371–384.
20 Huang, L., Massa, L., Karle, J., and Int, J. (1996) *J. Quant. Chem. Quant. Chem. Symp.*, **30**, 479–488.
21 Blessing, R. (ed.) (1990) *Studies of Electron Distributions in Molecules and Crystals, Transactions of the American Crystallographic Association*, vol. 26, Polycrystal Book Service, Dayton, OH.
22 Clinton, W.L., Galli, A.J., Henderson, G.A., Lamers, G.B., Massa, L.J., and Zarur, J. (1969) *Phys. Rev.*, **177**, 27–33.
23 Frishberg, C. (1986) *Int. J. Quantum Chem.*, **30**, 1–5.
24 Cohn, L., Frishberg, C., Lee, C., and Massa, L.J. (1985) *Int. J. Quantum Chem. Symp.*, **19**, 525–533.
25 Clinton, W.L. and Massa, L.J. (1972) *Int. J. Quantum Chem.*, **6**, 519–523.
26 Clinton, W.L. and Massa, L.J. (1972) *Phys. Rev. Lett.*, **29**, 1363–1366.
27 Clinton, W.L., Frishberg, C.A., Goldberg, M.J., Massa, L.J., and Oldfield, P.A. (1983) *Int. J. Quantum. Chem. Symp.*, **17**, 517–525.
28 Hammermesh, M. (1962) *Group Theory*, Addison-Wesley, Reading, MA.
29 Karle, I.L., Gibson, J.W., and Karle, J. (1970) *J. Am. Chem. Soc.*, **92**, 3755–3760.
30 Karle, I.L., Flippen-Anderson, J.L., Agarwalla, S., and Balaram, P. (1994) *Biopolymers*, **34**, 721–735.
31 Frisch, M.J., Frisch, A., Foresman, J.B. *et al.* (1994) Gaussian 94, Gaussian, Inc., Pittsburgh, PA.
32 Huang, L., Massa, L., and Karle, J. (2005) *Int. J. Quantum Chem.*, **103**, 808–817.
33 Karle, I.L., Perozzo, M.A., Mishra, V.K., and Balaram, P. (1998) *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 5501–5504.
34 Karle, I.L., Gopi, H.N., and Balaram, P. (2003) *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 13946–13951.
35 Karle, I.L., Gopi, H.N., and Balaram, P. (2002) *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 5160–5164.
36 Ravindra, G., Ranganayaki, R.S., Raghothama, S., Srinivasan, M., Gilardi, R.D., Karle, I.L., and Balaram, P. (2004) *Chem. Biodiver.*, **1**, 489–504.
37 Karle, I.L., Prasad, S., and Balaram, S. (2004) *J. Peptide Res.*, **63**, 175–180.
38 Gopi, H., Roy, R.S., Raghothama, S.R., and Karle, I. (2002) *Helv. Chim. Acta*, **85**, 3313–3330.
39 Karle, I.L., Awasthi, S.K., and Balaram, P. (1996) *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 8189–8193.
40 Karle, I., Gopi, H.N., and Balaram, P. (2001) *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 3716–3719.
41 Karle, I.L., Das, C., and Balaram, P. (2000) *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 3034–3037.
42 Vijayalakshmi, S., Rao, R.B., Karle, I.L., and Balaram, P. (2000) *Biopolymers*, **53**, 84–98.
43 Huang, L., Massa, L., and Karle, J. (2006) *Int. J. Quantum Chem.*, **106**, 447–457.
44 Hehre, W.J., Stewart, R.F., and Pople, J.A. (1969) *J. Chem. Phys.*, **51**, 2657–2664.
45 Collins, J.B., Schleyer, P.v.R., Binkley, J.S., and Pople, J.A. (1976) *J. Chem. Phys.*, **64**, 5142–5151.
46 Binkley, J.S., Pople, J.A., and Hehre, W.J. (1980) *J. Am. Chem. Soc.*, **102**, 939–947.
47 Gordon, M.S., Binkley, J.S., Pople, J.A., Pietro, W.J., and Hehre, W.J. (1982) *J. Am. Chem. Soc.*, **104**, 2797–2803.
48 Pietro, W.J., Francl, M.M., Hehre, W.J., Defrees, D.J., Pople, J.A., and Binkley, J.S. (1982) *J. Am. Chem. Soc.*, **104**, 5039–5048.
49 Dobbs, K.D. and Hehre, W.J. (1986) *J. Comput. Chem.*, **7**, 359–378.
50 Dobbs, K.D. and Hehre, W.J. (1987) *J. Comput. Chem.*, **8**, 861–879.
51 Dobbs, K.D. and Hehre, W.J. (1987) *J. Comput. Chem.*, **8**, 880–893.

52 Schaefer, A., Horn, H., and Ahlrichs, R. (1992) *J. Chem. Phys.*, **97**, 2571–2577.

53 Schaefer, A., Huber, C., and Ahlrichs, R. (1994) *J. Chem. Phys.*, **100**, 5829–5835.

54 Ditchfield, R., Hehre, W.J., and Pople, J.A. (1971) *J. Chem. Phys.*, **54**, 724–728.

55 Hehre, W.J., Ditchfield, R., and Pople, J.A. (1972) *J. Chem. Phys.*, **56**, 2257–2261.

56 Hariharan, P.C. and Pople, J.A. (1974) *Mol. Phys.*, **27**, 209–214.

57 Gordon, M.S. (1980) *Chem. Phys. Lett.*, **76**, 163–168.

58 Hariharan, P.C. and Pople, J.A. (1973) *Theo. Chim. Acta*, **28**, 213–222.

59 Blaudeau, J.P., McGrath, M.P., Curtiss, L.A., and Radom, L. (1997) *J. Chem. Phys.*, **107**, 5016–5021.

60 Francl, M.M., Pietro, W.J., Hehre, W.J., Binkley, J.S., DeFrees, D.J., Pople, J.A., and Gordon, M.S. (1982) *J. Chem. Phys.*, **77**, 3654–3665.

61 Binning, R.C. Jr. and Curtiss, L.A. (1990) *J. Comput. Chem.*, **11**, 1206–1216.

62 Rassolov, V.A., Pople, J.A., Ratner, M.A., and Windus, T.L. (1998) *J. Chem. Phys.*, **109**, 1223–1229.

63 Rassolov, V.A., Ratner, M.A., Pople, J.A., Redfern, P.C., and Curtiss, L.A. (2001) *J. Comput. Chem.*, **22**, 976–984.

64 Dunning, T.H. Jr. and Hay, P.J. (1976) *Modern Theoretical Chemistry III*, vol. 3 (ed. H.F. Schaefer), Plenum, New York, pp. 1–28.

65 Petersson, G.A. and Al-Laham, M.A. (1991) *J. Chem. Phys.*, **94**, 6081–6090.

66 Petersson, G.A., Bennett, A., Tensfeldt, T.G., Al-Laham, M.A., Shirley, W.A., and Mantzaris, J. (1988) *J. Chem. Phys.*, **89**, 2193–2218.

67 Woon, D.E. and Dunning, T.H. Jr. (1993) *J. Chem. Phys.*, **98**, 1358–1371.

68 Kendall, R.A., Dunning, T.H. Jr., and Harrison, R.J. (1992) *J. Chem. Phys.*, **96**, 6796–6806.

69 Dunning, T.H. Jr. (1989) *J. Chem. Phys.*, **90**, 1007–1023.

70 Peterson, K.A., Woon, D.E., and Dunning, T.H. Jr. (1994) *J. Chem. Phys.*, **100**, 7410–7415.

71 Wilson, A., Mourik, T.v., and Dunning, T.H. Jr. (1997) *J. Mol. Struct. (Theochem)*, **388**, 339–350.

72 Dewar, M. and Thiel, W. (1977) *J. Am. Chem. Soc.*, **99**, 4499–4450.

73 James, J.P. (2001) A major enhancement in computational chemistry accuracy: MOPAC 2002, Stewart Computational Chemistry, Fujitsu Computational Chemistry Seminars.

74 Roothan, C.C.J. (1951) *Rev. Mod. Phys.*, **23**, 69–89.

75 Kohn, W. and Sham, L.J. (1965) *Phys. Rev.*, **140**, A1133–A1138.

76 Pople, J.A., Seeger, R., and Krishnan, R. (1977) *Int. J. Quantum Chem. Symp.*, **11**, 149–163.

77 Moller, C. and Plesset, M.S. (1934) *Phys. Rev.*, **46**, 618–622.

78 Bartlett, R.J. and Purvis, G.D. (1978) *Int. J. Quantum Chem.*, **14**, 561–581.

79 Sanger, F. and Tuppy, H. (1951) *Biochem. J.*, **49**, 463–481.

80 Hodgkin, D. (1970) *Verh. Schweiz. Naturforsch. Ges.*, **150**, 93–101.

81 Gursky, O., Badger, J., Li, Y., and Caspar, D. (1992) *Biophys. J.*, **63**, 1210–1220.

82 Huang, L., Massa, L., and Karle, J. (2005) *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 12690–12693.

83 Wahl, M.C., Rao, S.T., and Sundaralingam, M. (1996) *Biophys. J.*, **70**, 2857–2866.

84 Leonard, G.A., Hambley, T.W., and Mcauleyhecht, K. (1993) *Acta Crystallogr., Sect. D: Biol. Crystallogr*, **49**, 458–467.

85 Soler-Lopez, M., Malinina, L., Tereshko, V., Zarytova, V., and Subirana, J.A. (2002) *J. Biol. Inorg. Chem.*, **7**, 533–538.

86 Vargason, J.M., Henderson, K., and Ho, P.S. (2001) *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 7265–7270.

87 Tari, L.W. and Secco, A.S. (1995) *Nucleic Acids Res.*, **23**, 2065–2073.

88 Valls, N., Uson, I., and Subiriana, C.J.A. (2004) *J. Am. Chem. Soc.*, **126**, 7812–7816.

89 Qiu, H., Dewan, J.C., and Seeman, N.C. (1997) *J. Mol. Biol.*, **267**, 881–898.

90 Rozenberg, H., Rabinovich, D., Frolow, F., and Hegde, R.S. (1998) *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 15194–15199.

91 Nunn, C.M. and Neidle, S.J. (1995) *Med. Chem.*, **38**, 2317–2325.

92 Egli, M., Williams, L.D., Gao, Q., and Rich, A. (1991) *Biochemistry*, **30**, 11388–11402.

93 Shakked, Z., Rabinovich, D., Kennard, O., Cruse, W.B., Salisbury, S.A., and Viswamitra, M.A. (1983) *J. Mol. Biol.*, **166**, 183–201.

94 Huang, L., Massa, L., and Karle, J. (2005) *Biochemistry*, **44**, 16747–16752.

95 Basavappa, R. and Sigler, P.B. (1991) *EMBO J.*, **10**, 3105–3111.

96 Huang, L., Massa, L., and Karle, J. (2006) *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 1233–1237.

97 Russell, R., Murray, J., Lentzen, G., Haddad, J., and Mobashery, S. (2003) *J. Am. Chem. Soc.*, **125**, 3410–3411.

98 Huang, L., Massa, L., and Karle, J. (2007) *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 4261–4266.

99 Delacoux, F., Fichard, A., Geourjon, C., Garrone, R., and Ruggiero, F. (1998) *J. Biol. Chem.*, **273**, 15069.

100 Huang, L., Massa, L., and Karle, J. (2007) *J. Chem. Theory Comput.*, **3**, 1337–1341.

101 Huang, L., Massa, L., and Karle, J. (2008) *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 1849–1854.

102 Frisch, M.J., Trucks, G.W., Schlegel, H.B. *et al.* (2003) Gaussian 03, Gaussian, Inc., Pittsburgh, PA.

103 Zhang, X., Green, T.J., Tsao, J., Qiu, S., and Luo, M. (2008) *J. Virol.*, **82**, 674–682.

104 (2007) HyperChem 8.0.3 for Windows, Hypercube, Inc, Gainesville, FL.

105 Huang, L., Massa, L., and Karle, J. (2009) *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 1731–1736.

106 Stillinger, F.H. *et al.* (1995) *Mathematical Challenges from Theoretical/Computational Chemistry*, National Academy Press, Washington.

107 Lipscomb, W.N. (1972) *Trans. Am. Cryst. Assoc.*, **8**, 79–92.

108 Provenzano, A. (2009) *Oncol. Times*, **31**, 3–4.