1/26/94

Protein Simulations using Techniques Suitable for Very Large Systems: the Ccll Multipole Method for Nonbond interactions antl the Newton-Euler Inverse Mass Operator Method for Internal Coordinate Dynamics

Alan M. Mathiowetz,[†#] Abhinandan Jain,[‡] Naoki Karasawa,[†] and William A. Goddard III[†*]

Materials and Molecular Simulation Center, Beckman Insitute (139-74)
California Institute of Technology, Pasadena, California 91125

Two new methods developed for molecular dynamics simulations of very large proteins are applied to a series of proteins ranging up to the protein capsid of tomato busily stunt virus (TBSV).

For molecular dynamics of very large proteins and polymers, it is useful to carry out the dynamics using internal coordinates (say, torsions only) rather than Cartesian coordinates. This allows larger time steps, eliminates problems with the classical description of high energy modes, and focuses on the important degrees of freedom. The resulting equation of motion has the form

$$\mathcal{M}(\theta)\ddot{\theta} + C(\theta,\dot{\theta}) = T(\theta)$$

where for T is the vector of generalized forces, $\mathcal{M}(\theta)$ is the moments of inertia tensor, $\ddot{\theta}$ is the vector of torsions and C is a vector containing Coriolis forces and nonbond forces. The problem is that to calculate the acceleration vector $\ddot{\theta}$ from M, C and T requires inverting $M(0)$, an order N^3 calculation. Since the number of degrees of freedom might

[†#] Current address: Sterling Winthrop, Inc., 1'250 South Collegeville Road, 1'. 0. Box 5000, Collegeville, Pennsylvania 19426-0900.

[‡] Jet 1'repulsion Laboratory

[†] Division of Chemistry and Chemical Engineering (CN 8921)

[†*] Materials and Molecular Simulation Center, Beckman Institute (139-74). To whom correspondence Should be addressed.

be 300,000 for a million atom system, solving these equations every time step is impractical, restricting internal coordinate methods to small systems. The new method, Newton-Euler Inverse Mass Operator (NEIMO) dynamics, constructs the torsional accelerations vector $\ddot{\theta} = M^{-1}(T - C)$ directly by an order N process, allowing internal-coordinate dynamics to be solved for super large (million atom) systems. The first use of the NEIMO method for molecular clynall-lies of proteins is presented here.

A second serious difficulty for large proteins is calculation of the nonbond forms. We report here the first application to proteins of the new Cell Multipole Method (CMM) to evaluate the Coulomb and van der Waals interactions. The cost of CMM scales linearly with the number of particles while retaining an accuracy significantly better than standard 110111.)01)(1 methods (involving; cutoffs).

Results for NEIMO and CMM are given for simulations of a wide range of peptide and protein systems, including the protein capsid of TBSV with **488,000** atoms. The computational time fox NEIMO and CMM are demonstrated to scale linearly with size. With NEIMO the dynamics time steps can be as large as '20 fs (for small peptides), much larger than possible with standard Cartesian coordinate dynamics.

For TBSV we considered both the normal form and the high pH form in which the Ca^{2+} ions are removed. These talc.ulaticn)s lead to a contraction of the protein for both forms (probably because of ignoring the RNA core not observed in the X-ray).

1. 1 ntroduction

Molecular dynamics simulations have become invaluable for such diverse tasks as building protein models from crystallographic data[1] and determining the relative free energy of binding for a variety of drug molecules to a common receptor.[2] Despite the advances, many problems of chemistry and biology seem completely outside the reach of current methodologies. For example, starting with the X-ray diffraction structure for the protein capsid of poliovirus, we would like to use molecular dynamics to predict the structure of the RNA inside the protein capsid, a simulation involving long-term simulations of over 1,000,000 particles. Major advances in computer hardware (including vector processing supercomputers, RISC workstations, and massively parallel supercomputers) have

allowed the extension of current methods to larger and more complex systems. Even more important are the advances in software which involve optimization of architectures,[3] improved efficiency in calculating interatomic forces,[4,5] and development of techniques for allowing larger time steps in molecular dynamics simulations.[6-12]

Molecular dynamics simulations typically involve numerical integration of Newton's equations of motion,

$$M\ddot{x} = -F \tag{1}$$

Time steps for the integration must be sufficiently small that the fastest modes are handled accurately. Systems containing explicit hydrogen atoms typically require time steps of about 1 femtosecond (1 fs = 10^{-15} S) for accurate results.

The most popular approach for increasing time steps is to fix the fastest degrees of freedom (bond stretches and angles) and to solve the equations of motion for the slower (torsional) degrees of freedom. Such an approach is especially justified for studies of large biological molecules, where bond lengths and angles vary little from one structure to another and nearly all important conformational transitions are due to torsional motions. [An alternative approach for increasing time steps is to separate short and long-range forces and use different time steps for the different forces.[6]]

The SHAKE algorithm[7] has become the standard approach for doing molecular dynamics with fixed bond lengths. It can also be used to hold angles fixed, but this is less effective.[29] SHAKE is a modification of the Verlet algorithm for integrating the equations of motion for the $3n - 6$ Cartesian coordinates degrees of freedom in an n-particle systems. Particle velocities are calculated first for the unconstrained system, and then modified to meet each constraint. An iterative process is required to meet all the constraints concurrently. The SHAKE algorithm has been successfully used for time steps up to 4 fs,[9,10] enabling a speedup in computational time that is partially balanced by the costs of iteratively solving the constraint equations.[9]

An alternative to the SHAKE methodology of solving Cartesian coordinate dynamics with constraints is to *solve the equations of motion directly for the internal degrees of*

freedom. This leads to equations of the form

$$\mathcal{M}(\theta)\ddot{\theta} - 1\ \mathcal{C}(\theta,\dot{\theta}) = \mathcal{T}(\theta) \tag{2}$$

where $\mathcal{T}(\theta)$ is a vector containing all t orques (or other generalized forces), $\mathcal{C}$ is a vector describing nonbond forces and external fields, $\mathcal{M}$ is the moment of inertia tensor (the mass matrix), and $\ddot{\theta}$ is the vector of angular accelerations (generalized accelerations). At a ny particular time step, M, $\mathcal{C}$, and $\mathcal{T}$ *arc* known and $\ddot{\theta}$ must be calculated to obtain the θ and $\dot{\theta}$ for the next time step. When only torsional degrees of freedom are allowed, solutions to (2) automati cally fulfill the desired bond length and/or angle constraints, so their efficiency is not limited to a secondary constraint-solving step. Indeed, Mazur et al.[14] were able to simulate accurately a small polypeptide, (Ala)g, with time steps as large as 13 fs, a significant improvement over the S}1 AKE algorithm. The problem is that for $\mathcal{N}$ degrees of freedom M *is* an $\mathcal{N}$ by $\mathcal{N}$ matrix and solving (2) requires a time proportional to $\mathcal{N}^3$, which becomes prohibitive for large systems.

Recently, Jain et al.[11,12] developed an alternative method for solving the equations of motion for internal coordinates. This new *Newton-Euler Inverse Mass Operator* (NEIMO) m eth od, does not require direct manipulation of matrices, and leads to computational times proportional to $\mathcal{N}$ rather than $\mathcal{N}^3$. ' Jhe methodology was (1 CW1O]KK1 for spacecraft dynamics, but in a separate report, Jain *et al.*[12] described how the method could be applied to molecular dynamics. This report presents the first implementation of the NEIMO method fox molecular systems. We have studied the dynamics of polypeptide systems and find that we are able to talc.ulatc accurately the dynamics of some systems with time steps as large as 20 fs. Because the computational costs using NEIMO arc rigorously proportional to $\mathcal{N}$, it, can be applied to very large systems. Actual application of NEIMO to systems as large as the tomato bushy stunt virus crystal structure,[24] (with an asymmetric unit of over 8000 atoms distributed along three chains tot alling nearly 900 residues) snow that the costs are indeed proportional to $\mathcal{N}$.

11. Methodology

A. NEIMO

In standard molecular dynamics calculations the independent variables are the $3n$ cartesian degrees of freedom, leading to Newton's equations of motion in the form

$$m_i \ddot{x}_{\alpha i} = F_{\alpha i}. \tag{1}$$

Here, m_i is the mass of particle i, $\ddot{x}_{\alpha i}$ is the α component of the acceleration for particle i, and $F_{\alpha i}$ is the α component of the force for particle i. At each time step the unknown acceleration is calculated from (1) for each of the $3n$ Cartesian degrees of freedom by dividing by m_i,

$$\ddot{x}_{\alpha i} = \frac{1}{m_i} F_{\alpha i}. \tag{2}$$

in internal coordinates the dynamical equations of motion are

$$\mathcal{M}(\theta)\ddot{\theta} + \mathcal{C}(\theta, \dot{\theta}) = T(\theta). \tag{3}$$

where θ is the set of generalized internal coordinates (e.g., torsion angles), $\mathcal{C}$ is the set of nonlinear forces (Coriolis plus nonbond), T is the set of generalized forces (torques in the case of torsional degrees of freedom), and $\mathcal{M}(\theta)$ is the moment of inertia tensor (mass matrix). For a system with $\mathcal{N}$ internal degrees of freedom, the $\mathcal{N}$ degrees of freedom are coupled, leading to off-diagonal elements in the mass matrix, M, with a nonlinear dependence on θ. Thus solving (3) for $\ddot{\theta}$ requires computing

$$\ddot{\theta} = \mathcal{M}^{-1}(\theta)[T(\theta) - \mathcal{C}(\theta, \dot{\theta})]. \tag{4}$$

at each time step.[4] The computational cost of solving this matrix equation is proportional to $\mathcal{N}^3$, which becomes prohibitive for large molecules.

Recently, Jain, *et al.*[12] developed a recursive algorithm for solving the equations of motion (3) which computes the right hand side of (4) *without* explicitly solving the $\mathcal{N} \times \mathcal{N}$ matrix equations in (3). Instead this NEIMO method uses *spatial operator algebra* in a recursive approach to calculate (4) directly in a procedure where the *computational effort*

is rigorously proportional to $\mathcal{N}$, making a whole new class of very large molecular systems available for study by internal-coordinate molecular dynamics.

The NEIMO methodology has been developed for general multibody systems configured as serial chains, topological trees, or Closed-loop systems.[11,12] Our first implementation for molecular systems reported here is for serial chain and tree topologies. (This includes all proteins without disulfide linkages or prosthetic groups having multiple attachment sites). Extensions to Closed-loop topologies and periodic systems have since been completed.[31]

NEIMO USES the concepts Of "clusters" and "hinges" to describe a molecular system. A *cluster* is an atom or group of atoms that moves as a rigid unit; this could be a single atom, a multiple-atom group such as a methylene group, a phenyl ring, or even an entire domain of a protein. A hinge describes the relative orientation between two connected clusters; in a molecular system, each hinge is a bond connecting two adjacent clusters. There are six possible degrees of freedom (dof) for each hinge. Special cases include torsions-only (1 dof) and all-angles (5 clef). Here we will concentrate on the torsions-only case, with each hinge limited to a single torsional dof. In addition to the internal dof, each connected chain of molecules is referenced to an absolute reference coordinate system. This is done by considering one cluster as the base and using a hinge with the full six degrees of freedom to describe the absolute orientation and position in space for this cluster.

The relationship between adjacent clusters is described in terms of "parents" and "children." Each cluster can have one or more attached child clusters; and is the parent of each of these children. In a topological tree, outward branching proceeds with each cluster having zero, one, or more children, but each child having only one parent cluster. Clusters at the far extent of each branch are termed "tips" and have no children. A serial chain is a linear polymer having the base at one end and a single tip cluster at the other. Between the base and tip clusters, each cluster has a unique parent and unique child. In a protein, most of the C_α atoms are branch points with two children, and the outermost cluster Of every sidechain is a tip. These concepts are illustrated in Figure 1, where the pentapeptide Met-enkephalin is shown with the hinges numbered. Hinge 0, which connects the base

cluster to the reference frame, is not shown. Each cluster has a unique hinge connecting to its parent cluster. The clusters and hinges of Met-enkephalin are described in Table 1. The torsional degree of freedom for each hinge (other than hinge O) is defined in terms of a *specific* dihedral angle. These dihedrals are also listed in Table 1, using the standard nomenclature of protein dihedrals.

The NEIMO method uses spatial operator algebra to simplify and solve the equations of motion for multibody systems.[12] Using spatial notation, the Newton-Euler recursive equations of motion for a tree-topology system are:

$$V(k) = \phi^*(p, k)V(p) + H^*(k)\dot{\theta} \tag{5a}$$

$$\alpha(k) = \phi^*(p, k)\alpha(p) + H^*(k)\ddot{\theta}(k) + a(k) \tag{5b}$$

and

$$f(k) = \sum_c \phi(k, c)f(c) + M(k)\alpha(k) + b(k) + \hat{f}_c(k) \tag{6a}$$

$$T(k) = H(k)f(k) \tag{6b}$$

(where * indicates transpose).

These equations describe the relationship between a cluster k and its parent (p) and child (c) clusters, as well as the relationship between spatial variables [e.g. spatial velocity $V(k)$] and generalized variables [e.g. generalized velocity $\dot{\theta}(k)$]. Equations 5a and 5b define the relationships between velocities and accelerations in the space fixed coordinate system, $V(k)$ and $\alpha(k)$, in terms of the succession of body fixed $V(p)$ and $\alpha(p)$. This starts with k = base and proceeds to all tips (that is, it proceeds from parent (p) to child (c)). Likewise, Equations 6a and 6b define the forces and torques in the space fixed coordinate system, $f(k)$ and $T(k)$, in terms of the inter-cluster interaction forces $f(c)$ plus the forces derived from non-bonded interactions, $\hat{f}_c$. This proceeds from k = tips to k = base (that is from child to parent)

The *spatial transformation matrix* $\phi(k, c)$ transforms spatial force quantities from the frame of reference of the child cluster c to its parent cluster, k; its transpose, $\phi^*(k, c)$ transforms velocities and accelerations from parent to child. The *hinge matrix* $H^*(k)$

describes the spatial velocity across the k hinge.[11,12] $H^*(k)$ is a 6 x dof dimensional matrix whose form depends upon the nature of the hinge motion (dof = 1 for torsions-only). The spatial velocities derived are used[11,12] to calculate the *Coriolis accelerations* $a(k)$ and the *spatial gyroscopic forces*, $b(k)$.

The NEIMO method allows one to use Cartesian forces as well as torques to calculate the dynamic%. For our calculations, we have used Cartesian forces exclusively, as these are already calculated by BIOGRAF. Future implementations of NEIMO will use Cartesian forces only for nonbonded interactions (which are Cartesian in nature) while using torques derived directly from the torsional potentials.

Spatial variables are specified in terms of the six degrees of freedom (three angular and three linear) of each cluster. The spatial velocity V(k) combines angular and linear velocities of the duster while the spatial force j(k) combines angular forces (moments or torques) and linear forces (Cartesian forces). *Spatial operators* can be used to express these recursive relationships very concisely. For instance, using spatial operators, the equation for v'(k) becomes

$$V = \mathcal{E}_\phi^* V + H^* \dot{\theta}. \tag{7}$$

The spatial operator H^* is a $6n\mathcal{N}$ by $6n\mathcal{N}$ block diagonal matrix defined by $H^* = diag\{H^*(1) \text{“ . “ } 11\text{”(?1)}\}$. The other spatial operators are defined similarly,[12] leading to the following factored expression for the mass matrix:

$$\mathcal{M} = H\phi M \phi^* H^*.$$

The NEIMO method for solving the equations of motion is based upon expressions for an alternative factorization of the mass matrix and its inverse. Using spatial operators, the *Innovations Operator Factorization* of the mass matrix[12] has the form

$$\mathcal{M} = [I + H\phi K]\, D\, [I + H\phi K]^*. \tag{8}$$

while the inverse of the mass matrix has the form[12]

$$\mathcal{M}^{-1} = [I - H\psi K]^* D^{-1} [I - H\psi K] \tag{9}$$

Therefore, the generalized accelerations $\ddot{\theta}$ in *(4)* are calculated by inverting the clef(k) x dof(k) matrices, $D(k)$, rather than by inverting the entire $\mathcal{N}$ x $\mathcal{N}$ mass matrix, M. For torsions-only, clef = 1.

With NEIMO the computation of $\ddot{\theta}$ is carried out in several recursive steps, each of which is linear in $\mathcal{N}$.

2. Velocity (V) step: an outward recursion from base to tips to calculate the spatial velocities, $V(k)$, from the geometry and hinge velocities, $\dot{\theta}(k)$, as in Equation 5a.

ii. Spatial Inertia (MKD) Step: Calculation of a number of dof(k) x dof(k) matrices, $P(k), D(k)$, etc.. related to the forces. This proceeds from tip to base (child to parent).

$$
\begin{aligned}
P(k) &= \sum_{c} \phi(k, c)\, P^{A}(c) \\
D(k) &= H(k)\, P(k) H^{*}(k) \\
G(k) &= P(k) H^{*}(k) D^{-1}(k) \\
K(p, k) &= \phi(p, k) G(k) \\
\bar{\tau}(k) &= I - G(k) H(k) \\
P^{A}(k) &= \bar{\tau}(k) P(k) \\
\psi(p, k) &= \phi(p, k) \bar{\tau}(k) \\
z(k) &= \sum_{c} \phi(k, c) z^{A}(c) + P(k) a(k) + b(k) + \hat{f}_{c}(k) \\
\epsilon(k) &= T(k) - H(k) z(k) \\
\nu(k) &= D^{-1}(k) \epsilon(k) \\
z^{A}(k) &= z(k) + G(k) \epsilon(k)
\end{aligned}
\tag{10}
$$

iii. Torsional Acceleration ($\ddot{\theta}$) Step: Calculates the angular acceleration, $\ddot{\theta}$, in terms of the cluster accelerations, $\alpha(k)$, and Coriolis accelerations, $a(k)$, effective torques, c, mass inverses, v, etc.

$$
\begin{aligned}
\alpha^{A}(k) &= \phi^{*}(p, k) \alpha(p) \\
\ddot{\theta}(k) &= \nu(k) - G^{*}(k)\, \alpha^{A}(k) \\
\alpha(k) &= \alpha^{A}(k) + H^{*}(k)\, \ddot{\theta}(k) + a(k)
\end{aligned}
\tag{11}
$$

iv. New velocities and coord inates ($\dot{\theta}$ θR). The accelerations $\ddot{\theta}$ are 11s('(1 to update the torsional velocities ($\dot{\theta}$), torsional angle (θ), and Cartesian coordinat es (R) of the system using the integrator described below. This step is done simultaneously with $\ddot{\theta}\theta R$.

The equations of motion were integrated using the "leapfrog" Verlet algorithm.[8] The Verlet algorithm calculates accelerations and velocities at alternating half time steps. Since the accelerations in NEIMO dynamics are not independent of velocities, the half time step separation of accelerations and velocit ies must be modified. As described in detail in the Appendix we solved iteratively for the velocities at integer time steps (very fast). A major virtue of th e Verlet algorithm is that it requires only a single calculation Of the forces at eachtime step. In simulations of large systems, the force calculation consumes the vast majority of computational time, so that methods requiring only a single force calculation are preferable to methods which require two or more force calculations per time step, such as the Gear predictor-corrector algorithm.[15] Other integration schemes are being investigated for use in NEIMO dynamics, but all results presented here use the leapfrog Verlet algorithm.

The NEIMO calculations presented here were performed using a version of the program written to work with the BIOG RAF/I'OLYG R AF program from Molecular Simulations, Inc[16] All calculations were performedon iris PowerSeries and Iris Indigo workstations from Silicon Graphics, Inc.

B. CMM

The Cell Multipole Method (CMM) is described in reference 4 (the CMM module described in reference 4 was adapted to BIOGRAF/P OLYG RAF). With CMM we place the molecule in a box and divide the box into eight children cells, each child cell into eight grandchildren cells, etc., until there are about 4 particles in the smallest cdl (the microcell). Thus for TBSV the full virus is placed in a box having sides of 341.8 Å and a hierarchy of 6 levels is used (262,144 level 6 cells). The charge, dipole moments, and quadrupole moments (both Coulomb and vdW) are calculated for each microcell and used to ol btain the momen its of the parent cells. In describing the Coulomb and vdW interactions for the atoms in some microcell, we explicitly calculate the interactions with each of the

particles in the same cell and in the 26 adjacent cells; this is denoted as the near field, V_{near}. The interactions with all other particles use the multipole fields. The cells are grouped so that the fields from larger (higher level) cells are used for regions farther from the cell of interest. These multipole fields are expanded in a Taylor series about the center of each cell, allowing rapid calculation of the energy and forces for each particle in the cell of interest. The total multipole field is denoted, V_{far}. Thus the potential energy is written as

$$V_{CMM}(R) = V_{near}(R) + V_{far}(R) \tag{12}$$

Each step i the process is rigorously linear in n (the number of particles) for a constant particle density.

For systems such as TBSV the computational time for V_{near} and V_{far} are approximately the same. However Wc find that V_{far} is relatively constant from step to step so that V_{far} needs be updated only every 50 time steps. The net result is that the total computational cost is that of calculating the near field (about 50 interactions per particles).

For TBSV the total calculation of all Coulomb and vdW interactions between all particles in one asymmetric unit (8083 atoms) with all particles of the whole protein (484,980 atoms) requires only 1.86 times the time (92.6 s vs. 49.8 s) for the interactions within the asymmetric unit alone. This is .003% of the time estimated to do all nonbond interactions (3.2 **x** 106s).

C. TBSV

The crystal structure of TBSV was obtained by assuming exact icosahedral symmetry for the protein capsid. in order to use BIOGRAF for calculating the valence forces and organizing the input and output, we added a symmetry mapping module (SYMMAP). Thus on each iteration the process was as follows:

1. usc BIOGRAF to calculate Cartesian valence forces (DREIDING force field) for the 8083 atoms asymmetric unit
2. usc SYMMAP to obtain the 484,980 atoms of the full protein capsid
3. USC CMM to calculate the Coulomb and vdW forces of all *484,980* atoms on the 8083 atoms of the asymmetry unit

4. use NEIMO with the valence forces from (1) and the Coulomb and vdW forces from (3) to calculate the accelerations, and thus the velocities and coordinates for the next time step.

5. return to (1).

Calculations on a simple workstation are not really practical for long term dynamics of viruses but we were able to perform 5000-step simulations (e.g. 10 ps using 2 fs time steps) in roughly 3 days (58 hours of CPU time on one processor of an SGI4D/380 workstation).

III. Results

NEIMO calculations were carried out on a wide variety of peptide and protein systems, ranging from the five-residue peptide Met-enkephalin (denoted MEnk) to the tomato bushy stunt virus (TBSV) protomer, which contains three proteins totaling 893 residues (8083 atoms). Table 2 contains a list of the ten systems studied. The two peptides (MEnk and Ala9) were built using the Peptide Builder of BIOGRAF,[16] which uses standard amino acid geometries. They were initially configured as alpha helices, but were minimized to a local potential energy minimum using conjugate gradients minimization. As in all calculations reported here, the DREIDING force field[25] was used for these minimizations. No solvent or counterions were used, but the dielectric constant was taken as distance-dependent ($\epsilon = r$). This provides a crude representation of the electrostatic shielding of aqueous solvent. For these small peptides, no nonbond cutoff was used; i.e., all possible pairs were included in the van der Waals and electrostatic calculations.

The initial conformations of the eight proteins were derived from the X-ray crystal structures listed in Table 2. All metal ions, solvent molecules, and disulfide bridges were removed, leaving only protein chains conforming to a tree topology. (As mentioned above, sidechain aromatic rings and proline rings are treated as single clusters.) Hydrogen atoms were then added to heteroatoms using the BIOGRAF hydrogen builder). As for the peptides, the DREIDING force field was used to energy-minimize these conformations. The large size of the proteins precluded the inclusion of all possible nonbond pairs, a number close to $\frac{1}{2} n^2$ for an n-atom protein. Therefore, CMM was used to calculate the van der Waals and electrostatic interactions.

A. Timing

Timing results for the ten systems are shown in Table 3. The times represent the average of 100 dynamics steps run on an iris Indigo (R3000) workstation. Times are given for both the NEIMO calculations and the nonbond calculations, the latter of which consumes the vast majority of CPU time, even when a very fast method such as CMM is used. The NEIMO timing is shown to be rigorously proportional to $\mathcal{N}$ for the proteins with over 400 atoms. For these systems, the NEIMO calculations take up less than .5% of the total CPU time. Since NEIMO scales linearly with size, there is no longer a practical limit to the size of system which can be simulated using internal-coordinate dynamics. Calculation of the nonbonded interactions is the limiting factor, as it is for Cartesian dynamics calculations.

As indicated in Table 3 and Figure 2, the CMM calculations are proportional to n, leading to times that are $n/500$ times faster than the exact calculations. The CMM calculations are not exactly proportional to n because protein systems are not homogeneous. Shape and density variations cause variations in the number of atoms per microcell. However, these variations are not themselves proportional to n, so the n-proportionality holds. These applications use the original CMM program developed for testing the algorithm.[4] It has since been optimized and parallelized.[30]

B. Energy Fluctuations

A primary advantage of internal-coordinate methods of molecular dynamics is the ability to use larger time steps than the 1 fs step size typically required for Cartesian molecular dynamics. A good measure of accuracy for dynamics calculations is the fluctuation in total energy. In microcanonical dynamics, the total energy of the system

$$E = K + V \tag{13}$$

should be constant, even though its component potential energy, V, and kinetic energy, K, fluctuate. The energy fluctuation $\mathcal{E}$ is defined by

$$\mathcal{E} = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T} \tag{14}$$

where k_B is the Boltzmann constant and T is the temperature of the simulation.

It is common practice to keep the temperature of a microcanonical dynamics simulation roughly constant by periodically scaling the velocities. Other calculations which must be done periodically, such as updating a list of nonbond pairs within a given cutoff distance or reassigning atoms to cells in CMM can be done at the same time the velocities are rescaled. This is particularly important for large systems, where calculation of all nonbonded interactions for every time step is prohibitive. Under such conditions, where nonbonds and velocities are updated periodically, the total energy, E, does not remain constant throughout the entire time of the simulation. Thus the energy fluctuation $\mathcal{E}$ from (7) no longer provides an accurate measure of the dynamics because the reference E is different in each period. Application of (7) then would lead to $\langle E^2 \rangle$ diverging from $\langle E \rangle^2$. Instead, we use the average fluctuation, $\langle \mathcal{E} \rangle$, determined by calculating $\mathcal{E}$ during each 0.100 ps interval, and averaging. If the total calculation has N_i 0.100 ps intervals, $\langle \mathcal{E} \rangle$ is defined by

$$\langle \mathcal{E} \rangle = \frac{1}{N_i} \sum_{i=1}^{N_i} \mathcal{E}_i, \tag{15}$$

where $\mathcal{E}_i$ is the energy fluctuation Calculated during the 2-th interval. In such calculations, time steps should be chosen so that they given an integral number of dynamics steps per 0.100 ps - for instance, a time step of 3,0303 fs is used, rather than 3.0 fs.

B.1 Met-enkephalin

Figure 3 shows the values of $\mathcal{E}$ obtained from 1 picosecond (10^{-12} s) simulations of the pentapeptide Met-enkephalin (NH_3^+-Tyr-Gly-Gly-Phe-Met-COO^-) for NEIMO (AT) and Cartesian (C) dynamics simulations at time steps ranging from 1 fs to 20 fs.

For Cartesian dynamics simulations the initial fluctuations were significantly higher and we equilibrated (for about 1 picosecond) using 1 fs time steps *before* starting the calculation of $\mathcal{E}$. The NEIMO simulations did not require an equilibration phase. Cartesian dynamics simulations using time steps greater than 3 fs led to exaggerated particle motions from one time step to the next and the energy quickly diverged. For Cartesian dynamics of large systems, our experience is that time steps must be restricted to 1 fs for robust performance. Even for the small peptide Met-enkephalin, a . ? fs time step gives rise to

energy fluctuations 11101'(! than 10 times as large as a 1 fs simulation.

In contrast, NEIMO dynamics simulations arc quite stable. With time steps as large as 18 fs, we found small fluctuations, smaller even th an the Cartesian dynamics simulat ion with a 1 fs time step. A fairer comparison might be to divide the energy fluctuations by the numl)er of degrees of freedom. For Met -enkephalin, $\mathcal{N}$ = 28 (22 dihedral angles plus the six degrees of freedom for the base body), while the number of degrees of freedom in Cartesian dynamics is $3n$ - G, or 138. The scaled fluctuations, $\mathcal{E}*$, are also shown in Figure 3 and are labeled with an asterisk (N* and C*). NEI MO time steps as large as 12 fs gave smaller scaled fluctuations than the 1 fs Cartesian simulations.

B.2 $(A\,lA)_9$

Similar results were obtained for Hillc-residue polyalanine, $(Ala)_9$. Cartesian dynami cs were reliable only at 1 fs and 2 fs time Steps. The 3 fs simulation did not diverge, but the flue.tuatiolls were extremel y large. The scaled fluctuations, $\mathcal{E}*$, were very similar for 1 fs and 2 fs Cartesian dynamics of both peptides. The NEI MO simulations of $(Ala)_9$ gave larger values of $\mathcal{E}$ and $\mathcal{E}*$ than for Met-c]lkq)lja]i]l at almost every time step, but the fluctuations did not diverge until time steps larger than 30 fs were used. It is likely that $(Ala)_9$ is able to tolerate such large time steps because it has no light sidechain c.lusters (which would have higher rotational velocities). We used the united-atom option in the DREIDING force field so that the CH_3 units of the Alanine sidechains were treated as single particles which did not rotate independently. However, the tyrosine, phenylalanine, and methionine sidechains of Met-cnkc])llalin all contain individual clusters with low moments of inertia. As indicated below in the analysis of Met-clllic]dlalin dihedral angle fluctuations, the long, unbranched methionine sidechain is particularly flexible.

As the simulations are carried out for longer periods of time, the flue.tuaticnls $\mathcal{E}$ gradually increased. For instance, a 1 ps NEIMO simulation of Met-enkephalin using a 5 fs time step leads to a value of $\mathcal{E}$ lcss than 0.0001 kcal/mol. The same simulation run for 5 ps leads to $\mathcal{E}$ = 0.0042 kcal/mol, even though each 0.1 ps stretch of the simulation has $\mathcal{E} <$ 0.0004 kcal/mol, and the average fluctuation over the 50 0.1 ps stretches was only 0.0001 kcal/mol. Over 25])s, the simulation leads to an overall $\mathcal{E}$ of ().()36() kcal/mol, even though

the average 0.1 ps stretch had $\mathcal{E}$ = 0.0005 kcal/mol. This discrepancy is caused by very slow fluctuations in the total energy which cause $\langle E^2 \rangle$ to slowly diverge from $\langle E \rangle^2$. The cause of this long-term flue.tuaiioll is unknown. Possibly due to time asymmetry of the new integrator.

In order to compare NEIMO directly to the matrix method of Mazur, *et al*,[14] the quantity δ_E (defined below) was calculated from simulations of $(Ala)_9$ at time steps ranging from 1 fs to 20 fs (sew Figure 4), For each time step, the simulation was run for 4.0 ps during which the velocities were rescaled, when necessary, to equilibrate the system. At the end of the 4.0 ps run, 110 additional steps were run. The first ten of these were discarded, but the final 100 steps were used to determine δ_E, which is defined by

$$\delta_E = \frac{\sqrt{\langle \Delta E \rangle}}{\langle E \rangle}. \tag{16}$$

$\langle E \rangle$ is the average energy during the 100 steps, and $\sqrt{\langle \Delta E \rangle}$ is the root-mean-square clevi-ation in the energy. Mazur et al. reported simulations on $(Ala)_9$ using a variety of models including some containing explicit hydrogens. The DREIDING/NEIMO talc.ulaticnl corresponds to their third model: united atoms are used rather than explicit hydrogens, and all bond lengths and angles are fixed. Only dihedral degrees of freedom are allowed plus the six degrees of freedom of the base body, for a total of 32 degrees Of freedom. Mazur et al. obtained a value of δ_E = 0.8 X 1 0^{-6} using time steps of 0.5 fs. The magnitude of δ_E increased linearly with increasing time steps, but they were able to achieve their desire level of accuracy, $\delta_E \approx 10^{-2}$, using time steps as large as 13 fs. NEIMO simulations using a 0.5 fs time step had a larger value of δ_E = 4.0 × 10- ‘, but time steps as large as 15 fs gave $\delta_E \approx$ 10-”2, as can be seen in Figure 4. These results are very consistent with the results of Mazur, *et al.*, even though they used different force field (a combination of CHARMm[26] and ECEPP[27]) and a different integration scheme.

B.3 Avian Pancreatic Polypeptide

Although **time** steps of 15 fs and longer are clearly possible for NEIMO simulations Of small peptides such as Met,-cIIIi(:l)l]ali]l and $(Ala)_9$, such time steps are too large for large polypeptides and proteins, using the original program. Avian pancreatic polypeptide

(d'l'), a 36 resiit ue hormone peptide, is a very interesting case because it is one of the smallest known polypeptides to fold into a stable globular form. Figure 5 Snows the alpha carbon trace of aPP, which has two helices: an α helix and a collag;ml-like polyproline helix[17] Hydrophobic sidechains line tl ie cleft between the two helices, allowing for unusual stability in a peptide this size.

Figure 6 shows $\mathcal{E}$ and $\mathcal{E}^*$ using different time steps for 1 ps simulations of aPP. NEIMO simulations of aPP break down when time steps above 10 fs are used. Although time steps as large as 9 fs give values of $\mathcal{E}$ as good or better than the 1 fs Cartesian simulation, the scaled fluctuations, $\mathcal{E}^*$, are approximately equal for 6 fs NEIMO and 1 fs Cartesian cases. Several factors may muse folded polypeptides and proteins to have substantially larger fluctuations than small peptides at large time steps. C o mplex secondary structure elements such as helices, turns, and beta. sheets, are held together by hydrogen bonds, whit.]1 are short-range interactions. Large time steps may cause rapid destabilization of these hydrogen bond networks. In general, nonbonded fore.cs such as van der Waals, electrostatics, and hydrogen bonding arc Cartesian in nature and can flue.tuatc substantially with respect to dihedral angle rotations. This effect is particularly great in the densely-packed interior of globular proteins, where self-c.ollisicnls occur very quickly, Much larger time steps can be used in NEIMO simulations of large low-density polymer systems.[31]

The fastest dynamical modes in the NEIMO model arc those with the smallest spatial inertia. In protein systems, these are clusters with explicit hydrogens, where rotation of the hinge moves only the hydrogen atoms. For instance, the hydroxyl group of Tyrosine forms a two-atom cluster. Rotation of the hinge between the aromatic ring C_ζ and the hydroxyl O_η modifies only the hydroxyl hydrogen coordinates. These arc the fastest degrees of freedom in the system. With NEIMO wc can hold fixed these dihedrals by counting the outer OH cluster as part of the parent cluster and then treating the hydroxyl and aromatic ring of tyrosine as a single cluster. This "Rigid 11" model removes the fastest degrees of freedom of the system and enables even longer NEIMO time steps. This is seen clearly in Figure 7, where the 18 hydroxyl and amino groups of aPP have been incorporated with their parent clusters. Although the scaled fluctuations, $\mathcal{E}^*$, are very similar for small time

steps, the standard model blows up when time steps above 10 fs are used, while the "Rigid H" fluctuations increase only slowly above this point. Simulations using even longer time steps displayed the same gradual increase in fluctuations, without the sharp jump in $\mathcal{E}^*$ for time steps above 10 fs. The "Rigid H" model should be useful for studies focussed primarily on large-scale motions, where the hydrogen-bonding interactions of these sidechain groups are less important and the advantage of longer time steps is pre-eminent.

Detailed studies of protein systems require the inclusion of solvent, which plays an important role in stabilizing the native conformation of most proteins. Solvent includes both water (and/or lipids in the case of membrane-bound proteins) and ionic charges, which may be present to stabilize charged groups on the protein. In order to test the ability of NEIMO simulations to include such factors, we ran calculations where NEIMO dynamics were used to solve the equations of motion for the protein, while standard Cartesian dynamics equations were solved simultaneously for counterions. Avian pancreatic polypeptide (aPP) was used as a test system. Oppositely-charged groups within 10 Å of each other were considered paired and were not given counterions. This left eight unpaired charges, which were then neutralized by adding counterions (five Na^+ and three Cl^-). The counterion locations were first optimized by minimizing their energies, then simulations were run for 2 ps using various time steps. The first picosecond was used for equilibrating the counterion motions and the next picosecond was used to determine $\mathcal{E}$. The results are shown in Figure 7, along with the results from standard and "Rigid H" simulations of the protein alone. The addition of counterions increases the energy fluctuation substantially, but time steps as large as 8-10 fs are still practical. This is a great improvement over simulations where all atoms are treated with Cartesian-space molecular dynamics.

Figure 8 shows the variation in $\langle\mathcal{E}\rangle$ during 5 ps simulations of aPP. In these calculations, the CMM was used for the nonbond calculations. The $\mathcal{E}$ was calculated during 0.1 ps intervals, during which the average kinetic energy was calculated and the farfield contribution to the CMM energy was held constant.[4] At the end of each 0.1 ps interval, the velocities were rescaled if necessary, the CMM farfield was recalculated, and the $\mathcal{E}$ was recorded. At the end of the 5 ps simulations, the $\mathcal{E}$ values were averaged to give $\langle\mathcal{E}\rangle$. These

values are plotted in Figure 8. For very short time steps (1 and 2 fs), the $\langle \mathcal{E} \rangle$ values are much larger than the $\mathcal{E}$ values from the 1 ps simulations in Figure 6. At large time steps, however, the results are very consistent with the shorter simulations.

B.4 Large Proteins

Figure 9 snows the average value of $\mathcal{E}^*$ during 5 ps simulations of several of the proteins in 'J'able 2. Clearly the energy fluct uations in NEIMO dynamics simulations, even when scaled by the number of degrees of freedom, increase with protein size. This is in contrast to the fluctuations during Cartesian dynamics simulations, which are roughly constant when divided by the number of degrees of freedom. Soi ne of the inherent difficulties of doing internal-coordinate dynami cs for dense protein systems are discussed above. Figure 9 allows that t) ıese problems increase with protein size. The results here indicate that NEIMO dynamics simulations of typical protein systems (1 000-10,000 atoms) should be used with 1 or 2 fs time steps if a high degree of accuracy is required. However, in some cases it may be valuable to increase the time step despite the loss in accuracy, in order to increase the time span of the simulation. Such simulations would include studies of large-scale protein motions such as hinge bending or local folding and unfolding. The accuracy of the NEIMO simulations for large time steps should improve as our implementation evolves. This has already been seen in recent work.[31]

C. Dihedral Distributions

Analysis of energy fluctuations indicates that the NEIMO method accurately solves it equations of motion for molecular systems. However internal coordinate dynamics produces a different sequence of molecular motions than Cartesian dynami cs (which includes the additional bond and angle degrees of freedom). in order to determine the relationship between NEIMO and Cartesian simulations for the dynamics of the Met-enkephalin peptides, we computed the distribution of dihedral angles during these simulations. Cartesian and NEIMO dynamics simulations were run at a temperature of 300 K for 5.() ps, during which the dihedral angles were output every O.]])S. The Cartesian dynamics calculations had a 1 fs time step while the NEIMO calculations were run at a variety of time steps. Figure 10 shows the resulting distributions from simulations of Met-enkephalin. The num-

b ering of the dihedral angles is shown in Figu e 1 and further identified in Table I. The top graph in Figure 1() shows the distribution from Cartesian dynamics and the bottom sl ows the distribution from a NEIMO clyjlall'lies simulation; both simulations used a 1fs time step. The distributions from the two simulations are very similar, with the backbone ω dihedrals (6, 9, 12, and 17) showing the least flexibility, as would be expected, and the methionine sidechain dihedrals showing the great.cst variation during the simulation. The aver age values for each dihedral, θ, can be calculated from such distributions. Because dihedral angles have a periodicity of 2π (3600), the average cannot be calculated directly, but is derived from the average cosine and sine[28]:

$$\langle\theta\rangle = \arctan\left(\langle\sin\theta\rangle / \langle\cos\theta\rangle\right) . \tag{17}$$

Once $\langle\theta\rangle$ is known, the standard deviations can be calculated easily for AT time steps:

$$\sigma = \left[\sum_{i=1}^{N} \Delta(\delta\theta)^2 \right]^{1/2}, \tag{18}$$

where

$$\delta\theta_i = (\theta_i - \langle\theta\rangle)$$
$$\pi < \delta\theta_i < \pi . \tag{19}$$

Because of the periodicity of dihedral angles, equation (8) can always be enforced by appropriate additions or subtractions of 2π

The average values, (0), and standard deviations, σ, for the distributions in Figure 10 are shown in Figure 11. The average values are also shown in ~'able 4, and are compared to the initial conformation. The NEIMO results are very similar to the results from the! Cartesian simulations, indicating that the reduction in the number of degrees of freedom does not, in general, affect the torsional flexibility of the molecules. There arc! two exceptions to this here: χ^1 of Met 5 undergoes a transition from roughly 30° to - 60° (300°) in the Cartesian simulation, but remains near 45° in the NEIMO simulation. Secondly, the ψ angle of Gly 2 is rotated from - 60° to 60° in the Cartesian simulation, but remains near - 60° duri ng the N EIMO calculation. Apparently fixing the angle terms increases the barriers to rotation sufficiently to prevent these transitions during 5 ps NEI MO simulation

at 300 K. The rotational transition of Met 5 χ^1 did occur after approximately 40 ps of a 50 ps NEIMO simulation using 5 *fs* time steps. A 600 K NEIMO simulation using 5 fs time steps saw both transitions occur by 20 ps, but the temperature was high enough that fill'tller transitions continued in both directions over these harriers. It is important to note that the NEIMO formalism explicitly includes the capacity for bond stretches and angle bends between clusters, but the current implementation uses only the dihedral degrees of freedom. Use of canonical dynamics also increases the rates for such torsional transitions. The torsional barriers are expected to increase when bonds and angles are fixed. This effect can quite simply be built into the force field by calculating the barriers for fixed and flexible bonds and angles and then adjusting the harriers for the torsional-only calculations to include this.

A slightly different view of the average dihedrals from 10 different NEIMO simulations is given in Figure 12. The simulations were identical except for the time step, which ranged from 1 fs to 10 fs (chosen to given an integer number of dynamics steps per 0.100 ps). It is clear that the results are quite consistent for time steps up to 10 fs. Only the two outer sidechain dihedrals χ^1 and χ^2 of Met 5 have significantly different distributions for different time steps. χ^1 has $\langle\theta\rangle \approx 145°$ for 8, 9, and 10 fs time step simulations, but (t?) $\approx 90°$ for the smaller time steps. It is possible that the larger time steps occasionally enable the molecule to jump over rotational energy barriers which cannot be cleared by simulations using smaller time steps which, in effect, calculate energies and forces at more points along the trajectory.

In order to quantify the dihedral distributions, wc represented cinch distribution]] by a gaussian, using the average, (0), and standard deviation, σ, from the 50 datapoints:

$$\Psi(\theta, \langle\theta\rangle, \sigma) = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^{1/2} e^{-\delta\theta^2/4\sigma^2}. \tag{20}$$

These gaussians were normalized as

$$\int_{-\pi}^{\pi} P(\theta)\, d\theta = \int_{\pi}^{\pi} [\Psi]^2\, d\theta = 1, \tag{21}$$

[The constant in equation (20) is appropriate for nonperiodic variables and would lead to a total probability in equation (21) that is not normalized if σ were so large that the

probability is non-zero for every value of θ; this did not occur for any of the distributions we have analyzed.]

Distributions from two different simulations can be compared by calculating the overlap, S_{12}, of the functions Ψ_1 and Ψ_2:

$$S_{12} = \int_{-\pi}^{\pi} \Psi_1 \Psi_2 d\theta. \tag{22}$$

If Ψ_1 and Ψ_2 are defined as

$$\Psi_1 = \left[\frac{2\alpha}{\pi}\right]^{1/4} e^{-\alpha(\delta\theta_1)^2}, \quad \Psi_2 = \left[\frac{2\beta}{\pi}\right]^{1/4} e^{-\beta(\delta\theta_2)^2}, \tag{23}$$

where $\alpha = 1/4\sigma_1^2$ and $\beta = 1/4\sigma_2^2$, and $\delta\theta_1$ and $\delta\theta_2$ are defined as in equation (19) for $\langle\theta\rangle_1$ and $\langle\theta\rangle_2$, then the product of these functions is also a gaussian:

$$\Psi_1 \Psi_2 = \left[\frac{4\alpha\beta}{\pi^2}\right]^{1/4} K_{12} e^{-(\alpha+\beta)\delta\theta_{12}}. \tag{24}$$

Here, $\delta\theta_{12}$ is defined as usual from (O) I 2, where

$$\langle\theta\rangle_{12} = \frac{\alpha\langle\theta\rangle_1 + \beta\langle\theta\rangle_2}{\alpha+\beta}. \tag{25}$$

The constant in equations (24) is

$$K_{12} = \exp\left[\frac{(\alpha\langle\theta\rangle_1 + \beta\langle\theta\rangle_2)^2}{\alpha+\beta} - \left(\alpha\langle\theta\rangle_1^2 + \beta\langle\theta\rangle_2^2\right)\right]. \tag{26}$$

Inserting equation **(24)** into equation (22) gives a formula for the overlap:

$$S_{12} = \left[\frac{4\alpha\beta}{(\alpha+\beta)^2}\right]^{1/4} K_{12}. \tag{27}$$

S_{12} equals 1 if the two distribution functions are identical and equals O if there is no overlap.

The overlaps from the 10 NEIM () simulations plotted in Figure 12 are shown in Figure] 3. Each line represents the overlap between the 1 fs time step simulations and one of the simulations with a larger time step. A second figure, Figure 14, specifically snows the overlaps between the 1 fs simulation and the 2, 5, and 10 fs simulations at a higher

resolution. As expected, there is almost 100% overlap among the NEIMO simulations, which indicates clearly that the molecular dynamics are very consistent across a range of time steps of 1 fs to 10 fs. The one exceptions to these are χ dihedrals of Met 5 and the ω of Gly 2. The relatively small overlap of the latter is due to the very small value of σ (0.3°) for the 1 fs NEIMO simulation. The discrepancy in the methionine sidechains is also due primarily to differences in σ rather than $\langle\theta\rangle$ for the smaller time step simulations. At larger time steps, however, both σ and $\langle\theta\rangle$ differ.

Overlaps between the dihedral distribution from the Cartesian simulation, and those from the NEIMO simulations, are shown in Figure 15. Here, the overlap is quite small for χ^2 of Met 5 and the ψ backbone dihedral of Gly 2, as indicated by the large differences in $\langle\theta\rangle$ note above. A third very-low overlap is seen for the ω of Gly 2. This difference is completely hidden in Figure 11 since it is due entirely to the extremely low value of σ in the NEIMO simulations. The value is so low, in fact, that it dots not appear in Figure 11 for the 1 fs NEIMO simulation. The overlaps are greater than 65% for 19 of the 22 dihedrals for every NEIMO time step. Excluding the methionine residue, overlaps are greater than 90% for 13 of the 16 dihedrals.

1).1 Tomato Bushy Stunt Virus

The linear-in-n scaling of CMM makes molecular dynamics calculations possible for million-atom systems.[4] In addition, the linear-in-$\mathcal{N}$ cost of NEIMO (see Table 3) means that there is no longer a restriction on the size of the system for which internal coordinate dynamics are practical (the computational time is dominated by calculation of the nonbonded interactions). Thus it is now computationally possible to perform molecular dynamics calculations on systems as large as icosahedral viruses, such as rhinoviruses or the tomato bushy stunt virus (TBSV). A typical virus of this type, having a protein coat of roughly 7×10^6 D surrounding an RNA strand of 1.5×10^6 D[24] contains on the order of a million atoms. of course, practical calculations for long term dynamics still requires supercomputers. However we have used the implementatons of NEIMO and CMM on a Silicon Graphics workstation (4D/380) to examine short dynamics studies (up to 10 ps) for several systems.

TBSV is an RNA virus composed of 180 identical coat proteins arranged in $T = 3$ icosahedral symmetry. The virus has been crystallized and the structure of the asymmetric unit, containing three copies of the coat protein, has been determined from 2.9Å X-ray data by symmetry averaging.[24] The three copies of the coat protein in the asymmetric unit have slightly different conformations which are designated A, B, and C. While all three conformations contain RNA-binding (R), surface (S), and projecting (P) domains, the R domain (residues 1-101) is completely unresolved in the A and B conformations while in the C conformation, residues 67-101 have an ordered structure and are resolved. The viral RNA (molecular weight 1.5×10^6) lacks icosahedral symmetry and hence structural data on the core region is not available experimentally.

The asymmetric unit of TBSV contains six calcium ions, Ca^{2+}, arranged as three pairs located at the A-B, B-C, and A-C interfaces. Each pair of calcium cations binds in a negatively charged pocket at the interface between adjacent S domains; the pocket is formed by five aspartic acid sidechains contributed by the two proteins. It is postulated that the interaction between these Asp residues and the Ca^{2+} ions plays a major role in stabilizing the viral coat.[32] If the Ca^{2+} ions are removed, the virus expands as the pH is raised above 7.[32] The hydrodynamic radius of the virus can expand by as much as 10%, but there is no loss of mass and the process is reversible. A low- resolution (8 Å) crystal structure was determined for the expanded conformation of the virus[32] and indicated that expansion occurred by relative motions perpendicular to the interfaces where Ca^{2+} ions bind in the unexpanded conformation. However, no atomic details were available from this low resolution data.

In order to investigate the expansion phenomenon, we carried out molecular dynamics calculations on the viral coat proteins with the calcium cations (the "pH7" model) and without the calcium cations ("NoCa").[34] The model systems included all resolved residues from the asymmetric unit plus counterions, Na^+ and Cl^- (and Ca^{2+} for pH7), for a total of 8138 atoms. We did not attempt to simulate either the RNA or the unresolved RNA-binding regions of the coat proteins. Through the use of the transformation matrices in the crystal structure (Brookhaven Protein Database structure 2TBV), the coordinates

were generated for the entire viral coat containing 180 proteins and 488,280 atoms. On a one-processor SGI 4D/380 workstation, the current implementation (NICIMO and CMM in conjunction with Biograf) requires 42 sec per time step (assuming icosahedral symmetry). Thus on an SGI workstation it would require about 42 min per time step to do the full capsid without symmetry. In order to carry out simulations for Several picoseconds, we restricted the dynamics by requiring icosahedral symmetry. That is, only the atoms of the three proteins of the asymmetric unit were considered as independent. However, it was practical to include all nonbonded interactions using CMM (Coulomb and van der Waals) between the asymmetric unit and all other 177 coat proteins (with no cutoffs). The coordinates of the entire viral coat were updated (all 488,280 atomic positions) after each dynamic step so that we considered the dynamics of the full protein capsid. Molecular dynamics calculations have been reported cm another similar size virus (with about 8000 atoms in the asymmetric unit), by Çağm, Holder, and Pettitt on HRV-14[33] also using icosahedral constraints (they used rotationally symmetric boundary conditions).

D.2 Methodology

In an attempt to simulate the expansion effect observed for high pH, we developed two different models of the TBSV. The first contains the protein atoms and calcium ions as they appear in the protein database coordinate file (2TBV)[24], with hydrogen atoms added to nitrogen, oxygen and sulfur atoms to allow for hydrogen bonding. In addition, Na^+ and Cl^- ions were added to balance the charges of unpaired acidic and basic residues, respectively. This structure is termed the "pH7" model. The second representation is the "NoCa" model, in which the six Ca^{2+} ions were removed and the free aspartic acid residues were allowed to form salt bridges with basic residues, or were given Na^+ counterions. In this model, the 15 Asp residues are no longer held together by interactions with Ca^{2+}, but are free to move independently. This is believed to be the major factor in the expansion of the virus particle.[32] Inclusion of explicit waters in a calculation of this type improves the accuracy but also greatly increases the computational cost. We partially corrected for the exclusion of water by modifying the system electrostatics in two ways: we used a distance-dependent dielectric constant ($\epsilon = \epsilon_0 r_{ij}$) and placed Na^+ and Cl^- counterions

near charged amino acids which had no oppositely-charged amino acids or Ca^{2+} ions within 10Å. The pH7 model required 25 Cl^- and 24 Na^+ in addition to its 6 Ca^{2+}. The NoCa model required 22 Cl^- and 33 Na^+. Both models were charge neutral and had the same number of atoms (8138). The S and P domains of the TBSV asymmetric unit were resolved independently in the X-ray studies, leading to mismatches in the hinge region (residues ~273-275). The crystal structure (2TBV) lists alternate S and 1' coordinates for the residues in the overlap region. For our calculations, we averaged the coordinates of the two alternates and re-optimized the structure by energy minimization.

In order to accurately model the capsid environment, the nonbonded forces acting on the asymmetric unit included all nonbond interactions with all other 177 proteins. This was made possible by the use of CMM[4], an extremely fast and accurate method for calculating nonbonds in large systems. CMM divides the simulation space into a hierarchy of cubic cells, the smallest of which contains, ideally, 4 or 5 atoms and the largest of which contains the entire system. For the asymmetric unit alone, four levels of cells were needed. There were 4096 (8^4) cells at level 4, measuring 6.397 Å on a side. Since this unit is rather flat, 81.4% of these cells are empty, leaving 762 populated cells with an average of 10.7 atoms per cell. When the asymmetric unit was expanded into the full 180 protein capsid, CMM used six levels for the 488,820 atom system. At level 6, there were 262,144 cells, each 5.340 Å on a side. 87.5% of these are empty, leaving 8.0 atoms per populated cell. Since the dimensions and the average population for the full system are better than those in simulations of the isolated asymmetric unit, it was faster to calculate the nonbonds for the entire 180 protein coat than for the asymmetric unit alone (see below)!

D.3 Results

Timing results for the CMM molecular dynamics calculations are shown in Figure 17, in terms of CPU seconds on one processor of an SGI 4D/380 workstation. The total charge, dipole, and quadrupole of each cell, collectively termed the "farfield," varies slowly with time so that it need not be recalculated every time step. We considered two cases, labeled "Update1" and "Update50," the latter referring to calculations in which the farfield was updated only every 50 steps. Also shown is the difference between calculations using only

the nonbonds of the three- protein asymmetric unit, labeled "NB3," and those including interactions with the entire 180-l)rotcill capsid, labeled "NB1 SO." The Update50 calculations are actually faster for NB180 than NB3 (37.2 versus 41.0 s), because the calculation is dominated by the nearfield interactions. The average cell in the NB180 case has 5.4 atoms versus **6.4** for the NB3 **case** which means that fewer pairwise interactions need to be calculated. The farfield interaction takes longer to calculate in the NB180 case, but the effect is not significant because of the hierarchical CMM approach. Only when the farfield itself must be updated every step, i.e., the Update1 calculations, dots the size of the system make a significant difference. In such calculations, including the entire capsid increases the time of the nonbond calculations from 49.8 S to 92.6 s.

Figure 17 also shows the total time required for NEIMO calculations, including the time required to calculate accelerations and velocities, and to update the system coordinates. NEIMO calculations for the three protein chains in the asymmetric unit ($\mathcal{N} = 4335$) require 5.3 s per dynamics step. Figure 18 shows the average scaled energy fluctuation, $\langle \mathcal{E} \rangle^*$, versus time step for 1.0 ps simulations of the PH7 model of TBSV. $\mathcal{E}$ and $\langle \mathcal{E} \rangle$ are defined in Equations 14 and 15, respectively. The scaled fluctuation, $\langle \mathcal{E} \rangle^*$, is $\mathcal{E}$ divided by the number of degrees of freedom ($\mathcal{N}$ in NEIMO and $3n$ -- 6 in Cartesian dynamics). The fluctuations in Figure 18 are larger than those in Figure 9 because these talc.ulatiolls included Ca^{2+}, Na^+, and Cl^- ions which had to be sim ulated using Cartesian dynamics. The NEIMO implementation allows for NEIMO dynamics of large systems and Cartesian dynamics of individual particles to be handled simultaneously. Figure 18 indicates that the energy fluctuations did not vary much between different nonbond methods. Fluctuations using the entire capsid, with the farfield updated every step, i. e., NB180/Update1, were not measured but are unlikely to provide! a substantial improvement. 111 every case, there was typically a 4- to 5-fold increase in the fluctuation for each 1 fs increase in the time step.

Cartesian dynamics with 1 fs time steps generally give a value of 0.0001" kcal/mol for $\langle \mathcal{E} \rangle^*$. This value was matched by NEIMO simulations with a 1 fs time step, but exceeded in simulations using larger time steps. However, as discussed above, the computational

speedup obtained from using large time steps may be worth the loss in accuracy. We used a time step of 2 fs for our NEIMO simulations of TBSV. This allowed us a nearly twofold speedup over Cartesian dynamics while maintaining reasonably low energy flue.t,uat,iol]s. As is clear from Figure 17, the NEIMO computational time is small compared to the nonbonded calculations, so the speedup vs. Cartesian is essentially linear with increasing time step.

The two model systems,])117 and NoCa, were initially optimized using Cartesian space conjugate gradients minimization. The CMM method was used at the NB180/Update50 level. The radius of gyration of the viral capsid system was calculated every 50 steps, when the farfield was updated. The radius of gyration is defined as:

$$R_{gyr} = \left[\frac{\sum_i^n m_i (x_i - x_{cm})^2 + (y_i - y_{cm})^2 + (z_i - z_{cm})^2}{\sum_i^n m_i} \right]^{1/2} \qquad (30)$$

where the coordinates and mass of each particle i are (x_i, y_i, z_i) and m_i, respectively, and the coordinates of the center of mass are (x_{cm}, y_{cm}, z_{cm}). Both structures contracted very little (about 0.08%) during the minimization. At the end of the minimization (converged at 0.01 kcal/mol$Å^2$), the pH 7 structure was almost 700 kcal/mol lower in energy than NoCa (-5801.6 kcal/mol vs. - 5135.6), despite containing identical numbers of atoms. This energy difference is almost entirely due to the electrostatic energy term and indicates the large stabilizing energy of the Ca^{2+} ions.

Molecular dynamics calculations were carried out on the different minimized structures, again using NB180/Update50 for the nonbonds. The farfield was updated every 0.1 ps. Two different Cartesian dynamics methods were used:

(i) microcanonical dynamics (NVE) with temperature scaling and

(ii) Nosh canonical dynamics (NVT).[36]

In addition, for NVE we carried out NEIMO dynamics with a time step of 2 fs. Figure 19 shows the radius of gyration of pH7 during the first 2.0 ps of dynamics. Both Cartesian simulations show an initial expansion phase followed by a longer contraction. The NEIMO simulation shows no expansion phase, but its contraction phase closely resembles that of the Cartesian NVE simulation, with roughly the same slope, contracting until approximately

1.8])s, when it levels out. The canonical dynamics simulation, in contrast, shows 110 similar leveling out through the first 2.0])S,

Longer simulations were run using Cartesian NVT dynamics and NEIMO dynamics on both the])117 and NoCa models. The NEIMO dynamics simulations were twice as fast, since 2 fs time steps were used. Figure 20 shows the radius of gyration of the pH7 and NoCa models during the first 4 .(I ps of NEIMO and Cartesian NVT simulations. In the NEIMO simulations, the NoCa model undergoes a rapid expansion during the first 2.0 ps, then an even sharper contraction. The pH7 model has no such expansion phase but does have a gradually increasing contraction rate. Both of these simulations show far more variation in the radius of gyration than the corresponding Cartesian simulations. However, in both simulations, the NoCa model initially has a larger radius of gyration than for pH7, but eventually becomes smaller. Although the curves of the radii cross after about 3.9 ps in the Cartesian simulation, the energy curves do not cross, as shown in Figure 21. The energy plot indicates that the NoCa model is less stable, as it undergoes larger energy fluctuations after 3.0 ps. These fluctuations continue until the end of a 5.0 ps simulation (data not shown). The pH7 model is relatively stable. For the NEIMO simulations, the contraction rate is comparatively exaggerated, but the energies do not show such large fluctuations. The NoCa model has a potential energy around -3850 kcal/mol while the potential energy for pH7 remains near -4550 kcal/mol. Note that these energies are substantially lower than in the Cartesian simulation because the numerous bond and angle degrees of freedom remain at their minimum potential energy values. Therefore, the approximately 700 kcal/mol differential between pH 7 and NoCa is relatively constant, even though the! radii change significantly.

The current simulations do not reproduce the 1 0% expansion expected for the NoCa model on the basis of the experimental data.[32] However, the NoCa model is substantially higher in energy (700 kcal/mol), indicating that it might be driven to expand in more extensive calculations. The NEIMO simulations show substantial contraction of both the pH7 and NoCa models. This is likely due to ignoring the RNA in the interior of the virus.

As these new methods (NEIMO and CMM) are optimized for parallel supercomputers,

we expect to carry out calculations on the full virus, including RNA. This could be most valuable since experimental techniques provide little structural data about the RNA region.

IV. conclusions

The NEIMO and CMM methods have now been successfully applied to polypeptide and protein systems. NEIMO is extremely fast compared to other internal-coordinate dynamics methods. NEIMO and CMM scales linearly with the number of degrees of freedom making them practical for super large systems. For increasingly large systems, the NEIMO computational requirements grow more slowly than those of energy calculations. Molecular dynamics including only torsional degrees of freedom allow much larger time steps than simulations including all possible degrees of freedom. NEIMO calculations of peptides indicate that time steps as large as 20 femtoseconds can be used for these small Systems. Time steps of this size are not yet possible for large polypeptides and proteins, as judged by the criterion of total energy fluctuations. However, time steps of 5 fs and longer can be used for large systems without danger of energy divergence. Such calculations should be useful for conformational analyses of extremely large systems such as viruses. As these methods are refined, it is likely that the energy conservation of larger systems will be improved.

The dynamics of polypeptides are accurately modeled by NEIMO. Analyses of dihedral angle fluctuations show that NEIMO dynamics simulations produce conformational fluctuations very similar to those arising from Cartesian dynamics simulations. The few exceptions to this in simulations of Met-enkephalin appear to be cases where rotational barrier is sufficiently higher for fixed bonds and angles that they are traversed in the Cartesian dynamics simulation but not for NEIMO simulations at the same temperature. Such problems can be eliminated by using torsional barrier based on the adiabatic energy curves. Thus we believe that CMM and NEIMO can now be used for simulations on very large molecules, such as viruses.

Acknowledgments

We wish to thank Dr. Guillermo Rodriguez of JPL and Dr. N. Vaidehi of Caltech for important contributions to the development of the current NEIMO implementation and to

Dr. Naoki Karasawa for the implementation of CMM with BioGraf. AMM acknowledges a National Research Service Award/NIH Predoctoral Traineeship in Biotechnology. We thank Mr. K. 'J'. Lim for constructing Figure 16.

We thank DOE-AICD for funding this research. The facilities of the MSC are also supported by grants from NSF (CHE 91-100289), NSF-ACR, DOE-CHBIOL, Allied-Signal Corp., Asahi Chemical, Asahi Glass, BP America, Chevron, BF Goodrich, Vestar, Xerox and Beckman Institute.

This work has been partially performed at the Jet Propulsion laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

APPENDIX

For Cartesian dynamics we generally use the "leapfrog" formulation of the Verlet algorithm. In this approach, the coordinates z(") at time step $f = nh$, are used to calculate the forces at this time step, $F^{(n)}$ which are related to the accelerations by

$$\ddot{x}^{(n)} = -\frac{1}{m} F^{(n)}.$$

These are combined with the velocities at step $n - \frac{1}{2}$, $v^{(n-\frac{1}{2})}$, to calculate the velocity at $n + \frac{1}{2}$,

$$v^{(n+\frac{1}{2})} = v^{(n-\frac{1}{2})} + \frac{h}{m} F^{(n)} \tag{A.1}$$

which is in terms used to calculate the new coordinates

$$x^{(n+1)} = x^{(n)} + hv^{(n+\frac{1}{2})} \tag{A.2}$$

This is initiated with

$$v^{(\frac{1}{2})} = x^{(0)} + v^{(0)} + \frac{h}{2m} F^{(0)} \tag{A.3}$$

For NEIMO dynamics this algorithm is more complicated because the acceleration $\ddot{\theta}^{(n)}$ depends explicitly 011 both the velocities $\dot{\theta}^{(n)}$ and coordinates $\theta^{(n)}$ at time step 71. For this paper, we used the following iterative procedure. We estimate the velocities $\dot{\theta}^n$ from the previously determined velocities:

$$\dot{\theta}^n = 1.5\dot{\theta}^{n-\frac{1}{2}} - 0.5\dot{\theta}^{n-\frac{3}{2}}. \tag{A.4}$$

This allows us to calculate the NEIMO accelerations $\ddot{\theta}^n$ by solving the spatial operator (S0) equations,[11,12]

$$\ddot{\theta}^n = SO\left(\mathbf{F}^n, \mathbf{T}^n, \theta^n, \dot{\theta}^n\right) \tag{A.5}$$

using the coordinates θ^n, the velocities $\dot{\theta}^n$, the torques $\mathbf{T}^n$, and nonbond forces $\mathbf{F}^n$. The accelerations are used to update the velocities as in A.]

$$\dot{\theta}^{n+\frac{1}{2}} = \dot{\theta}^{n-\frac{1}{2}} + h\ddot{\theta}^n. \tag{A.6}$$

Because $\dot{\theta}^n$ is estimated in equation (A.4), the $\ddot{\theta}^n$ from A.5 is inaccurate and errors could build up as the simulation progresses. In order to eliminate such errors, $\dot{\theta}^n$ is re-estimated from the $\dot{\theta}^{n+\frac{1}{2}}$ calculated in A.6 and the known $\dot{\theta}^{n-\frac{1}{2}}$:

$$\dot{\theta}^n = 0.5\dot{\theta}^{n+\frac{1}{2}} + 0.5\dot{\theta}^{n-\frac{1}{2}}. \qquad (A.7)$$

This process (A.5), (A.6), (A.7) is repeated until $\dot{\theta}^n$ converges, producing an accurate value for $\ddot{\theta}^n$. We find that sufficient convergence (based on maximal improvement to energy conservation) is generally reached after a single iteration, so that the effect on overall computational costs is minimal.

The converged values of $\ddot{\theta}^n$ from equation (A.5) give the converged values for $\dot{\theta}^{n+\frac{1}{2}}$ from A.6 which are then used to update the coordinates:

$$\theta^{n+1} = \theta^n + h\dot{\theta}^{n+\frac{1}{2}}. \qquad (A.5)$$

The dynamics step is completed using the new internal coordinates θ^{n+1} to update the Cartesian coordinates χ^{n+1} which are used for the Cartesian forces, $F^{(n+1)}$.

References

1. Brünger, A.T. Simulated annealing in crystallography. Annu. Rev. Phys. Chem. 42:197-223, 1991,
2. Beveridge, D. L. and Di Capua, F. M. Free-energy via molecular simulation: applications to chemical and biomolecular systems. Annu. Rev. Biophys. Biophys. Chem. 18:434-492, 1989.
3. Wilson, C.L., Young, W.S., Tobias, D.J. Molecular simulations on supercomputers. Int. J. Supercomp. Appl. 5:98-112, 1991.
4. Ding, H. Q., Karasawa, N., and Goddard, W. A. III. Atomic level simulations on a million particles - the Cell Multipole Method for coulomb and London nonbond interactions. J. Chem. Phys. 97:4309-4315, 1992.
5. Ding, H. Q., Karasawa, N., and Goddard, W.A. III. The reduced cell multipole method for coulomb interactions in periodic-systems with million-atom unit cells. Chem. Phys. Lett. 196:6-10, 1992.
6. Tuckerman, M.E. and Berne, B.J. Molecular dynamics in systems with multiple time scales: Systems with stiff and soft degrees of freedom and with short and long range forces. J. Chem. Phys. 95:8362-8364, 1991.
7. Ryckaert, J.-P., Ciccotti, G., Berendsen, H.J.C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comp. Phys. 23:327-341, 1977.
8. Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. Phys. Rev. 136:A405-411,1964. Verlet, L. Computer "Experiments" on Classical Fluids. 1. Thermodynamical Properties of Lennard-Jones Molecules. Phys. Rev. 159:98-103, 1967.
9. Hammonds, K.D. and Ryckaert, J.-P. On the convergence of the SHAKE algorithm. Comp. Phys. Comm. 62:336-351, 1991.
10. McCammon, J.A. and Harvey, S.C. "Dynamics of proteins and nucleic acids." Cambridge, U. K.: Cambridge University Press, 1987.
11. Rodriguez, G., Jain, A., Kreutz-Delgado, K. Spatial operator algebra for multibody system dynamics. J. Astronaut. Sci. 40:27-50, 1992.

12. Jain, A., Vaidehi, N., Rodriguez, G. A fast recursive algorithm for molecular dynamics simulation. J. Comp. Phys. 106:2 58-268, 1993.

13. Goldstein, H. "Classical Mechanics," 2nd. ed. Reading, Mass: Addison-Wesley, 1980.

14. Mazur A.I<., Dorofeev V.E., Abagyan R.A. Derivation and testing of explicit equations of motion for polymers described by internal coordinates. J. Comp. Phys. 92:261-272, 1991.

15. Gear, G.W. "Numerical Initial Value Problems in Ordinary Differential Equations." Englewood Cliffs, N. J.: Prentice-Hall, 1971.

16. BIOGRAF^{TM}/POLYGRAF^{TM}. 1s from Molecular Simulations, Inc., Burlington, Mass.

17. Blundell, 'T.]., Pitts, J.E., Tickle, I. J., Wood, S.P., Wu, C.-W. X-ray analysis (1.4 Å resolution) of avian pancreatic polypeptide: small globular protein hormone. Proc. Natl. Acad. Sci., USA 78:4175-4179, 1981.

18. Hendrickson, W.A. and Teeter, M.M. Structure of the hydrophobic protein crambin determined directly from the anomolous scattering of sulphur. Nature 290:107-113, 1981.

19 Collyer, C. A., Guss, J.M., Sugimura, Y., Yoshizaki, F., Freeman, H.C. Crystal structure of plastocyanin from a green alga, *enteromorpha prolifera*. J. Mol. Biol. 211:617-632, 1990.

20. Herzberg, O. and James, M. N. G. Refined crystal structure of troponin C from turkey skeletal muscle at 2.0 Å resolution. J. Mol. Biol. 203:761-779, 1988.

21. Fujinaga, M., Delbaere, L.T.J., Brayer, G.])., James, M. N. G. Refined structure of α-lytic protease at 1.7 Å resolution. Analysis of hydrogen bonding and solvent structure. J. Mol. Biol. 184:479-502, 1985.

22. Eriksson, E.A., Jones, 'J'. A., Liljas, A. Refined structure of human carbonic anhydrase II at '2.0 Å resolution. Proteins 4:174-282, 1988.

23. Rees, 1"). C., Lewis, hf., Lipscomb, W.N. Refined crystal structure of carboxypeptidase A at 1.5 Å resolution. J. Mol. Biol. 168:367-387, 19S3.

24. Harrison, S.C., Olson, A. J., Schutt, C.E., Winkler, F.K., Bricogne, G. Tomato bushy

stunt virus at 2.9 Å resolution. Nature 276:36 8-373,1978.

25. Mayo,S.L.,Olafson, 11.1)., and Goddard,W.A.III,DREIDING:a generic fore'cficld for molecular simulations. J. Phys. Chem. 94:8897-8909, 1990.
26. 1]1'001{s. 11.1{., Bruccoleri,R.E., Olafson,B.D, States, D. J.,Swaminathan,XDs., Karplus, M. CH AR Mm a program for macromolecular energy, minimization, and dynami cs calculations. J. Comp. Chem. 4:187 -217, 1983.
27. Sippl,M.M.,Némethy, G., Scheraga, 11 .A. Intermolecular potentials from crystal data. 6. Determination of empirical potentials for O-H= O= C hydrogen bonds from packing configurations. J. Phys. Chem. 88:6231-6233, 1984.
28. Lambert, M.H. and Scheraga, H. A. Pattern recognition in the prediction of protein structure. 111. An importance- sampling minimization procedure. J. Comp. Chem. 10:817-831, 1989.
29. Van Gunsteren, W.F. and Karplus, M. Effect of constraints on the dynamics of macromolecules. Macromolecules 15:1528-1544, 1982.
30. Lim, K.T., and Goddard, W.A. 111, to be published.
31. Vaidehi, N., Mathiowetz, A. M., Jain, A., and Goddar d, W. A. III, to be published.
32. Robinson, 1.1<, and Harrison, S.C. Structure of the expanded state of the tomato bushy stunt virus.
33. Çağın 'I'., Holder, M., and Pettitt, B.M. A Method for Modeling Icosahedral Virions: Rotational Symmetry Boundary Conditions. J. Comp. Chem. 12:627-634, 1991.
34. A . Mathiowetz, PhD Thesis, Chemistry, October 1992, "Dynamics and Stochastic Protein Simulations: From Peptides to Viruses."
35. Nosè 1984
36. Çağın 'J'., Karasawa, N., Dasgupta, S., and Goddard, W. A. III. Thermodynamic and Elastic Properties of Polyethylene at Elevated Temperatures, Mat. Res. Sot. Symp. Proc. 278:61, 1992. Also Cagin, 'J'., Goddard> W. A. 111, and Ary, M. I,. Thermodynamic and Elastic of Polyethylene at Elevated Temperatures, Mat. lies. Sot.. Symp. Proc. 278:62, 1992.

Table 1. The clusters, hinges, and related dihedrals of Met-enkephalin, shown in Figure 1.

Cluster	Hinge	Dihedral	Residue
$-NH_3^+$	0	(none)	Tyr 1
- c l I -	1	ϕ	
- cl12-	2	χ^1	
- C611'1 -	3	χ^2	
- OII	4	χ^6	
(c o) -	5	ψ	
- (NH)-	6	ω	Gly 2
- c l 1 2 -	7	ϕ	
- (CO)–	8	ψ	
–(NH)-	9	ω	Gly 3
$-CH_2-$	10	ϕ	
- (c~o)-	11	ψ	
- (NH)-	12	ω	Phe 4
- c l I -	13	ϕ	
- CH_2-	14	χ^1	
--C_6H_5	15	χ^2	
-(CO)-	16	ψ	
- (A'H)-	17	ω	Met 5
- c l l -	18	ϕ	
- c o O -	19	ψ	
-C.132-	20	χ^1	
- cl12--	21	χ^2	
- SII	22	χ^3	

Table 2. Proteins and peptides used in NEIMO simulations. The structures listed are the initial Protein Database files, except for the peptides "MEnk" and "Ala9," which were created using the BIOGRAF peptide builder.

Protein	structure	Ref.	Residues	atoms	$\mathcal{N}$
Met-Enkephalin	MEnk	-	5	48	28
$(Ala)_9$	Ala9		9	57	32
Avian Pancreatic Polypeptide	1ppt	17	36	368	192
Crambin	1crn	18	46	402	216
Plastocyanin	7pcy	19	98	857	460
Troponin-C	5tnc	20	161	1514	857
Alpha-Lytic Protease	2alp	21	198	1748	959
Carbonic Anhydrase	2ca2	**22**	256	**2482**	1305
Carboxypeptidase A_α	4cpa	**23**	**307**	2986	1581
Tomato Bushy Stunt Virus	2tbv	**24**	**893**	8083[a]	4335

[a] The TBSV simulations also include 6 Ca^{+}, 25 Cl^{-}, and 24 Na^{+} (see Section III.D.2) for a total of 8138 atoms per unit. Including all 60 units this leads to 488,280 atoms.

Table 3. Times per time step for 100 steps of dynamics for various protein/peptides systems. The average times per time step of the NEIMO calculation and the nonbond calculation are given, along with the NEIMO time divided by $\mathcal{N}$, and the nonbond time divided by the number of atoms, n.

	Number of	NEIMO		Nonbonds		
Protein	Atoms	Time[a] (s)	Time/$\mathcal{N}$ (ms)	Method	Time (s)	Time/n (11")s)
MEnk	48	0.011	0.393	All NB	0.044	0.92
Ala9	57	0.012	0.375	All NB	0.061	1.07
1ppt	368	0.06'4	0.438	All NB	1.933	5.25
1 Cl'I]	402	0.102	0.472	All NB	2.322	5.78
7pcy	857	0.220	0.478	All NB	10.121	11.81
1ppt	368	0.084	0.438	CMM	1.408	3.83
1crn	402	0.102	0.472	CMM	1.950	4.85
7pcy	857	0.220	0.478	CMM	3.541	4.13
5tnc	1514	0.411	0.480	CMM	9.180	6.06
2alp	1748	0.460	0.480	CMM	11.153	5.26
2ca2	2482	0.629	0.482	CMM	15.612	6.29
5cpa	2986	0.762	0.482	CMM	22.733	7.61
2tbv	8083	2.094	0.483	CMM	55.439	6.86

[a] Timing resolution is 0.01 sec.

Table 4. The average values of the Met-enkephalin dihedrals from 5 ps NEIMO ($\langle\theta\rangle_N$) and Cartesian ($\langle\theta\rangle_C$) dynamics simulations, compared to the initial values θ_0 and compared to each other.

Dihedral	θ_0	$\langle\theta\rangle_N$	$\delta\theta_{N0}$	$\langle\theta\rangle_C$	$\delta\theta_{C0}$	$\delta\theta_{CN}$
1	186.3	191.1	4.8	189.6	3.3	- 1.5
2	70.6	66.6	-4.0	76.1	5.5	9.5
3	107.4	99.9	-7.5	104.6	-2.8	4.7
4	178.0	178.0	0.0	180.2	'2.2	2.2
5	300.1	306.8	6.7	301.4	1.3	-5.4
6	185.1	183.4	-1.7	186.0	2.6	2.6
7	311.5	272.0	-39.5	279.4	-32.1	7.4
8	306.8	300.8	-6.0	55.8	109.0	115.0
9	182.1	177.7	-4.4	173.0	-9.1	-4.7
10	294.8	271.7	-23.1	263.4	-31.4	-8.3
11	353.2	303.8	-49,4	309.8	-43.4	6.0
12	173.4	173.9	0.5	172.3	-1.1	-1.6
13	246.1	'241.6	-4.5	254.7	-8.6	13.1
14	61.3	65.5	4.2	71.2	9.9	5.7
15	72.0	92.8	20.8	99.8	27.0	7.6
16	351,7	320.0	-31.7	320.4	-31.3	0.4
17	184.0	177.2	-6.8	176.3	-7.7	-0.9
18	238.8	241.0	2.2	245.6	6.8	4.6
19	116.3	128.8	12.5	116.0	-().3	-12.8
20	33.3	46.0	12.7	289.8	-103.5	-116.4
21	70.1	89.2	19.1	113.5	43.4	24.3
22	82.0	117.2	35.2	159.4	77.4	-39.8

Figure Captions

Figure 1. The peptide Met-enkephalin with its hinges numbered (see text and Table 1). Bonds which are not numbered are held fixed. Clusters are units which remain fixed during dynamics, such as the phenyl group of Tyr 1, located between hinges 3 and 4. The Last cluster is the N-terminal amino group.

Figure 2. (a) A plot of computational time vs. protein size (n = number of atoms) for nonbond calculations and NEIMO. (b) CPU time/n vs. protein size. Times are given in CPU seconds per dynamics step, as determined on an SGI Indigo R3000 workstation,

Figure 3. Energy fluctuations, $\mathcal{E}$, for NEIMO (N) and Cartesian (C) dynamics simulations of Met-enkephalin (48 atoms, 23 clusters). Simulations were run for 1 ps using time steps ranging from 1 to 20 fs. N* and C* are the scaled fluctuations, $\mathcal{E}^*$, where $\mathcal{E}$ is divided by the number of degrees of freedom: $\mathcal{N}$ for NEIMO simulations and $3n - 6$ for Cartesian coordinates. For Cartesian dynamics, the system was equilibrated for 1 ps before calculating $\mathcal{E}$.

Figure 4. $\delta_E = \sqrt{\Delta E}/\langle E \rangle$ for 100 time steps of NEIMO dynamics on $(Ala)_9$.

Figure 5. The alpha carbon trace of Avian pancreatic polypeptide (aPP). From the crystal structure 1PPT.[17]

Figure 6. Energy fluctuations, $\mathcal{E}$, for NEIMO (N) and Cartesian (C) dynamics simulations of avian pancreatic polypeptide (aPP). Simulations were run for 1 ps using time steps ranging from 1 to 15 fs. Time steps above 11 fs caused the energy to diverge. N* and C* are the scaled energy fluctuations, $\mathcal{E}^*$.

Figure 7. Scaled energy fluctuations, $\mathcal{E}^*$, for 1 ps NEIMO simulations of aPP. "Rigid H" differs from "Normal" NMHh40 in that hinges which rotate only hydrogen atoms are held fixed. The "Counterions" simulation used the standard NEIMO method for the protein, but concurrently solved the Cartesian equations of motion for counterions ($5\ Na^+$ and $3\ Cl^-$) added to neutralize unpaired charges.

Figure 8. The average energy fluctuations, $\langle \mathcal{E} \rangle$, during 5 ps simulations of avian pancre-

atic polypept ide. Fluctuations in NE IMO (N) and Cartesian (C) dynamics were determined at 0.1 ps intervals during the course of the simulation, after which velocities could be rescaled and the CMM nonbond farfield talc.ulatiml was updated.

Figure 9. $\langle \mathcal{E} \rangle^*$ vs. protein size for 5.0 ps simulations. Values are given for Cartesian dynamics using 1 fs time steps and NEIMO dynamics usi ng time steps ranging from 1 fs to 5 fs.

Figure 10. During 5 ps molecular dynamics simulations of Met-cl)kqdlalill, the 22 dihedral angles were written out at 0.1 ps intervals. The fifty Values for each dihedral are plotted here for Cartesian and NEIMO dynamics simulations using 1 fs time steps.

Figure 11. The average dihedrals from the distributions in Figure 10 are shown here with error bars indicati ng $\pm\sigma$, the standard deviations.

Figure 12. The average dihedrals from NEIMO simulations using time steps ranging from 1 to 10 fs.

Figure 13. The overlaps S_{12} between 1 fs and 2-10 fs NEIMO simulations of Met-enkephalin.

Figure 14. The overlaps S_{12} between 1 fs and 2, 5, and 10 fs NEIMO simulations of Met-cmkcphalin shown at higher resolution than Figure 13.

Figure 15. The **overlaps** S_{12} between dihedral distributions from Cartesian dynamics versus those from N EIMO dynamics simulations with time steps ranging from 1 fs.

Figure 16. The strut.tum of tomato bushy stunt virus (TBSV). Shown is the van der Waal surface for the entire protein. Each of the three independent chains is given a different color (red for chain A, green for chain B, blue for chain C). We thank K. T. Lim for constructing this figure.

Figure 17. CPU times for CMM calculations, NEIMO acceleration calculations, and overhead. Overhead includes coordinate updat ing for NE I M O and other parts of Biograf.

Figure 18. Energy fluctuations, $\langle \mathcal{E} \rangle^*$ for 1.0 ps of NEIMO simulations.

Figure 19. The TBSV radius during 2.0 ps of Cartesian and NEIMO dynamics simulations.

Figure 20. The radius of the pH7 and NoCa models of TBSV during 4.0 ps of Cartesian (NVT) and NEIMO dynamics.

Figure 21. Potential energy during the 4.0 ps Cartesian canonical dynamics simulations.

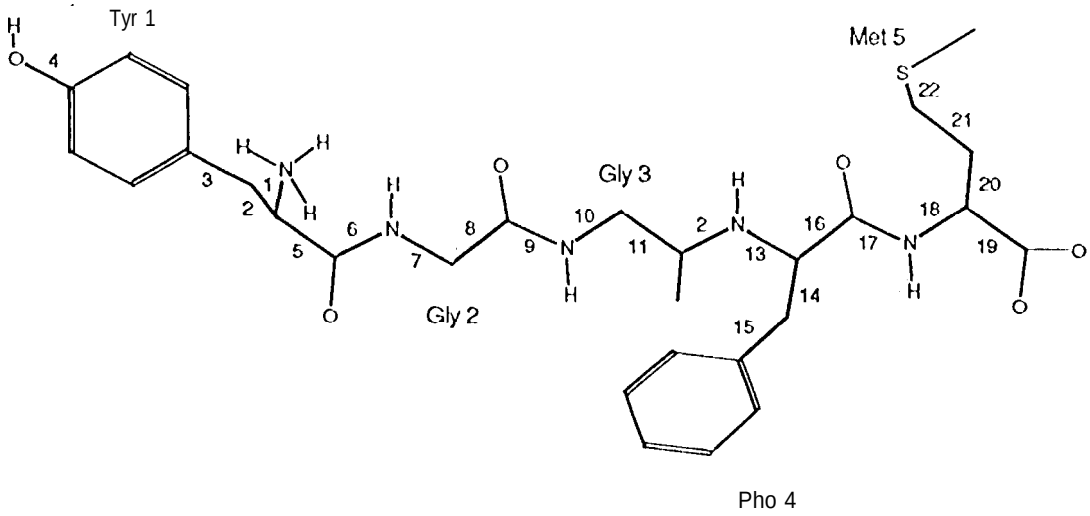

Figure 1:

Figure 2:

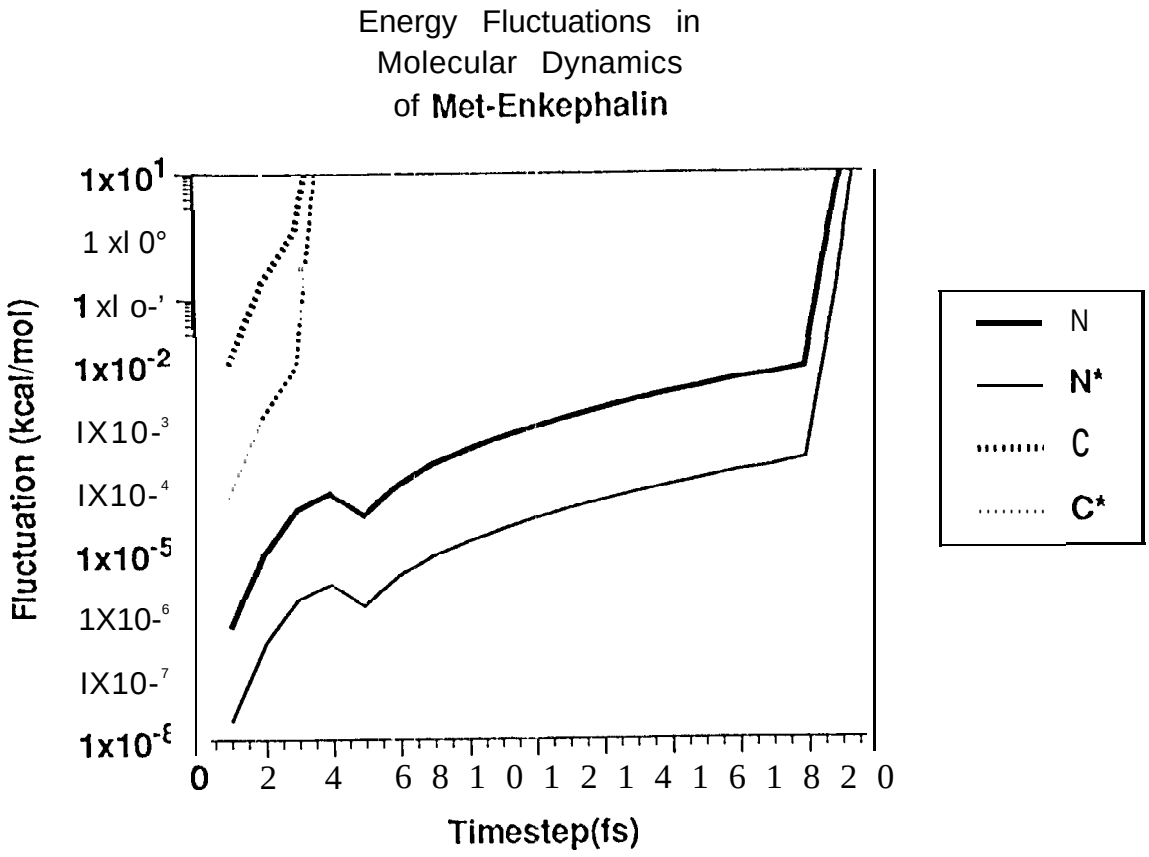
Energy Fluctuations in
Molecular Dynamics
of Met-Enkephalin
Fluctuation (kcal/mol)
1x10^1
1 xl 0°
1 xl o-'
1x10^-2
IX10-^3
IX10-^4
1x10^-5
1X10-^6
IX10-^7
1x10^-8
0 2 4 6 8 1 0 1 2 1 4 1 6 1 8 2 0
Timestep(fs)
N
N*
C
C*

Figure 3:

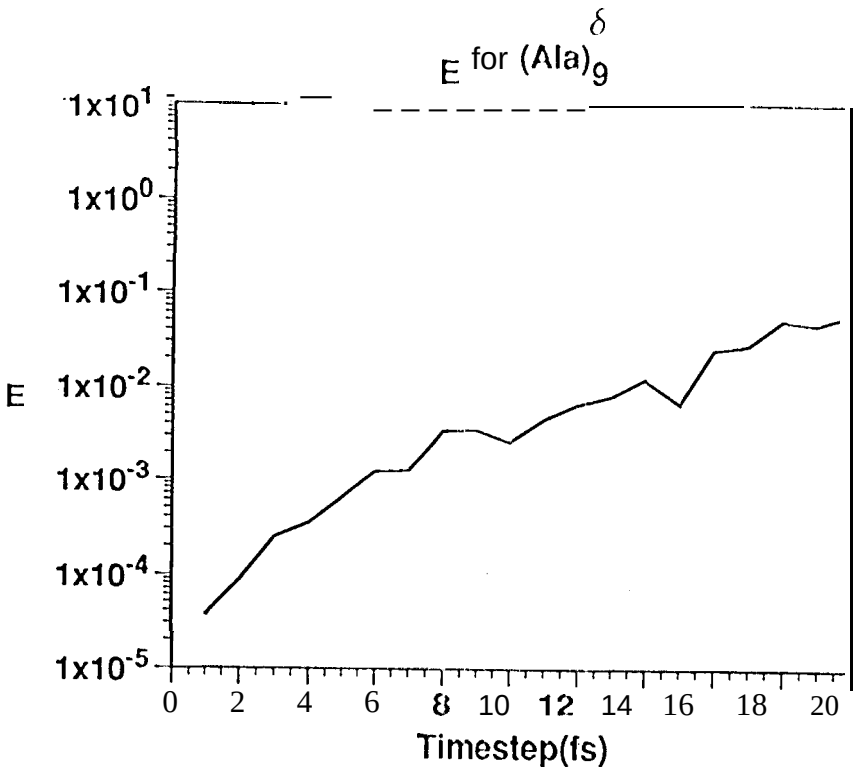

Figure 4:

Figure 5:

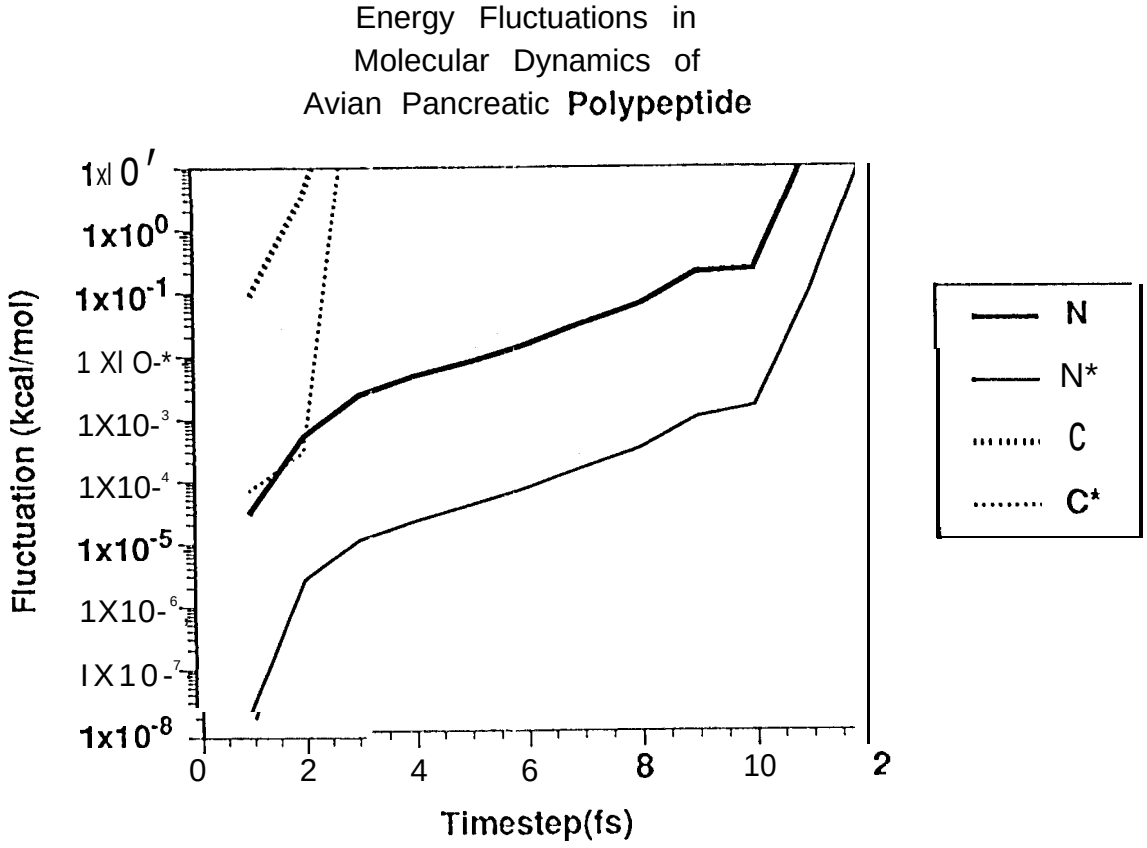
Energy Fluctuations in
Molecular Dynamics of
Avian Pancreatic Polypeptide
Fluctuation (kcal/mol)
1x10^0
1x10^-1
1X10^-3
1X10^-4
1x10^-5
1X10^-6
1X10^-7
1x10^-8
0
2
4
6
8
10
Timestep(fs)
N
N*
C
C*

Figure 6:

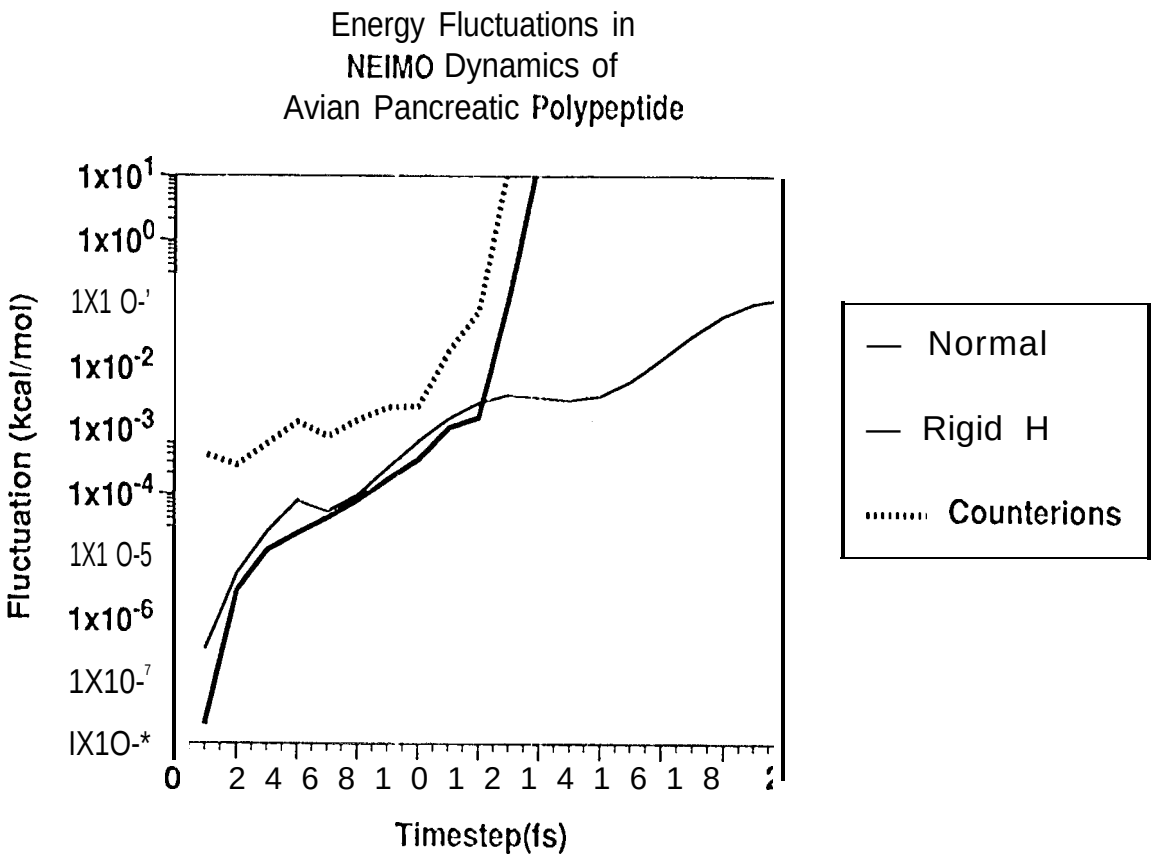
Energy Fluctuations in
NEIMO Dynamics of
Avian Pancreatic Polypeptide
Fluctuation (kcal/mol)
1x10^1
1x10^0
1X1 0-'
1x10^-2
1x10^-3
1x10^-4
1X1 0-5
1x10^-6
1X10-^7
IX1O-*
0
2 4 6 8 1 0 1 2 1 4 1 6 1 8
Timestep(fs)
Normal
Rigid H
Counterions

Figure 7:

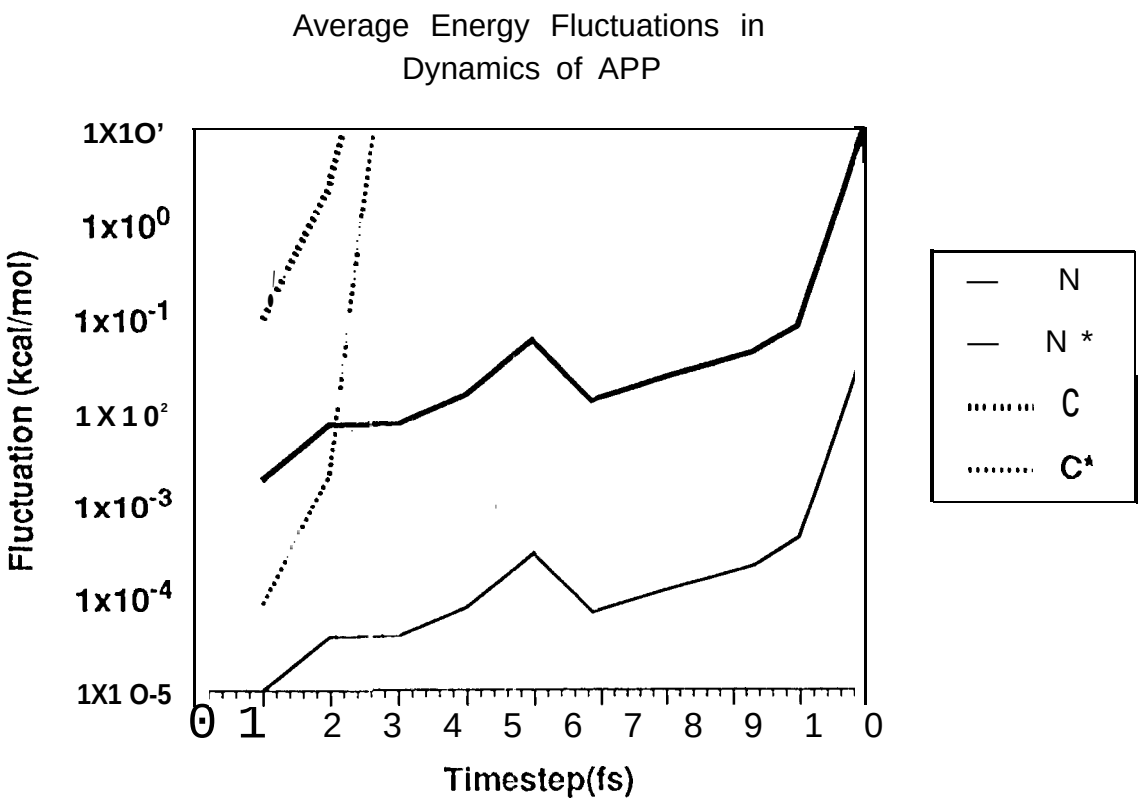
Average Energy Fluctuations in Dynamics of APP
Fluctuation (kcal/mol)
1X1O'
1x10^0
1x10^-1
1 X 1 0^2
1x10^-3
1x10^-4
1X1 O-5
0 1 2 3 4 5 6 7 8 9 1 0
Timestep(fs)
N
N *
C
C*

Figure 8:

Figure 9:

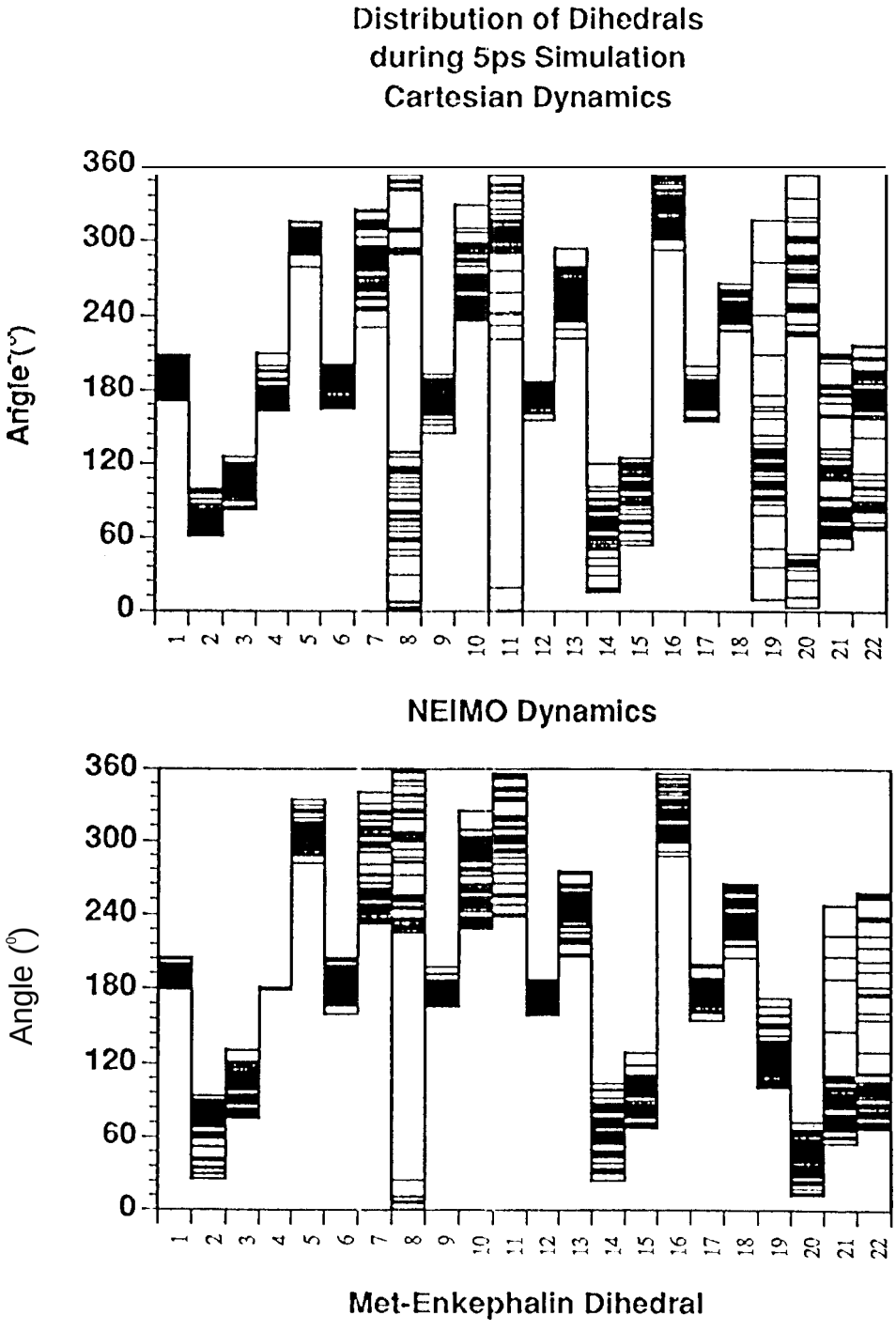
Distribution of Dihedrals
during 5ps Simulation
Cartesian Dynamics
Angle(°)
360
300
240
180
120
60
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
NEIMO Dynamics
Angle (°)
360
300
240
180
120
60
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Met-Enkephalin Dihedral

Figure 10:

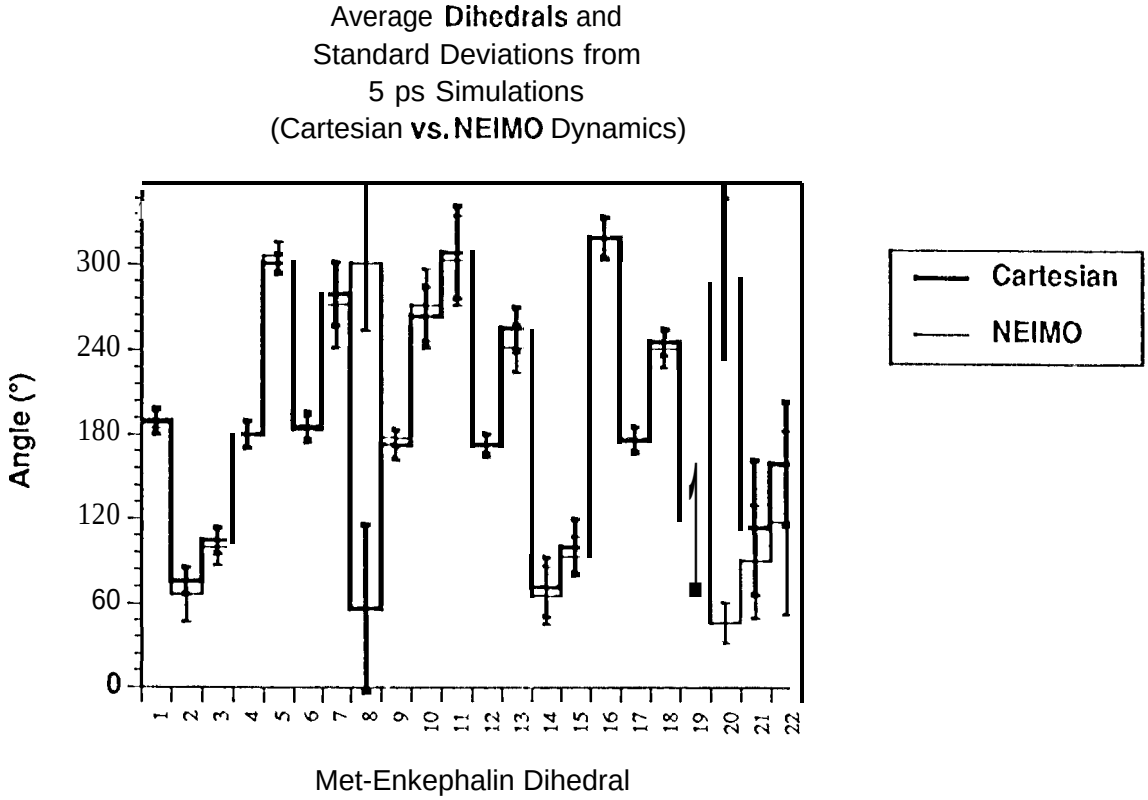
Average Dihedrals and
Standard Deviations from
5 ps Simulations
(Cartesian vs. NEIMO Dynamics)
Angle (°)
300
240
180
120
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Met-Enkephalin Dihedral
Cartesian
NEIMO

Figure 11:

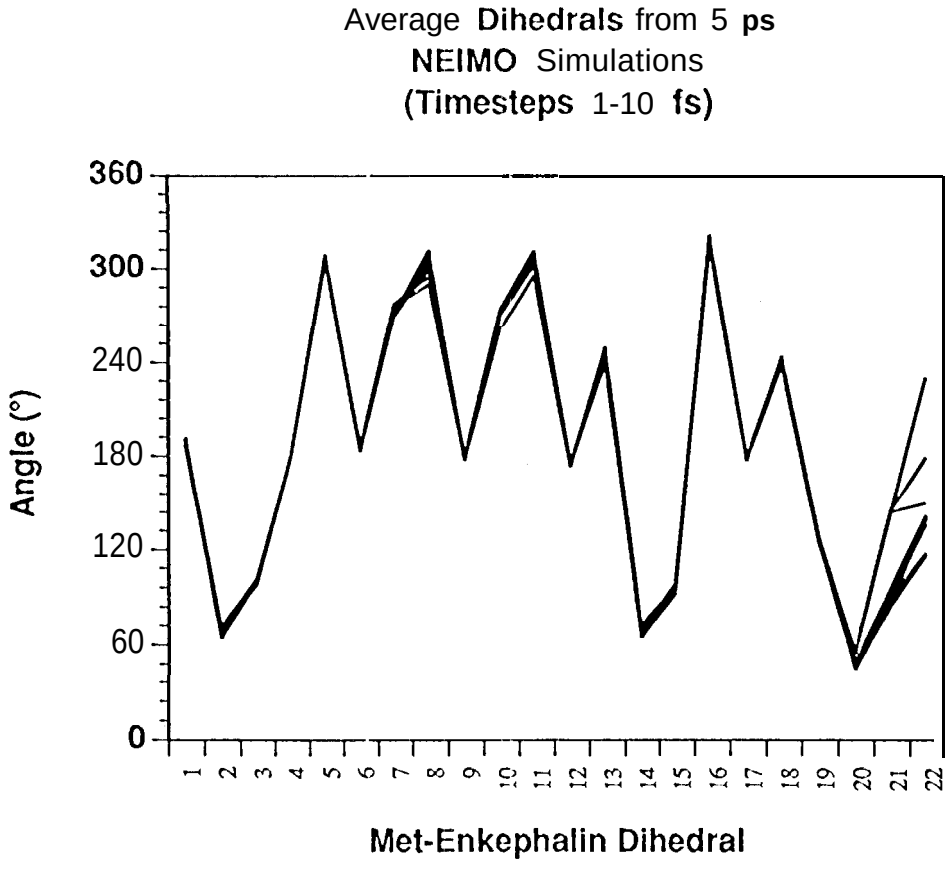
Average Dihedrals from 5 ps
NEIMO Simulations
(Timesteps 1-10 fs)
360
300
240
180
120
60
0
Angle (°)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Met-Enkephalin Dihedral

Figure 12:

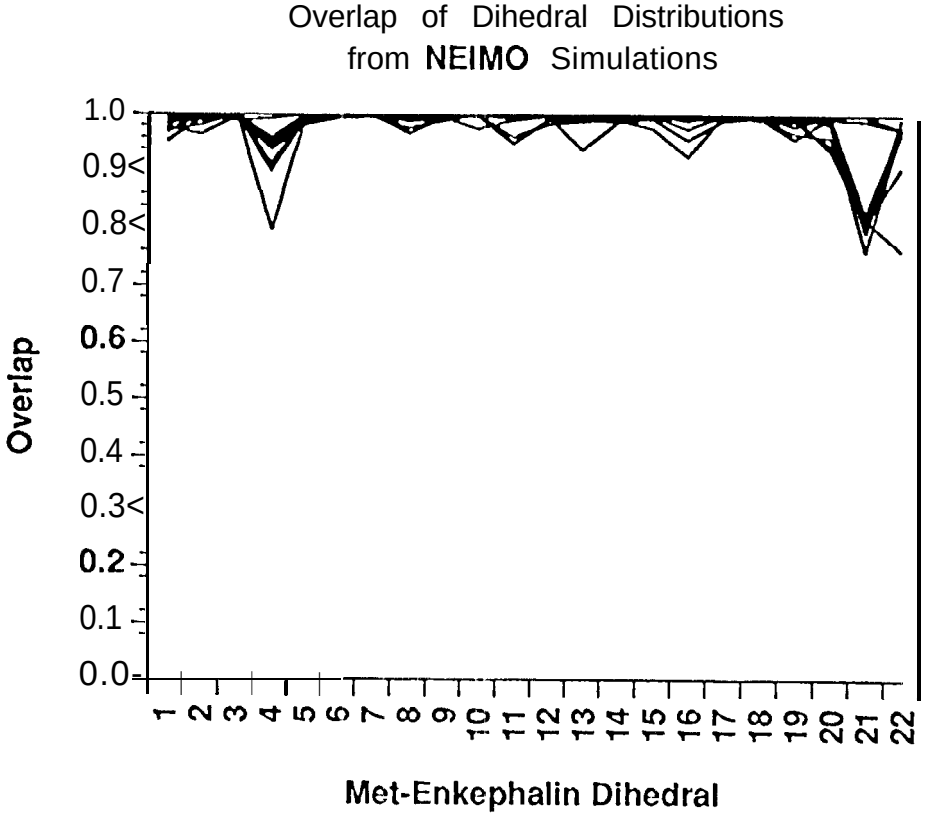
Overlap of Dihedral Distributions
from NEIMO Simulations
Overlap
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Met-Enkephalin Dihedral

Figure 13:

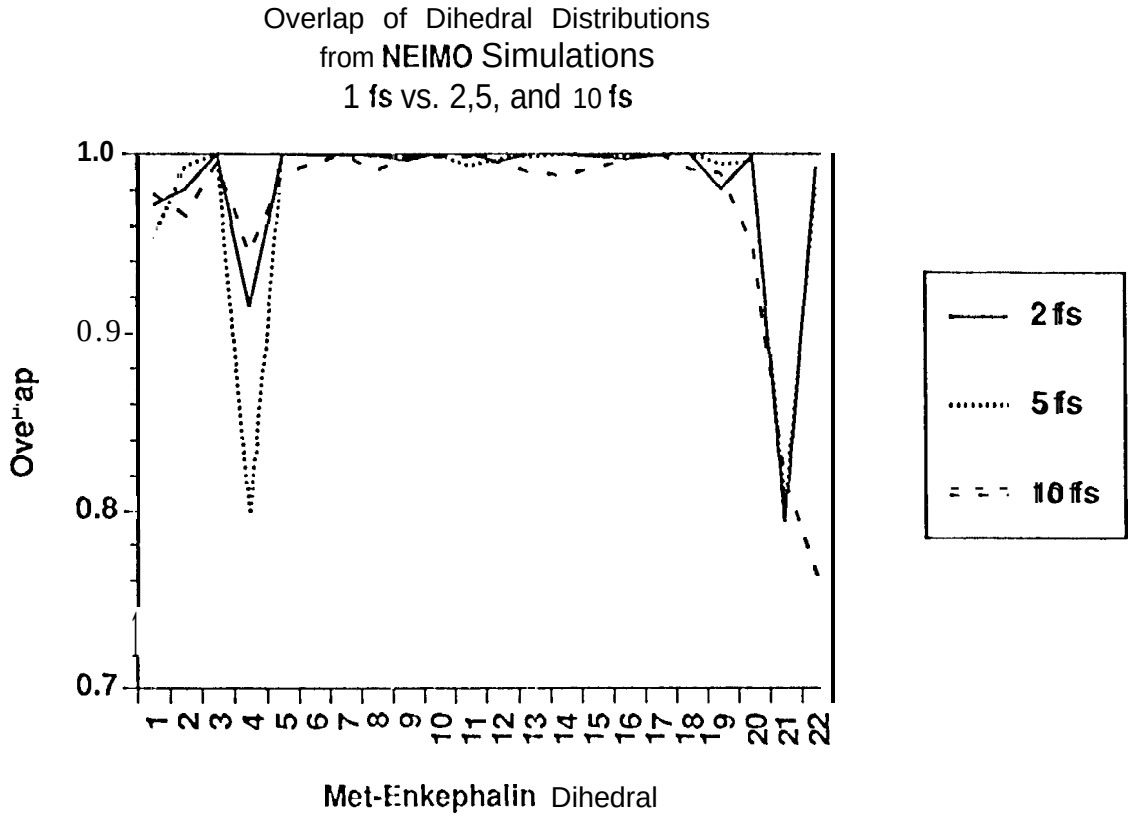
Overlap of Dihedral Distributions
from NEIMO Simulations
1 fs vs. 2,5, and 10 fs
1.0
0.9
0.8
0.7
Overlap
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Met-Enkephalin Dihedral
2 fs
5 fs
10 fs

Figure 14:

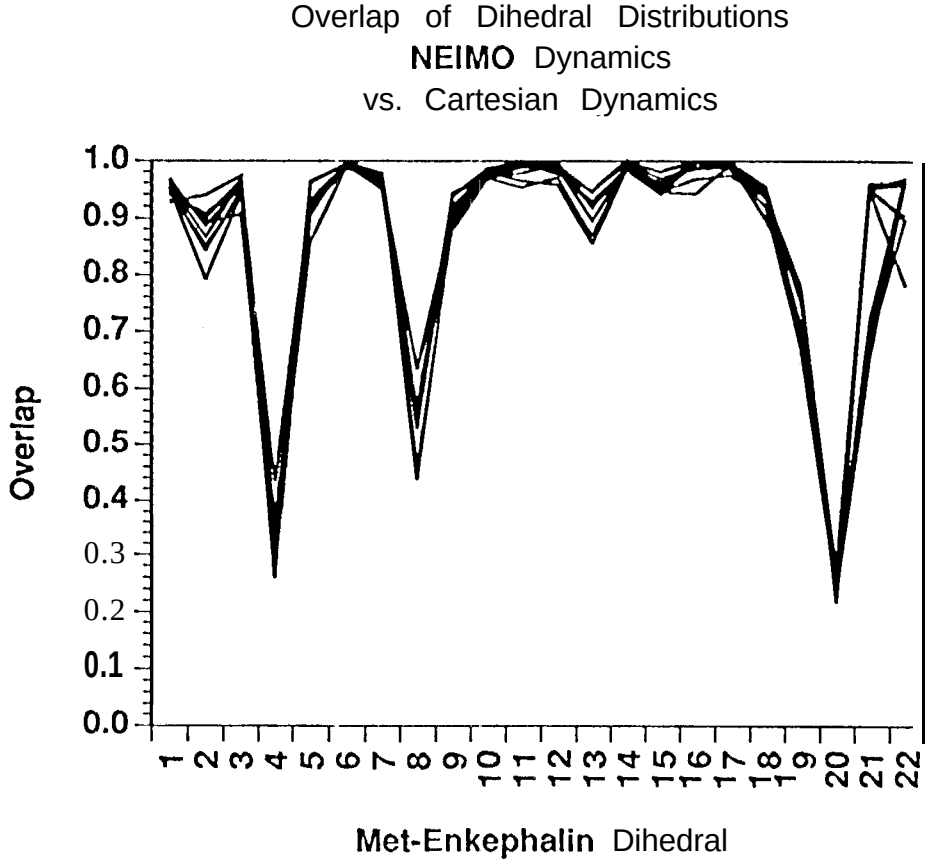
Overlap of Dihedral Distributions
NEIMO Dynamics
vs. Cartesian Dynamics
Overlap
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Met-Enkephalin Dihedral

Figure 15:

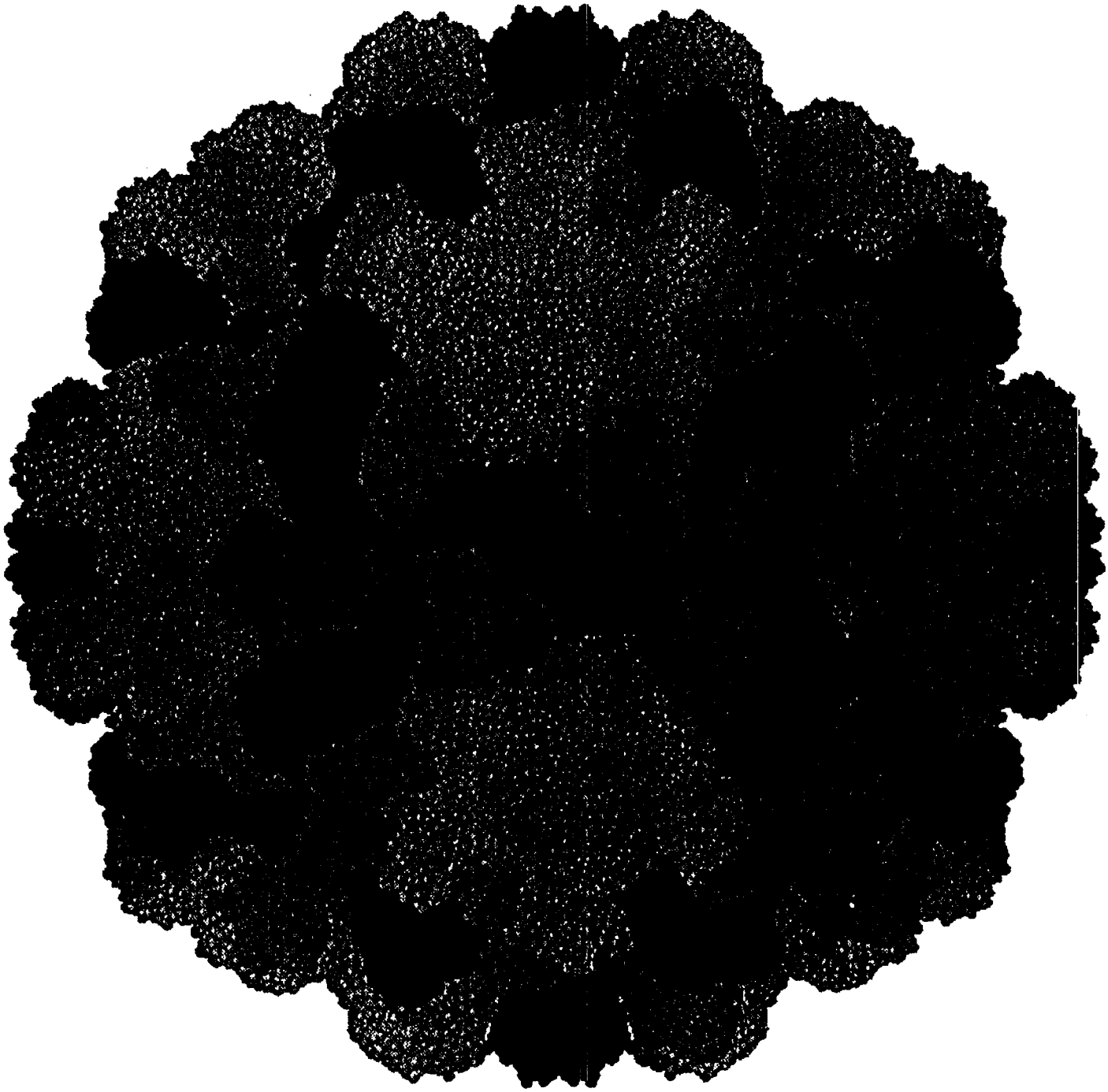

Figure 16

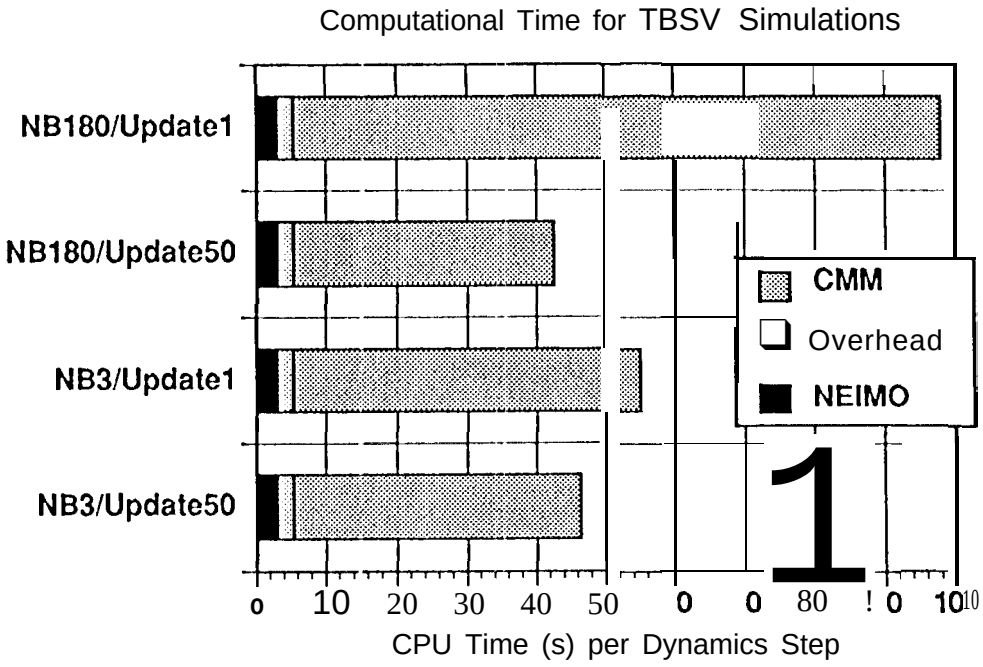
Computational Time for TBSV Simulations
NB180/Update1
NB180/Update50
NB3/Update1
NB3/Update50
CMM
Overhead
NEIMO
0
10
20
30
40
50
0
0
80
! 0
10
CPU Time (s) per Dynamics Step

Figure 17:

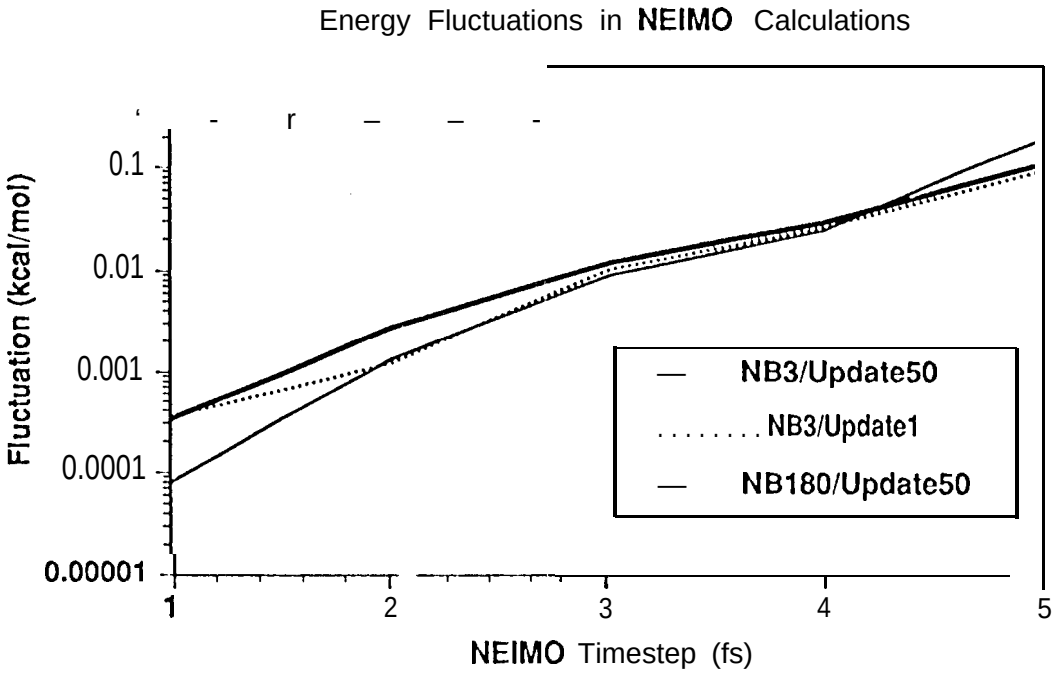
Energy Fluctuations in NEIMO Calculations
Fluctuation (kcal/mol)
0.1
0.01
0.001
0.0001
0.00001
1
2
3
4
5
NEIMO Timestep (fs)
NB3/Update50
NB3/Update1
NB180/Update50

Figure 18:

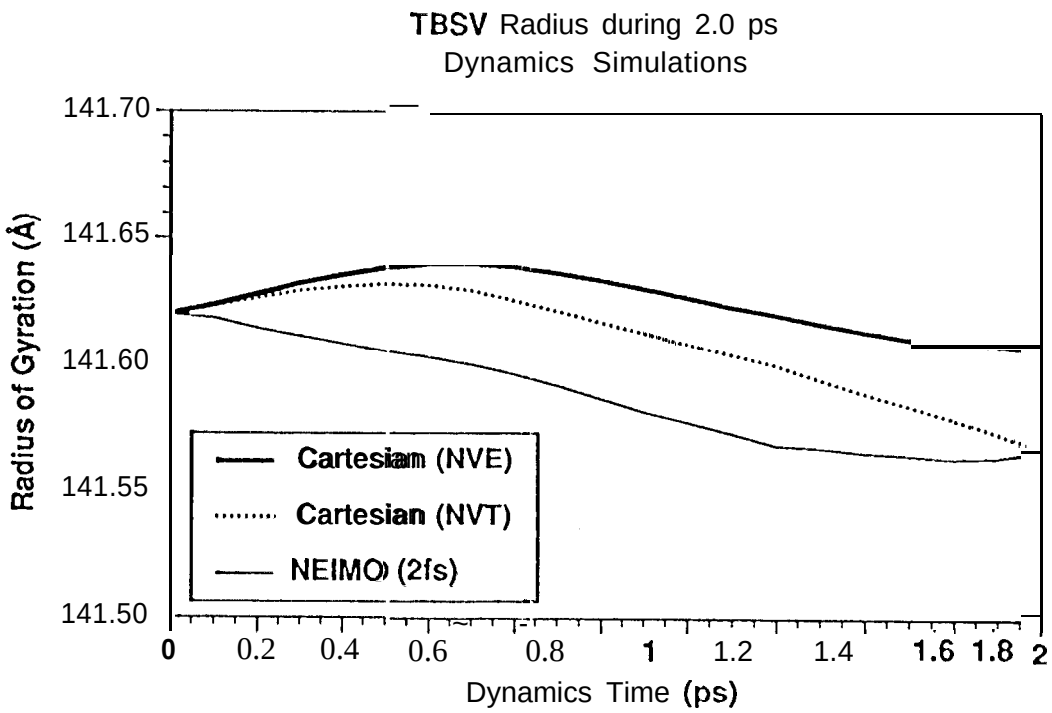
TBSV Radius during 2.0 ps
Dynamics Simulations
141.70
141.65
141.60
141.55
141.50
Radius of Gyration (Å)
Cartesian (NVE)
Cartesian (NVT)
NEIMO (2fs)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Dynamics Time (ps)

Figure 19:

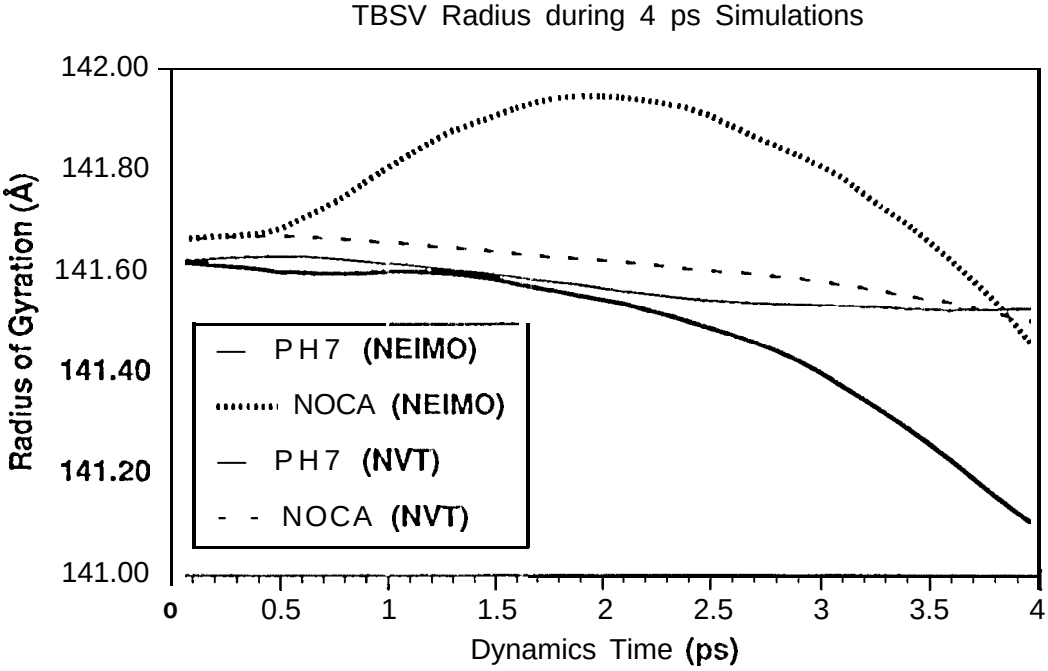
TBSV Radius during 4 ps Simulations
142.00
141.80
141.60
141.40
141.20
141.00
Radius of Gyration (Å)
PH7 (NEIMO)
NOCA (NEIMO)
PH7 (NVT)
NOCA (NVT)
0
0.5
1
1.5
2
2.5
3
3.5
4
Dynamics Time (ps)

Figure 20:

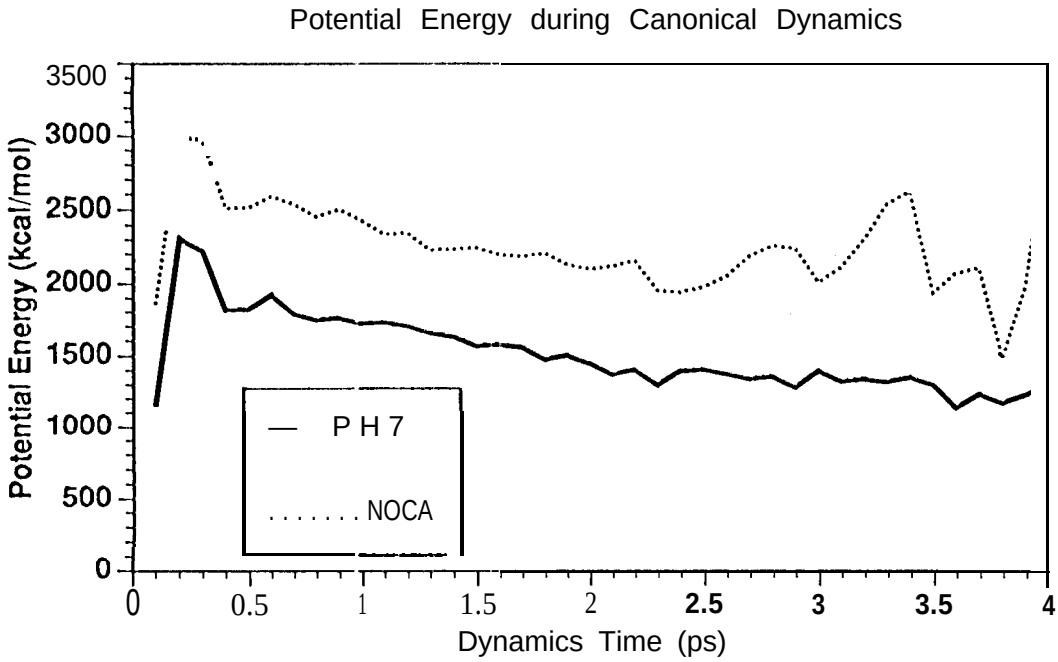
Potential Energy during Canonical Dynamics
Potential Energy (kcal/mol)
3500
3000
2500
2000
1500
1000
500
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Dynamics Time (ps)
— P H 7
. . . . . NOCA

Figure 21: