Crystal Structure Prediction using CRYSTALG program

Yelena Arnautova

Baker Laboratory of Chemistry and Chemical Biology, Cornell University

Problem of crystal structure prediction:

- theoretical importance

- practical importance for design and industrial processing of molecular materials (pharmaceuticals, pigments, optical materials, explosives)

> Polimorphism is the ability of a particular substance to exist in several different crystal forms. Polymorphs of a compound differ in their physical and biological properties.

The solubility and rate of dissolution can vary, which influences the bioavailability.

Differences in density, hardness, and crystal morphology can influence the handling properties in the production process.

Color is dependent on the crystal packing and is essential for pigments.

The existence of different polymorphs can cause problems with patent protection.

The stability of crystal structure at a given temperature and pressure is determined by its free energy

$$G(T,P) = U(T,P) + pV - TS$$

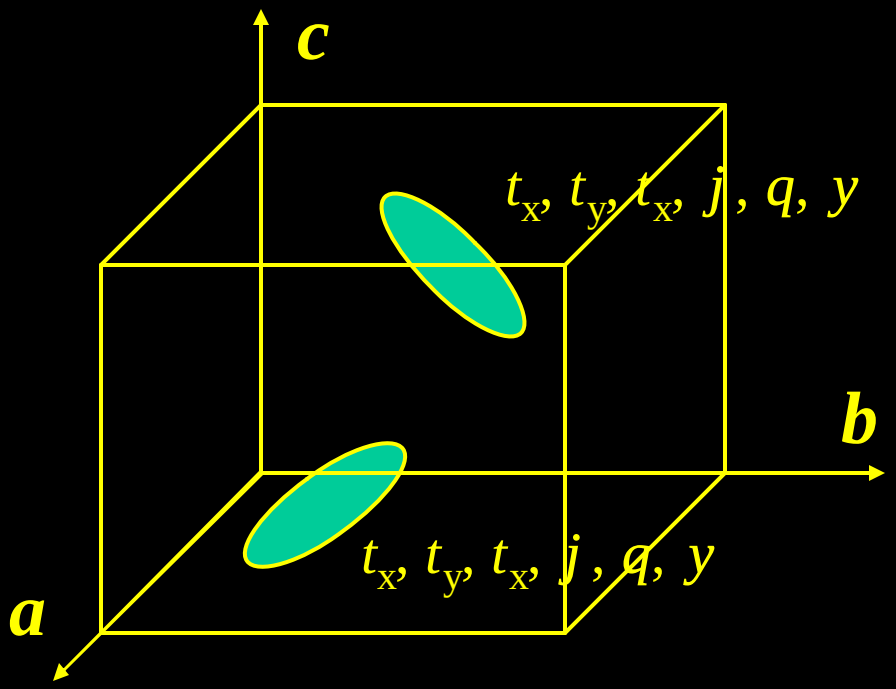
Thermodynamic hypothesis:

The observed structure corresponds to the global minimum of the free energy (G)

The thermodynamic stability is not always the decisive factor because kinetics plays an important role, influenced by crystallization conditions.

We can assume that the observed structure is among structures having a low free energy. The latter can be approximated by the potential energy *U* (thermal effects may contribute as much as 2 kcal/mol).

Crystal structure prediction is search for lowest energy minima of potential energy (which correspond to possible polymorphs) given only the atomic connectivity for an organic compound.



$E=E(a,b,c,a,b,g,t_x^1,t_y^1,t_z^1,f^1,q^1,y^1,..$

$6+Z(6+N)$ variables,

where Z is number of molecules in the unit cell,

N is number of torsional angels in the case of flexible molecule.

Global optimization of potential energy (E) = multiple minima problem

PROBLEM: huge number of local minima = a very efficient global optimization method is needed.

A good potential function is a necessary element for any algorithm for crysta
tructure prediction. On the other hand, crystal structure prediction algorithm
based on global optimization is a very useful tool for assessing the quality of a
given potential.

Good (predictive) potential function:

a) Should reproduce the experimental structure within a certain accuracy
b) If experimental value of sublimation enthalpy is known, it should be
eproduced as well.
c) The structures corresponding to the lowest-energy minima found for the
otential should represent plausible structures, and one of them, preferably the
global minimum, should correspond to the observed most stable experimental
tructure

Most potentials are derived based on a series of local minimizations instead of a
global optimization, therefore they may not satisfy criterion (c).

Blind test of crystal structure prediction of small organic molecules

(organized by the Cambridge Crystallographic Data Centre)

May 1999 – *Acta Crystallographica*, **2000**, *B*56, 697-714

May 2001 - *Acta Crystallographica*, **2002,** in press

Global optimization methods for crystal structure prediction

- Self-Consistent Basin-to-Deformed-Basin Mapping (SCBDBM) Method

Deformation-based method developed in Scheraga group in 1998. It has been successfully applied to many different systems such as proteins, Lennard-Jones clusters and crystals.

- Conformation-Family Monte Carlo (CFMC)

A variant of Monte-Carlo Minimization (MCM). It has been recently developed in our laboratory in application to proteins and crystals.

Monte Carlo-Minimization (MCM)

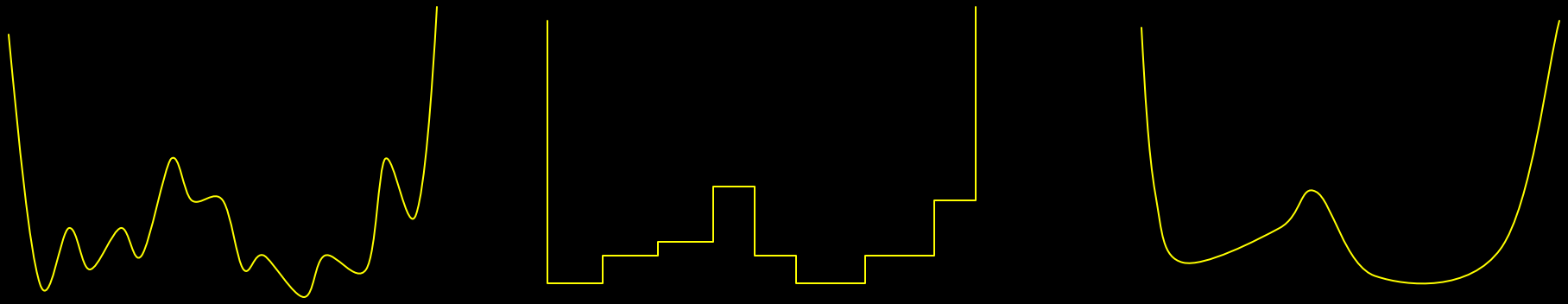
1. Select a conformation at random, and minimize the potential energy.
2. Select any variable at random.
3. Make a random change in the variable.
4. Minimize the energy of the new conformation.
5. Compare E_{new} to E_{old} by means of the Metropolis criterion.
6. Iterate.

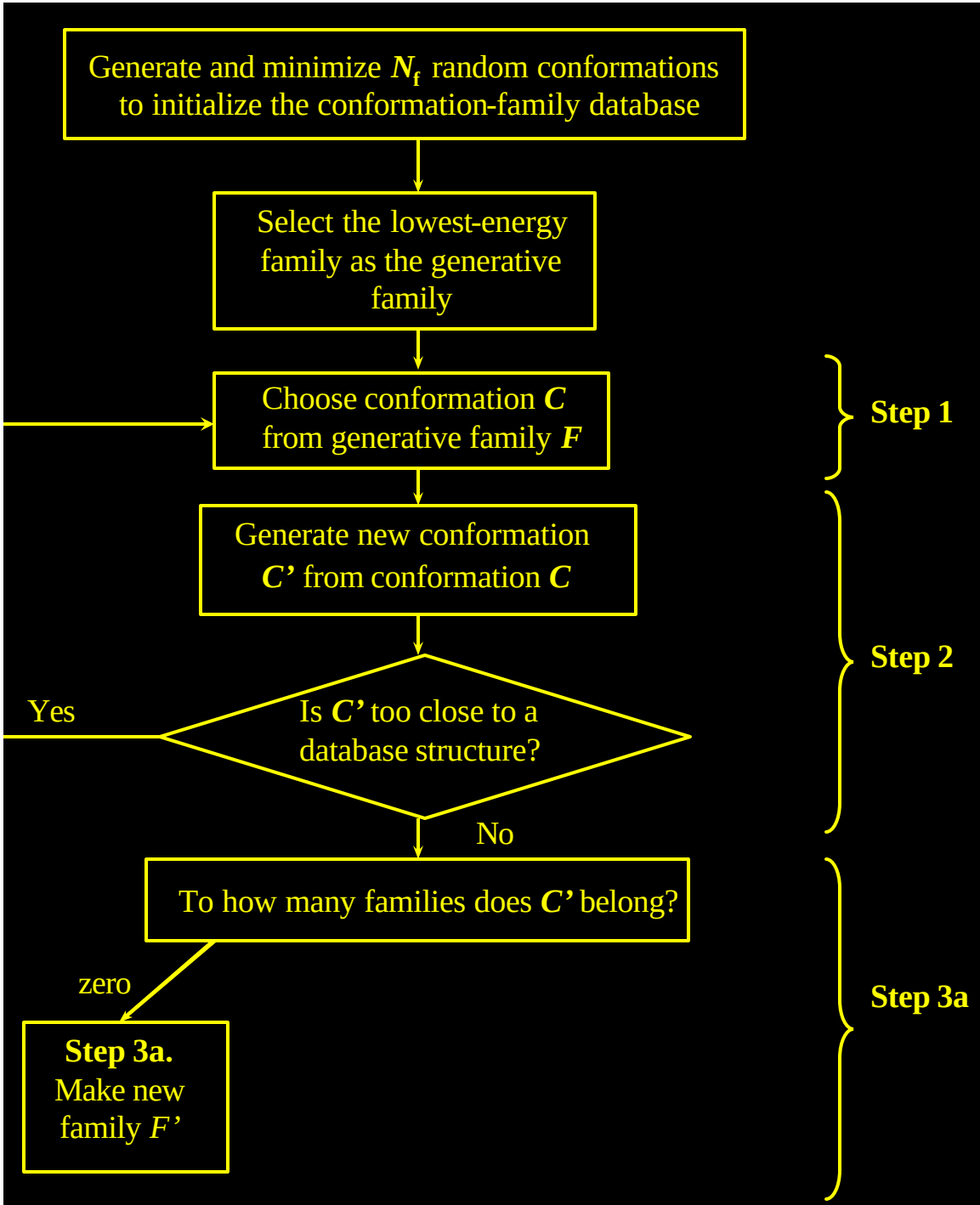
Conformation-Family Monte Carlo (CFMC)

Helvet.Chim. Acta **2000**, 83 (9), 2214-2230; *PNAS* **2001**, 98 (22),12351-12356

The central element of the CFMC method is the structure family database

- Uses Metropolis criterion to move between families
- Uses Boltzmann distribution to choose conformation from a family
- Does not move between structures, but between families
- It is equivalent to smoothed “staircase deformation” of a potential function
- It has been successfully applied for protein structure prediction and performs as well as the Conformational Space Annealing (CSA)



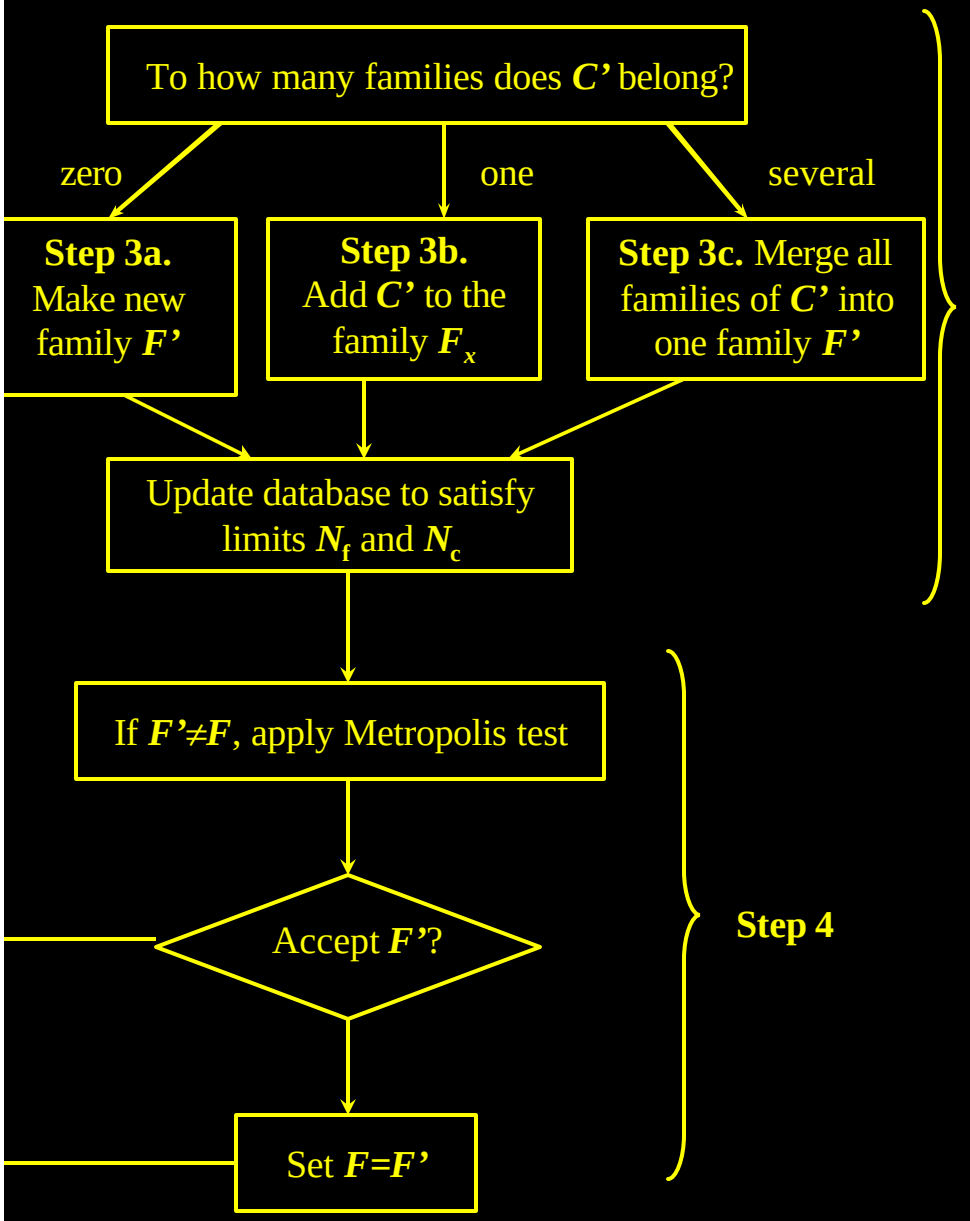


CFMC algorithm

Step 1. A conformation ***C*** is chosen at random from the generative family ***F***, with a probability proportional to its Boltzmann weight.

Step 2. This conformation ***C*** is modified to yield a new conformation ***C´***.

Step 3a. If the new conformation ***C´*** is unconnected to any family in the database, a new family ***F´*** is created whose sole member is ***C´***. If the number of families in the database exceeds the limit N_f, the family with the highest energy is eliminated.



Step 3b. If the new conformation C' connected to exactly one family F' in t database, the conformation is added to t family. If $N > N_c$, the conformation with t highest energy is eliminated.

Step 3c. If the new conformation C' connected to more than one family in t database, these families are merged into single family F'. Two families are merged if has lower energy than every conformation the higher-energy family and C' has a lov energy than at least one conformation of t lower-energy family.

Step 4. If the new family $F' \neq F$, a Metropo criterion is applied to determine whether to make the generative family. F' becomes the new generati family if it has a lower energy than F, or if Boltzmann factor $\exp(-b\Delta E/kT)$ > randon generated number in the interval (0,1)

The Methods for Producing New Conformations

There are two general classes of moves used in the CFMC method:

- *internal* (or local) moves, intended to improve the low-energy conformations within a family
- *external* (global) moves, intended to search the conformational space for new families.

Families and structures are always chosen according to Boltzmann distribution

In order to avoid atomic clashes all newly generated structures are expanded before local minimization.

All newly generated structures are locally minimized.

CFMC moves

	internal moves	external moves
1. Perturbation of translations of all molecules within unit cell by	DPERT_2	D_PERT1
2. Perturbation of all rotations of molecules within unit cell by	D_PERT2A	D_PERT1A
3.	Systematic search in rotations for one molecule	All degrees of freedom are perturbed
4. Averaging. Every variable is calculated according to: $\nu^*_i = \nu^{(1)}_i \cdot x + \nu^{(2)}_i \cdot (1 - x)$	All variables between two structures from the same family	All variables between two structures from different families

Input data

- Covalent structure of target molecule
- Force field parameters
- Number of molecules per unit cell, as a parameter

Variables in the calculation (all vary independently)

- Internal degrees of freedom (torsional angles)
- Intermolecular distances (translational vectors of molecules)
- Euler angles for molecular rotations
- Components a_x, b_x, b_y, c_x, c_y, c_z of the lattice vectors (basis vector ***a*** of unit cell coincides with the ***x*** axis, vector ***b*** lies in the (***x***,***y***) plane, and lattice vectors form a right-handed system

Potential Energy

$$E_{tot} = E_{el} + E_{nb} + E_{tor}$$

E_{el} obtained from Coulomb formula with Ewald summation. Point charges on atoms and additional sites

E_{nb} calculated with Lennard-Jones 6-12 or Buckingham 6-exp potential function

E_{tot} calculated using real part of third-order Fourier expansion

$$E_{tor} = \sum_{i=1}^{3} k_m (1 - \cos mw)$$

(w, k_1, k_2, k_3 obtained by fitting E_{tor} to *difference* between *ab initio* and molecular mechanic (E_{el}+E_{nb}) profiles)

Structure comparison for identification of a family

The following values are being calculated and stored for all structures from the database:

- total energy E_{tot} and volume of the unit cell $V_{unit\ cell}$
- parameters of the unit cell $a, b, c, \alpha, \beta, \gamma$
- set of shortest interatomic distances within a cutoff radius r_d sorted according to their value

Structure comparison is carried out using the similarity measure calculated according to the formula:

$$R = |E_i - E_j|/\Delta E + |V_i - V_j|/\Delta V + \max|r_i^k - r_j^k|/\Delta r,$$

where ΔE, ΔV, Δr are preset values of the largest deviations of energies, volumes and interatomic distances, allowed for equivalent structures.

In order to speed up calculations, all local minimizations (except final ones) are carried out with relatively low accuracy.

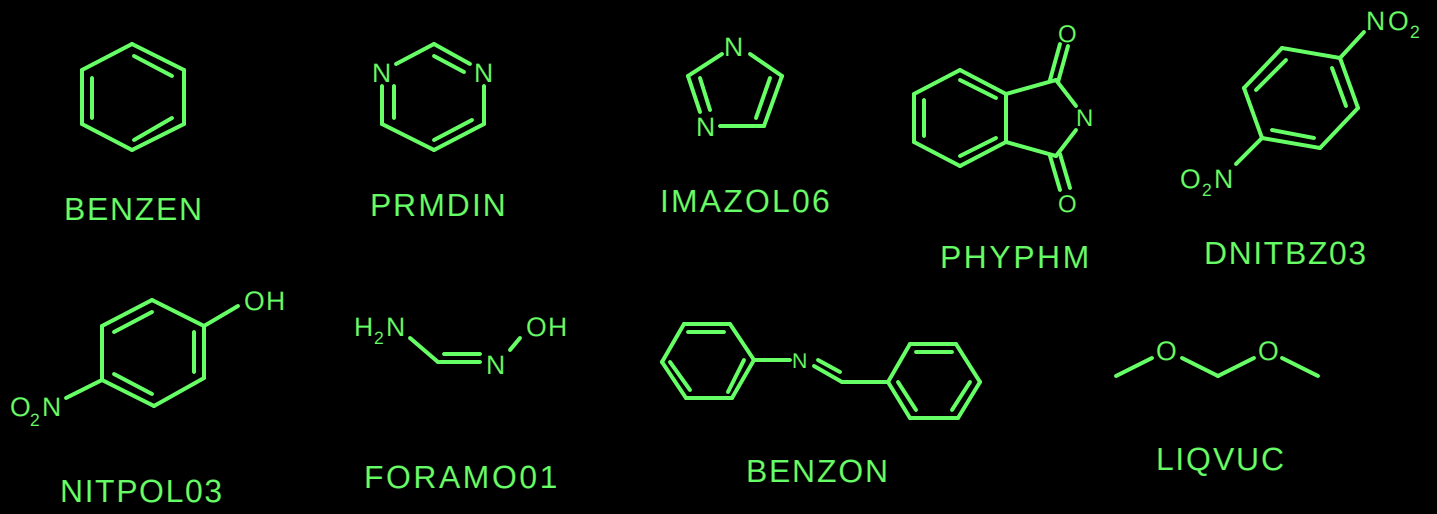
The geometrical focusing is achieved by gradual lowering R cutoff for structural differences.

Application to Crystal Structure Prediction

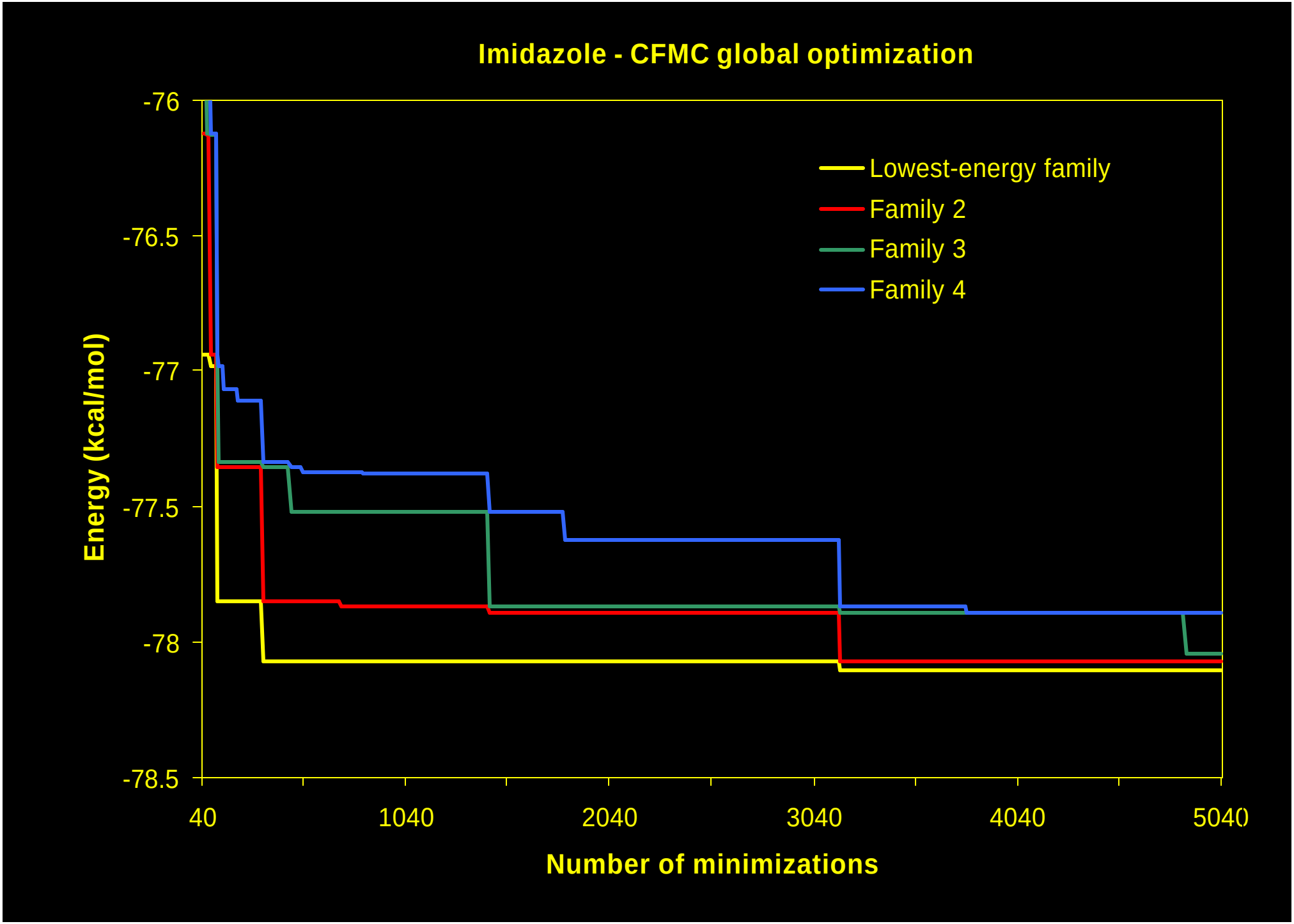
PNAS, **2001**, 98 (22), 12351-12356

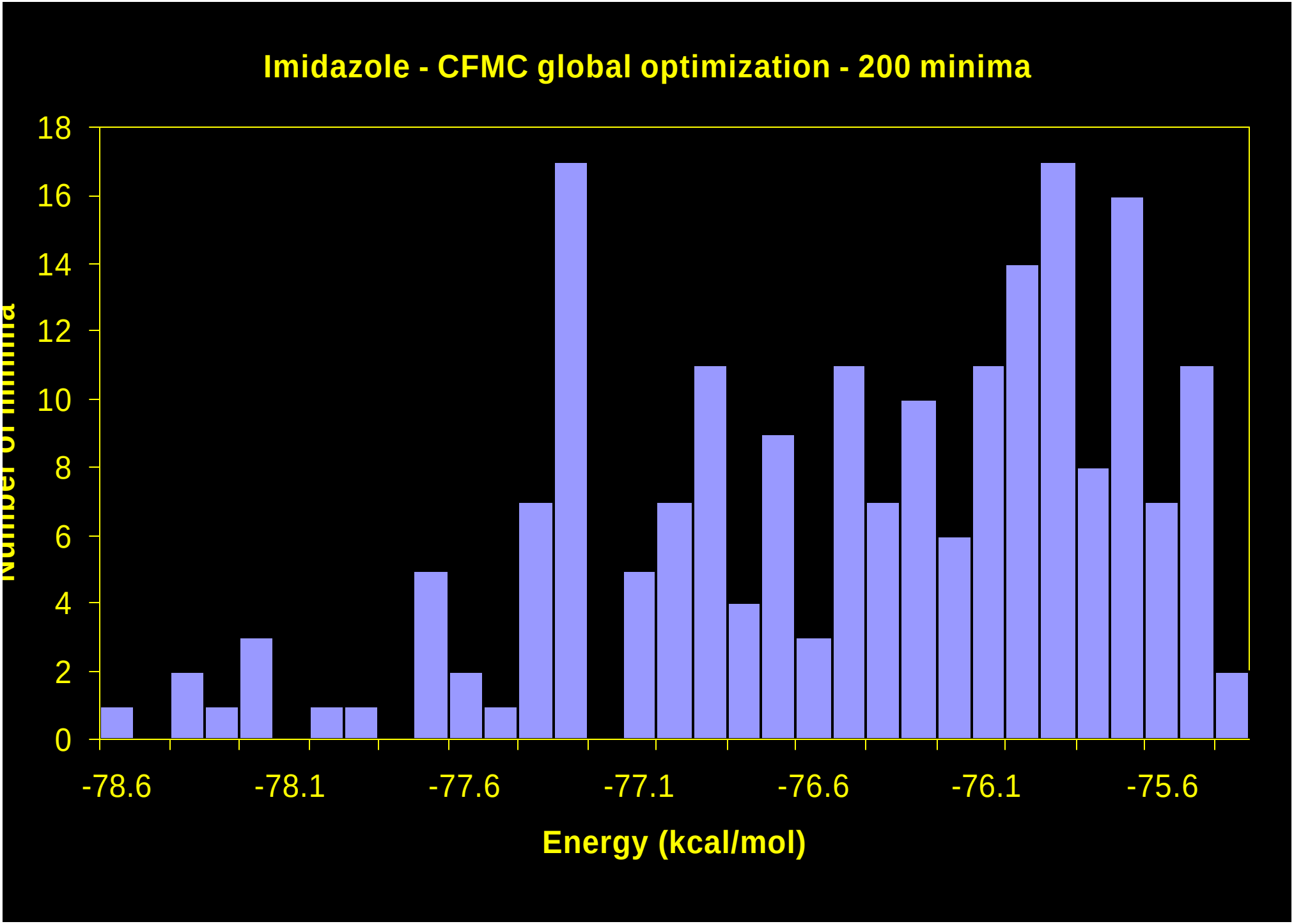
Two different potentials have been tested: AMBER and W99.

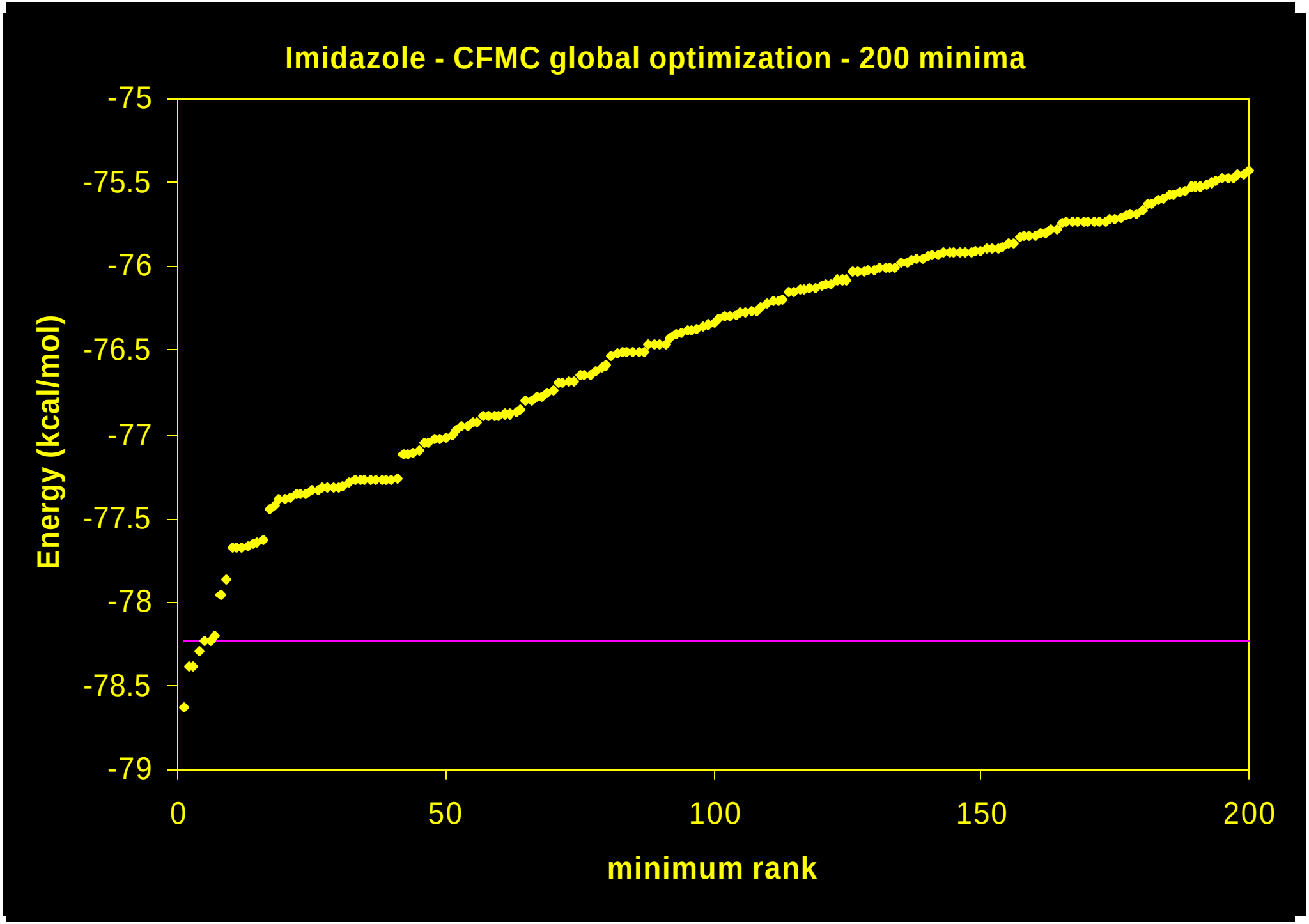
Set of nine rigid and flexible organic molecules was used:



None of the above potentials was perfect, however W99 force field performed slightly better, i.e., it is closer to satisfy criteria (a)-(c) discussed earlier.

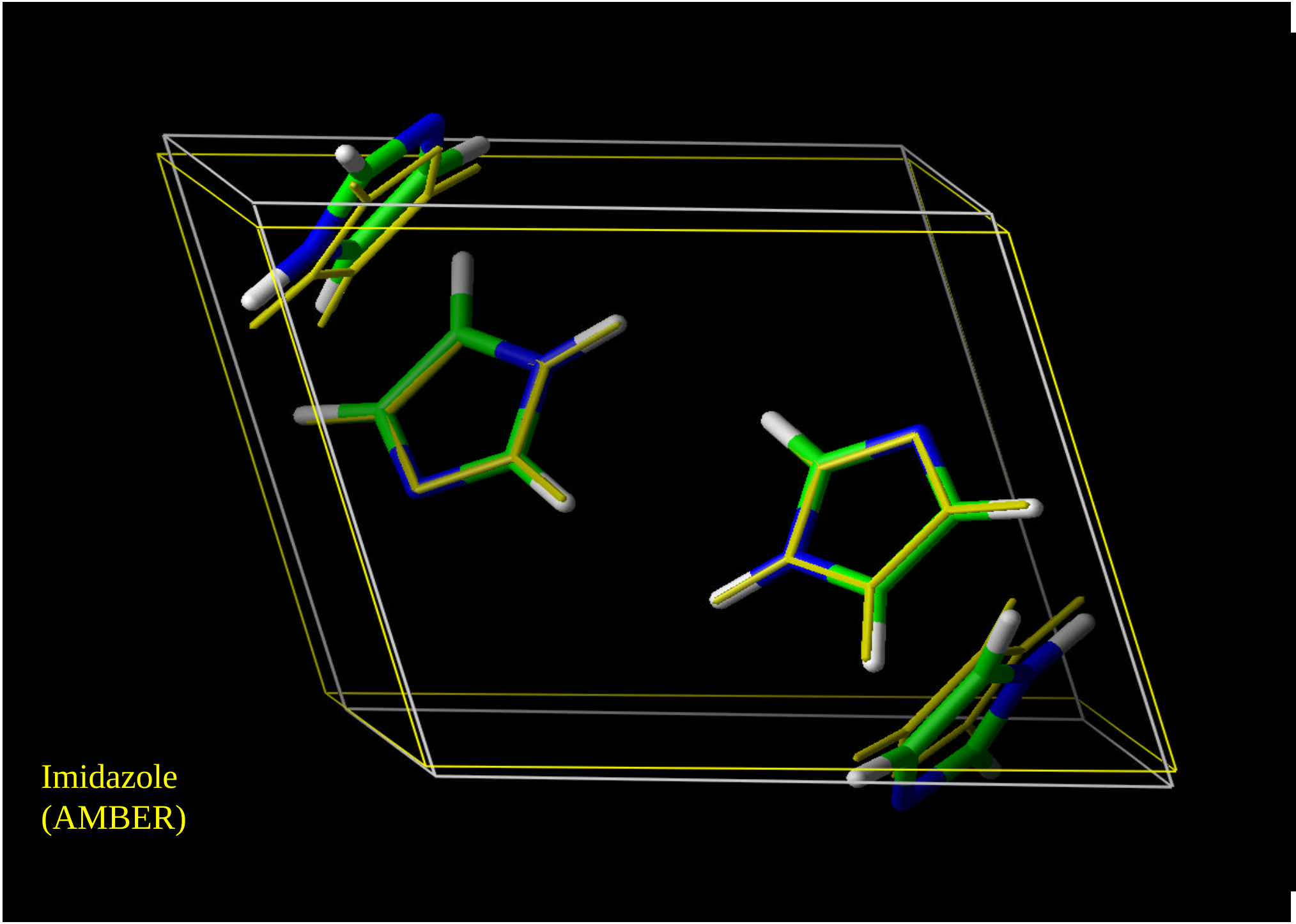

Imidazole - CFMC global optimization
-76
-76.5
-77
-77.5
-78
-78.5
Energy (kcal/mol)
40
1040
2040
3040
4040
5040
Number of minimizations
Lowest-energy family
Family 2
Family 3
Family 4

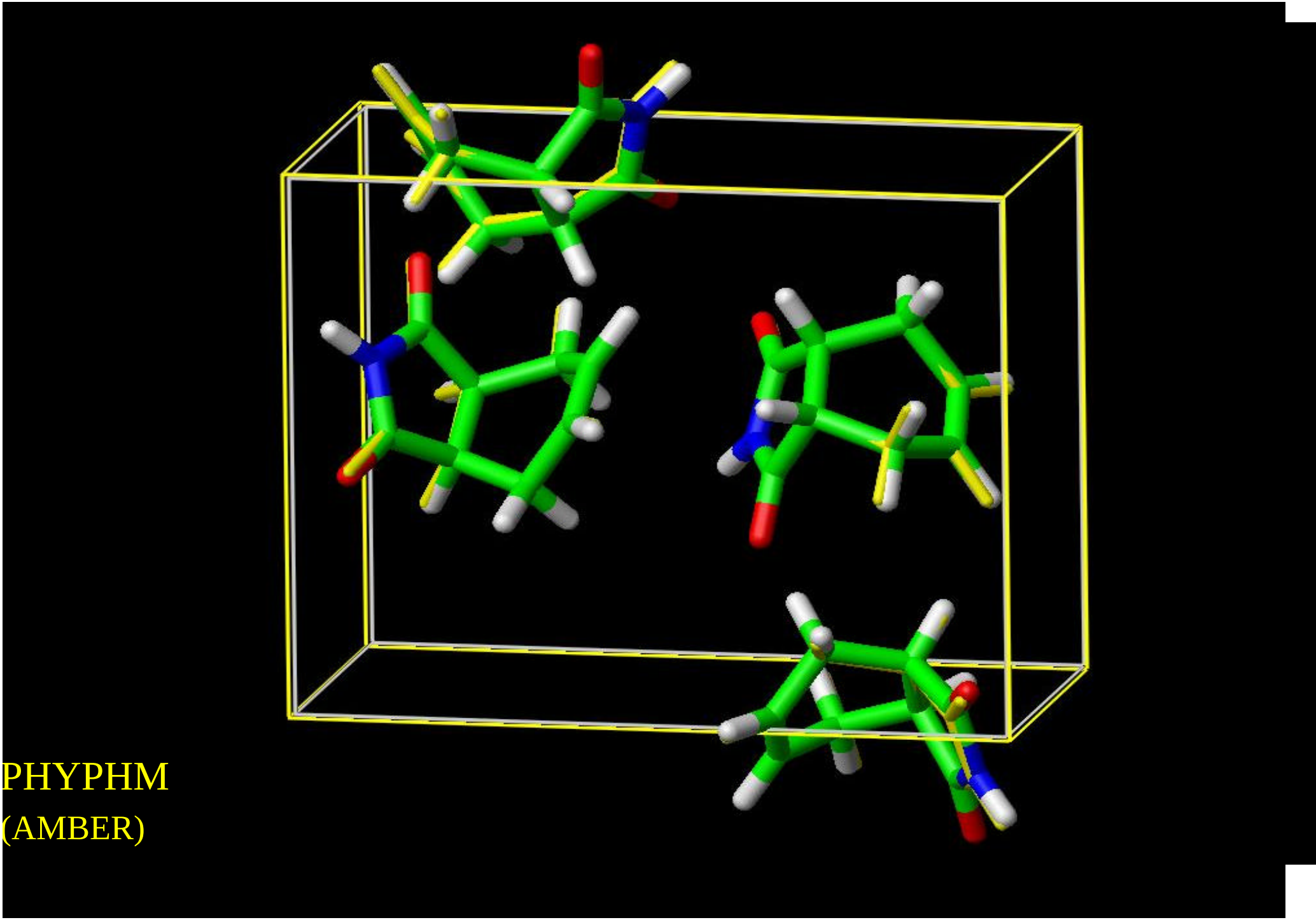

Imidazole - CFMC global optimization - 200 minima
Number of minima
0
2
4
6
8
10
12
14
16
18
-78.6
-78.1
-77.6
-77.1
-76.6
-76.1
-75.6
Energy (kcal/mol)


Imidazole - CFMC global optimization - 200 minima
Energy (kcal/mol)
-75
-75.5
-76
-76.5
-77
-77.5
-78
-78.5
-79
0
50
100
150
200
minimum rank

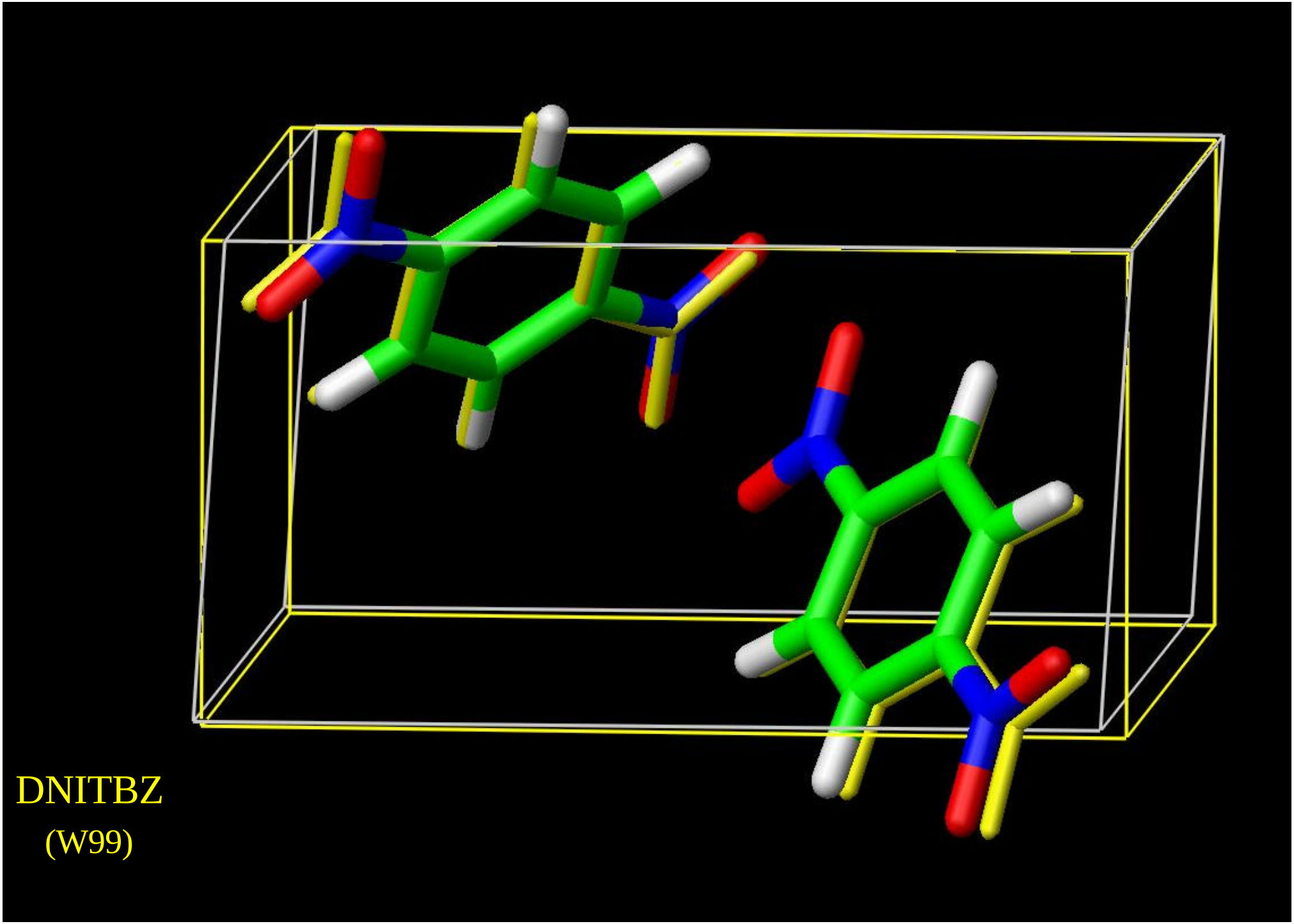
	W99		AMBER	
	rank	ΔE	rank	ΔE
BENZEN	2	0.19	4	0.14
PRMDIN	11	0.11	83*	0.55
IMAZOL06	24	0.79	2	0.02
PHYPHM	1	0.00	1	0.00
DNITBZ03	1	0.00	1	0.00
NITPOL03	19*	1.44	9*	0.87
FORAMO01	>200*	3.82	>200*	6.68
LIQVUC	2	0.04	5	0.52

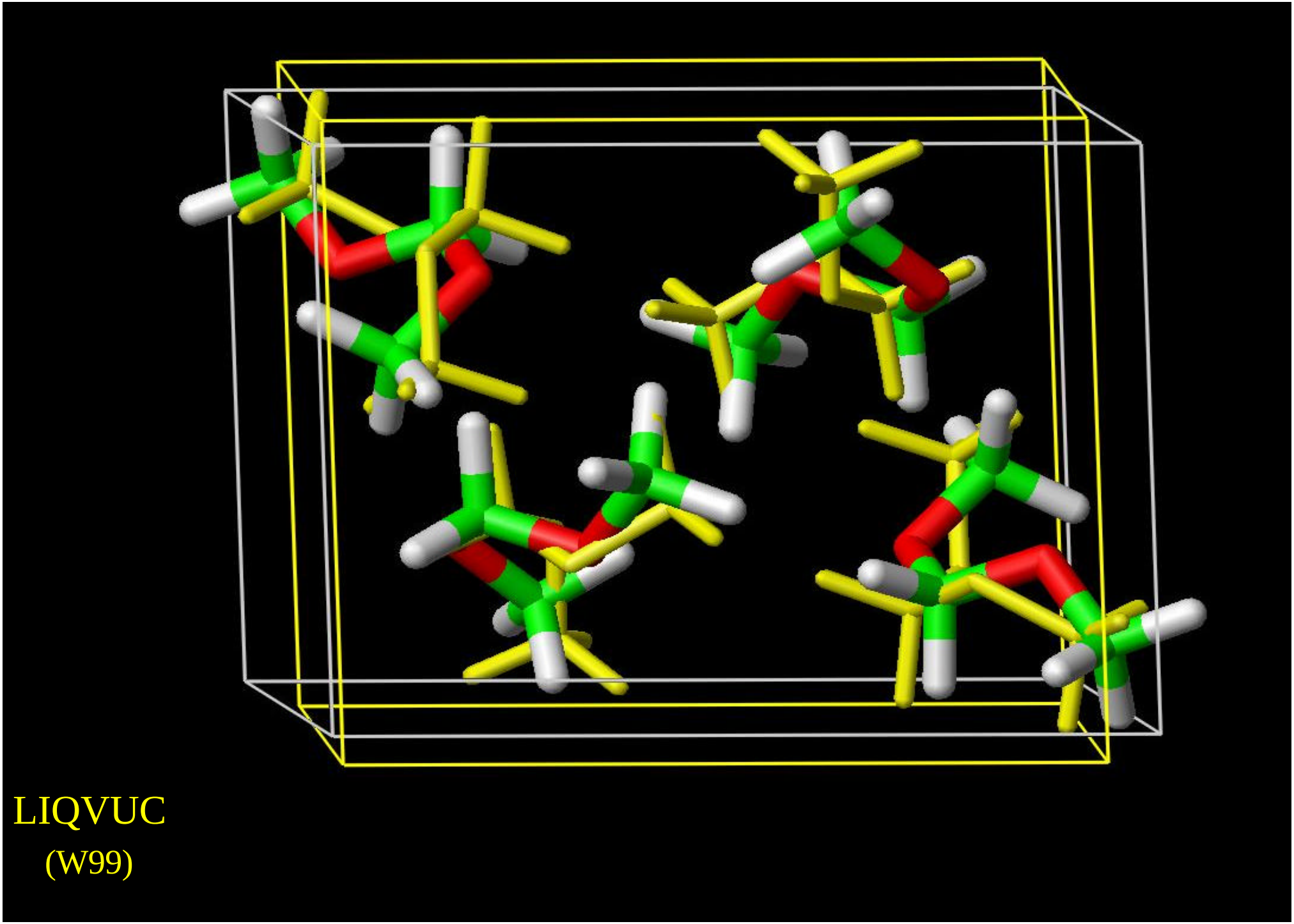
* - experimental structure was not found during the global search


Imidazole
(AMBER)



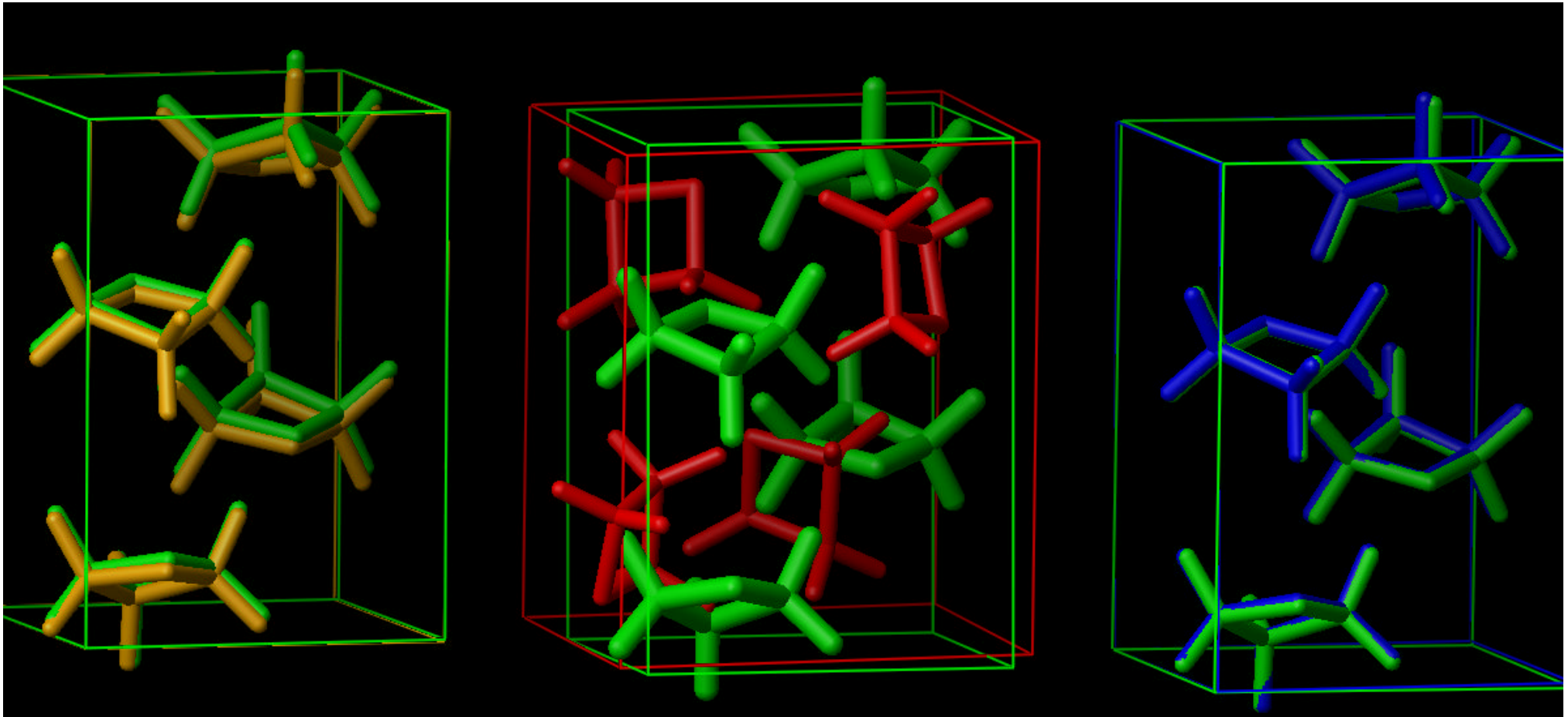
PHYPHM
(AMBER)


DNITBZ
(W99)



LIQVUC

(W99)



(a)

Minimized experimental structure (green, no 13 above global minimum)

(b)

Global minimum

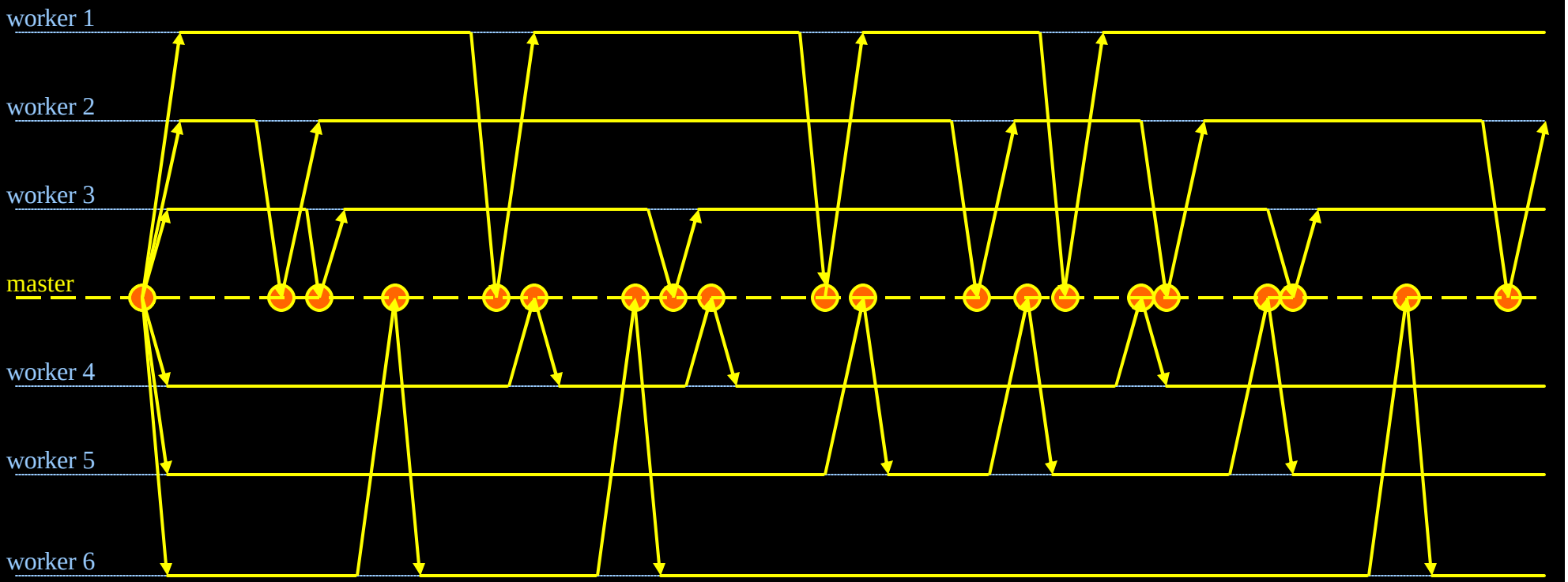
(c)

Global minimum

(optimized)

Propylene oxide – example of potential optimization

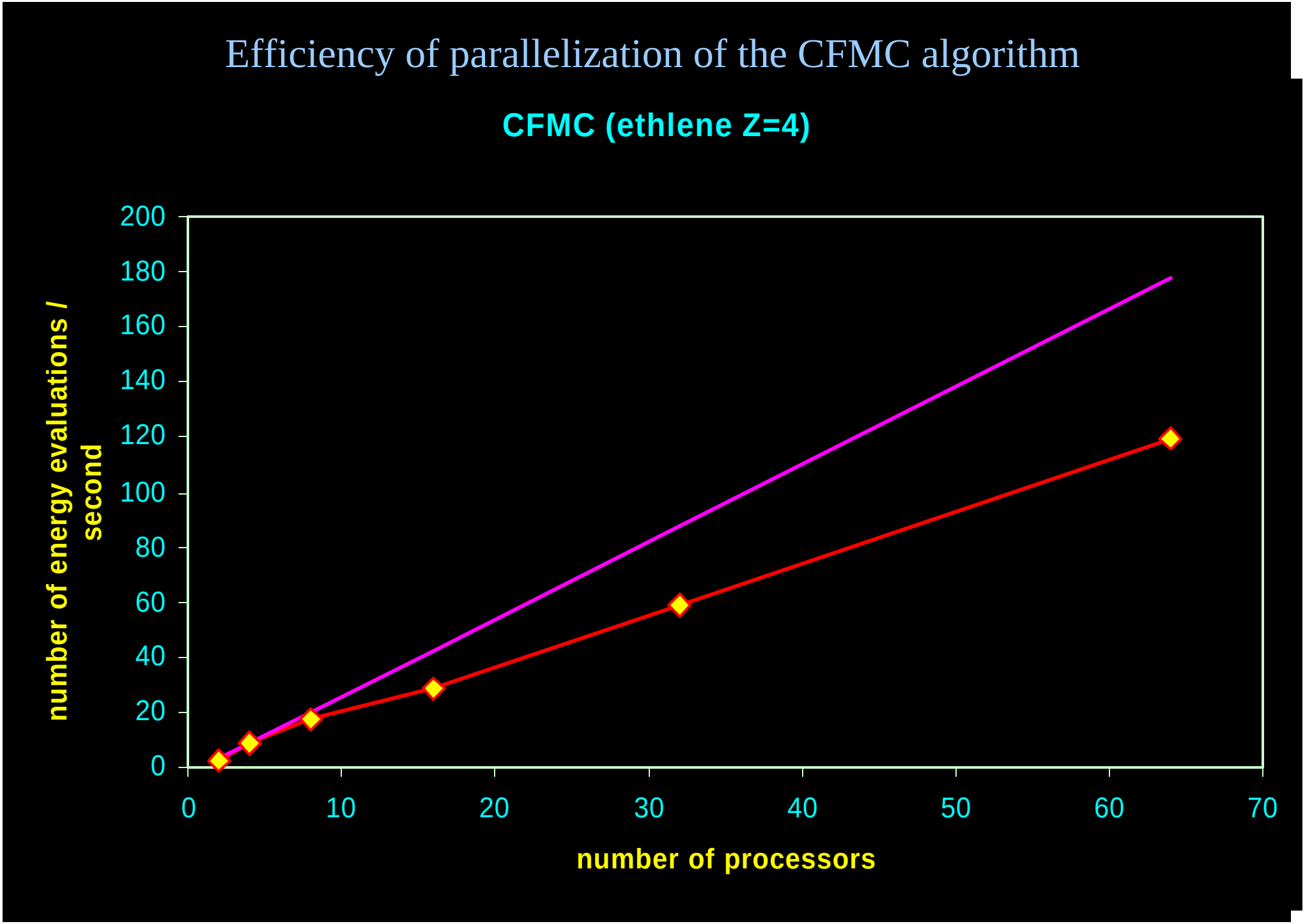
Massive parallelization



The CFMC algorithm has been parallelized at the *coarse-grain* level – all local minimizations are carried out in parallel by workers. The master is maintaining the database of structures and it is making all decisions within the algorithm.

MPI has been used for communication.

Efficiency of parallelization of the CFMC algorithm



Efficiency of parallelization of the CFMC algorithm