# An Evolutionary-inspired Probabilistic Search Algorithm to Structurally Characterize the Native State of a Novel Protein Sequence

*Shehu lab* www.cs.gmu.edu/~ashehu
{ssaleh2, bolson3, amarda}@gmu.edu

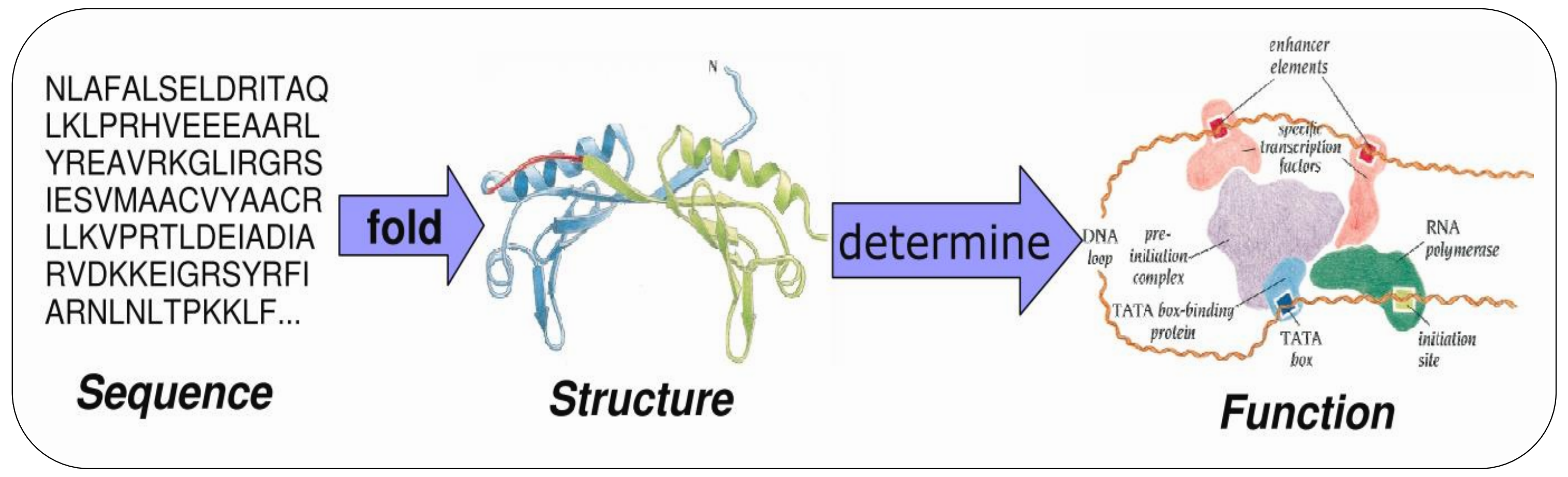
**Sameh Saleh[1], Brian Olson, and Amarda Shehu[1,2,3]**

[1]Department of Computer Science, [2]Department of Bioinf. & Comp Biol., [3]Department of Bioengineering

## Abstract

Obtaining a structural characterization of the biologically active (native) state of a protein is a long standing problem in computational biology. The high dimensionality of the conformational space and ruggedness of the associated energy surface are key challenges to algorithms in search of an ensemble of low-energy decoy conformations relevant for the native state. As the native structure does not often correspond to the global minimum energy, diversity is key. We present a memetic evolutionary algorithm to sample a diverse ensemble of conformations that represent low-energy local minima in the protein energy surface. Conformations in the algorithm are members of an evolving population. The molecular fragment replacement technique is employed to obtain children from parent conformations. A greedy search maps a child conformation to its nearest local minimum. Resulting minima and parent conformations are merged and truncated back to the initial population size based on potential energies. Results show that the additional minimization is key to obtaining a diverse ensemble of decoys, circumvent premature convergence to sub-optimal regions in the conformational space, and approach the native structure with lRMSDs comparable to state-of-the-art decoy sampling methods.

## Introduction



> *"[...] the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment."   Anfinsen, C. B. Science 181, 1973*

Experimental techniques that are devoted to resolving the native structure of a protein sequence cannot keep pace with the exponential explosion in the number of new protein sequences deposited to databases.

Determining the biologically-active structure of a protein sequence in-silico remains a central challenge in computational structural biology.

Exploring the protein conformational space in search of conformations that populate the protein native state is an NP-hard problem.

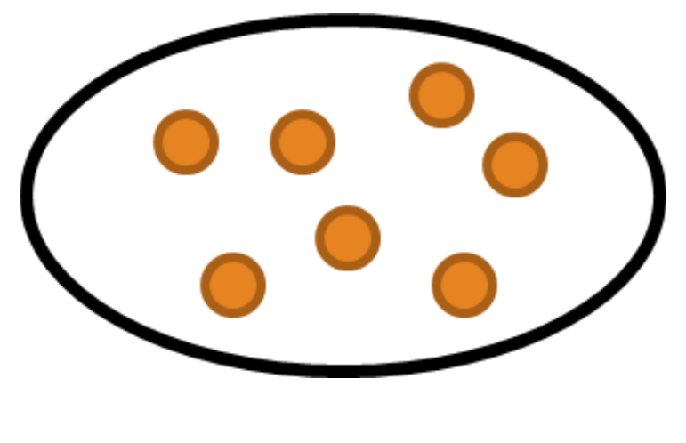
The protein conformational space is vast, continuous, and high-dimensional.

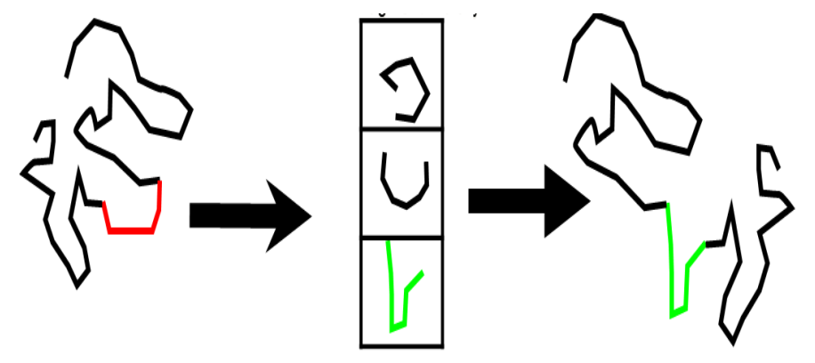
The protein energy surface is funnel-like, but rich in local minima of varying sizes.

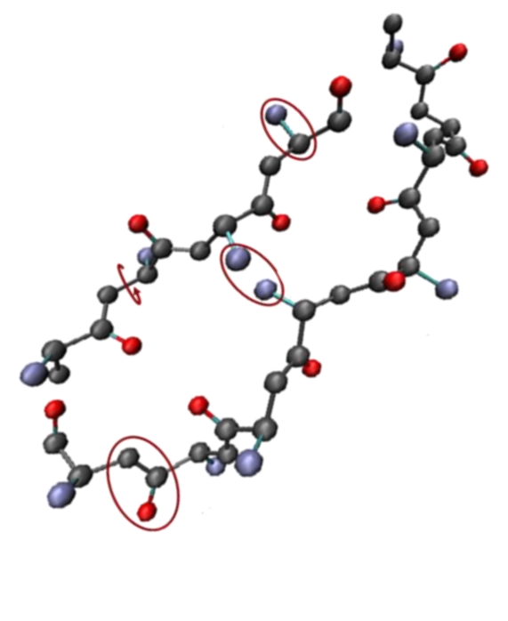
The protein native state is associated with the basin of the protein energy surface.

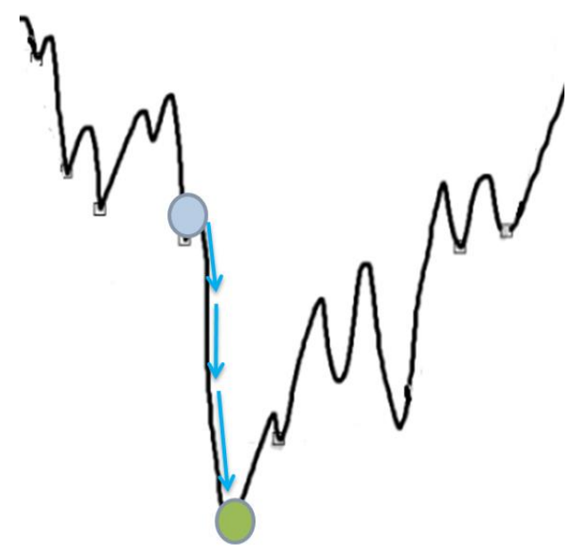
We propose to revisit evolutionary search strategies and combine them with the state-of-the-art fragment assembly in computational biophysics and the coarse-grained energy function in order to effective explore the protein conformational space and obtain lowest-energy conformations associated with the native state.

## Contributions


Evolving Population


Fragment Replacement

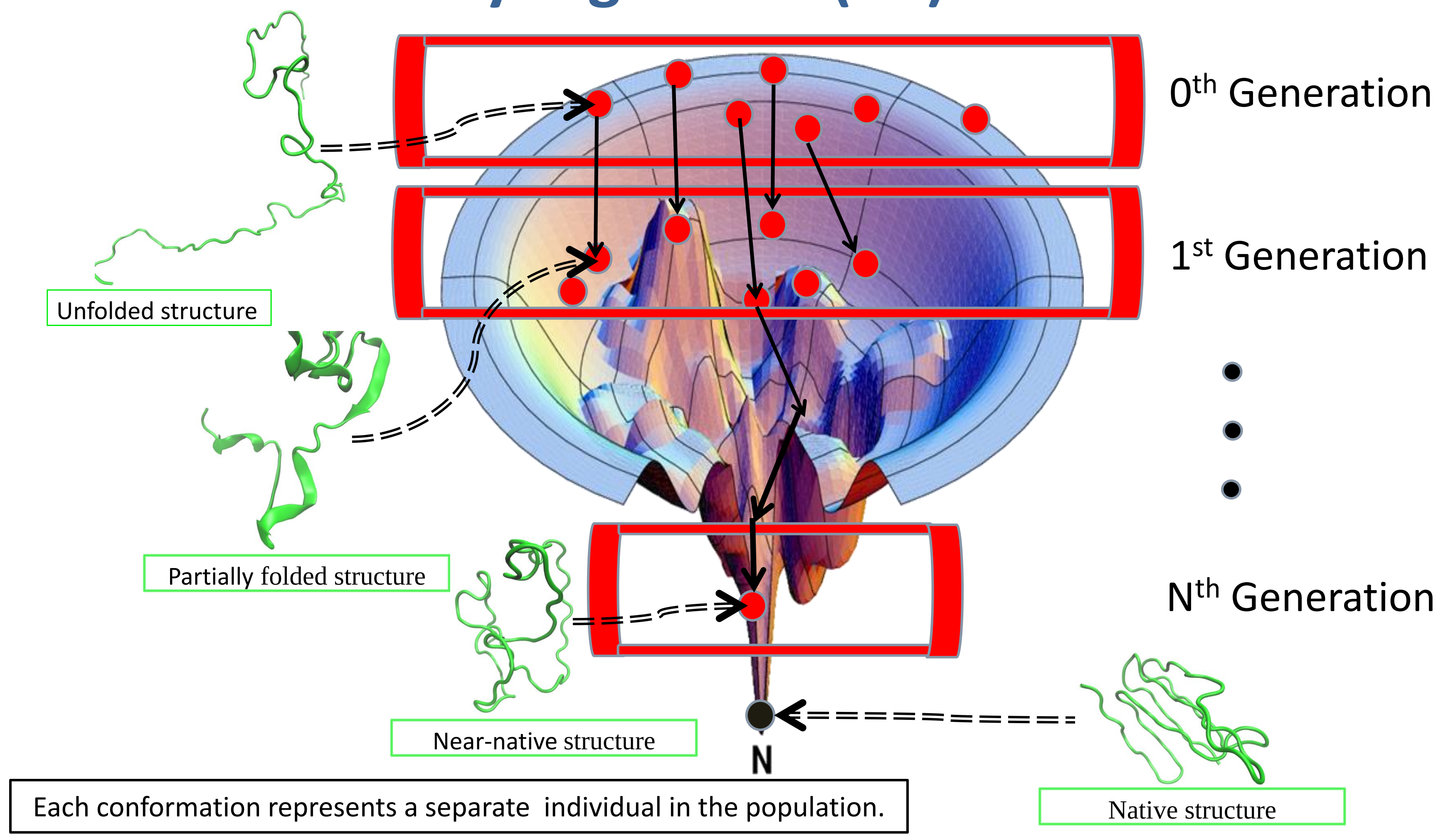

Coarse Graining


Local Greedy Search

## Acknowledgments

This work was partially supported by the Undergraduate Student Research Program to S.S., an NSF CCF Grant No. 1016995 to A.S., and an NSF Career Award Grant No. 1144106 to A.S.

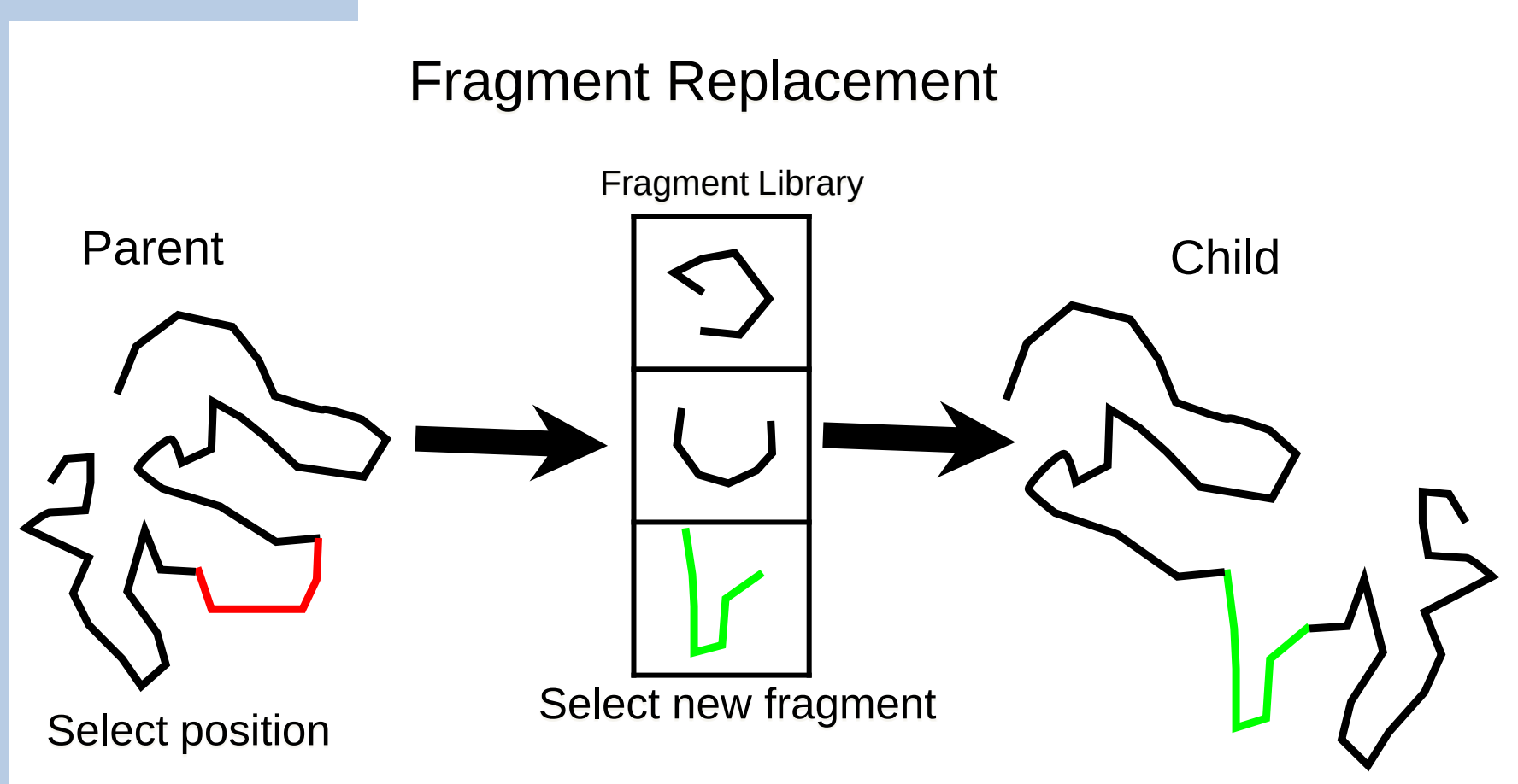
## The Evolutionary Algorithm (EA)



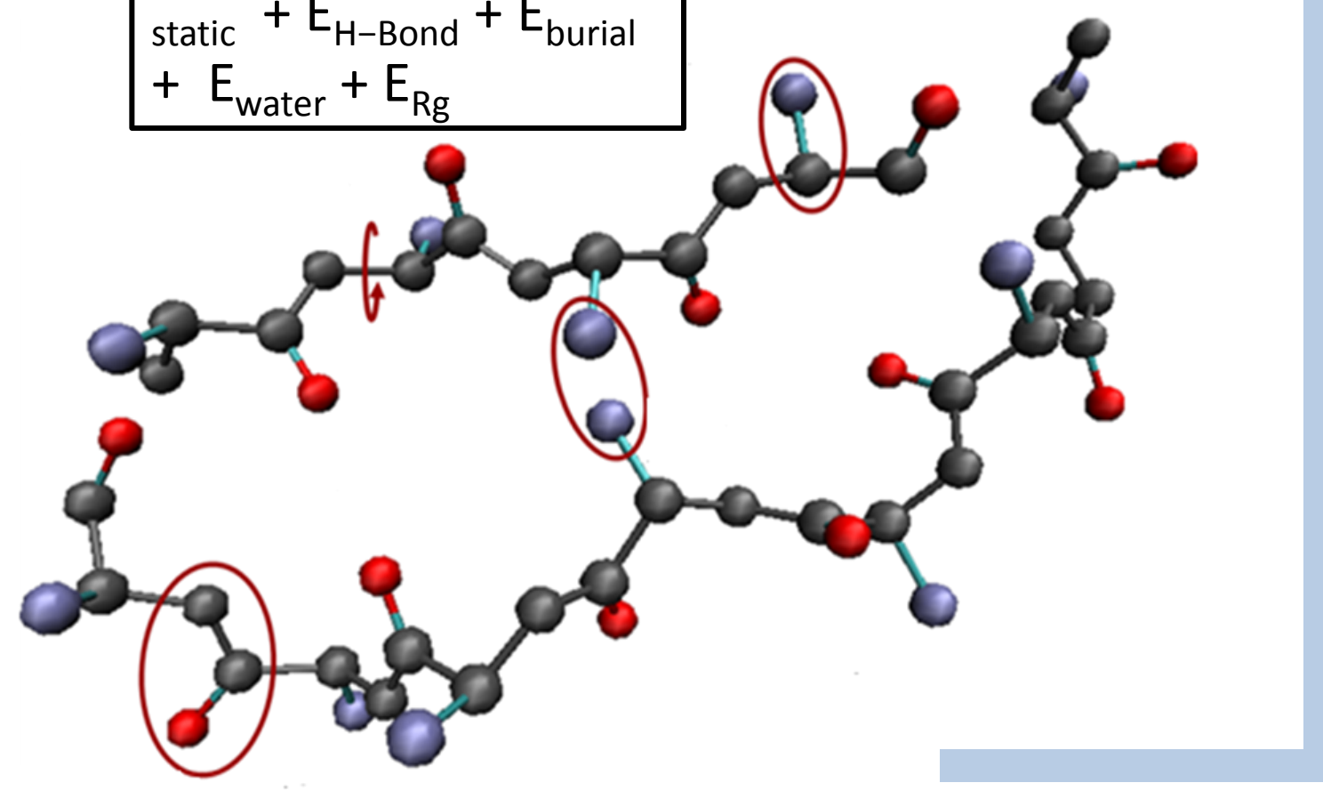
Each conformation represents a separate individual in the population.

From the *N* members of the population, *M* conformations are chosen to be copied and act as the children of the population. *N = 1000* was used. For *M*, *4000* was used for the EA. Such a selection was made using fitness-proportional selection. The stochastic implementation precludes that the conformations with the higher fitness are more likely to have children.

These children copies of the parent are modified through asexual reproduction to produce a child conformation. A fragment configuration in the parent is replaced with a configuration selected from a pre-built fragment configuration library to produce the child. The process is facilitated by fragment-based assembly, a state-of-the-art technique in computational biophysics that effectively obtains realistic conformations.



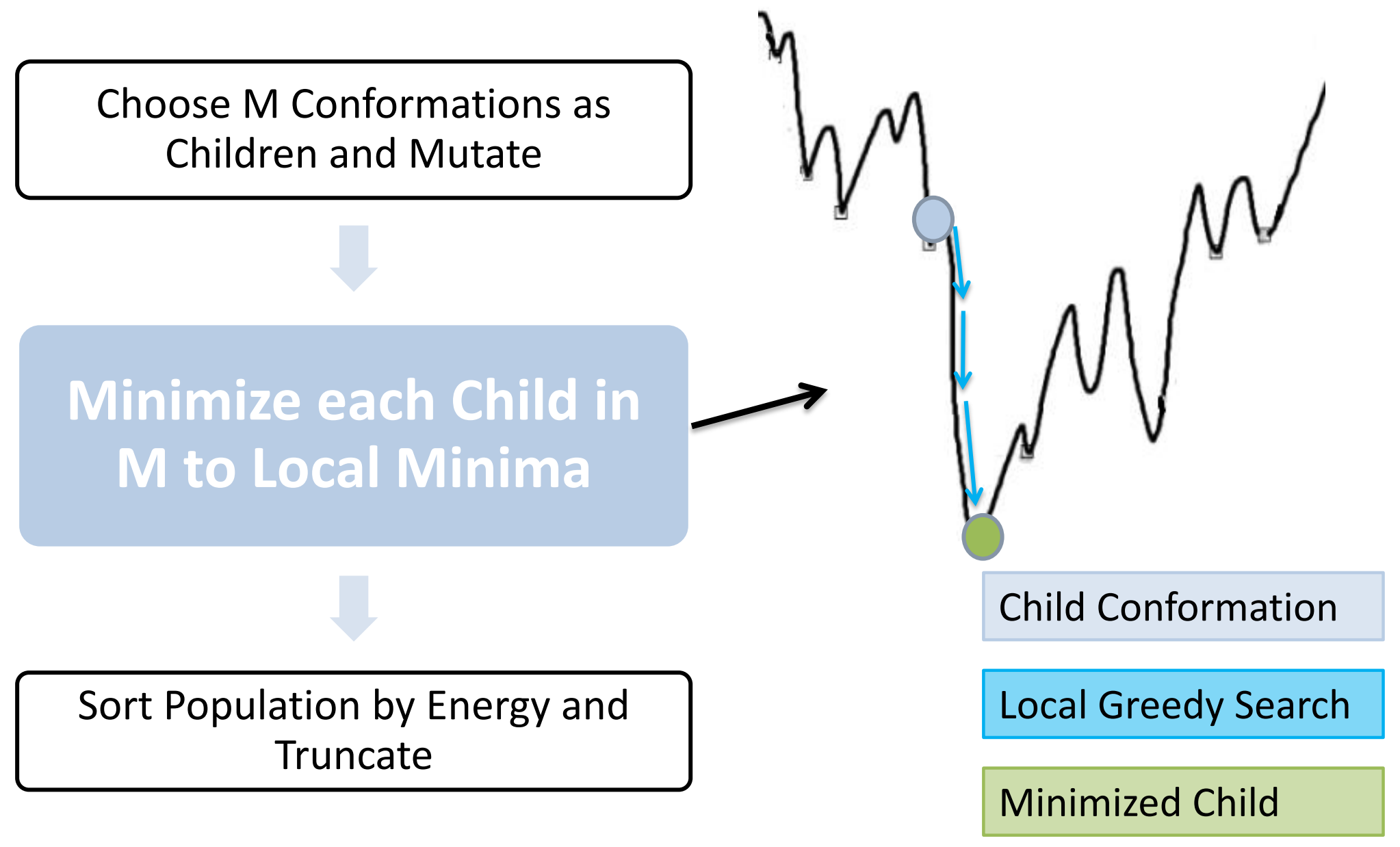
**EnergyAMW** = $E_{\text{electriostatic}} + E_{\text{H-Bond}} + E_{\text{burial}} + E_{\text{water}} + E_{\text{Rg}}$



The fitness of each individual is measured through a state-of-the-art energy function. The children are integrated into the original parent population, which is then sorted in descending order based on the energy fitness function Such a population is truncated to the original population size of *n* individuals and the process is repeated for each generation.

## The Memetic Evolutionary Algorithm (MEA)

The MEA employs an additional minimization step, implemented as a local greedy search, to map a child conformation sampled by the basic EA framework to a nearby local minimum. The corresponding conformation representing the minimum replaces the initial child conformation. Some of these minimized children will be members of the evolved population. In the next iteration, some may be selected to be parents where a mutation to get the new child will be equivalent to a jump out of the current minimum. This resetting is crucial to obtain new nearby minima in the energy surface, avoid convergence, and enhance conformation diversity.


Choose M Conformations as Children and Mutate

Minimize each Child in M to Local Minima

Sort Population by Energy and Truncate

Child Conformation

Local Greedy Search

Minimized Child

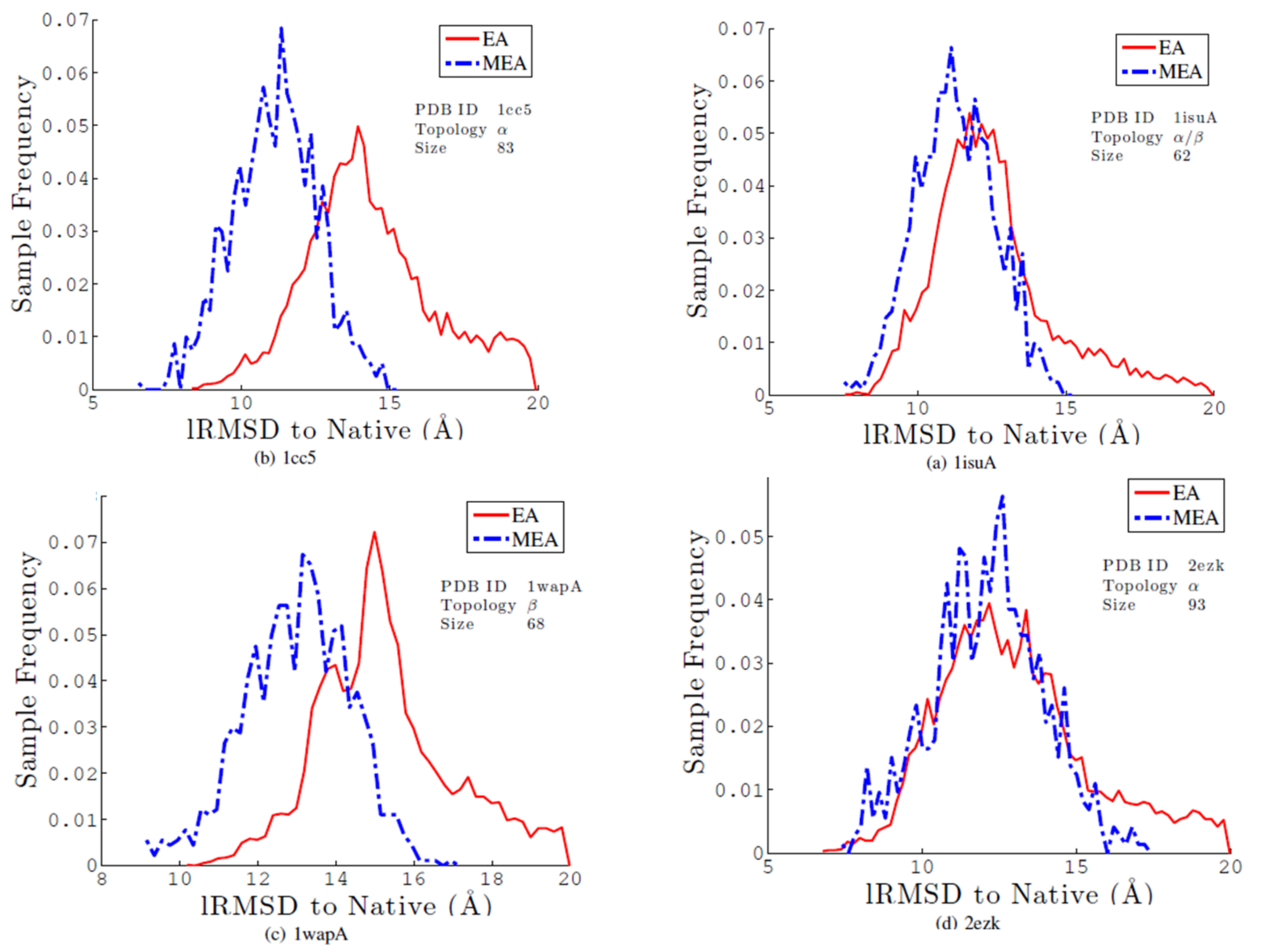
## Results

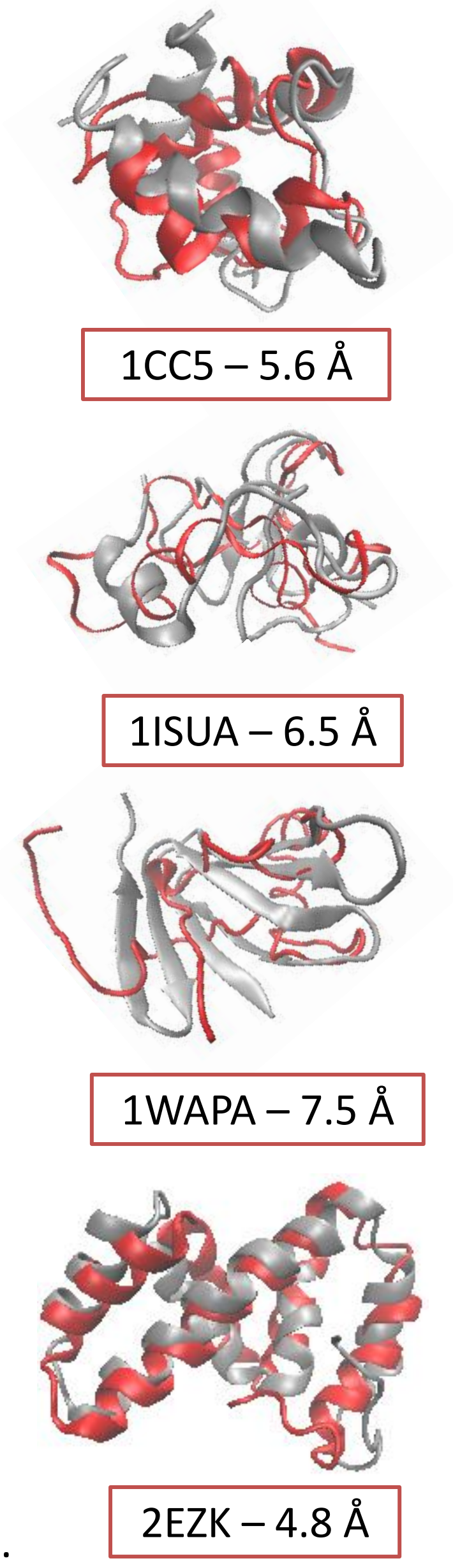
### Comparison to Other Groups

Each algorithm was run independently 3 times on eleven proteins. The proteins used represent a variety of structure types with different folds and different lengths. The efficacy of the algorithm is evaluated by calculating the root mean square deviation (RMSD) from the native structure. Columns 3 and 4 represent the simple EA and MEA, respectively. Column 5 represents, FeLTr, another local-minima sampling algorithm developed by our lab. Column 6 represents the published results by the Sosnick lab's ItFix algorithm. The MEA does better than the EA by at least 0.5Å in 10 of the 11 proteins. The MEA compares favorably with FeLTr in 6 of 11 proteins and in 4 of the remaining targets, does significantly better. Lastly, the MEA finds lower RMSD values for 5 of 11 targets, while ItFix finds lower RMSD values for 3 proteins; the methods are comparable for the remaining 3 proteins.

| | | | Avg(Min) lRMSD (Å) | | | |
|---|---|---|---|---|---|---|
| **PDB** | **Size** | **Fold** | **EA** | **MEA** | **FeLTr** | **ItFix** |
| **1dtdB** | 61 | α/β | 8.1(7.3) | 6.9(6.7) | 7.7(7.6) | 6.5 |
| **1isuA** | 62 | α/β | 7.5(7.1) | 6.5(6.1) | 6.8(6.7) | 6.5 |
| **1c8cA** | 64 | α/β | 8.5(8.5) | 7.2(6.8) | 6.5(6.0) | 3.7 |
| **1sap** | 66 | α/β | 8.0(7.2) | 6.7(6.2) | 7.1(6.5) | 4.6 |
| **1hz6A** | 67 | α/β | 6.7(5.5) | 6.2(6.0) | 6.6(6.6) | 3.8 |
| **1wapA** | 68 | β | 9.3(8.8) | 7.5(6.8) | 7.8(7.3) | 8 |
| **1fwp** | 69 | α/β | 7.6(7.1) | 6.8(6.6) | 6.8(6.4) | 8.1 |
| **1ail** | 70 | α | 4.8(4.0) | 3.5(3.4) | 4.7(4.5) | 5.4 |
| **1aoy** | 78 | α/β | 7.0(6.2) | 5.3(5.1) | 5.1(4.6) | 5.7 |
| **1cc5** | 83 | α | 7.1(6.4) | 5.7(5.5) | 6.4(6.2) | 6.5 |
| **2ezk** | 93 | α | 6.1(5.0) | 4.9(4.6) | 6.4(6.0) | 5.5 |

### EA vs. MEA RMSD Distribution





The frequencies of the lRMSDs to the experimentally determined native structure are given for each conformation sampled during all three runs for four representative target proteins. The solid red line represents the simple EA and the dashed blue line represents the EA with minimization. The conformation of the lowest RMSD using the MEA for each protein is visualized in red and is superimposed on the native conformation which is displayed in silver.

## Conclusions

The results show that the minimization step added in the MEA allows the algorithm to more effectively sample near-native conformations than the basic EA. The basic EA is nonetheless effective at optimizing the AMW energy function and reaches much lower-energy conformations than MEA since it is highly exploitative and converges rapidly to a few particular basins in the energy surface. The exploitation in MEA on the other hand, is limited by the greedy local search employed for minimization. The MEA quickly reaches a low-energy floor, but then a fragment replacement allows it to jump out of that local minima to a higher energy. Therefore, the MEA avoids convergence and explores a breadth of conformations around the low-energy floor. Hence, the population maintains diversity better in the MEA than the EA.

Comparison of MEA to other conformational search methods shows that the addition of the minimization step along with domain-specific techniques from the computational structural biology community make evolutionary search strategy comparable to other state-of-the-art decoy sampling methods for ab-initio protein structure prediction. Future work will investigate more advanced evolutionary search strategies that encourage greater diversity.

Implementation details: The presented results were obtained by running the two algorithms on a 2.66GHz Opteron processor with 8GB of memory for 30 to 50 hours depending on protein length. The method was implemented in C/C++.