

# Multiobjective Genetic Algorithms for Multiscaling Excited-State Direct Dynamics in Photochemistry

Kumara Sastry<sup>1</sup>, D.D. Johnson<sup>2</sup>, A. L. Thompson<sup>3</sup>,  
D. E. Goldberg<sup>1</sup>, T. J. Martinez<sup>3</sup>, J. Leiding<sup>3</sup>, J. Owens<sup>3</sup>

<sup>1</sup>Illinois Genetic Algorithms Laboratory, Industrial and Enterprise  
Systems Engineering

<sup>2</sup>Materials Science and Engineering

<sup>3</sup>Chemistry and Beckman Institute

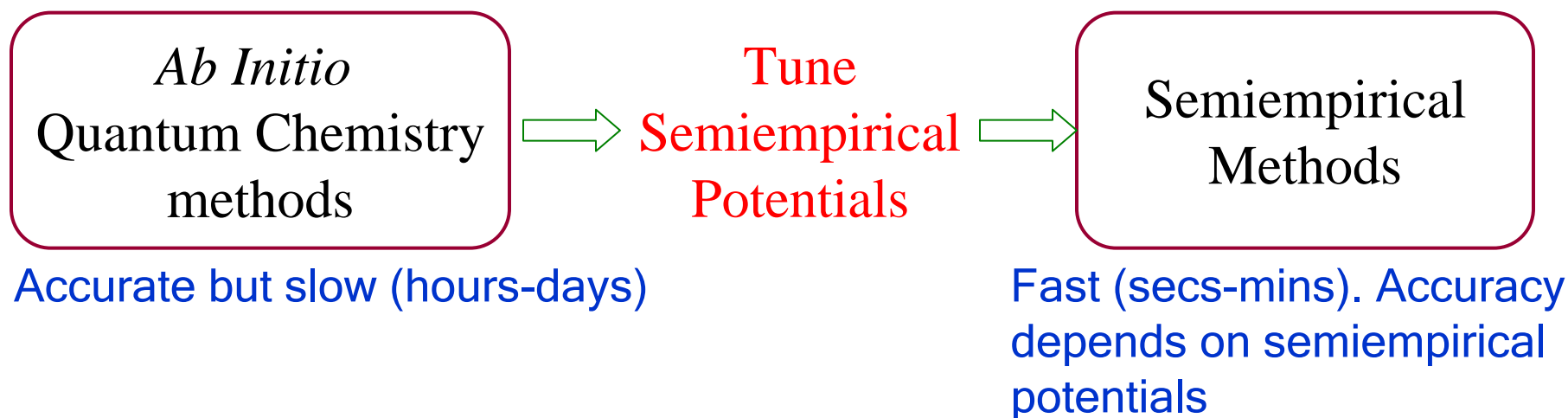
University of Illinois at Urbana-Champaign



Supported by AFOSR F49620-03-1-0129, NSF/DMR at MCC DMR-03-76550

# Multiscaling Photochemical Reaction Dynamics

- ❖ Multi-scale modeling is ubiquitous in science & engineering
  - ◆ Phenomena of interest are usually multiscale
  - ◆ Powerful modeling methodologies on single scale
- ❖ Multiscaling photochemical reaction dynamics

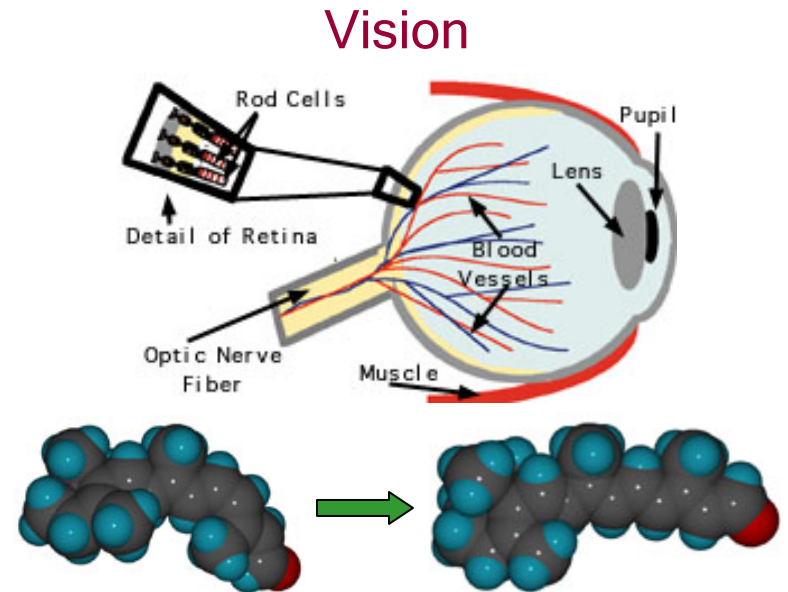
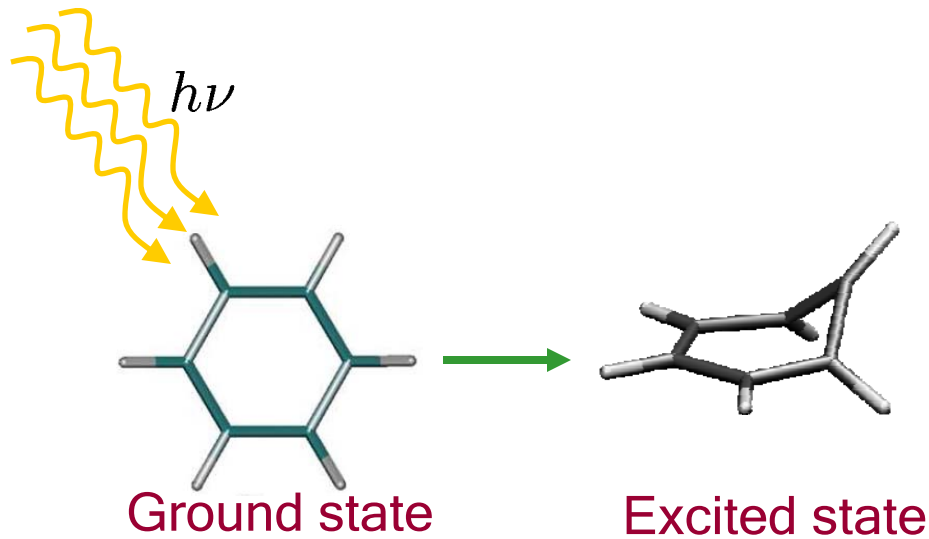


- ❖ Use multiobjective genetic algorithm for tuning semiempirical potentials for multiscaling reaction dynamics

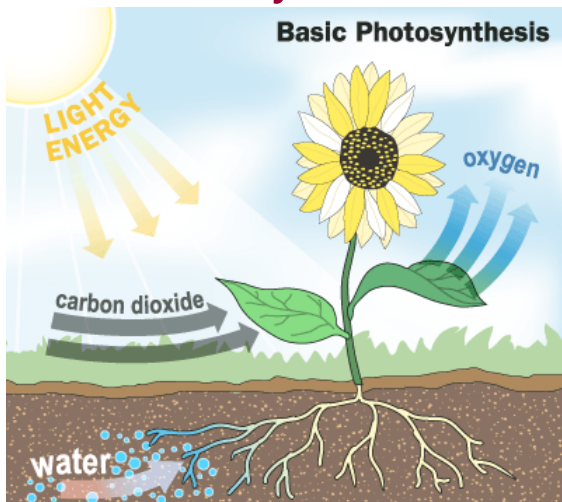
# Outline

- ❖ Introduction: Background and Purpose
- ❖ Method for multiscale reaction dynamics
  - ◆ Limitations of existing methods
- ❖ Problem formulation
- ❖ Overview of NSGA-II
- ❖ Results and Discussion
- ❖ Summary and Conclusions

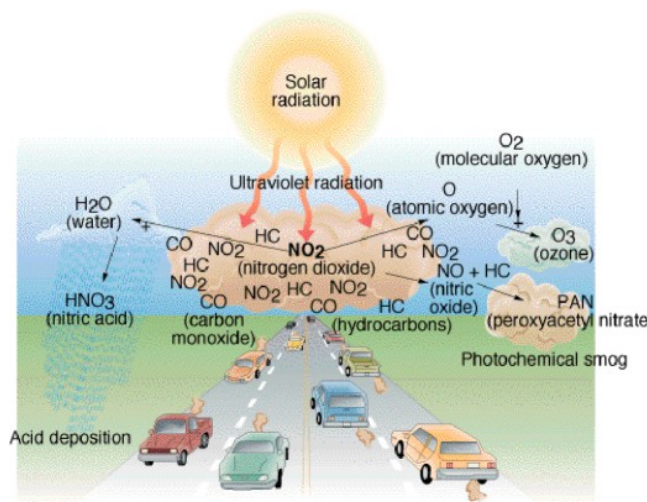
# Photochemical Reaction



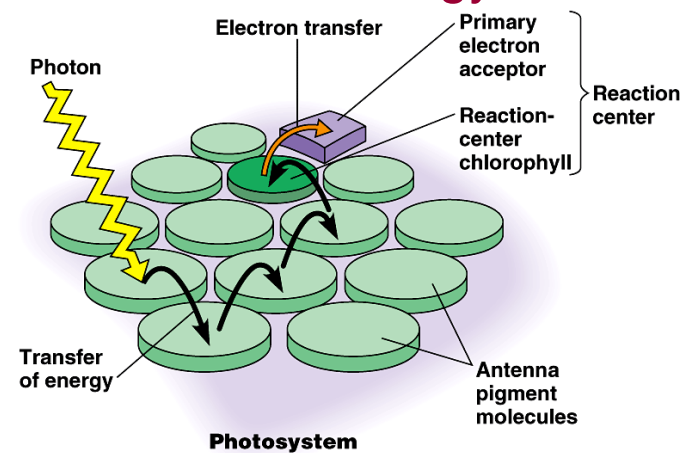
## Photosynthesis



## Pollution



## Solar energy



# Accurate Simulation of Reaction Dynamics Isn't Easy

## ❖ *Ab initio* quantum chemistry methods:

- ◆ *Ab initio* multiple spawning methods [Ben-Nun & Martinez, 2002]
- ◆ Solve nuclear and electronic Schrödinger's equations.
- ◆ **Accurate, but prohibitively expensive (hours-days)**

## ❖ Semiempirical methods:

- ◆ Solve Schrodinger's equations with expensive parts replaced with parameters.
- ◆ **Fast (secs-mins), accuracy depends on semiempirical potentials**
- ◆ Tuning semiempirical potentials is non-trivial
- ◆ Energy & shape of energy landscape matter
- ◆ Two objectives at the bare minimum
  - ★ Minimizing errors in energy and energy gradient

# Why Does This Matter?

- ❖ Multiscaling speeds all modeling of physical problems:
  - ◆ Solids, fluids, thermodynamics, kinetics, etc.,
  - ◆ Example: GP used for multi-timescaling Cu-Co alloy kinetics [Sastry, et al (2006), *Physical Review B*]
- ❖ Here we use MOGA to enable fast and accurate modeling
  - ◆ Retain *ab initio* accuracy, but exponentially faster
- ❖ Enabling technology: Science and Synthesis
  - ◆ Fast, accurate models permit larger quantity of scientific studies
  - ◆ Fast, accurate models permit synthesis via repeated analysis
- ❖ This study potentially enables:
  - ◆ Biophysical basis of vision
  - ◆ Biophysical basis of photosynthesis
  - ◆ Protein folding and drug design
  - ◆ Rapid design of functional materials (zeolites, LCDs, etc.,)

# Methodology: Limited *Ab Initio* and Experimental Results to Tune Semiempirical Parameters

- ❖ Perform *ab initio* computations for a *few* configurations
  - ◆ Both excited- and ground-state configurations
  - ◆ Augmented with experimental measurements
- ❖ Standard parameter sets don't yield accurate potential energy surfaces (PESs)
  - ◆ Example: AM1, PM3, MNDO, CNDO, INDO, etc.
  - ◆ Accurately describe ground-state properties
  - ◆ Yields wrong description of excited-state dynamics
- ❖ **Parameter sets need to be reoptimized**
  - ◆ Maintain accurate description of ground-state properties
  - ◆ Yield globally accurate PES and hence physical dynamics

# Current Reparameterization Methods Fall Short

- ❖ Staged single objective optimization
  - ◆ First minimize error in energies
  - ◆ Subsequently minimize weighted error in energy and gradient
- ❖ Reparameterization involves multiple objectives
  - ◆ Don't know the weights of different objectives
- ❖ Reparameterization is highly multimodal
  - ◆ Local search gets stuck in low-quality optima
- ❖ Current methods still fall short
  - ◆ Often doesn't yield globally accurate PES
  - ◆ Yield *uninterpretable* semiempirical potentials
  - ◆ Semiempirical potentials are not *transferable*
    - ★ Use parameters optimized for simple molecules in complex environments without complete reoptimization.

# Fitness: Errors in Energy and Energy Gradient

- ❖ Choose a few ground- and excited-state configurations
- ❖ Fitness #1: Error in energy and geometry
  - ◆ For each configuration, compute energy and geometry
    - ★ Via *ab initio* and semiempirical methods

$$f_1 = \sum_1^{n_c} |\Delta E_{ai} - \Delta E_{se}| + \text{Differences in geometry}$$

- ❖ Fitness #2: Error in energy gradient
  - ◆ For each configuration compute energy gradient
    - ★ Via *ab initio* and semiempirical methods

$$f_2 = \sum_1^{n_c} |\nabla E_{ai} - \nabla E_{se}|$$

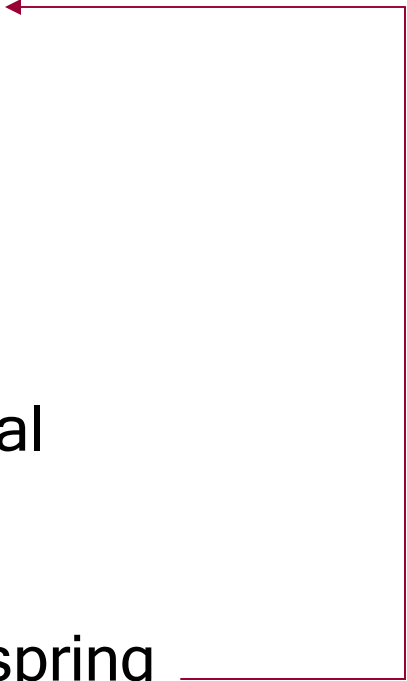
# Chromosome: Real-Valued Encoding of Semiempirical Parameters

- ❖ Consists of 11 semiempirical parameters for carbon
  - ◆ Most important parameters affecting excited-state PES
- ❖ Semiempirical parameters for hydrogen is not reoptimized
  - ◆ Set to their PM3 values
- ❖ Core-core repulsion parameters are not optimized
  - ◆ Set to their PM3 values
- ❖ Real-valued encoding of chromosomes
- ❖ Variable ranges: 20-50% around PM3 values
  - ◆ Retain reasonable representation of ground state PES

# Multiobjective GA: NSGAI with Binary Tournament, SBX, and Polynomial Mutation

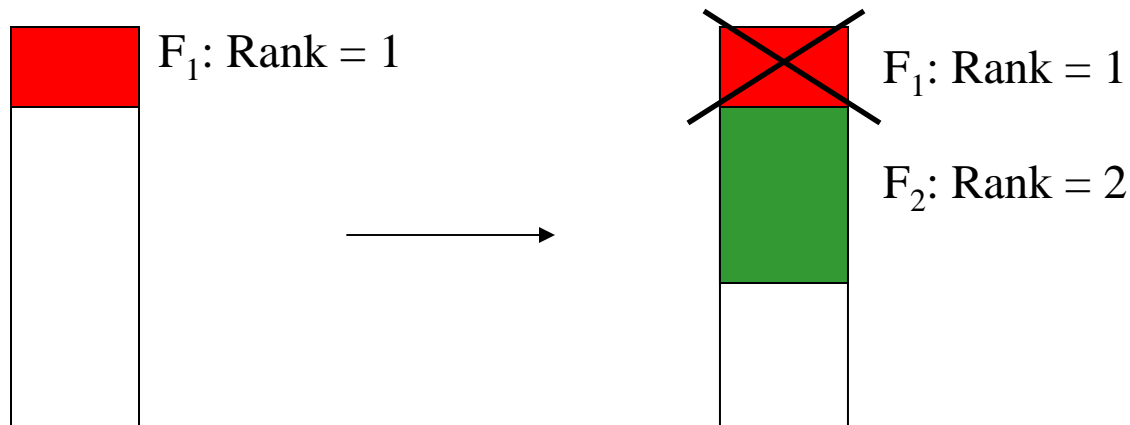
- ❖ Non-dominated sorting GA-II (NSGA-II) [Deb *et al*, 2000]
- ❖ Binary tournament selection ( $s = 2$ ) [Goldberg, Deb, & Korb, 1989]
- ❖ Simulated binary crossover (SBX) ( $\eta_c = 5$ ,  $p_c = 0.9$ ) [Deb & Agarwal, 1995; Deb & Kumar, 1995]
- ❖ Polynomial Mutation ( $\eta_m = 10$ ,  $p_m = 0.1$ ) [Deb *et al*, 2000]
- ❖ Results reported are best over 5, 10, and 30 NSGA-II runs

# Overview of NSGA-II

- ❖ Initialize Population
  - ❖ Evaluate fitness of individuals
  - ❖ Selection: “Survival of the non-dominated”
    - ◆ Non-dominated sorting
    - ◆ Individual comparison
  - ❖ Recombination: Combine traits of parents
  - ❖ Mutation: Random walk around an individual
  - ❖ Evaluate offspring solutions
  - ❖ Replacement: Best among parents and offspring
- 

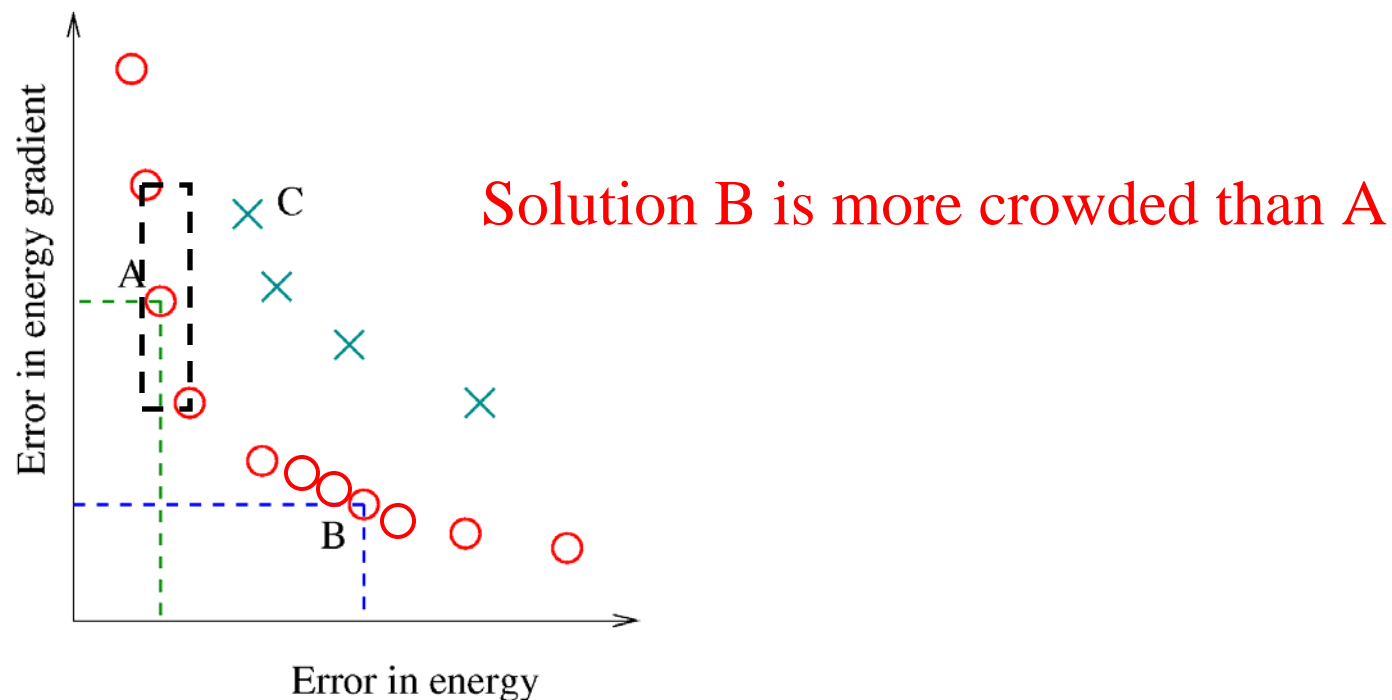
# Non-dominated Sorting Procedure

- ❖ Identify the best non-dominated set
  - ◆ A set of solutions that are not dominated by any individual in the population
- ❖ Discard them from the population temporarily
- ❖ Identify the next best non-dominated set
- ❖ Continue till all solutions are classified



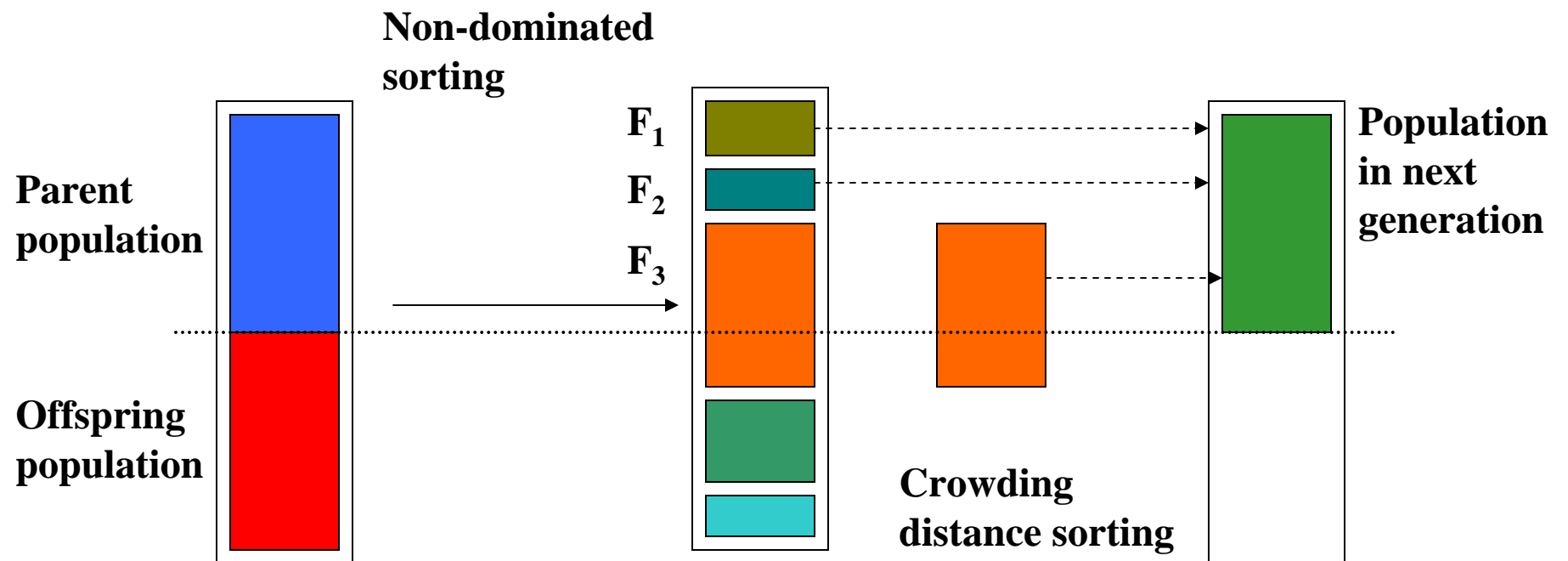
# Crowding In NSGA-II for Niche Preservation

- ❖ Crowding (niche-preservation) in objective space
- ❖ Each solution is assigned a crowding distance
  - ◆ Crowding distance = front density in the neighborhood
  - ◆ Distance of each solution from its nearest neighbors



# Elitist Replacement in NSGA-II

- ❖ Combine parent and offspring population
- ❖ **Select better ranking individuals and use crowding distance to break the tie**



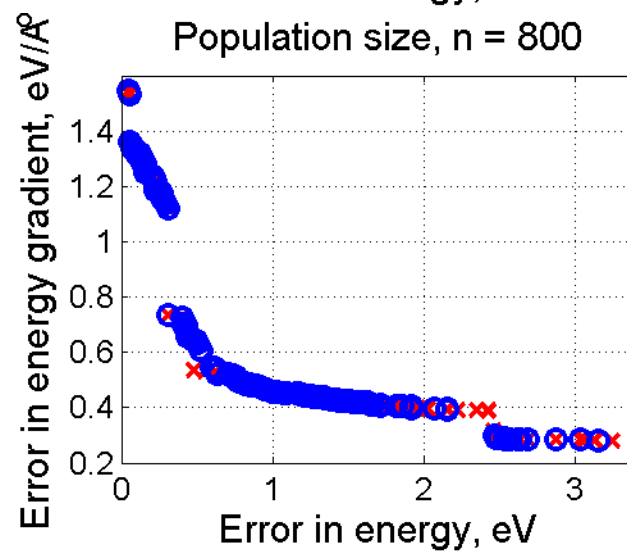
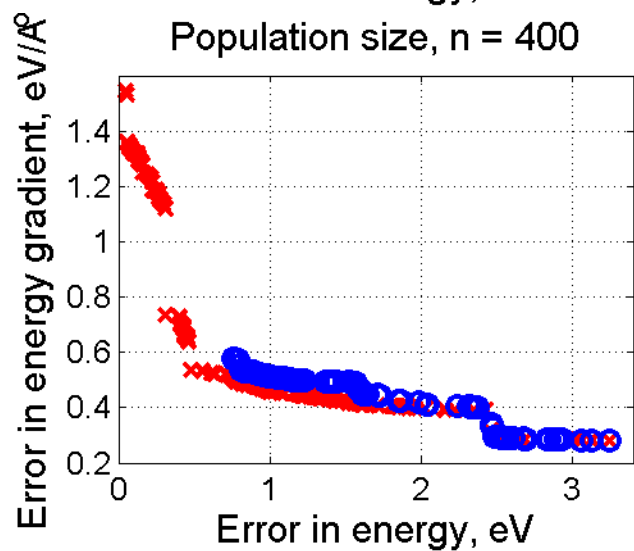
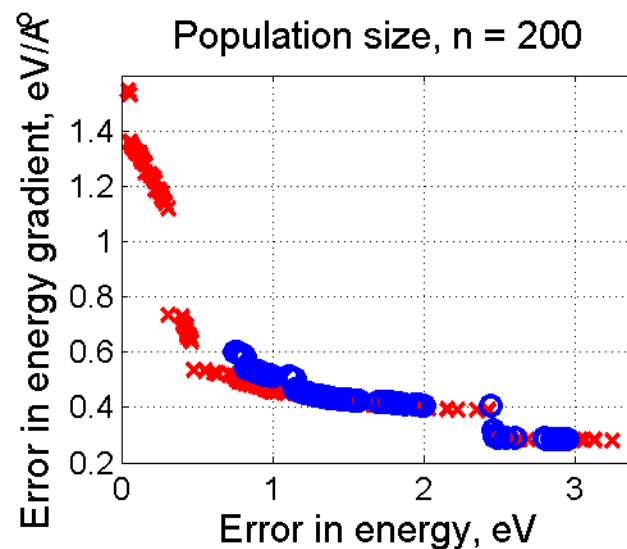
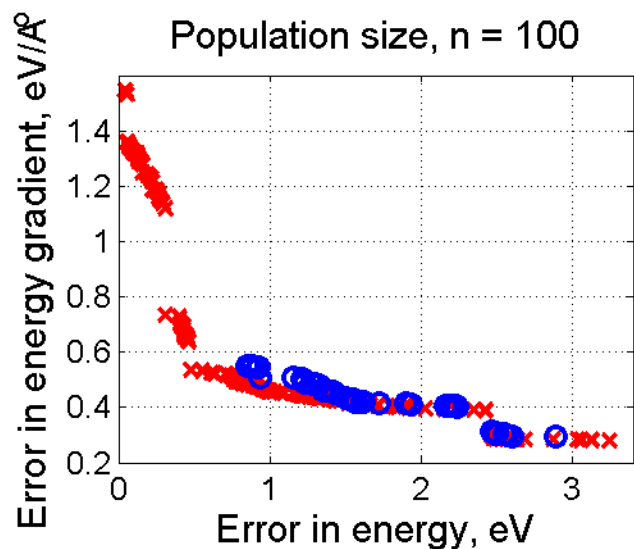
# Test Molecules: Ethylene and Benzene

- ❖ Tune semiempirical potentials for ethylene and benzene
- ❖ Fundamental building blocks of organic molecules
- ❖ Play important role in photochemistry of aromatic systems
- ❖ Extensively studied both theoretically and experimentally
- ❖ Simple and thus amenable to rapid analysis
  - ◆ Verification using *ab initio* results
  - ◆ Exhaustive dynamics simulations
  - ◆ Transferability to more complex molecules
  - ◆ Expect less complex results thus easy interpretability
- ❖ Complex enough and thus can test for validity of semiempirical methods on untested, yet critical configurations

# Population Size of 800 Yields Good Solutions

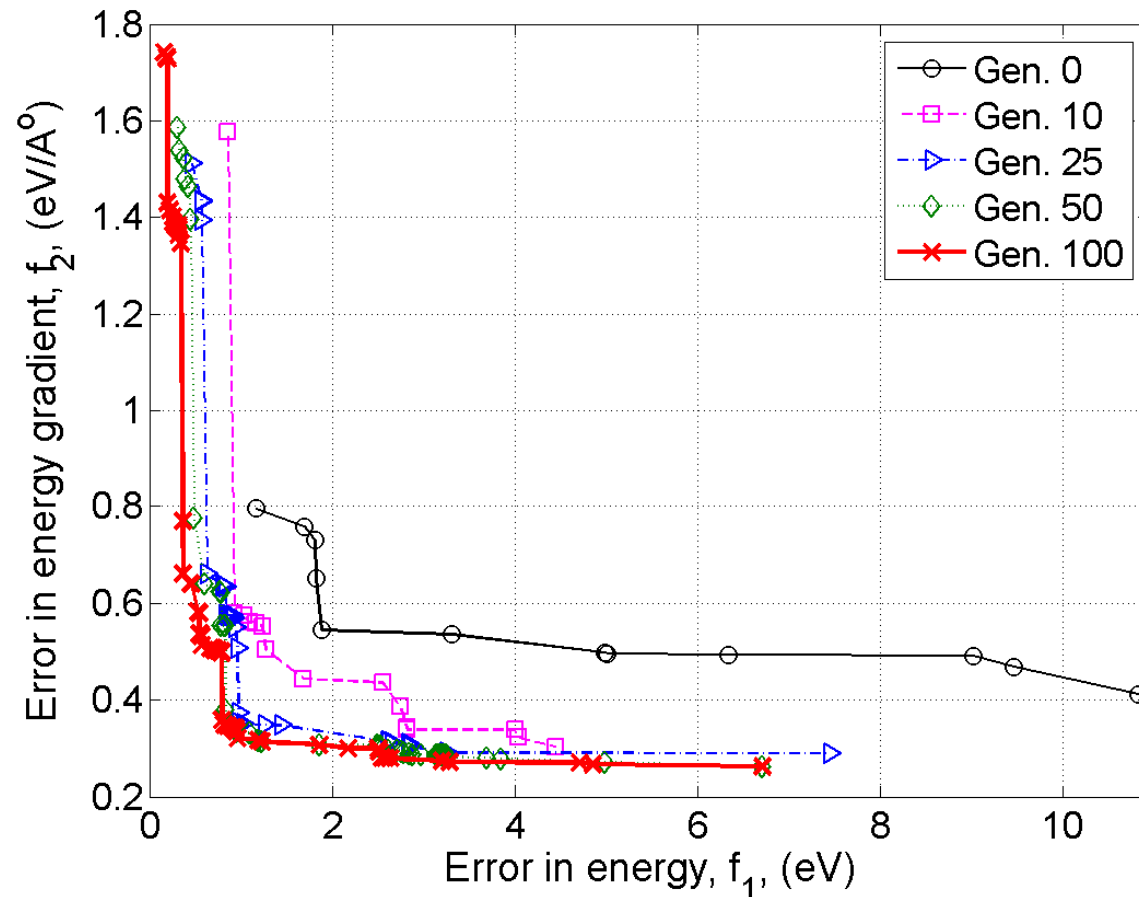
- ❖ 5 independent runs with  $n = 2000$  for 200 generations
  - ◆ Best non-dominated set *assumed* to be **true Pareto front**
- ❖ Mahfoud's population-sizing model suggests  $n = 750$ 
  - ◆ Maintain *at least* one copy of each optimum with 98% probability
- ❖ Empirical results agree with the model prediction
  - ◆ 10 independent runs with  $n = 50, 100, 200, 400,$  and 800
  - ◆ Fixed total number of evaluations at 80,000
  - ◆ With  $n = 800$ , NSGA-II finds almost all Pareto-optimal solutions
- ❖ **Suggests operators are appropriate for the search problem**

# Population Size of 800 Yields Good Solutions

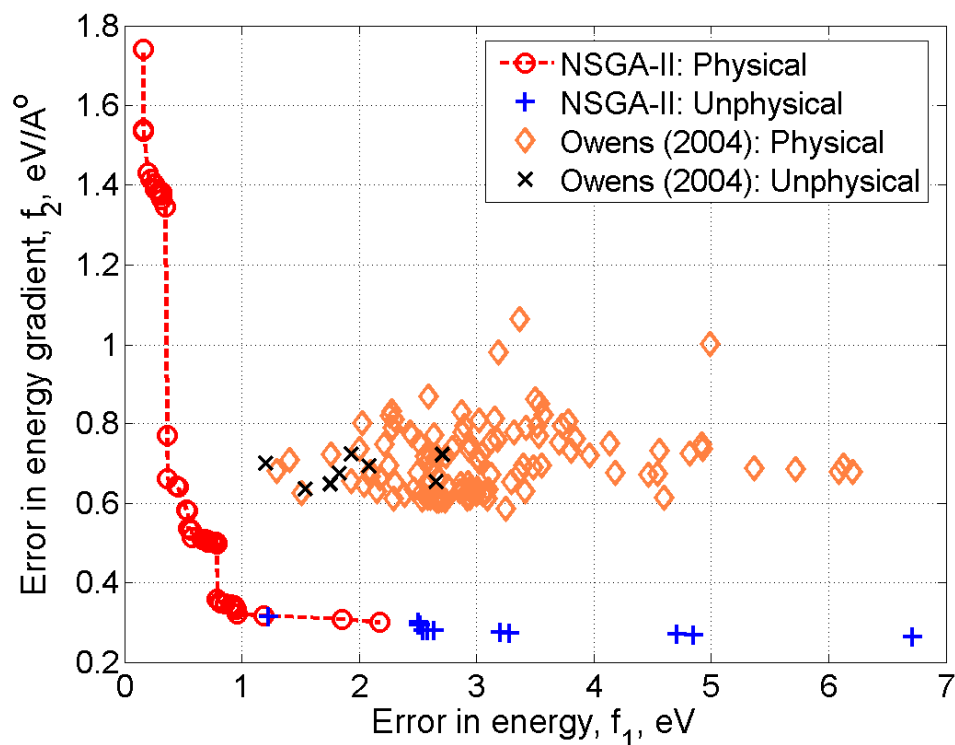


# Run Duration of 100 Generations Is Appropriate

- ❖ 10 independent runs with  $n = 800$ 
  - ◆ Rapid improvement up to gen. 25
  - ◆ Gradual improvement up to gen. 100



# Ethylene: GA Finds Physical and Accurate PES

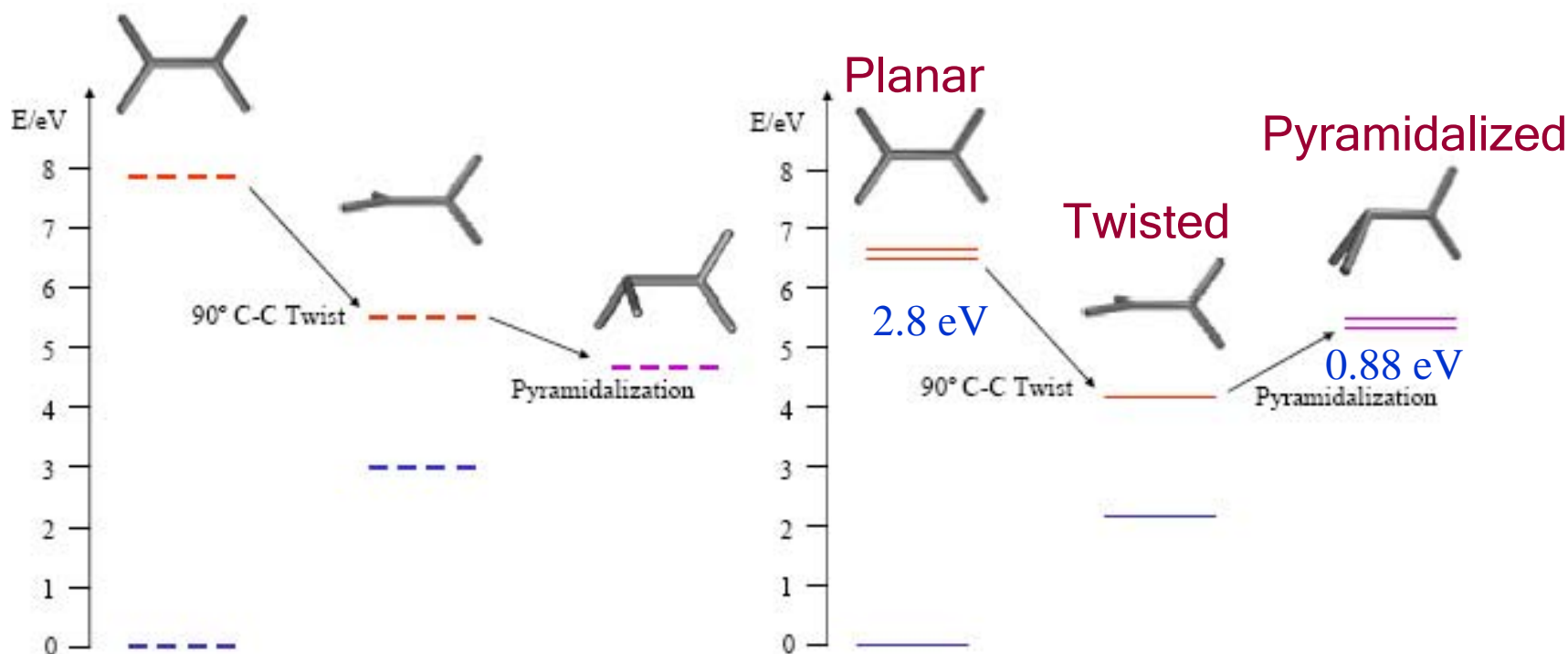


- ❖ 226% lower error in energy
- ❖ 32.5% lower error in energy gradient
- ❖ All solutions below 1.2eV error in energy yield globally accurate PES

- ❖ Significant reduction in errors
- ❖ Globally accurate potential energy surfaces
  - ◆ Resulting in **physical** reaction dynamics
- ❖ Evidence of transferability: “Holy Grail” in molecular dynamics

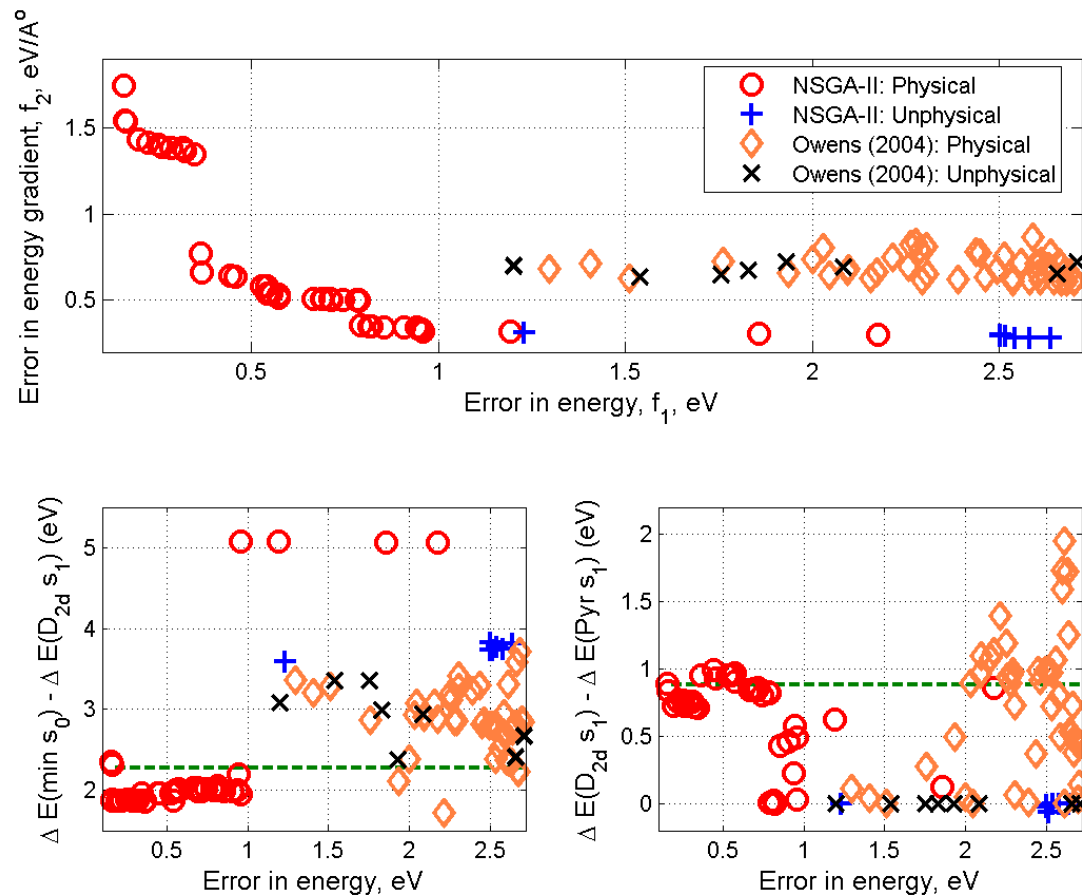
# GA Optimized Semiempirical Potentials are Physical

- ❖ Dynamics agree with *ab initio* results
- ❖ Energetics on untested, yet critical configurations
  - ◆ *cis-trans* isomerization in ethylene
  - ◆ AM1, PM3, and other parameter sets yield **wrong** energetics
  - ◆ **GA yields results consistent with *ab initio* results**

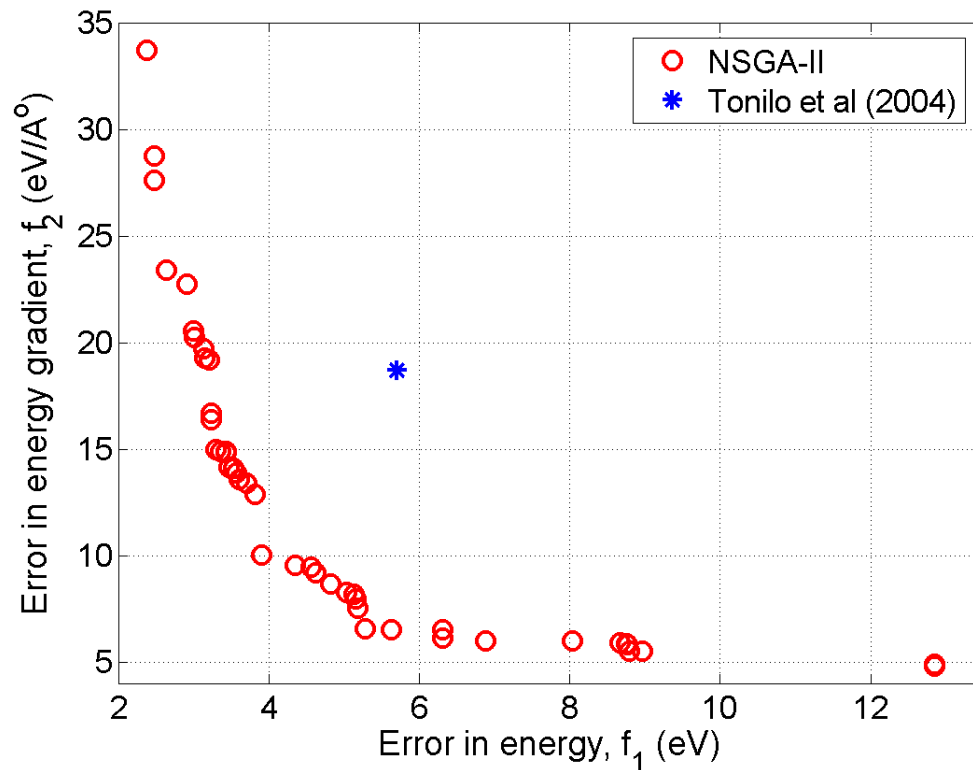


# GA Optimized Semiempirical Potentials are Physical

- ❖ Energy difference between planar and twisted geometry should be greater than zero (ideally  $\sim 2.8$  eV)
- ❖ Energy difference between pyramidalized and twisted geometry should be greater than zero (ideally  $\sim 0.88$  eV)



# Benzene: GA Finds Physical and Accurate PES



- ❖ 46% lower error in energy
- ❖ 86.5% lower error in energy gradient
- ❖ All solutions below 8eV error in energy yield globally accurate PES

- ❖ Significant reduction in errors
- ❖ Globally accurate potential energy surfaces
  - ◆ Resulting in **physical** reaction dynamics
- ❖ Evidence of transferability: “Holy Grail” in molecular dynamics

# Summary of Key Results

- ❖ Yields multiple parameter sets that are **up to 226% lower energy error and 87% lower gradient error**
- ❖ Enables  **$10^2$ - $10^5$**  increase in simulation time even for simple molecules
- ❖  **$10$ - $10^3$**  times faster than the current methodology for tuning semiempirical potentials
- ❖ **Observed transferability is a very important to chemists**
  - ◆ Enables accurate simulations without complete reoptimization
  - ◆ "Holy Grail" for two decades in chemistry & materials science.
- ❖ **Pareto analysis using rBOA and symbolic regression via GP**
  - ◆ Interpretable semiempirical potentials
  - ◆ New insight into multiplicity of models and why they exist.

# Conclusions

- ❖ Broadly applicable in chemistry and materials science
  - ◆ Analogous applicability when multiscaling phenomena is involved: Solids, fluids, thermodynamics, kinetics, etc.
- ❖ Facilitates fast and accurate materials modeling
  - ◆ Alloys: Kinetics simulations with *ab initio* accuracy.  $10^4$ - $10^7$  times faster than current methods.
  - ◆ Chemistry: Reaction-dynamics simulations with *ab initio* accuracy.  $10^2$ - $10^5$  times faster than current methods.
- ❖ Lead potentially to new drugs, new materials, fundamental understanding of complex chemical phenomena
  - ◆ Science: Biophysical basis of vision, and photosynthesis
  - ◆ Synthesis: Pharmaceuticals, functional materials