Supported by the NSF Division of Materials Research

The Materials Computation Center

Duane D. Johnson and Richard M. Martin (PIs)

Funded by NSF DMR 03-25939

Multiscale Modeling Methods for Materials Science and Quantum Chemistry

Genetic Programming: Machine-Learning Method for Multiscale Modeling

D.D. Johnson, D.E. Goldberg, and P. Bellon
Students: Kumara Sastry (MSE/GE), Jia Ye (MSE)
Departments of Materials Science and Engineering and General Engineering
University of Illinois at Urbana-Champaign

Multiscaling via Symbolic Regression

Overview

Multiscale simulations by coupling traditional methods have proven inadequate due to ranges of scales, detailed information needed from finer scales, and the prohibitively large numbers of variables then required. So, for multiscale simulations (spatial and temporal) we must provide data from finer (atomic) scales that is reliable, avoids the need for finding "hidden variables" at various scales, and is computational inexpensive.

Abstract

As such, we employ *Symbolic-Regression* via *Genetic-Programming* – a *Genetic Algorithm* that evolves computer programs – to represent the atomic-scale details needed to simulate processes at time and lengths pertinent to experiment, or even to reveal pertinent ***correlations*** that determine the relevant physics or chemistry at differing scales.

We provide two recent examples involving regression of:

i) Regress the constitutive behavior for an aluminum alloy,
Here a correlation between stress/strain/strain-rate is extracted from experimental data.

ii) Diffusion barriers for multiscale kinetics on alloy surfaces.
A bottleneck for multi-timescale thermally-activated dynamics is computing the potential energy surface (PES). GP regresses symbolically a mapping of ALL saddle-point barriers from a FEW via molecular dynamics, avoiding explicit calculation of all barriers.

First, what is a Genetic Programming (GP)?

A Genetic Program is a genetic algorithm that *evolves* computer programs, requiring:

Representation: programs represented by trees
- Internal nodes contain functions
  - e.g., {+, -, *, /, ^, log, exp, sin, AND, if-then-else, for}
- Leaf nodes contain terminals
  - e.g., Problem variables, constants, Random numbers

Fitness function: Quality measure of the program

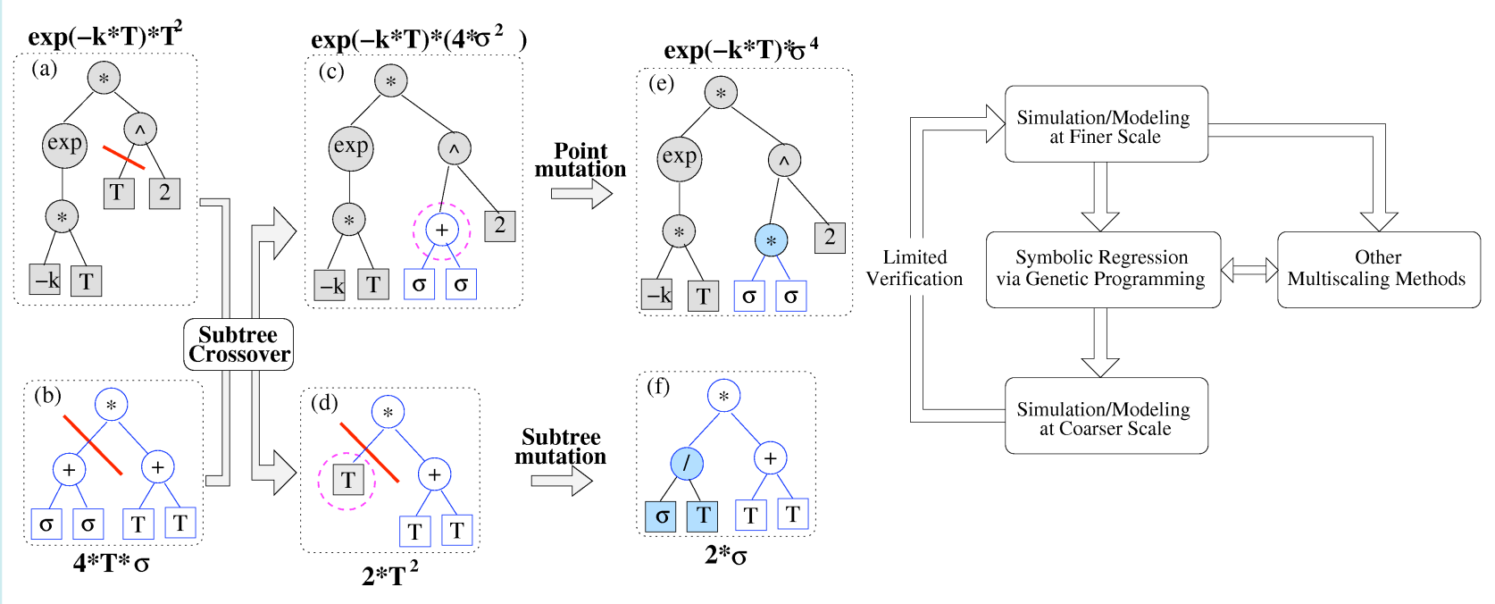
Population: Candidate programs (set of individuals)

Genetic operators:
- **Selection:** "Survival of the fittest".
- **Recombination:** Combine parents to create offspring.
- **Mutation:** Small random modification of offspring.

- Getting the Problems 'Basis Functions'

Using these operations a *tree-like* code is self-generated and provides machine-learned "basis functions" and their "coefficients" (by fitting to some measure of fitness, e.g., comparing calculated and GP-derived diffusion barriers).

– Example "leaf of the tree" (term in basis) created via the above "genetic operators", where (a) and (b) leaves created (e) and (f).

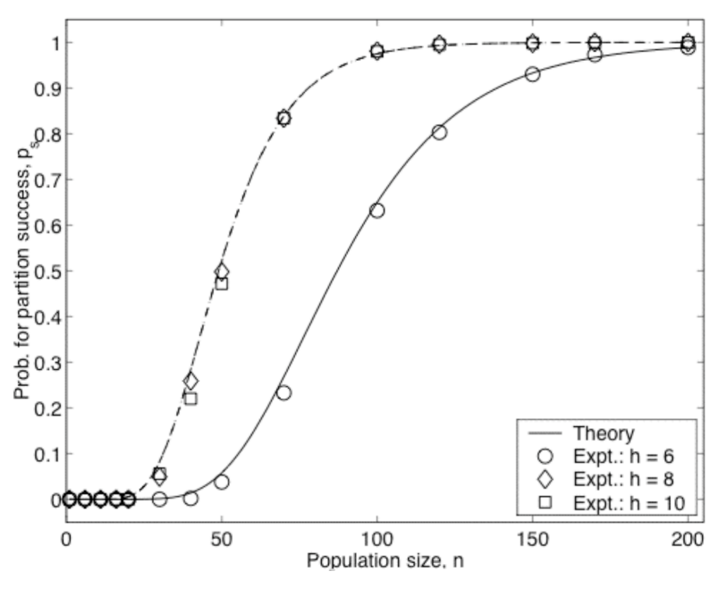

exp(−k*T)*T²
(a)
exp(−k*T)*(4*σ²)
(c)
exp(−k*T)*σ⁴
(e)
Point mutation
Subtree Crossover
(b)
4*T*σ
(d)
2*T²
Subtree mutation
(f)
2*σ
Simulation/Modeling at Finer Scale
Limited Verification
Symbolic Regression via Genetic Programming
Other Multiscaling Methods
Simulation/Modeling at Coarser Scale

- Getting the Problems 'Optimal Population Size'

Analytic Estimate of Population Size vs. Empirical Results: Population size (no. of solutions kept to evolve) is a critical factor to ensure reliable solution.

Shown is the probability that *at least one copy of all raw subcomponents appear in population vs population size*, n, for different tree sizes λ=2^h, for the later diffusion example.

Finding: population of 150-200 is enough.



- Getting the Problems 'Measure of Fitness'

Problem-dependent choice: e.g., for diffusion, choose *weighted* (w_i) *least-squares fit* of GP-derived vs. M calculated barriers, where $w_i = |\Delta E_{MD}|^{-1}$ as lower-energy barriers are more accessible than high-energy ones. Fit could to experimental data, too.

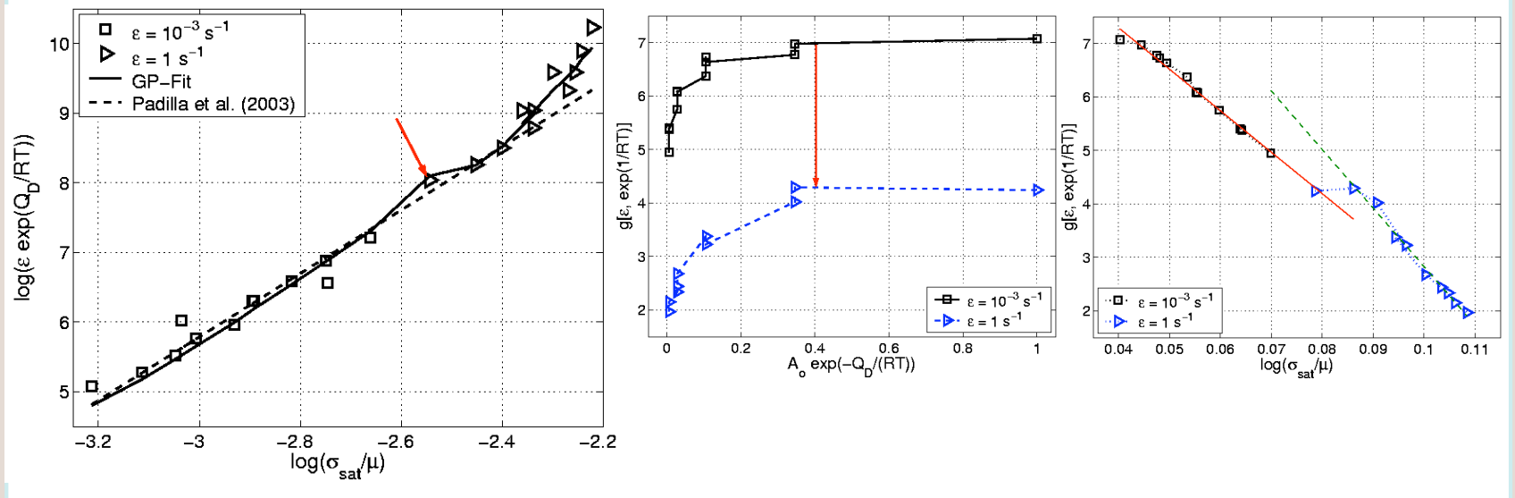
$$fitness, f = \frac{1}{M}\sum_{j=1}^{M} w_i \left|\Delta E_{GP}(\vec{x}_i) - \Delta E_{MD}(\vec{x}_i)\right|$$

1. Evolving Constitutive Relations

Goal: Evolve constitutive "law" between macroscopic variables from stress-strain data with multiple strain-rates for use in continuum finite-element modeling.

Flow stress vs. temperature-compensated strain rate for AA7055 Aluminum [Padilla, et al. (2004)].

- GP fits both low- and high strain-rate data well by introducing (effectively) a step-function between different strain-rate even though no knowledge of two sets of strain-rate data were indicated to GP.



– Automatically identified transition point via a complex relation, *g*, which models a step function between strain-rates involved.

- GP identifies "law" with two competing mechanisms
  - 5-power law modeling known creep mechanism
  - 4-power law for as-yet-unknown 'creep' mechanism.

$$\dot{\epsilon} = \frac{c_O}{g\left[\dot{\epsilon}, \exp\left(\frac{1}{RT}\right)\right]} \exp\left(\frac{1}{RT}\right)\left(\frac{\sigma}{\mu}\right)^4 \left[1 - \frac{\sigma}{\mu}\right]$$

Kumara Sastry, D.D. Johnson, D.E. Goldberg, and P. Bellon,
Int. J. of MultiScale Computational Engineering 2 (2), 239-256 (2004).

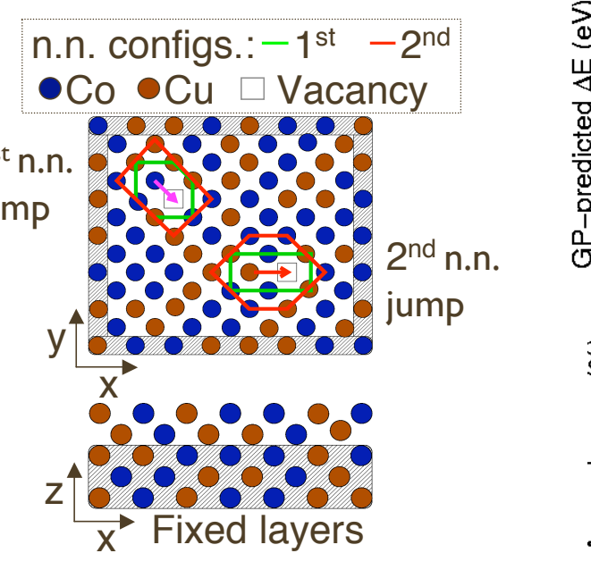
2. Multi-Timescale Kinetics Modeling

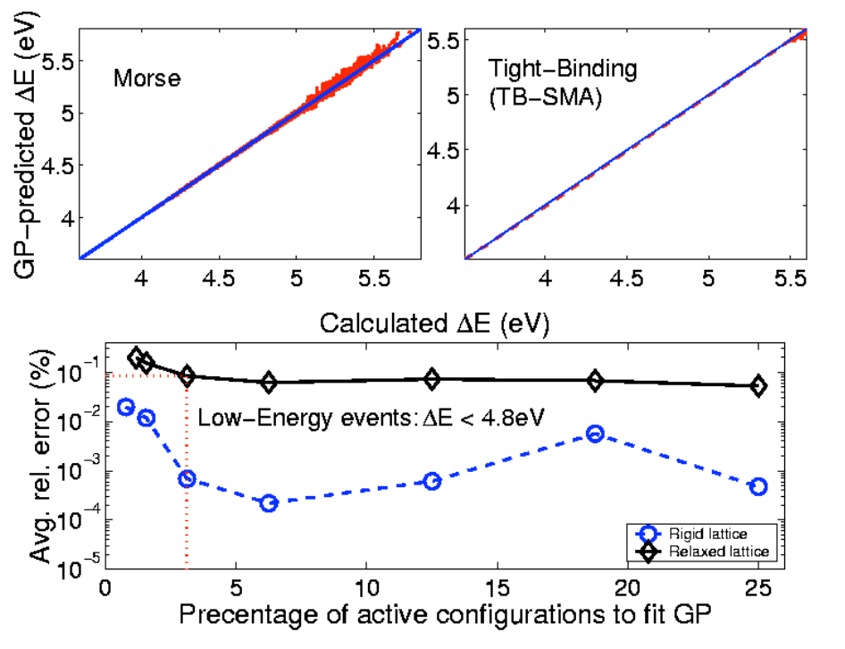
Goal: To advance dynamics simulation to experimentally relevant time scales (seconds) by regressing the diffusion barriers on the PES as an *in-line function*.

- Molecular Dynamic (MD) or Kinetic Monte Carlo (KMC) based methods fall ***short 3–9 orders of magnitude in real time***.
  - Unless ALL the *diffusion barriers* are known in a 'look-up' table.
  - Table KMC has 10^9 increase in "simulated time" over MD at 300K.
- Our new "Symbolically-Regressed" KMC (sr-KMC)
  - Use MD to get *some* barriers.
  - *Machine learn* via GP *all* barriers as a regressed *in-line function call*, i.e. "table-look-up" KMC is replaced by function.

Application: Surface-vacancy-assisted diffusion in segregating $Cu_xCo_{1-x.}$

- Using Molecular Dynamics based on density-functional, tight-binding, or empirical potentials, we calculate M (un)relaxed saddle-point energies $\Delta E(x_i)$ for atoms surrounding a vacancy with first and second neighbor environment denoted by 0 or 1 (for binary alloys) in a vector $\{x_i\}$.
- GP evolves *in-line barrier function* and *predicts* remaining unknown barriers.
- Newly predicted low-energy barriers are calculated directly by MD as *verification step*. If correct, use barrier function. If not correct, now have new barrier in a M+1 learning set. Repeat cycle (MD is +99.9% of step).





- GP predicts all barriers with ~0.1% error from explicit calculations of only <3% of the barriers. (Standard basis-set regressions *fail.)*
- GP symbolic-regression approach yields:
  - 10^2 decrease in CPU time for barrier calculations.
  - 10^2 decrease in CPU over table-look-ups (in-line function call).
  - 10^4–10^7 less CPU time per time-step vs. *on-the-fly* methods (note that each barrier calculation requires 10 s with empirical potential, 1800 s for tight-binding, and first-principles even more).
- (Future) Could combine with pattern-recognition methods *(e.g., T. Rahman et al.)*, or temperature-accelerated MD, to model more complex cooperative dynamics.
- (Current) Utilize the GP in-line table function obtain from tight-binding potential in a kinetic Monte Carlo simulation for this surface alloy vacancy-assisted diffusion.

K. Sastry, H.A. Abbass, D.E. Goldberg, D.D. Johnson, "Sub-structural Niching in Estimation Distribution Algorithms," Genetic and Evolutionary Computation Conference (2005).

Kumara Sastry, D.D. Johnson, D.E. Goldberg, and P. Bellon, "Genetic programming for multitimescale modeling," Phys. Rev. B 72, 085438-9 (2005).
chosen by the AIP Editors as focused article of frontier research in
***Virtual Journal of Nanoscale Science & Technology*, Vol 12, Issue 9 (2005).**

Summary

Our results indicate that GP-based symbolic regression is an effective and promising tool for multiscaling. The flexibility of GP makes it readily amenable to hybridization with other multiscaling methods leading to enhanced scalability and applicability to more complex problems. Unlike traditional regression, GP adaptively *evolves* both the functional relation and regression constants for transferring key information from finer to coarser scales, and is inherently parallel.

Ab Initio Accurate Semiempirical Quantum Chemistry Potentials via Multi-Objective GAs

D.D. Johnson, T.J. Martinez, and D.E. Goldberg,
Students: Kumara Sastry (MSE/GE) and Alexis L. Thompson (Chemistry)
Departments of Materials Science and Engineering, Chemistry, and General Engineering
University of Illinois at Urbana-Champaign

Ab Initio Accurate Semiempirical Potentials Excited-State Reaction Chemistry

Recently, use of genetic algorithms to fit empirical potentials has grown in interest to build in more problem specific information cheaply. For example, developing an *accurate empirical potential from database of high-level quantum-chemistry results* is done by serial fitting to minimize error in *energy differences between ground-state and excited states* and then *error in the energy derivative differences*. Typically, however, the fitting is done in a serial fashion (first on error of energy difference, then on error in derivatives), which is not a global search. Moreover, the genetic algorithms used are not so-called *competent GAs* developed from optimization theory, which lead to bad scaling and inefficient performance.

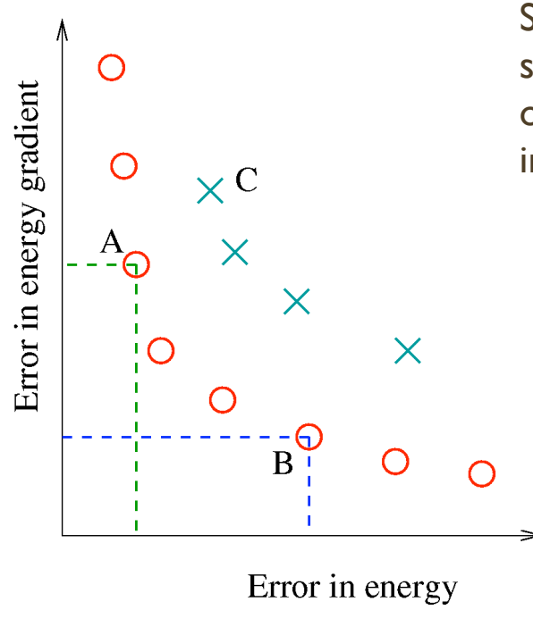
Here we explore the use of *Non-Dominant, Multi-Objective Minimization using Genetic Algorithm* to **re**parameterize semi-empirical quantum-chemistry potentials over a global search domain using the concepts of Pareto optimization fronts.

Goal: Functional augmentation and rapid multi-objective *re*parameterization of semi-empirical methods to obtain reliable pathways for excited-state reaction chemistry.

- *Ab Initio* methods: accurate, but highly expensive.
- Semi-Empirical (SE) methods: approximate, but very inexpensive.
  - *Reparameterization* based on few *ab initio* calculated data sets involving excitations of a molecule, rather than low-energy (Born-Oppenheimer) states, e.g. use MNDO-PM3 Hamiltonian and find the MNDO parameters specific to particular molecular system.
  - Involves optimization of multiple objectives, such as fitting simultaneously limited *ab initio* energy and energy-gradients of various chemical excited-states or conformations.
  - (Future) Augmentation of functions may be needed.
- Propose: Multi-objective GAs for reparameterization
  - *Non-dominate solutions* represent physically allowed solutions, whereas dominant solutions can lead to unphysical solutions.
  - *Obtain* set of ***Pareto*** *non-dominate solutions* in parallel, not serially.
  - *Avoid* potentially irrelevant pathways, arising from SE-forms, so as to reproduce more accurate reaction paths.
  - (Future) Use Genetic Programming for functional augmentation, e.g., symbolic regression of core-core repulsions.
- Advantages of GA/GP Multi-Objective Optimizations, method is:
  - robust, and yields good quality solutions quickly, reliably, and accurately,
  - converges rapidly to Pareto-optimal ones,
  - maintain diverse populations,
  - suited to finding diverse solutions,
  - niche-preserving methods may be employed,
  - implicitly parallel search method, unlike applications of classic methods.

- What is Non-Dominant Solutions on Pareto-Optimal Front?

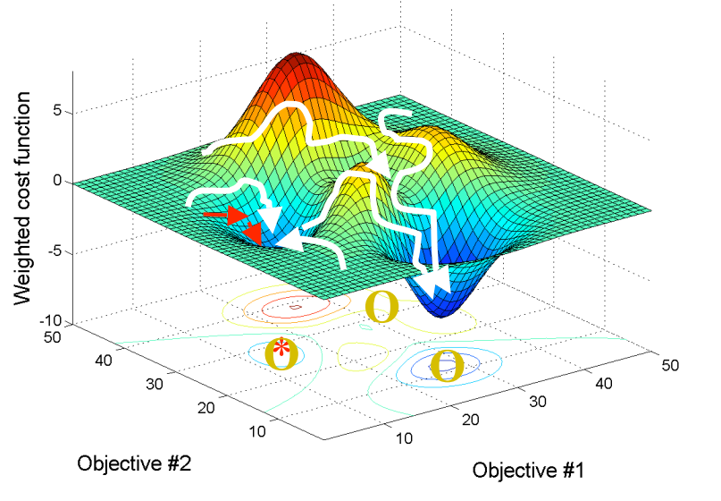
Using a MNDO method for Benzene C_6H_6 requires *11 parameters*, if the H parameters are fixed. To fit accurately CASPT2 results for two objectives (energy and energy-gradient errors) on the excited-state potential energy surface (Frank-Condon region), the 11 parameters are globally optimized keeping a population of solutions to evolve and the solutions at the 'nose' of the Pareto are accepted as 'best' solutions.



Semi-empirical potential parameterizations lead to differing solutions, or competing solutions. Using GA/GP we can find optimal potentials and avoid pathways from dominating but irrelevant solutions.
- e.g., solution C is dominate over B in *error in energy*.
- solutions A and B are non-dominate in multiobjectives.
- ***Pareto front*** solutions are denoted by red circles and all represent potentially "good" solutions all with different parameter sets for the empirical potential
- For physical reasons, the ***"nose" of Pareto front*** (between A and B) gives the most optimal solution desired for quantum-chemistry applications.

GA/GP multiobjective optimization avoids falling into local minimum of fitness, as marked by *red arrows and star*. Whereas the global optimized solutions on the Pareto front are shown in *dark blue*.

Also, avoids possible irrelevant and unphysical reaction paths possible from semiempirical potential forms.

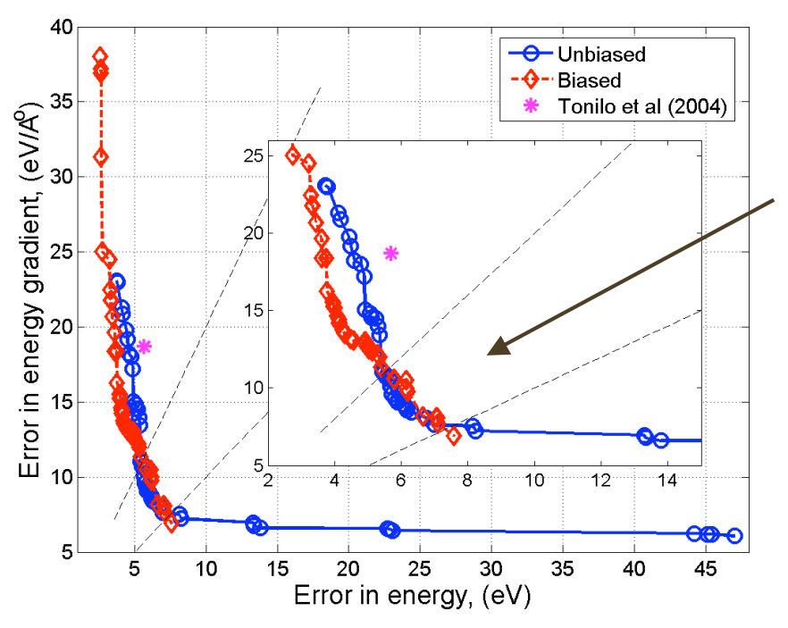


- Biasing the Multi-Objective Search

Weights can be assigned to each objective to bias search and speed up global search. For example, *error in energy* can easier weighted as more important to minimize than the *error in energy-gradient.* , even if both objectives are obtain via an analytic formula.

Such weighting is an important parameter for control of time to solution.

- (Un)Biased GA Multiobjective Optimization of Benzene
  - Biasing (here factor of 2) the *error in energy* over *error in energy-gradient* yields rapid advance of Pareto front and physical solutions.
  - Unbiased, if left to evolve long enough, reaches biased solutions, but early solutions may yield unphysical excited-state reactions.
  - (Un)Biased solutions on the Pareto front consistently better than all previous parameterizations, including using standard GA optimization, e.g., from Martinez and coworkers, see Toniolo, et al. (2004).

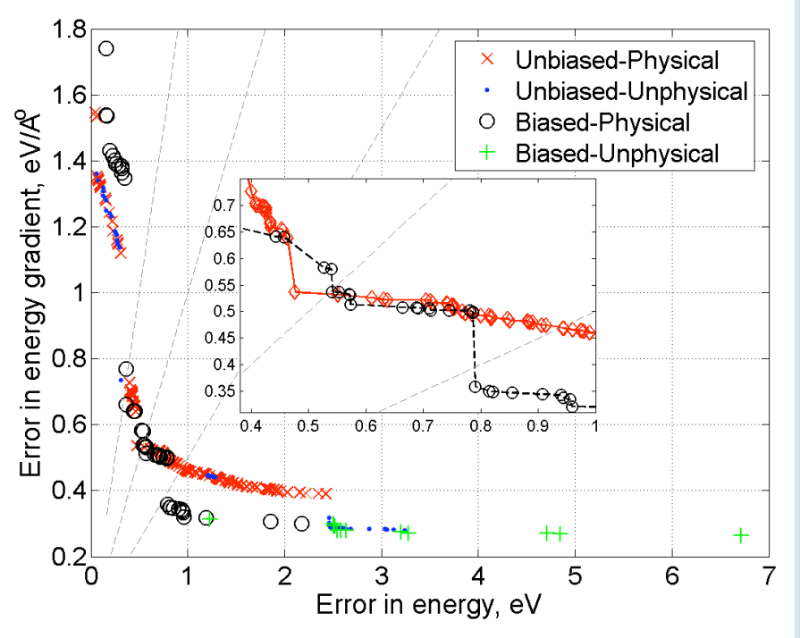


Re-parameterized MNDO Hamiltonian yields relatively accurate excited-state potential energy surfaces.

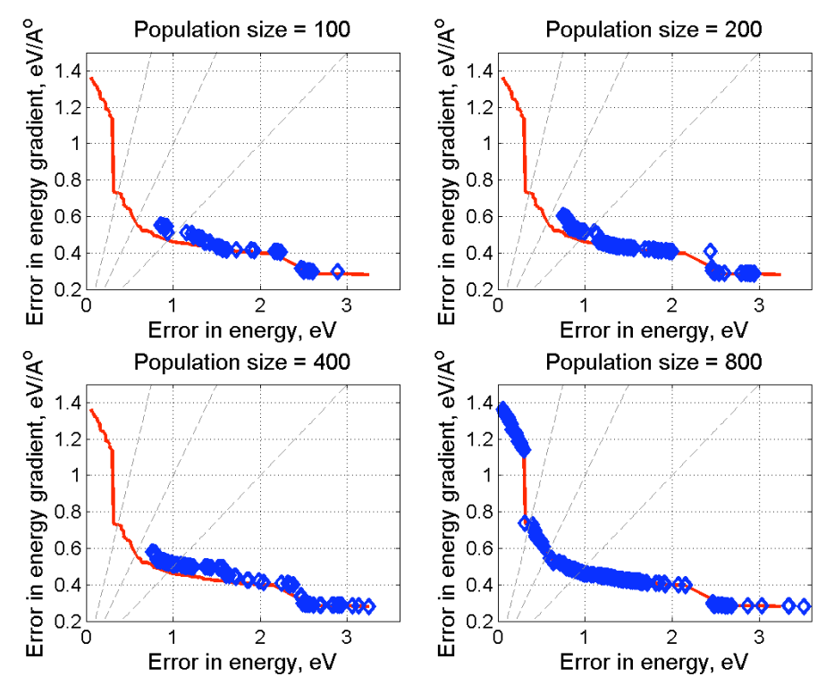
- GA-MO-dervied MNDO S2/S1 conical intersections agree well with CASPT2, even though only included x=0 reaction coordinate in fitting.
- Molecular geometry for excited-states also agree well.

- (Un)Biased GA Multiobjective Optimization of Ethylene, C_2H_4.
  - Found similar results to Benzene: Biased solutions on Pareto front often better than unbiased and always physical. But near the nose all solutions are physical.
  - We find that the historical MNDO parameters are a set yielding almost unphysical solutions (see figure near 2.5 eV on *error in energy*).
  - GA-MO-derived MNDO S2/S1 conical intersections agree well with CASPT2, with only x=0 reaction coordinate included in fitting.
  - Molecular geometry for excited-states also agree well.

Transferability of the MNDO parameters: Amazingly we find that a Benzene set of parameters may be used for Ethylene and provide a solution near a Pareto set found by direct optimization.



- Population Analysis for Ethylene, C_2H_4
  - Must maintain large enough population to obtain full Pareto front but not so large as to waste computational resources because each solution is a full MNDO run for the set of molecular configurations used in fitting!



- **Red Line** is ***Pareto front*** for large population > 1000.
- Analytic estimate suggests ~760 is required to find population size.
- Figure show that until ~800 the Pareto front is not found.
- For Benzene, only about ~150 is required for the population size.

Summary

- We find that *non-dominant, multi-objective reparameterization of empirical Hamiltonians using Genetic Algorithms* is an effective tool for developing *ab initio accurate empirical potential* based upon databases from high-level quantum-chemistry methods.
- *Excited-state properties* (reaction paths and structures) are in very good agreement with direct CASPT2 calculations.
- We find that parameters sets from one molecular system is *transferable* to a similar molecular system, opening the possibility of addressing more complex molecular interactions.

Future Directions

- We will investigate the use of *Genetic Programming* to machine-learn *new and more accurate empirical potential functional forms.*
- e.g. We will start with the original MNDO Hamiltonian and *machine-learn in a molecular-specific way a GP-MNDO Hamiltonian*.
- With this *GP-MNDO Hamiltonian* we can perform nearly *ab initio* accurate global searches of reaction pathways, which later may be studied with higher-level methods for reactions of interest.

Acknowledgements

- We thank ILLIGAL (Illinois Genetic Algorithms Lab) for use of their parallel cluster for the MO-GA optimization.
- This multidisciplinary effort was made possible only via support of the MCC and the National Science Foundation (Divisions of Chemistry and Materials Research).