

# Finding Many Stable Molecular Arrangements

Conformational Searching with Genetic Algorithms

---

Evan Curtin

December 2, 2016

University of Illinois at Urbana-Champaign

# Outline

1. Background Information
2. The Genetic Algorithm
3. Finding Low Energy Conformers of Dipeptides
4. Concluding Remarks

*First-Principles Molecular Structure Search with a Genetic Algorithm*

Supady, A.<sup>P1</sup>; Blum, V.<sup>1</sup>; Baldauf, C. J.<sup>1,2</sup> Chem. Inf. Model. 2015, 55 (11), 23382348.

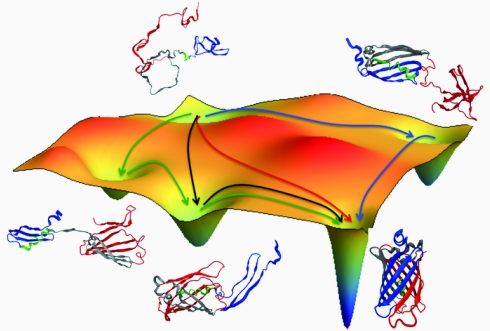
1. Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin
2. Department of Mechanical Engineering & Materials Science, Duke University

## Background Information

---

# The Problem

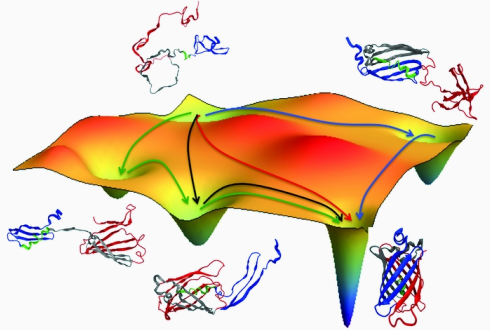
- Protein folding is important



Reddy, G.; Liu, Z.; Thirumalai, D. Proc. Natl. Acad. Sci. 2012, 109 (44), 17832-17838.

# The Problem

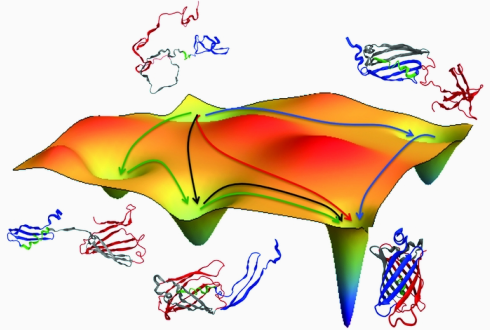
- Protein folding is important
- Peptides are building blocks of protein



Reddy, G.; Liu, Z.; Thirumalai, D. Proc. Natl. Acad. Sci. 2012, 109 (44), 1783217838.

# The Problem

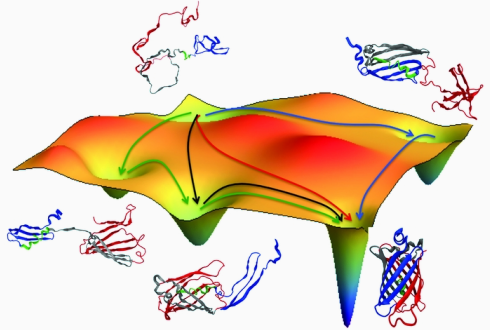
- Protein folding is important
- Peptides are building blocks of protein
- Peptide conformations → protein folding



Reddy, G.; Liu, Z.; Thirumalai, D. Proc. Natl. Acad. Sci. 2012, 109 (44), 1783217838.

# The Problem

- Protein folding is important
- Peptides are building blocks of protein
- Peptide conformations → protein folding
- How do we understand peptide conformations?



Reddy, G.; Liu, Z.; Thirumalai, D. Proc. Natl. Acad. Sci. 2012, 109 (44), 1783217838.



# The Problem

**Computational methods require knowledge of molecular structure**

We need to find the lowest energy structure

# The Problem

Computational methods require knowledge of molecular structure

We need to find the lowest energy structure

The potential energy surface (PES) is high dimensional and has many minima

We can't tell for sure if we've found the **global** minimum



# The Problem

Computational methods require knowledge of molecular structure

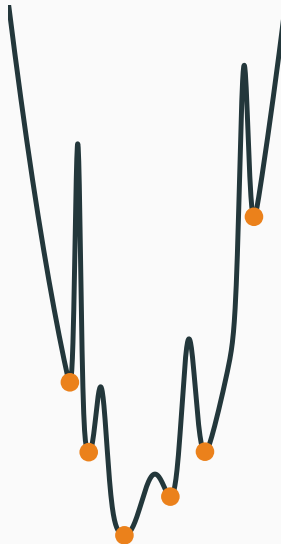
We need to find the lowest energy structure

The potential energy surface (PES) is high dimensional and has many minima

We can't tell for sure if we've found the **global** minimum

We may need information about one or more low-energy conformations

- Ok, let's find them all!



# The Problem

Computational methods require knowledge of molecular structure

We need to find the lowest energy structure

The potential energy surface (PES) is high dimensional and has many minima

We can't tell for sure if we've found the **global** minimum

We may need information about one or more low-energy conformations

- Ok, let's find them all!
- Under a cutoff



# Possible Solutions

- Many techniques are well established

Method	Implented in
grid-based	CEASAR, <b>Open Babel</b> , <b>Confab</b> , MacroModel, MOE
rule-based	<b>ALFA</b> , <b>CONFECT</b> , CORINA, ROTATE, <b>COSMOS</b> , OMEGA
population-based	<b>Balloon</b> , <b>Cyndi</b>
basin-hopping	<b>ASE</b> , <b>GMIN</b> , <b>TINKER SCAN</b>

# Possible Solutions

- Many techniques are well established
- None are perfect

Method	Implented in
grid-based	CEASAR, <b>Open Babel</b> , <b>Confab</b> , MacroModel, MOE
rule-based	<b>ALFA</b> , <b>CONFECT</b> , CORINA, ROTATE, <b>COSMOS</b> , OMEGA
population-based	<b>Balloon</b> , <b>Cyndi</b>
basin-hopping	<b>ASE</b> , <b>GMIN</b> , <b>TINKER SCAN</b>

## What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures

## What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures
2. Minimize # of geometry optimizations



## What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures
2. Minimize # of geometry optimizations
3. Find many low energy conformations

## What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures
2. Minimize # of geometry optimizations
3. Find many low energy conformations
4. Minimize human bias

## What Algorithmic Properties do we want for conformer search?

1. Accurate energies & Structures
2. Minimize # of geometry optimizations
3. Find many low energy conformations
4. Minimize human bias
5. Parallel

# The Genetic Algorithm

---

- At least 23 years

Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. J. Comput. Chem. 1993, 14 (11), 14071414.

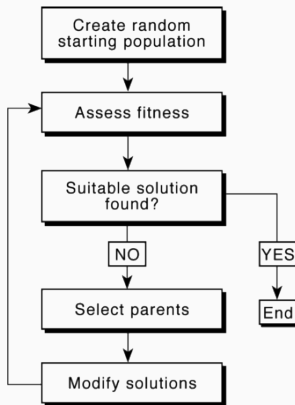
- Flexible ligand docking (Most cited: 7428)

Morris, G. et al. J. Comput. Chem. 1998, 19 (14), 16391662.

- Precombustion CO<sub>2</sub> adsorbing MOFs (October 14<sup>th</sup>)

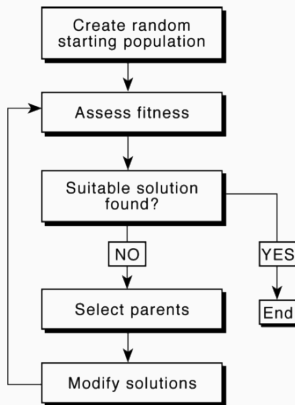
Chung, Y. G.; Gomez-Gualdron, D. A.; Li, P.; Leperi, K. T.; Deria, P.; Zhang, H.; Vermeulen, N. A.; Stoddart, J. F.; You, F.; Hupp, J. T.; Farha, O. K.; Snurr, R. Q. Sci. Adv. 2016, 2 (10)

- Inspired by biological evolution



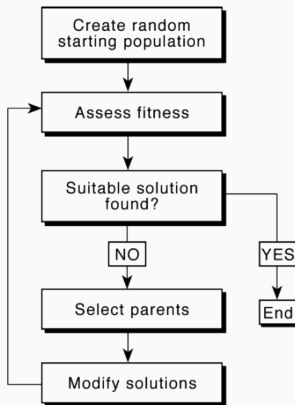
# Outline

- Inspired by biological evolution
- Evolve a population over generations



# Outline

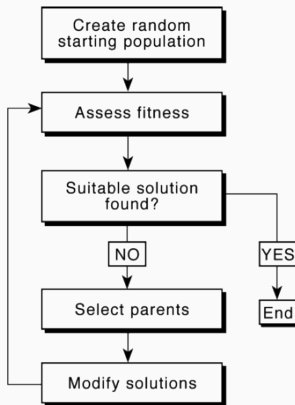
- Inspired by biological evolution
- Evolve a population over generations
- Survival of the fittest





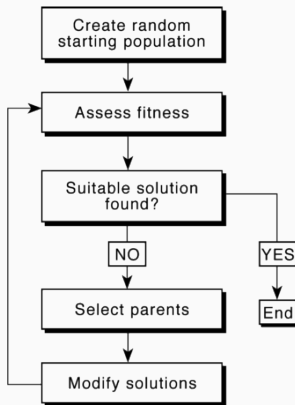
# Outline

- Inspired by biological evolution
- Evolve a population over generations
- Survival of the fittest
- Requirements:



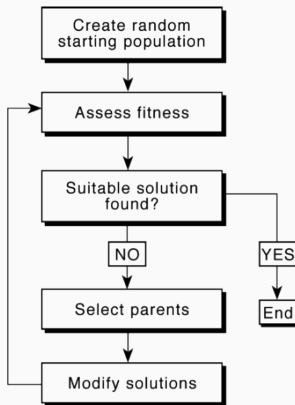
# Outline

- Inspired by biological evolution
- Evolve a population over generations
- Survival of the fittest
- Requirements:
  - Represent individuals as vector



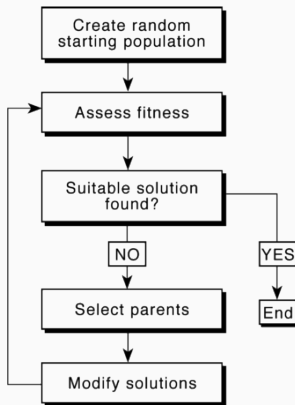
# Outline

- Inspired by biological evolution
- Evolve a population over generations
- Survival of the fittest
- Requirements:
  - Represent individuals as vector
  - Fitness function



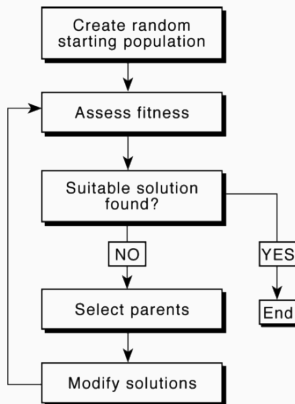
# Outline

- Inspired by biological evolution
- Evolve a population over generations
- Survival of the fittest
- Requirements:
  - Represent individuals as vector
  - Fitness function
- $V = (x_1 \ y_1 \ z_1 \ x_2 \ y_2 \ z_2 \ \cdots \ x_N \ y_N \ z_N)$



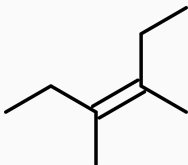
# Outline

- Inspired by biological evolution
- Evolve a population over generations
- Survival of the fittest
- Requirements:
  - Represent individuals as vector
  - Fitness function
- $V = (x_1 \ y_1 \ z_1 \ x_2 \ y_2 \ z_2 \ \dots \ x_N \ y_N \ z_N)$
- $F = \frac{E_{\max} - E}{E_{\max} - E_{\min}}$

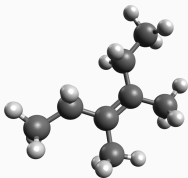


# The Many Representations of a Molecule

2D Image



3D Image



Name

(3Z)-3,4-Dimethyl-  
3-hexene

SMILES

CCC(C)=C(C)CC

InChI

1S/C8H16/c1-5-  
7(3)8(4)6-2/h5-  
6H2,1-4H3/b8-7-

Cartesian Coords

	x	y	z
C	0.90	-0.25	0.02
C	2.35	0.15	-0.17
C	2.91	1.30	-0.67

Internal Coords

		r		$\theta$
C				
C	1	1.51		
C	2	1.38	1	131

Internal Coords are the natural choice here

1. Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.
2. <http://www.chemspider.com/Chemical-Structure.2298795.html>

# Selecting Parents

## Roulette Wheel Method

1. Reinforce good characteristics

# Selecting Parents

## Roulette Wheel Method

1. Reinforce good characteristics
2. Still give losers a chance



# Selecting Parents

## Roulette Wheel Method

1. Reinforce good characteristics
2. Still give losers a chance
3. 'Breed' pairs of winners

# Selecting Parents

## Roulette Wheel Method

1. Reinforce good characteristics
2. Still give losers a chance
3. 'Breed' pairs of winners

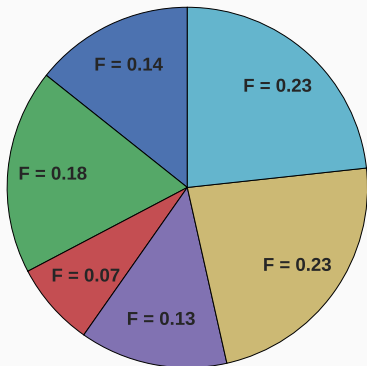


Sum of Fitness Scores = 1.0

# Selecting Parents

## Roulette Wheel Method

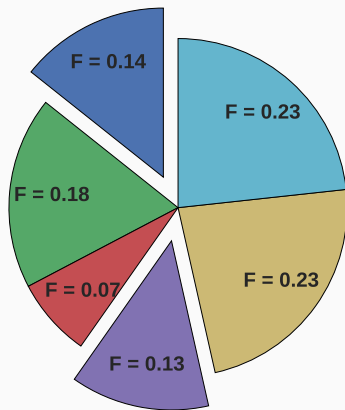
1. Reinforce good characteristics
2. Still give losers a chance
3. 'Breed' pairs of winners



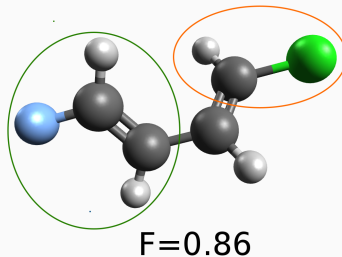
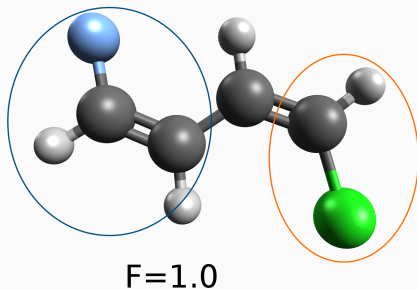
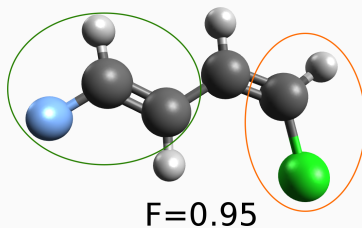
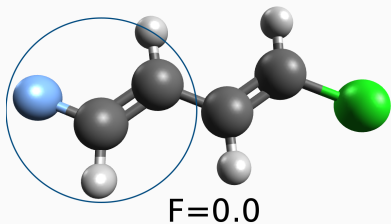
# Selecting Parents

## Roulette Wheel Method

1. Reinforce good characteristics
2. Still give losers a chance
3. 'Breed' pairs of winners

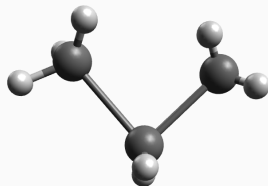


# The Next Generation



# The Whole Algorithm

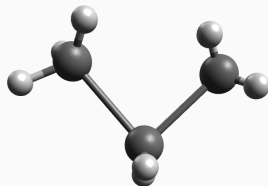
1. Generate N random, sensible geometries



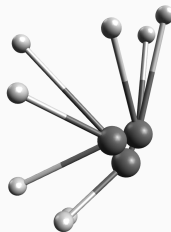
sensible

# The Whole Algorithm

1. Generate N random, sensible geometries



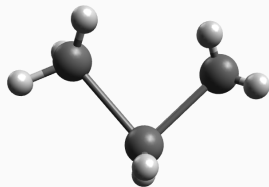
sensible



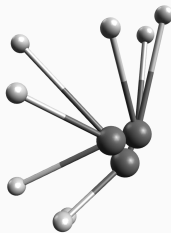
utter nonsense

# The Whole Algorithm

1. Generate N random, sensible geometries
2. Optimize



sensible

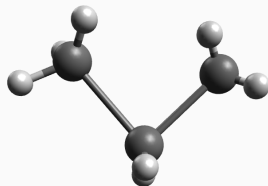


utter nonsense

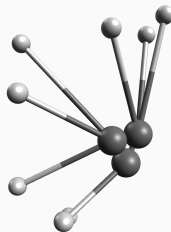


# The Whole Algorithm

1. Generate N random, sensible geometries
2. Optimize
3. Select Parents



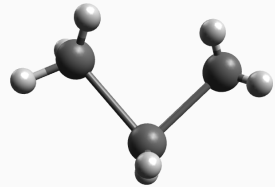
sensible



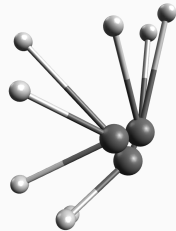
utter nonsense

# The Whole Algorithm

1. Generate N random, sensible geometries
2. Optimize
3. Select Parents
4. Crossover & Mutate



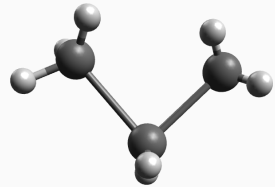
sensible



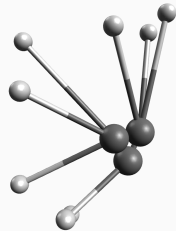
utter nonsense

# The Whole Algorithm

1. Generate N random, sensible geometries
2. Optimize
3. Select Parents
4. Crossover & Mutate
5. Add Children to population



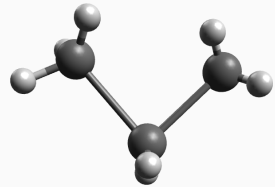
sensible



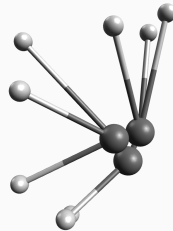
utter nonsense

# The Whole Algorithm

1. Generate N random, sensible geometries
2. Optimize
3. Select Parents
4. Crossover & Mutate
5. Add Children to population
6. Remove the unfit



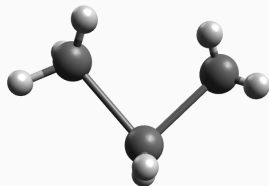
sensible



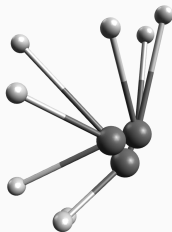
utter nonsense

# The Whole Algorithm

1. Generate N random, sensible geometries
2. Optimize
3. Select Parents
4. Crossover & Mutate
5. Add Children to population
6. Remove the unfit
7. If converged:
  - Done!



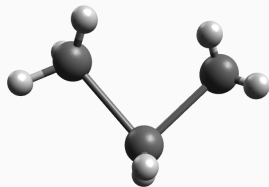
sensible



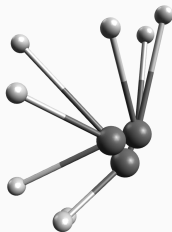
utter nonsense

# The Whole Algorithm

1. Generate N random, sensible geometries
2. Optimize
3. Select Parents
4. Crossover & Mutate
5. Add Children to population
6. Remove the unfit
7. If converged:
  - Done!Otherwise:
  - Go to 2



sensible

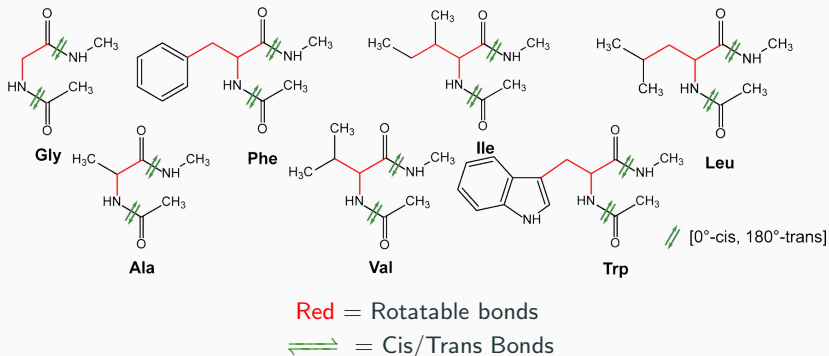


utter nonsense

# Finding Low Energy Conformers of Dipeptides

---

# "Dipeptide" Structures



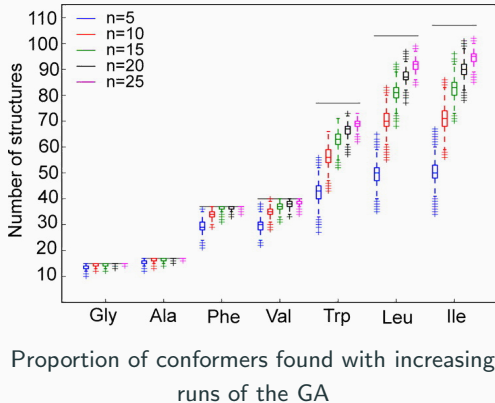
*"We use the term dipeptide for amino acids with an acetylated N terminus and an amino-methylated C terminus"*



	Molecule	N	# Rotatable +	# Conformers
			# Cis/Trans Bonds	
• GA beats other methods if space is large	Gly	19	2 + 2	15
	Ala	22	2 + 2	28
	Phe	32	4 + 2	64
• Space gets large <b>fast</b>	Val	28	3 + 2	60
	Trp	36	4 + 2	141
	Leu	31	4 + 2	183
	Ile	31	4 + 2	176

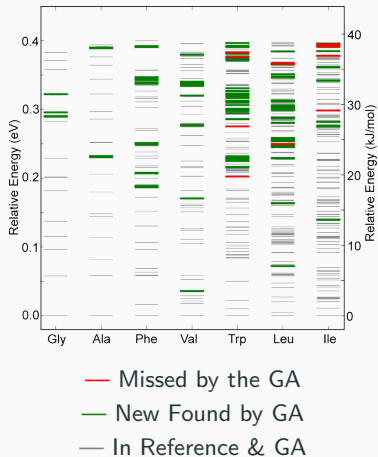
# Coverage

- Smaller systems are reliably sampled
- As # of conformers increases, miss more and more
- Is there a pattern to what is missed?



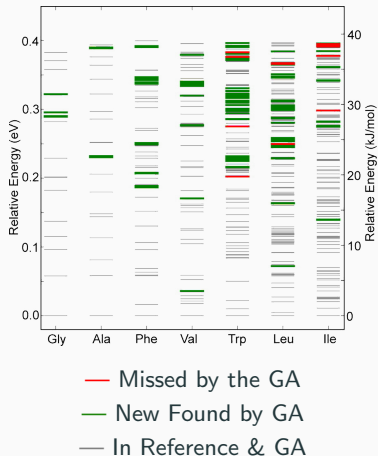
# Coverage

- Most misses are very high energy



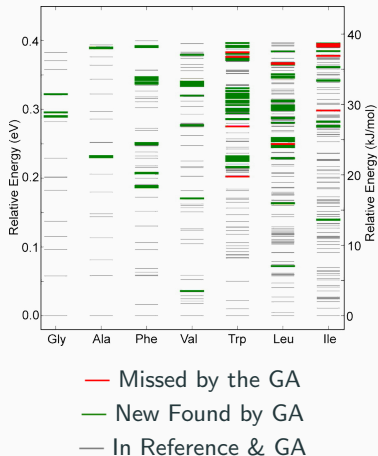
# Coverage

- Most misses are very high energy
- Algorithm favors low energy areas of the space

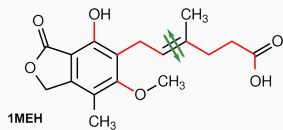


# Coverage

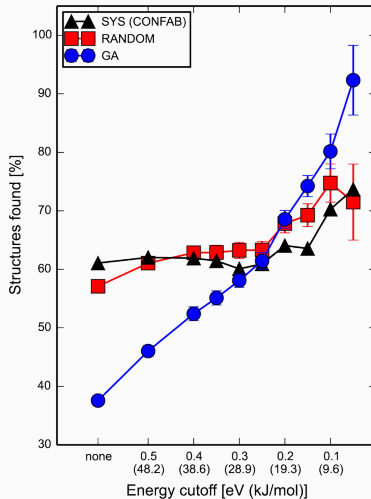
- Most misses are very high energy
- Algorithm favors low energy areas of the space
- Features low in energy are favored and recombined



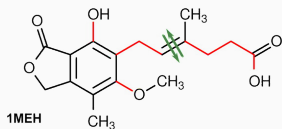
# Energy Cutoff



Mycophenolic Acid

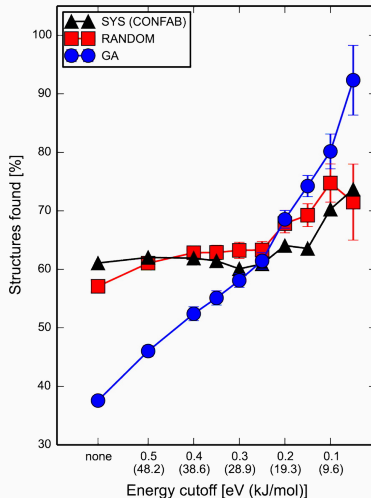


# Energy Cutoff

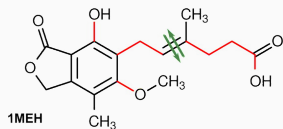


Mycophenolic Acid

- GA is more sensitive to energy cutoff

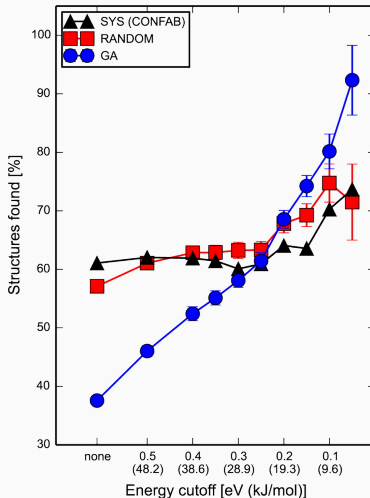


# Energy Cutoff



Mycophenolic Acid

- GA is more sensitive to energy cutoff
- For finding low energy ensemble, GA outperforms purely stochastic/deterministic method





## Concluding Remarks

---

# Concluding Remarks

## Review

- Conformational searching is expensive

# Concluding Remarks

## Review

- Conformational searching is expensive
- The Genetic Algorithm is a guided global search

## Review

- Conformational searching is expensive
- The Genetic Algorithm is a guided global search
- It shines when asked to find many low energy solutions

## Review

- Conformational searching is expensive
- The Genetic Algorithm is a guided global search
- It shines when asked to find many low energy solutions
- GA can be used with any electronic structure package

## Review

- Conformational searching is expensive
- The Genetic Algorithm is a guided global search
- It shines when asked to find many low energy solutions
- GA can be used with any electronic structure package
- This one is available under the GNU Lesser General Public License:  
<https://github.com/adrianasupady/fafoom>

**Questions?**

- Geometry optimization step makes the algorithm more Lamarckian (Jean Baptiste Larmarck, [1744-1829])



# Genetic Algorithm Parameters

Geometry Optimization: DFT PBE + VdW, *tier1* basis in FHI-aims<sup>1</sup>.  
Convergence at 0.005 eV / Å

	parameter	value
molecule	SMILES	<chem>CC(=O)N[C@H](C(=O)NC)[C@H](CC)C</chem>
	distance_cutoff_1	1.2 Å
	distance_cutoff_2	2.0 Å
	rmsd_cutoff_uniq	0.2 Å
	chiral	true
run settings	max_iter	10
	iter_limit_conv	10
	energy_diff_conv	0.001 eV
GA settings	popsize	5
	energy_var	0.001 eV
	selection	roulette wheel
	fitness_sum_limit	1.2
	prob_for_crossing	0.95
	cross_trial	20
	prob_for_mut_cistrans	0.5
	prob_for_mut_rot	0.5
	max_mutations_cistrans	1
	max_mutations_torsions	2
	mut_trial	100

GA Parameters for Isoleucine Dipeptide<sup>2</sup>

(1) Blum, V. et. al., M. Comput. Phys. Commun. 2009, 180 (11), 21752196.

(2) Supady, A.; Blum, V.; Baldauf, C. J. Chem. Inf. Model. 2015, 55 (11), 23382348.