

Gradient-Based Optimization for Molecular Design

Weitao Yang, Duke University



Funding

DARPA

NSF

NIH-Pitt CMLD

Inverse Molecular Design Team at Duke



Prof. David Beratan



Shahar Keinen



Nan Jiang



Xiangqian Hu



Prof. Mingliang Wang
Shenzhen Univ



Prof. M. Therien



Aaron Virshup



Julia Conteras-Garcia



Dequan Xiao
Postdoc at Yale

Ongoing Collaboration with **Univ. of Pitt** on **Library Design**



Peter Wipf



Kay Brummond

C. Riderspacher



Inverse Molecular Design Team at Duke



Prof. David Beratan

Xiangqian Hu

Ongoing Collaboration with **Univ. of Pitt** on Library Design



Peter Wipf

Two Directions of Inquiry

1. **The Forward Study:** Specify a system or model and investigate its behavior (energies, dynamics, linear and nonlinear responses, free energies, phase diagrams,...).
 - Rigorous classical, quantum, and statistical mechanics foundation.
 - Rich tradition in theoretical and computational development.
 - Most studies are of this type.

Two Directions of Inquiry

As “forward” theories and computer technology advance, we can begin to address ...

- 2. The Inverse Design:** Specify a property, and design/search a system that optimizes that property.
- Clearly very important.
 - Previous examples: laser pulses (Yijing Yan, Rabitz), protein sequence, drug design
 - Bad News: Challenging.
 - Good News: Encouraging development, could be most useful.

The Aim of Inverse Design

Specify a property, and design/search a system that optimizes that property.

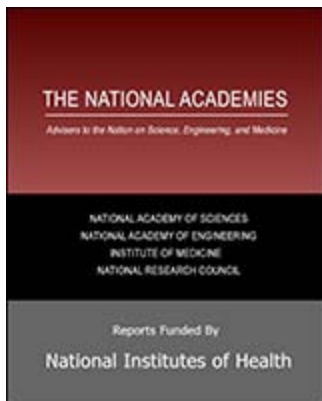
- General
- Efficient

Challenges in Inverse Design

- The properties, as objects of optimization, are very diverse.
- The variables of design now are in the vast chemical/material or biological space.
- The space is discrete.
- Direct enumeration/combinatory approach is severe limited.

On the size of chemical space

- Reymond and co-workers enumerated a virtual library of all organic compounds (within certain synthetic constraints) of **13 heavy atoms** or less and composed of H, C, N, O, S and/or Cl(10-12). The GDB13 database contains nearly 1 billion compounds
- Synthetically realizable small-molecule compounds (stable organic molecules of 500 Daltons or less) have been variously estimated to number between 10^{30} and 10^{160} .



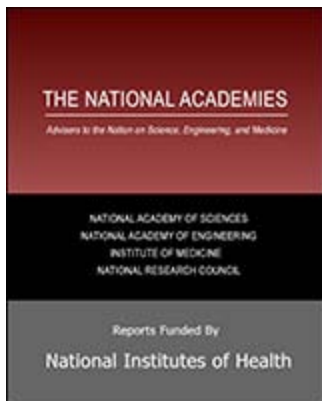
Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology

NAS, 1999

Vision 2020: Computational Needs of the Chemical Industry

Computational "Grand Challenges" for Materials and Process Design in the Chemical Enterprise

- A. Reliable prediction of biological activity from chemical structure
- B. Reliable prediction of environmental fate from chemical structure
- C. Design of efficient catalysts for chemical processes
- D. Design of efficient processes in chemical plants from an understanding of microscopic molecular behavior
- E. Design of a material with a given set of target properties



Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology

NAS, 1999

Vision 2020: Computational Needs of the Chemical Industry

“Grand Challenge E in Table 1 is extremely difficult ...

... “Holy Grail” of materials design ... to solve the problem of ***going backwards from a set of desired properties to realistic chemical structures*** ...

These efforts ... have, so far, only had limited success. Much work needs to be done ...”

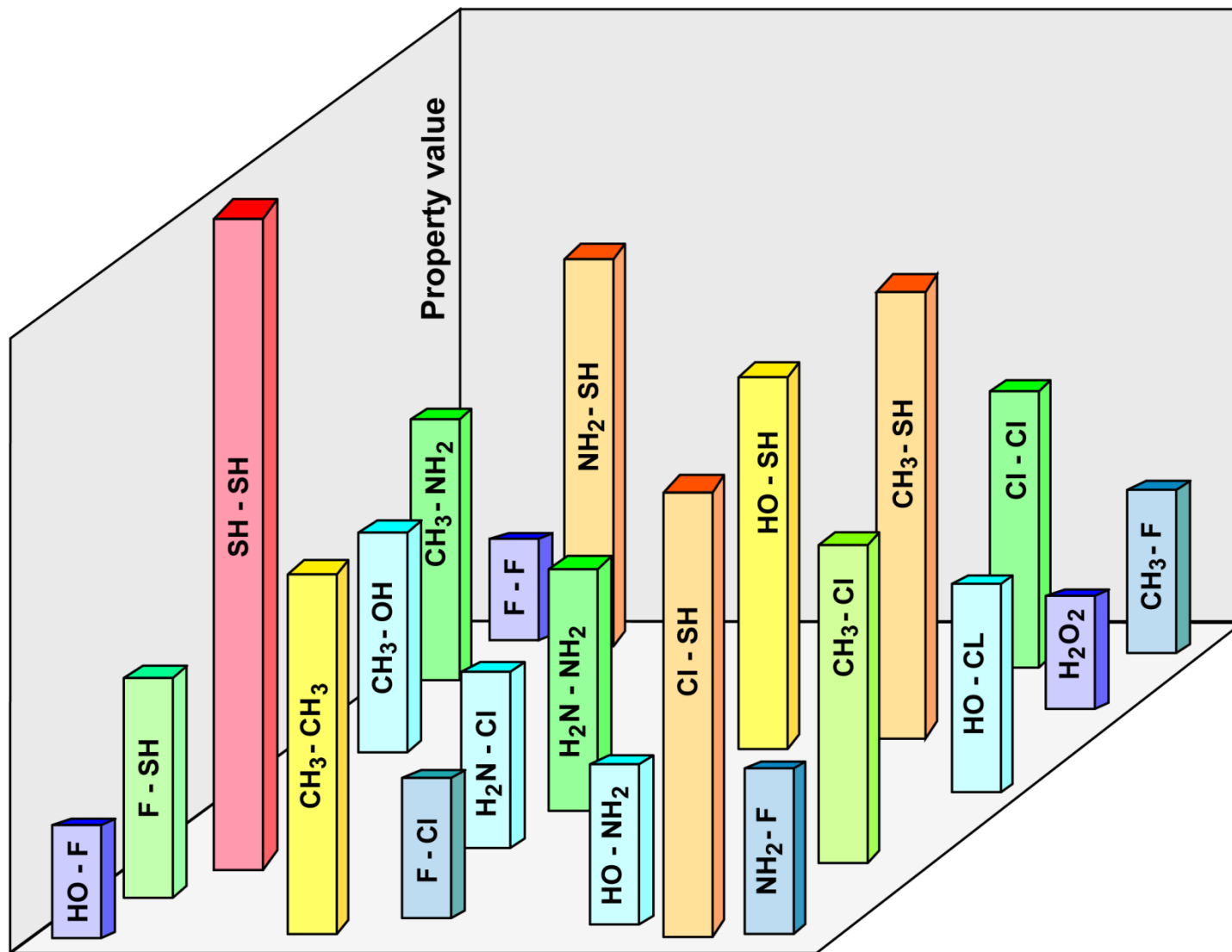
How can we explore molecular space to achieve optimization of the target property?

--the inverse design challenge

Outline: Designing Molecules/Materials

- The Discreteness of Chemical Structures
- Ideas for Navigating Chemical Space
- Linear Combination of Atomical Potentials (LCAP)
- LCAP in DFT & Semiempirical QM frameworks:
Examples
- Gradient-Directed Monte Carlo approach for molecular design -- when the surface is rugged

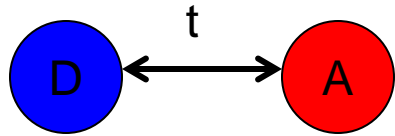
Discrete Molecular Space



Key Questions

- Can a continuous property surface be established?
(continuous optimization is much more efficient)
- How rugged are property surfaces as a function of structure?

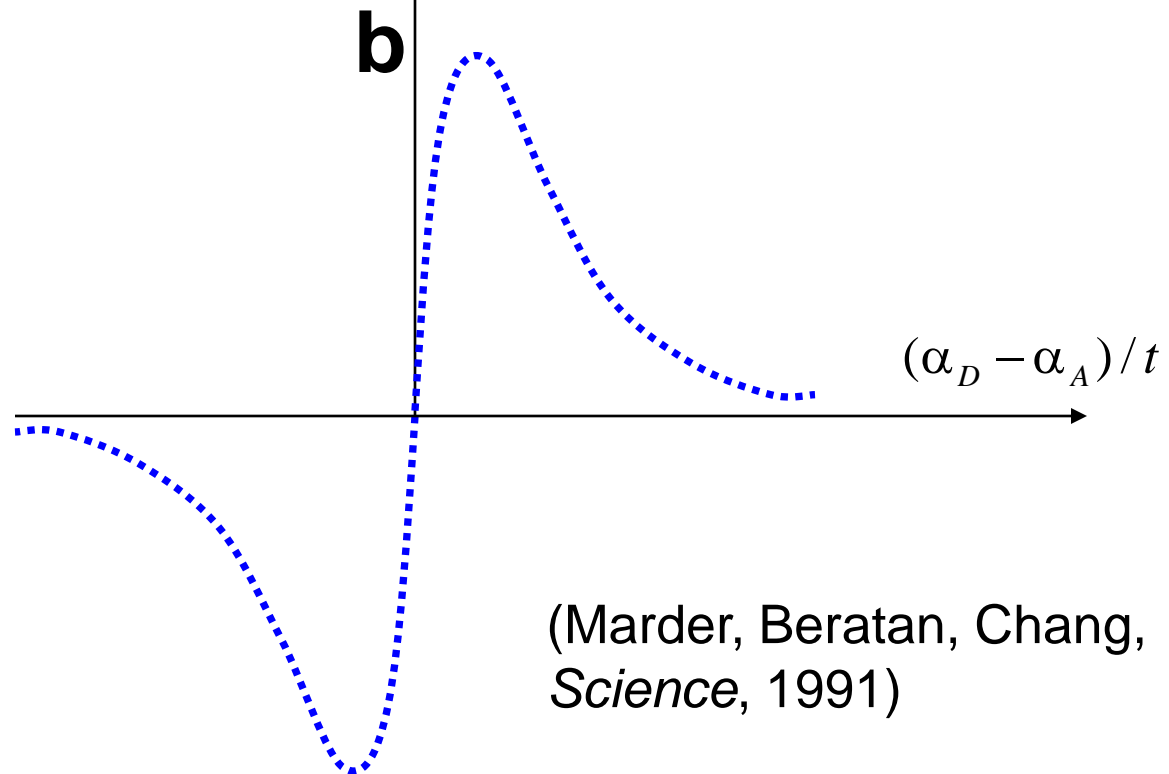
Earlier idea: explore the link between the Hamiltonian & the Property



$$H = \begin{bmatrix} \alpha_D & t \\ t & \alpha_A \end{bmatrix}$$

$$\beta \sim \frac{\mu_{ge}^2 \Delta \mu_{ge}}{\Delta E_{ge}}$$

Property varies smoothly with Hamiltonian, suggesting design strategies.



(Marder, Beratan, Chang, *Science*, 1991)

New Idea: Focus on the potential

Observations:

- atom types and positions define $V(\mathbf{r})$ uniquely
- $V(\mathbf{r})$ (and N) determines **everything!**
- Hohenberg-Kohn ('64): $\rho(\mathbf{r}) \square V(\mathbf{r})$
- Yang, Ayers and Wu ('02-'04): Potential functionals – using $V(\mathbf{r})$ as the basic variable
 1. in **DFT formulation** for solving the v -representability problem
 2. in **DFT computation** for the Optimized Effective Potential (OEP).

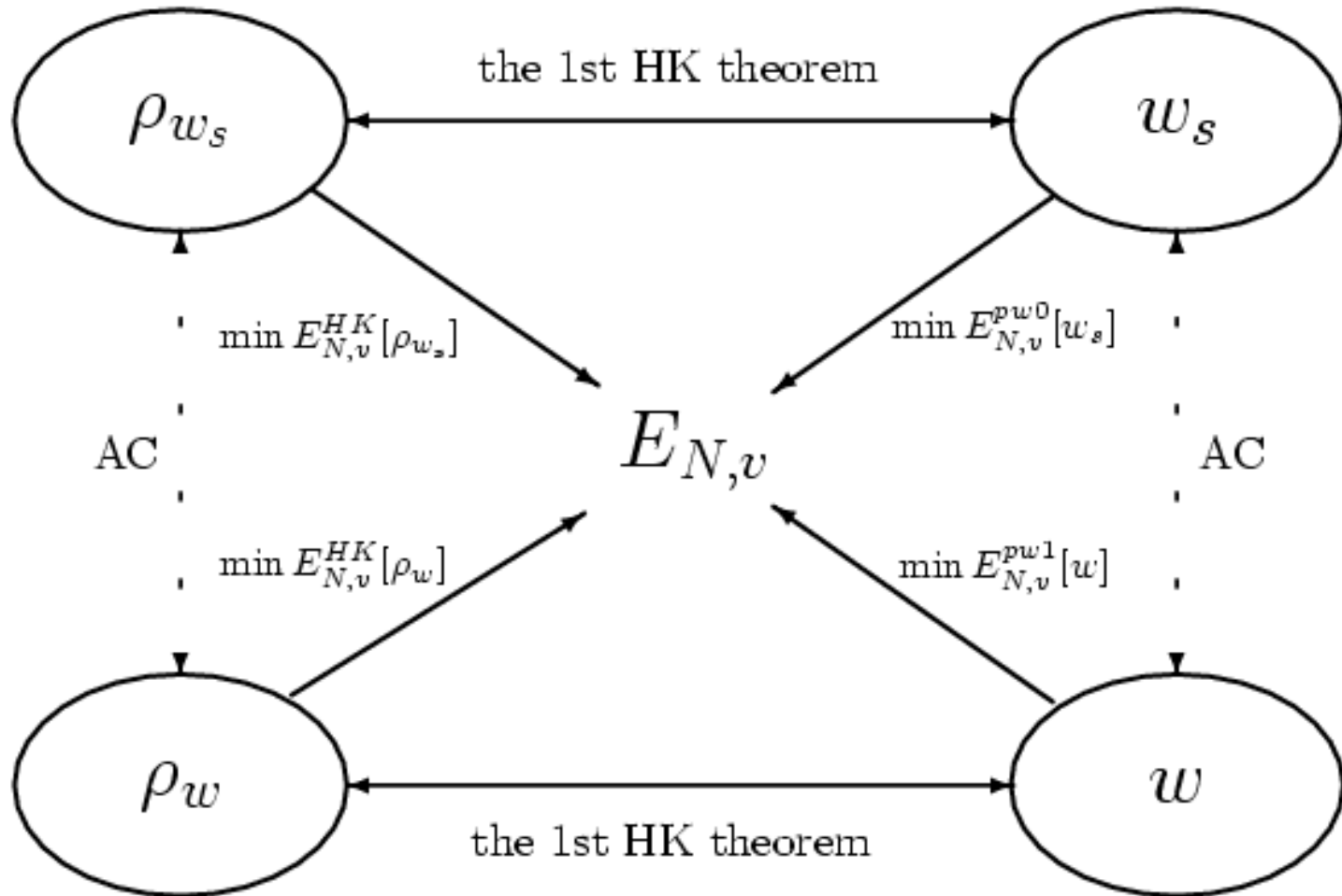
$v(\mathbf{r})$ is the dual of $\rho(\mathbf{r})$

the function	its dual	the linear functional
$ \Psi\rangle$	$\langle\Psi $	$\langle\Psi \Psi\rangle = \int d\mathbf{r}\Psi^*(\mathbf{r})\Psi(\mathbf{r}) < \infty$
$\rho(\mathbf{r})$	$v(\mathbf{r})$	$\int d\mathbf{r}\rho(\mathbf{r})v(\mathbf{r}) < \infty$

Density Functional vs. Potential Functional

Yang, Ayers and Wu , PRL 04

Non-interacting
Kohn-Sham



Physical
Hohenberg
-Kohn

Summary for the potential functionals (Yang, Ayers and Wu, PRL 2004)

- **Solution to the v -representability:** In the potential space, the v -representability is no longer a constraint.
- **Theoretical foundation for OEP:** The variational principle in terms of the Kohn-Sham potential (Yang & Wu, PRL 02)

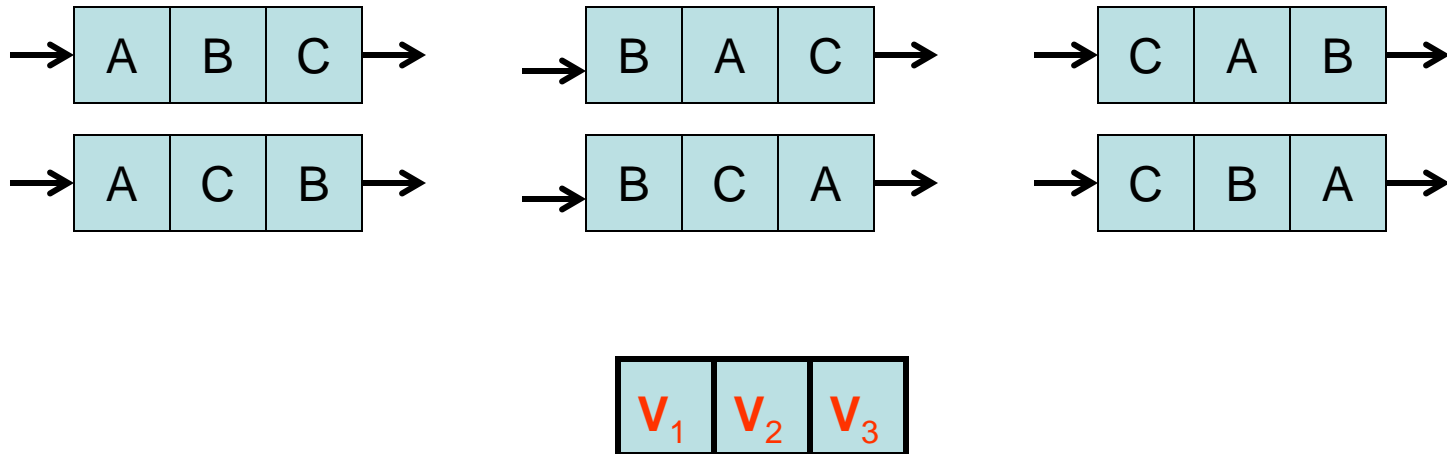
The target: $E[\{\psi_i\}]$

The variable: $v_s^\sigma(\mathbf{r})$

$$v_s^\sigma(\mathbf{r}) = v_{ext}(\mathbf{r}) + v_o(\mathbf{r}) + \sum_t b_t^\sigma g_t(\mathbf{r})$$

Another advantage: Structure vs. $V(r)$ Complexity

- Molecular complexity grows as $N!$
- Complexity of “external potential” $v(r)$ grows only as N :



Plan

- Start with a target property.
- Find an appropriate method to compute the property from a given $v(\mathbf{r})$.
- Use $v(\mathbf{r})$ as the optimization variable.

The electron-nucleus attraction potential

$$v(\mathbf{r}) = - \sum_A \frac{Z_A e^2}{|\mathbf{r} - \mathbf{R}_A|}$$

A Challenge

- any molecule can be expressed by a $v(\mathbf{r})$
- not all potentials come from a molecule, or Chemistry-representable (C-representable).

The Electron-Nucleus Attraction

$$v(\mathbf{r}) = -\sum_A \frac{Z_A e^2}{|\mathbf{r} - \mathbf{R}_A|}$$

Our Strategy

--Define a continuous potential function that smoothly interpolates among atom/group types.

LCAP: Linear Combination of Atomic Potentials

$$V(\mathbf{r}) = \sum_{R,g} C_{R,g} v_{R,g}(\mathbf{r})$$

The diagram illustrates the equation $V(\mathbf{r}) = \sum_{R,g} C_{R,g} v_{R,g}(\mathbf{r})$ with three explanatory boxes below it. An arrow points from the 'Site index' box to the R in the summation index. Another arrow points from the 'Group index' box to the g in the summation index. A third arrow points from the 'Coulomb's law e/n attraction for integer Z' box to the $v_{R,g}(\mathbf{r})$ term.

Site index

Group index

Coulomb's law e/n attraction for integer Z

--Seek the optimal molecule by optimizing coefficients in the potential.

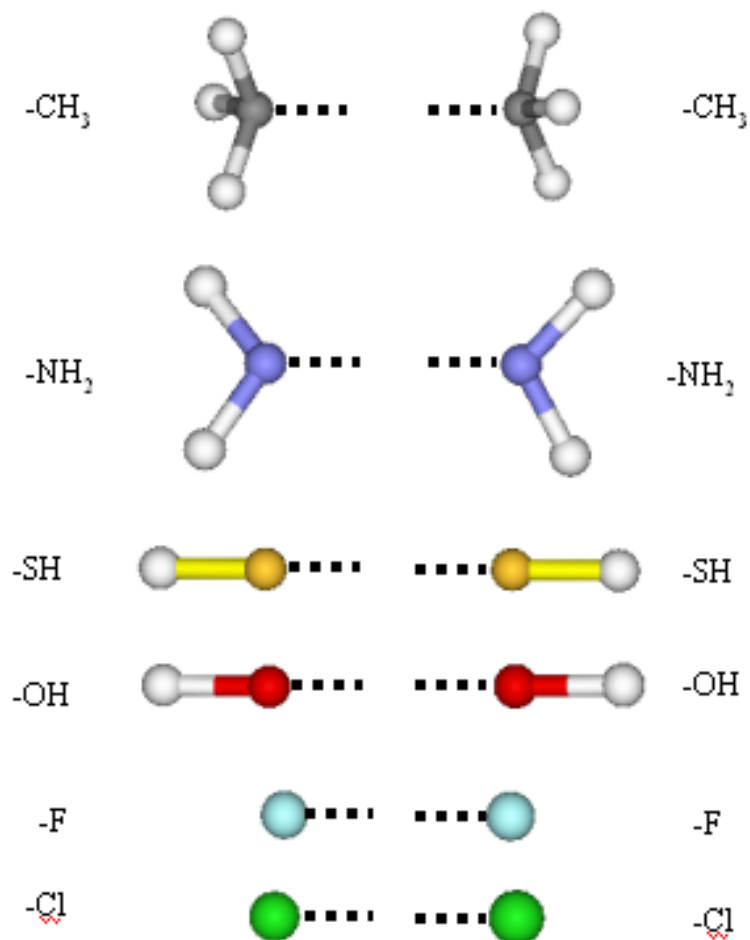
L.C.A.P.

$$V(\mathbf{r}) = \sum_{R,g} C_{R,g} V_{R,g}(\mathbf{r})$$

- Remember LCAO?
- Directly linked to chemistry/molecules
- Properties optimized by varying the C's
- Follow property gradients to optimize C's

A few examples

- Choose a designable framework.
- At each designable site, a potential is constructed from a LCAP.
- Optimize the property with respect to all of the coefficients in the LCAP.



Ultrasoft pseudopotential - plane-wave DFT scheme

- Plane-wave basis set independent of atom types or positions
- No basis set superposition error
- Orthogonal basis set
- USPPs for fewer basis functions

Example: Polarizability (α) optimization

- Finite field

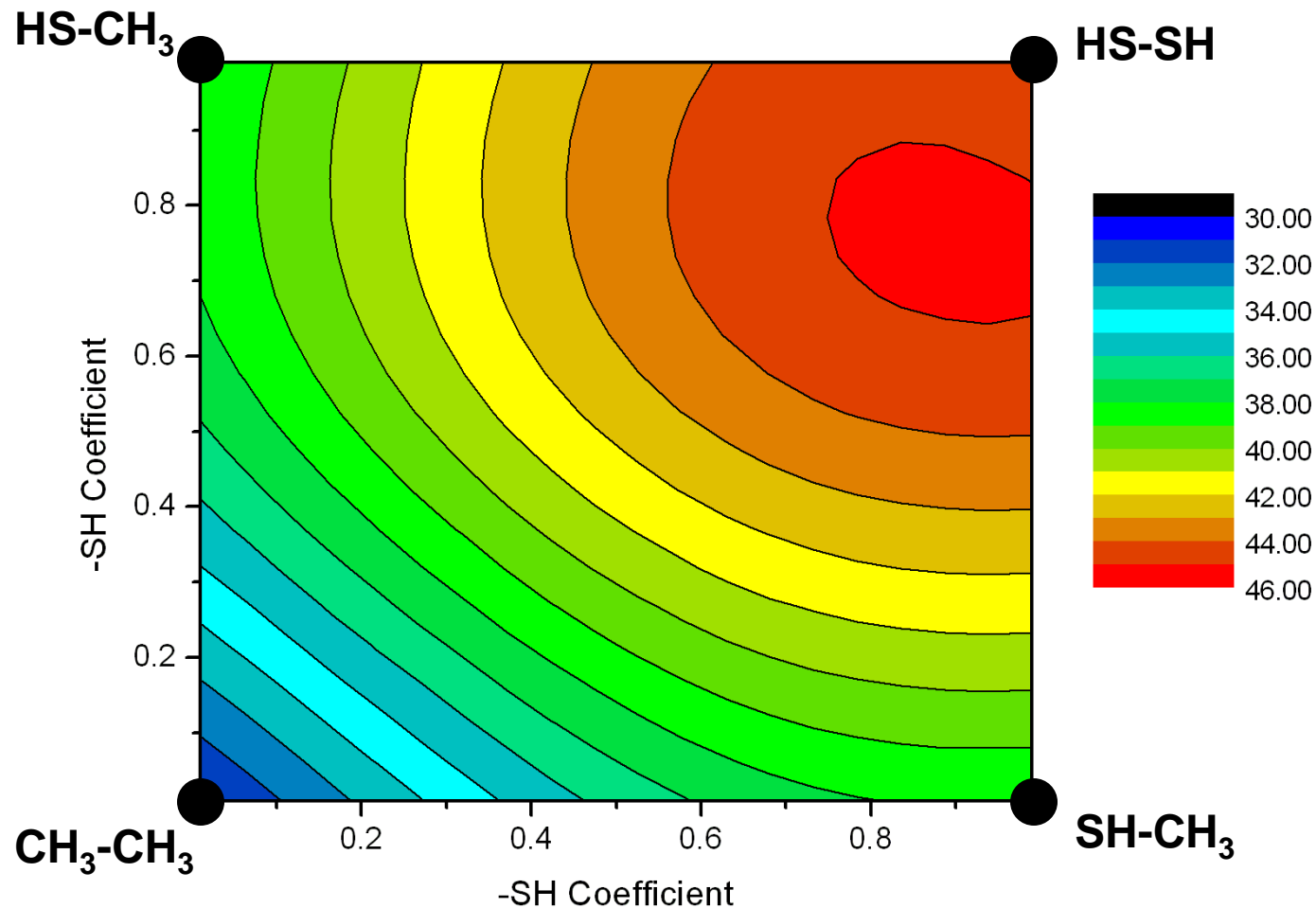
$$\alpha_{ij} = -\frac{1}{F^2} [E_{tot}(+F_i) + E_{tot}(-F_i) - 2E(0)]$$

- Derivative of polarizability with respect to $C_{R,g}$

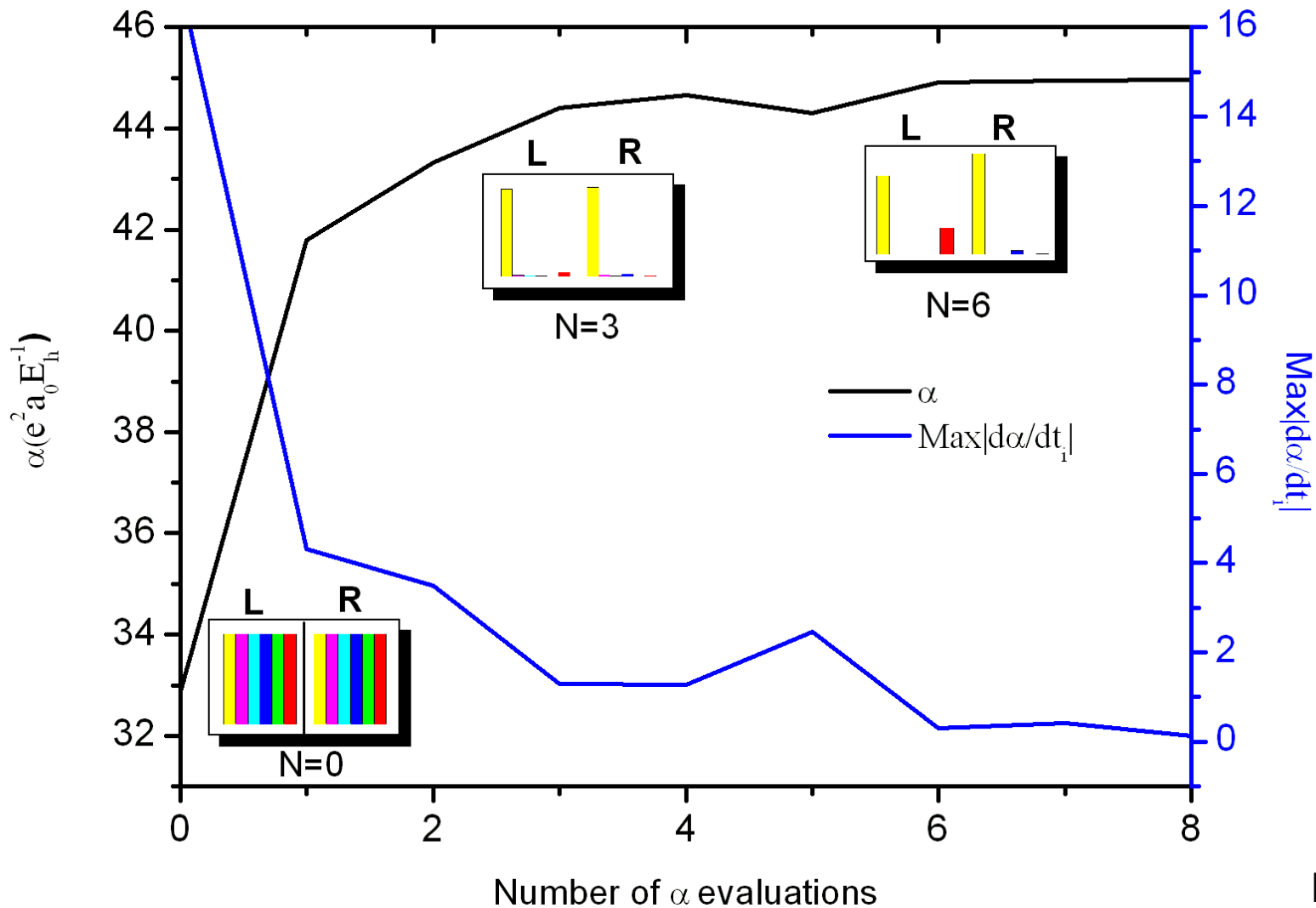
$$\frac{\partial \alpha}{\partial C_{R,g}} = -\frac{1}{F^2} \left[\frac{\partial E(+F)}{\partial C_{R,g}} + \frac{\partial E(-F)}{\partial C_{R,g}} - 2 \frac{\partial E(0)}{\partial C_{R,g}} \right]$$

- Use analytical energy derivative in the optimization

Two designable sites and two functional groups (-CH₃, -SH): Electronic Polarizability

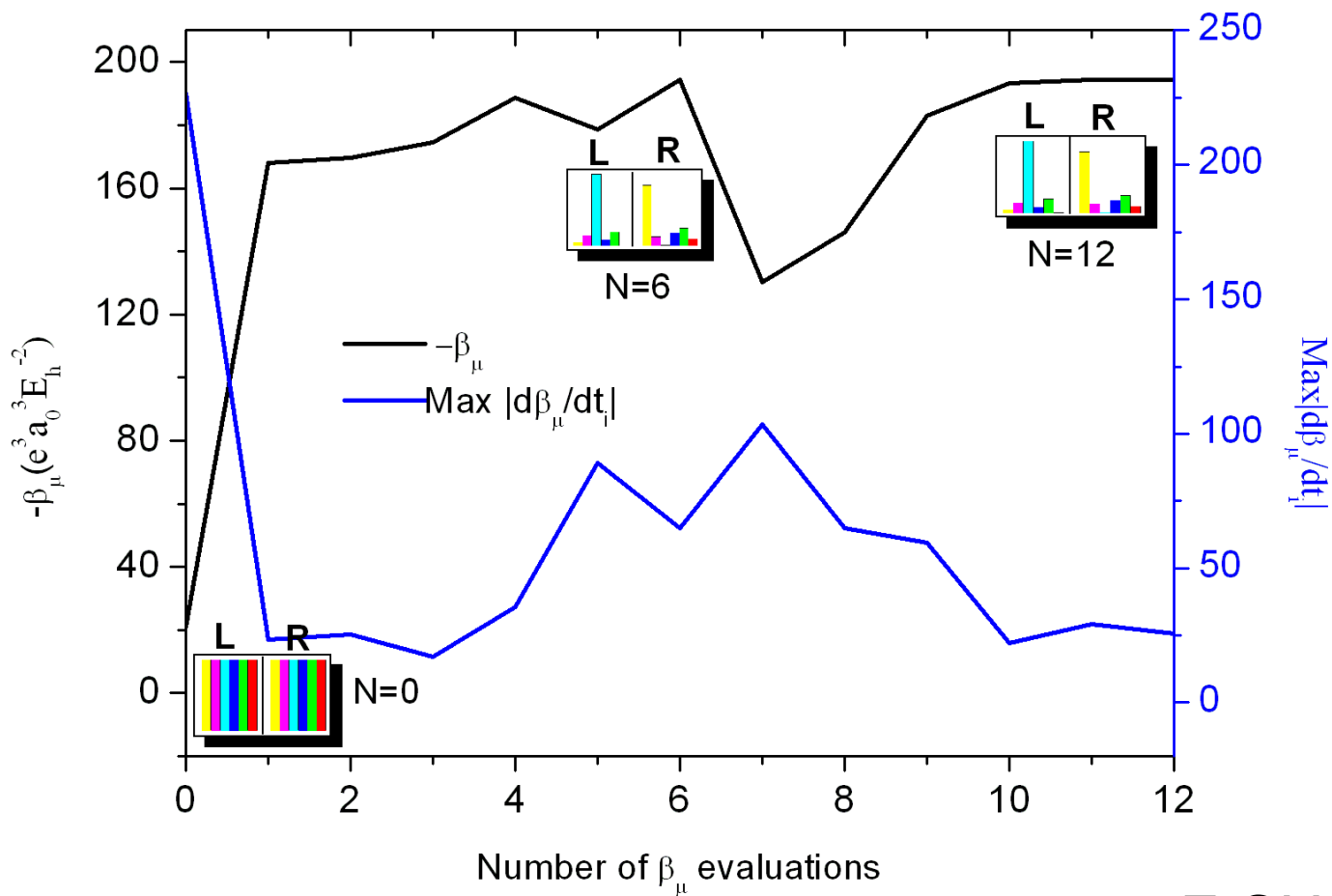


X-Y with six choices per site (21 unique structures). Optimization is complete in a few steps.



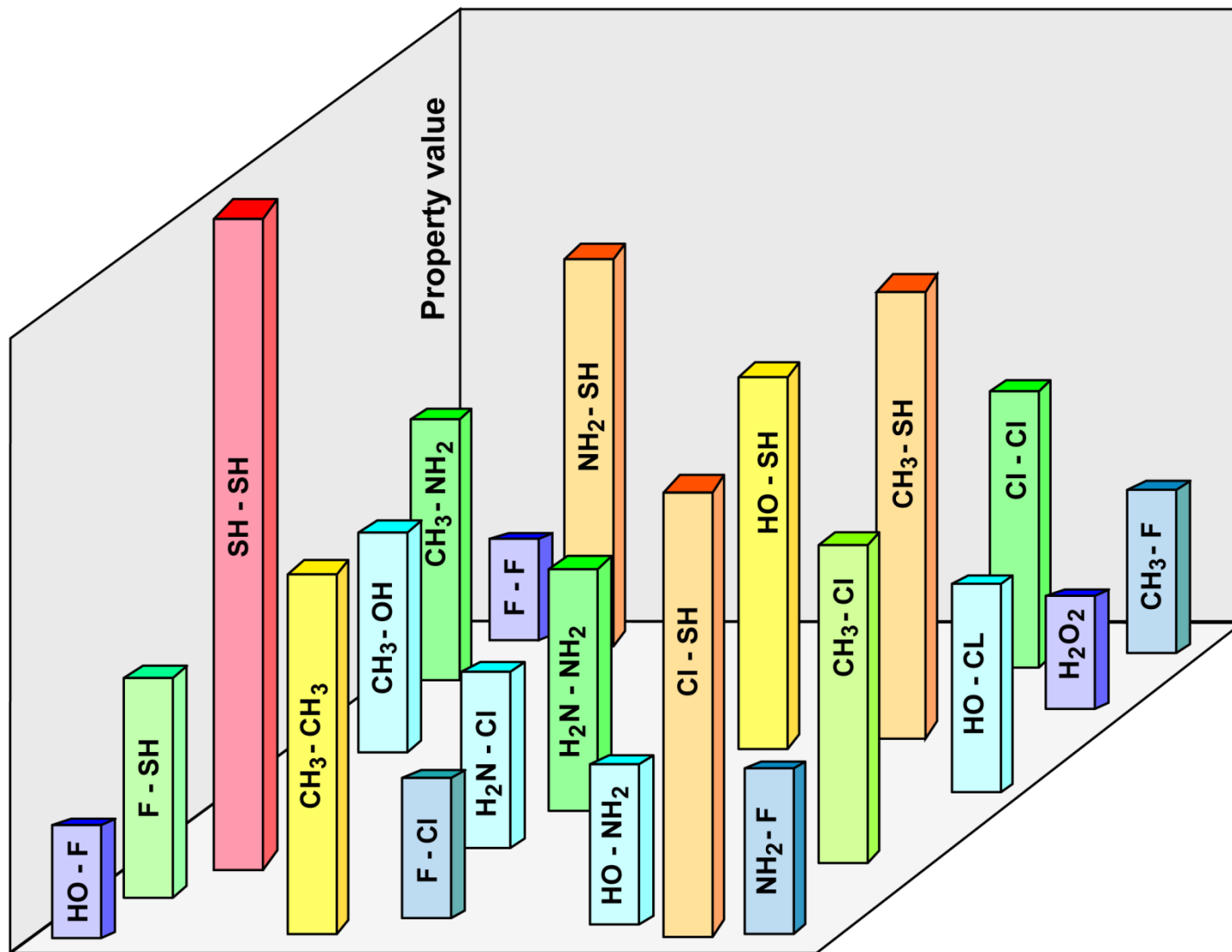
HS-SH

... and for β

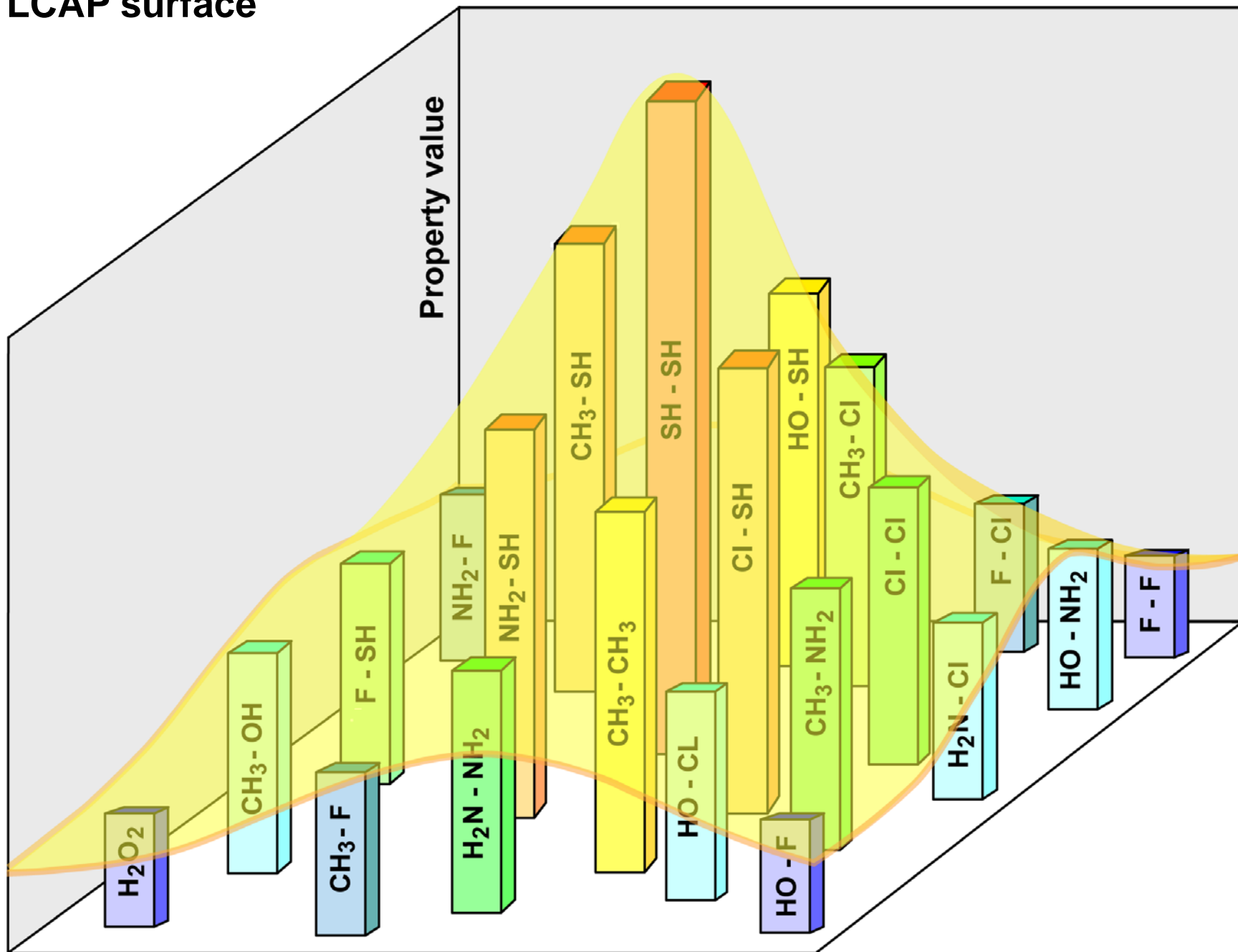


F-SH

Discrete Molecular Space



LCAP surface



π -electron (Hückel)-LCAP



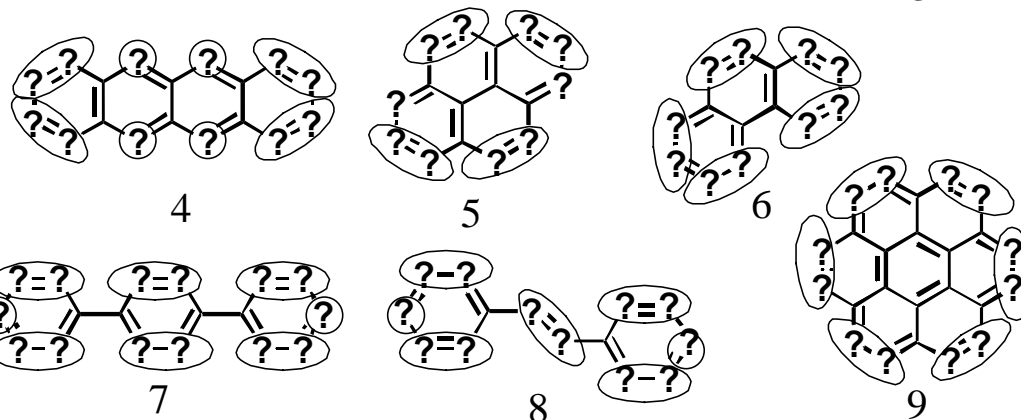
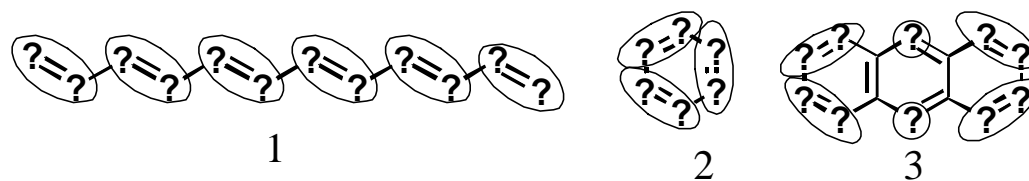
Dequan Xiao

Site energy:
$$H_{ii} = \sum_{\mu} \lambda_{i\mu} h_{\mu}$$

Interaction energy:
$$H_{ij} = \sum_{\nu} \sum_{\mu} \lambda_{i\mu} \lambda_{j\nu} t_{\mu\nu}$$

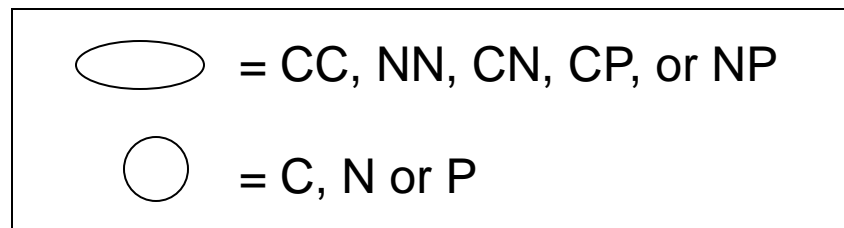
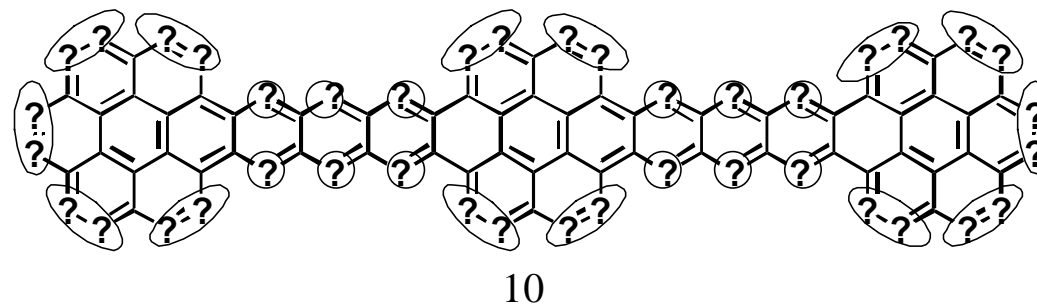
Constraint:
$$\sum_{\mu} \lambda_{i\mu} = 1$$

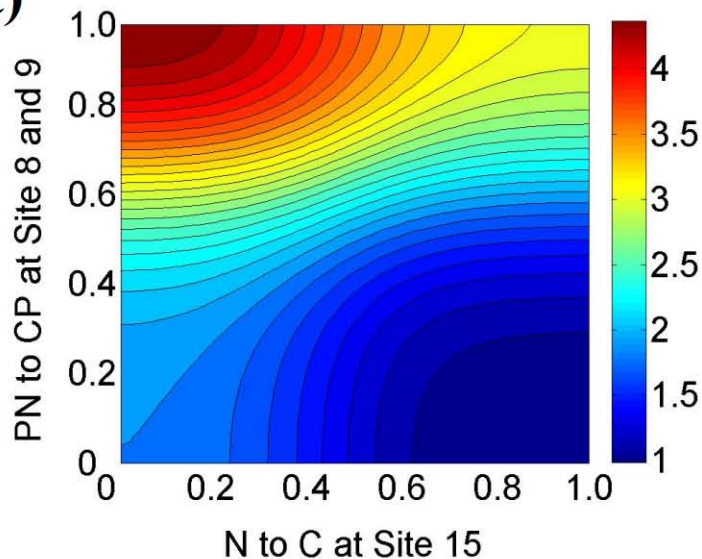
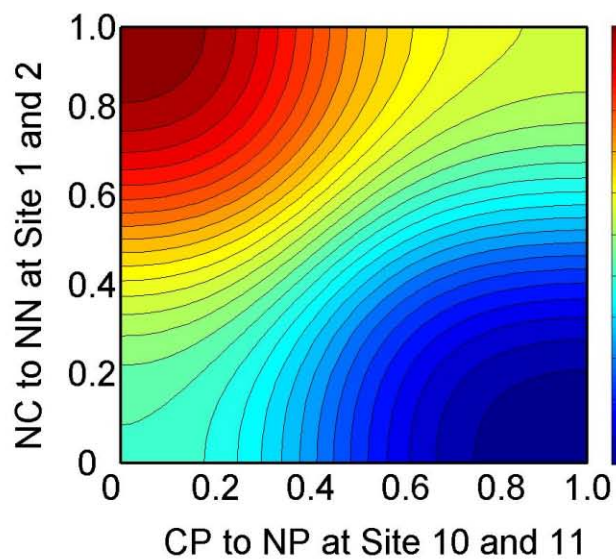
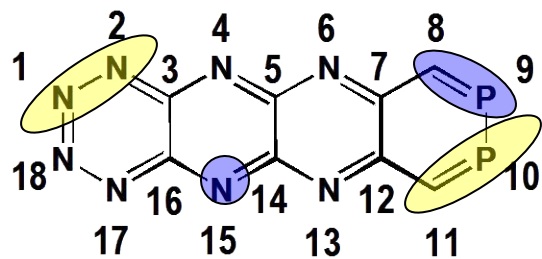
Some Designable π -frameworks



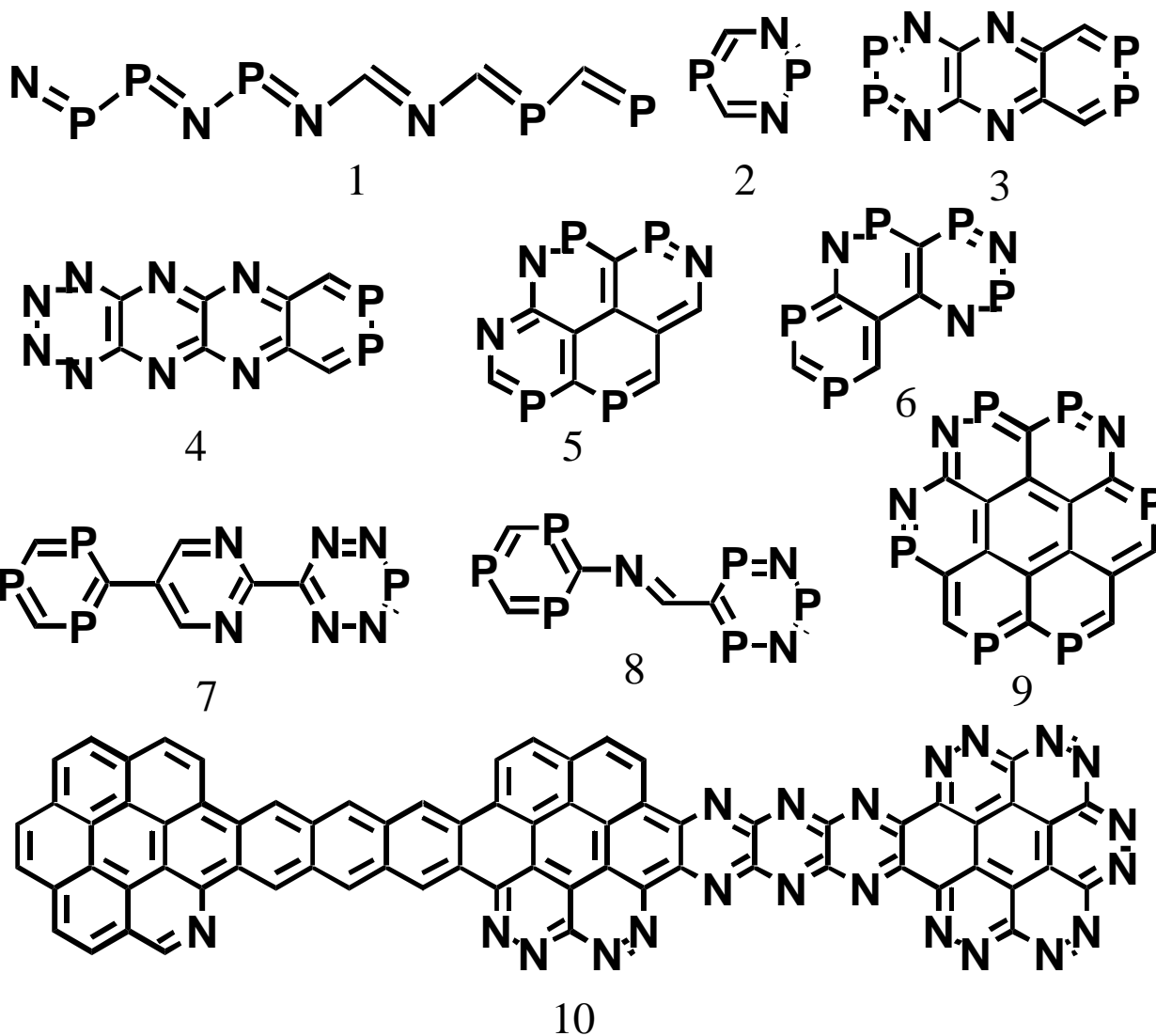
$\sim 10^5$ molecules

$\sim 10^{16}$ molecules





Optimized structures for p-contribution to beta



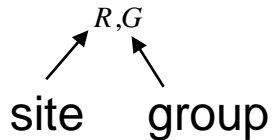
LCAP in a Semiempirical (AM1) Framework



Shahar Keinen

- Definition of alchemical potential

$$V(r) = \sum_{R,G} \lambda_{R,G} V_{R,g}(r) \quad \sum_i \lambda_R = 1$$



 site group

Linear Combination of Atomic Potential (LCAP)

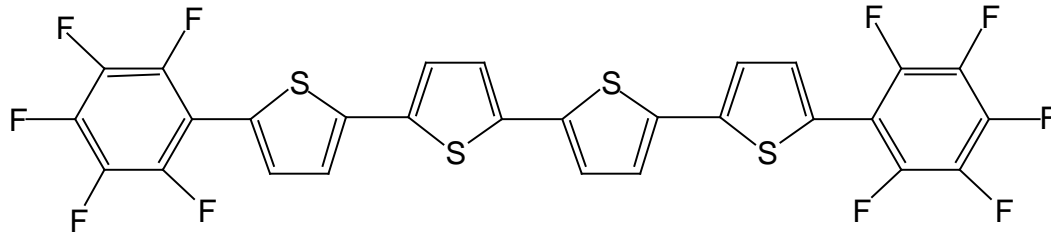
$$\phi_x = \sum_i \lambda_i \phi_i \quad \text{Matching Atomic Orbitals}$$

- e.g., X: F → I

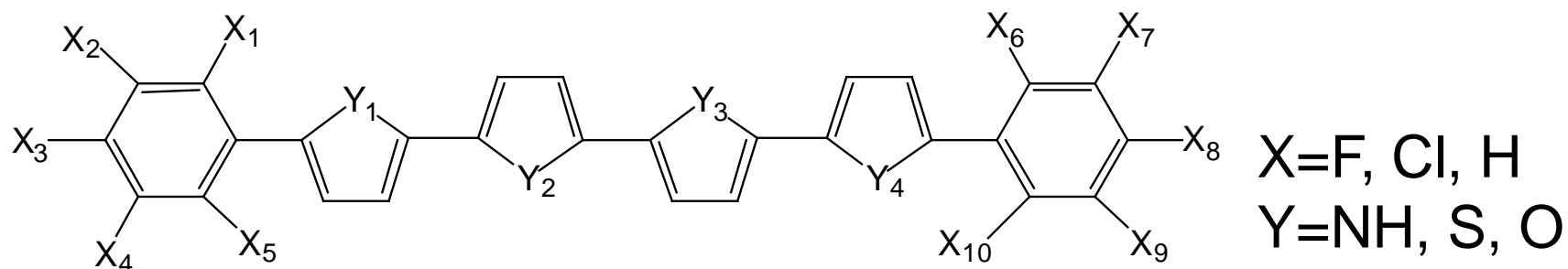
$$Z_x = \lambda_1 Z_F + \lambda_2 Z_I$$

$$\phi_{x,s} = \lambda_1 \phi_{s,F} + \lambda_2 \phi_{s,I}$$

n-type semiconductors

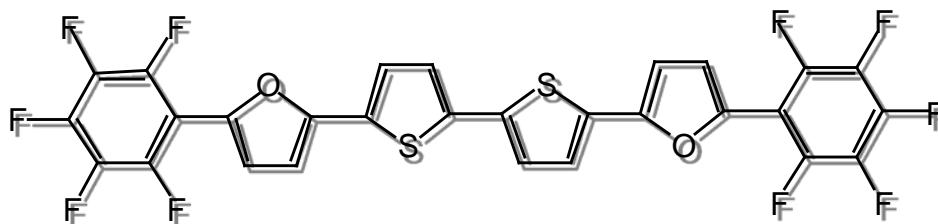


- While searching for p-type semiconductors, the Marks lab discovered this n-type semiconductor
- n-type semiconductors are of interest for “plastic electronics”
- Improved n-type properties correlate with lowered LUMO energy

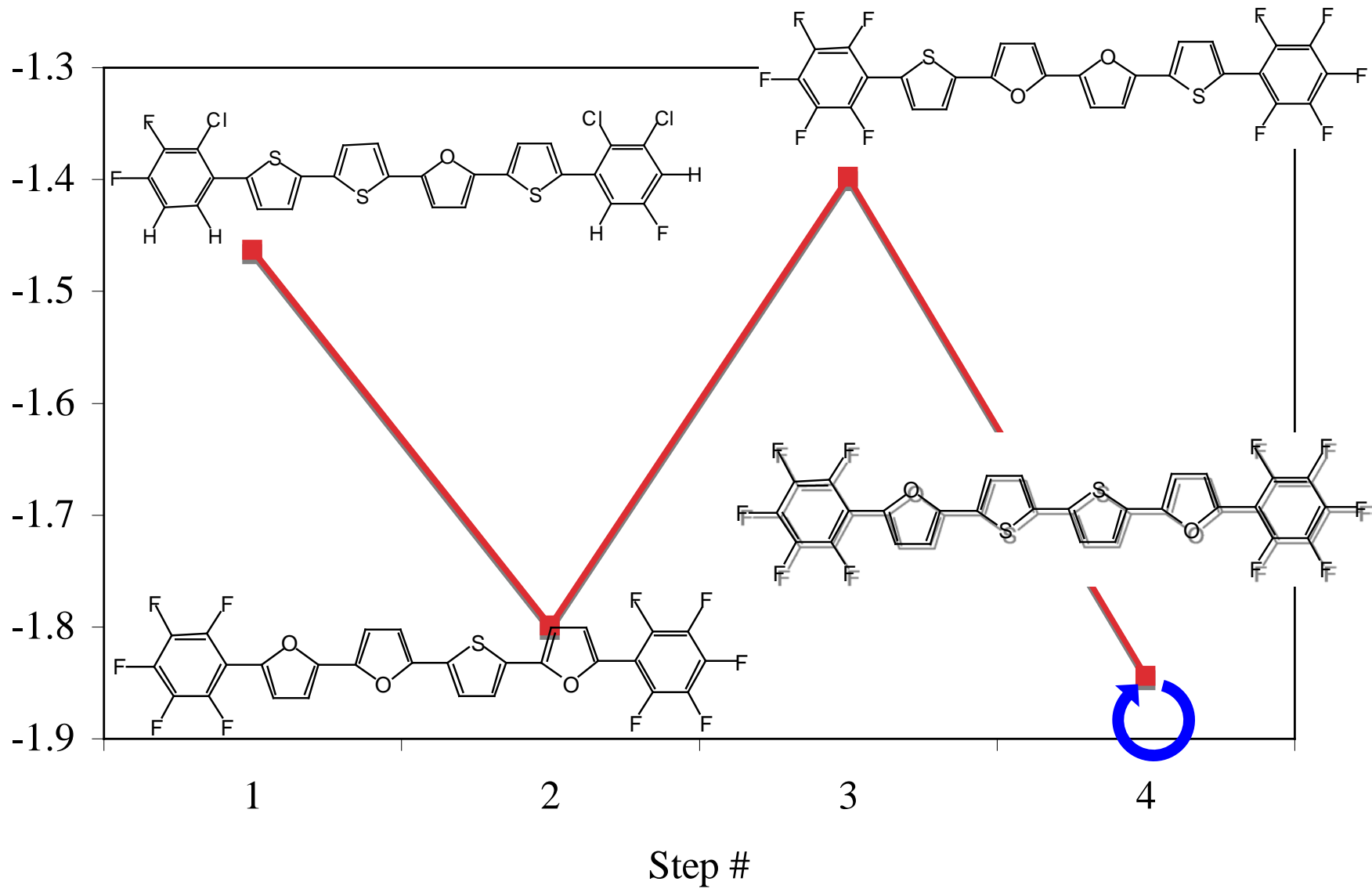


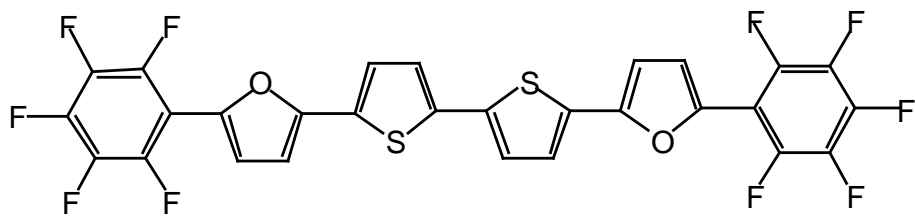
- 4,782,969 possible molecules
- Enumeration not accessible
- Assuming that the molecules are frozen is poor - the molecules twist.
- Use an optimized geometry LCAP search for lowered LUMO
- 19 of 19 searches (all starting from random initial molecule) found this optimized molecule:

$$E_{\text{LUMO}} = -1.844\text{eV}$$

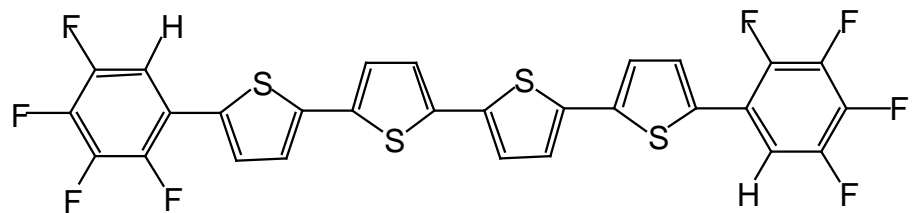


Search profile

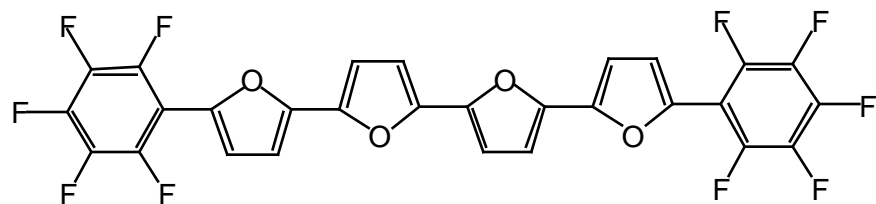




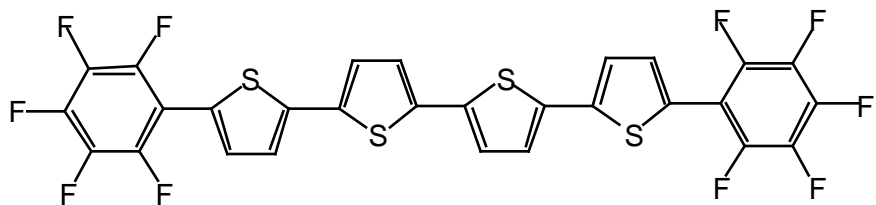
Planar, $E_{\text{LUMO}} = -1.8442\text{eV}$



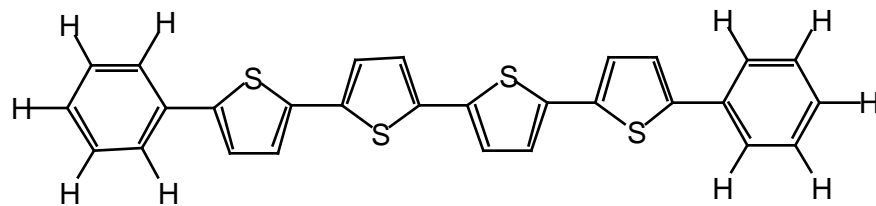
Planar, $E_{\text{LUMO}} = -1.8434\text{eV}$



Planar, $E_{\text{LUMO}} = -1.769\text{eV}$

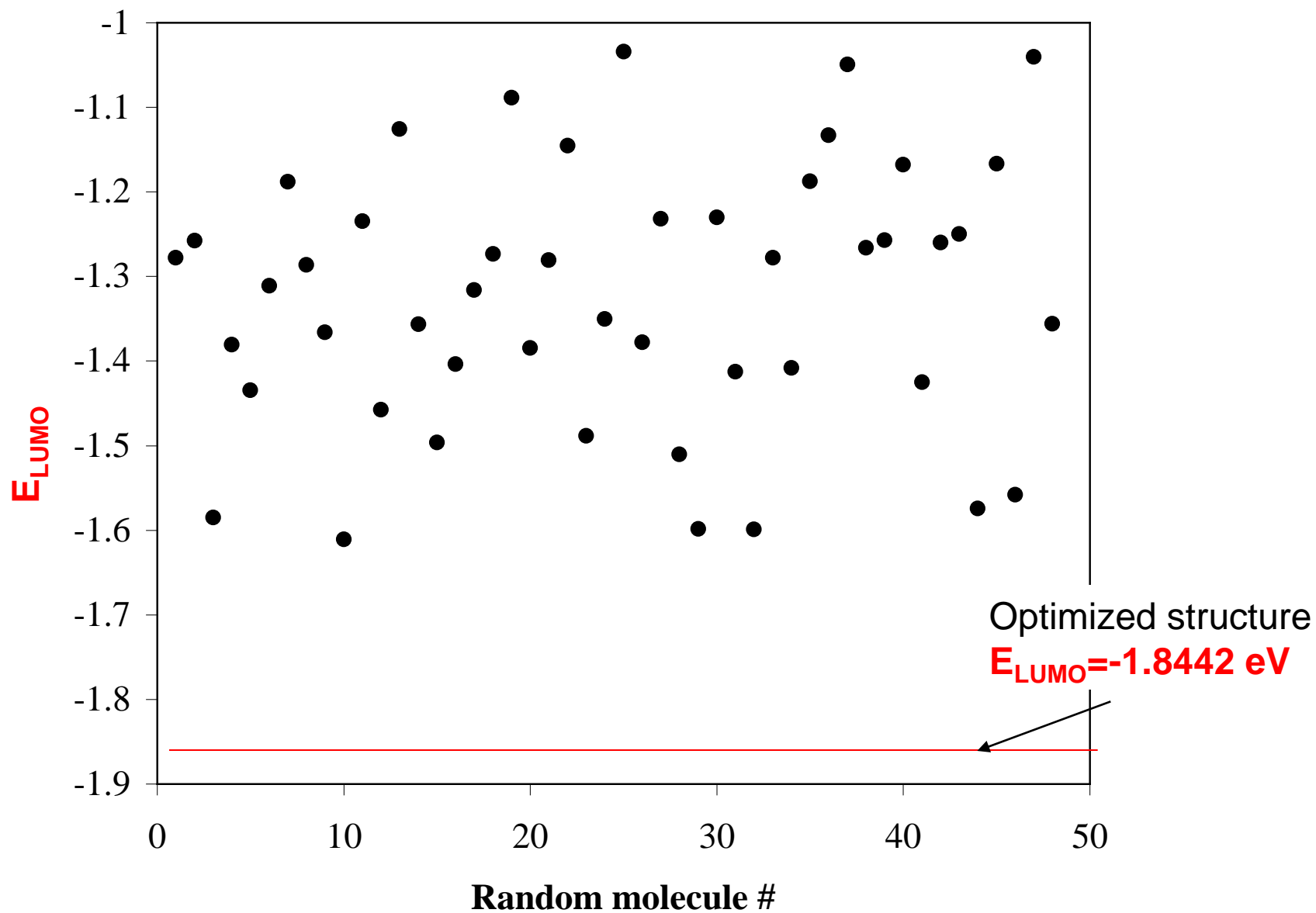


Twisted, $E_{\text{LUMO}} = -1.609\text{eV}$



Planar, $E_{\text{LUMO}} = -1.397\text{eV}$

Randomly chosen structures in this family:

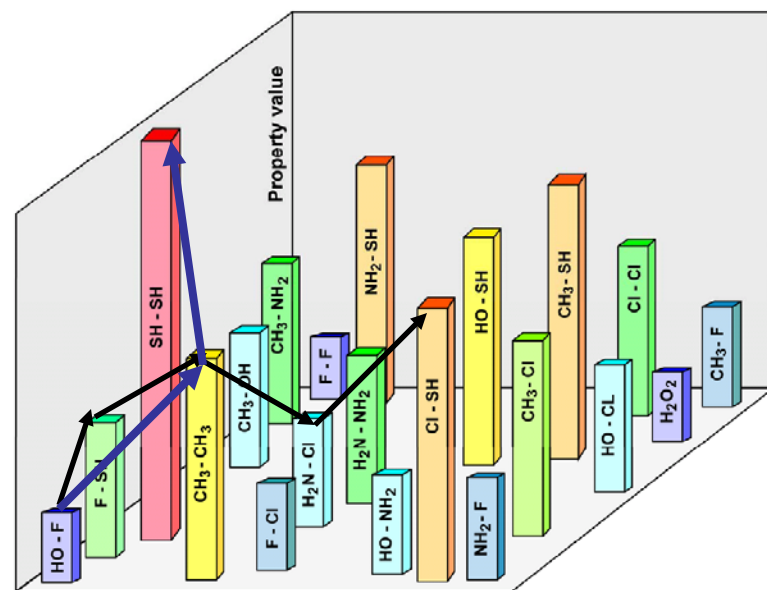


LCAP

- The LCAP approach maps an intrinsically discrete molecular space onto a set of continuous variables, making efficient optimization possible.
- The LCAP approach can be implemented with classical or quantum Hamiltonians; Many kinds of property optimization can be explored with this scheme.
- The LCAP approach appears to provide a promising theoretical framework to address broader challenges in molecular design.
- Continuous optimization schemes do not efficiently explore the LCAP property surface, when it is **rugged**.

Gradient-Directed Monte Carlo (GDMC)

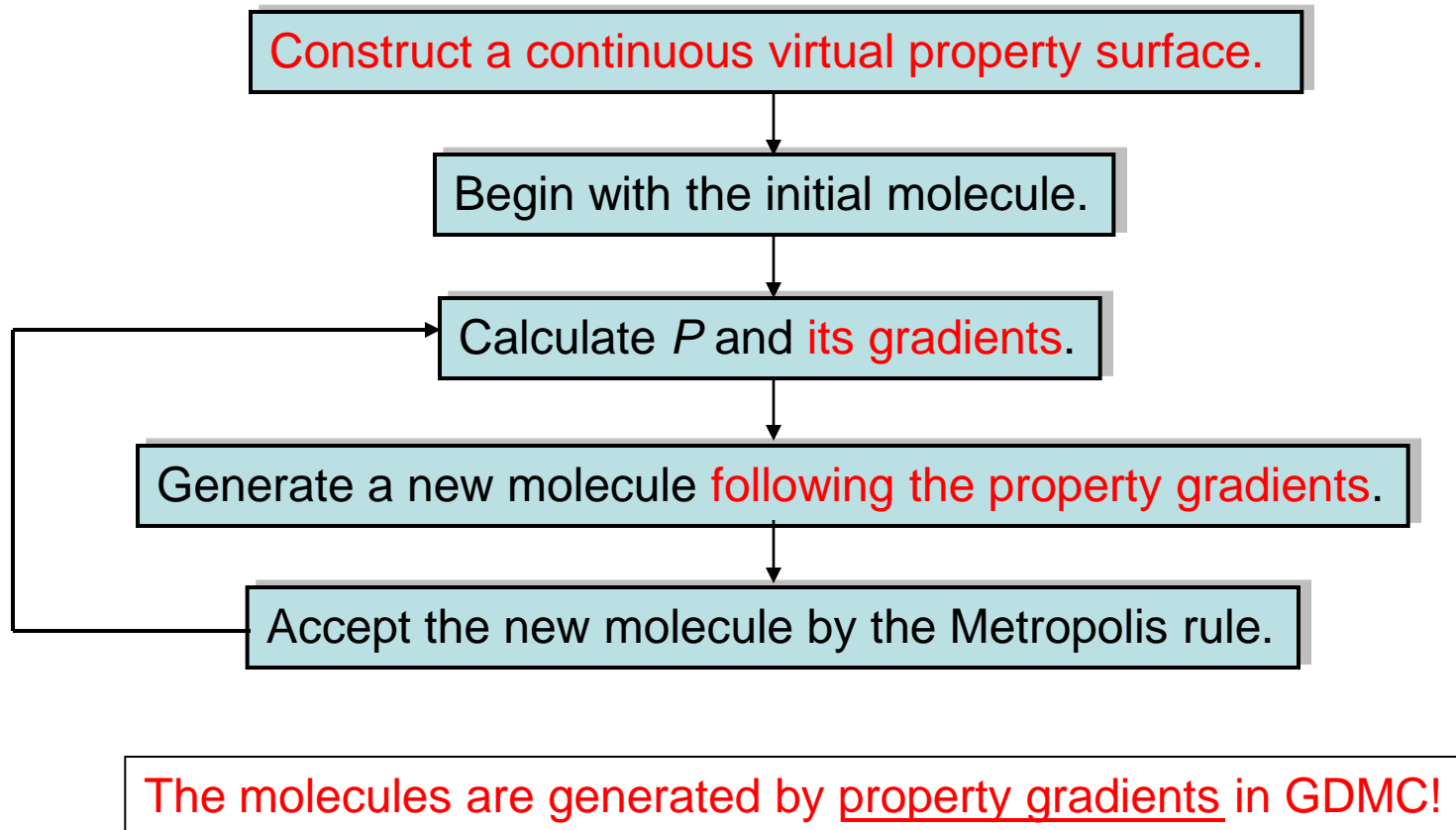
- GDMC uses the property gradients to jump between discrete molecules.
- A new molecule is generated at each step.
- When the search algorithm is trapped in a local optima, random MC moves are helpful to overcome local barriers.
- GDMC saves computational time by not searching “intermediate” states of continuous property surface.



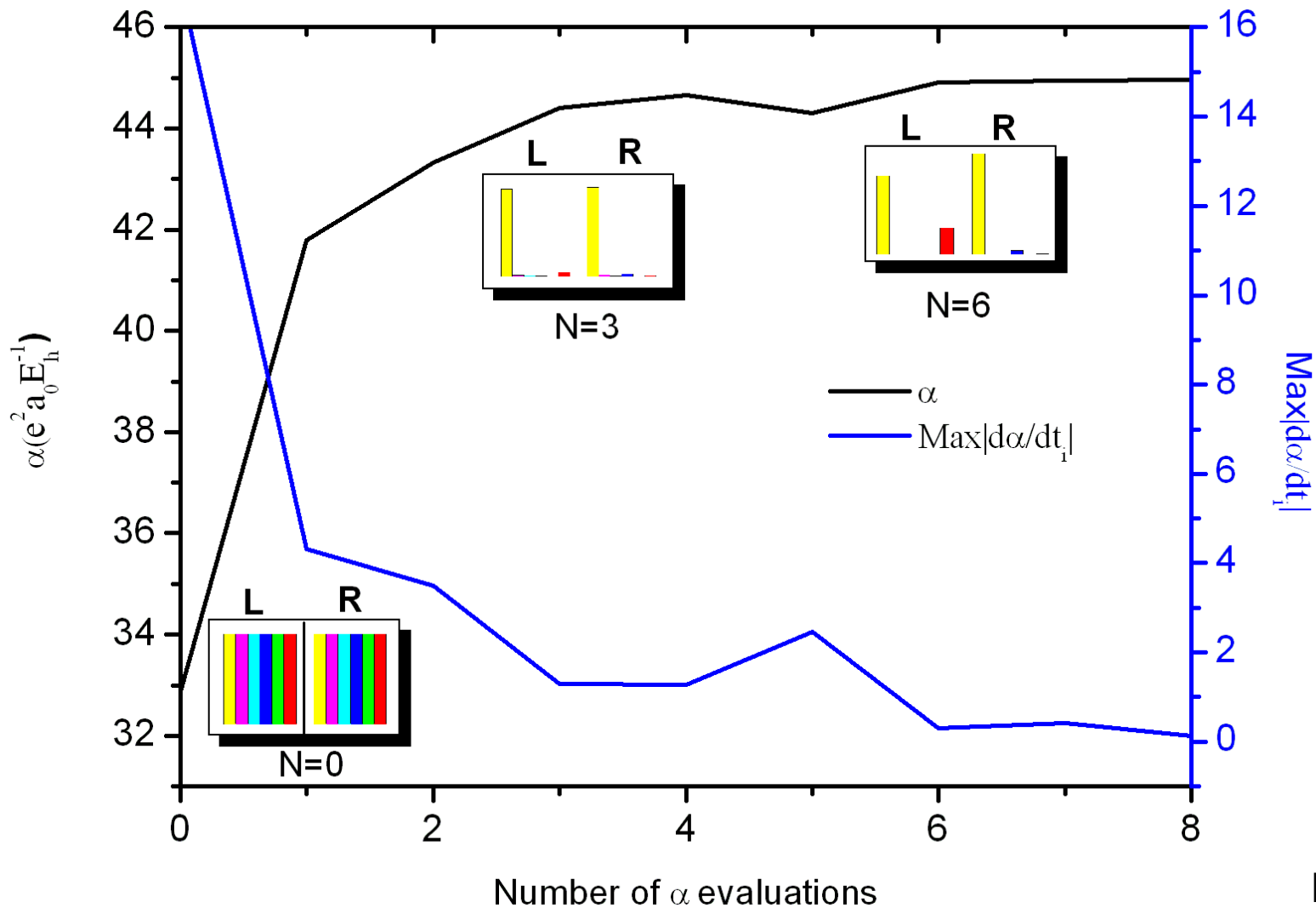
Discrete space
Semi-stochastic search

- X. Hu, D. Beratan, and W. Yang, *JCP*, 2008
- X. Hu, D. Beratan, and W. Yang, *JCP*, 2009

Flowchart of GDMC Procedure



X-Y with six choices per site (21 unique structures). Optimization is complete in a few steps.

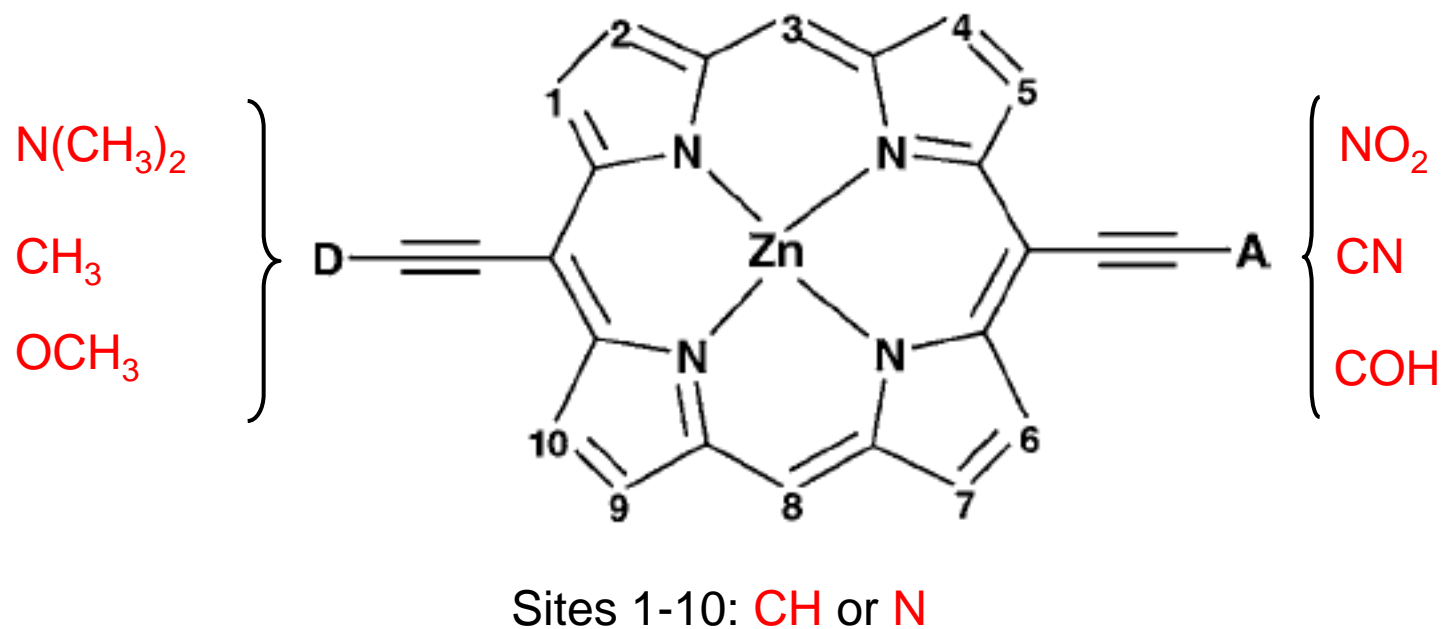


HS-SH

Example 1: Use of GDMC-LCAP to design new NLO molecules

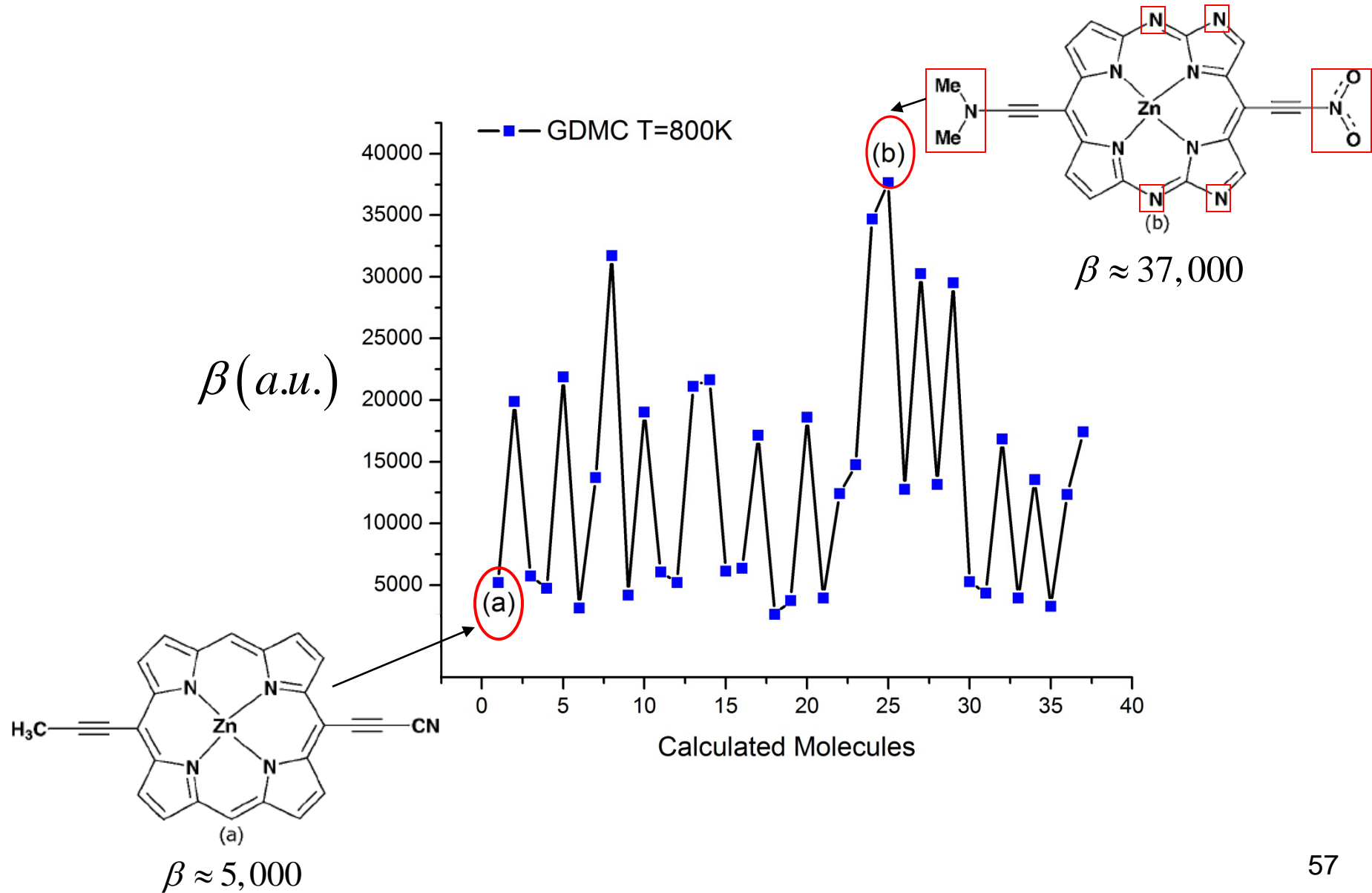
- Nonlinear optical (NLO) molecules are important for optoelectronic materials and devices.
- Electronic first hyperpolarizabilities (β) of molecules determine their nonlinear optical processes.
- Quantum-mechanical approaches are necessary to predict β .
- Porphyrin-based compounds with a donor-bridge-acceptor motif have large β .

Porphyrin-based framework:

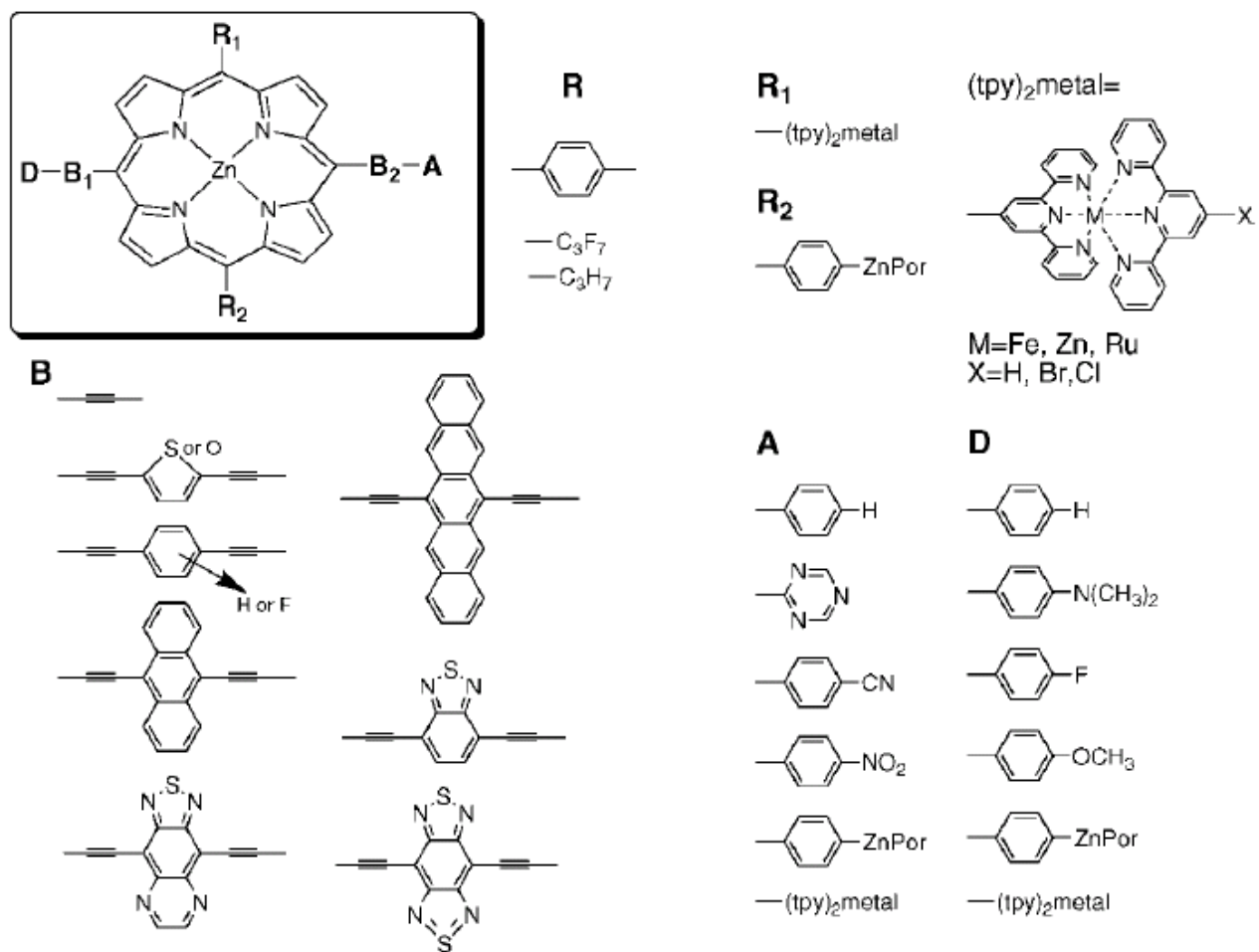


The number of possible molecules is $3 \times 2^{10} \times 3 = 9,216!$

The optimum structure was found after less than 40 molecular calculations!

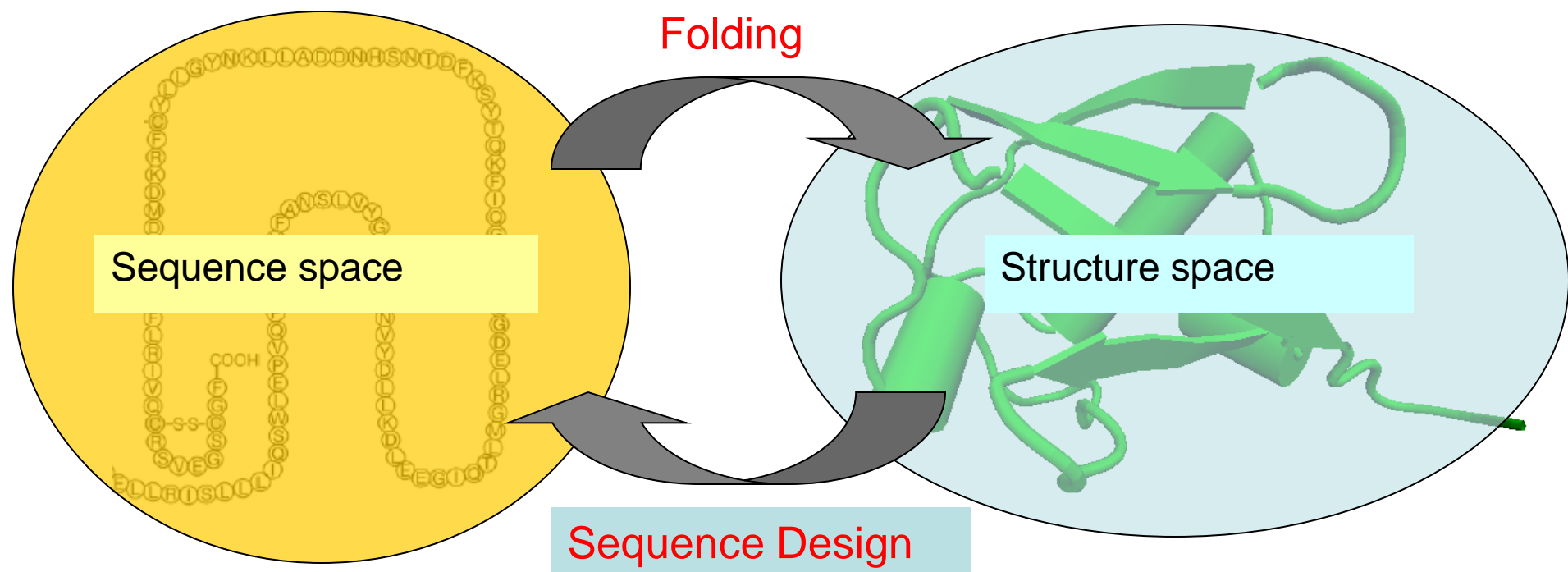


Another application of the GDMC-LCAP method for NLO



The number of possible molecules is 940,800!

Example 2: Use of GDMC for protein design



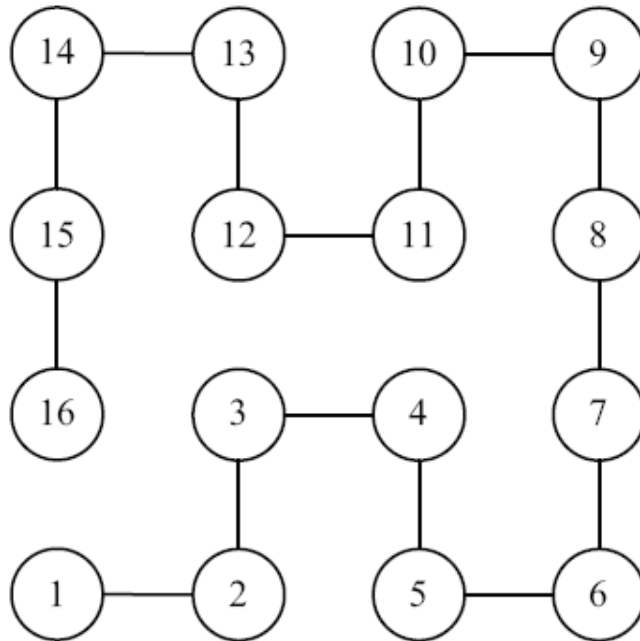
Both spaces are enormous!

Aim: for a given protein conformation, what are the optimum sequences with the lowest energies?

Example 2-1:

A simple 2D HP model for sequence design using GDMC

A 4x4 2D protein lattice conformation



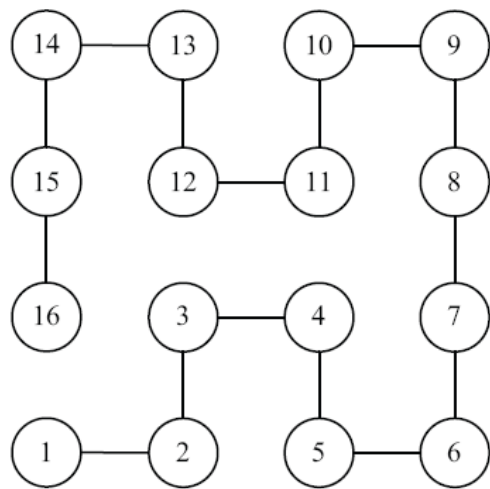
For site i ,

$$S_i = \begin{cases} 0: \text{polar residue (P)} \\ 1: \text{hydrophobic residue (H)} \end{cases}$$

Sites 1-16: either **H** or **P** residue

Number of possible sequences is 2^{16} !

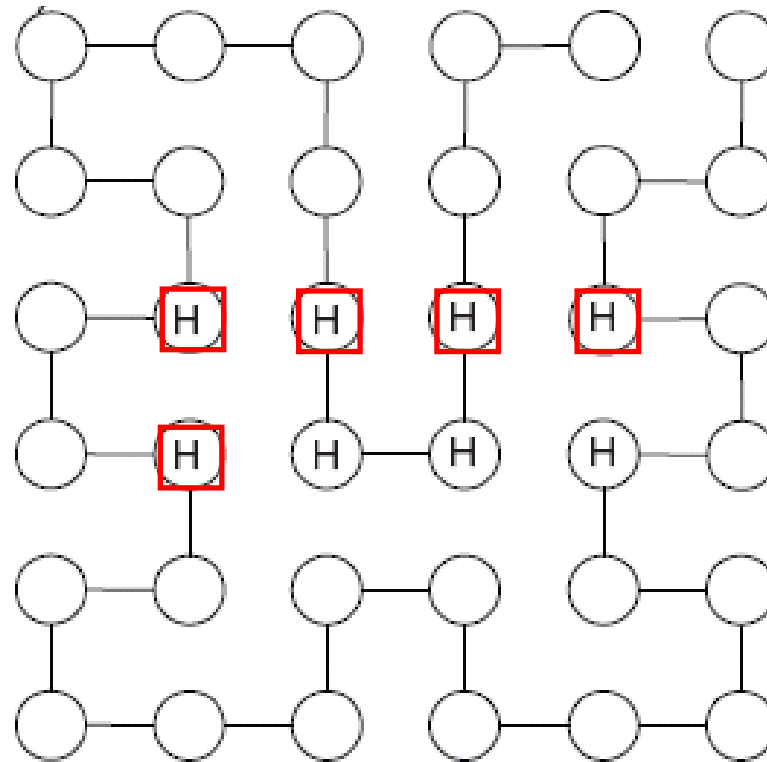
(a) GDMC optimization results for the 4x4 2D protein lattice model:



N_H	E_{\min}^{exact}	E_{\min}^{MC}	MC Step	E_{\min}^{GDMC}	GDMC Step
6	-12.2	-11.9	286	-12.2	2
7	-13.5	-13.5	540	-13.5	2
8	-14.8	-14.8	185	-14.8	3
9	-16.1	-15.8	149	-16.1	7

- The exact global minima are known for this example.
- GDMC always found the global minima.
- Less than 10 sequence calculations are sufficient in GDMC!

(b) A 6×6 2D protein lattice conformation with $N_H=8$:



Number of possible sequences is 2^{36} !

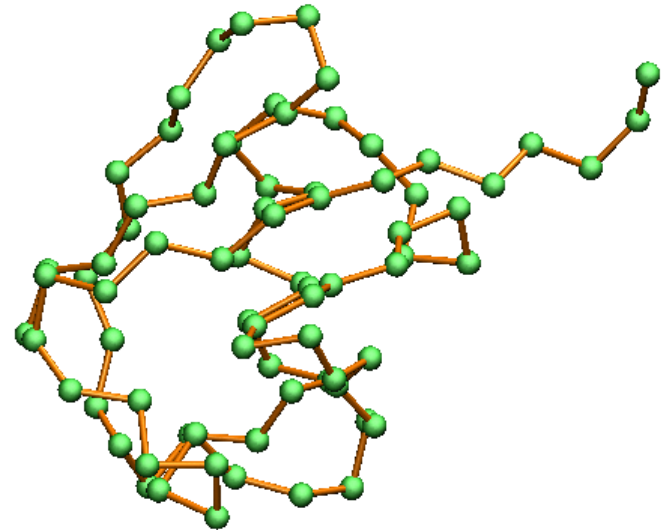
- Four degenerate sequences with the global minimum energy were obtained.
- The positions of five H residues are conserved for the four minima.
- These five adjacent positions form a favorable hydrophobic core.

Example 2-2:

Realistic protein sequence design using RosettaDesign and GDMC

Ubiquitin scaffold:

- 74 residues
- 19 amino acids per site
- 11,076 possible rotamers per site

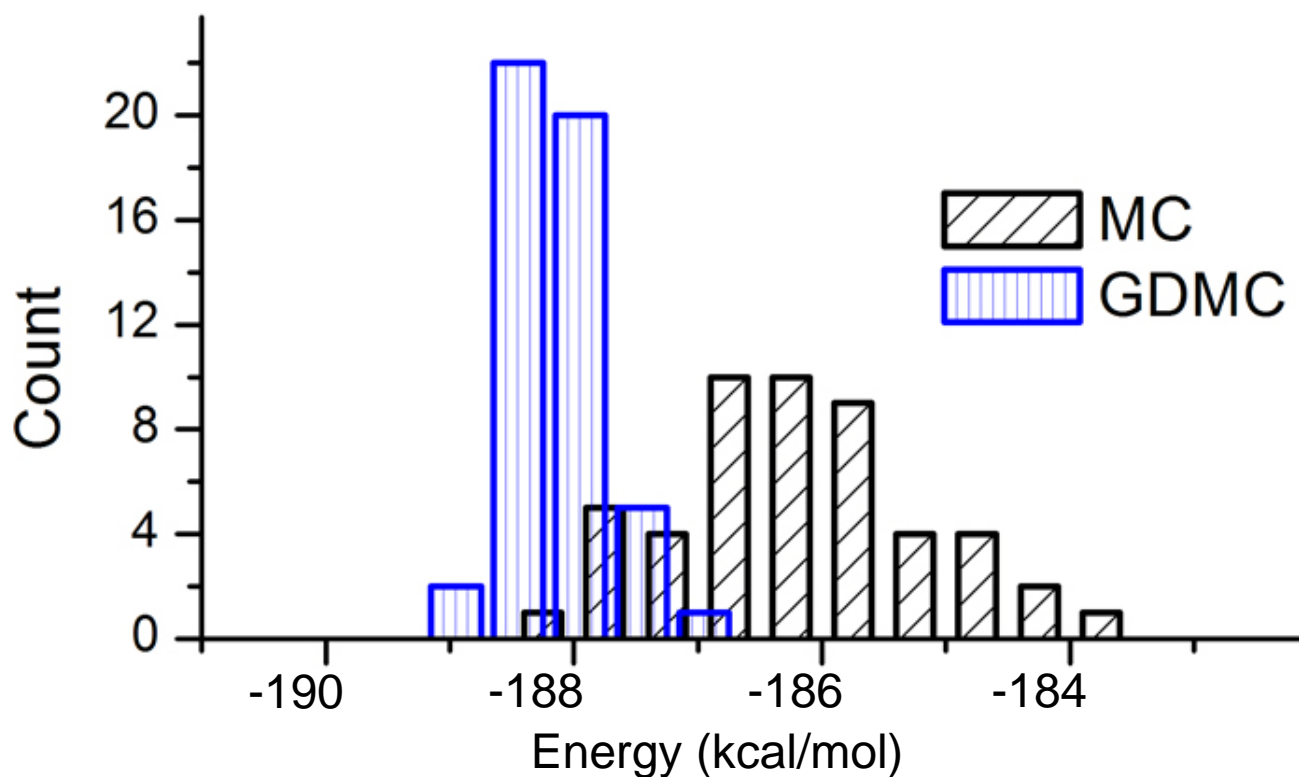


The number of possible sequence is $11,076^{74}$!

X. Hu, D. Beratan, and W. Yang, *J. Comp. Chem.*, 2010, 31, 2164

GDMC obtains more sequences with lower energies than MC.

Fifty independent GDMC and MC optimizations beginning with random sequences were performed.



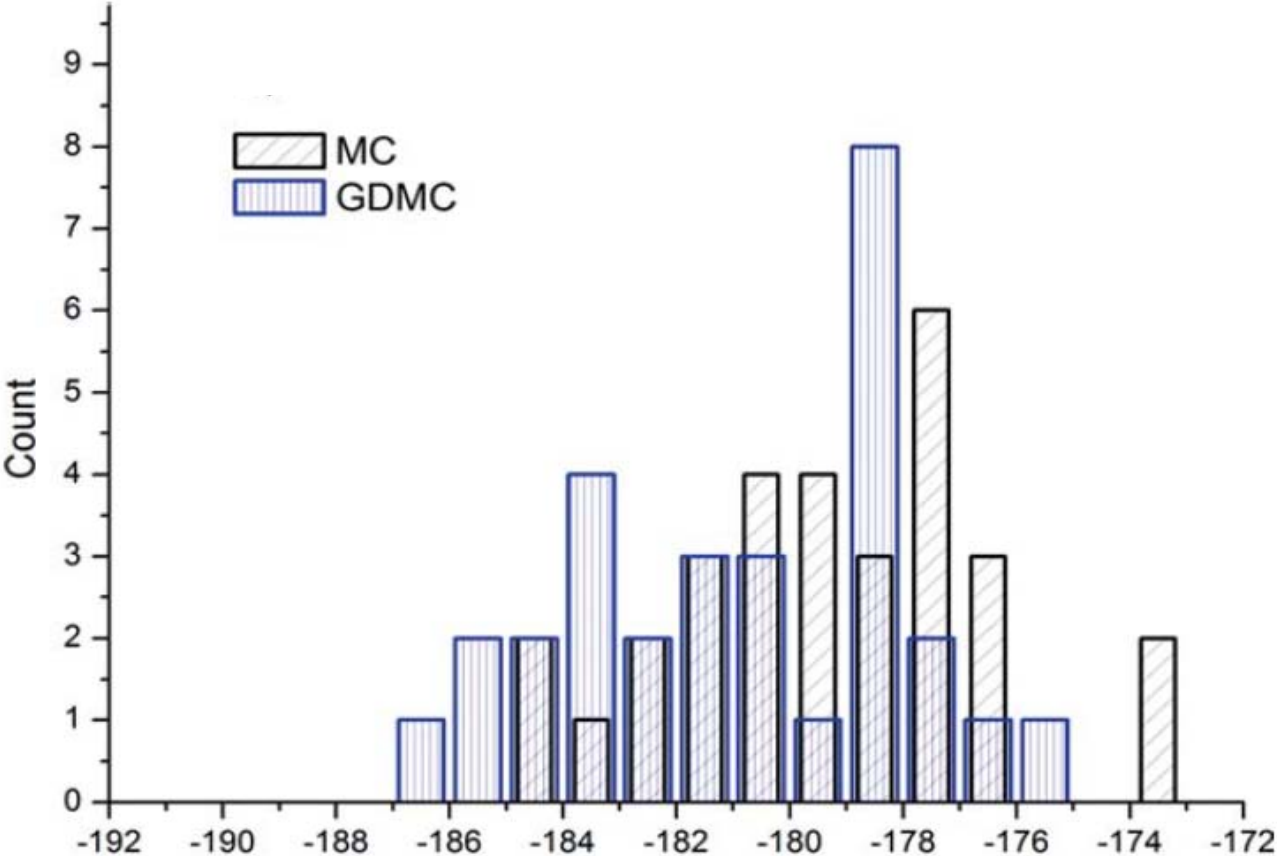
More tests with fixed/flexible backbone support GDMC can obtain a better sequence ensemble than MC using Rosetta scoring function.

Table 1. The Average Energy Values (kcal/mol) of Optimal Sequences for Different Optimizations.

Methods	Fixed backbone		Flexible backbone
	Native sequence	Random sequence	Native sequence
MC	-160.99 (0.1)	-163.98 (0.5)	-180.92 (9)
MC×5	-186.89 (0.4)	-186.59 (1.2)	N/A
MC×200	-187.90 (2.8)	-187.33 (27)	-179.22 (116)
GDMC	-187.63 (2.0)	-188.12 (22)	-180.58 (35)

*The computational cost (unit: hours) is shown in parentheses.

MC can be easily trapped in local minimum when native sequences are relatively stable while GDMC can explore much broader sequence space.



Example 3: Use of GDMC for protein folding

Can we fold a set of 2D HP sequences using GDMC?

ID	No. of residues	Sequence ^a
1	20	HRHRPHHRPHRPHHRPHRH
2	24	HNRRHRPHRPHRPHRPHRPHRH
3	25	RRHRPHHRPPRHNRRPPRHNRRPHH
4	36	RRRHNRRPHHRPPRRHNHNHNHNRRHNRRPPRHNRRPHRPH
5	48	RRHRPHHRPHHRPPRRHNHNHNHNHNRRPPRRRHNRRPHRPHRPHRPHHNHNHN
6	50	HNRRHRPHRPHHNHNRRHRPPRRHRPPRRHRPPRRHRPHHNHNRRHRPHRPHRH
7	60	RRHNHNRRHNHNHNHNRRPHHNHNHNHNHNRRPHHNHNHNHNHNRRPPRHNHNHNHNRRHRPH
8	64	HNHNHNHNHNHNHNRRHRPHRPHHRPHRPHRPHRPHRPHRPHRPHRPHRPHRPHRPHRPHHNHNHNHNHNHNHNHN
9	85	HNHNHRPPRHNHNHNHNHNHNHNRRPPRRRHNHNHNHNHNHNHNRRPHHNHNHNHNHNHNRRPPRHNHNHNHNHN
		HNHNHRPPRPHRPHRPHRPHRPHRH
10	100	RRPPRRHRPHHRPPRRHNHNRRHNHNHNRRHNRRPHRPHRPHRPHHNHNHNRRHNHNHNHNHNRRHNRRHNHNHNHN
		HRPPRRPPRRRHNHNHNHNRRHRPHRPHRPHRPPRRRPHRH

GDMC with numerical gradients can outperform other global optimization methods when folding 2D HP sequences.

ID	MC ^a		GA ^a		CI ^b			MC-pull moves ^c			GDMC-pull moves ^c			Duration ^d ($\times 10^3$)
	E^c	Time ^f	E^c	Time ^f	E^c	Time ^f	Hits ^g	E^c	Time ^f	Hits ^g	E^c	Time ^f	Hits ^g	
1	-9	292 443	-9	30 492	-9	171	5	-9	149	5	-9	496	5	10
2	-9	2 492 221	-9	30 491	-9	1425	5	-8	144	5	-9	1248	4	10
3	-8	2 694 572	-8	20 400	-8	1132	5	-8	674	4	-8	1683	5	30
4	-13	6 557 189	-14	301 339	-14	40 237	2	-14	8895	5	-14	4238	5	300
5	-20	9 201 755	-22	126 547	-23	204 928	1	-23	29 269	5	-23	43 434	5	300
6	-21	15 151 203	-21	592 887	-21	13 464	5	-19	170 047	3	-21	76 349	1	300
7	-33	8 262 338	-34	111 400	-35	361 533	1	-36	198 486	1	-36	194 950	2 ^h	500
8	-35	7 848 952	-38	97 220	-40	461 099	1	-42	76 401	4	-42	61 233	3	500
9	-52	233 039	5	-53	4 302 404	2	5000
10	-47	295 270	2	-48 ⁱ	1 441 692	1	5000

Conclusions

- We developed LCAP to optimize molecular properties at the quantum mechanical level.
- When the property surface of LCAP is smooth, the continuous optimization algorithms are extremely efficient.
- When the property surface of LCAP is rugged, we developed an efficient global optimization approach - GDMC.
- GDMC can be applied to many complex problems such as inhibitor design and protein design.

Inverse Molecular Design Team at Duke



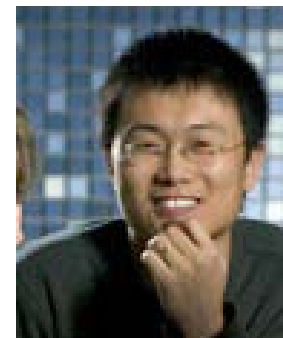
Prof. David Beratan



Shahar Keinen



Nan Jiang



Xiangqian Hu



Prof. Mingliang Wang
Shenzhen Univ



Prof. M. Therien



Aaron Virshup



Julia Conteras-Garcia



Dequan Xiao
Postdoc at Yale

Ongoing Collaboration with **Univ. of Pitt** on **Library Design**



Peter Wipf



Kay Brummond

C. Riderspacher

We are thinking inversely...



Prof. David Beratan



Shahar Keinen



Nan Jiang



Xiangqian Hu



Prof. Mingliang Wang
Shenzhen Univ



Prof. M. Therien



Aaron Virshup



Julia Conteras-Garcia



Dequan Xiao
Postdoc at Yale

Ongoing Collaboration with **Univ. of Pitt** on **Library Design**



Peter Wipf



Kay Brummond