

Learning Scoring Functions for Chemical Expert Systems

C.-A. Azencott, M. A. Kayala, and P. Baldi

Institute for Genomics and Bioinformatics
Donald Bren School of Information and Computer Sciences

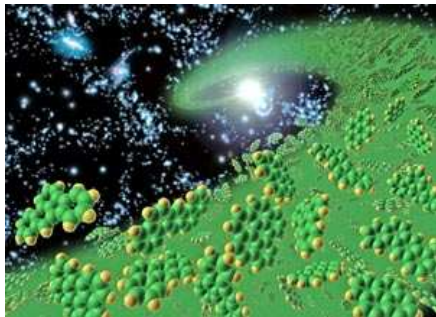


237th ACS Meeting

March 24, 2009

Goals

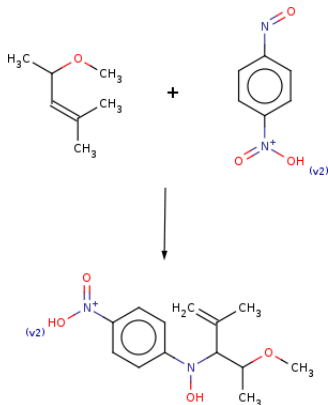
- Chemical space exploration
- Reproduce Augment problem-solving ability of organic chemists



Reaction Simulator

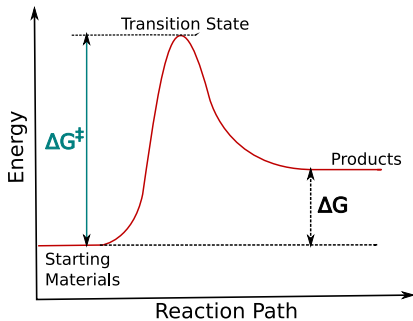
- Input:
 - Starting materials
 - Conditions (pH, temperature)
- Output:
 - Products
 - Energy diagram

→ Reaction favorability **scoring function?**



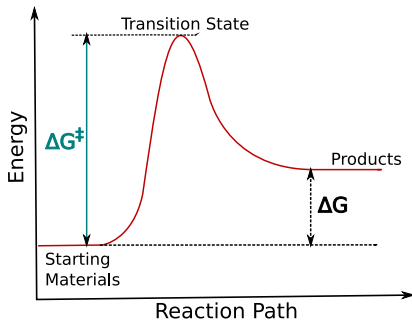
Energy of Activation ΔG^\ddagger

- Measure of **reaction favorability**
- Directly related to reaction rates



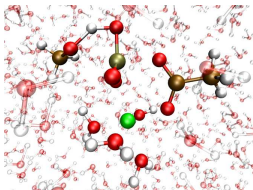
Energy of Activation ΔG^\ddagger

- Measure of **reaction favorability**
- Directly related to reaction rates



- How can we compute ΔG^\ddagger ?
 - Direct computation not possible
 - Laboratory experiment or QM/MM simulation
 - Infer with **Machine Learning** → Data

ΔG^\ddagger numerical data



- Experimental data
- Quantum Mechanical / Molecular Modeling

Problems:

- Time Consuming
- Expensive
- Very little available

→ What can we do?

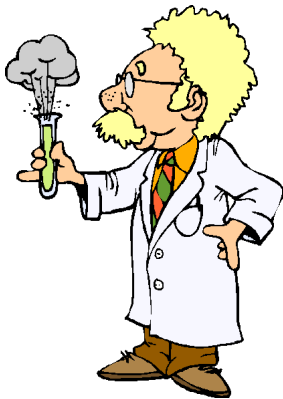
How do chemists solve problems?

Without:

- performing experiments
- producing numerical ΔG^\ddagger

chemists can (almost) always:

- tell which of several possible reactions is **more favorable**
- produce **relative rates**



How do chemists solve problems?

Without:

- performing experiments
- producing numerical ΔG^\ddagger

chemists can (almost) always:

- tell which of several possible reactions is **more favorable**
- produce **relative rates**



→ How do we get this **qualitative knowledge**?

Jonathan H. Chen: <http://www.reactionexplorer.org>

- Rule-based expert system
- Basic undergraduate organic chemistry

ChemDB / Reaction Explorer / Reaction Explorer: Setup

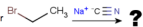
[Reaction Drills](#) [Synthesis Explorer](#) [Mechanism Explorer](#) [User Records](#) [Help](#)

Reaction Explorer: Organic Chemistry Tutorials

Reaction Explorer is an interactive system for learning and practicing reactions, syntheses and mechanisms in organic chemistry, with advanced support for the automatic generation of random problems, curved-arrow mechanism diagrams, and inquiry-based learning.

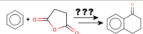
Reaction Drills: Single-Step Reaction Completion

Reaction Drills produce a series of (ungraded) reaction equations with one hidden component (reactant, reagent or product) and asks you to mentally "fill-in-the-blank" to test and train your skill, similar to flashcards.



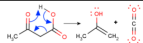
Synthesis Explorer: Multi-Step Synthesis Design

The primary interface for solving multi-step synthesis design problems and for freely exploring different reactant and reagent combinations.



Mechanism Explorer: Arrow-Pushing Mechanism Diagrams

Reactions support curved-arrow mechanism diagram viewing and exploration to not only show "what" a reaction will produce, but to also explain "how" the reaction will proceed.



Jonathan H. Chen: <http://www.reactionexplorer.org>

- Rule-based expert system
- Basic undergraduate organic chemistry

→ Create pairs of

- elementary steps
- ordered by favorability

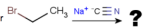
ChemDB / Reaction Explorer / Reaction Explorer: Setup
Reaction Drills [Synthesis Explorer](#) [Mechanism Explorer](#) [User Records](#) [Help](#)

Reaction Explorer: Organic Chemistry Tutorials

Reaction Explorer is an interactive system for learning and practicing reactions, syntheses and mechanisms in organic chemistry, with advanced support for the automatic generation of random problems, curved-arrow mechanism diagrams, and inquiry-based learning.

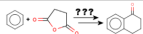
Reaction Drills: Single-Step Reaction Completion

Reaction Drills produce a series of (ungraded) reaction equations with one hidden component (reactant, reagent or product) and asks you to mentally "fill-in-the-blank" to test and train your skill, similar to flashcards.



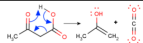
Synthesis Explorer: Multi-Step Synthesis Design

The primary interface for solving multi-step synthesis design problems and for freely exploring different reactant and reagent combinations.



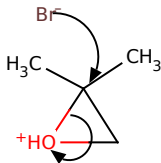
Mechanism Explorer: Arrow-Pushing Mechanism Diagrams

Reactions support curved-arrow mechanism diagram viewing and exploration to not only show "what" a reaction will produce, but to also explain "how" the reaction will proceed.

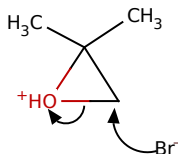


Approach 1: Priority Ordering

- Rules priority \rightarrow order elementary movements
- Most favorable electron movement

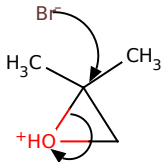


- Less favorable electron movement

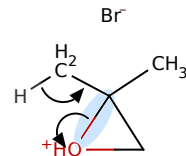
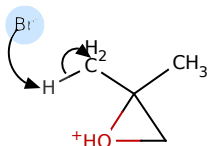
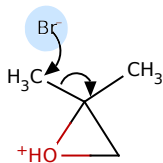


Approach 2: Implicitly Unfavorable Reactions

- From what simple elementary movement **should** happen



- Deduce what elementary movements **should not** happen

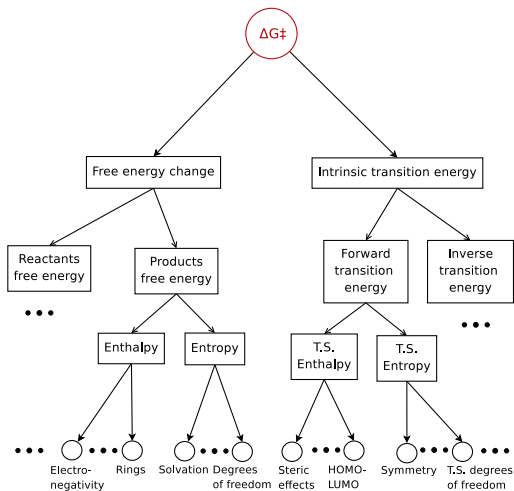


CHNO + Halides, Ionic Reactions.

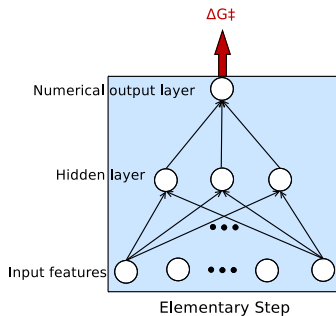
Data Set	Number of Pairs
Priority ordered	3,457
Implicitly unfavored	475,684 (from 16,806 most favorable reactions)

Feature Representation, f : Elementary Step $\rightarrow \mathbb{R}^n$

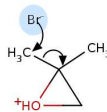
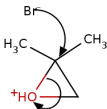
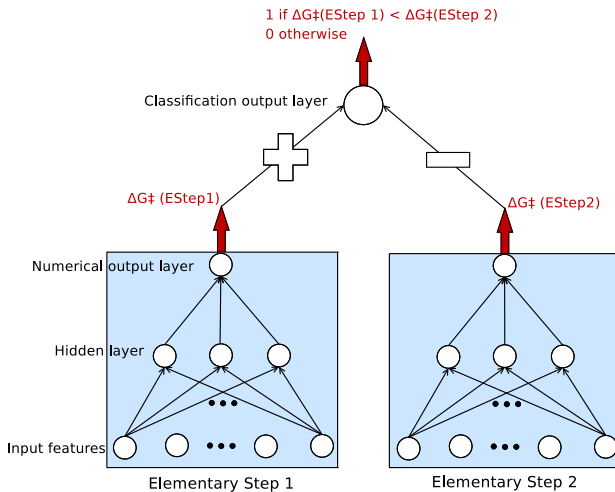
- Electron movement between pair of source / sink orbitals
- Principled choice of 150 features



Neural Network Architecture I



Neural Network Architecture II



Classification Performance

Cross-Validated accuracy:

Data Set	Architecture	Accuracy
Priority ordered	Perceptron (no hidden nodes)	92.8%
	15 hidden nodes	93.7%
Implicitly unfavored	Perceptron (no hidden nodes)	98.8%
	15 hidden nodes	99.0%
Altogether	Perceptron (no hidden nodes)	98.6%
	15 hidden nodes	99.0%

Performance in the Simulator

Reaction Mixture [Help](#)

Molecule / SMILES	Quantity
	<input type="text"/>
<chem>CC([O-])</chem>	<input type="text" value="100"/>
	<input type="text"/>
<chem>CC(=O)Cl</chem>	<input type="text" value="100"/>

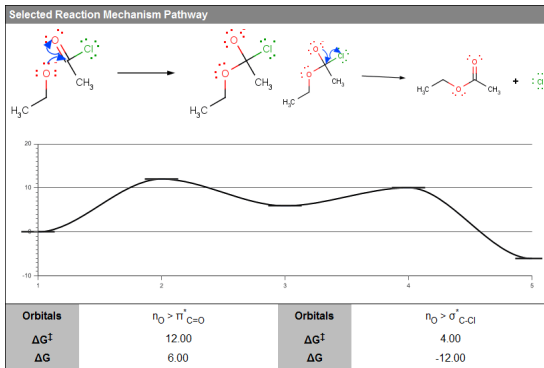
Performance in the Simulator II

Reaction Mixture		Help
Molecule / SMILES	Quantity	
<input type="text" value="CC[O-]"/>	<input type="text" value="50"/>	
<input type="text" value="CC(=O)Cl"/>	<input type="text" value="50"/>	
<input type="text" value="CCOC(C)[O-]Cl"/>	<input type="text" value="50"/>	

Performance in the Simulator III

Reaction Mixture Help	
Molecule / SMILES	Quantity
<input type="text" value="CC[O-]"/>	<input type="text" value="4"/>
<input type="text" value="CC(=O)Cl"/>	<input type="text" value="4"/>
<input type="text" value="CCOC(C)(O-)Cl"/>	<input type="text" value="9"/>
<input type="text" value="CCOC(=O)C"/>	<input type="text" value="87"/>

Performance in the Simulator IV



Conclusion

- Lack of **quantitative** data
- Compensate with **qualitative** knowledge
- Applied to the prediction of ΔG^\ddagger
 - Ionic Reactions
 - C, H, N, O + halides
 - Qualitative trends / ranking order of reactivities necessary to **solve relevant chemistry problems**
- Future Work
 - Expand **coverage**
 - **Validation** on external reaction databases (e.g. SPRESI)

Thanks

- Matthew A. Kayala
- Jonathan H. Chen
Reaction Simulation Expert System for Synthetic Organic Chemistry, talk Thursday, March 26th at 10:30 am (CINF General Papers)
- Prof. Pierre Baldi
- The Baldi Lab

- OpenEye Academic Software License
- ChemAxon Academic Software License

- NIH/NLM Biomedical Informatics Training Program
- NSF Grants 0321390 and 0513376
- The Dreyfus Foundation Special Grant Program in Chemical Sciences