Learning Scoring Functions for Chemical Expert Systems

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

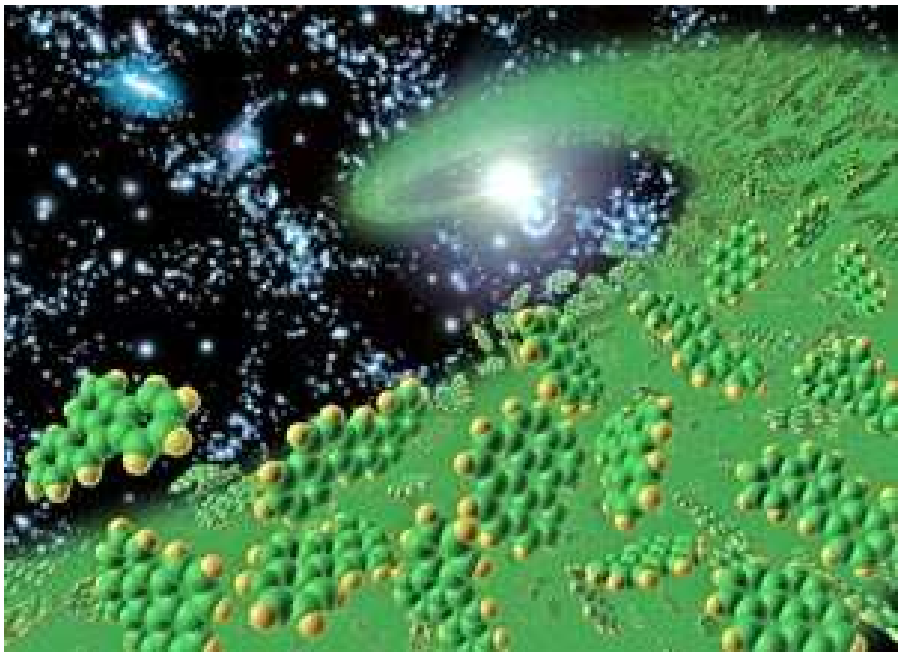
Institute for Genomics and Bioinformatics
Donald Bren School of Information and Computer Sciences

237th ACS Meeting

March 24, 2009

Goals

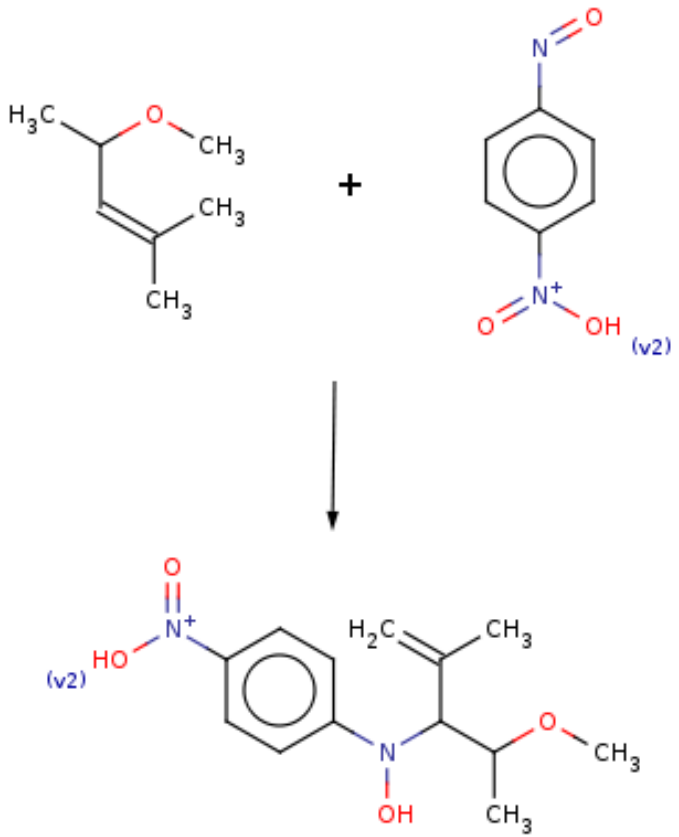
- Chemical space exploration
- ~~Reproduce~~ Augment problem-solving ability of organic chemists



Reaction Simulator

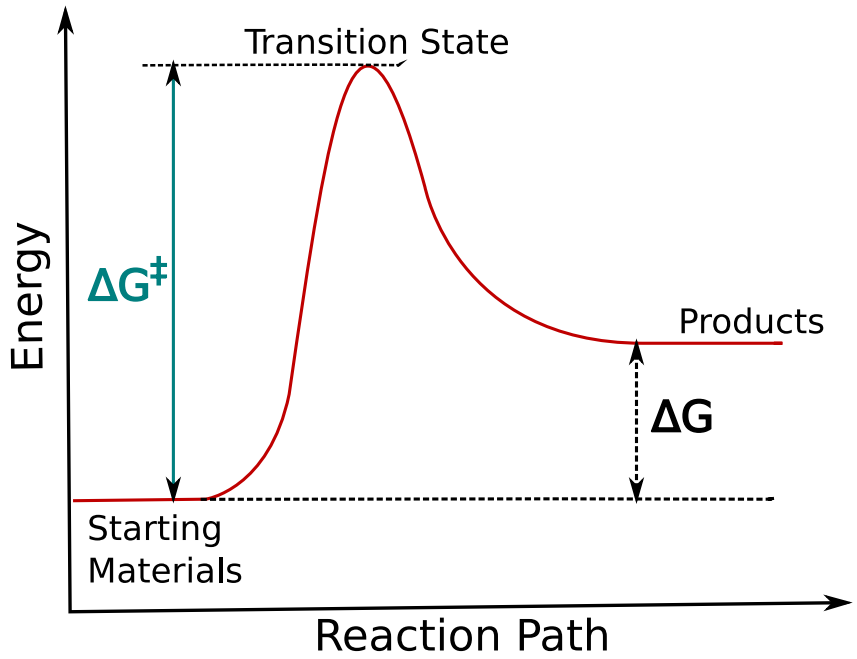
- Input:
  - Starting materials
  - Conditions (pH, temperature)
- Output:
  - Products
  - Energy diagram

→ Reaction favorability scoring function?



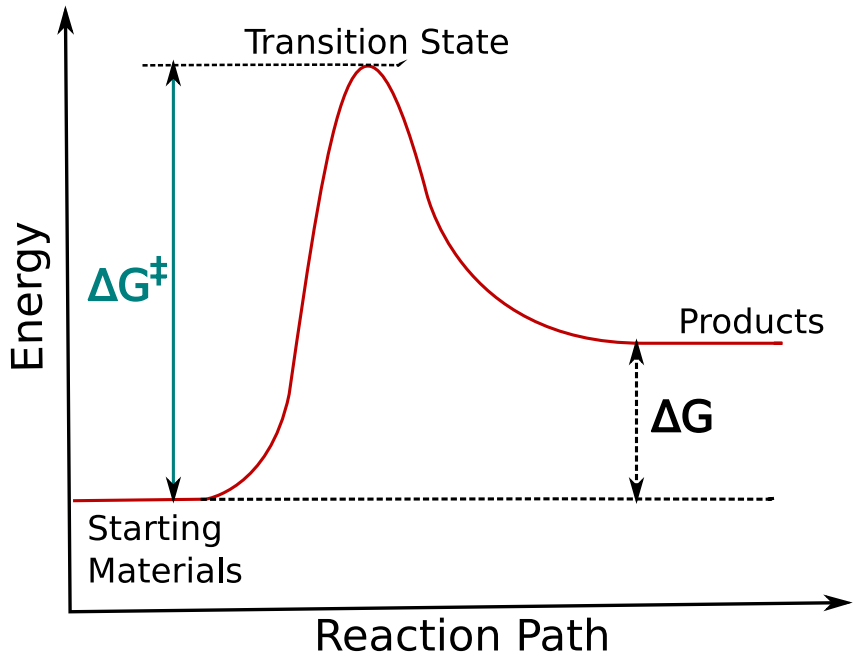
Energy of Activation $\Delta G^{\ddagger}$

- Measure of reaction favorability
- Directly related to reaction rates



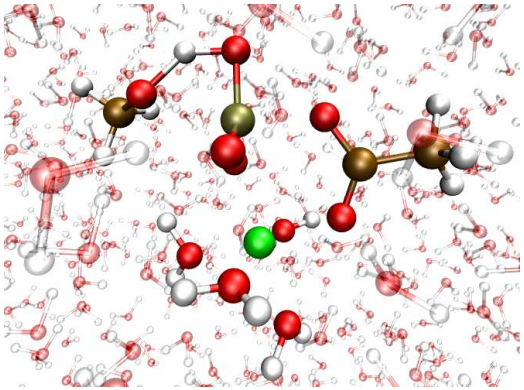
Energy of Activation $\Delta G^{\ddagger}$

- Measure of reaction favorability
- Directly related to reaction rates



- How can we compute $\Delta G^{\ddagger}$?
  - Direct computation not possible
  - Laboratory experiment or QM/MM simulation
  - Infer with Machine Learning $\rightarrow$ Data

$\Delta G^{\ddagger}$ numerical data



- Experimental data
- Quantum Mechanical / Molecular Modeling

Problems:

- Time Consuming
- Expensive
- Very little available

$\longrightarrow$ What can we do?

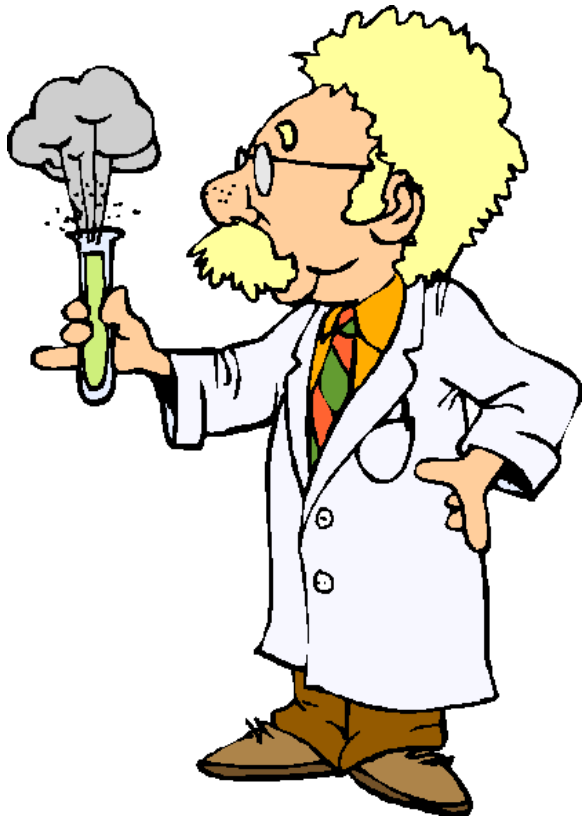
How do chemists solve problems?

Without:

- performing experiments
- producing numerical $\Delta G^{\ddagger}$

chemists can (almost) always:

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



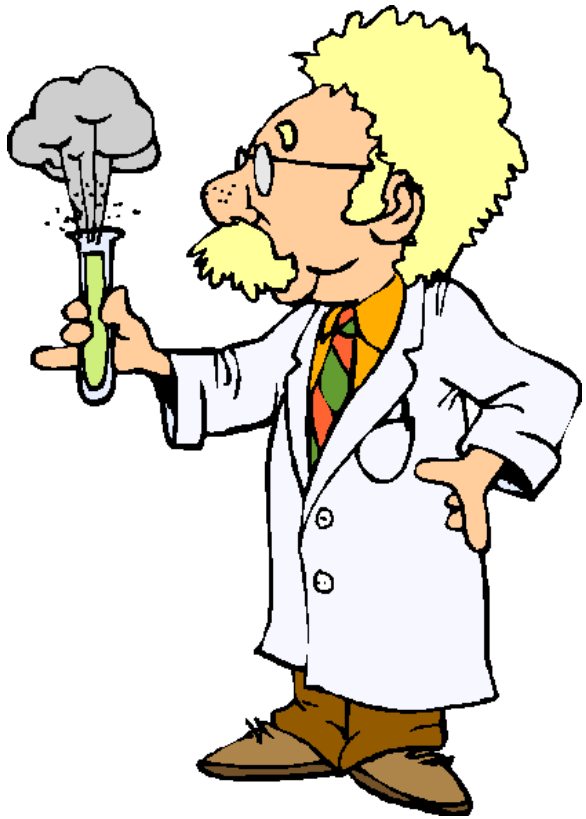
How do chemists solve problems?

Without:

- performing experiments
- producing numerical $\Delta G^{\ddagger}$

chemists can (almost) always:

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

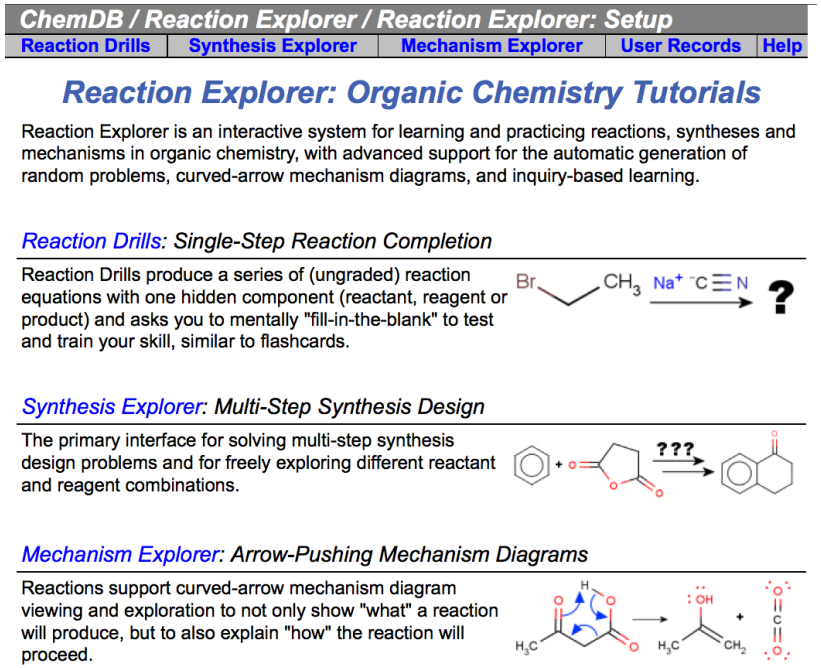


$\rightarrow$ How do we get this qualitative knowledge?

Reaction Explorer

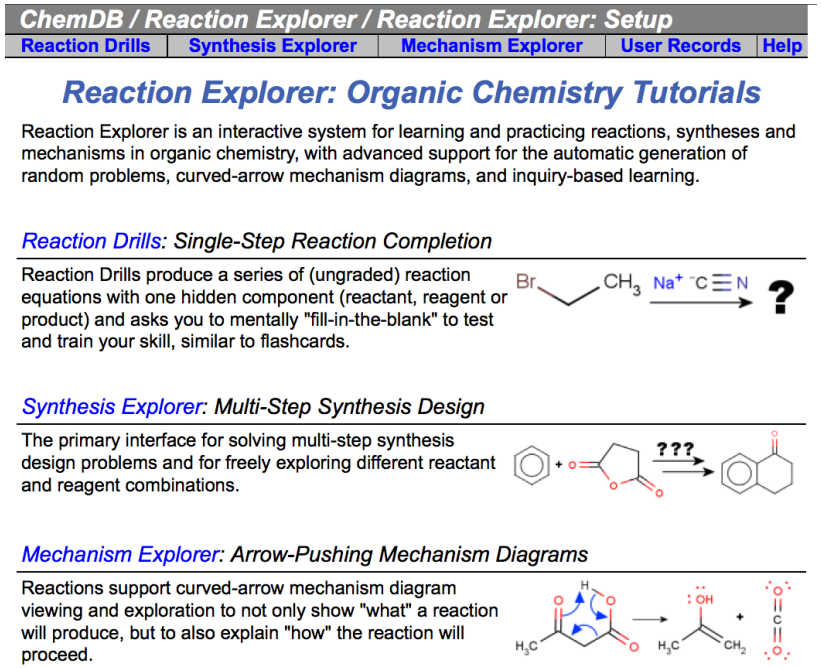
Jonathan H. Chen: `http://www.reactionexplorer.org`

- Rule-based expert system
- Basic undergraduate organic chemistry



Reaction Explorer

Jonathan H. Chen: `http://www.reactionexplorer.org`



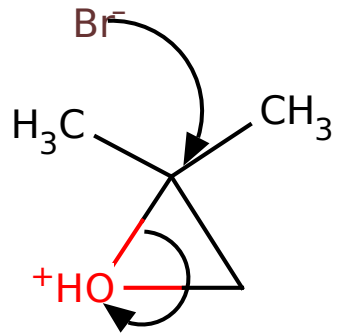
- Rule-based expert system
- Basic undergraduate organic chemistry

$\longrightarrow$ Create pairs of

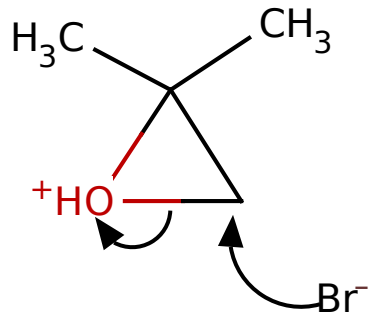
- elementary steps
- ordered by favorability

Approach 1: Priority Ordering

- Rules priority $\longrightarrow$ order elementary movements
- Most favorable electron movement

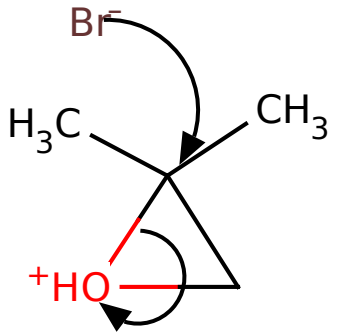


- Less favorable electron movement

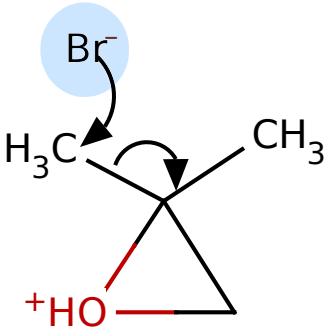


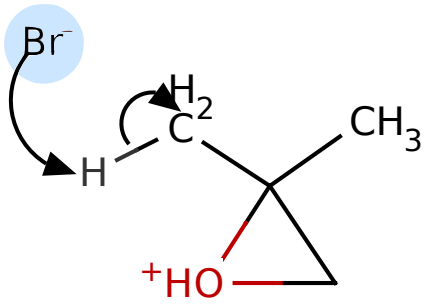
Approach 2: Implicitly Unfavorable Reactions

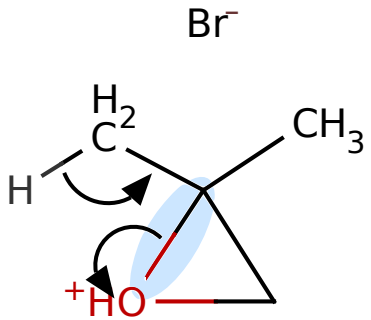
- From what simple elementary movement **should** happen



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







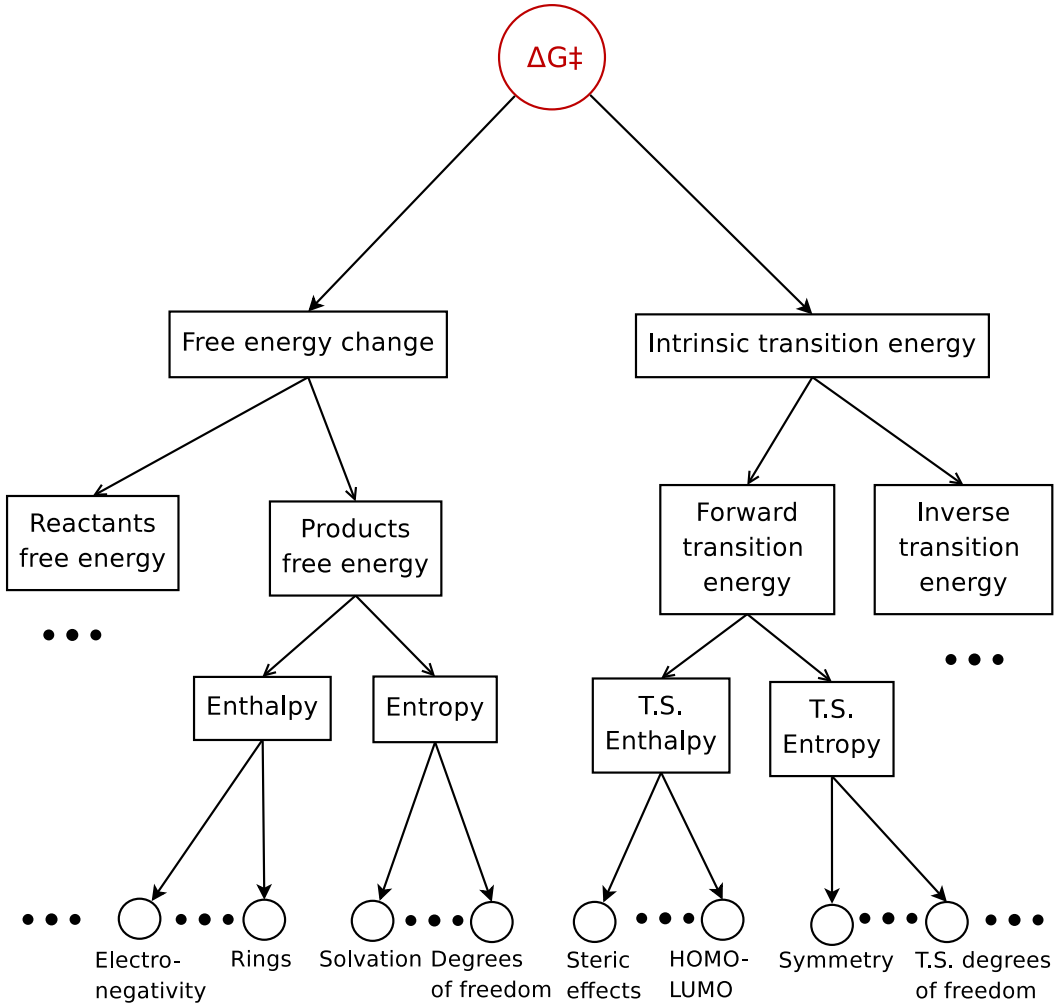
Data Sets

CHNO + Halides, Ionic Reactions.

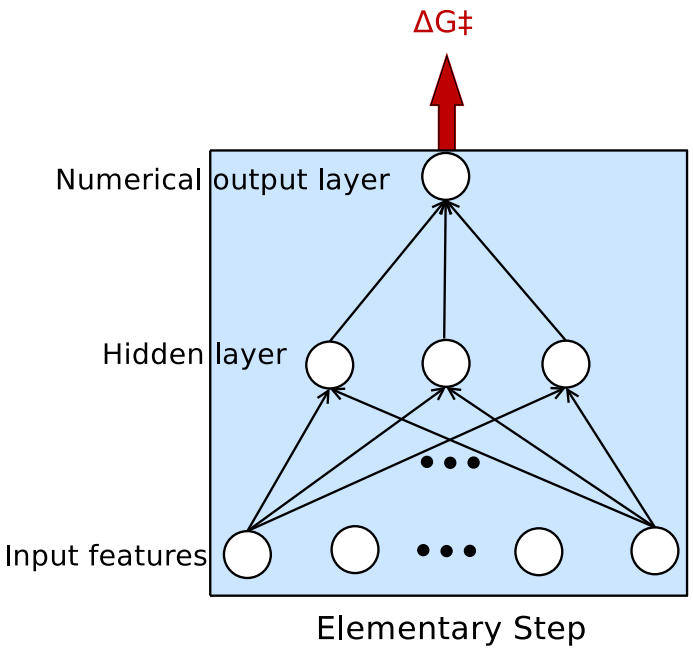
Data Set	Number of Pairs
Priority ordered	$3,457$
Implicitly unfavored	$475,684$ (from $16,806$ most favorable reactions)

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

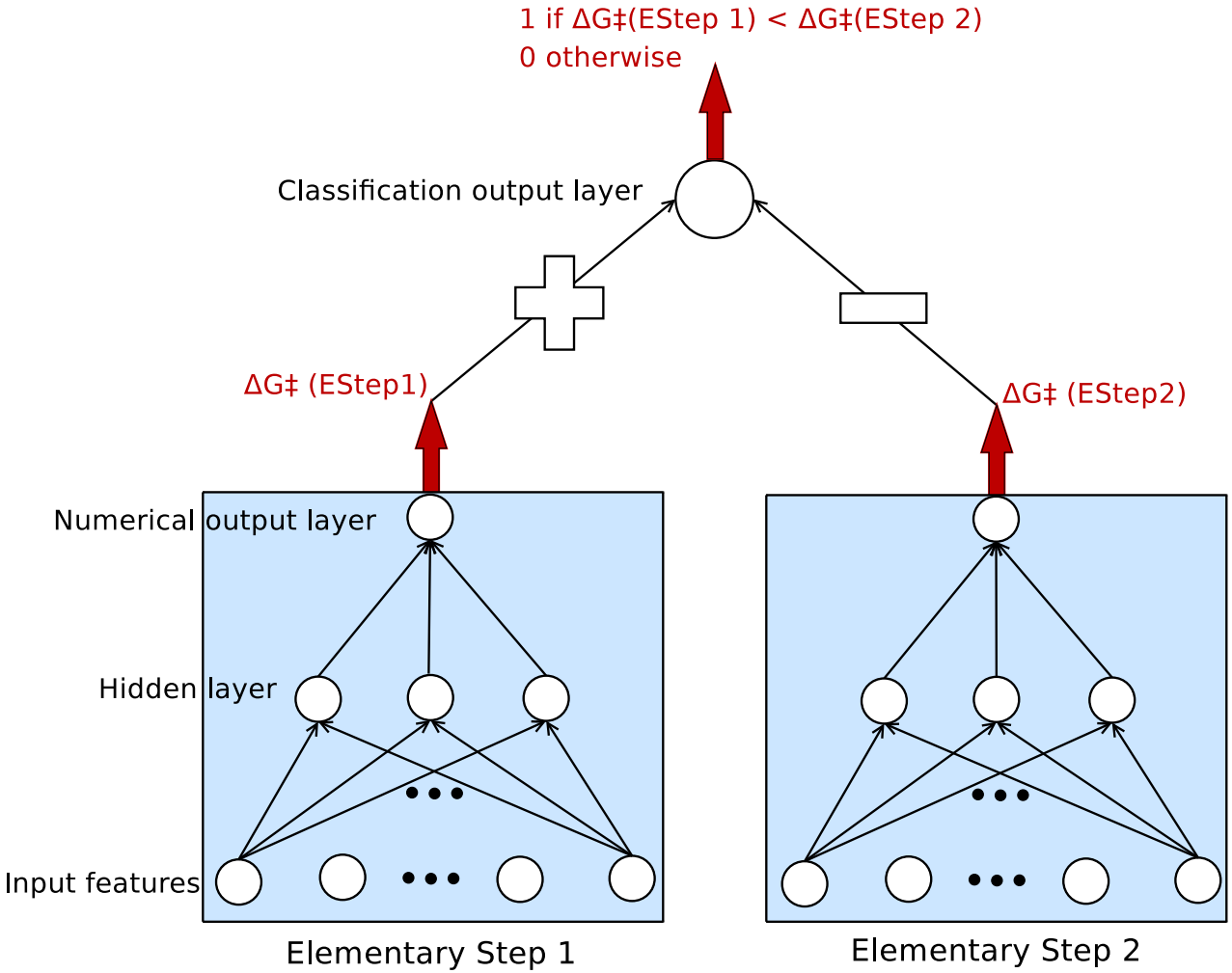
- Electron movement between pair of source / sink orbitals
- Principled choice of 150 features

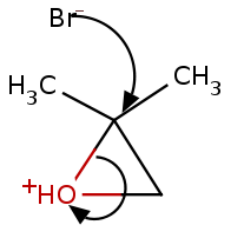


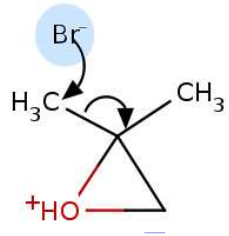
Neural Network Architecture I



Neural Network Architecture II





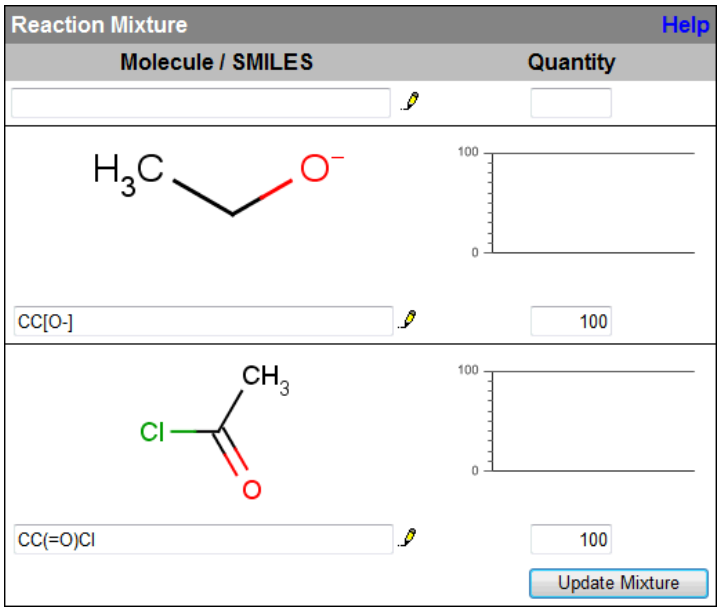


Classification Performance

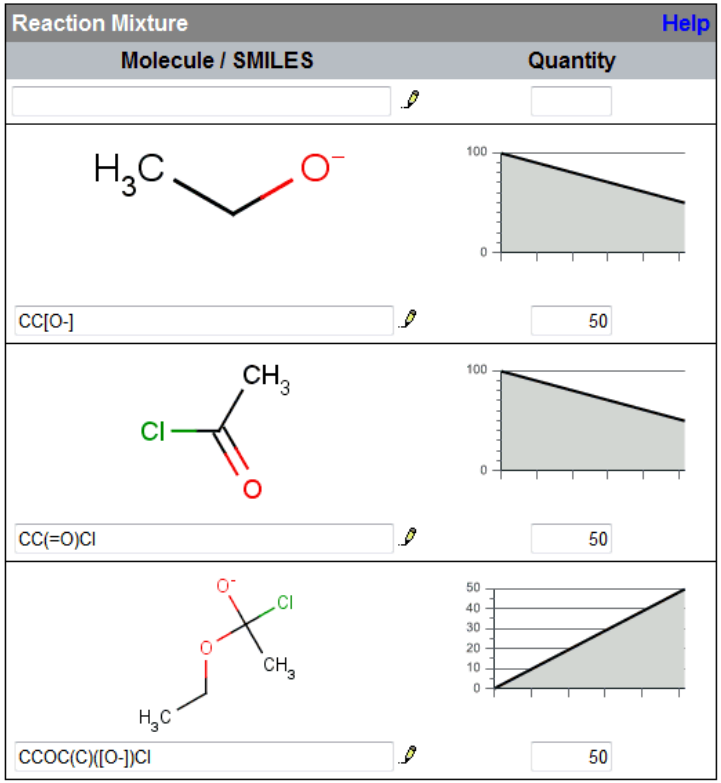
Cross-Validated accuracy:

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%**

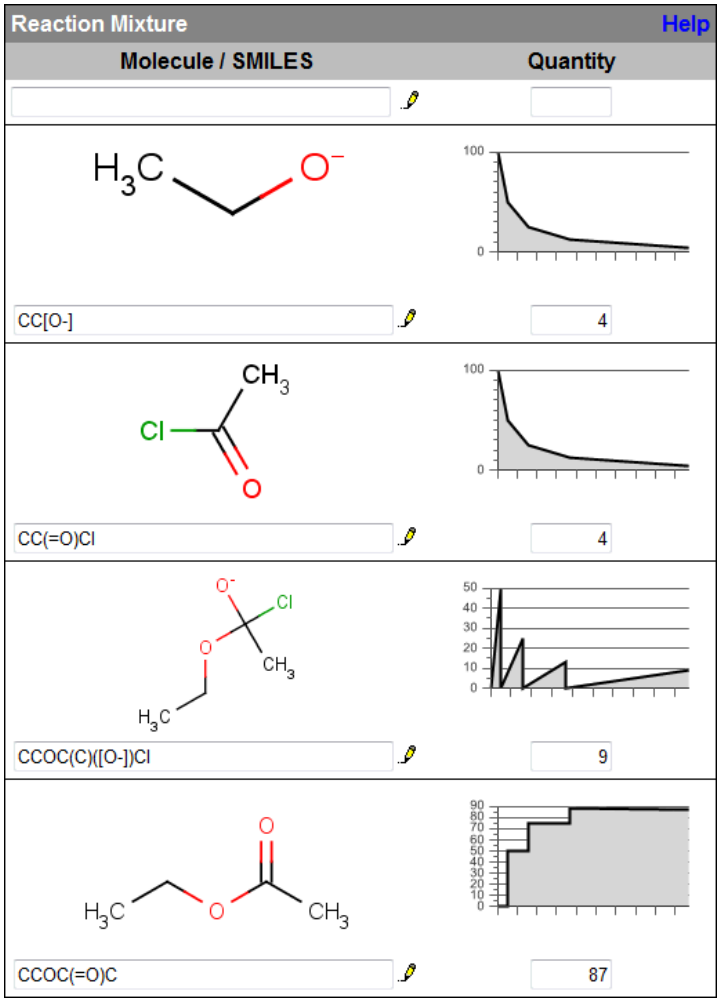
Performance in the Simulator I



Performance in the Simulator II

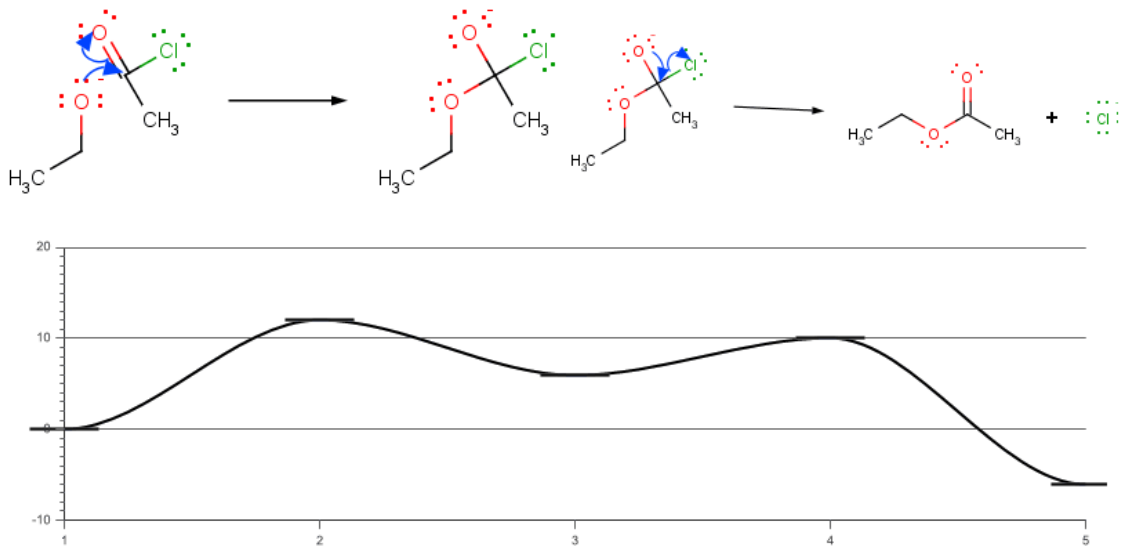


Performance in the Simulator III



Performance in the Simulator IV

Selected Reaction Mechanism Pathway



Orbitals	$n_O > \pi^*_{C=O}$	Orbitals	$n_O > \sigma^*_{C-Cl}$
$\Delta G^{\ddagger}$	12.00	$\Delta G^{\ddagger}$	4.00
ΔG	6.00	ΔG	-12.00

Conclusion

- Lack of quantitative data
- Compensate with qualitative knowledge
- Applied to the prediction of $\Delta G^{\ddagger}$
  - 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 Acadedmic Software License

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