Data Handling and Informatics
Tools for Model-Based Discovery

Jim Caruthers, Nick Delgass, Sam Midkiff, Venkat Venkatsubramanian and Gary Blau

with

Stephen Stamatis, Leif Delgass Bala Krishnamurthy, Tanu Malik, Jun Cao, Hongang Wang,
Shou-Huan Hsu, Sumo Nandi, Seza Orcun and Steve Dunlop

and

Tom Manz, Grisha Medvedev, Jesmin Haq, Krista Novstrup, Ayush Goyal, Gowri
Krishnamurthy, Abhijit Phatak, Shalini Sharma, Khampee Phomphree, Fabio Riberio and
Mahdi Abu Omar

Chemical Engineering, Chemistry, Computer Graphics Technology, Computer Science,
Electrical & Computer Engineering, Industrial Engineering, ITaP, Cyber Center, Envision
Center and Center for Catalyst Design

Supported by: DOE Office of Basic Science,
Indiana’s 21st Century Research and Development Fund
ExxonMobil
Cummins
Equistar Chemicals
Purdue University
Current Data Archiving Methods

- Stored in Windows nested file folders
- Can not be searched
- At the end of a PhD thesis
- Limited data into Excel spreadsheet
- Copy/paste into other programs
- Metadata incomplete
Current situation is barely manageable, but just think about scaling-up with high throughput data.

1\textsuperscript{st} Group + 2\textsuperscript{nd} Group + \ldots + n\textsuperscript{th} Group

\[ \text{x10} \quad \text{x10} \quad \text{x10} \]

- How can we integrate data from different groups?
- How does one ensure data persistence?
- How does one assign intellectual ownership of group data?
- How can this be done for a small research effort like the battery community?
• How can CyberInfrastructure aid in the extraction of useful knowledge from the flood of data?
• Requirements
  – Single time of ingress
  – Databases, not folders, that are ontologically enabled
    (i.e. can be searched with words/concepts that have chemical meaning)
  – Analysis programs integrate with database
  – Advanced visualization tools for human processing of information
  – Must be low friction - the researcher can focus on chemistry not IT tools

SciAether™

Science: the process of systematically generating knowledge from data

Aether: the magical substance postulated by the late 19th century physicists that supported all physical processes
Outline

• Data Ingress – e-Lab Notebook
• Database – ontologically enabled
• Integrated analysis environment
• Analysis tools
• Visualization
• Computer-aided discovery
Requirements of e-Lab Notebook

1. Intuitive interface that creates connections in the database
2. Ability to easily create new templates
3. eLN has to be able to work offline
4. Interface should allow integration with 3rd party software (e.g. Chemsketch, etc.)
5. Ability to attach raw/binary data from instruments
6. Interface should have ability to enter symbols and equations
7. The e-Lab Notebook should freeze all data at the end of the day and time stamp the data – legal IP protocol
8. Data provenance must be archived

Data must only be ingressed a single time – no copying from paper notebook
Welcome to the ELN Client

1. Use the Browse button to select your template file

2. Click on the Load Data button

File for template: N:\Personal\ELN Archives\CCD_template1_blank_arylOxi_de_expsV2.2.zip cpt

Load Data  Browse  Exit Client
### Single Site Polymerization Project

<table>
<thead>
<tr>
<th>Request Date and Time</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Template Unique Identifier</td>
<td>template2</td>
<td></td>
</tr>
<tr>
<td>Template Name</td>
<td>Single Site Polymerization Project</td>
<td></td>
</tr>
<tr>
<td>Template Version</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>Template Description</td>
<td>Single Site Polymerization Computation Template designed for Tony Mann</td>
<td></td>
</tr>
<tr>
<td>Requester</td>
<td>offline_test_user</td>
<td></td>
</tr>
<tr>
<td>Requester Type</td>
<td>Edit</td>
<td></td>
</tr>
<tr>
<td>Requester ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name(s) of Owner(s)</td>
<td>offline_test_user</td>
<td></td>
</tr>
<tr>
<td>Date Modified/Created</td>
<td>11/7/2006 1:07 PM</td>
<td></td>
</tr>
<tr>
<td>Record Number</td>
<td>now</td>
<td></td>
</tr>
<tr>
<td>Parent Record Number</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

Tabbed layout for Easy Navigation

Color Coded fields tell user what is required
Several ways of identifying the catalyst
Including InChI and SMILES

Catalyst structure drawing made with chemsketch

Catalyst structure also attached in chemsketch format for easy editing
# Kinetic Data from NMR

<table>
<thead>
<tr>
<th>Performed on</th>
<th>2/9/2007 5:09 PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performed by</td>
<td>Nicholas Travia</td>
</tr>
<tr>
<td>Catalyst Batch Name</td>
<td>1</td>
</tr>
<tr>
<td>Instrument</td>
<td>Inova 300</td>
</tr>
<tr>
<td>Internal Standard</td>
<td>PH2CH2</td>
</tr>
<tr>
<td>Solvent</td>
<td>Toluene-d8</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>25</td>
</tr>
<tr>
<td>Kinetic Data</td>
<td>MQ-NET-1240.csv</td>
</tr>
<tr>
<td>Image of Kinetic Data</td>
<td>MQ-NET-1240.PNG</td>
</tr>
</tbody>
</table>

**Metadata and other details about the experiment**

**Raw data attached**

**Short Summary**

**Graph Image**
Custom Parser automatically fills out fields from Gaussian Log files
Outline

• Data Ingress – e-Lab Notebook
• Database – ontologically enabled
• Integrated analysis environment
• Analysis tools
• Visualization
• Computer-aided discovery
Consider a database with 10,000 or more records

Example Query: Find all polymerization data for all non-styrenic olefin monomers for which the kinetics were measured via NMR in toluene for bridged Group IV catalysts.

SQL needs to understand that 1-hexene is a nonstyrenic olefin

SQL needs to understand that Ti and Zr are Group IV metals & what is a bridged ligand
Ontology: Defines relationships between vocabulary words

Example ontology for aryloxide chemistry

Protégé screenshot

- Web Ontology Language (OWL) is W3C standard
- Encodes logic in Description Logic (DL) format
- Use Protégé as OWL editor
Ontological-SQL Query Engine

Design: A semantic analysis layer interacts with a data retrieval layer

ANTLR = ANother Tool for Language Recognition: Parser generator
Racer, Jess: classification of concepts and instances in ontology
OWL = Web Ontology Language

AntLR = ANother Tool for Language Recognition: Parser generator
Racer, Jess: classification of concepts and instances in ontology
OWL = Web Ontology Language
Outline

• Data Ingress – e-Lab Notebook
• Database – ontologically enabled
• Integrated analysis environment
• Analysis tools
• Visualization
• Computer-aided discovery
<table>
<thead>
<tr>
<th>Image</th>
<th>( \theta_{O \text{-Cone Angle}} )</th>
<th>( \theta_{\text{Predicted}} )</th>
<th>( \theta_{\text{Experimental}} )</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>117.3</td>
<td>0.36</td>
<td>0.34</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>116.2</td>
<td>0.38</td>
<td>0.42</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>65.6</td>
<td>0.41</td>
<td>0.41</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>65.6</td>
<td>0.24</td>
<td>0.27</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>65.5</td>
<td>0.27</td>
<td>0.27</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>65.44</td>
<td>0.27</td>
<td>0.27</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>66.4</td>
<td>0.41</td>
<td>0.38</td>
<td>36</td>
</tr>
<tr>
<td>8</td>
<td>66.4</td>
<td>0.44</td>
<td>0.46</td>
<td>36</td>
</tr>
<tr>
<td>9</td>
<td>57</td>
<td>0.51</td>
<td>0.51</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>111.3</td>
<td>0.78</td>
<td>0.78</td>
<td>26</td>
</tr>
<tr>
<td>11</td>
<td>114.8</td>
<td>0.99</td>
<td>0.92</td>
<td>26</td>
</tr>
<tr>
<td>12</td>
<td>97.9</td>
<td>0.41</td>
<td>0.74</td>
<td>29</td>
</tr>
<tr>
<td>13</td>
<td>98.9</td>
<td>0.32</td>
<td>0.28</td>
<td>29</td>
</tr>
<tr>
<td>14</td>
<td>92.0</td>
<td>0.27</td>
<td>0.27</td>
<td>28</td>
</tr>
<tr>
<td>15</td>
<td>110.1</td>
<td>1.35</td>
<td>1.36</td>
<td>14</td>
</tr>
<tr>
<td>16</td>
<td>65.7</td>
<td>0.25</td>
<td>0.24</td>
<td>29</td>
</tr>
<tr>
<td>17</td>
<td>66.4</td>
<td>0.18</td>
<td>0.2</td>
<td>30</td>
</tr>
<tr>
<td>Image</td>
<td>dG_Cone_Angle</td>
<td>lg_Predicted</td>
<td>lg_Experimental</td>
<td>CPE</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>--------------</td>
<td>----------------</td>
<td>-----</td>
</tr>
<tr>
<td>0</td>
<td>97.0</td>
<td>0.23</td>
<td>0.22</td>
<td>35</td>
</tr>
<tr>
<td>1</td>
<td>112.3</td>
<td>0.36</td>
<td>0.34</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>116.2</td>
<td>0.38</td>
<td>0.32</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>55.6</td>
<td>0.39</td>
<td>0.48</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>60.6</td>
<td>0.24</td>
<td>0.27</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>56.5</td>
<td>0.27</td>
<td>0.24</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>68.44</td>
<td>0.27</td>
<td>0.27</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>66.4</td>
<td>0.41</td>
<td>0.36</td>
<td>36</td>
</tr>
<tr>
<td>8</td>
<td>65.7</td>
<td>0.44</td>
<td>0.46</td>
<td>36</td>
</tr>
<tr>
<td>9</td>
<td>57</td>
<td>0.51</td>
<td>0.51</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>111.9</td>
<td>0.28</td>
<td>0.28</td>
<td>26</td>
</tr>
<tr>
<td>11</td>
<td>114.3</td>
<td>0.99</td>
<td>0.92</td>
<td>26</td>
</tr>
<tr>
<td>12</td>
<td>97.9</td>
<td>0.41</td>
<td>0.24</td>
<td>29</td>
</tr>
<tr>
<td>13</td>
<td>58.9</td>
<td>0.32</td>
<td>0.28</td>
<td>29</td>
</tr>
<tr>
<td>14</td>
<td>62.8</td>
<td>0.27</td>
<td>0.27</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td>140.3</td>
<td>1.35</td>
<td>1.36</td>
<td>14</td>
</tr>
<tr>
<td>16</td>
<td>66.7</td>
<td>0.25</td>
<td>0.24</td>
<td>29</td>
</tr>
<tr>
<td>17</td>
<td>66.4</td>
<td>0.18</td>
<td>0.2</td>
<td>36</td>
</tr>
</tbody>
</table>
### CCD Database Query

**Project:** Single site polymerization Experimental  
**Back**

**Selected Fields**
- kp_Predicted  
- kp_Experimental  
- C  
- ObE  
- CPE  
- Activity  
- Mol. Weight  
- PDI  
- Temperature

**Search Constraints**
- Temperature < 10 AND (PDI < 2)

**Example:** (Molecular Weight < 10000) AND (PDI < 2)

**Fields**
- kp_Predicted  
- kp_Experimental  
- C  
- ObE  
- CPE  
- Activity  
- Mol. Weight  
- PDI  
- Temperature

<table>
<thead>
<tr>
<th>Image</th>
<th>kp_Predicted</th>
<th>kp_Experimental</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>117.3</td>
<td>0.36</td>
<td>0.54</td>
</tr>
<tr>
<td>2</td>
<td>116.2</td>
<td>0.38</td>
<td>0.42</td>
</tr>
<tr>
<td>3</td>
<td>66.6</td>
<td>0.39</td>
<td>0.41</td>
</tr>
<tr>
<td>4</td>
<td>66.6</td>
<td>0.24</td>
<td>0.27</td>
</tr>
<tr>
<td>5</td>
<td>66.5</td>
<td>0.27</td>
<td>0.24</td>
</tr>
<tr>
<td>6</td>
<td>66.44</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>7</td>
<td>66.4</td>
<td>0.41</td>
<td>0.30</td>
</tr>
<tr>
<td>8</td>
<td>66.7</td>
<td>0.44</td>
<td>0.46</td>
</tr>
<tr>
<td>9</td>
<td>57</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>10</td>
<td>111.3</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>11</td>
<td>114.8</td>
<td>0.99</td>
<td>0.92</td>
</tr>
<tr>
<td>12</td>
<td>97.9</td>
<td>0.41</td>
<td>0.74</td>
</tr>
<tr>
<td>13</td>
<td>66.9</td>
<td>0.32</td>
<td>0.28</td>
</tr>
<tr>
<td>14</td>
<td>67.6</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>15</td>
<td>140.1</td>
<td>1.35</td>
<td>1.36</td>
</tr>
<tr>
<td>16</td>
<td>66.7</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>17</td>
<td>66.4</td>
<td>0.18</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**CCD Graph**

![Graph](image)
• **Data Ingress – e-Lab Notebook**

• **Database – ontologically enabled**

• **Integrated analysis environment**

• **Analysis tools**
  - Commercial Packages *(MatLab, JMP, etc.)*
  - Personal Codes *(MatLab, C++, Fortran, etc.)*
  - Nonlinear Bayesian Statistics
  - Domain specific tools

• **Visualization**

• **Computer-aided discovery**
Parameter Estimation

- Expert knowledge (Prior probability distribution)

\[ p(\theta) = p(\theta_1) p(\theta_2) \cdots p(\theta_n) \]

- Likelihood function, \( L(\text{data}|\theta) \)

\[ L(\text{data} | \theta) = p(e_1) p(e_2) \cdots p(e_n) \]

- Both expert knowledge and data fitting are important

- How to compromise these two different types of information to obtain the most reasonable parameter estimates?

The larger, the better
Revisit the Example

\[ C_i = C_0 \exp(-kt_i) + \varepsilon_i \quad C_0 = 2000, \; k = 0.23, \; \gamma = 1.3 \]

- Non-designed data set

- Designed data set
Outline

• Data Ingress – e-Lab Notebook
• Database – ontologically enabled
• Integrated analysis environment
• Analysis tools
• Visualization
• Computer-aided discovery
Natural Representation of Data
Summary

- **SciAether** – prototype cyberinfrastructure
  - Initially developed for catalysis science
  - Can be expanded to include a wide range of chemistry/materials/biology research
- **Scaleable components**

  ![Diagram of components: e-Lab Notebook, Ontologically Enabled Database, Linked Analysis Environment, Advanced Visualization]

  **Low Friction**

- **Minimum system requirements**
  - **PC with Windows XP**
  - **Microsoft Office**
  - **Database (MS Access or DB2 or ...)**
  - **ChemSketch – freeware version**
  - **Analysis software (MatLab, etc. – user’s choice)**

  | Academic Cost | 1,200               | 100 (free) | 50 (free) | free | ??? (free) |
Outline

• Data Ingress – e-Lab Notebook
• Database – ontologically enabled
• Integrated analysis environment
• Analysis tools
• Visualization
• Computer-aided discovery
Does it work?
1- Hexene Polymerization by Titanium Catalysts with Phenoxy based ligands

A large number of available substituted phenols allow tunability of steric and electronic variation of the catalyst.

\[
\begin{align*}
    \text{M} &= \text{Ti, Zr, Hf} \\
    \text{C.I.} &= [\text{MeB(C}_6\text{F}_5)_3]^- \\
    &\quad [\text{B(C}_6\text{F}_5)_4]^-
\end{align*}
\]

\[
\begin{align*}
    \text{L}_1 &= \{ \text{structures} \} \\
    \text{L}_2 \text{ ligand} &\quad \text{Steric}
\end{align*}
\]
Effect of Aryloxide Ligand on Propagation Rate for Titanium Catalyst

Batch Polymerization

\[
\ln\left(\frac{[M]}{[M]_0}\right) = -k_{obs} \times t
\]

-3.5 -3 -2.5 -2 -1.5 -1 -0.5 0 1000 2000 3000 4000 5000 6000 7000 8000 9000

Time (s)

\[k_{obs} = 1.4 \times 10^{-3} \text{ s}^{-1}\]
\[k_{obs} = 7.1 \times 10^{-4} \text{ s}^{-1}\]

[Hex] = 1.0 M
[Ti] = 0.0050 M
[B] = 0.0050 M
Toluene-d8, 0°C

James M. Caruthers  NSF Battery Workshop, MIT, Cambridge, MA  September 8-9, 2008
Homo-polymerization Kinetics

* Initiation

\[ C + M \xrightarrow{k_i} R_1 \]

* Propagation

\[ R_i + M \xrightarrow{k_p} R_{i+1} \]

* 2,1-misinsertion

\[ R_i + M \xrightarrow{k_{p2,1}} P_{i+1} \]

* Propagation after 2,1-misinsertion

\[ P_i + M \xrightarrow{k_{p2,1,1,2}} R_{i+1} \]

* Chain transfer to monomer

\[ R_i \xrightarrow{k_{im}} D_i + R_i \]

\[ P_i \xrightarrow{k_{im}} D_i + R_i \]

* β hydride chain transfer

\[ R_i \xrightarrow{k_i \beta} D_i + C \]

\[ P_i \xrightarrow{k_i \beta} D_i + C \]

where

- \( C \) activated catalyst
- \( M \) monomer
- \( R_i \) living polymer chain
- \( P_i \) dormant polymer chain
- \( D_i \) terminated polymer chain (\( i = 1,2, \ldots \))

Population Balances

* For active sites

\[ \frac{dC}{dt} = -k_i \cdot C \cdot M + k_{i\beta} \cdot \left( \sum_{1}^{\infty} R_i + \sum_{2}^{\infty} P_i \right) \]

* For living chains of unit length

\[ \frac{dR_i}{dt} = k_i \cdot C \cdot M - (k_p + k_{p2,1}) \cdot M \cdot R_i + k_{iM} \cdot M \cdot \left( \sum_{1}^{\infty} R_i + \sum_{2}^{\infty} P_i \right) - k_{i\beta} \cdot R_i \]

* For living chains with length \( i \)

\[ \frac{dR_i}{dt} = k_p \left( R_{i-1} - R_i \right) \cdot M - k_{p2,1} \cdot R_i \cdot M + k_{p2,1,1,2} \cdot P_{i-1} \cdot M - \left( k_{iM} \cdot M - k_{i\beta} \right) \cdot R_i \]

* For monomer

\[ \frac{dM}{dt} = -k_i \cdot C \cdot M - (k_p + k_{p2,1} \cdot M) \cdot \sum_{1}^{\infty} R_i - k_{p2,1,1,2} \cdot M \cdot \sum_{2}^{\infty} P_i \]

\[ -k_{iM} \left( \sum_{2}^{\infty} R_i + \sum_{2}^{\infty} P_i \right) \cdot M + k_{i\beta} \cdot R_i \]

* For living chains after 2,1-misinsertion with length \( i \)

\[ \frac{dP_i}{dt} = k_{p2,1} \cdot R_{i-1} \cdot M - k_{p2,1,1,2} \cdot P_i \cdot M - (k_{iM} \cdot M - k_{i\beta}) \cdot P_i \]

* For terminated chains of length \( i \)

\[ \frac{dD_i}{dt} = k_{iM} \cdot M \cdot \left( R_i + P_i \right) + k_{i\beta} \cdot (R_i + P_i) \]
DFT Simulation of Propagation Step

- Backside insertion of 1-hexen into [CpTi(OC₆H₃Me₂-2,6)Me⁺][MeB(C₆F₅)₃⁻]; OLYP/LALNL2DZ level calculation

<table>
<thead>
<tr>
<th>State</th>
<th>Energy (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I – Reactants</td>
<td>E_I = 0</td>
</tr>
<tr>
<td>II – TS1</td>
<td>E_II = 17.9</td>
</tr>
<tr>
<td>III – Coordinated π-complex</td>
<td>E_III = 12.0</td>
</tr>
<tr>
<td>IV – TS2</td>
<td>E_IV = 13.2</td>
</tr>
<tr>
<td>V – Products</td>
<td>E_V = -18.8</td>
</tr>
</tbody>
</table>

Conclusion: adsorption is the rate determining step
Postulated mechanics for propagation: Adsorption step

Descriptor #1 – ArO ligand cone angle

Descriptor #2 – CI binding energy

CI B.E. = \( E \left( \begin{array}{c} \text{Cp}\text{*} \\ \text{ArO} \end{array} \right) + E \left( \text{CH}_3 \right) \) - \( E \left( \begin{array}{c} \text{Cp}\text{*} \\ \text{ArO} \end{array} \right) - E \left( \text{CI}^- \right) \)
**$k_p$ Prediction Model**

*(Manz, et al. JACS-Communication, 2007)*

\[ k_{\text{pred}} = k_0 e^{-\frac{E_a}{RT}} = \gamma a_0 e^{-\frac{E_0}{RT}} e^{-\frac{\alpha EIPS}{RT}} \]

\[ \gamma = 1 - \sin^2(\frac{\theta_{Cp}}{4}) - \sin^2(\frac{\theta_{OAr}}{4}) - f \]

$\alpha = 0.300$, $f = 0.187$, and the following values of $A$ (M$^{-1}$ s$^{-1}$) according to catalyst family: (A) $3.01 \times 10^8$ (■), (B) $5.22 \times 10^7$ (●), (C) $A = 2.65 \times 10^7$ (◆), (D) $6.88 \times 10^5$ (★), and (E) $1.77 \times 10^7$ (▲).

**Insight from analysis**

- **Postulate ‘hot’ catalysis**
- **Synthesized, and catalyst ~2x hotter than reported catalysts**
Future Plans

• System has been designed and implemented using commercial software development tools/practices/people

• Expanding usage of SciAether to other groups at Purdue

• Additional capabilities under-development
  • Template Designer for eLN
  • Drop down menus for eLN
  • Expanded ontologies with user GUI for addition of terms
  • Connect to other databases like PubChem via Web Services
  • Direct connection to eLN from analysis environment
  • 3D visualization inside of linked analysis environment

• Looking for a few development partners

www.sciaether.org
caruther@ecn.purdue.due