

Collaboratory for Multi-Scale Chemical Science

Status as of January 2003 / Quarterly Report for Q1 of FY 2003

Project Staff

Larry Rahn-SNL*, Christine Yang, Carmen Pancerella, Wendy Koegler, David Leahy, Michael Lee, Renata McCoy, Theresa Windus-PNL*, James D. Myers, Karen Schuchardt, Brett Didier, Eric Stephan, Carina Lansing, Al Wagner-ANL*, Branko Ruscic, Michael Minkoff, Sandra Bittner, Gregor von Laszewski, Reinhardt Pinzon, Sandeep Nijsure, Kaizar Amin, Baoshan Wang, William Pitz-LLNL*, David R. Montoya-LANL*, Lili Xu, Yen-Ling Ho, Thomas C. Allison-NIST*, William H. Green, Jr.-MIT*, Michael Frenklach-UCB*

* denotes Institutional Point of Contact

Summary

During this performance period the pilot Collaboratory for Multi-Scale Chemical Sciences (CMCS) implemented a prototype of the Version 1 software. Multiple science areas have made data available to the CMCS structure through program modification and data translators. These include data from NIST, NWChem, Ecce, GRI-Mech, Chemkin, and HCT. The first version of Active Thermochemical Tables (ATcT) has been developed and demonstrated in the CMCS Portal. The CMCS team produced seven use-case driven demonstrations illuminating the central aspects of the project, and presented the Main Demonstration at SC2002 in two theater presentations. Presentations of the SC2002 demonstrations were also made to team members not able to attend SC2002 with a focus on updating our vision for scientific applications. We focus this summary report on three highlights of our progress. These include successful implementation and demonstration of 1) the infrastructure integrating SAM, CHEF Web Portal, DAV, and other technologies, 2) data pedigree and browsing concepts, and 3) Active Thermochemical Tables development and integration into the infrastructure.

This progress has established a baseline that is helping the project team to prioritize the tasks required to implement Version 1.0 production features of the CMCS software as well as the planning required for future infrastructure and application development. Progress has also been made in the planning for support of our first external user groups, and for the upcoming Peer Review.

Progress

This document summarizes the work done over the first quarter during FY03 of the CMCS project. First, a summary listing of project activities is presented, then more a detailed discussion of accomplishments is presented in the form of project highlights.

Summary of CMCS Project Activities – October 2002 through December 2002

- Updated CMCS Pedigree document to include information to be used for searching data (10/2002)

- Attended CHEF Workshop at University of Michigan, Ann Arbor, MI (10/2002)
- Participated in Workshop on Data Derivation and Provenance in Chicago, IL (10/2002)
- Defined XML schema based on portal objects schema (10/2002)
- Started development of XSLT translators (10/2002)
- Presented and refined demos for SC2002 (10/2002)
- Integrated CHEF into CMCS portal (10/2002)
- Developed electronic notebook content (10/2002)
- Developed pedigree browser (11/2002)
- Developed basic 2D plotting applet (11/2002)
- Presented demonstrations of CMCS capability in SciDAC booth at SC2002 in Baltimore, MD (11/2002)
- Presented CMCS SC2002 Demos to project team (12/2002)
- Drafted plan for starting collaboration with PrIME project (12/2002)

Project Management, Structure and Planning

The CMCS leadership coordinated the presentation of CMCS capabilities at SC2002 with the BES SciDAC applications posters and presentations in the SciDAC booth. The CTO and CIO along with working group leads actively coordinated the decisions and tasks leading up to the successful presentation of capabilities at SC2002. Decisions to build rather than buy portal software, and the intensely focused efforts leading up to SC2002 lead to resource distribution issues within the project that will be addressed by the POC team, and discussions with the sponsor.

CMCS Highlights

The CMCS Chemical Informatics Portal

Chemical Informatics Portal Demonstration

At the SC2002 Conference, the CMCS team gave the first public demonstrations of the Multi-scale Chemical Sciences portal and infrastructure. The portal serves as the web interface for the adaptable informatics infrastructure being developed by the CMCS team and piloted within the chemical science community. The data infrastructure takes advantage of a variety of standards and open-source information technologies to provide an unprecedented ability to share data, data pedigree, and project information within groups and across communities. The portal, which can easily be enhanced and customized through the inclusion of new 'portlets', includes real-time collaboration capabilities, search and notification tools, and a pedigree browser. To support the chemistry community, the CMCS team has integrated a variety of powerful chemistry applications, data viewers, and data translators. .

The demonstrations at SC2002 followed the work of a hypothetical kineticist who uses the portal to discover a recent electronic structure calculation for a key species in a combustion model he is in the process of developing. The kineticist is able re-optimize the overall thermochemistry used in his model using a portlet to access a remote Active Thermochemical Tables (ATcT) web service. The portlet and infrastructure automate the production of several derived and translated data sets that are required to analyze the new

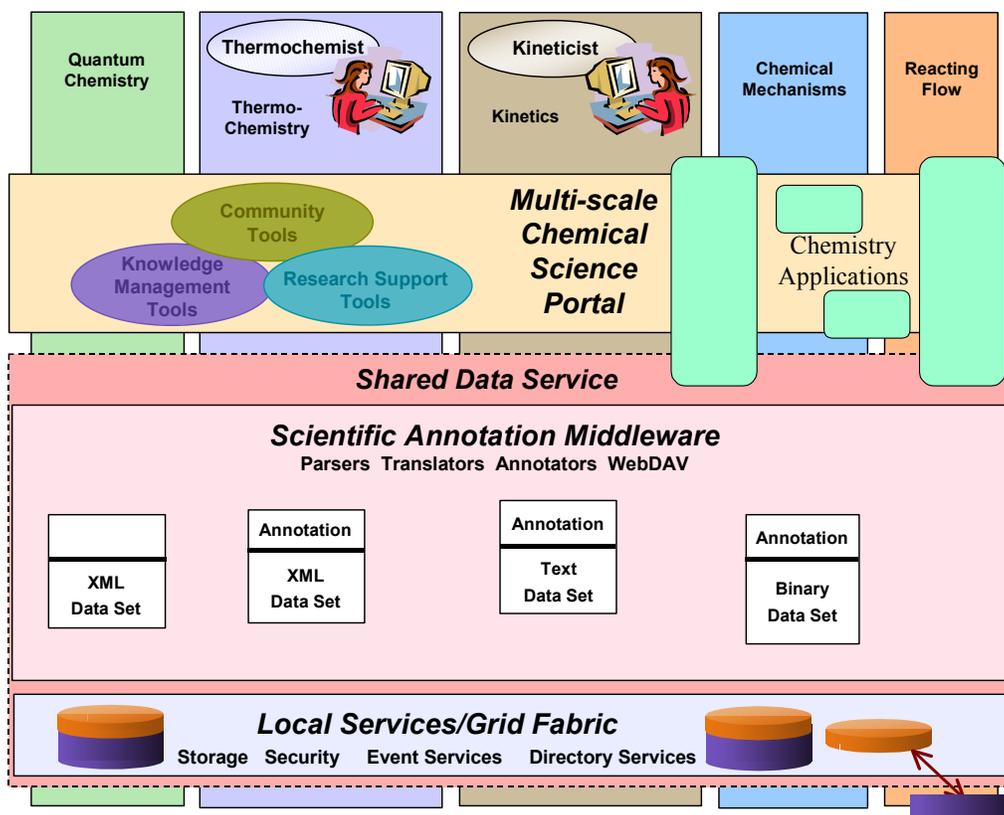


Figure 1. A diagram representing the major conceptual elements of the CMCS Informatics Infrastructure.

input and to format it for use in subsequent applications. The kineticist's collaborators stay informed via access to the kineticist's notes in an electronic notebook and the automatically generated pedigree information on his new results. Compared with the current process, the kineticist spends less time in discovering data, translating and transforming it, gathering applications, and documenting his work.

CMCS Infrastructure Overview

The diagram in Figure 1 provides an overview of the intended CMCS pilot community and the high-level CMCS architecture. The diagram shows how a CMCS user interacts primarily with the top layer, the CMCS portal and chemistry applications. The applications can appear within the portal or provide their own user interfaces and interact directly with the underlying metadata/data and other CMCS services. The portal provides an array of functionality to support group and community processes, with an emphasis on simplifying the discovery and use of data. The shared data service, shown as the second layer, provides configurable capabilities for automating the generation of metadata, translating data between standard formats, and federating multiple data stores. At the lowest layer, the portal can take advantage of existing distributed services for security, event management, and data storage.

Technical Specifications

The CMCS portal and infrastructure have been developed using Java and web standards. The portal itself is built upon an extension of the open-source Jetspeed portal

environment developed by the CHEF (CompreHensive collaborativE Framework) project at the University of Michigan. Jetspeed is in turn based on the Java Servlet standard and a variety of other open-source tools including Tomcat, Turbine, and Velocity. CHEF extends the basic portal capabilities of Jetspeed to include a variety of synchronous and asynchronous collaboration tools based on standard collaborative service programming interfaces.

A primary feature of the CMCS portal is the CMCS Explorer portlet. CMCS Explorer was developed to provide a rich interface to the CMCS data/metadata service that provides hierarchical (file-system-like) and pedigree-based browsing, data upload and download, metadata-based searching, data viewing, and access to dynamically generated translations of data. Another capability available within CMCS Explorer is the ability to register interest in specific types of data in order to receive email notification when such data is uploaded to CMCS.

The underlying data/metadata service uses the standard WebDAV protocol and is based on software developed within the Scientific Annotation Middleware (SAM) project. SAM also produces the Java Messaging Service (JMS) events used by the CMCS Notification Email Daemon (NED) to implement the notification service. At present, the open-source MySQL database is used as the low-level data and metadata store.

CMCS Pilot Use

The CMCS team is currently working with potential pilot groups to define the general and chemistry specific functionality required to support them effectively. Towards this end, several quantum chemistry, thermochemistry, and kinetics data stores are being integrated into the CMCS data service. Applications such as the Extensible Computational Chemistry Environment (Ecce) and ATcT have been extended to use the CMCS infrastructure to store data and to generate CMCS pedigree information. Third-party programming interfaces and user tutorials are also being developed. These activities are expected to culminate in a Version 1 CMCS portal that will be in active pilot use beginning this spring.

Data Pedigree

Pedigree Browsing Concept Demonstrated in CMCS Portal

The CMCS project has demonstrated a portal-based mechanism for searching and browsing the pedigree of chemical science data as part of a general metadata and pedigree management capability. Data pedigree, sometimes referred to as data provenance, documents the inputs required to create a data set, thus providing a “line of ancestors”. This allows for the categorization and tracing of the origins of scientific data, within projects and potentially across chemical scales a to the data’s ultimate origins in experimental measurements or theoretical calculations. CMCS has defined a core set of pedigree relationships such as “has inputs” and “is part of” and has adopted the Dublin Core schema for pedigree related metadata such as “creator”, “creation date”, “publication date”. CMCS has also defined a core set of chemistry-related metadata such as “chemical formula”, “Chemical Abstracts Service (CAS) number”, and various

chemical properties. Additional pedigree relationships and metadata can be defined by users as desired.

The Pedigree Browser

All pedigree relationships and metadata are stored in CMCS's WebDAV-based data repository as properties. As shown in Figure 1, the CMCS Pedigree Browser allows configurable subsets of this information to be visualized within the CMCS portal. The Browser is integrated with other portal tools and can be used to show the pedigree of items in shared project folders or search results. The Browser displays relationships as live HTML links, enabling users to quickly follow pedigree relationships, browsing, for example, from a data object to one of its inputs and from there to information about the program used to create that input.

Pedigree Examples

Hundreds of data sets of many different types, representing several chemistry-related databases, have been annotated with pedigree properties and included in demonstrations at the SC 2002 Conference and elsewhere. For example, browsing a data file returned by a search query, one might discover that it was developed as part of the Gas Research Institute Mechanism (GRI-Mech) Project. One could then browse to the project web page, to the overall GRI-Mech data collection and other project data, and to the literature

Nov 13, 2002 01:02 pm

Collaboratory for Multi-Scale Chemical Science

My Workspace CMCS team CMCS Dev

Address: http://cmcs.ca.sandia.gov:10080/slide/files/projects/ReactingFlow/Flamemaster_output

Folders Search Notify Pedigree

Select pedigree:	Pedigree Properties	Pedigree Values
CMCS Standard Pedigree <input checked="" type="checkbox"/>	Publisher	Elsevier
All Dublin Core Metadata <input type="checkbox"/>	Resource Type	text
All CMCS Metadata <input type="checkbox"/>	Modification Date	2002-10-29
All CMCS Experimental Metadata <input type="checkbox"/>	Inputs	Flamemaster Input Run 1 Input http://cmcs.ca.sandia.gov:10080/files/projects/ReactingFlow/Shock_Tl
All properties except DAV properties <input type="checkbox"/>	Keywords	kinetics
All properties <input type="checkbox"/>		DIEZKI, H.K. ADOMEIT, G. SHOCK-TUBE INVESTIGATION OF SELF-IGNITION OF N-HEPTANE AIR MIXTURE UNDER ENGINE RELEVANT CONDITIONS COMBUSTION AND FLAME, v. 93(#4) pp. 421-433 JUN 1993

Pedigree Browse

Data is linked to projects, references, inputs, and outputs.

Figure 1. A screenshot of the Pedigree Browser running in the CMCS Portal. The pedigree properties shown include a link to the Flamemaster input files used to create the data, and information about the paper in which the data were published.

references for the data. The demonstration with the longest pedigree data trail involves following GRI-Mech pedigree links to underlying Active Tables thermochemical data from which they derive, and links from these data to the Ecce/NWChem computations of molecular properties used to create them. This demonstrates the ability to track information to its ultimate origin, across physical scales and scientific applications. Another pedigree demonstration shows how electronic notebook entries about the data will also appear within the Browser as an additional type of pedigree information.

Automating the Generation of Pedigree Data

To be visible to CMCS tools, pedigree data must be stored as WebDAV properties. Properties can be generated manually using any WebDAV browser, e.g. DAV Explorer, or, in the near future, through a CMCS web form. Alternately, applications can record pedigree properties directly using the WebDAV protocol and/or by using the CMCS pedigree Java API, as in done in the Extensible Computational Chemistry Environment (ECCE) and the Electronic Laboratory Notebook (ELN). The final, and most transparent, option is to define how properties should be created from file content during upload to the CMCS repository. This capability is derived from CMCS's use of the Scientific Annotation Middleware (SAM) WebDAV service. SAM allows registration of XML-based binary format description (BFD) and XSLT templates that describe how information should be extracted from binary, ASCII, and XML files to automatically generate properties.

External collaborations and interactions

In developing these capabilities, the CMCS team has collaborated closely with the SAM project and has held ongoing discussions with Earth Systems Grid II team members. In addition, Carmen Pancerella, Larry Rahn, and Jim Myers presented CMCS pedigree concepts and participated in general metadata discussions at the Workshop on Data Provenance and Data Derivation, Chicago, IL, October 2002.

Active Thermochemical Tables

ATcT Web Service Demonstrated in CMCS Portal

The development of the first version of Active Thermochemical Tables (ATcT) and the successful demonstration of this application within the Collaboratory for Multi-scale Chemical Science (CMCS) Web-based portal are significant steps towards the CMCS goal to enhance chemical science. ATcT are a novel scientific application, centered on a distinctively different paradigm of how to obtain reliable thermochemistry. As opposed to conventional thermochemical tables, which are based on sequentially developing thermochemical values for the chemical species of interest, Active Thermochemical Tables are based on the thermochemical network approach.^{i,ii} Several successful presentations and demonstrations of ATcT and its integration have been conducted, including at Supercomputing 2002 and Grid2002. The broad availability of the ATcT approach and data products promises a new paradigm in the chemical science community with rapidly accessible, accurate, reliable, and self-consistent thermochemistry data. This data is crucial in kinetics and chemical mechanism research, in predictive modeling of chemical systems, and in many industrial applications.

ATcT web services were demonstrated as a use case for a kineticist who is seeking thermochemical data for methyl peroxy radical (CH_3OO) in the context of investigations

of combustion modeling in a Homogeneous Charge Compression Ignition (HCCI) engine. A search within the CMCS portal reveals recent high-quality calculations that provide needed bond energies. The CH₃OO dataset is integrated it into his workspace of ATcT thermochemistry tables, established in earlier work and stored in his CMCS workspace. As shown in Fig. 1, the kineticist activates the ATcT portlet by clicking on Active Tables capability shown on the left bar of his HCCI team workspace. He imports his updated ATcT data into the Active Tables service. He is then able to run the Active Tables optimization that produces a state-of-the-art value for the enthalpy of formation of methyl peroxy radical. The resulting thermochemical network displayed by ATcT is shown in Fig. 1. Then he can export to CMCS an optimized table of thermochemistry data for methyl peroxy. The exported data automatically inherit the Active Table's pedigree, including new pedigree information provided by the kineticist, namely, the reference to the calculation.

Additional features of the demonstration included an automatically generated translation (via CMCS middleware) of the new Active Tables data from XML into the commonly used JANAF format and some resulting automatic operations. Additional features of the

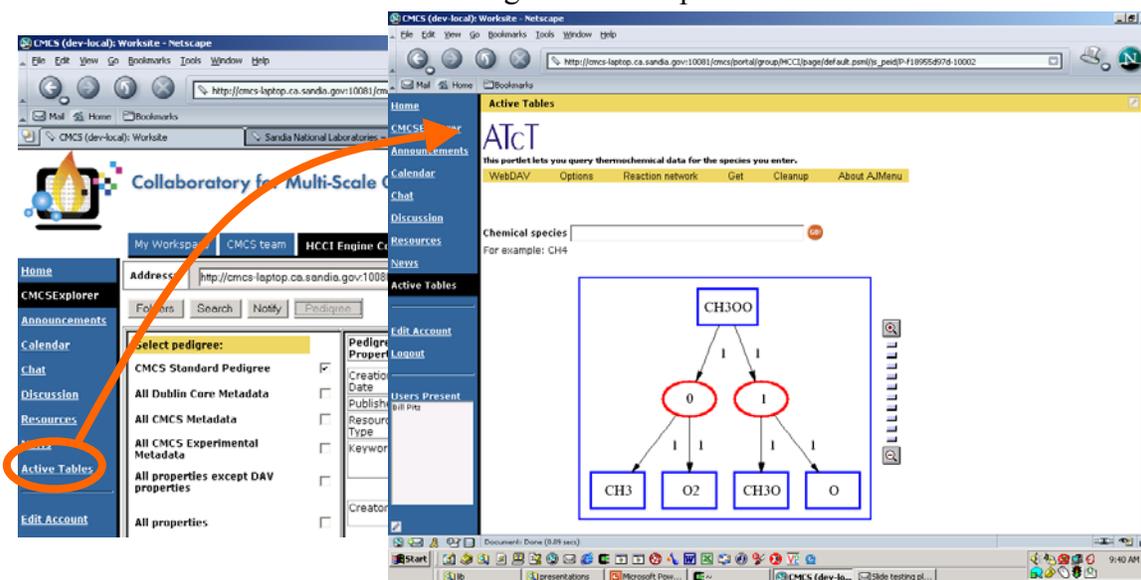


Figure 1. The Active Tables (ATcT) Web service is available from the HCCI Project Team workspace in the CMCS Portal (shown from view of pedigree of computational data used provide new inputs into Active Tables).

demonstration included automatically generated translations of the new ATcT data. An event generated by the appearance ATcT data in the CMCS repository triggers automated creation of derived data sets including a 'NASA5' polynomial fit to the data, which is required by common combustion kinetics modeling software, and tabular data that can be visualized within the CMCS portal to assess the quality of the fit to the ATcT data.

Active Tables development

ATcT has several functional parts: the computational kernel, the underlying libraries defining the thermochemical network, a user interface, and a Web services framework. The ATcT kernel is quite complex. The 1.0 Beta release has nearly 40,000 lines of Fortran 95 code. The developed capabilities include a several approaches to calculate the

partition function for the chemical species of interest and a simultaneous solution of the network via χ^2 minimization, resulting in a set of preferred thermochemical values for the chemical species involved. It also includes a large set of user-selectable options (for example, which ATcT libraries to access and a selection of output data parameters such as temperature schedule, output units, etc.), as well as internal versioning of data contained within the ATcT libraries. During the demonstrations, the released version performed flawlessly, exactly as designed and expected.

The underlying data needed by ATcT is organized in a number of Libraries and Notes. Libraries are large collections of data, typically generated by thermochemical committees who oversee and anoint the scientific soundness of the content. Notes are lighter versions of Libraries, typically associated with individual users or collaborative workgroups. Another library (JANAF Library) contains a large number of species from the popular JANAFⁱⁱⁱ tabulation (almost all of its gas-phase species). The Gurvich Library contains a similar (albeit smaller) set of tabulated enthalpies extracted from the popular Russian tabulation of Gurvich et al.^{iv} Similarly, a slightly smaller library contains a tabulation of CODATA-recommended^v selected enthalpies of key chemical species. In addition, a small library (Pitz Notes) that contains a selection of networked data (relating to methyl-peroxy radical) that were directly used in the demonstrations was also assembled.

Also developed were the ATcT GUI and the infrastructure aspects of the distributed architecture that were needed to efficiently interface the ATcT kernel with the CMCS infrastructure. An integral part of this architecture is the exposure of the complex ATcT Fortran program through sophisticated, possibly distributed, Web services. These services allow easier integration within a collaborative framework as anticipated within the overall SciDAC project goals. The necessary interfaces were developed as part of this service. Among others, these allow the transformation of data needed and produced by the native ATcT component to an XML format that can be reused with ease within the rest of the CMCS framework.

As the development of the services is independent of the presentation technology, the access mechanism to the functionality of the ATcT program has been provided through a web portal. Related portlets were developed to provide a menu allowing the user of the Active Thermochemical Table component to access suitable tasks such as displaying a graph of chemical reactions or querying thermochemical data on a chemical species.

Some of the technologies used within this project have been developed and explored as part of other projects, including the DOE SciDAC CoG Kit project and its interactions with the NSF Alliance Portal Expedition, and CHEF (CompreHensive collaborative Framework, U. Mich). Additional collaborative contributions occurred through the SAM (Scientific Annotation Middleware project, DOE NC program) project and a variety of open source technologies, some of which are listed in CMCS Portal Infrastructure Highlight.

Some Scientific Ramifications of the current ATcT development

Ruscic et al.^{vi,vii} have recently convincingly shown that the generally accepted bond dissociation energy in water is too high by $\sim 2 \text{ kJ mol}^{-1}$ and proposed its revision. This study left open a small discrepancy of $\sim 20 \text{ cm}^{-1} = 0.2 \text{ kJ mol}^{-1}$ between the photoionization measurements and the value obtained by Rydberg-tagging techniques.

Subsequently, Ruscic et al.^{viii} have addressed the resolution of this remaining small discrepancy by using the present network approach (in a “manual” mode at the time). The network approach shows convincingly that the Rydberg tagging result is slightly off. While the final result of the analysis of this problem via ATcT is no different than the initial “manual” approach, its implementation through ATcT is clearly more elegant and substantially more efficient, making the “manual” approach obsolescent.

Additionally, the process of construction of the critically-evaluated networked data in the Main Library and its interim harmonizations and solutions is not only providing new and improved thermochemical quantities for several very basic (“key”) chemical species, but also providing pointers to future experimental measurements that will “tighten” the network and bring improvements to the derived thermochemistry (one good example is the planned re-examination of ozone via photoionization).

Publications and references

Gregor von Laszewski, Branko Ruscic, Patrick Wagstrom, Sriram Krishnan, Kaizar Amin, Sandeep Nijsure, Reinhardt Pinzon, Melita L. Morton, Sandra Bittner, Mike Minkoff, Al Wagner, and John C. Hewson. *A Grid Service Based Active Thermochemical Table Framework in Third International Workshop on Grid Computing, Lecture Notes in Computer Science*, Baltimore, MD, 18 November 2002.

ⁱ B. Ruscic, J. V. Michael, P. C. Redfern, L. A. Curtiss, and K. Raghavachari, *J. Phys. Chem. A* **102**, 10889 (1998)

ⁱⁱ B. Ruscic, M. Litorja, and R. L. Asher, *J. Phys. Chem. A* **103**, 8625 (1999)

ⁱⁱⁱ M. W. Chase, C. A. Davies, J. R. Downey, Jr., D. J. Frurip, R. A. McDonald, and A. N. Syverud, "*JANAF Thermochemical Tables*", 3rd ed.; *J. Phys. Chem. Ref. Data* **14**, Suppl. 1 (1985)

^{iv} L. V. Gurvich, I. V. Veyts, and C. B. Alcock, "*Thermodynamic Properties of Individual Substances*", Vol. 1, Parts 1 and 2, Hemisphere, New York, 1989; *id.*, Vol. 2, Parts 1 and 2, Hemisphere, New York, 1991.

^v J. D. Cox, D. D. Wagman, and V. A. Medvedev, "CODATA Key Values for Thermodynamics", Hemisphere, New York, 1989.

^{vi} B. Ruscic, D. Feller, D. A. Dixon, K. A. Peterson, L. B. Harding, R. L. Asher, and A. F. Wagner, *J. Phys. Chem. A* **105**, 1 (2001)

^{vii} B. Ruscic, A. F. Wagner, L. B. Harding, R. L. Asher, D. Feller, D. A. Dixon, K. A. Peterson, Y. Song, X. Qian, C.-Y. Ng, J. Liu, W. Chen, and D. W. Schwenke, *J. Phys. Chem. A* **106**, 2727 (2002)

^{viii} B. Ruscic, R. Pinzon, and M. L. Morton, to be published