CONF-9509139--10

LA-UR- 95-4054

Title: A PLUG-AND-PLAY APPROACH TO AUTOMATED DATA INTERPRETATION: THE DATA INTERPRETATION MODULE (DIM)

RECEIVED
JAN 16 1996
OSTI

Author(s): B. K. Den Hartog, J. W. Elling, S. M. Mniszewski

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Submitted to: American Chemical Society
Atlanta, GA
September 1995

MASTER

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED BS

Form No. 836 R5
ST 2629 10/91

A PLUG-AND-PLAY APPROACH TO AUTOMATED DATA INTERPRETATION: THE DATA INTERPRETATION MODULE (DIM)

B.K. Den Hartog, J.W. Elling, S.M. Mniszewski, Measurement Technologies Group, Engineering Sciences and Applications Division, Los Alamos National Laboratory, Los Alamos, NM 87545.

Abstract

The Contaminant Analysis Automation (CAA) Project's automated analysis laboratory provides a "plug-and-play" reusable infrastructure for many types of environmental assays.

As a sample progresses through sample preparation to sample analysis and finally to data interpretation, increasing expertise and judgment are needed at each step. The Data Interpretation Module (DIM) echoes the automation's plug-and-play philosophy as a reusable engine and architecture for handling both the uncertainty and knowledge required for interpreting contaminant sample data.

This presentation describes the implementation and performance of the DIM in interpreting polychlorinated biphenyl (PCB) gas chromatogram and shows the DIM architecture's reusability for other applications.

A PLUG-AND-PLAY APPROACH TO AUTOMATED DATA INTERPRETATION: THE DATA INTERPRETATION MODULE (DIM)

B.K. Den Hartog, J.W. Elling, S.M. Mniszewski, Measurement Technologies Group, Engineering Sciences and Applications Division, Los Alamos National Laboratory, Los Alamos, NM 87545.

The need for automated environmental data interpretation

The United States Department of Energy (DOE) faces a daunting environmental remediation problem. But before remediating any location suspected of contamination, the site must be characterized by laboratory analysis. In 1995, the DOE needed two million characterizations. With increasing emphasis on restoring the environment, the combined DOE and Department of Defense (DoD) plans estimate the need for 50 million sample analyses in the year 2000 [1]. Sample analysis accounts for $30 billion dollars of the current $160 billion dollar annual environmental remediation business: by the year 2000, sample analysis could grow to $55 billion dollars of the predicted $220 billion dollar annual industry total [2]. The DOE effort requires a variety of characterizations, including organics, metals, and radionuclides. To attack this immediate and immense problem, the DOE is developing advanced technology as well as applying today's methods.

Contaminant Analysis Automation (CAA) [3] is the DOE advanced technology project for characterization. CAA automates the entire chemistry laboratory process (Figure 1). The CAA automated laboratory reduces turnaround time by switching among chemical procedures and by traveling to the characterization site to eliminate transportation delays. By separating the reusable infrastructure from analysis-specific parts of the automated laboratory, parts can be replaced to perform the analysis at hand.

The reusable infrastructure (Figure 2) includes the user interface, the controller, a database and communications protocols, as well as laboratory necessities installed in a trailer. By adding specialized equipment that follows the standard communications and control protocols [4], the laboratory is ready for a specific chemical analysis. These standardized instruments, called Standard Laboratory Modules™ (SLMs™), can be plugged into the CAA controller, human interface, database, and communications without rewriting the framework, permitting work to continue with a minimum of disruption. Completing the automated laboratory analysis, the Data Interpretation Module (DIM) [5] interprets the data

and reports the nature of the sample. Just as CAA is reusable for many analyses, the DIM framework is similarly adaptable.

A specific analysis: Polychlorinated Biphenyls (PCBs)

CAA's first chemical analysis for automation is EPA SW-846 Method 8080 for the detection and reporting of hydrocarbons in soil. The specific hydrocarbons are three polychlorinated biphenyls (PCBs) mixtures known as Aroclor® 1242, Aroclor 1254 and Aroclor 1260. The differences in their names refer to the differing weight percentage chlorine in each Aroclor.

A sample travels through the laboratory in three steps: (1) sample preparation, (2) sample analysis, and (3) data interpretation. Sample preparation chemically and physically prepares the sample for chemical measurement. For PCBs in soil, sample preparation converts a beaker of dirt into a liquid sample suitable for injection into a gas chromatography instrument. SLMs such as sonicators and concentrators prepare the sample, coordinated by the controller through standardized communication protocols and standardized behavior.

Sample analysis produces data from the liquid sample by injecting the liquid sample into a gas chromatography instrument. Analytical Instrument Modules (AIMs)--such as the gas chromatography instrument for PCBs--measure the sample and produce data, coordinated by the controller through standardized communication protocols and standardized behavior. AIMs are generally more complex than SLMs, and assessing AIM behavior is more complicated. The PCB analysis measures the PCB content of a sample through gas chromatography with electron capture detection. Each PCB is a mixture of 50 to 70 chlorinated biphenyl compounds. Each of these compounds creates a peak in the chromatogram. A chromatogram displays the amount of a material detected; the time scale separates the materials by volatility.

Data interpretation converts the data into a quantitative assessment of the sample. Data Interpretation Modules (DIMs)--including the PCB DIM--identify and quantify targets using automated techniques and information about the techniques and using standardized communication and accepted behaviors. The gas chromatography instrument is calibrated by producing chromatograms of fifteen standard samples before each batch of samples is analyzed. These standard chromatograms are five concentrations of each of the three Aroclors (Figure 3). By comparing the chromatogram of an unknown sample to these fifteen standards, a chemist reports the quantity of each of the three Aroclors believed

present. Errors in manual quantitation of PCB chromatograms typically range from 20% to 600% [6].

Complexities arise in both manual and automated interpretations in at least three ways: first, the presence of multiple Aroclors in a sample causes their signals to interfere with one another. Second, sites containing PCBs often also contain oil, and oil contamination produces nonspecific responses which obscure the PCB signal (Figure 4). Third, systematic and specific weathering degrade the PCB, causing difficulty in comparing its chromatogram to standards.

The difficulty of automation increases as the sample passes from sample preparation to sample analysis and to data interpretation, reflected in the expertise and training of the persons traditionally performing the task. The DIM faces uncertainty in how to do its task and, at the same time, requires much knowledge about its task.

Data interpretation: Is the data analyzable?

Data interpretation requires (1) analyzable data, (2) assurance that the laboratory is not contaminated and is running within the range of a valid calibration, and (3) a way to determine and report the contents of the data.

To address whether the data is analyzable, the Analysis Assessment System (AAS) (Figure 5) studies the chromatogram from the gas chromatography instrument, forming a table of symptoms such as peak tailing or clipping. The AAS examines the symptom table through its expert system [7], diagnosing whether the gas chromatography instrument is running reliably. The AAS is vital to an AIM--that is, an integrated, automated gas chromatography instrument in the case of PCB analysis. Without runtime feedback, a over-concentrated sample can contaminate the instrument's column or a septum can start leaking, rendering the output chromatograms useless, but an autoloading instrument will continue to run samples from its tray and waste the samples. With performance feedback from the Analysis Assessment System, the gas chromatography instrument can stop operation and prevent costly delays or possibly complete loss of samples.

Data interpretation: Is the laboratory reliable?

The DIM assists the chemist and the controller with Quality Assurance and Quality Control by reporting contents of blanks and controls, and by flagging over-concentrated samples. The chemist and the controller monitor the performance to ensure that a

calibration is still valid and that contamination of the laboratory has not occurred.

Data interpretation: What is in the data?

Once the analysis assessment system confirms the AIM was running correctly, and the chemist and controller verify that the laboratory is in calibration, the DIM begins processing the data.

The PCB DIM uses five pattern recognition techniques, each with its own strengths, to quantify Aroclors 1242, 1254 and 1260 in a gas chromatogram.

Figure 6 shows the DIM user interface. A row of logbooks, opened with a point-and-click interface, allows the user to look at details when desired. The object-oriented design results in objects: that is, icons with actions associated with them. The user connects the interfaces by clicking objects of the controller, database and Matlab® connections. Buttons open palettes containing the objects such as models, samples (Figure 7), and methods needed to describe and analyze a chromatogram. The user selects a gas chromatography sample object on the DIM screen and views its description by clicking the object. The user enters the sample type (such as unknown or blank), and selects an analysis strategy to run on it the sample. In Figure 6, the file 016f0101 is being analyzed as an unknown, using the analysis "Aroclor1".

The strategy named Aroclor1 includes objects with plugged-in routines for preprocessing method, the five pattern recognition techniques, combiner logic for selecting the best answer from the techniques, and final answer procedures for archiving and reporting results. The preprocessing object's plug-in Matlab baseline removal program executes. The graphics and messages from external packages such as Matlab and those generated from the DIM are displayed and logged.

Once the preprocessing completes, the pattern recognition techniques begin to execute. Colors change as objects pass from *idle* to *starting*, and so on until they are *done*, making the status of the system visible at a glance from across a laboratory.

The PCB DIM uses a neural network, peak area analysis, principle component regression, iterative multiple linear regression, and a peak interpolation method as pattern recognition techniques. Pattern recognition techniques return identity and quantity information for each Aroclor. Some techniques return only identity data; others, only quantity (Figure 8). The combiner object is a small expert system whose rules and procedures synthesize a final result

from the pattern recognition techniques' results. It is the combiner's duty to make sense of these results.

The combiner passes the final result to the answer object. The answer object archives data, converts return codes according to the sample type for use in an integrated system (e.g., if a blank contains an Aroclor, the DIM reports contamination to the controller) and passes the result and return code to an integrated system. The user can examine results from techniques, combiners, and the answer object, or view the original chromatogram.

For a single Aroclor without oil contamination, the combiner logic is trivial: every pattern recognition technique works well. The power of the DIM shows in the more difficult cases. The combiner currently handles mixtures, singles and blanks without oil. The DIM's ability to interpret mixtures contrasts to the inability of peak area analysis, which emulates the way a chemist would interpret the chromatogram (Figure 9). Despite the confusion in the peak area method's result, the final report on sample 034f0101 is 400 ppb Aroclor 1242 and 800 ppb Aroclor 1254--exactly as the sample was created.

DIM Architecture: How is all of this reusable among analyses?

The DIM has three layers: (1) a commercial expert system shell, (2) a base DIM that provides the services and types of objects and interfaces needed for every DIM application, and (3) an analysis-specific DIM that has had specific pattern recognition and preprocessing techniques and combiner logic inserted. The commercial expert system shell, G2, features real-time responsiveness, graphical user interface tools, object oriented modularity, fuzzy logic for handling uncertainty, and supports procedures as well as rules for precise control by the programmer.

The base DIM contains the user interface through which the chemist can build and execute analysis-specific data interpretation strategies as well as examine raw data, calibration models, and intermediate and final results. The base DIM interfaces to the rest of the automated laboratory, connecting across computers to a controller and to an object oriented database. However, it requires neither as the DIM can operate in either standalone mode or integrated with the automated laboratory. The base DIM provides connections to external signal processing packages such as Matlab and analytical chemistry laboratory software such as Target3®.

Analysis strategies consist of sample, preprocessor, method, combiner, and answer objects. The base DIM provides generic

versions of these and the environment for running them. Using an object-oriented hierarchy, all these objects inherit from the general purpose DIM-object class. DIM-objects contain information and operations generic to all parts of a strategy. Fields in a DIM-object contain state and diary information. Using the state and diary information, the base DIM connects selected DIM-objects (preprocessors, pattern recognition methods, etc.) and executes them while logging progress and errors, and triggering alarms (Figure 10).

The analysis-specific DIM has customized sample, preprocessor, method, combiner, answer and instrument objects. For the PCB DIM, a sample label contains the sample type, the name of the data file, and the requested chemical analysis. The preprocessor points to baseline removal code in Matlab signal processing software. A method object is created to point to each of the pattern recognition methods developed by experts especially for Aroclor interpretation. Individual methods provide results in standard format to the DIM. The results from the methods are combined and the most reliable answer produced by the combiner object. The report object determines return codes, archives data, and passes return code and results to the controller if necessary.

Ongoing work: combiner logic, multiple input DIMs, and new applications

Work continues on the combiner logic for oil contaminated PCB chromatograms. Once again, the plug-and-play framework approach allows the combiner architecture to be used for other analyses as well. Later work will focus on more general types of contamination and on weathering effects.

Data fusion is the study of selecting or synthesizing the best answer among many views and it describes the DIM's approach perfectly. A DIM easily accepts multiple types of input, viewing it as additional information in combining technique results. A GC/MS sample with its own preprocessing and pattern recognition methods (Figure 11) and those results could, for instance, be combined with GC result data to produce a final result.

The DIM provides the framework for similar applications in biological, pharmaceutical, and other chemical processes that previously required both too much knowledge and accommodating too much uncertainty to attempt automation.

Summary and Conclusion

The Data Interpretation Module (DIM) is (1) a reusable framework for data analysis, and (2) a method for interpreting the

presence of PCBs in gas chromatograms. Its reusability comes from a modular, hierarchical approach and standardized interfaces. Preliminary trials indicate the PCB DIM's success at interpreting mixtures of Aroclors, as well as single Aroclors, and blanks in the absence of oil contamination. Work continues to expand the DIM's performance to interpreting contaminated and weathered data. This will be accomplished in a way to provide a reusable framework for use in other analyses as well.

Acknowledgments

This work is supported by the U.S. Department of Energy under the Defense Programs Technology Transfer Initiative Program. Portions of this work were performed under a Cooperative Research and Development Agreement with Varian Chromatography Systems Business, Lockheed Missiles and Space Company, and NeuralWare, Inc.

Standard Laboratory Module and SLM are trademarks of SciBus Analytical, Inc.

Aroclor is a registered trademark of Monsanto Corporation.

G2 is a registered trademark of Gensym Corporation.

Matlab is a registered trademark of MathWorks.

Target3 is a registered trademark of ThruPut Systems.

References

[1] *Report on DOE Analytical Capacity Available to meet EM Environmental Sampling and Analysis Needs for FY 93-99* (DOE/EM-0223) Analytical Services Division (EM-263), Office of Compliance and Program Coordination, Office of Environmental Management, U.S. DOE.

[2] *Marketing Survey and Cost Analysis of Automated and Manual Analysis Methods within the US Analytical Chemistry Industry.* Communication by Design. Report to Pacific Northwest Laboratory, May 30, 1995.

[3] J.M.Griesmeyer, T.D.Urenda, R.M.Pacetti, J.J.Ferguson, *A standard control system for Modular Automation of Chemistry*. Laboratory Robotics and Automation. **6 [2]**: 79-84 (1994)

[4] T.H.Erkkila, R.M.Hollen, T.J.Beugelsdijk, *The Standard Laboratory Module: An Integrated Approach to Standardization in the Automated Laboratory*. Laboratory Robotics and Automation. **6 [2]**: 57-64 (1994)

[5] J.W.Elling, S.M.Mniszewski,J.D.Zahrt,L.N.Klatt, *Automated Chromatographic Data Interpretation Using an Expert System*. Journal of Chromatographic Science, Vol. 32 No. 6. (1994)

[6] R.P.Eganhouse,*Sources and magnitudes of bias associated with determination of polychlorinated biphenyls in environmental samples.* Analytical Chemistry. **63**: 2130-2135 (1991)
[7] A.P.Levis, R.G.Timpany, J.W.Elling, et al, *Application of Knowledge-based Network Processing to Automated Gas Chromatography Interpretation* Society of Photo-Optical Instrumentation Engineers, Proceedings, Volume 2492. (1995)

Figures

(1) Scope of automation in the CAA Project
(2) CAA architecture
(3) Chromatograms of Aroclor 1242, 1254, 1260
(4) Chromatogram of Aroclor with oil present
(5) Analysis assessment description
(6) DIM screen showing PCB analysis and MLR detail
(7) DIM screen showing available sample palette
(8) DIM screen showing method result subworkspaces
(9) DIM screen showing final results for a mixture
(10) DIM screen showing alarm response
(11) DIM screen showing two input samples

Fig 1

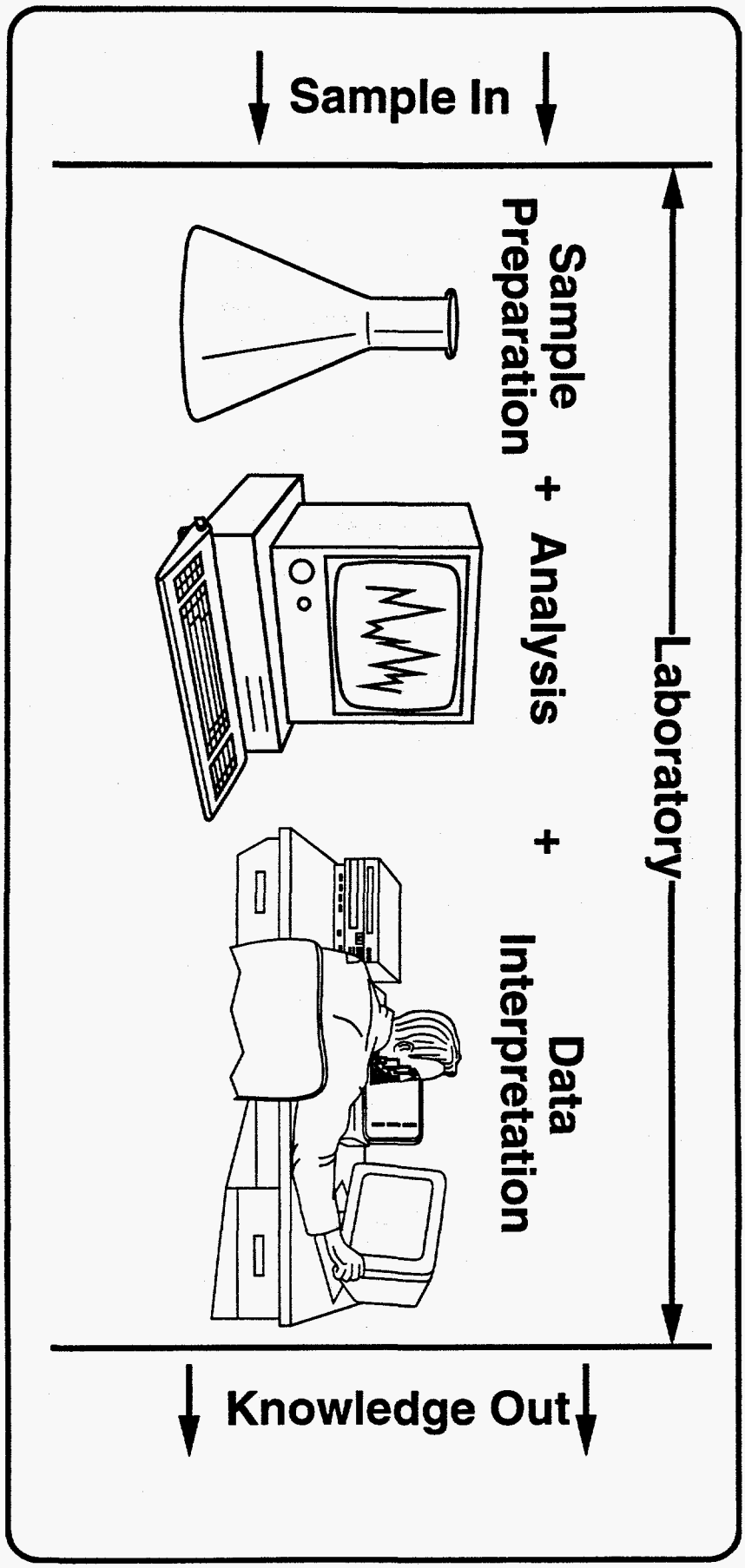

Sample In
Laboratory
Sample Preparation
+
Analysis
+
Data Interpretation
Knowledge Out

Fig 2

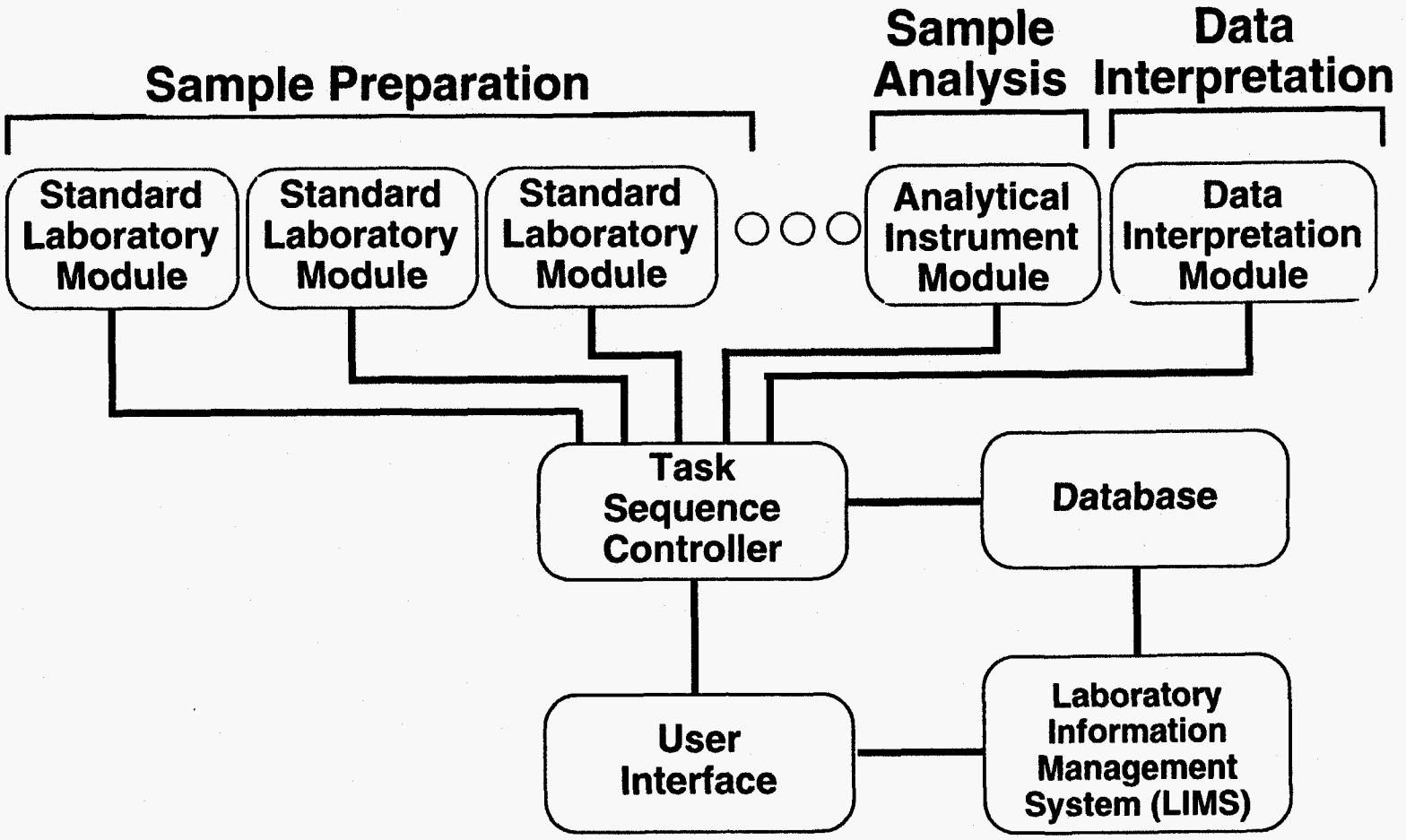

Sample Preparation
Sample Analysis
Data Interpretation
Standard Laboratory Module
Standard Laboratory Module
Standard Laboratory Module
Analytical Instrument Module
Data Interpretation Module
Task Sequence Controller
Database
User Interface
Laboratory Information Management System (LIMS)

Fig 3

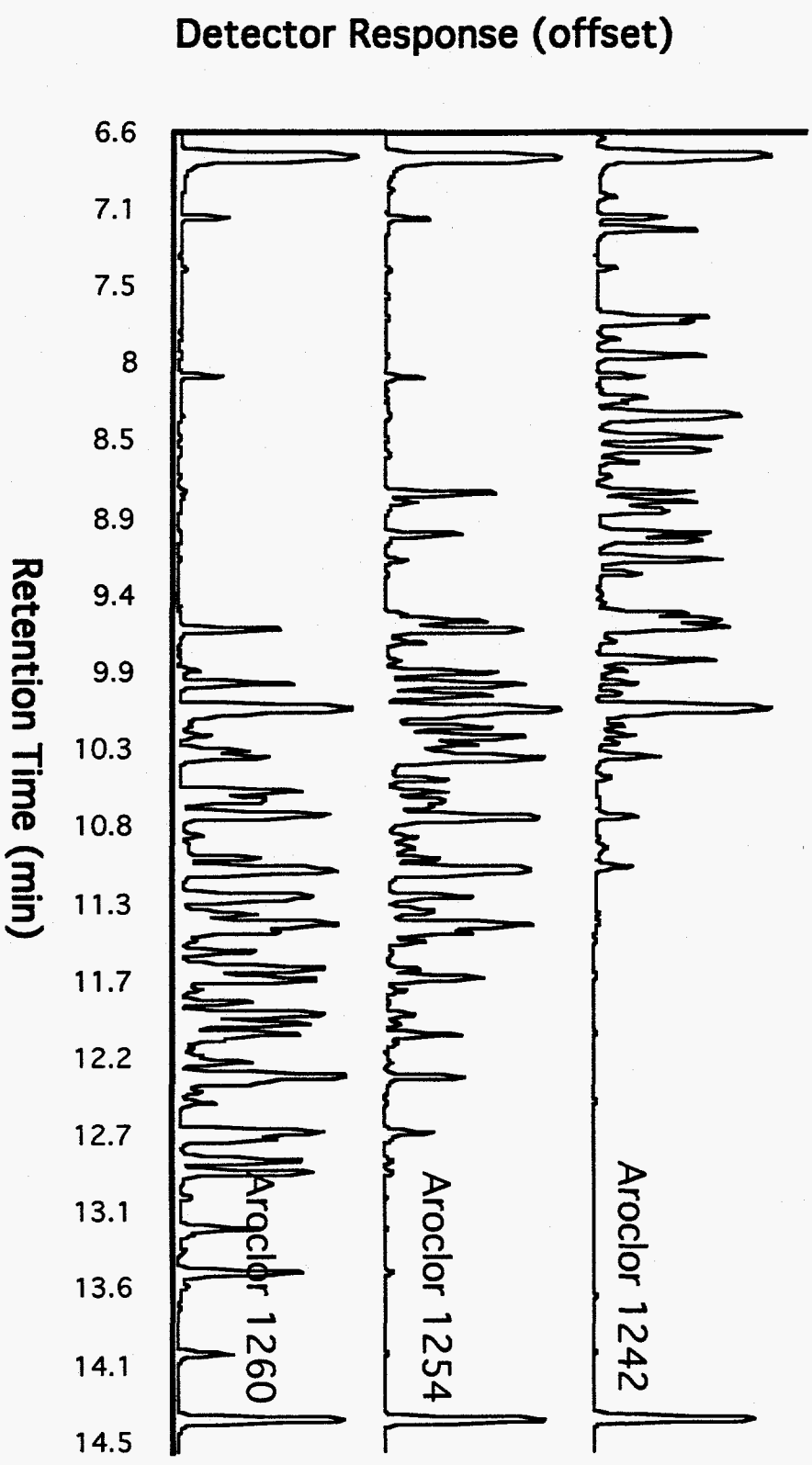

Chromatograms of Aroclor Standards
Detector Response (offset)
Retention Time (min)
6.6
7.1
7.5
8
8.5
8.9
9.4
9.9
10.3
10.8
11.3
11.7
12.2
12.7
13.1
13.6
14.1
14.5
Aroclor 1242
Aroclor 1254
Aroclor 1260

Fig 4

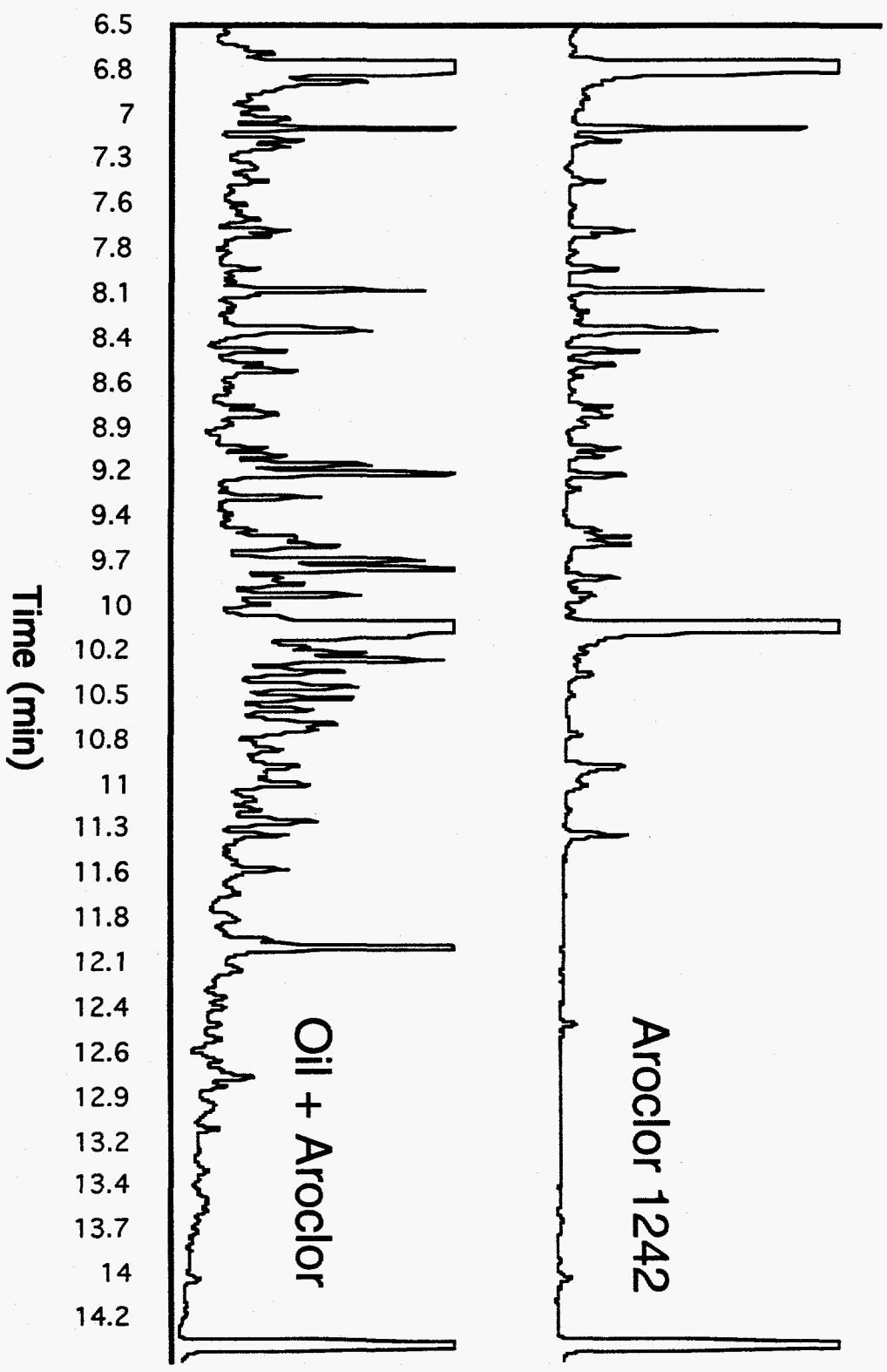

Detector Response
Time (min)
6.5
6.8
7
7.3
7.6
7.8
8.1
8.4
8.6
8.9
9.2
9.4
9.7
10
10.2
10.5
10.8
11
11.3
11.6
11.8
12.1
12.4
12.6
12.9
13.2
13.4
13.7
14
14.2
Oil + Aroclor
Aroclor 1242

Fig5

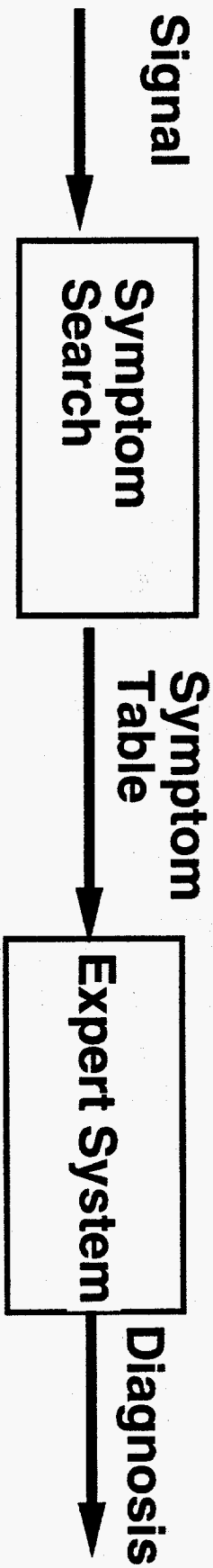
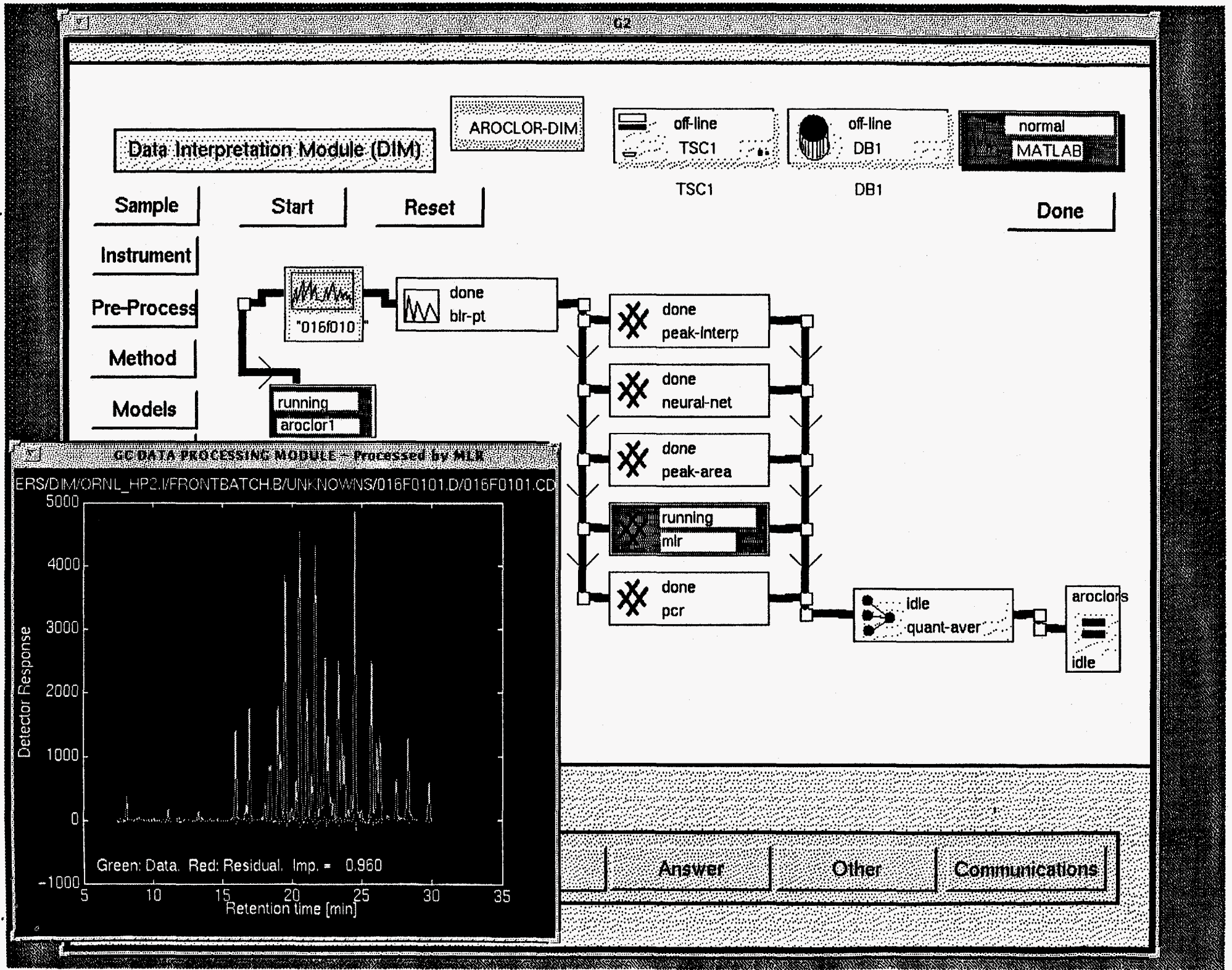
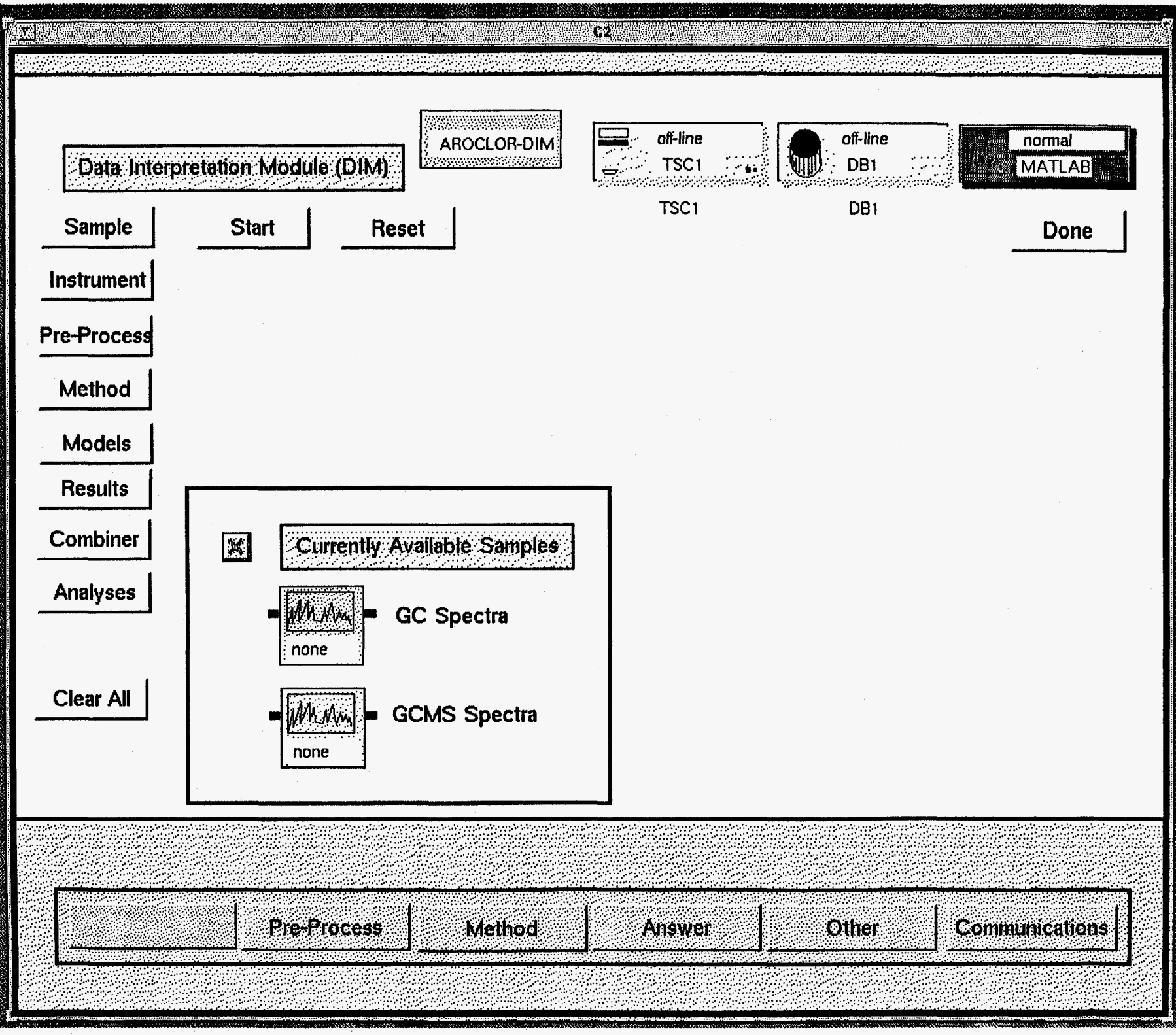
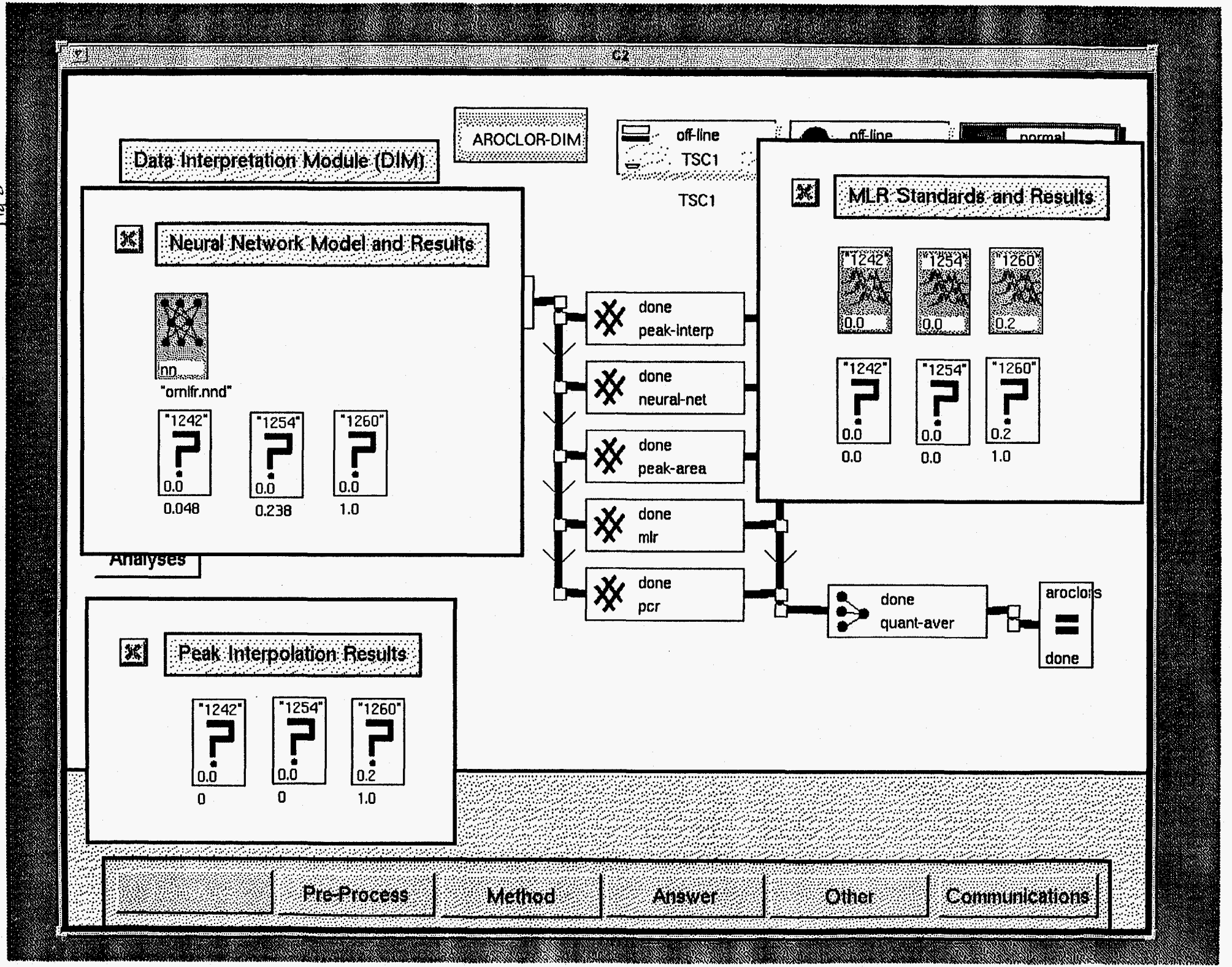

Signal
Symptom Search
Symptom Table
Expert System
Diagnosis

Fig 6


G2
Data Interpretation Module (DIM)
AROCLOR-DIM
off-line
TSC1
off-line
DB1
normal
MATLAB
TSC1
DB1
Sample
Start
Reset
Done
Instrument
Pre-Process
Method
Models
"016f010"
done
blr-pt
running
aroclor1
done
peak-interp
done
neural-net
done
peak-area
running
mlr
done
pcr
idle
quant-aver
aroclors
idle
GC DATA PROCESSING MODULE - Processed by MLR
ERS/DIM/ORNL_HP2.I/FRONTBATCH.B/UNKNOWNS/016F0101.D/016F0101.CD
Detector Response
5000
4000
3000
2000
1000
0
-1000
5
10
15
20
25
30
35
Retention time [min]
Green: Data. Red: Residual. Imp. = 0.960
Answer
Other
Communications


C2
Data Interpretation Module (DIM)
AROCLOR-DIM
off-line
TSC1
TSC1
off-line
DB1
DB1
normal
MATLAB
Sample
Start
Reset
Done
Instrument
Pre-Process
Method
Models
Results
Combiner
Analyses
Clear All
Currently Available Samples
GC Spectra
none
GCMS Spectra
none
Pre-Process
Method
Answer
Other
Communications


C2
AROCLOR-DIM
off-line
TSC1
TSC1
off-line
normal
Data Interpretation Module (DIM)
Neural Network Model and Results
nn
"ornlfr.nnd"
"1242"
0.0
0.048
"1254"
0.0
0.238
"1260"
0.0
1.0
Analyses
done
peak-interp
done
neural-net
done
peak-area
done
mlr
done
pcr
MLR Standards and Results
"1242"
0.0
"1254"
0.0
"1260"
0.2
"1242"
0.0
0.0
"1254"
0.0
0.0
"1260"
0.2
1.0
done
quant-aver
aroclors
done
Peak Interpolation Results
"1242"
0.0
0
"1254"
0.0
0
"1260"
0.2
1.0
Pre-Process
Method
Answer
Other
Communications

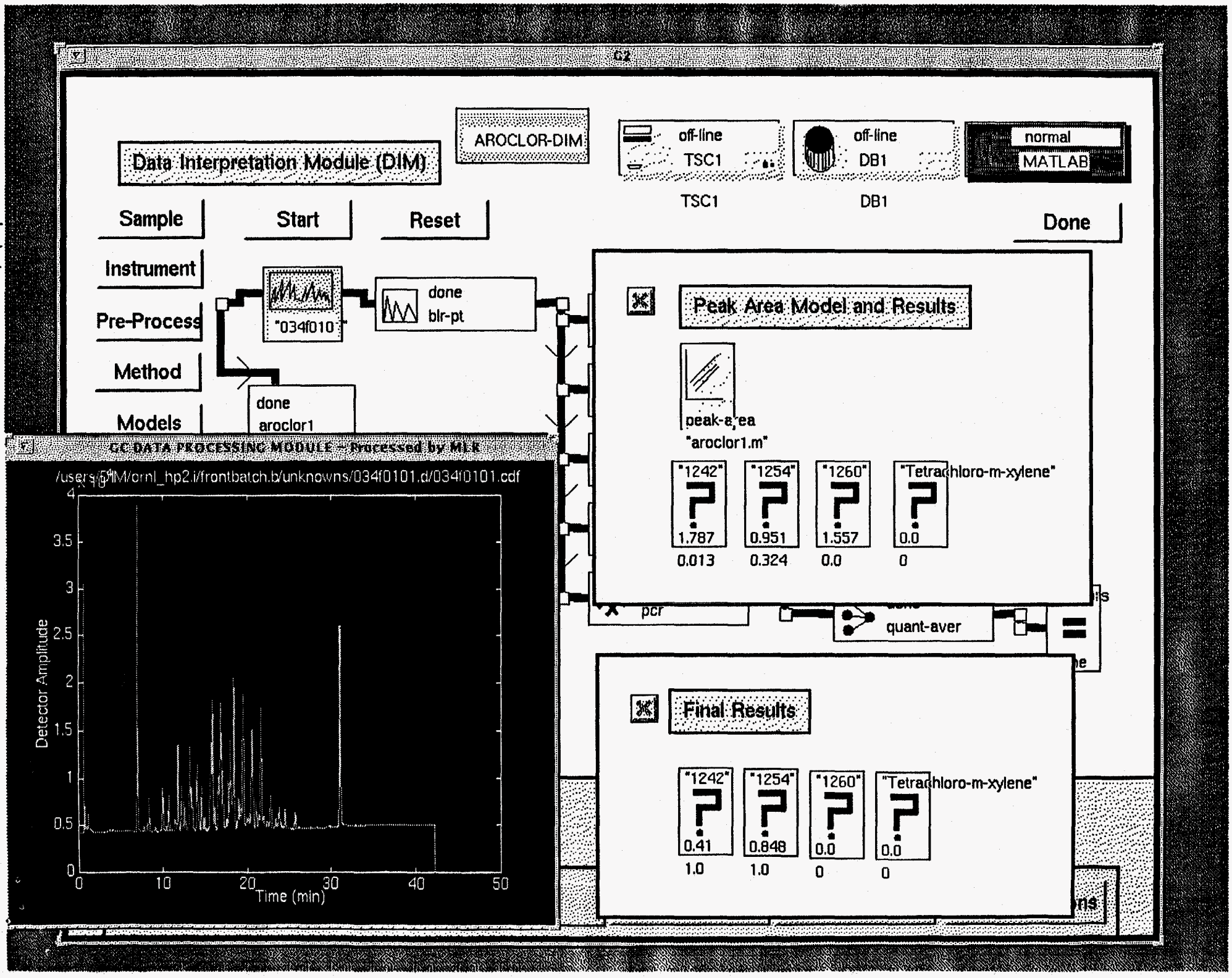

G2
Data Interpretation Module (DIM)
AROCLOR-DIM
off-line
TSC1
off-line
DB1
normal
MATLAB
TSC1
DB1
Sample
Start
Reset
Done
Instrument
Pre-Process
Method
Models
"034f010"
done
blr-pt
done
aroclor1
Peak Area Model and Results
peak-area
"aroclor1.m"
"1242"
1.787
0.013
"1254"
0.951
0.324
"1260"
1.557
0.0
"Tetrachloro-m-xylene"
0.0
0
pcr
quant-aver
Final Results
"1242"
0.41
1.0
"1254"
0.848
1.0
"1260"
0.0
0
"Tetrachloro-m-xylene"
0.0
0
GC DATA PROCESSING MODULE – Processed by MLR
/users/5IM/ornl_hp2.i/frontbatch.b/unknowns/034f0101.d/034f0101.cdf
Detector Amplitude
Time (min)
0
0.5
1
1.5
2
2.5
3
3.5
4
10
20
30
40
50

Fig 9

Fig 10

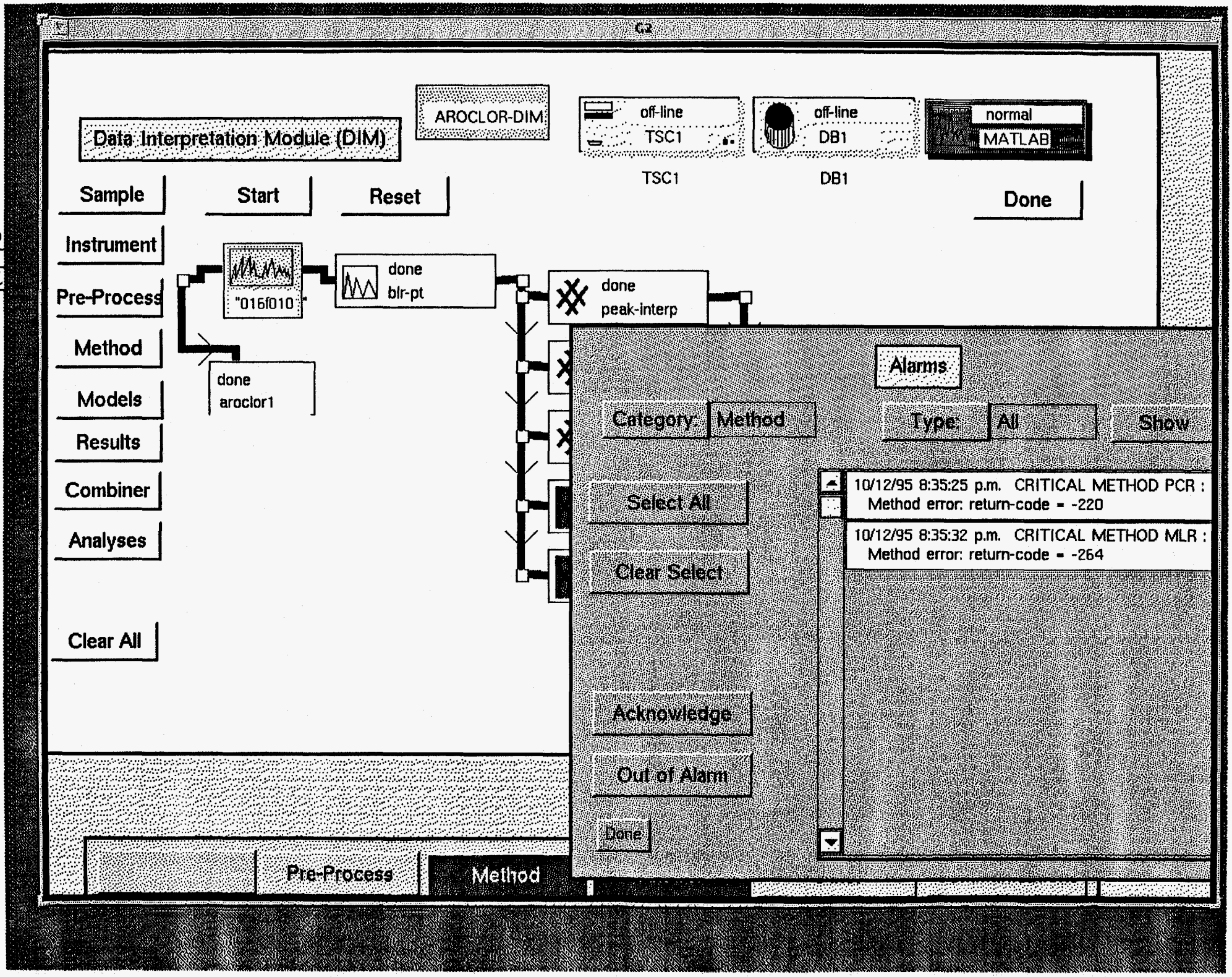
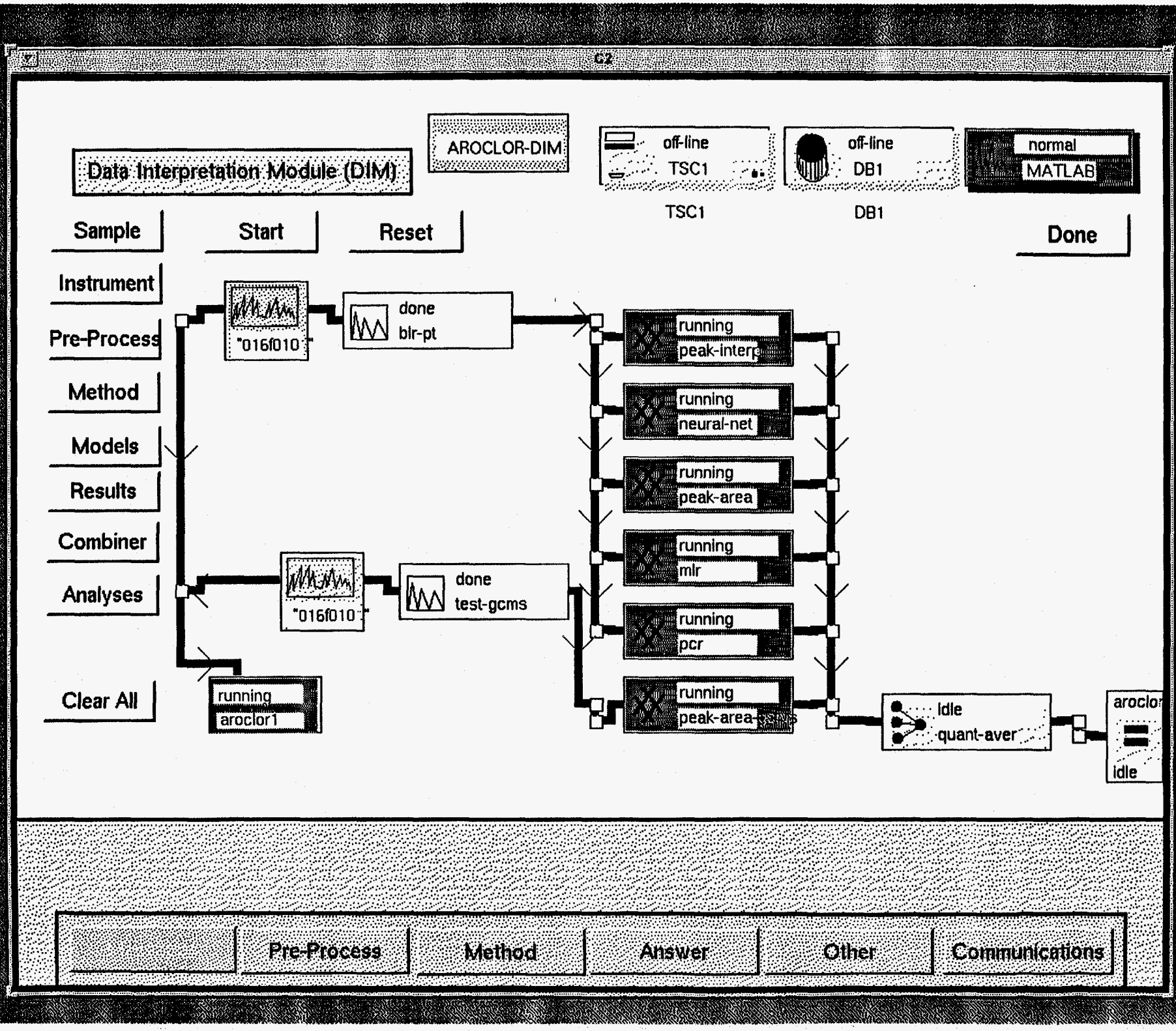

G2
Data Interpretation Module (DIM)
AROCLOR-DIM
off-line
TSC1
TSC1
off-line
DB1
DB1
normal
MATLAB
Done
Sample
Start
Reset
Instrument
Pre-Process
Method
Models
Results
Combiner
Analyses
Clear All
"016f010"
done
blr-pt
done
peak-interp
done
aroclor1
Alarms
Category:
Method
Type:
All
Show
Select All
Clear Select
10/12/95 8:35:25 p.m. CRITICAL METHOD PCR :
Method error: return-code = -220
10/12/95 8:35:32 p.m. CRITICAL METHOD MLR :
Method error: return-code = -264
Acknowledge
Out of Alarm
Done
Pre-Process
Method

Fig 11


C2
Data Interpretation Module (DIM)
AROCLOR-DIM
off-line
TSC1
off-line
DB1
normal
MATLAB
TSC1
DB1
Sample
Start
Reset
Done
Instrument
Pre-Process
Method
Models
Results
Combiner
Analyses
Clear All
"016f010"
done
blr-pt
running
peak-interp
running
neural-net
running
peak-area
running
mlr
running
pcr
running
peak-area
"016f010"
done
test-gcms
running
aroclor1
idle
quant-aver
idle
Pre-Process
Method
Answer
Other
Communications