Technical Note: SN-505

Streamlined Chemometric Model Development Using the Symbion Suite

W. M. Doyle and M. A. Power
Symbion Systems, Inc.
Tustin, CA 92780

ABSTRACT

This technical note describes the efficient and secure development and deployment of the multivariate chemometric calibrations required by process analytical instrumentation. The automated data collection capabilities of Symbion DX, RX, LX, and LRX integrated with the efficient multivariate modeling environment provided by Symbion QT facilitates rapid model prototyping, optimization, validation and deployment. Instrumental data acquired by Symbion DX, RX, LX, and LRX is directly available to Symbion QT. There is no need to separately import or reenter data. Model development proceeds in a logical sequence guided by an intuitive "workflow" environment. Final deployment is straight forward using Symbion DX, RX, or RTM.

Introduction

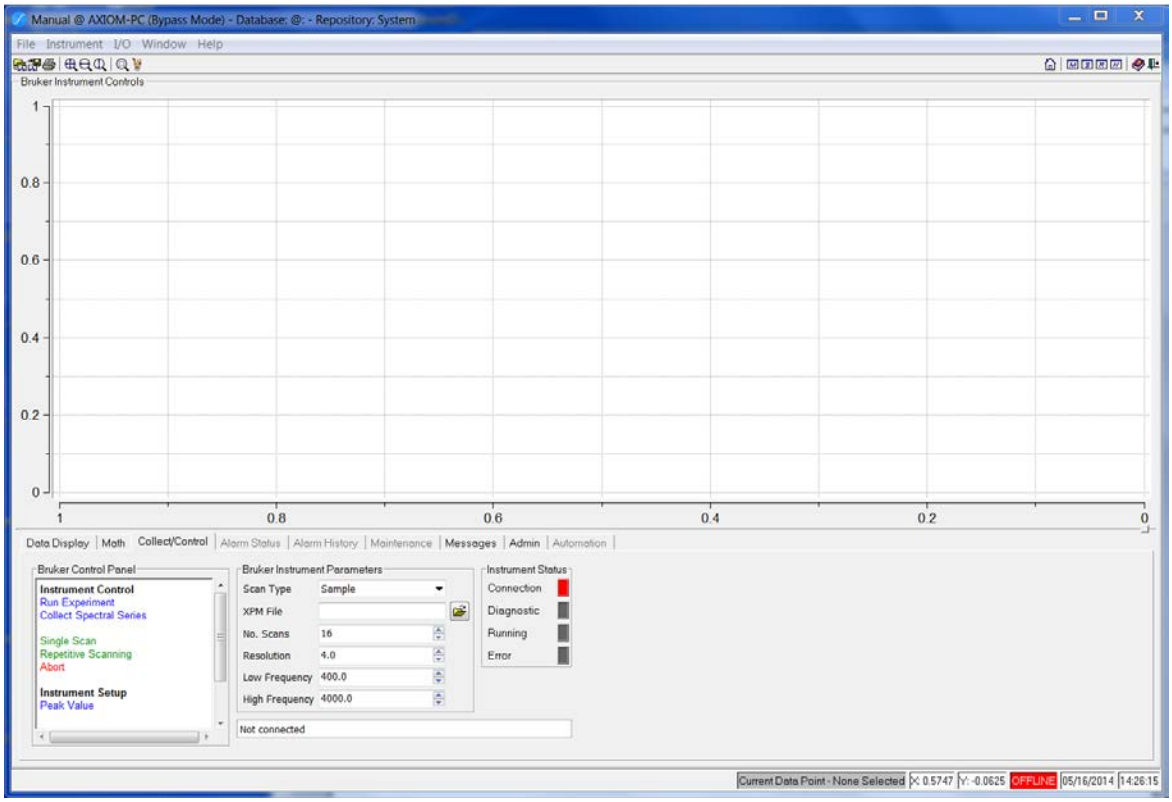
Historically, the development of multivariate chemometric calibrations has involved several discrete steps. Typically, calibration and validation samples are measured with an analytical instrument. The resulting data is then transferred to chemometric software where the calibration is developed and validated. Reference values for the calibration and validation samples must also be transferred into the chemometric software. The calibration produced is then installed in the software operating the process analyzer.

Since a given application can require hundreds, or even thousands, of calibration and validation samples the clerical task of accurately correlating all of the reference values with the respective samples is prone to error. Often the majority of the time required to develop a calibration is spent on data management rather than model development and optimization. Symbion provides significant, time saving efficiencies by eliminating the need for intermediate data transfer steps, by automating the process of data compilation, and by providing an efficient, workflow-based environment for calibration development and deployment.

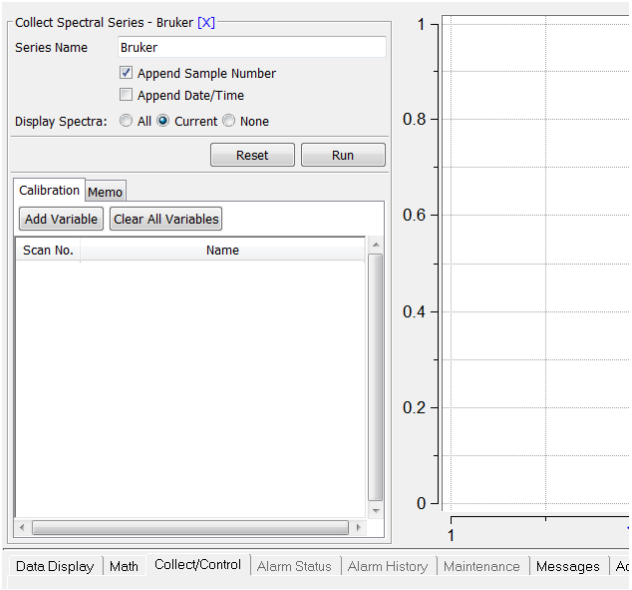
Part 1: Acquiring Data

Data acquired on any instrument can be used to build models using Symbion QT as long as the data is available in .spc, format or as either an OPUS or xy file. However, there is a unique benefit in acquiring data using one of the Symbion instrument control and analysis programs (Symbion DX, RX, LX, or LRX). In this approach, sample reference values (eg. Concentration) need only be entered once – at the time of data collection. This information is then stored along with the spectra and automatically read into QT to build the calibration matrix. To illustrate the process, we will build a calibration using mixtures of three common non-polar chemicals: cyclohexane, toluene, and hexane. These were analyzed in the mid-infrared using an Axiom Analytical DPR-207 ATR probe coupled to a Bruker IR Cube FTIR spectrometer. Data acquisition was under the control of Symbion DX.

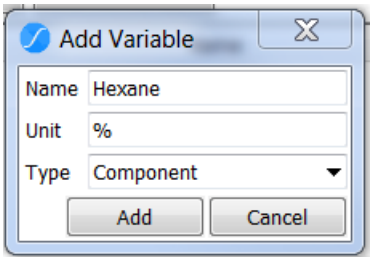
To start, we open the Symbion instrument control window. This is shown below for the case of a Bruker FTIR spectrometer.



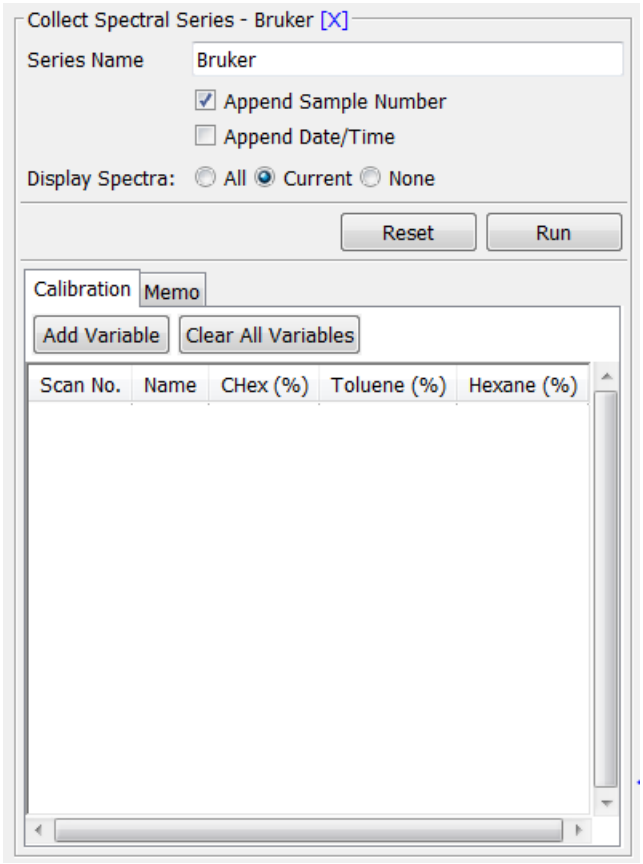
The panel at the left provides a selection of data acquisition choices. For the present illustration, we will select "Collect Spectral Series". The following pane appears.



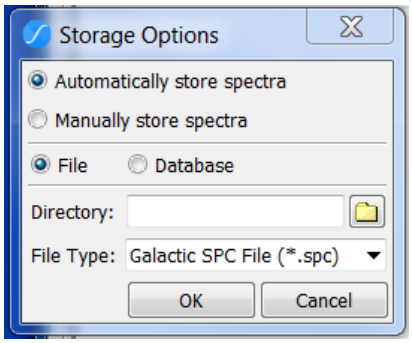
To start the process, we select “Add Variable”. The following appears.



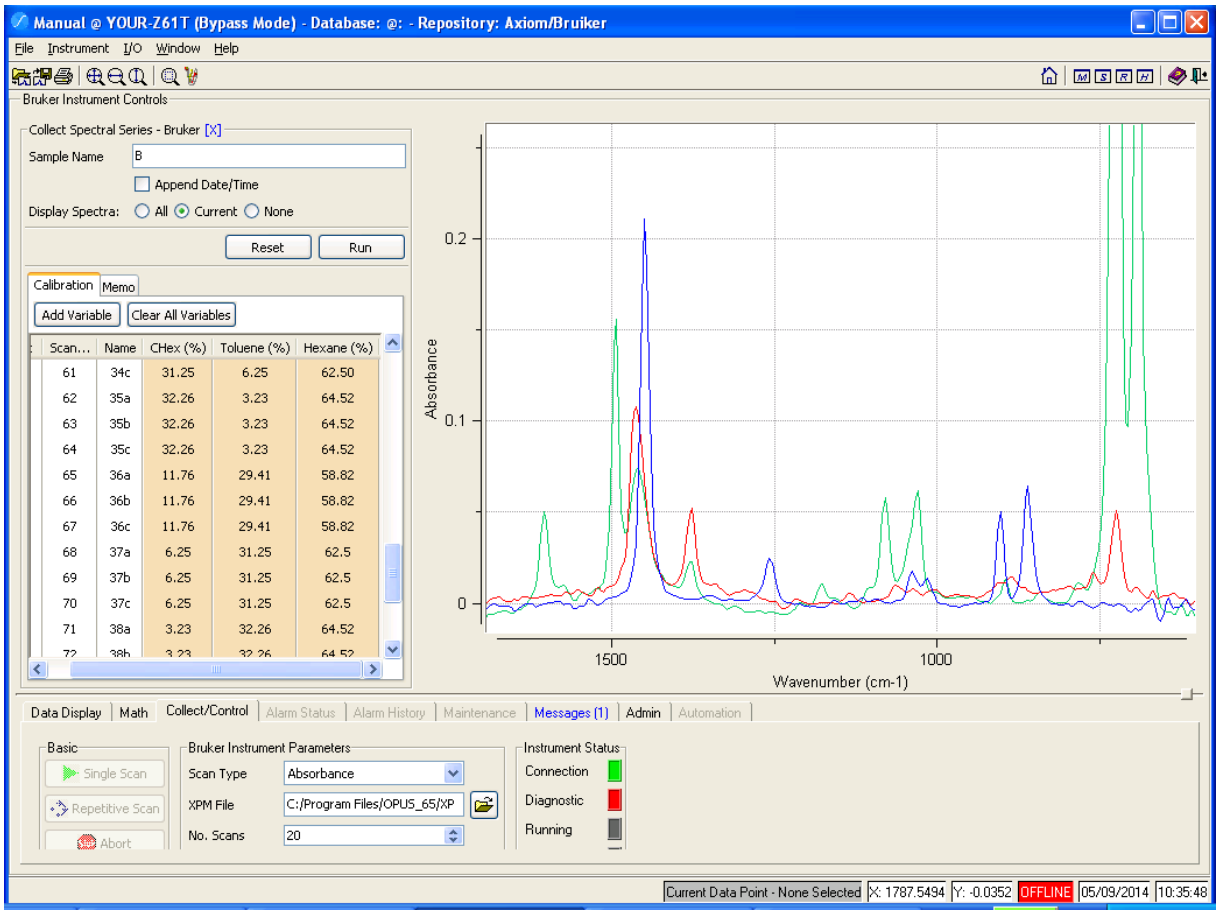
Here we have typed in the component name, Hexane, and the concentration units, %. When we press “Add” this information will be added to the data table. We repeat this step for each of the intended variables. The result is a formatted data entry table.



At this point we can decide whether to append a sample number or Date/Time stamp. The latter is appropriate if collecting continuous data for subsequent characterization. For this illustration we will use discrete samples. As each sample is loaded into the instrument, we simply press “Run”. The “Storage Options” pane will appear, giving us the choice of storing the data in the database or a specified Directory in the file system.



Pressing OK initiates data collection. We now enter the sample name and concentration values for the current sample. The example below shows a portion of the completed data table for our 82 samples. Here, we have chosen to display the current spectrum as well as two other manually selected spectra.

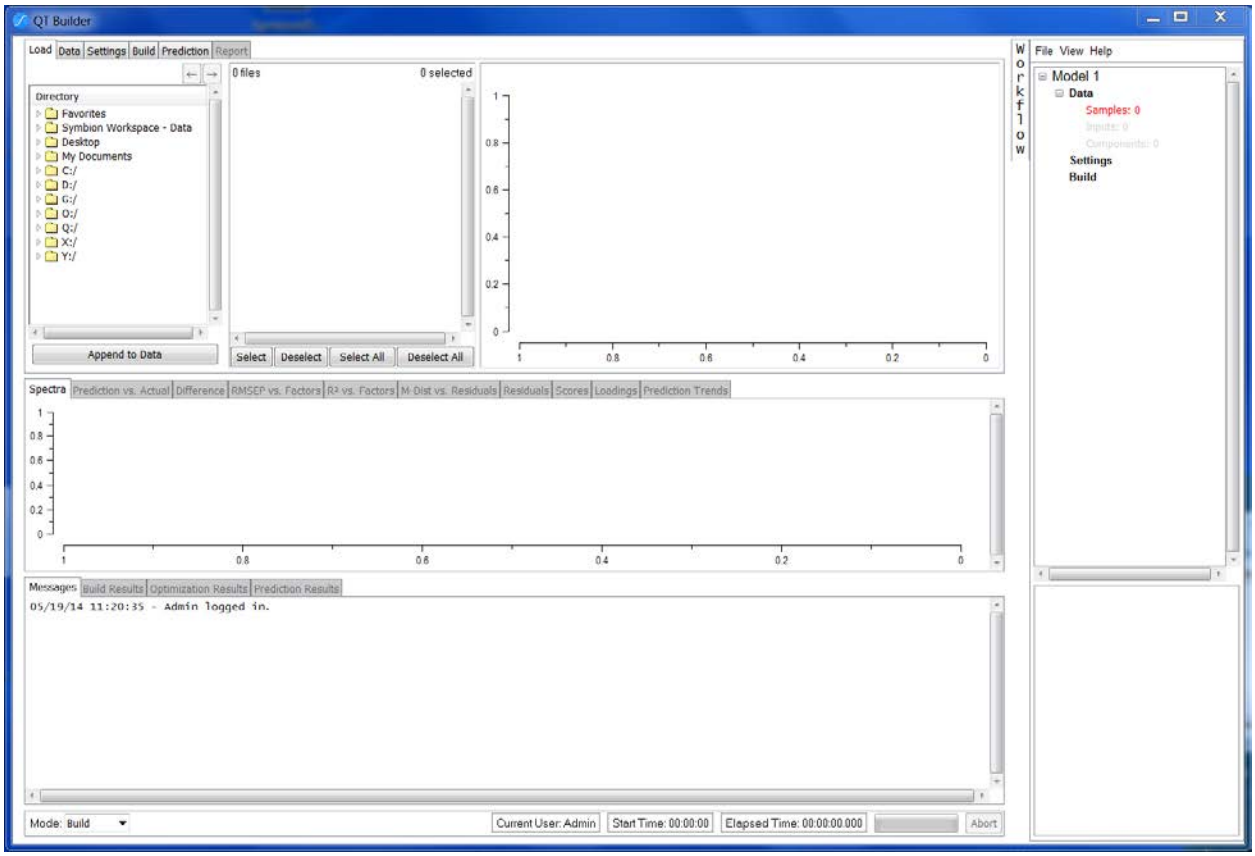


When we have completed the data collection process, we can exit the “Collect Spectral Series” mode by pressing the “X” at the top of the pane. The concentration values and other attributes will be stored in the designated location along with the associated spectra.

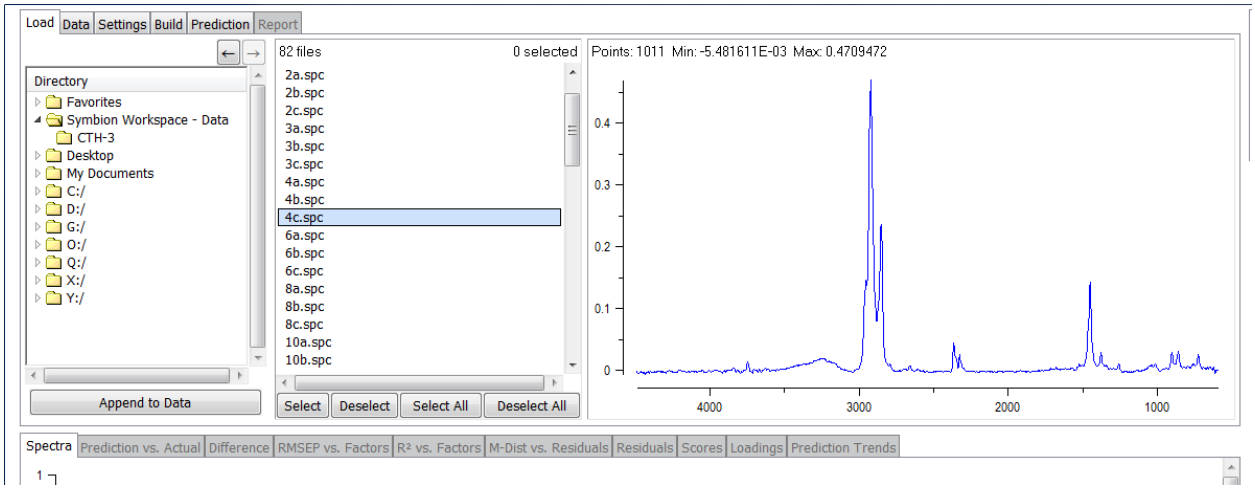
Part 2: Building a Model

2.1 Loading Data

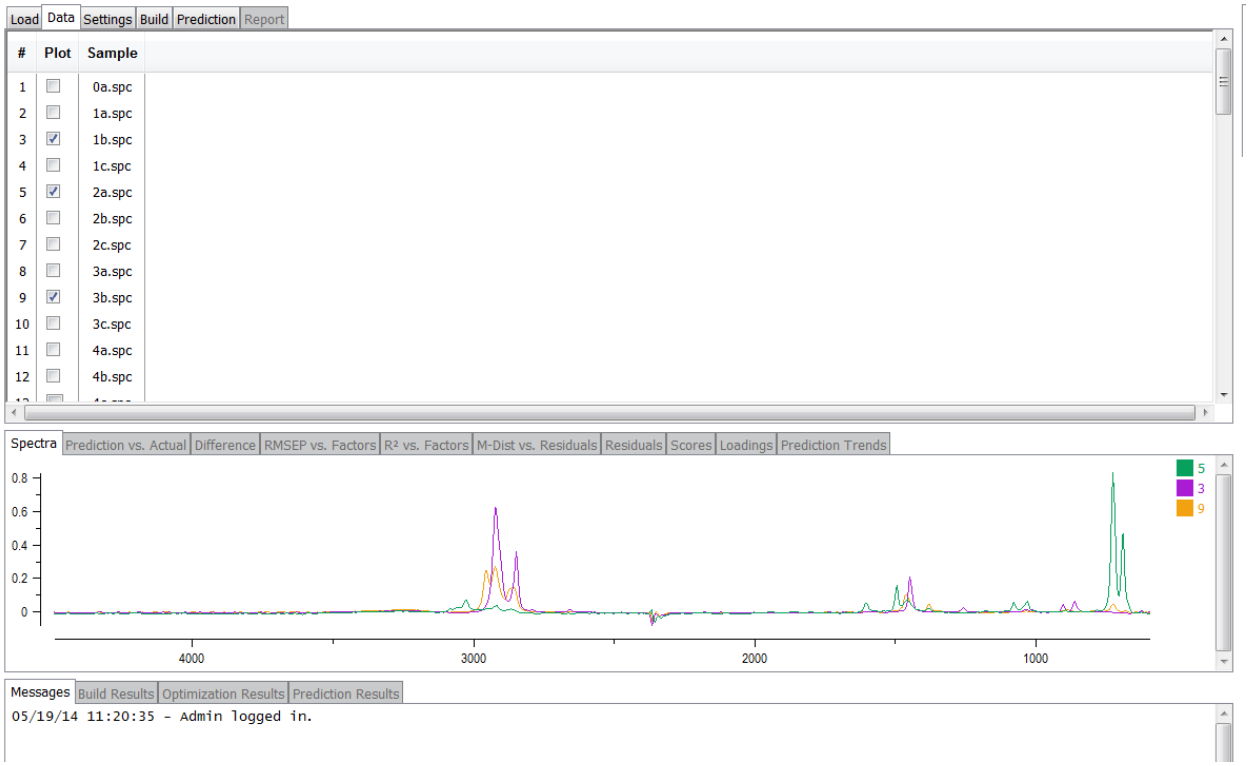
On logging into Symbion QT, you will be taken to the Main Screen. The flat layout provides ready – and easily understood – access to all of QT's major functions.



The first task is to load in the files to be used for calibration. This is facilitated by the navigation panel at the upper left. For our example (see below), we have navigated to the folder "CTH-3", which was created during the data acquisition step, above. The list box shows all of the files that QT recognizes as data files, in this case 82 spectra. We can quickly preview any of the data in the folder by using the left click mouse function, as illustrated below.

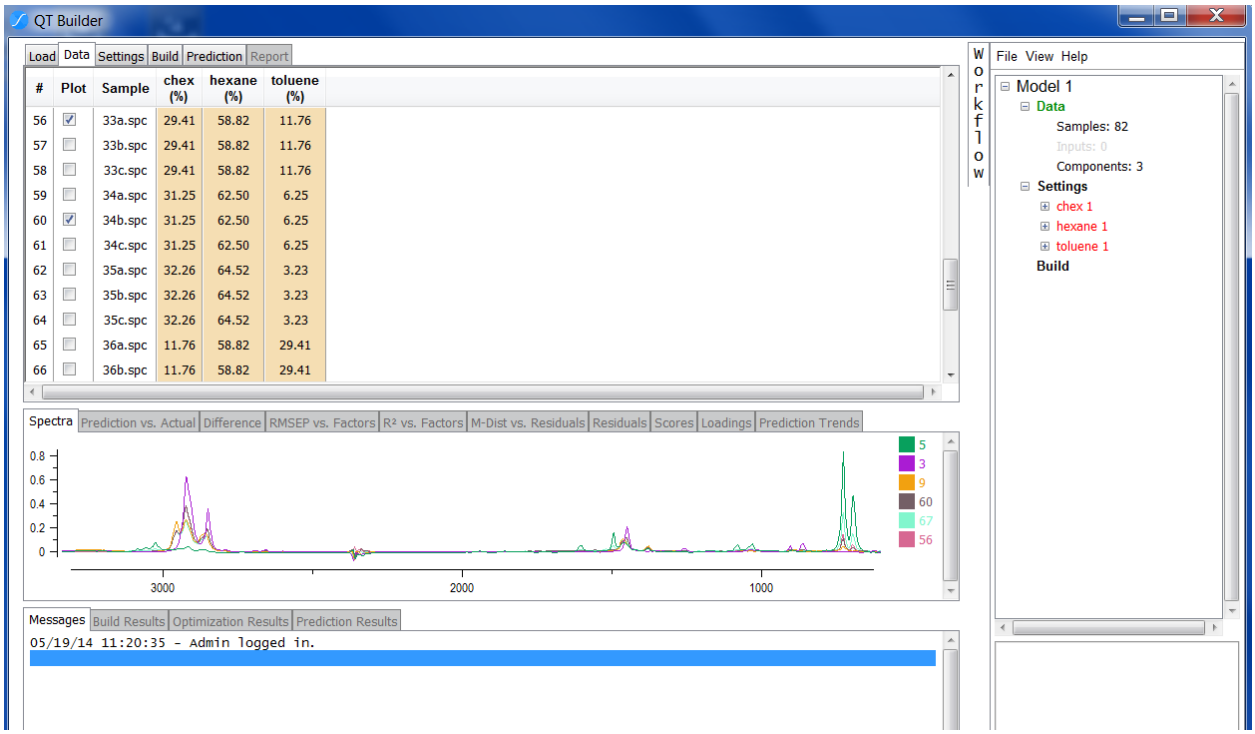


To select spectra for use by QT, simply highlight and select the desired files, or press “Select All”. Then press “Append to Data”. The display will now switch to the “Data” tab and the selected sample files will be listed. Here you can check any of the files that you wish to display.



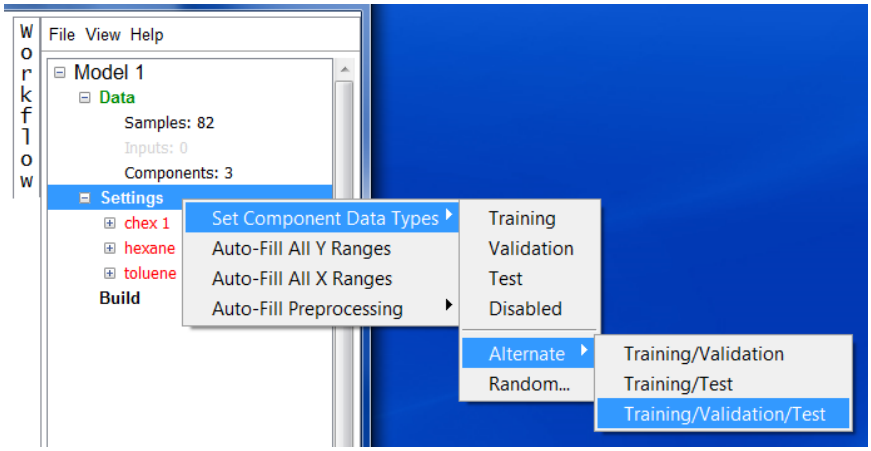
2.2 Creating Variables

The next step is to add the variables (concentration information) for the components present. This is where the power resulting from the synergy between QT and the Symbion data acquisition programs comes to the fore. By right clicking in the region to the right of the Sample column you will be given several options for adding the variable information. Simply select “Build variables from Symbion data files”. The variable names and concentration values are automatically loaded into the data table, as shown below. Here we have selected a few additional spectra for display.



2.3 Model Settings

The first step in specifying settings is to indicate which files are to be used for training, validation, and test. This can be done individually or automatically. For automatic selection, simply right click on the word “Settings” in the Workflow panel and choose “Set Component Data Types”. You will be given the choices shown below.



For our illustration, we will choose “Training/Validation/Test”. Returning to the data tab, we can see which spectra have been designated for each purpose.

Load | Data | Settings | Build | Prediction | Report

#	Plot	Types (hexane 1)	Sample	chex (%)	hexane (%)	toluene (%)
1	☐	Training	0a.spc	0	0	0
2	☐	Validation	1a.spc	100	0	0
3	☐	Test	1b.spc	100	0	0
4	☐	Training	1c.spc	100	0	0
5	☐	Validation	2a.spc	0	0	100
6	☐	Test	2b.spc	0	0	100
7	☐	Training	2c.spc	0	0	100
8	☐	Validation	3a.spc	0	100	0
9	☐	Test	3b.spc	0	100	0

To continue, Right click on Settings in the Workflow region and Pull-down to *Auto fill all Y Ranges* (component output ranges). This will look for the minimum and maximum value in each component column and automatically place the min and max value into the settings for each component. These are shown in the Settings table.

Load | Data | Settings | Build | Prediction | Report

☐ Show Limits

#	Component	Unit	Y range	Min	Max	Max Factors	Region	Min	Max	Pre-Processing	
1	chex	%	1	0	100	20	1 of 1				Preview
2	hexane	%	1	0	100	20	1 of 1				Preview
3	toluene	%	1	0	100	20	1 of 1				Preview

Similarly, right click on Settings in the Workflow region and Pull-down to Auto-Fill all X ranges (regions of interest). This will look for the minimum and maximum value in the x-values of the data files and automatically place the min and max value into the settings for each component.

Finally, right click on Settings in the Workflow region and Pull-down to Auto-Fill Preprocessing-Baseline (2pt). This will automatically remove offset and slope from the baseline. The Settings table now looks like this.

Load | Data | Settings | Build | Prediction | Report

☐ Show Limits

#	Component	Unit	Y range	Min	Max	Max Factors	Region	Min	Max	Pre-Processing	
1	chex	%	1	0	100	20	1 of 1	597	4494	Baseline(2pt)	Preview
2	hexane	%	1	0	100	20	1 of 1	597	4494	Baseline(2pt)	Preview
3	toluene	%	1	0	100	20	1 of 1	597	4494	Baseline(2pt)	Preview

The number of factors is the last value to be set. This must be less than the number of training or validation samples. If this condition is violated, there will be an indication in red on any given model.

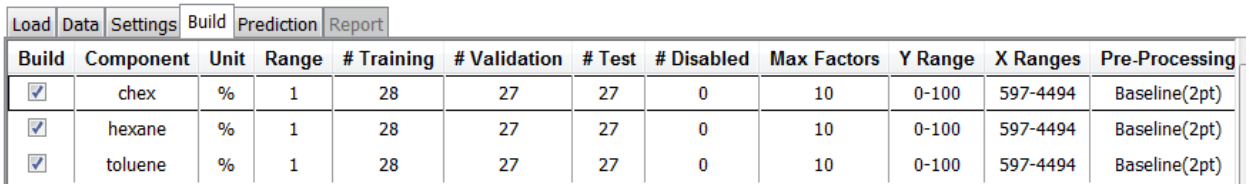
If you expand the data types, you can see that we have 27 samples for training and 28 for validation. While these are greater than 20, it still is advisable to restrict the maximum number of factors so as not to model noise. In our example, we know that we have three components plus at least two possible interferences, H2O vapor and CO2

vapor. So, to be safe, we will set the maximum number of factors to 10. We simply type this value into the table in place of 20.

Up to this point, we have applied the same sample settings to each of the three components. As we will see later, it may be advantageous to use different settings for each of the components.

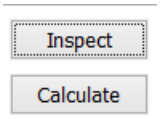
2.4 Building and Validating a Model

To start a build, first select the Build tab and check the models to be built.

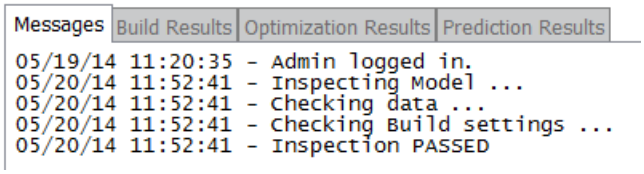


Build	Component	Unit	Range	# Training	# Validation	# Test	# Disabled	Max Factors	Y Range	X Ranges	Pre-Processing
☑	chex	%	1	28	27	27	0	10	0-100	597-4494	Baseline(2pt)
☑	hexane	%	1	28	27	27	0	10	0-100	597-4494	Baseline(2pt)
☑	toluene	%	1	28	27	27	0	10	0-100	597-4494	Baseline(2pt)

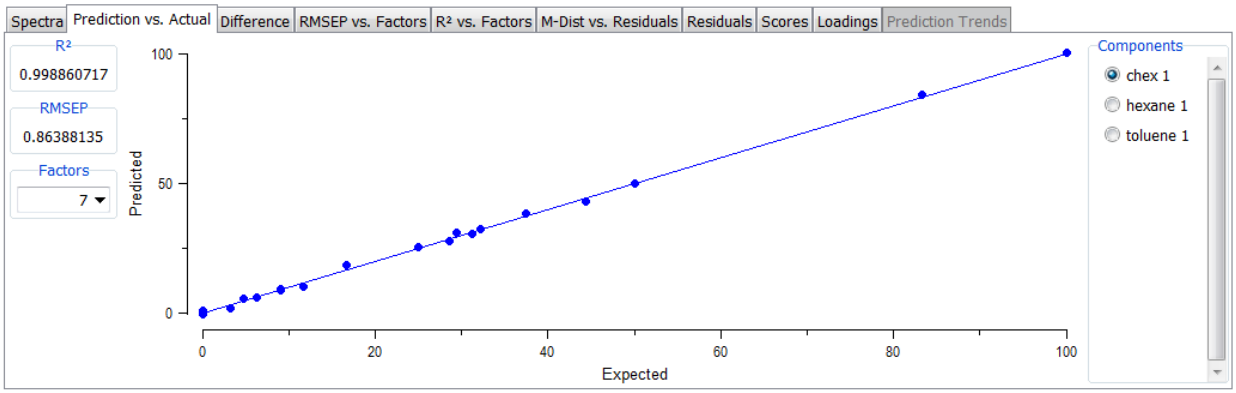
Next press the *Inspect* button to the right of the build summary table.

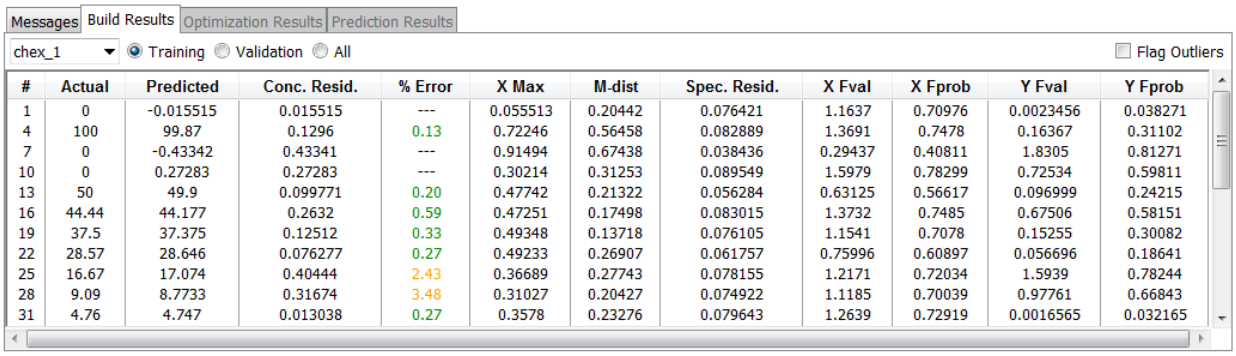


An indication of whether the inspection passed or failed will occur in the lower pane.



If the Inspect step passed, press the *Calculate* button to build and validate all selected models. After this is completed, the lower two panes will show the Prediction vs. Actual plots and the Build statistics.



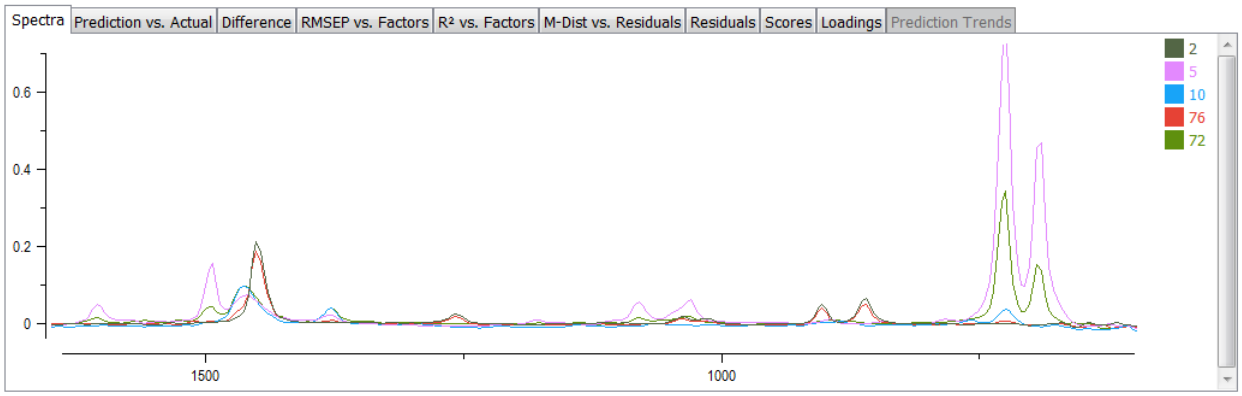


#	Actual	Predicted	Conc. Resid.	% Error	X Max	M-dist	Spec. Resid.	X Fval	X Fprob	Y Fval	Y Fprob
1	0	-0.015515	0.015515	---	0.055513	0.20442	0.076421	1.1637	0.70976	0.0023456	0.038271
4	100	99.87	0.1296	0.13	0.72246	0.56458	0.082889	1.3691	0.7478	0.16367	0.31102
7	0	-0.43342	0.43341	---	0.91494	0.67438	0.038436	0.29437	0.40811	1.8305	0.81271
10	0	0.27283	0.27283	---	0.30214	0.31253	0.089549	1.5979	0.78299	0.72534	0.59811
13	50	49.9	0.099771	0.20	0.47742	0.21322	0.056284	0.63125	0.56617	0.096999	0.24215
16	44.44	44.177	0.2632	0.59	0.47251	0.17498	0.083015	1.3732	0.7485	0.67506	0.58151
19	37.5	37.375	0.12512	0.33	0.49348	0.13718	0.076105	1.1541	0.7078	0.15255	0.30082
22	28.57	28.646	0.076277	0.27	0.49233	0.26907	0.061757	0.75996	0.60897	0.056696	0.18641
25	16.67	17.074	0.40444	2.43	0.36689	0.27743	0.078155	1.2171	0.72034	1.5939	0.78244
28	9.09	8.7733	0.31674	3.48	0.31027	0.20427	0.074922	1.1185	0.70039	0.97761	0.66843
31	4.76	4.747	0.013038	0.27	0.3578	0.23276	0.079643	1.2639	0.72919	0.0016565	0.032165

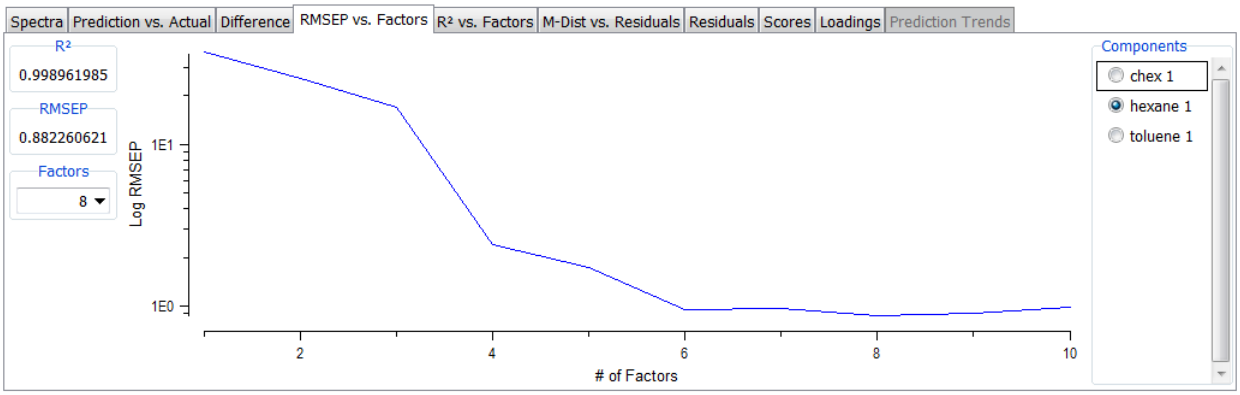
The user can check the goodness of fit by the level of linearity shown in the above plot. The percent error of prediction for each sample can be assessed in the lower pane. The percent error is color-coded so that the user can quickly determine which samples are outliers. The results shown are for cyclohexane. Checking on the other two components, in turn, will provide the corresponding results.

Note that the above model used 7 factors. You can quickly test the effect of using a different number of factors by using the ”Factors” pull-down at the left to change the number. If you change the number of factors, the program will quickly revise its calculations and provide the new result.

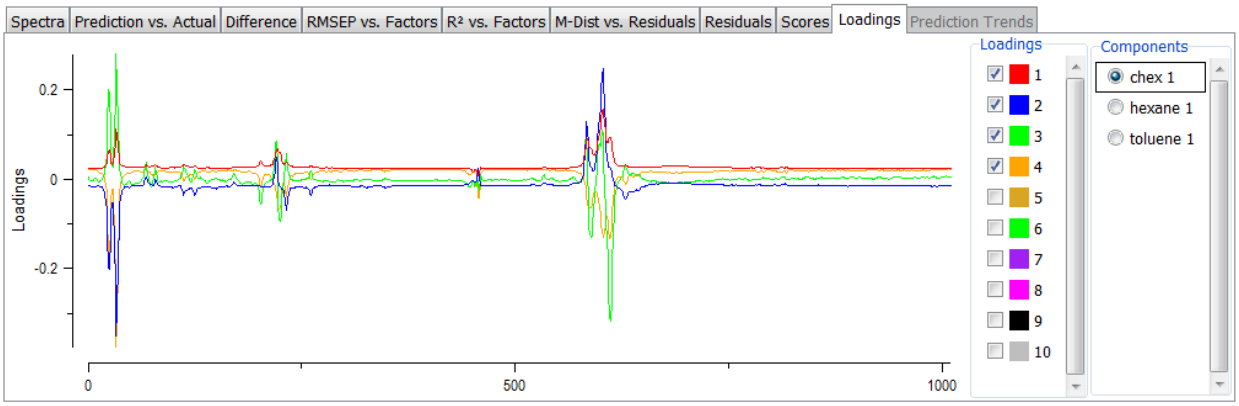
The tabs across the top of the graphic display enable you to inspect your results in a variety of ways. For example, selecting “Spectra” displays any of the spectra that are checked on in the data table.



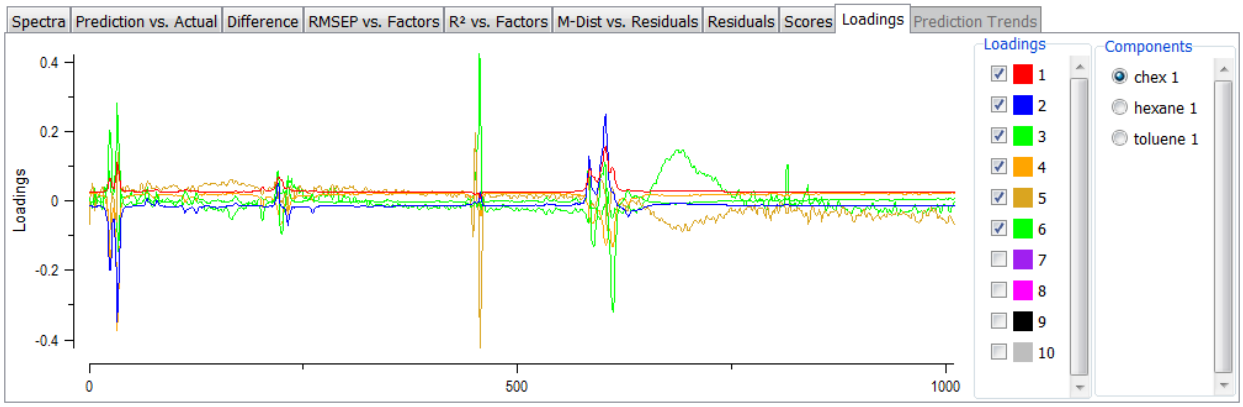
The identification of the chosen spectra is indicated by the color coded numbers at the right. The contributions of the successive factors to the goodness of fit can be evaluated by selecting either “R^2 vs. Factors” or “RMSEP vs. Factors”. Since the latter is plotted on a log scale, it provides a more sensitive measure. In either case, individual plots are provided for each component.



Another way to evaluate the contributions of the factors is to look at the loadings for each of the components.



Here we see that the first four factors (checked on) contribute their maximum variance in the regions where our given components are known to absorb. However, if we check on the next two factors, we find that these contribute primarily in the regions characterized by CO2 and water absorption.



Before attempting to optimize our model design, we will save the current model. To do this simply use the File pull down at the upper right, navigate to the desired directory, and provide a name for the model – in this case “CTH Model 1”.

To obtain a comprehensive report of both the settings and results of the build, we select the “Report” tab at the top of the QT screen, right click in the open area, and select “Generate”. If desired, the report can be copied into a Word document.

Examining the report for our current calibration, we find that the program used 7, 8, and 9, factors respectively for the three components to generate the results. Here is the validation portion of the report.

CTH Model 1 Build Results (Validation):

Component	Range	Max X	Factors	R^2	RMSEP	Outliers	Max Error	Avg. Error
chex	1	0.895347774	7	0.99886	0.86388	0	1.00 (23)	0.38
hexane	1	0.895347774	8	0.99896	0.88226	0	1.42 (53)	0.39
toluene	1	0.895347774	9	0.99953	0.58151	0	0.49 (68)	0.21

Note that the number in ( ) under “Max Error” indicates the sample for which the maximum error occurred.

2.5 Model Optimization

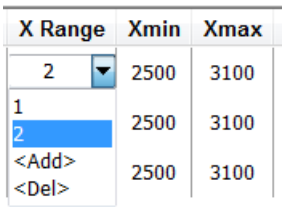
Since we know that the primary effects of CO2 and H20 are at the higher frequencies, we will modify our settings to eliminate this region. On inspection of the spectra, we decide to set X max at 1580 cm-1. The new results are as follows:

CTH Model 2 Build Results (Validation):

Component	Range	Max X	Factors	R^2	RMSEP	Outliers	Max Error	Avg. Error
chex	1	0.852787733	5	0.99878	0.89452	0	1.89 (47)	0.63
hexane	1	0.852787733	9	0.99522	1.89357	0	1.66 (59)	0.49
toluene	1	0.852787733	7	0.99947	0.61716	0	0.80 (68)	0.30

We see that excluding the high frequency region actually degraded the results, although the program did use fewer factors for cyclohexane and toluene. The substantially worse result for hexane is understandable since hexane is a weak absorber in the low frequency region.

To further refine our model, we can add a second spectral region to the analysis. To do this we go to the Settings tab and left click on the X range for the first component. We see the pull-down:



This enable us to add, delete, or change ranges. Here we have added a second range from 2500 to 3100 cm-1 for all three components. Once again building the calibration, we obtain the following result:

CTH Model 3 Build Results (Validation):

Component	Range	Max X	Factors	R^2	RMSEP	Outliers	Max Error	Avg. Error
chex	1	0.852787733	5	0.99930	0.67594	0	1.50 (47)	0.43
hexane	1	0.852787733	8	0.99864	1.01095	0	1.29 (56)	0.45
toluene	1	0.852787733	8	0.99961	0.52499	0	0.61 (65)	0.27

This result is an improvement over Model 2 but is still not as good as model 1 for which we used the whole spectrum.

After inspecting all of our results to date, we decided to use two ranges for cyclohexane, the full spectrum for hexane, and only a narrow low frequency region for toluene. The result is as follows:

CTH Model 4 Build Results (Validation):

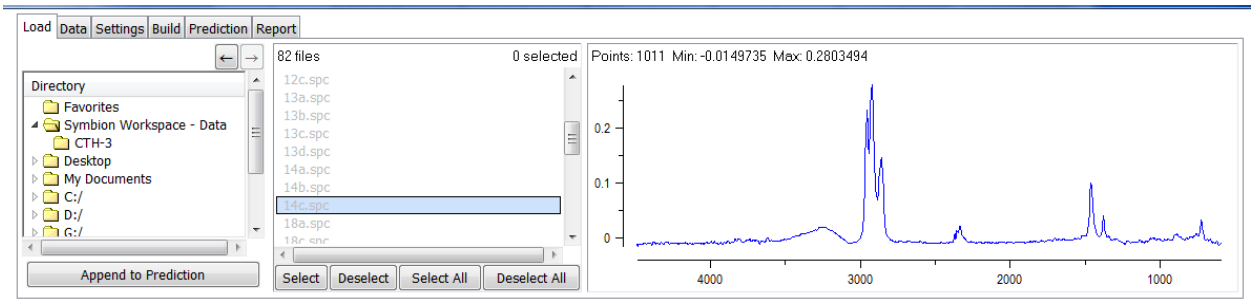
Component	Range	Max X	Factors	R^2	RMSEP	Outliers	Max Error	Avg. Error
chex	1	0.851022542	6	0.99937	0.64489	0	0.94 (80)	0.40
hexane	1	0.899917841	8	0.99896	0.88338	1	1.58 (53)	0.36
toluene	1	0.831566036	4	0.99938	0.66554	0	1.36 (38)	0.43

Note that the RMSEP for toluene is somewhat worse than for the previous model. However, the current model was obtained with only 4 factors. It thus can expect it to be more robust than Model 3.

The above example illustrates the ease of model development with Symbion QT as well as the ability to rapidly optimize and evaluate multiple models.

Part 3: Predicting a Data Set

Once a model has been developed, it can be used to predict the values of an appropriate data set. To do this, we first select “Predict” from the Mode selection pull down at the lower left of the QT screen. Then using the Load tab at the upper left we navigate to the data set of interest. You can preview the spectra by using your mouse to highlight individual file names.



Next, select the desired data files or simply press “Select All”. Then press “Append to Prediction”. The program will then switch to the prediction tab and a list of spectra will be displayed. Next, we press “Predict” at the upper right. A table of predicted values and a set of prediction trend plots will appear.