Written Tutorial for Running PAT: Existing Dataset

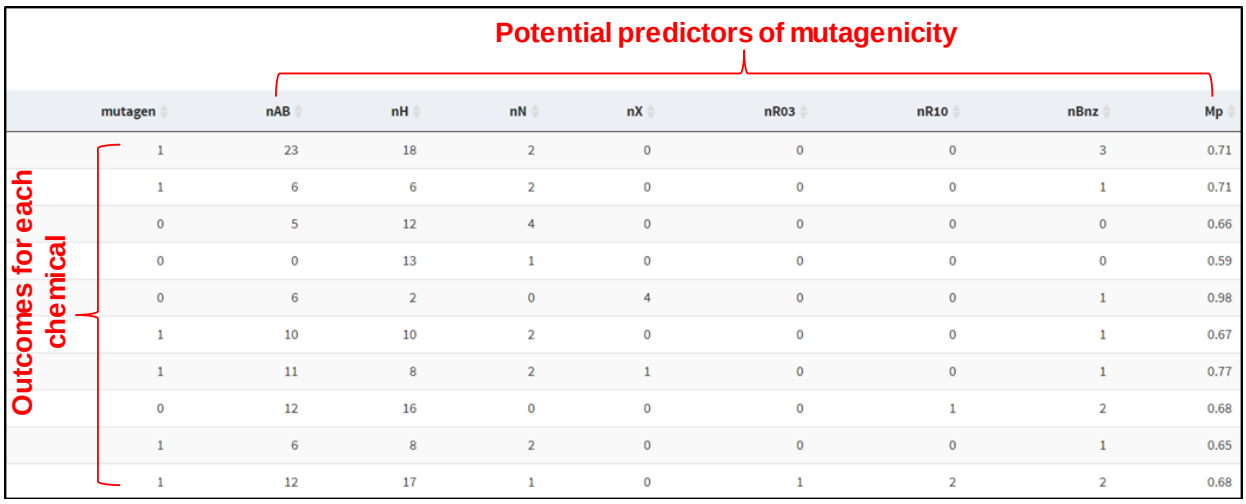
Video Link

Table of Contents

1. Data Selection and Formatting

In this tutorial, we will demonstrate how to run the Predictive Analytics Toolkit (PAT) using a dataset provided on the PAT website.

Figure 1 provides a snapshot of the mutagens dataset, which is a modified version of a dataset from the "QSARdata" R package described at this web address. Briefly, these data are the Ames mutagenicity test results (binary classification of mutagenicity; see Ames et al. 1973) of various chemicals as collected by Kazius et al. (2005). For purposes of this tutorial, the modified dataset used here employs roughly 700 of the original 4000+ chemicals, as well as only 8 of the original 1500+ chemical attributes used as potential predictors of mutagenicity (see Figure 1 for a snapshot of this dataset).



mutagen	nAB	nH	nN	nX	nR03	nR10	nBnz	Mp
1	23	18	2	0	0	0	3	0.71
1	6	6	2	0	0	0	1	0.71
0	5	12	4	0	0	0	0	0.66
0	0	13	1	0	0	0	0	0.59
0	6	2	0	4	0	0	1	0.98
1	10	10	2	0	0	0	1	0.67
1	11	8	2	1	0	0	1	0.77
0	12	16	0	0	0	1	2	0.68
1	6	8	2	0	0	0	1	0.65
1	12	17	1	0	1	2	2	0.68

Figure 1. Example subset of the 'mutagens' dataset provided on the PAT website.

The first column of the dataset represents the outcome to be predicted in binary format, where 0 = is not a mutagen and 1 = is a mutagen. The subsequent columns are chemical attributes that are potential predictors of the outcome. The rows are the data for each chemical. The predictive algorithms within PAT will draw from the data in the potential predictor columns to predict the mutagenicity of each chemical. This structure of the mutagens dataset (i.e., a column of binary outcomes with other columns being potential predictors of the outcome) is the required general structure of any dataset input to PAT.

The following step-by-step guide describes how to conduct an analysis of the mutagens example dataset in PAT:

1. On the main page of the website, select the drop-down arrow next to Data (see Figure 2; in the Select data source box, under Samples). Select the mutagens option (1).

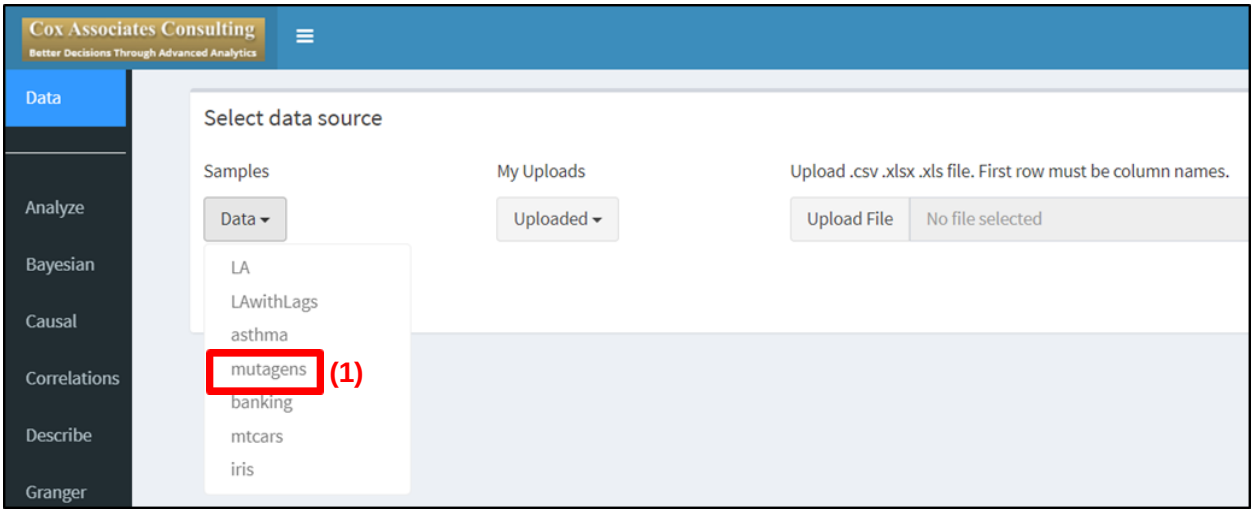


Figure 2. Location of the 'mutagens' dataset that is used within this tutorial.

2. As shown in Figure 3, additional options, as well as a preview of the selected dataset, will display in the lower part of the tool.
    - Under the Optional: Select columns heading (1; see Figure 3), users can select individual columns to be included in the analysis, rather than using all columns in the dataset (by clicking into the field and choosing from a drop-down list). For this tutorial, all columns of the dataset will be used, and so this box is left empty.
    - Additionally, users can choose to discretize individual columns under the Optional: Select integer/character variables to make discrete heading (2; see Figure 3). These functionalities are not employed in this tutorial, but are described further in the User-provided Dataset Tutorial.
3. Under the Show 10 entries selection box (3), a preview of the selected data is displayed (see Figure 3). The number of rows displayed at one time can be modified by selecting the desired number of rows from the drop-down menu between Show and entries (3). The section of rows can be navigated using the numbers next to Previous and Next below the bottom right corner of the table (4).

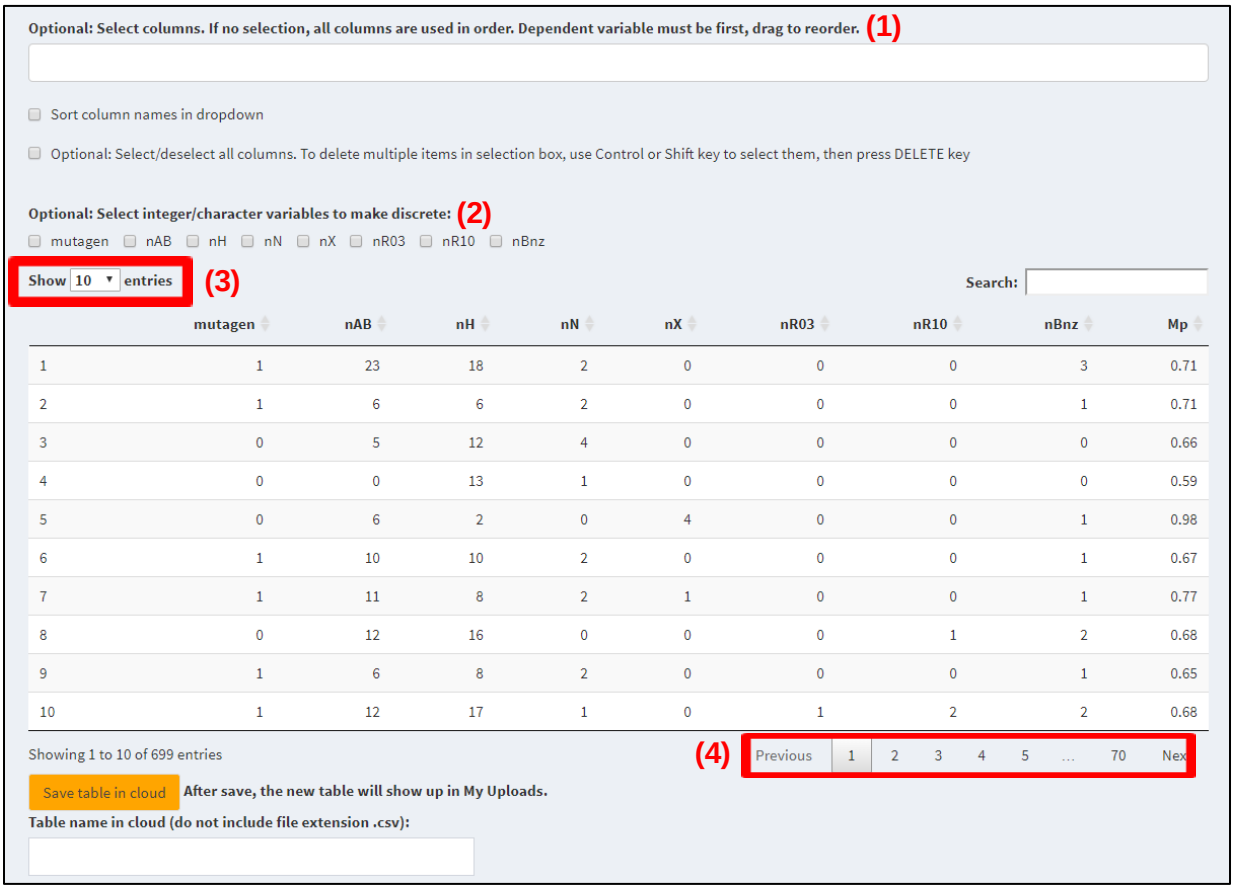


Figure 3. Additional options and preview of the input 'mutagens' dataset and associated navigation buttons.

4. Figure 4 illustrates how a dataset along with any modifications can be saved to the website by selecting the Save table in cloud button (1) below the data preview section (see Figure 4

or the User-provided Dataset Tutorial for more information). The bottom table of the webpage displays the data type of each of the columns of the dataset (2). Possible data types are numeric (i.e., integer or decimal numbers) and factor (text).

5. To move to the next step, in the far-left ribbon of the website, select the Predict tab (3; see Figure 4). For this tutorial, you will not have made any changes to the mutagens dataset at this point.
    a. Please note that the other tabs present in this ribbon provide functionalities for which user documentation has not yet been developed.

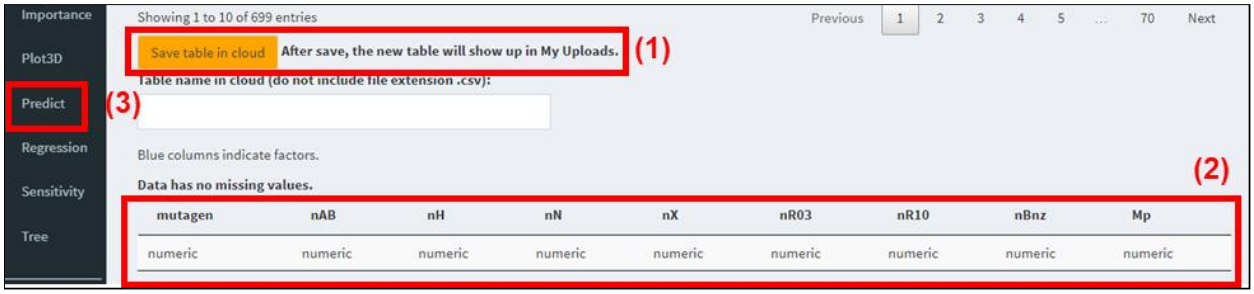


Figure 4. Location of the 'Save table in cloud' button, the input data types table, and the location of the 'Predict' tab.

2. Prediction Modeling Options

6. As shown in Figure 5, the Predict Options section will appear, which allows users to configure several options for the prediction analysis (see Figure 5).

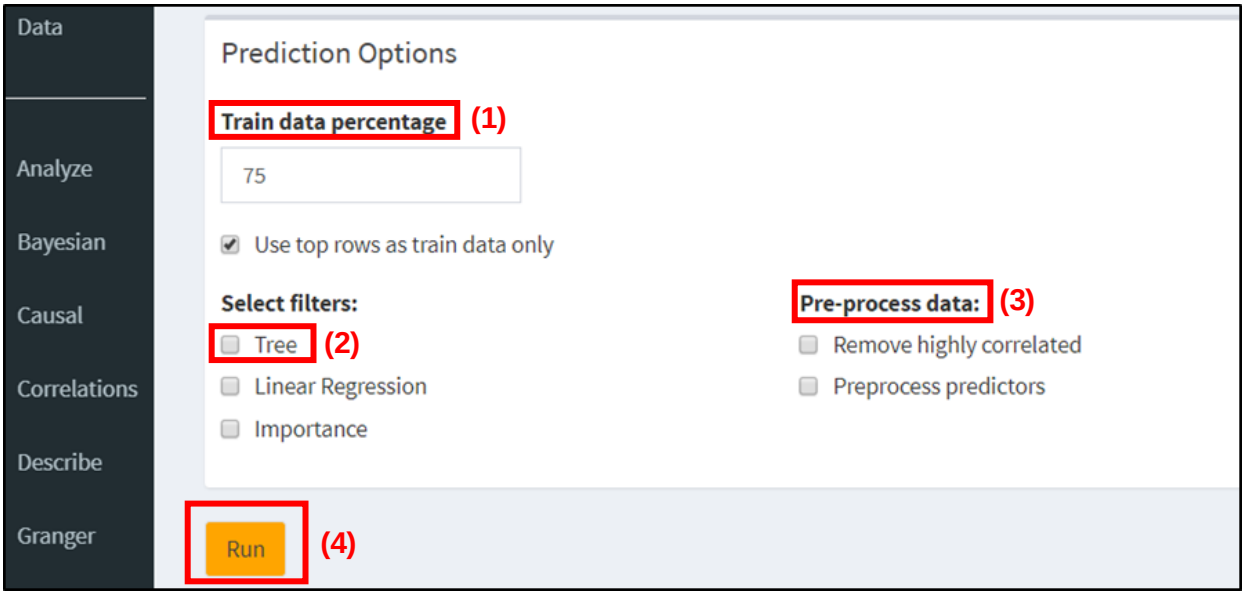


Figure 5. Locations of the 'Train data percentage' box, prediction modeling filters, and pre-processing options.

7. In the Train data percentage box (1), users specify the percentage of the input data that the predictive analytics models will be trained on, with the remaining percentage of the data

being used to test the predictive models that are developed. Generally, a majority of the data (conventionally, two-thirds of the data) are used for training (see this website for further discussion on splitting a dataset into training and testing sets). For the purposes of this tutorial, leave the default value of 75 as the training percentage. The Use top rows as train data only box can be selected if users wish to only train the models using rows from the top percentage of the dataset. For details on the usefulness of this functionality, see the User-provided Dataset Tutorial. Uncheck this box for this tutorial so that the training data subset will be selected randomly from throughout the whole dataset.

8. The options under the Select filters heading can be used to have PAT automatically select columns (potential predictors) that are the strongest predictors when developing these three models (CART trees, linear regression models, or randomForest importance analyses) and disregard the other columns. By default, PAT will use all columns in the prediction modeling, which can impact runtime. Use of these filters can decrease the number of columns that need to be processed, and thus their use is encouraged for datasets with thousands of columns. For the purposes of this tutorial, select only the Tree filter option (2).
9. Under the Pre-process data heading (3), users can select whether to apply two types of pre-processing to the dataset. The first option, Remove highly correlated, removes potential predictors that are highly correlated with each other or are duplicates of other columns (as they do not provide any additional predictive power to the models). It also removes columns that do not contain any information (e.g., columns that are all zeros) as well as columns whose rows all contain the same value. The second option, Preprocess predictors, takes columns that are on very different scales and standardizes them (by calculating distance from the central value in units of standard deviation), while also removing low-variance predictors. These are two standard options in most machine learning programs, and so are included in PAT to help with reproducibility of previous results. For more information on these functions, users are directed to this web page. For the purposes of this tutorial, select both of the pre-processing options.
10. Once all of the Prediction Options have been set, click the orange Run button (4). A blue box will appear with the text Calculating, please wait... The run time for PAT may be several minutes, depending on the size of the dataset and which pre-processing filters were selected. For the mutagens dataset initialized with the settings in this tutorial, the run time will be roughly 5 minutes. Once PAT has finished processing, the blue box will disappear.

3. Prediction Modeling Outputs

By default, six prediction models are developed from the data input to PAT. Outputs for user analysis include tabular and graphical representations of the results of these predictive models, as well as diagnostic information for each of the models. The definition of each of the default algorithms used for model building is provided below:

- earth: The R programming language name for Multivariate Adaptive Regression Splines (which is trademarked and licensed to Salford Systems).
- rpart: Recursive partitioning.
- ctree: Classification tree.

- rf: Random Forest.
- gbm: Gradient boosting machines.
- glmboost: Boosted generalized linear model.

A description of the default PAT outputs is provided in Section 3.1 through Section 3.6. Note that the PAT outputs presented in the screenshots in these sections may differ slightly from outputs of PAT when run on users' computers due to inherent randomness in the underlying algorithms.

3.1. PAT Run Synthesis

The outputs of PAT are displayed in several sections under the Prediction output heading. As shown in Figure 6, the first section displays a synthesis of the user selections that were applied to the modeling (e.g., the name of the dependent variable, the train data percentage, filters ("Tree") that were applied).

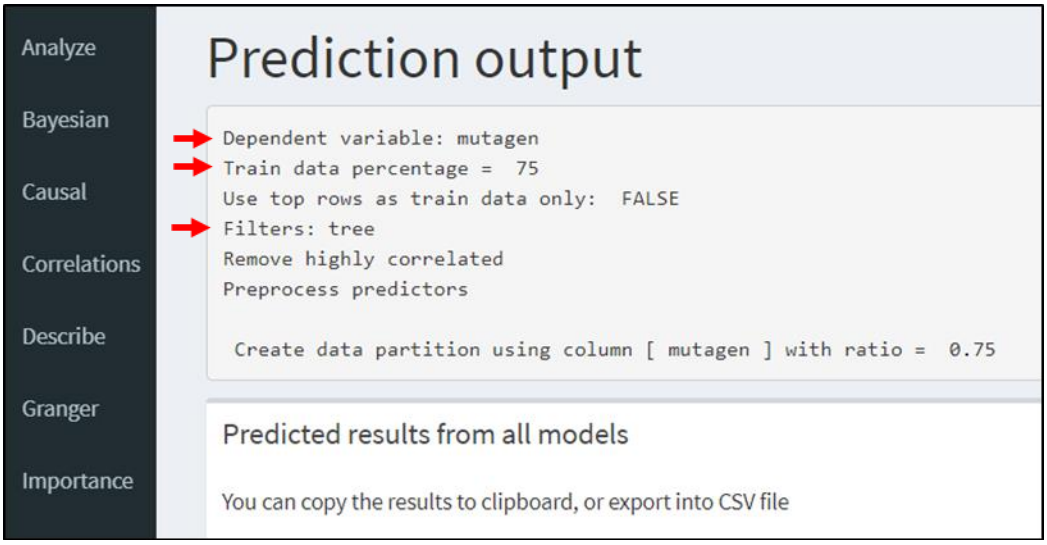


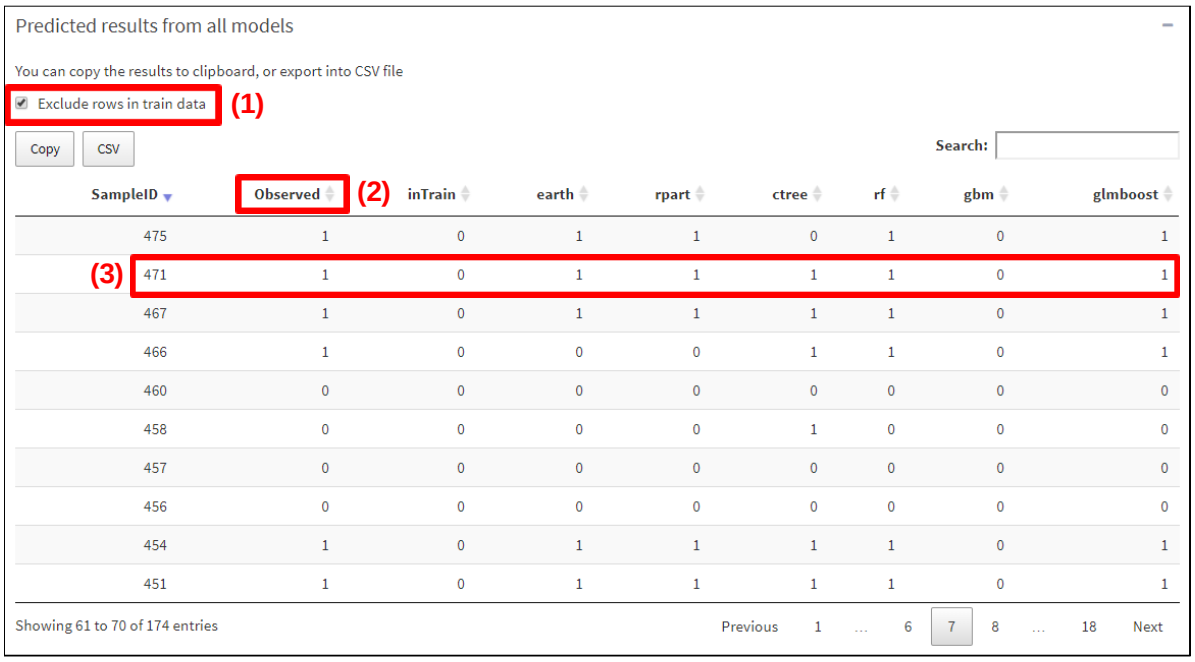
Figure 6. Screenshot of the summary of modeling parameters used within the PAT run.

3.2. Results from Individual Models

Binary classifications are output from each of the predictive models in PAT (i.e., mutagenic or non-mutagenic classifications for each chemical of the example dataset). These binary results are based on model-produced probabilities that a specific outcome will be true (mutagenic) in tandem with a threshold probability that has to be met for a predicted outcome to be classified as true. In PAT, a probability of 0.5 or higher is converted into a positive outcome (i.e., 1, or mutagenic), while probabilities below this are converted to negative outcomes (i.e., 0, or non-mutagenic).

Under the Predicted results from all models heading (shown in **Error! Reference source not found.**) is a table that contains a comprehensive accounting of the mutagenicity prediction for each chemical from all of the prediction models).

- Select the Exclude rows in train data (1) option so that the chemicals used in training the prediction models are not displayed in this table (the InTrain column, which denotes whether a chemical was used in training the models, will then display all zeros).
- The Observed (2) column denotes whether the chemical represented in the given row was observed to be mutagenic (represented by a one) or non-mutagenic (represented by a zero). These delineations are from the input dataset and are not output from the models.
- The earth, rpart, ctree, rf, gbm, and glmboost columns denote whether or not the given model predicted the chemical to be a mutagen or not. For example, the row (chemical) highlighted in red (3) in Error! Reference source not found. was observed to be mutagenic, and all of the prediction models except for gbm correctly predicted that the chemical was mutagenic.



SampleID	Observed	inTrain	earth	rpart	ctree	rf	gbm	glmboost
475	1	0	1	1	0	1	0	1
471	1	0	1	1	1	1	0	1
467	1	0	1	1	1	1	0	1
466	1	0	0	0	1	1	0	1
460	0	0	0	0	0	0	0	0
458	0	0	0	0	1	0	0	0
457	0	0	0	0	0	0	0	0
456	0	0	0	0	0	0	0	0
454	1	0	1	1	1	1	0	1
451	1	0	1	1	1	1	0	1

Figure 7. Example snapshot of results from the individual prediction models generated by PAT.

3.3. Confusion Matrices

Under the Confusion Matrix heading (Figure 8), confusion matrices are presented for each of the prediction models. Confusion matrices show at a glance, both quantitatively and qualitatively, the numbers of correctly (green lobes) and incorrectly (yellow lobes) predicted results from the different prediction models.

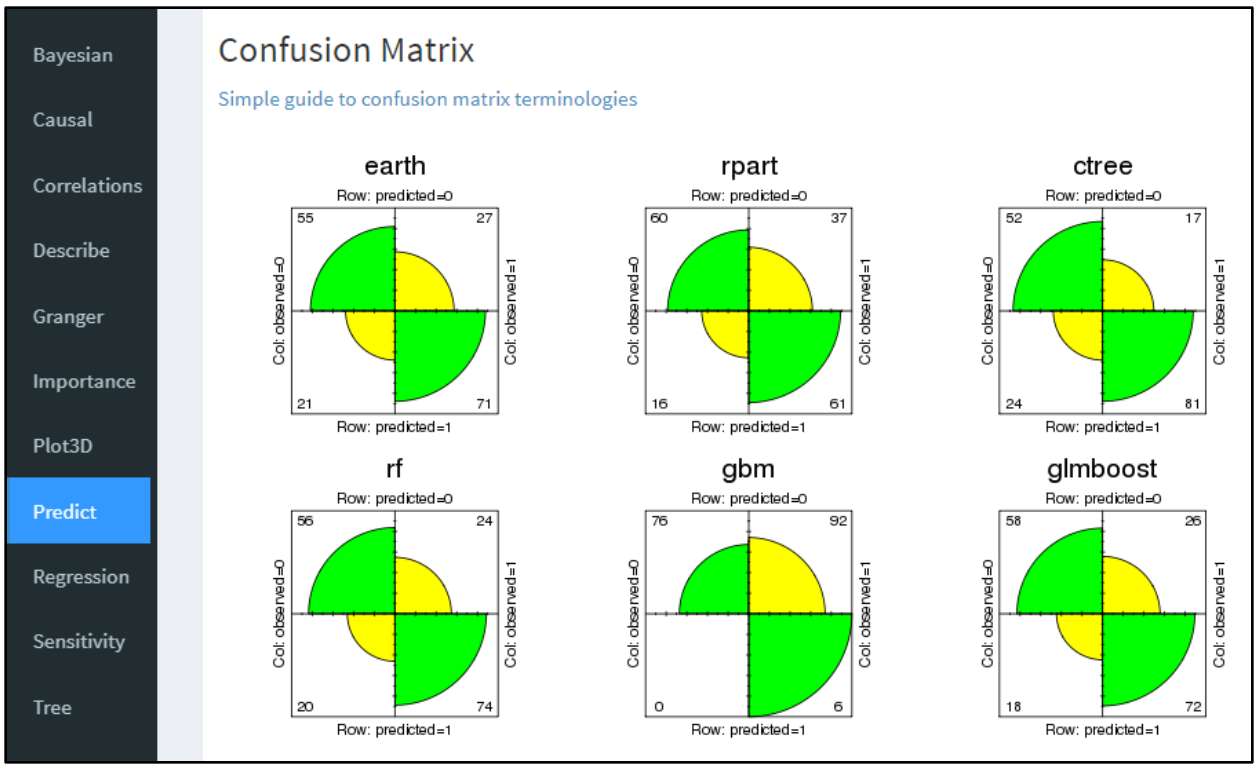


Figure 8. Screenshot of the confusion matrices generated by PAT for each prediction model.

Figure 9 provides a schematic of how to interpret these confusion matrices. The area of each lobe is proportional to the square root of the cell frequencies in each confusion matrix.

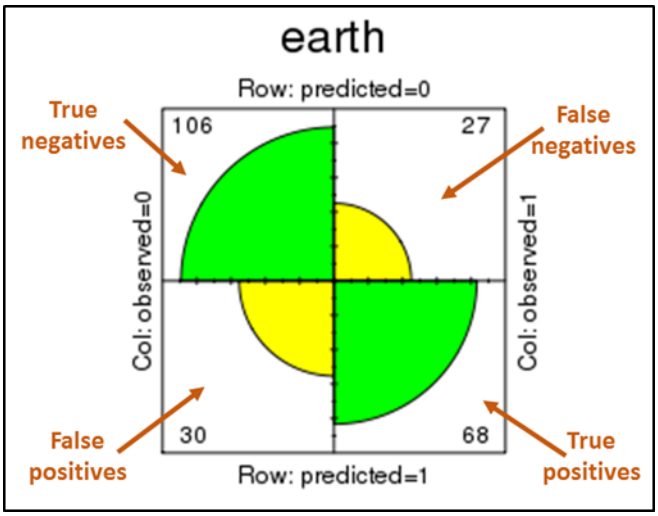
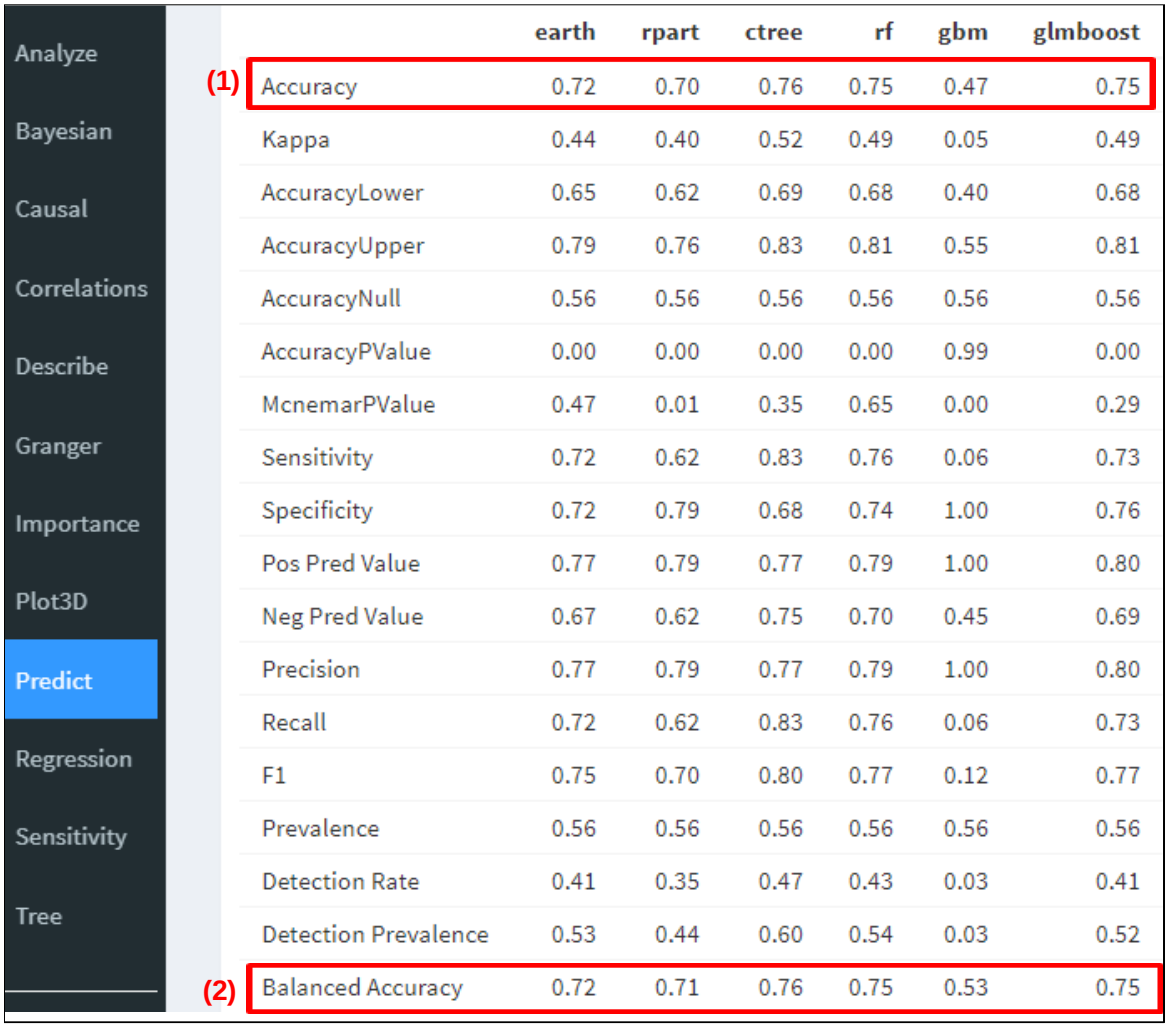


Figure 9. Diagram explaining how to interpret the lobes of a confusion matrix.

A particular model that performs well in predicting a given outcome (in this case, mutagenicity of a chemical) will display a confusion matrix that has relatively large green lobes (many true positives and true negatives) while having relatively small yellow lobes (few false positives and false negatives). Figure 8 shows that the rf model performs the best of all the models, while the gbm model is the least predictive.

A table numerically summarizing the performance of each of the predictive models is presented below the confusion matrices. As shown in Figure 10, the two key metrics from this table are the Accuracy (1) and Balanced Accuracy (2).



		earth	rpart	ctree	rf	gbm	glmboost
(1)	Accuracy	0.72	0.70	0.76	0.75	0.47	0.75
	Kappa	0.44	0.40	0.52	0.49	0.05	0.49
	AccuracyLower	0.65	0.62	0.69	0.68	0.40	0.68
	AccuracyUpper	0.79	0.76	0.83	0.81	0.55	0.81
	AccuracyNull	0.56	0.56	0.56	0.56	0.56	0.56
	AccuracyPValue	0.00	0.00	0.00	0.00	0.99	0.00
	McnemarPValue	0.47	0.01	0.35	0.65	0.00	0.29
	Sensitivity	0.72	0.62	0.83	0.76	0.06	0.73
	Specificity	0.72	0.79	0.68	0.74	1.00	0.76
	Pos Pred Value	0.77	0.79	0.77	0.79	1.00	0.80
	Neg Pred Value	0.67	0.62	0.75	0.70	0.45	0.69
	Precision	0.77	0.79	0.77	0.79	1.00	0.80
	Recall	0.72	0.62	0.83	0.76	0.06	0.73
	F1	0.75	0.70	0.80	0.77	0.12	0.77
	Prevalence	0.56	0.56	0.56	0.56	0.56	0.56
	Detection Rate	0.41	0.35	0.47	0.43	0.03	0.41
	Detection Prevalence	0.53	0.44	0.60	0.54	0.03	0.52
(2)	Balanced Accuracy	0.72	0.71	0.76	0.75	0.53	0.75

Figure 10. Screenshot of the table summarizing the numerical performance of each predictive model.

The accuracy of each of the models developed using the training dataset is provided in the Accuracy row (where the accuracy is calculated as the number of correctly predicted outcomes from the testing dataset compared to the total number of outcomes in the testing dataset).

The data from the input dataset that is selected for training may contain a heavier weighting of positive or negative outcomes, depending on the makeup of the full dataset. An input dataset being skewed like this has the possibility of making a predictive model *appear* better at predicting an outcome than it actually is, given that the actual outcomes it is predicting are skewed towards one outcome. To account for the possibility of the sample training data being skewed towards one outcome, the training dataset is rebalanced via 'up-sampling' and/or 'down-sampling' of the minority and majority outcomes, respectively, such that an equal number of outcomes of each classification are included in the training dataset. The predictive models are then re-developed

using these 'balanced' training data, and the accuracy of the predictive models using this balanced version of the training dataset is presented in the Balanced Accuracy row. Therefore, depending on the makeup of the original dataset, the unbalanced training dataset may differ greatly from the balanced dataset. See the description of the downSample function at this webpage for further details.

As a rule of thumb, look for a Balanced Accuracy of 0.7 or more for determining whether a prediction model is adept at generating accurate predictions for a given dataset. Figure 10 shows that the earth, rpart, ctree, rf, and glmboost models fall within this range, while the gbm model does not. As a quick look at a suite of prediction models, determining which model has the highest Balanced Accuracy is a good way of determining which model provides the best prediction results for a given dataset (in this case, this is the ctree model). Descriptions of the other metrics provided in this table can be found in the R documentation for the CARET package (at this web address) under the section for the confusionMatrix function (see also this web address).

3.4. Areas Under Curve (AUC)

Figure 11 illustrates the Areas Under Curve (AUC) section, which displays in tabular and bar-graph format the 'Area Under the Curve' for each of the predictive models.

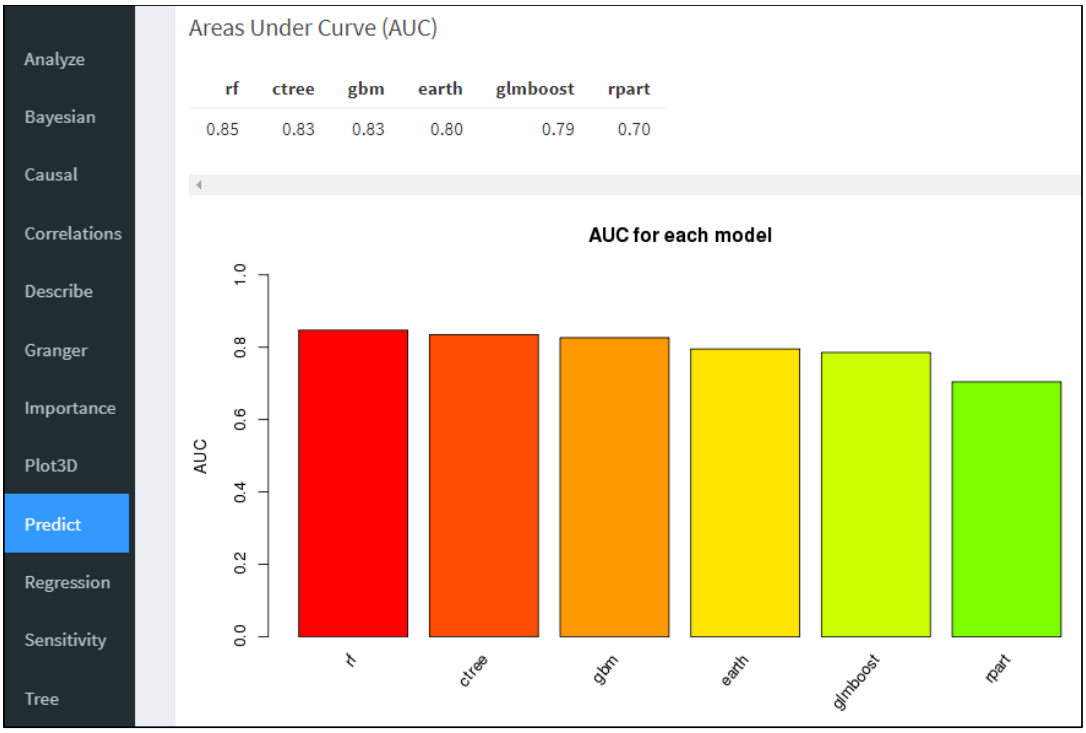


Figure 11. Screenshot showing the Areas Under Curve section within the PAT output.

As noted in Section 3.2, PAT converts outcome probabilities for each chemical to a binary (true or false) classification using a probability threshold of 0.5. The rate of true-positive outcomes and false-positive outcomes will vary depending on the value of this threshold.

- The AUC value describes the relationship between the true-positive and false-positive rates given all possible values of the binary classification threshold.
- AUC values closer to 1.0 indicate prediction models that maximize true positives while minimizing false positives.
- AUC values of 0.5 denote prediction models that are no better at classifying binary outcomes than random guessing.
- From Figure 11, it can be seen that all 6 of the predictive models provide better predictive power than random guessing (all AUC values are over 0.5). For more information on how an AUC metric is generated, readers are referred to this web page.

3.5. Observed vs. Predicted Heatmap

Another method for visualizing the relationships between observed versus model-predicted outcomes is presented in the Observed vs. Predicted Heatmap (see **Error! Reference source not found.**

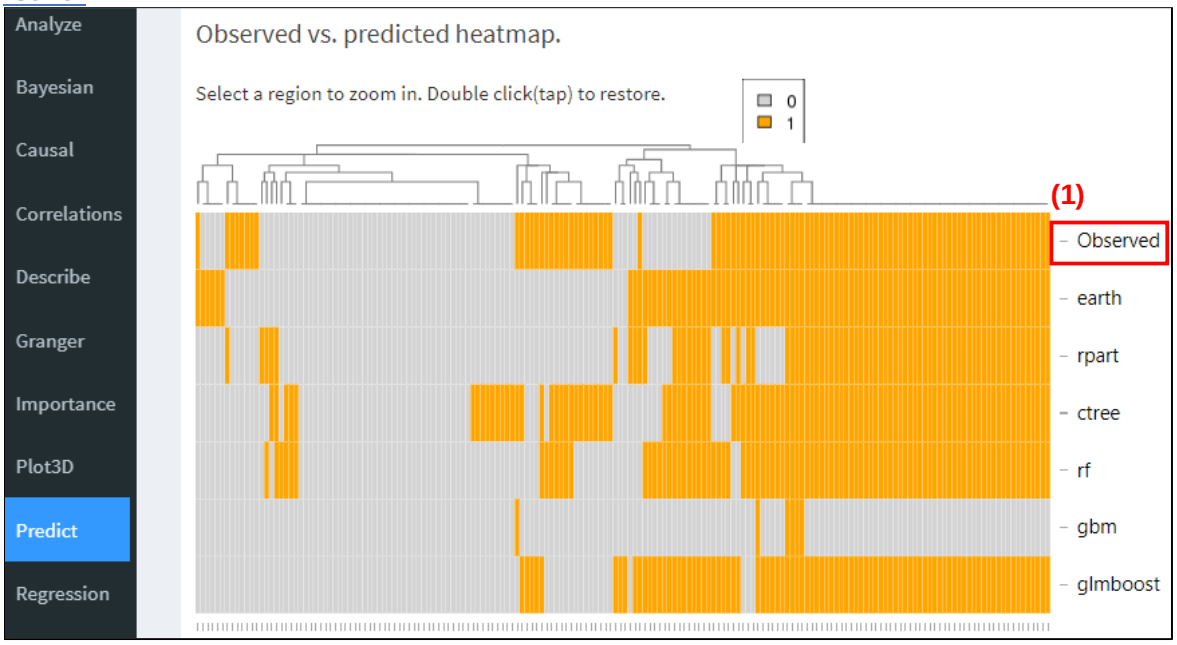


Figure 12).

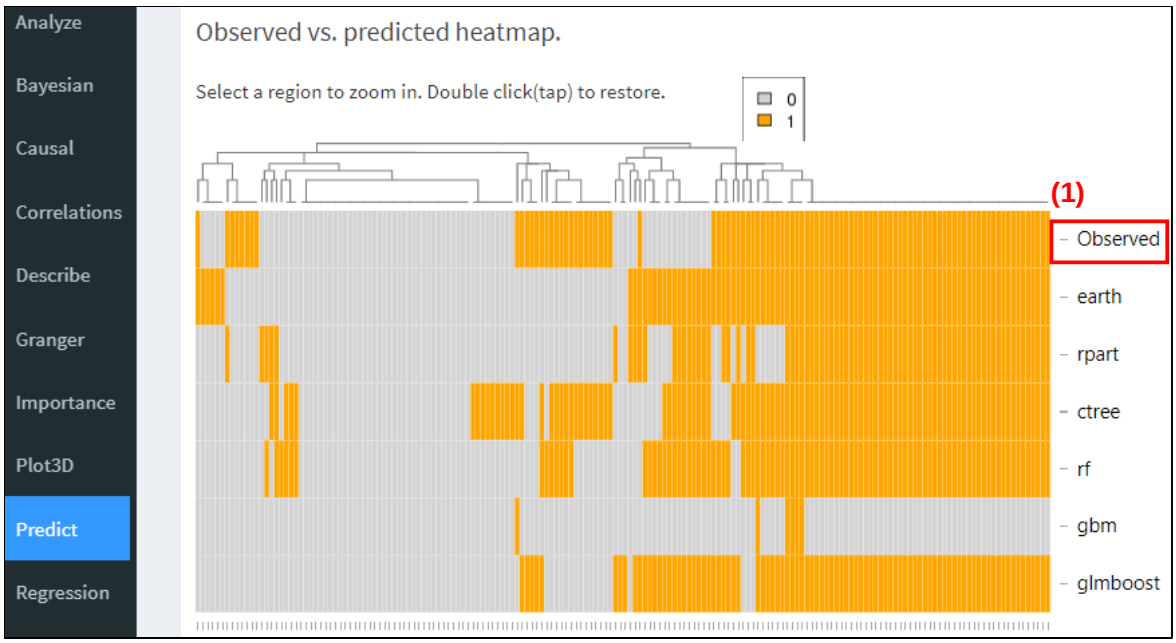


Figure 12. Screenshot showing the Observed vs. Predicted Heatmap generated by PAT.

The columns of this diagram are separate chemicals from the input dataset. The first row (1) denotes the Observed outcome and the other rows represent results from each of the predictive models. Gray cells denote a negative outcome (non-mutagenic) and orange cells denote a positive outcome (mutagenic).

The columns are structured using dendrogram organization such that, *generally*, chemicals that were both observed and predicted by the models to be non-mutagenic are on the left-hand side of the diagram, while those that were observed and predicted to be mutagenic are on the right-hand side. This organization scheme enables rapid visual identification of which chemicals were most difficult for the models to predict accurately, as well as insight into the accuracy of the models. Columns that are entirely orange denote chemicals that were observed and predicted by all models to be mutagenic, and entirely gray columns denote chemicals observed and predicted by all models to be non-mutagenic.

Users can hover their cursor over cells of the diagram and an information box will appear enumerating the name of the row, the column number, the predicted/observed value, and the sample ID. Clicking and drawing a box over a region of interest of the diagram will cause the

diagram to zoom in to that region, enabling easier inspection of individual cells (see

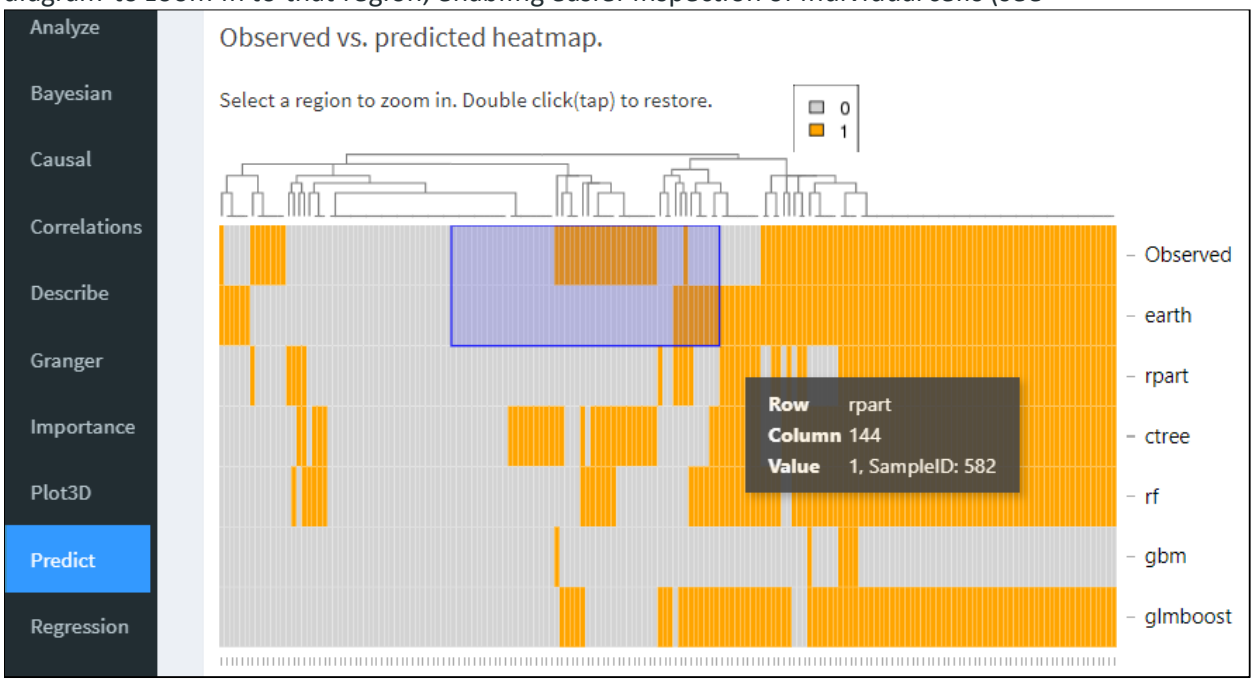


Figure 13). To restore the diagram to the original zoom level, double click anywhere in the diagram.

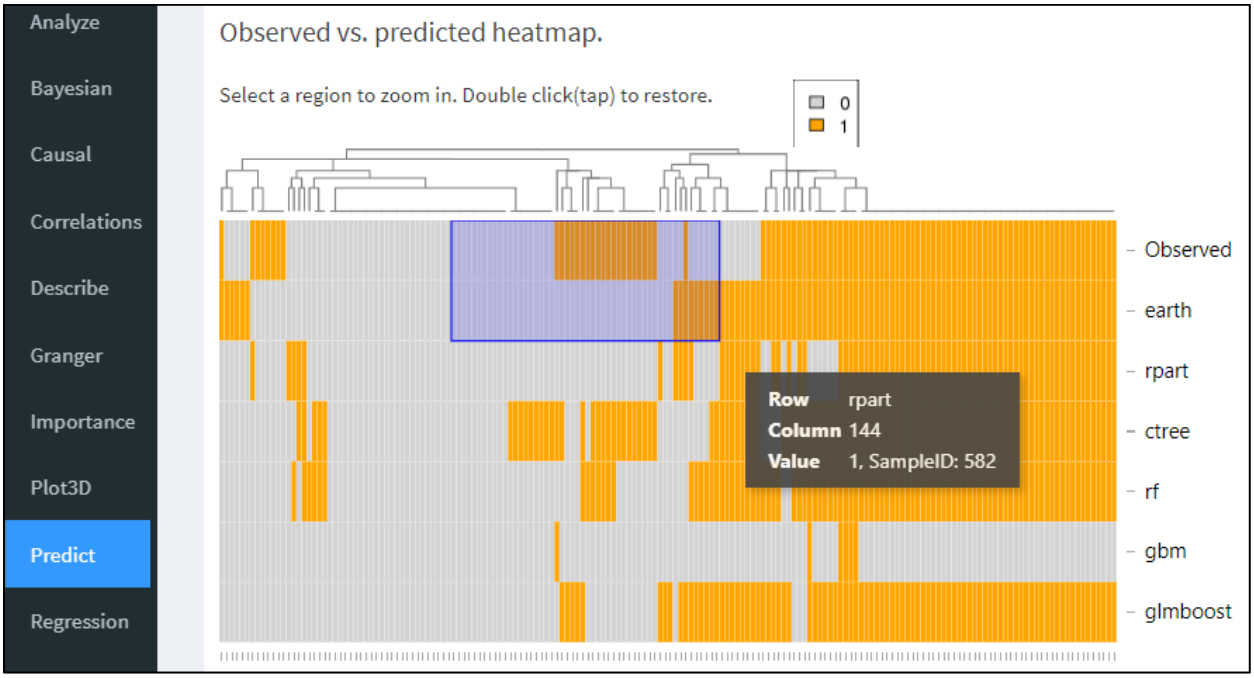


Figure 13. Demonstration of how to zoom into a region of interest in the Observed vs. predicted heatmap.

3.6. Other Output for Each Model

Several other figures generated by PAT from the predictive models are displayed under the Observed vs. Predicted Heatmap (see Figure 14). These outputs describe some of the more technical aspects of each of the developed models, and their interpretation is discussed in the User-provided Dataset Tutorial.

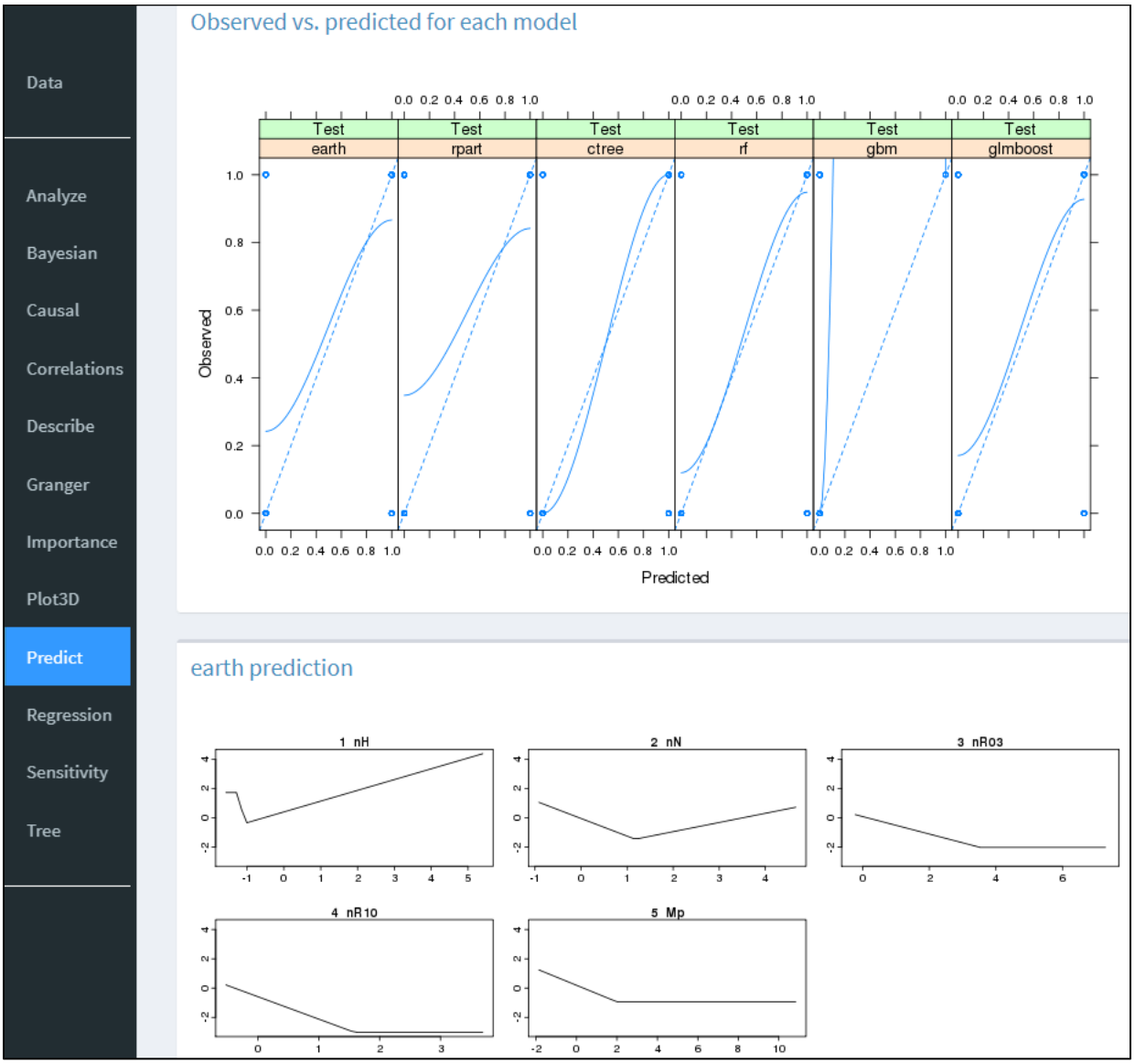


Figure 14. Screenshot of a subset of the additional model-specific outputs not covered by this tutorial.

4. References

Ames, B.N., Gurney, E.G., Miller, J.A., and Bartsch, H. (1972). Carcinogens as Frameshift Mutagens: Metabolites and Derivatives of 2-Acetylaminofluorene and Other Aromatic Amine Carcinogens. *Proceedings on the National Academy of Sciences*, 69(11): 3128-3132.

Kazius, J., McGuire, R., and Bursi, R. (2005). Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1): 312-320.