Supplementary Material for:

**Predicting Mouse Liver Microsomal Stability with “Pruned” Machine-Learning Models and Public Data**

Alexander L. Perryman,a Thomas P. Stratton,b Sean Ekins,c,d and Joel S. Freundlicha,b,\*

a Division of Infectious Disease, Department of Medicine, and the Ruy V. Lourenço Center for the Study of Emerging and Re-emerging Pathogens, Rutgers University–New Jersey Medical School, Newark, New Jersey 07103, United States

b Department of Pharmacology & Physiology, Rutgers University-New Jersey Medical School, Newark, New Jersey 07103, United States

c Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, North Carolina 27526, United States

d Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, California 94010, United States

Corresponding author:

Joel S. Freundlich, Ph.D.

Departments of Pharmacology & Physiology and Medicine

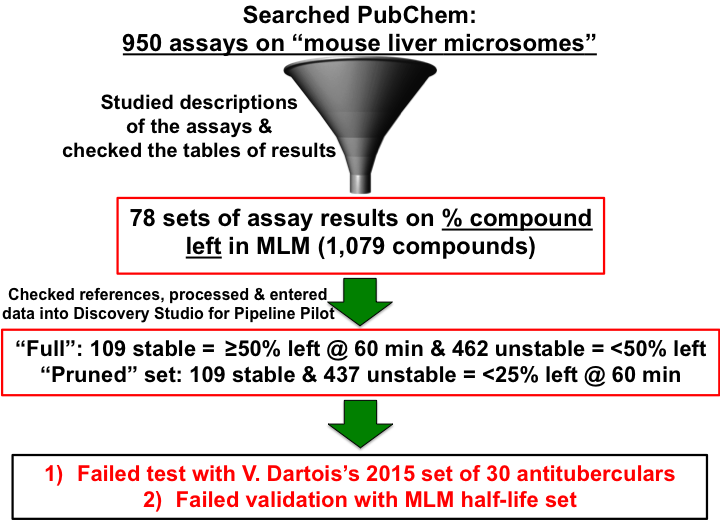
The Ruy V. Lourenço Center for Emerging and Re-emerging Pathogens

Rutgers University-New Jersey Medical School; Medical Sciences Building, I-503

185 South Orange Ave.

Newark, NJ 07103

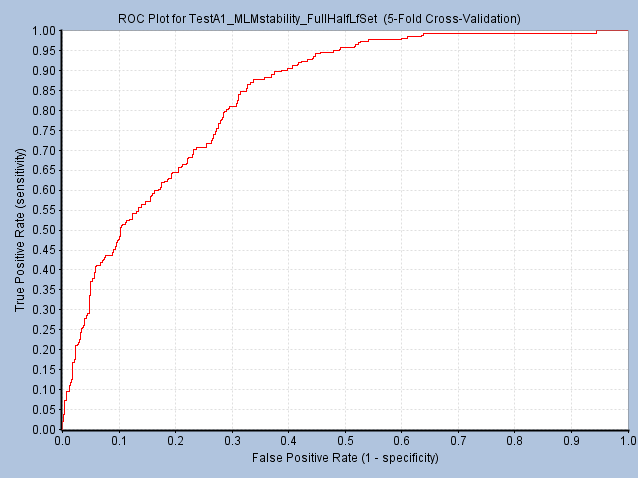
Phone: 973-972-7165; Fax: 973-972-1141; E-mail [freundjs@rutgers.edu](mailto:freundjs@rutgers.edu)

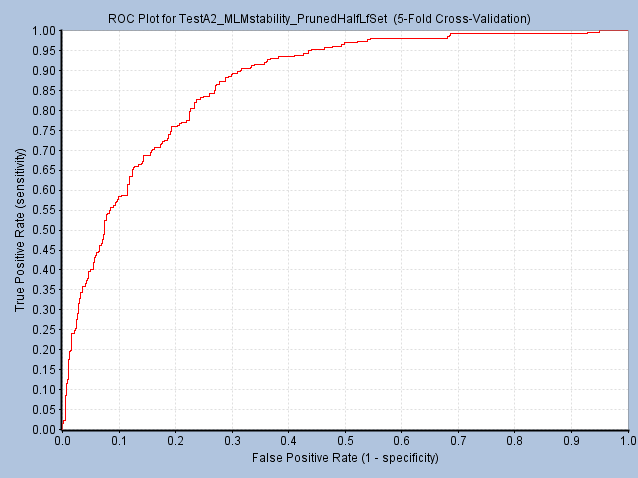


**Figure S1. Workflow used to build, test, and validate a different set of Bayesian models for predicting mouse liver microsomal stability (% compound left).** The PubChem BioAssay database was queried for data on the stability that different small molecule compounds displayed in the presence of mouse liver microsomes (MLM). The summaries and tables of results on PubChem were investigated, to determine which assay results and structural data files (sdf) to obtain. The cited publication for each set of MLM stability results (reported as either the % compound remaining or the % metabolized after incubation for a specified length of time in MLM) was examined, to ensure that the primary literature presented the same set of MLM stability values with the same units as the comma separated value (csv) file that was obtained from PubChem. Compounds with ambiguous stability data were removed, such as those that had ≥50% left after incubations that were less than 60 min. Assays that were performed for <60 min were only used to find “unstable” compounds. The verified, unambiguous MLM stability data were entered into a spreadsheet that Discovery Studio 4.0 (BIOVIA) created from the sdf files retrieved from PubChem. The “full” % compound left set (B1) of MLM data contained 109 compounds (19.1%) that had ≥50% of the original substance left after 60 min in MLM, which we defined as being “stable.” It contained 462 compounds (80.9%) that had <50% left after 60 min (or a shorter duration) in MLM, which we defined as “unstable”. The “pruned” % compound left set (B2) included the same 109 stable compounds (now 20.0%) as set B1, but it only included 437 “unstable” compounds (80.0%), since we “pruned” out (removed) the moderately unstable / moderately stable compounds that had 25 to 49% left after 60 min or 43 to 48% after 40 min. Two Bayesian classifier models were built—one using the “full” % compound left set and the other using the “pruned” % left set. Both Bayesian models were then tested with the Dartois 2015 set of 30 antitubercular compounds and performed poorly. A validation step was then performed by scoring the independent, “full” half-life set of MLM data (see **Figure 1**). Both % compound left Bayesians failed this external validation step, due to very poor ROC curves.

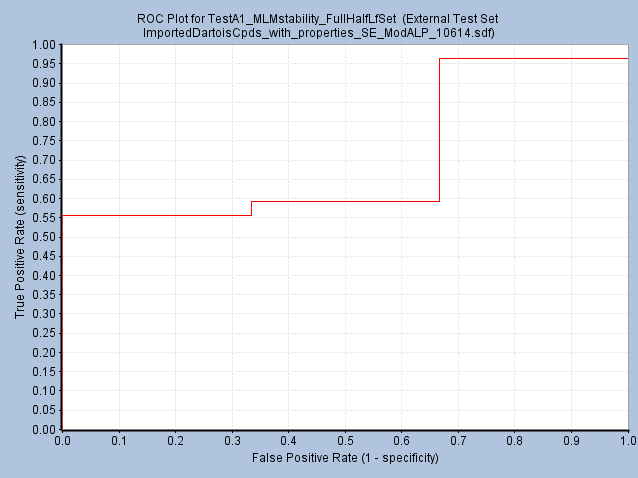
B

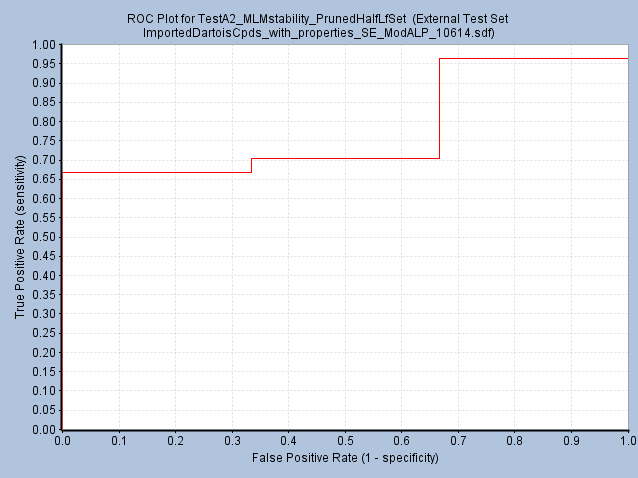
A





**Figure S2. Receiver Operating Characteristic (ROC) curves from internal five-fold cross-validation studies performed when creating the “full” and “pruned” half-life Bayesians to predict MLM stability.** When the Bayesian models were created in Pipeline Pilot 9.1 (BIOVIA), the default protocol involved performing five-fold cross-validation, in which 20% of the training set was randomly removed, the model was built with the remaining 80%, and the model was then evaluated on the 20% of data left out. The process was repeated five times. (**A**) The ROC curve, which plots the sensitivity (true positive rate) on the y-axis versus the false positive rate (1-specificity) on the x-axis, for the “full” half-life Bayesian is shown. (**B**) The slightly superior ROC curve for the “pruned” half-life Bayesian is displayed.

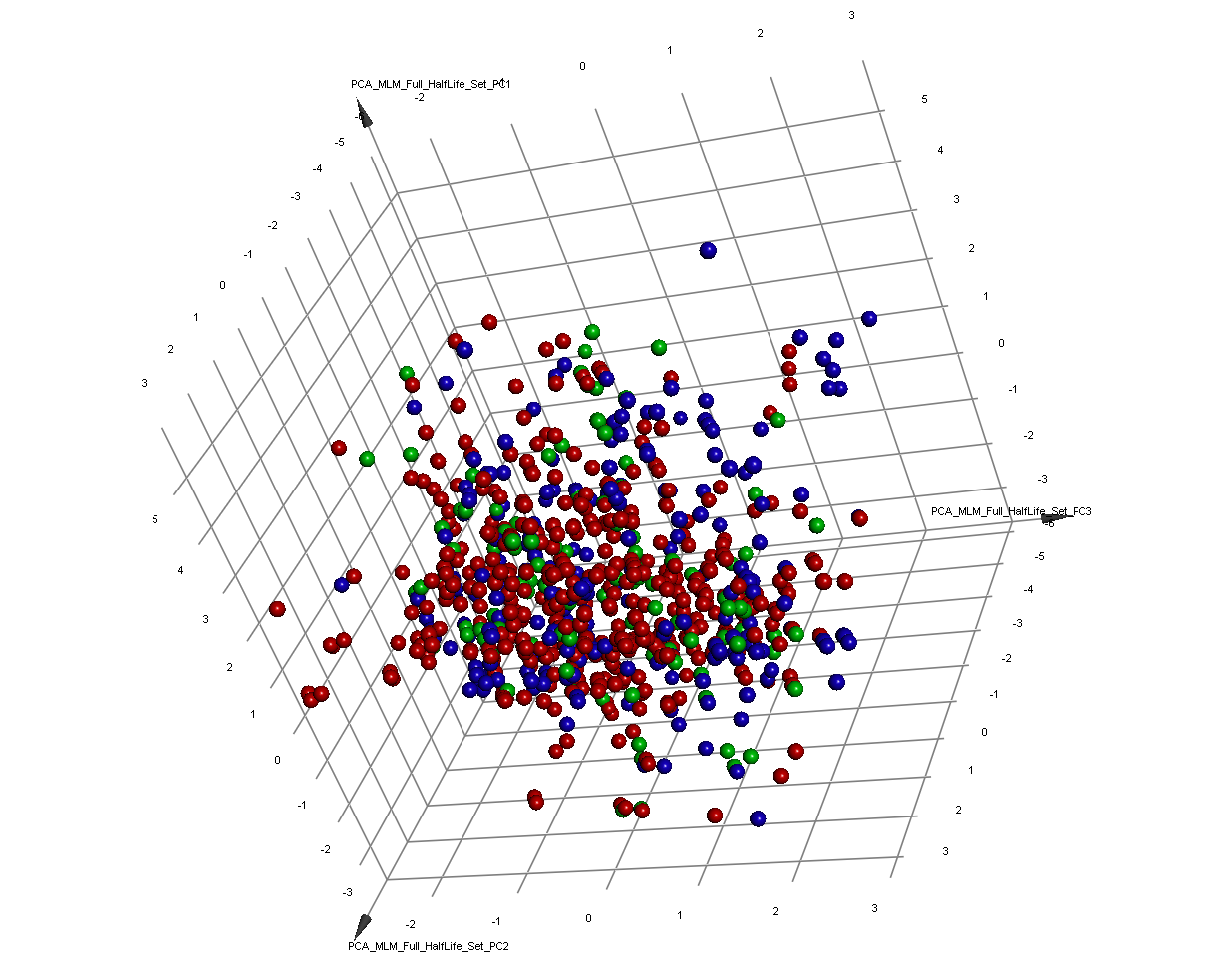




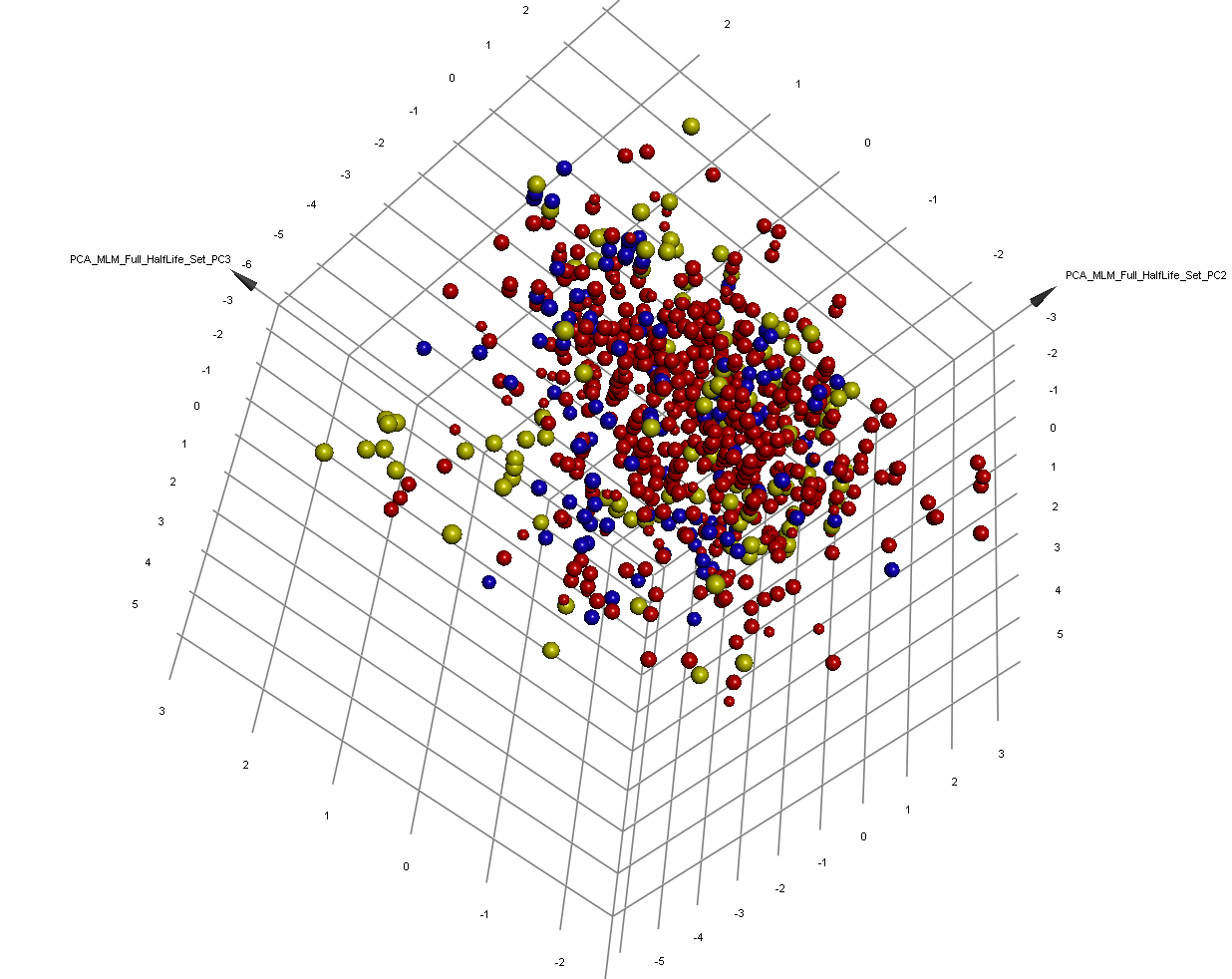
**Figure S3. Receiver Operating Characteristic (ROC) curves produced when testing the accuracy of the “full” and “pruned” half-life Bayesians by scoring the Dartois 2015 set of 30 antituberculars with known MLM stability.** The two MLM stability Bayesian models created (see **Figures 1** and **2**) were evaluated in Pipeline Pilot 9.1 (BIOVIA) with the independent Dartois 2015 set of 30 antitubercular compounds, of which 3 were “unstable”. (**A**) The ROC curve, which plots sensitivity versus 1-specificity, for the “full” half-life Bayesian is shown. (**B**) The ROC curve for the “pruned” half-life Bayesian is displayed. Since the pruned half-life Bayesian reached higher y-values quicker than the “full” half-life Bayesian (*i.e*., it detected a higher number of stable compounds, or true positives, amongst the compounds that had the top Bayesian scores), the “pruned” half-life Bayesian was more accurate in this external test.

A

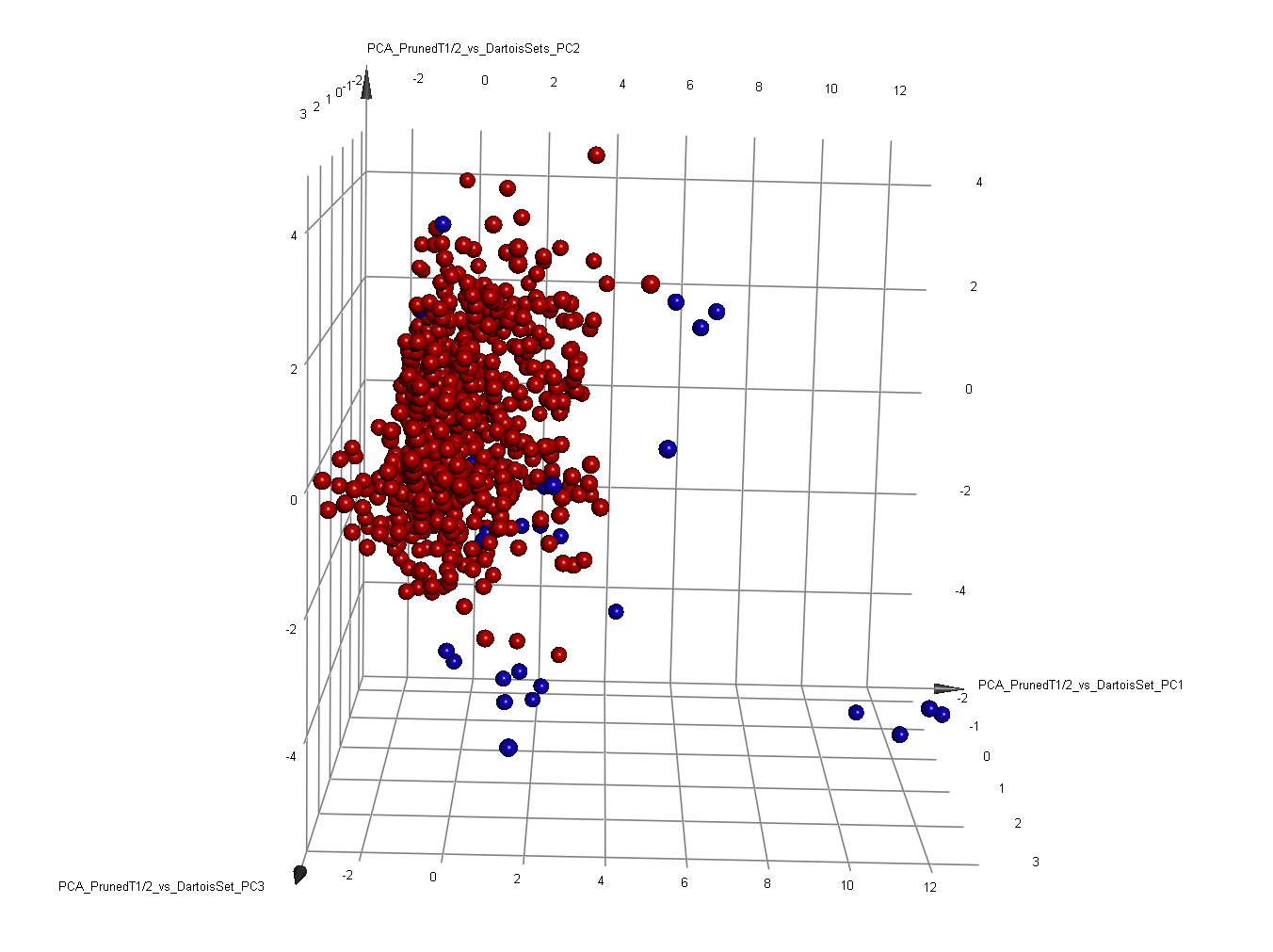
B



**Figure S4. Principal Component Analysis (PCA) comparing the chemical property space sampled by the full half-life set, the pruned half-life set, and the compounds that were removed to generate the pruned set.** The compounds contained in the MLM half-life training sets were characterized by performing a PCA in Pipeline Pilot 9.1, which was then visualized in Discovery Studio 4.0 (BIOVIA). PCA used eight interpretable descriptors (ALogP, molecular weight, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rings, number of aromatic rings, and molecular fractional polar surface area). Three principal components (PCs) explained 74% of the variance, while 4 PCs explained 86% of the variance. The top three PCs are plotted. “Stable” compounds, which had a t1/2 ≥ 60 min, are shown in blue. The moderately unstable compounds that were removed to generate the pruned half-life set (*i.e*., those with 30 ≤ t1/2 < 60 min) are rendered in green, and the rest of the “unstable” compounds are displayed in red. All three types of compounds are well dispersed throughout these three PCs. No significant clustering occurred.



**Figure S5. Principal Component Analysis (PCA) comparing the chemical space sampled by the “full” half-life set, the “pruned” half-life set, and the compounds that were removed to generate the “pruned” set.** The compounds contained in the MLM half-life training sets were characterized by performing a PCA in Pipeline Pilot 9.1 (BIOVIA), which was visualized in Discovery Studio 4.0 (BIOVIA). PCA used eight interpretable descriptors (ALogP, molecular weight, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rings, number of aromatic rings, and molecular fractional polar surface area). Three principal components (PCs) explain 74% of the variance, while 4 PCs explain 86% of the variance. The top three PCs are plotted. Very stable compounds, which had a t1/2 ≥120 min, are shown in yellow. The rest of the stable compounds, which had 60 ≤ t1/2 <120 min are in blue. The moderately unstable compounds that were removed to generate the “pruned” half-life set (*i.e*., those with 30 ≤ t1/2 <60 min) are rendered as red spheres with smaller radii, and the rest of the “unstable” compounds are displayed in red with larger radii. All four types of compounds are well dispersed throughout these three PCs. No significant clustering occurred.

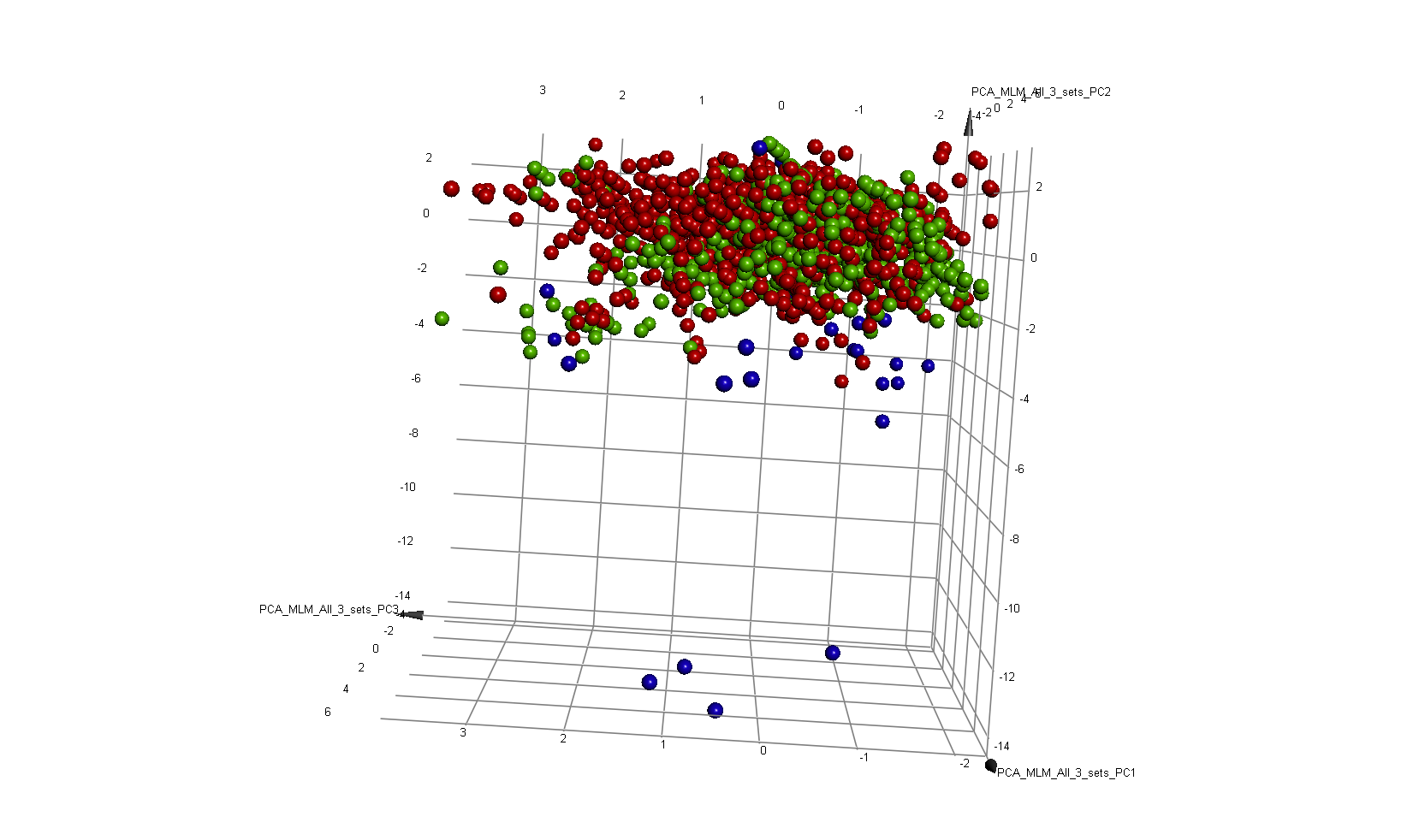


**Figure S6. Principal Component Analysis comparing the chemical space sampled by the pruned half-life set to the Dartois 2015 set of 30 antituberculars.** This PCA compared the physical properties of the training set that was used to create the best Bayesian model overall, the pruned half-life set (shown in red), to the independent test set of 30 antitubercular drugs (shown in blue). Three PCs explained 78% of the variance, while 4 PCs explained 87% of the variance. The top 3 PCs were plotted in Discovery Studio 4.0 (BIOVIA). These known antituberculars were generally in distinct areas of chemical space that were not thoroughly sampled by the pruned half-life set, yet the pruned half-life Bayesian predicted their MLM stability classification accurately.

**Commentary: Structural Comparison of the Half-life and Percent Compound Left Sets and the Effects of Removing Duplicate Compounds on Enrichment Factors**

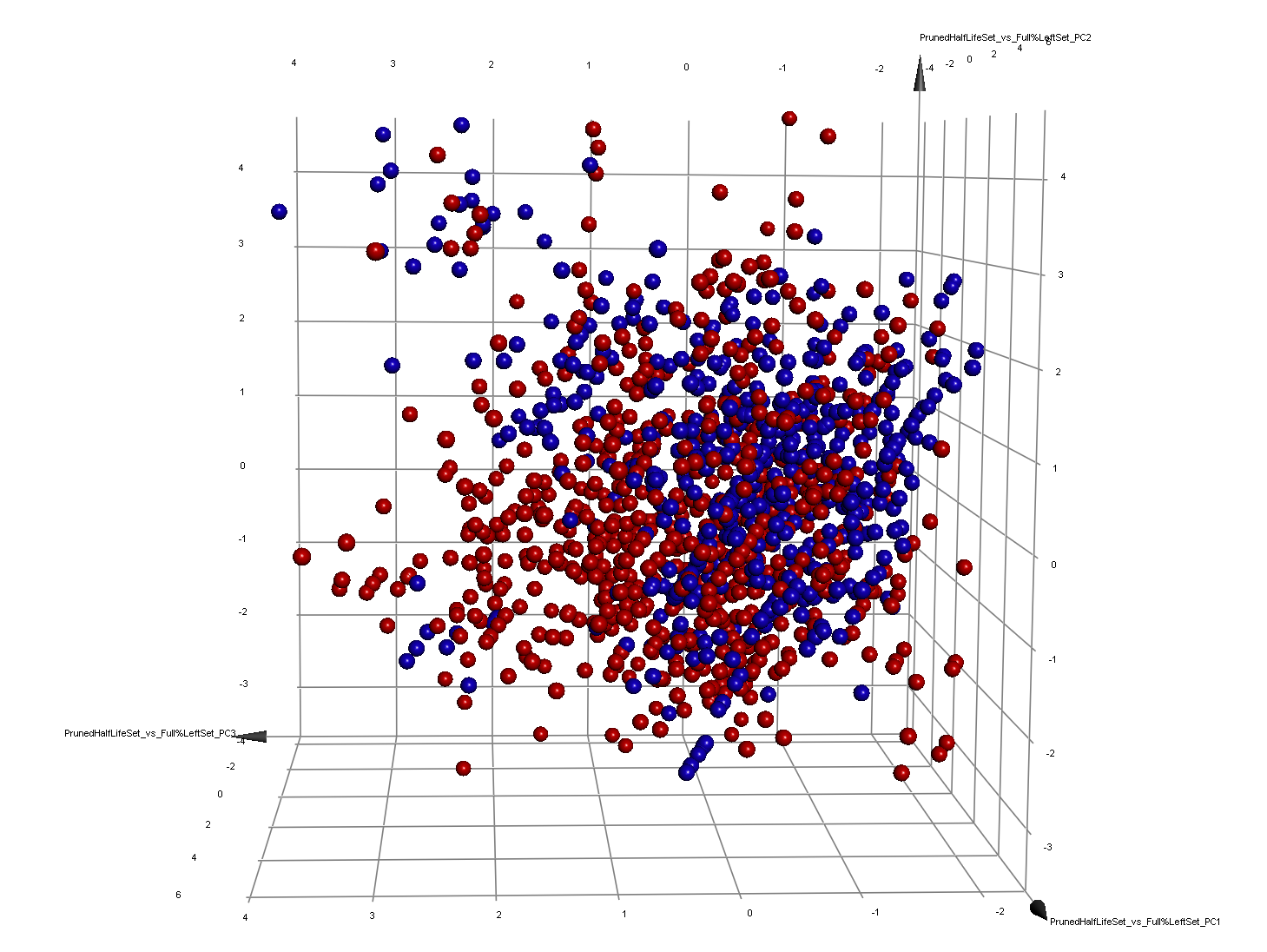
The full percent compound left set has an average of the closest distance of 0.649 to the full half-life set and 0.654 to the pruned half-life set. Similarly, for the compounds that received the top 50 Bayesian scores, the average of the closest distance values were 0.677 and 0.674 to the full and pruned half-life training sets, with minimum values of 0 and 0, and maximum values of 0.905 and 0.927, respectively. The average of the maximum Tanimoto similarity values displayed between each member of the full percent compound left set and every member of the full half-life set was 0.22, with a minimum of 0.13 and a maximum of 1.00. The same Tanimoto similarity values were observed when comparing the full percent compound left set and the pruned half-life set. Note that the observance of minimum closest distance values of 0 and maximum Tanimoto similarity values of 1.00 indicate that some compounds were present in both the half-life sets and the full percent compound left set. Although the duplicate compounds were removed within each individual half-life or percent compound left set, compounds that were present in both types of sets were not removed. 9 of the compounds in the full half-life set were present in the full percent compound left set. 9 out of 571 represents 1.6% of this validation set. If these 9 compounds are removed, then the next closest distance value was 0.455. Similarly, 8 compounds overlapped between the pruned half-life set and the full percent compound left set. If these 8 duplicates are removed, then the next closest distance value was 0.455, as well. Although there were 8 or 9 duplicates between the pruned or full half-life sets and the full percent compound left set, only one of these compounds was present in the top 50 compounds according to either model’s Bayesian scores: this duplicate was ranked as the 45th compound by the full half-life model and as the 48th compound by the pruned half-life model. When these duplicates were removed, it only slightly decreased the enrichment factor for the full half-life Bayesian from 3.46 to 3.35, since it predicted the 51st compound incorrectly. However, the pruned half-life Bayesian predicted the 51st top-scoring compound correctly; thus, removing that duplicate did not change its enrichment factor from 3.25. Unlike the trends in 2D similarity, with respect to physiochemical properties, a fair amount of similarity exists between these sets in PCA plots (Fig. S7 and Fig. S8).

Of the 8 or 9 compounds that were present in both the pruned or full half-life training sets and the percent compound left validation set, 3 of these compounds were present in the top-scoring results for the pruned half-life CDD Bayesian (ranked as 9th, 12th and 39th), and 3 were also present in the top-scoring results for the full half-life CDD Bayesian (ranked as 7th, 13th, and 34th). When these duplicate compounds were removed, the enrichment factors for the CDD Bayesians decreased to 1.05 for the full model and 0.84 for the pruned model (*i.e*., near or below random chance, respectively).

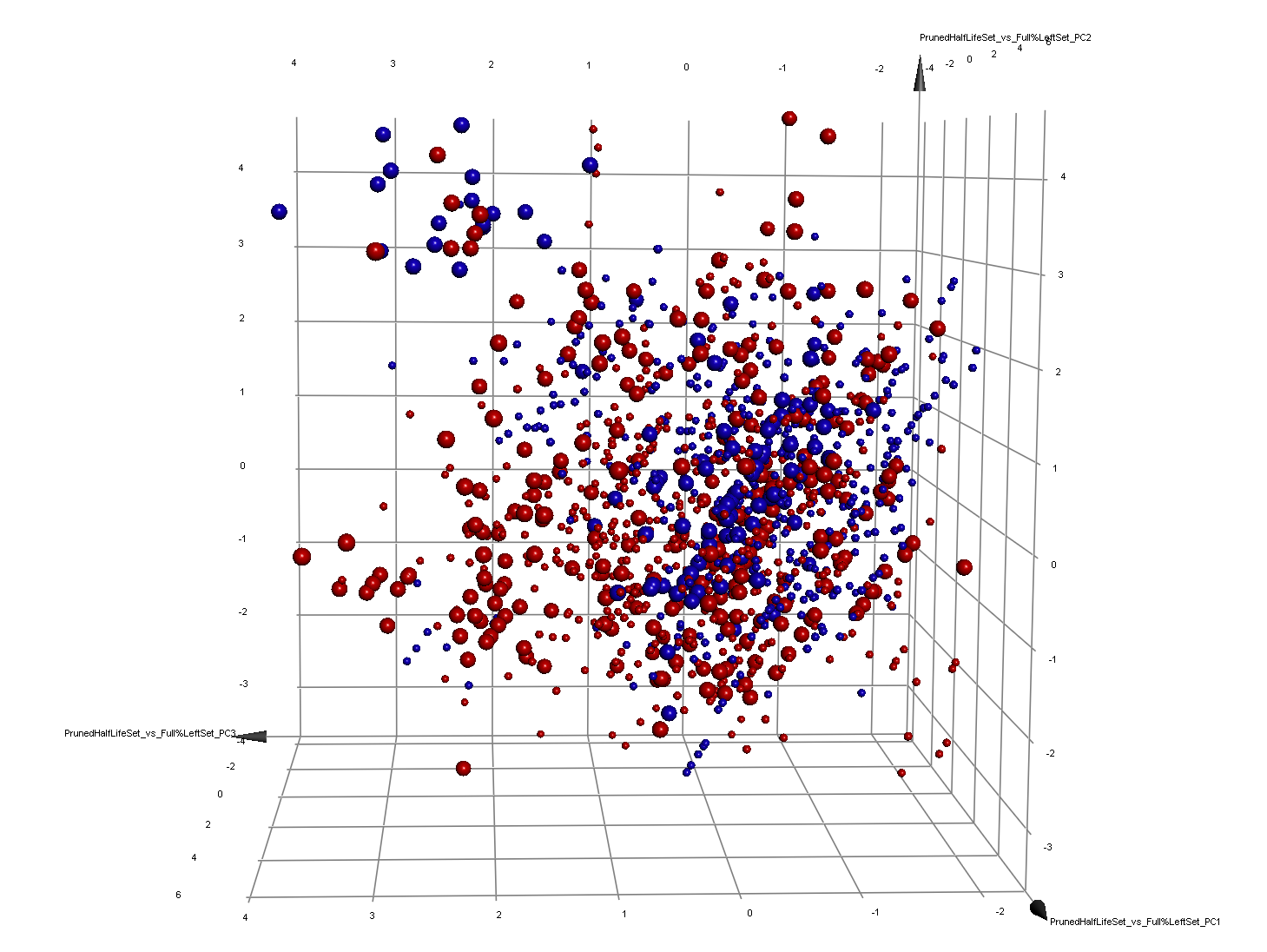


**Figure S7. Principal Component Analysis (PCA) comparing the chemical space sampled by the “full” half-life set, the “full” percent compound left set, and the Dartois 2015 set of known antitubercular drugs.** Three principal components (PCs) explain 76% of the variance, while 4 PCs explain 85% of the variance. The top three PCs were plotted in Discovery Studio 4.1 (BIOVIA). The “full” half-life set is shown in red, the “full” percent compound left set is in green, and the set of known antitubercular drugs is rendered in blue. These known antituberculars were generally in distinct areas of chemical space that were not thoroughly sampled by the “full” half-life set or the “full” % compound left set, yet the “full” and “pruned” half-life MLM Bayesian models predicted the stability classification of these antitubercular drugs accurately.

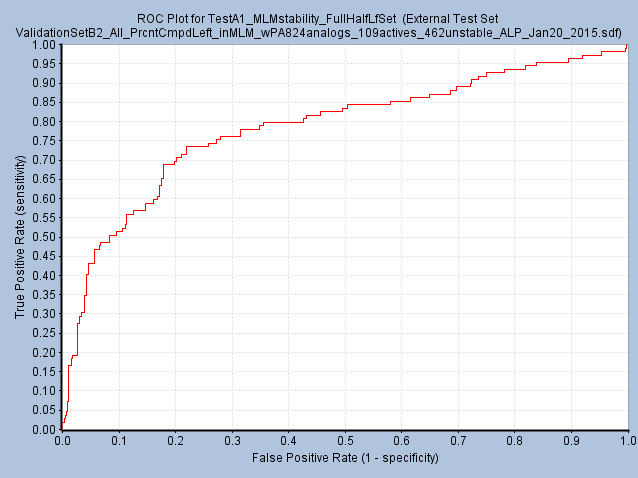
**A**

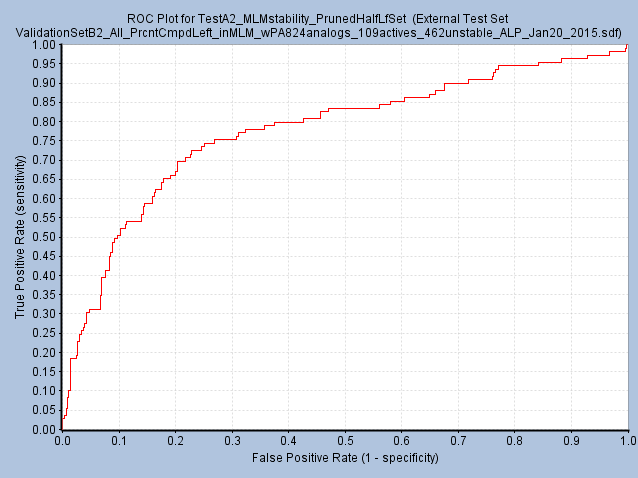


**B**



**Figure S8. Principal Component Analysis comparing the chemical space sampled by the pruned half-life training set to the full percent compound left validation set.** (**A**) This PCA compared the physical properties of the training set that was used to create the best Bayesian model overall, the pruned half-life set (in red), to the full % compound left set (shown in blue), which was used for external validation. Three PCs explain 74% of the variance, while 4 PCs explain 84% of the variance. The top 3 PCs are plotted. Both sets of compounds sample most of the regions within these 3 PCs, but the pruned half-life set covers this chemical space more thoroughly. In (**B**) the unstable compounds from both sets have smaller radii. The stable compounds are dispersed throughout these 3 PCs, without significant clustering that differentiates most stable from unstable compounds.

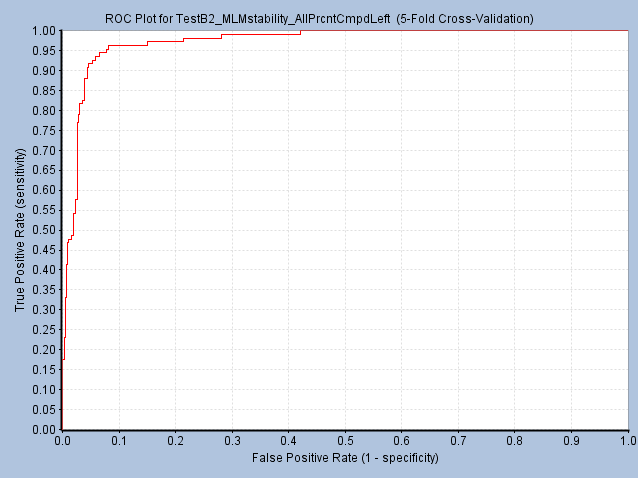


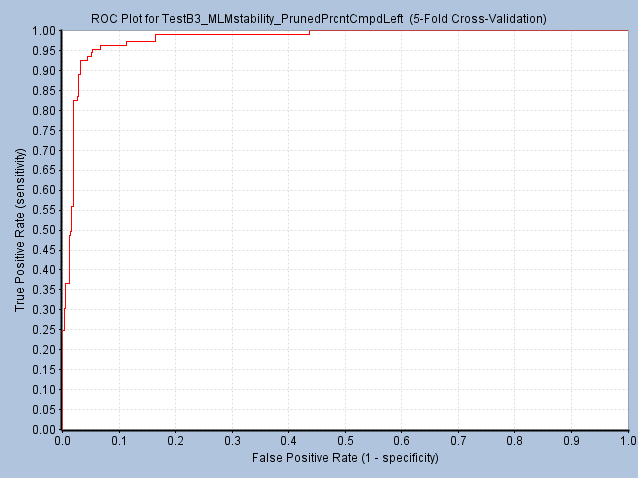


**Figure S9. Receiver Operating Characteristic (ROC) curves produced when validating the accuracy of the “full” and “pruned” half-life Bayesians by scoring the “full” % compound left set.** The two MLM stability Bayesian models created (see **Figures 1** and **S4**) were validated by testing their predictive power with another independent set of compounds that had known MLM stability values (from PubChem, see **Figure S1**). (**A**) The ROC curve, which plots sensitivity versus 1-specificity, for the “full” half-life Bayesian is shown. (**B**) The ROC curve for the “pruned” half-life Bayesian is presented. For this very challenging validation set, in which only 19% of the compounds are stable, both half-life Bayesians performed well and produced similar ROC curves.

A

B

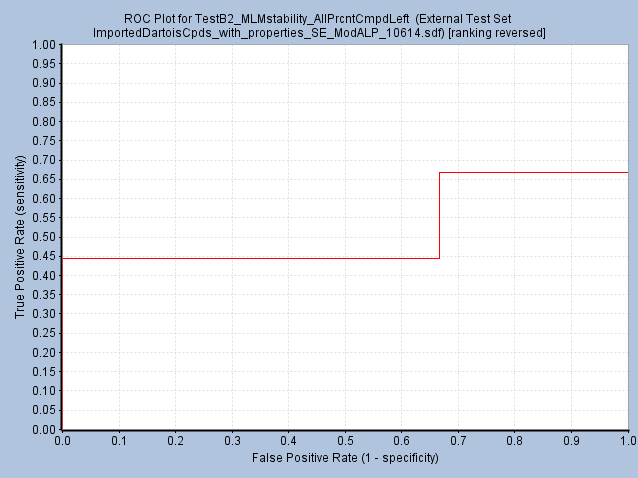


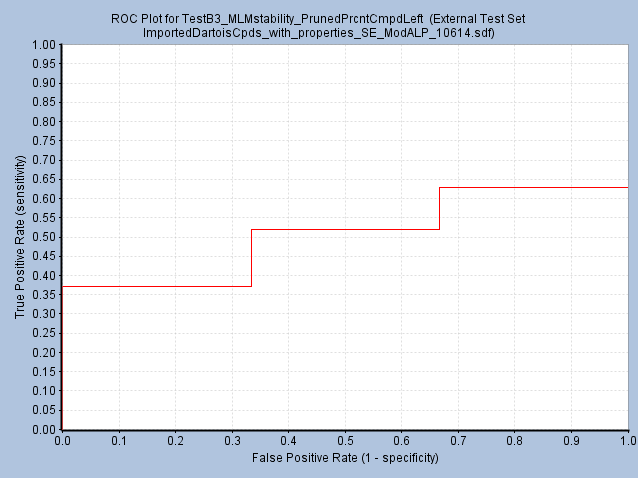


**Figure S10. Receiver Operating Characteristic (ROC) curves from internal five-fold cross-validation studies performed when creating the “full” and “pruned” percent compound left Bayesians to predict MLM stability.** When Bayesian models were created in Pipeline Pilot 9.1 (BIOVIA), the default protocol involved performing five-fold cross-validation, in which 20% of the training set was randomly removed, the model was built with the remaining 80%, and the model was then evaluated on the 20% of data that was left out. The process was repeated five times. (**A**) The ROC curve, which plots the sensitivity (true positive rate) on the y-axis versus the false positive rate (1-specificity) on the x-axis, for the “full” percent compound left Bayesian is shown. (**B**) The very similar ROC curve for the “pruned” percent compound left Bayesian is displayed. Both of these internal ROC curves are near perfect.

B

A





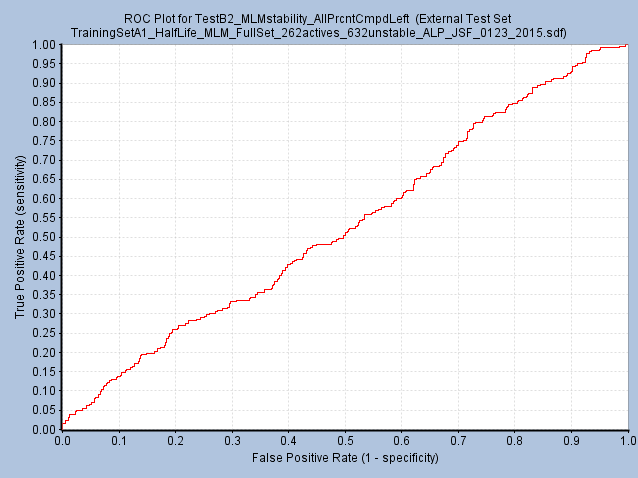
**Figure S11. Receiver Operating Characteristic (ROC) curves produced when testing the accuracy of the “full” and “pruned” percent compound left Bayesians by scoring the Dartois set of 30 antituberculars with known MLM stability.** The two MLM stability Bayesian models created (see **Figures S1** and **S10**) were evaluated with an independent set of 30 antitubercular compounds, of which 3 were “unstable”. (**A**) The ROC curve, which plots the sensitivity vs. 1-specificity, for the “full” compound left Bayesian is shown. (**B**) The ROC curve for the “pruned” percent compound left Bayesian is displayed. Both of these percent compound left Bayesians displayed much less predictive power than the half-life Bayesians in this independent test.

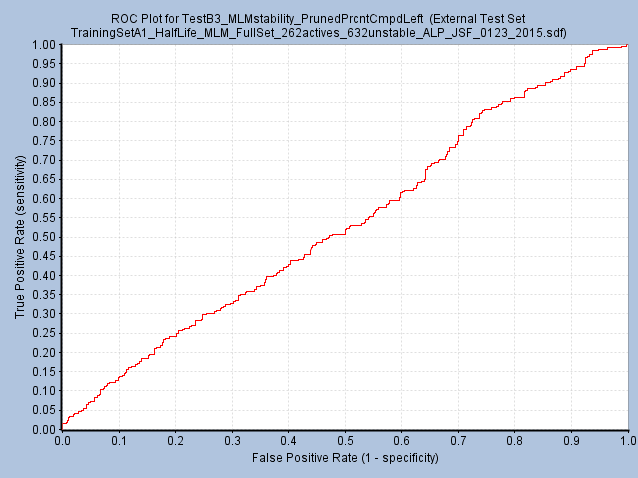
B

A

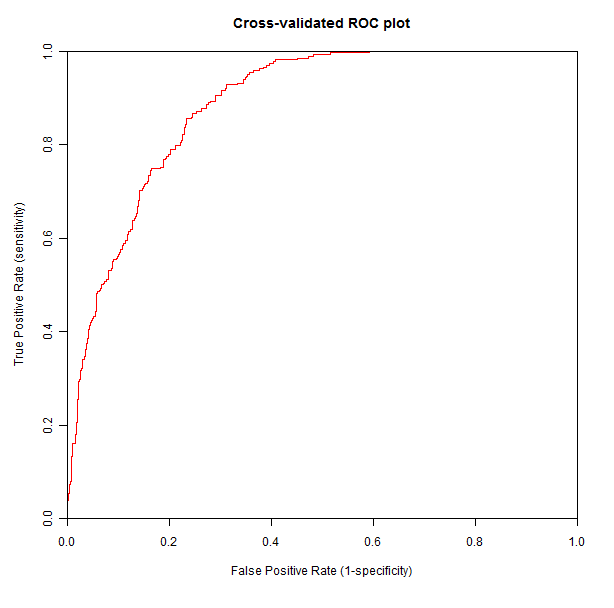
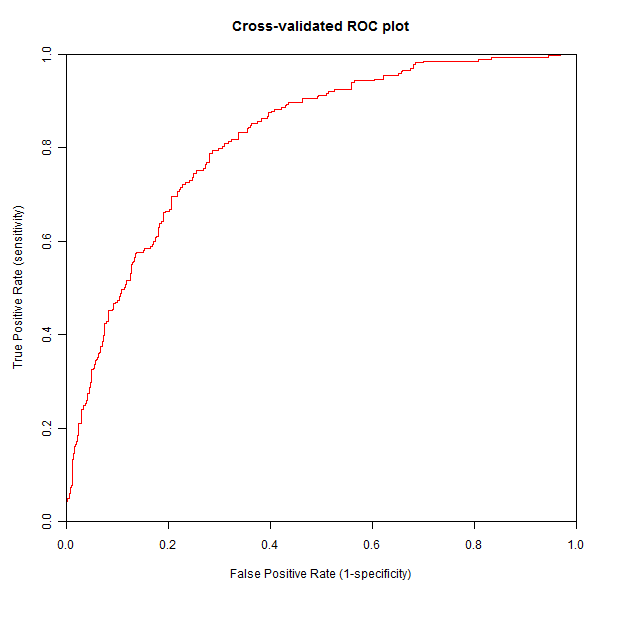
B

A





**Figure S12. Receiver Operating Characteristic (ROC) curves produced when assessing the accuracy of the “full” and “pruned” percent compound left Bayesians by scoring the independent, “full” half-life validation set.** The two MLM stability Bayesian models created (see **Figures S1** and **S10**) were investigated by testing their predictive power against another independent set of compounds that had known MLM stability values (from PubChem, see **Figure 1**). (**A**) The ROC curve, which plots sensitivity versus 1-specificity, for the “full” percent compound left Bayesian is shown. (**B**) The ROC curve for the “pruned” percent compound left Bayesian is presented. Both of these percent compound left Bayesians failed this validation study, by displaying ROC curves that were similar to random (*i.e*., the diagonal).



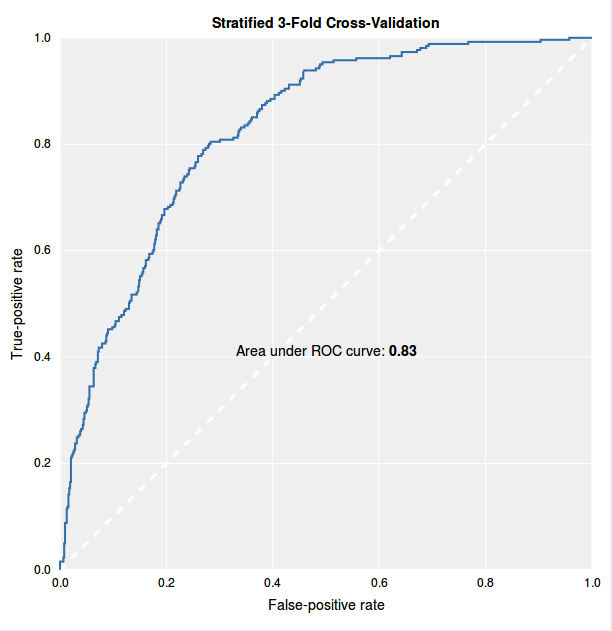
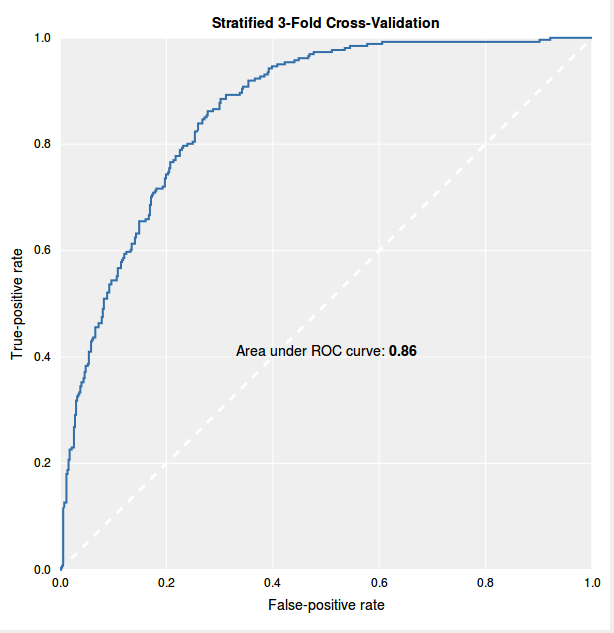
**Figure S13. Receiver Operating Characteristic (ROC) curves from internal five-fold cross-validation studies performed when creating “full” and “pruned” half-life Support Vector Machine models to predict MLM stability.** When the Support Vector Machine (SVM) models to predict MLM stability were created in Pipeline Pilot, five-fold cross-validation was performed. These SVM models were created using the same “full” and “pruned” MLM half-life training sets. The same set of 9 descriptors that were used to create the Bayesian models were also utilized to create these SVM models. (**A**) The ROC curve, which plots the sensitivity versus1-specificity, for the “full” half-life SVM model is shown. (**B**) The slightly superior ROC curve for the “pruned” half-life SVM model is displayed.

A) “Full” half-life SVM

B) “Pruned” half-life SVM

**B**

**A**

**Figure S14. Receiver Operating Characteristic (ROC) curves from internal three-fold cross-validation studies performed when using CDD to create “full” and “pruned” half-life Bayesian models to predict MLM stability.** The Bayesian modeling tools in CDD (www.collaborativedrug.com) use only a single descriptor—the open source version of FCFP\_6 fingerprints (which describe the 2D topology of compounds). The CDD Bayesians were described using three-fold cross-validation, in which 33% of the training set was removed, the model was trained on the remaining 66%, and the model was then evaluated by using it to score the 33% that was initially left out. This was repeated three times. These CDD Bayesian models were created using the same “full” and “pruned” MLM half-life training sets used for the other machine learning models. (**A**) The ROC curve, which plots the sensitivity versus 1-specificity, for the “full” half-life CDD Bayesian model is shown. It displayed an area under the ROC curve value of 0.83 (with 1.0 being a perfect model). (**B**) The slightly better ROC curve for the “pruned” half-life CDD Bayesian model is displayed. It produced an area under the ROC curve of 0.86.

**Table S-I. Internal statistics from five-fold cross-validation studies performed when creating different machine learning models to predict MLM stability**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Half-Life MLM** | **ROC**  **score** | **ROC**  **Rating a** | **Sensitivity**  **%** | **Specificity**  **%** | **Concordance**  **%** |
| Full  Half-Life  Bayesian | 0.835 | good | 92.7 | 72.2 | 78.2 |
| Pruned  Half-Life Bayesian | 0.870 | good | 93.1 | 80.9 | 85.1 |
| Full  Half-Life SVM | 0.819 | good |  |  |  |
| Pruned  Half-Life SVM | 0.885 | good |  |  |  |
| Full  Half-Life Random Forest | 0.817 | good |  |  |  |
| Pruned  Half-Life Random Forest | 0.828 | good |  |  |  |

Notes: (a) the “ROC rating” is a qualitative grading system that Pipeline Pilot outputs, which ranges from: fail < poor < fair < good < excellent. The internal sensitivity, specificity, and concordance values were only output for the Bayesian models (these scores for the other models were not available).

**Table S-II. External test statistics from evaluating the accuracy of different machine learning models that predict MLM stability by using them to score the Dartois 2015 set of 30 known antitubercular drugs.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Half-Life MLM** | **External ROC score a** | **External Sensitivity**  **%** | **External Specificity**  **%** | **External**  **Concor-dance**  **%** | **Stability**  **Hit Rate b** | **Unstable**  **True –**  **Filtered c** |
| Full  Half-Life  Bayesian | 0.704 | 81.5 | 33.3 | 76.7 | 22 / 24 | 1 / 3 |
| Pruned  Half-Life Bayesian | 0.778 | 81.5 | 33.3 | 76.7 | 22 / 24 | 1 / 3 |
| Full  Half-Life SVM |  | 22.2 | 66.7 | 26.7 | 6 / 7 | 2 / 3 |
| Pruned  Half-Life SVM |  | 59.3 | 66.7 | 60.0 | 16 / 17 | 2 / 3 |
| Full  Half-Life Random Forest | 0.716 | 70.4 | 33.3 | 66.7 | 19 / 21 | 1 / 3 |
| Pruned  Half-Life Random Forest | 0.531 | 74.1 | 33.3 | 70.0 | 20 / 22 | 1 / 3 |

Notes: (a) the “external ROC score” was not available for the SVM models (*i.e*., Pipeline Pilot did not produce that result for this type of machine learning model). (b) The “stability hit rate” is equivalent to the “positive predicted value” and is calculated by dividing the number of true positives by the sum of the number of true positives plus the number of false positives. True positives are compounds that were correctly predicted to be stable, while false positives are unstable compounds that were incorrectly classified as stable. (c) “Unstable true negatives filtered” corresponds to the number of correctly predicted unstable compounds divided by the total number of unstable compounds. In the Dartois set of 30 antitubercular drugs, only 3 compounds were unstable in MLM.

**Table S-III. External validation statistics from evaluating the accuracy of different machine learning models that predict MLM stability by using them to score the full percent compound left set of compounds.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Half-Life MLM** | **External ROC score a** | **External Sensitivity**  **%** | **External Specificity**  **%** | **External**  **Concor-dance**  **%** | **Stability**  **Hit Rate b** | **Unstable**  **True –**  **Filtered c** |
| Full  Half-Life  Bayesian | 0.785 | 83.5 | 49.8 | 56.2 | 91/323  28.2% | 230 |
| Pruned  Half-Life Bayesian | 0.777 | 83.5 | 44.8 | 55.2 | 91/346  26.3% | 207 |
| Full  Half-Life SVM |  | 10.1 | 92.2 | 76.5 | 11/47  23.4% | 426 |
| Pruned  Half-Life SVM |  | 23.9 | 83.8 | 72.3 | 26/101  25.7% | 387 |
| Full  Half-Life Random Forest | 0.560 | 25.7 | 70.8 | 62.2 | 28/163  17.2% | 327 |
| Pruned  Half-Life Random Forest | 0.507 | 26.6 | 67.7 | 59.9 | 29/178  16.3% | 313 |

Notes: (a) the “external ROC score” was not available for the SVM models (*i.e*., Pipeline Pilot did not produce that result for this type of machine learning model). (b) The “stability hit rate” is equivalent to the “positive predicted value” and is calculated by dividing the number of true positives by the sum of the number of true positives plus the number of false positives. True positives are compounds that were correctly predicted to be stable, while false positives are unstable compounds that were incorrectly classified as stable. (c) “Unstable true negatives filtered” corresponds to the number of correctly predicted unstable compounds (which should be divided by the total number of unstable compounds, 462).

**Table S-IV. External test statistics from evaluating the accuracy of different machine learning models (constructed using either 9 descriptors or just 1, FCFP\_6) by using them to score the Dartois 2015 set of 30 known antituberculars.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Half-Life Bayesian | External ROC score | External Sensitivity  % | External Specificity  % | External  Concor-dance  % | Stability  Hit Rate | Unstable  True –  Filtered |
| Full t1/2  with 9 a | 0.704 | 81.5 | 33.3 | 76.7 | 22 / 24 | 1 / 3 |
| Pruned t1/2  with 9 | 0.778 | 81.5 | 33.3 | 76.7 | 22 / 24 | 1 / 3 |
| Full t1/2  with 1 b | 0.790 | 81.5 | 33.3 | 76.7 | 22 / 24 | 1 / 3 |
| Pruned t1/2  with 1 | 0.815 | 81.5 | 33.3 | 76.7 | 22 / 24 | 1 / 3 |

Notes: (a) “With 9” indicates that all 9 descriptors were utilized when creating that Bayesian model, while (b) “with 1” means that only the FCFP\_6 fingerprints that describe 2D topology were used.

**Testing Different Sets of Run Parameters with the pruned and full Half-life Bayesians**

In addition to testing our full and pruned half-life training sets with different types of machine learning models, we also examined how a few different sets of run parameters would affect these Bayesian models. The default protocol for constructing Bayesians in Pipeline Pilot 9.1 involves using 10 bins when characterizing the descriptors. We investigated how “pruning” the half-life training set affected the accuracy of Bayesians that utilized FCFP\_6 fingerprints with 10 bins, FCFP\_12 (that characterize which functional groups are connected to each other, up to and including 12 spheres, or dimensions, of topology) with 10 bins, and FCFP\_12 with 20 bins. The internal statistics from five-fold cross-validation (see **Table S-V**) demonstrate that “pruning” the training set also increased the ROC score (0.873 vs. 0.838), specificity (80.9% vs. 72.2% and 81.5% vs. 73.1%) and concordance (85.8% vs. 78.9% and 86.2% vs. 79.4%), as compared to the full half-life Bayesians that were constructed using FCFP\_12 with 10 bins and FCFP\_12 with 20 bins, respectively. The high sensitivity values were maintained (at 95%) for these Bayesians that involved using different run parameters, as well.

In the external tests with the Dartois 2015 set of antituberculars, similar trends were observed (see **Table S-VI**). The pruned half-life Bayesians produced better external ROC scores (0.753 vs. 0.691 and 0.765 vs. 0.704) than the full Bayesians that were built using FCFP\_12 with 10 bins and FCFP\_12 with 20 bins, respectively. The pruned half-life Bayesian with FCFP\_12 and 10 bins also displayed a better external concordance (86.7% vs. 80.0%) and stability hit rate (25 out of 27 versus 23 out of 25) than the corresponding full Bayesian model. For the other metrics, the pruned and full Bayesians that utilized different run parameters were equivalent in this external test. In the external study with the full percent compound left validation set, the trends from the Bayesians built with these two new sets of run parameters were similar to the aforementioned trends displayed by the pruned and full Bayesians constructed with FCFP\_6 and 10 bins (see **Table S-VII**), except that these new run parameters improved the sensitivity values for the pruned models, instead of just maintaining it.

Observing similar trends in the internal five-fold cross-validation studies and external tests (with respect to “pruned” models showing some enhanced predictive power over the corresponding “full” Bayesian models) for Bayesians that were constructed with three different sets of run parameters further supports the hypothesis that “pruning” the training set can be a useful strategy in machine learning models. It also supports the robustness of the half-life training set that we curated and of our overall modeling approach.

**Table S-V. Internal statistics from five-fold cross-validation studies performed when using different sets of run parameters in Pipeline Pilot to create Bayesian models that predict MLM stability**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MLM stability Bayesian** | **ROC**  **score** | **ROC**  **rating a** | **Sensitivity**  **%** | **Specificity**  **%** | **Concordance**  **%** |
| Full  Half-Life  FCFP\_6 (10 bins) b | 0.835 | good | 92.7 | 72.2 | 78.2 |
| Pruned  Half-Life FCFP\_6 (10 bins) | 0.870 | good | 93.1 | 80.9 | 85.1 |
| Full  Half-Life  FCFP\_12 (10 bins) | 0.838 | good | 95.0 | 72.2 | 78.9 |
| Pruned  Half-Life FCFP\_12 (10 bins) | 0.873 | good | 95.0 | 80.9 | 85.8 |
| Full  Half-Life  FCFP\_12 20 bins | 0.838 | good | 94.7 | 73.1 | 79.4 |
| Pruned  Half-Life FCFP\_12 20 bins | 0.873 | good | 95.0 | 81.5 | 86.2 |

Notes: (a) the “ROC rating” is a qualitative grading system that Pipeline Pilot 9.1 (BIOVIA) outputs, which ranges from: fail < poor < fair < good < excellent. (b) The run parameters in parentheses correspond to the default settings in the “create Bayesian model” protocol in Pipeline Pilot. For each set of run parameters investigated, the “pruned” half-life training set always produced a more accurate model than the “full” half-life training set, according to these internal statistics from five-fold cross-validation.

**Table S-VI. External test statistics from evaluating the accuracy of MLM half-life Bayesians that were created using different sets of run parameters, by using these Bayesians to score the Dartois 2015 set of 30 known antitubercular drugs.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MLM stability Bayesian** | **External ROC score** | **External Sensitivity**  **%** | **External Specificity**  **%** | **External**  **Concor-dance**  **%** | **Stability**  **Hit Rate b** | **Unstable**  **True –**  **Filtered c** |
| Full  Half-Life  FCFP\_6 (10 bins) a | 0.704 | 81.5 | 33.3 | 76.7 | 22 / 24 | 1 / 3 |
| Pruned  Half-Life FCFP\_6 (10 bins) | 0.778 | 81.5 | 33.3 | 76.7 | 22 / 24 | 1 / 3 |
| Full  Half-Life  FCFP\_12 (10 bins) | 0.691 | 85.2 | 33.3 | 80.0 | 23 / 25 | 1 / 3 |
| Pruned  Half-Life FCFP\_12 (10 bins) | 0.753 | 92.6 | 33.3 | 86.7 | 25 / 27 | 1 / 3 |
| Full  Half-Life  FCFP\_12 20 bins | 0.704 | 85.2 | 33.3 | 80.0 | 23 / 25 | 1 / 3 |
| Pruned  Half-Life FCFP\_12 20 bins | 0.765 | 85.2 | 33.3 | 80.0 | 23 / 25 | 1 / 3 |

Notes: (a) the run parameters in parentheses correspond to the default settings in the “create Bayesian model” protocol in Pipeline Pilot 9.1 (BIOVIA). (b) The “stability hit rate” is equivalent to the “positive predicted value” and is calculated by dividing the number of true positives by the sum of the number of true positives plus the number of false positives. True positives are compounds that were correctly predicted to be stable, while false positives are unstable compounds that were incorrectly classified as stable. (c) “Unstable true negatives filtered” corresponds to the number of correctly predicted unstable compounds divided by the total number of unstable compounds. In the Dartois 2015 set of 30 antitubercular drugs, only 3 compounds were unstable in MLM. For each set of run parameters investigated, the “pruned” half-life Bayesian model always displayed similar or better accuracy (for each type of external test statistic) than the corresponding “full” half-life Bayesian.

**Table S-VII. External validation statistics from evaluating the accuracy of using different sets of run parameters to create MLM half-life Bayesians, by using them to score the “full” percent compound left set of compounds.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MLM Stability Bayesian** | **External ROC score** | **External Sensitivity**  **%** | **External Specificity**  **%** | **External**  **Concor-dance**  **%** | **Stability**  **Hit Rate b** | **Unstable**  **True –**  **Filtered c** |
| Full  Half-Life  FCFP\_6 (10 bins) a | 0.785 | 83.5 | 49.8 | 56.2 | 91/323  28% | 230 |
| Pruned  Half-Life FCFP\_6 (10 bins) | 0.777 | 83.5 | 44.8 | 55.2 | 91/346  26% | 207 |
| Full  Half-Life  FCFP\_12 (10 bins) | 0.789 | 85.3 | 42.4 | 50.6 | 93/359  26% | 196 |
| Pruned  Half-Life FCFP\_12 (10 bins) | 0.780 | 89.0 | 31.8 | 42.7 | 97/412  24% | 147 |
| Full  Half-Life  FCFP\_12 20 bins | 0.788 | 85.3 | 40.5 | 49.0 | 93/368  25% | 187 |
| Pruned  Half-Life FCFP\_12 20 bins | 0.779 | 87.2 | 37.7 | 47.1 | 95/383  25% | 174 |

Notes: (a) the run parameters in parentheses correspond to the default settings in the “create Bayesian model” protocol in Pipeline Pilot 9.1 (BIOVIA). (b) The “stability hit rate” is equivalent to the “positive predicted value” and is calculated by dividing the number of true positives by the sum of the number of true positives plus the number of false positives. True positives are compounds that were correctly predicted to be stable, while false positives are unstable compounds that were incorrectly classified as stable. (c) “Unstable true negatives filtered” corresponds to the number of correctly predicted unstable compounds (there were a total of 462 unstable compounds in the full percent compound left validation set). For each corresponding set of run parameters, the “pruned” half-life training set produced a Bayesian model that displayed better (or similar) sensitivity than the “full” half-life Bayesian, when scoring the “full” percent compound left validation set. However, for the other external statistics, the “full” half-life Bayesians displayed slightly better predictive power than the corresponding “pruned” half-life Bayesians, when scoring the “full” percent compound left validation set. These validation studies against the percent compound left set were the only cases in which the “pruned” half-life Bayesian models displayed less accuracy than the “full” half-life Bayesian models, according to some of the external statistics.

**Table S-VIII. External validation statistics from evaluating the accuracy of using different sets of run parameters to create MLM half-life Bayesians, by using them to score the “pruned” percent compound left set of compounds.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MLM Stability Bayesian** | **External ROC score** | **External Sensitivity**  **%** | **External Specificity**  **%** | **External**  **Concor-dance**  **%** | **Stability**  **Hit Rate b** | **Unstable**  **True –**  **Filtered c** |
| Full  Half-Life  FCFP\_6 (10 bins) a | 0.794 | 83.5 | 51.5 | 57.9 | 91/303  30% | 225 |
| Pruned  Half-Life FCFP\_6 (10 bins) | 0.785 | 83.5 | 46.2 | 53.7 | 91/326  28% | 202 |
| Full  Half-Life  FCFP\_12 (10 bins) | 0.798 | 85.3 | 43.7 | 52.0 | 93/339  27% | 191 |
| Pruned  Half-Life FCFP\_12 (10 bins) | 0.789 | 89.0 | 33.0 | 44.1 | 97/390  25% | 144 |
| Full  Half-Life  FCFP\_12 20 bins | 0.798 | 85.3 | 42.1 | 50.7 | 93/346  27% | 184 |
| Pruned  Half-Life FCFP\_12 20 bins | 0.788 | 87.2 | 39.1 | 48.7 | 95/361  26% | 171 |

Notes: (a) the run parameters in parentheses correspond to the default settings in the “create Bayesian model” protocol in Pipeline Pilot 9.1 (BIOVIA). (b) The “stability hit rate” is equivalent to the “positive predicted value” and is calculated by dividing the number of true positives by the sum of the number of true positives plus the number of false positives. (c) “Unstable true negatives filtered” corresponds to the number of correctly predicted unstable compounds (there were a total of 437 unstable compounds in the “pruned” percent compound left validation set, see **Figure S1**). In general, each particular type of Bayesian model displayed slightly more accuracy against the “pruned” percent compound left validation set than it produced against the more challenging “full” percent compound left validation set (see **Table S-VII**).