Confidence and Credibility in Cheminformatics: a Bayesian Approach to Uncertainty Quantification in Accumulation Curves

Stefanie Andersen
Jeremy Ash
Lenora Kepler

Introduction

In cheminformatics, a routine task is to determine what compounds are active in a given screen. There are far too many possibilities for researchers to test using wet lab techniques alone, and so many machine learning models have been developed to predict probable active compounds, allowing a researcher to narrow down what they will need to test in the lab. Here we will examine a common scenario in which 500 compounds are assessed for toxicity to a human T-cell line. In this data set, 50 compounds are actually active (toxic).

When a new machine learning model is developed, its marginal improvement over existing methods is tauted as proof that it performs better than previous methods. However, uncertainty is often unaccounted for in these comparisons. It is unclear whether small improvements are actually significant, and if they are, on what data sets these improvements hold.

Previous work by Jeremy Ash and Jaqueline Hughes-Oliver addresses this uncertainty by implementing a program that streamlines the fitting and assessment of these various machine learning models, while accounting for the uncertainty inherent in model prediction. Specifically, they quantify the performance of each model when using each of two descriptor sets, which provide different information about the chemical structure of a compound. The program runs coss-validation simulations to predict from a data set the binary ability of each descriptor-model (DM) combination to correctly find an active compound, and then graph the count of found actives from each when allowed to select increasingly large numbers.

Objectives

Our work aims to use bayesian methods to fit credible intervals to models of each descriptor-set/modeling-routine combination. We contrast these results with the confidence intervals created using frequentist methods, and assess the ability of both to differentiate between models.

Model Specification

Compounds sorted based on predicted probability of activity, such that a compound with index 1 is most likely to be active and that with n is the least.

Descriptor Sets: Burden, Pharmacophore
Models: Tree, SVM, RF, KNN, NNet, PLSLDA

θ_{ij} Probability of activity of prioritized compound j according to D-M combination i

X_{ij} Indicator of a hit for D-M combination i and prioritized compound j

$\sum_{j=1}^{n} \theta_{ij}$ Expected number of hits for a D-M combination with n compounds selected

Prior: $\theta_{ij} \sim Beta(a, b)$

Likelihood: $X_{ij}|\theta_{ij} \sim Bernoulli(\theta_{ij})$

Posterior: $\theta_{ij}|X_{ij} \sim Beta(a + X_{ij}, b + 1 - X_{ij})$

Model Development

Jeffrey's Prior: $\theta_{ij} \sim Beta(.5, .5)$

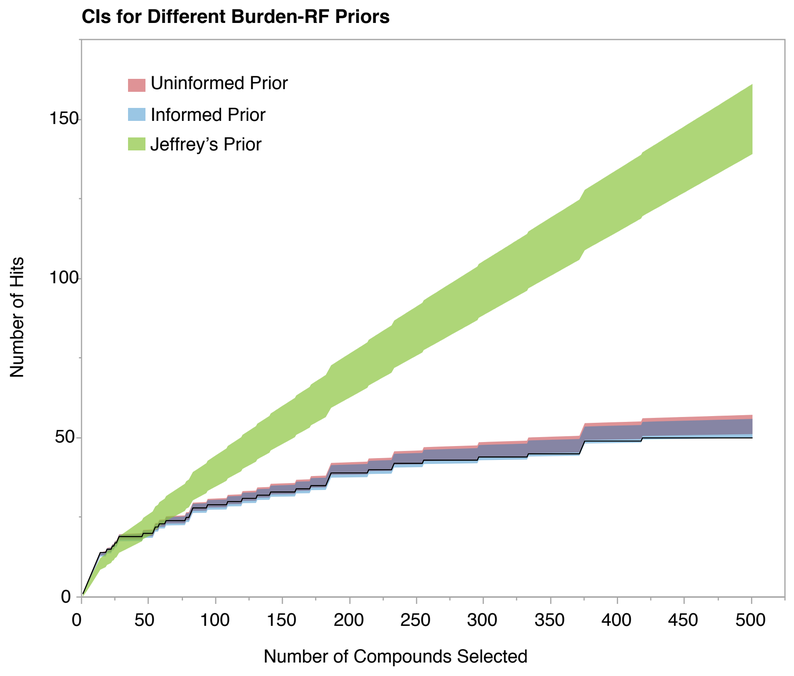
We wanted to first select an uninformative prior, and so we opted to use a Jeffrey's prior of *Beta*(.5, .5). However, because at each point we have a binomial distribution of only n=1, Jeffrey's prior actually ends up being far too informative, and we end up with credible intervals that do not match the predicted activity of the model.

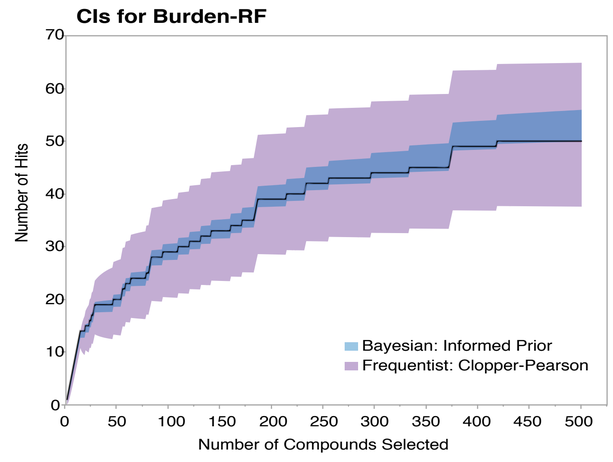
Uninformative Prior: $\theta_{ij} \sim Beta(.01, .01)$

To be truly uninformative, our model required that we use very small values for a and b. Using a Beta distribution with a = 0.01 and b = 0.01 produces much more accurate credible intervals.

Informative Prior: $\theta_{ij} \sim Beta(.01, .01 + .0005i)$

Though the confidence intervals using the uninformative prior appeared to correctly bound test predictions in which less than 100 compounds are selected, after the number of hits reached 50, the true total number of actives, the confidence intervals no longer covered the true predicted number of hits. To get a better fit, the ordering of the compounds must be taken into account. We thus modify b to increase as a function of index. This improves the fit and flattens the interval as compounds selected increases. The interval both covers the observed accumulation curve and stays within the bounds of the band created by frequentist confidence intervals





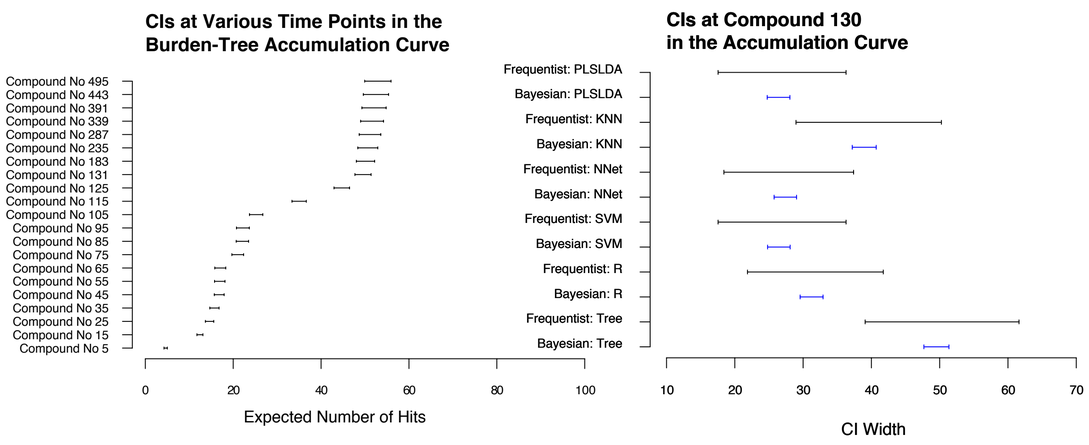
Left: Very small values for a and b are required to produce an uninformative prior; An informed prior taking index into account further improves credible intervals.

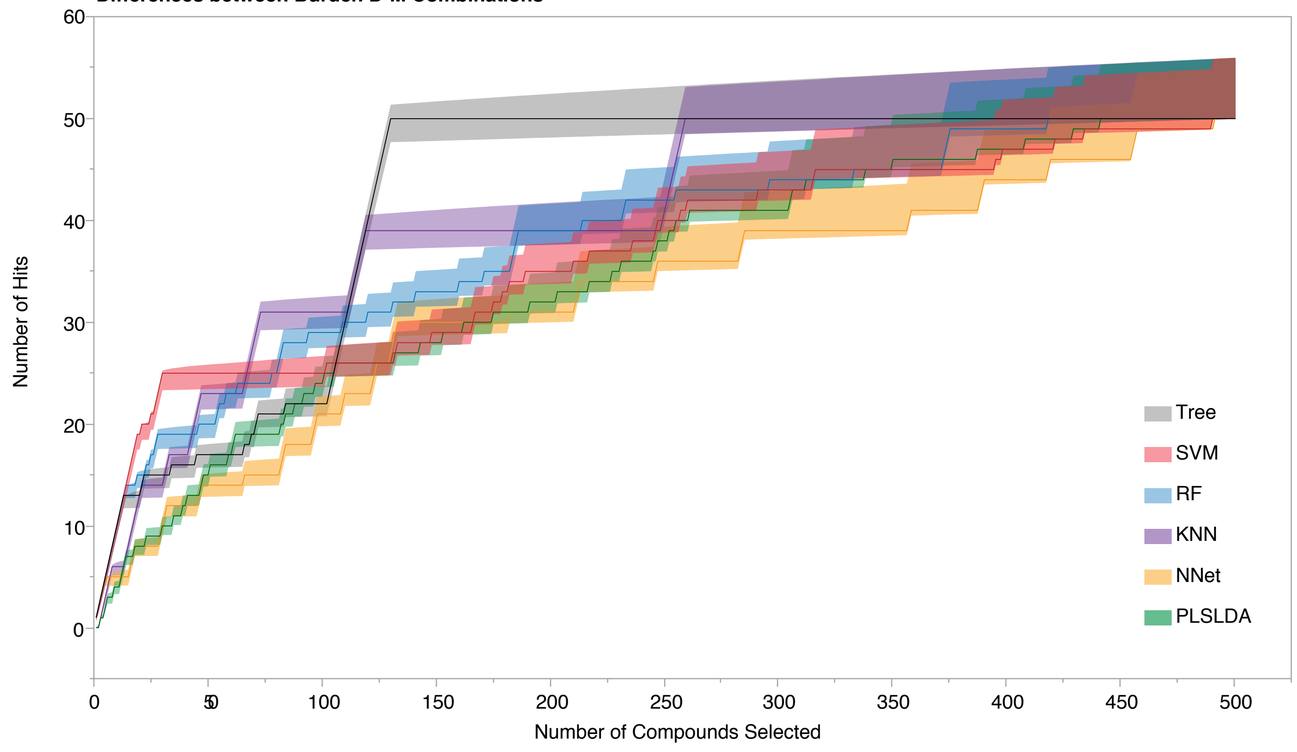
Right: An informed prior keeps the Bayesian credible interval within the 95% frequentist confidence interval.

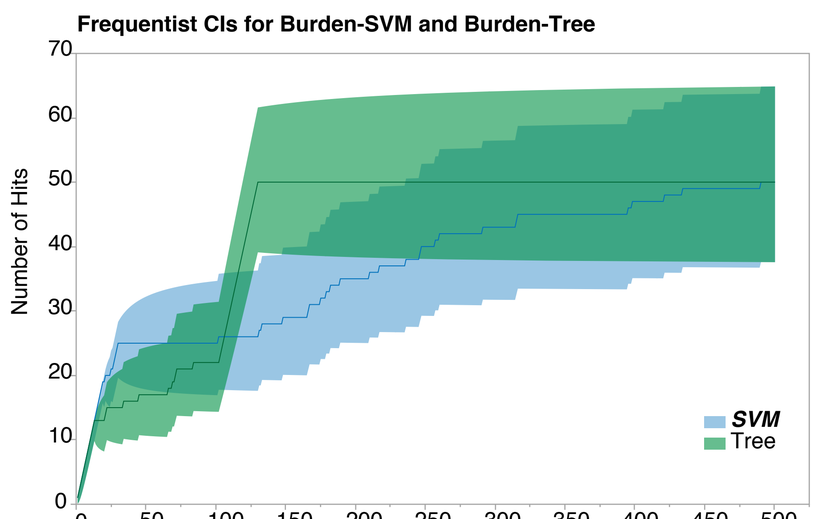
Emperical Interval Results

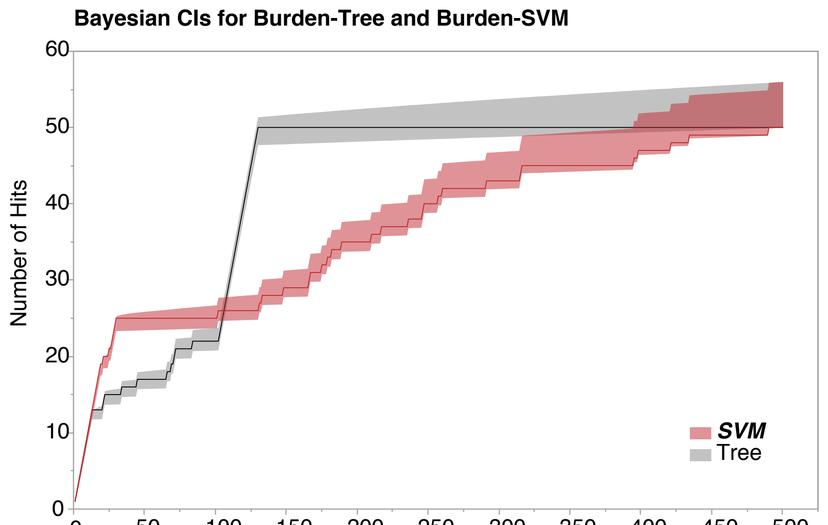
Summary measures for all point wise Frequentist 95% confidence intervals and Bayesian 95% credible intervals (informed prior model)

	Burden Numbers		Pharmacophores	
	Frequentist	Bayes	Frequentist	Bayes
Fraction of Signficant Differences	0.40	0.60	0.43	0.59
Average Interval Width	21.92	4.86	20.94	4.94
Average Distance from .5	0.30	0.31	0.33	0.33









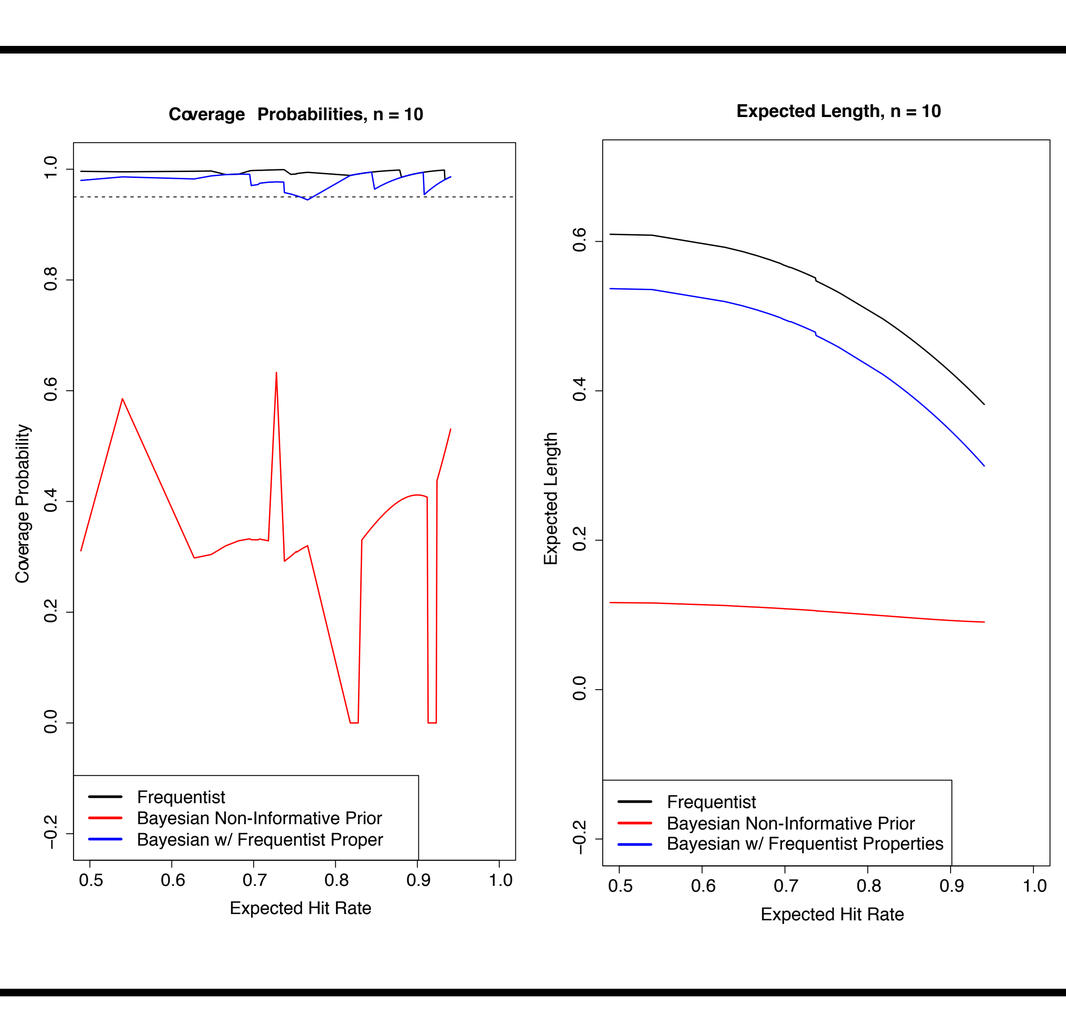
Left: Despite having very different accumulation curves, the confidence intervals created using frequentist models fail to find a significant difference between the two except for a short window at about 125 compounds selected, where Burden-Tree finds significantly more hits (50) than does Burden-SVM (25).

Right: With Bayesian credible intervals, there is clear difference in model prediction. Significantly, the Bayesian credible intervals reveal that while Burden-SVM performs significantly better when selecting between 10 and 100 compounds, Burden-Tree takes over to perform better when between 100 and 315 compounds are tested.

This has direct, real-world impact: if a chemical screen is expensive, a researcher might utilize the Burden-SVM model to maximize the number of actives found in small tests. However, that same model is insufficient if the goal of an experiment is instead to reveal all active compounds, with no limitation on test size.

Simulation Experiment

Simulation experiments were conducted to examine whether our Bayesian model had Frequentist properties that would enable its inferences to be extended to the larger population of accumulation curves. While, as expected, the Bayesian non-informative prior led to much tighter confidence intervals, they had very poor coverage. Though the coverage improved as the number of tests increased, the Frequentist models had much higher coverage probability, with intervals always above 95%



Simulation Setup

100 probabilities were sampled from a Beta distribution, simulating the probability of activity for the jth ranked compound, where j = 1, ...100. 100 distinct probability vectors were generated according to the following Beta distribution, where k = 1, ... 100 and j = 1, ... 100 :

$$\theta_{kj} \overset{\text{iid}}{\sim} Beta(1 + \frac{20(k-1)}{100}, 1)$$

This allowed us to simulate probability vectors that corresponded to a range of population accumulation curves - from ideal curves where the probability of identifying an active is close to one for all j (k = 100) to random curves where the probability of the jth ranked compound is drawn from a *Unif*(0, 1) (k = 1). At n tests, a 95% confidence/credible interval was estimated for the expected hit rate using three methods: Frequentist (Clopper Pearson), Bayesian non-informative prior, and a modified Bayesian model with good Frequentist properties. Since the expected hit rate was known during simulation, we computed coverage probabilities and expected lengths exactly (marginalizing over all possible data sets).

Conclusions

Using Bayesian analysis to calculate credible intervals for predicted model performance significantly decreased interval size such that comparison of Descriptor-Model combinations were meaningful. However, while narrow, the Bayesian credible intervals provided poor coverage. With the prior utilized in this study, it would likely make sense to use the more conservative frequentist confidence intervals,as they provide a more reliable interval estimate.

Future Directions

However, there is ample room to improve upon the Bayesian model. As demonstrated in the simulation experiments, there are priors that can be used to give interval estimates nearly identical to those created with the Frequentist Clopper-Pearson method. While still lacking one of the main advantages of a Bayesian model, decreased expected length, these credible intervals still have the advantage of a more intuitive interpretation and thus may be worthwhile. Additionally, there is still room for improvement in the Bayesian model, for instance by using a random slopes model.