Small Random Forest Models for Effective Chemogenomic Active Learning

Christin Rakers¹, Daniel Reker², J.B. Brown³*

¹ Institute of Transformative bio-Molecules (WPI-ITbM), Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8602, Japan

² Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology (MIT), 500 Main St, Cambridge, MA 02139, USA

³ Life Science Informatics Research Unit, Laboratory for Molecular Biosciences, Kyoto University Graduate School of Medicine, Kyoto 606-8501, Japan

(Received April 21, 2017; Accepted July 11, 2017)

The identification of new compound-protein interactions has long been the fundamental quest in the field of medicinal chemistry. With increasing amounts of biochemical data, advanced machine learning techniques such as active learning have been proven to be beneficial for building high-performance prediction models upon subsets of such complex data. In a recently published paper, chemogenomic active learning had been applied to the interaction spaces of kinases and G protein-coupled receptors featuring over 150,000 compound-protein interactions. Prediction models were actively trained based on random forest classification using 500 decision trees per experiment. In a new direction for chemogenomic active learning, we address the question of how forest size influences model evolution and performance. In addition to the original chemogenomic active learning findings that highly predictive models could be constructed from a small fraction of the available data, we find here that model complexity as viewed by forest size can be reduced to one-fourth or one-fifth of the previously investigated forest size while still maintaining reliable prediction performance. Thus, chemogenomic active learning can yield predictive models with reduced complexity based on only a fraction of the data available for model construction.

Key Words: Virtual screening, chemogenomics, computational chemistry, active learning, drug discovery, random forest

*jbbrown@kuhp.kyoto-u.ac.jp
1. Introduction

In search of new and potent drug candidates, continuous advancements in high-throughput screening technologies brought forth large amounts of experimental data that characterize compound-protein interactions (CPIs). Despite increasing efforts to identify new drugs and increased spending by the pharmaceutical industry whilst producing evermore data, the average number of new drugs brought to market is currently declining [1,2]. Nevertheless, these complex, human-intractable databases of molecular interactions might hold the key to accelerating drug discovery. Data mining approaches such as machine learning (ML) represent valuable computational tools to rationalize given activity spaces and derive structure-activity relationships (SARs) that can be used to predict new desired endpoints (e.g. molecules, targets, or interactions) [3,4]. Traditionally, computational studies that explore SARs either focus on leveraging molecular information about small molecules (ligand-based approaches) [5-9] or protein structures (receptor-based approaches) [10,11]. By combining these two worlds, chemogenomic (or proteochemometric) methods explore interaction spaces based on joint compound-protein descriptors and extrapolate on both target and chemical spaces while extending the model’s applicability domain [12-17].

Chemogenomically-derived prediction models not only allow for "one-sided" prediction of novel compounds or targets, but also for prediction of compound-protein interactions (CPIs) that are absent in the training matrix, and prediction of pairs of ligands and targets both outside of the training data [16,18]. Potential application areas of chemogenomic approaches therefore also include assessment of target selectivities, receptor deorphanising, and drug repurposing. The slow but steady increase in retro- and prospective studies using chemogenomic methodologies hints at its utility and benefit for different applications in medicinal chemistry and chemical biology [14,19-23]. Nevertheless, the sheer data volume, and sparseness and complexity of the compound-protein matrix often necessitate complex chemogenomic machine learning approaches. Advanced approaches, such as (deep layered) neural networks, and computationally sophisticated architectures, including GPU computing, are harnessed to manage and incorporate the evermore increasing amount of available data.

In recent years, active learning (AL) - a concept adapted to the field of medicinal chemistry approximately 15 years ago [24] – has re-attracted interest in the drug discovery community [25,26]. Based on actively updating or feedback-driven model training, AL approaches capitalize on implemented selection strategies that guide iterative sample selection [25,26]. A prominent strategy employed for active model training utilizes an explorative approach that is driven by prediction uncertainty (a “curious” sample selection strategy). Here, learning is enforced by selecting training samples that yield maximum prediction ambiguity under the expectation that such samples ultimately add knowledge to the model [26].

To date, several pro- and retrospective studies report successful applications of AL strategies throughout different fields of research [27-46]. In the area of drug discovery, active learning has been shown to efficiently derive high-performance prediction models based on small subsets of input data [32,34,39,46]. Furthermore, actively trained models not only reached significantly higher hit rates compared to experimental standards which frequently remain below 1 % in cases of unbiased chemical libraries [34,39,40,47-49], but such models also contributed to successful identification of novel bioactive compounds [33,36,42] and cancer rescue mutants of p53 [31]. Overall, AL approaches bear the potential to improve drug discovery processes by increasing hit rates, reducing the amount of time- and cost-intensive experimentation, and accelerate hit-to-lead processes through integration into a feedback-driven experimentation workflow [33,36,42,43,50].

In a recently published study, Reker et al. applied an active learning strategy on family-wide interaction spaces including more than 150,000 interaction data points between 308 biomolecular targets and approximately 100,000 ligands [50]. The authors set out to investigate the applicability and utility of active learning toward reducing the amount of data necessary for constructing highly predictive target family-wide chemogenomic models. Comparing active learning selection strategies in terms of resulting model performances characterized by Matthews correlation coefficients (MCC) [51], calculations showed that a “curious” (explorative) selection strategy based on maximum prediction variance outperformed random and greedy (exploitative) selection [50] (see Methods below for the formula and interpretation of MCC). Further, the curious AL strategy achieved highly predictive performances using less than a quarter of available data points, thereby indicating chemogenomic active learning’s beneficial applicability for drug discovery and chemical biology screening approaches. This further validates the generalized active learning concept as an adaptive data sampling technique to reduce the training data.
Apart from the selection strategy, the machine learning models and their associated parameters will influence learning behavior and predictive performance. Interestingly, these parameters can also directly affect resulting model complexity as well as computational costs for learning and prospective predictions. In this article, the influence of random forest size, i.e. the number of decision trees, on prediction model performance was explored (see Figure 1 for concept and workflow). The interaction spaces of G protein-coupled receptors (GPCRs) and kinases were actively modeled with the curios selection strategy, where the number of trees per experiment was treated as the key variable. Constructed prediction models were assessed regarding model performance (MCC), the evolution of picked samples, and stopping criteria for model training.

2. Materials & Methods

Datasets

Active learning approaches for classification were based on datasets from a previously published study [50] and attributes are summarized in the following. Datasets covered the ligand-target interaction space of human GPCRs and kinases (sources: GPCR-specific GLASS database [52], ChEMBL GPCR SARfari 3 and Kinase SARfari 5 [53]). Activities were given as IC50 values for Kinase SARfari 5 and as Ki values for the two GPCR datasets. The activity threshold for CPIs was set to 100 nM or stronger. The non-CPIs were defined by setting a threshold of 10 µM or weaker for Kinase SARfari 5 and GPCR SARfari 3, and a threshold of 1 µM or weaker for GPCR GLASS. Contradictions and data points between these ranges were discarded. The numbers of records retained and discarded are given in Table 1. Note it is possible for a given compound and protein pair to have multiple supporting records.

From these databases, interaction pairs, i.e. compound-protein interactions (CPIs) and non-interactions (non-CPIs), were extracted (Figure 1, Table 1). Data preparation resulted in three datasets containing 39,706, 47,602, and 69,960 compound-protein pairs (CPIs and non-CPIs) with 48%, 82%, and 71% positive (strong) interactions (CPIs) for Kinase SARfari 5, GPCR SARfari 3, and GPCR GLASS, respectively. The datasets covered the protein space of 98 kinases and 100 (GPCR SARfari 3) or 110 (GPCR GLASS) GPCRs.

Descriptors for active learning modeling were generated by translating compound structures into circular topological extended connectivity fingerprints (ECFP) [54] using a radius of 4 and a bit length of 4096 (OpenEye OEChem library). Vectorizing dipeptide frequencies of amino acid sequences generated target descriptors (of length 400). The input for a compound-protein pair is the concatenation of the descriptors and its (non-)interaction status.

For the pair of GPCR datasets, only two targets had identical FASTA sequences (Supplementary Figure S6). However, most sequences in one database had a highly homologous counterpart in the other database, as validated by clustering protein pairwise sequence similarity values derived from the Local Alignment Kernel (Supplementary Figure S7) [55]. 14195 ligands had identical full InChI string representations in the two datasets (Supplementary Figure S6).

<table>
<thead>
<tr>
<th>Data source</th>
<th>KS5</th>
<th>GS3</th>
<th>GLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioactivity type</td>
<td>IC50</td>
<td>Ki</td>
<td>Ki</td>
</tr>
<tr>
<td>Threshold for interaction</td>
<td>100 nM</td>
<td>100 nM</td>
<td>100 nM</td>
</tr>
<tr>
<td>Threshold for non-interaction</td>
<td>10 µM</td>
<td>10 µM</td>
<td>1 µM</td>
</tr>
<tr>
<td>Interaction records</td>
<td>25118 (52%)</td>
<td>49622 (84%)</td>
<td>130483 (76%)</td>
</tr>
<tr>
<td>Non-interaction records</td>
<td>22881</td>
<td>9690</td>
<td>40951</td>
</tr>
<tr>
<td>Discarded records</td>
<td>32567</td>
<td>40539</td>
<td>57893</td>
</tr>
<tr>
<td>Number of CPIs</td>
<td>19231 (48%)</td>
<td>39166 (82%)</td>
<td>49815 (71%)</td>
</tr>
<tr>
<td>Number of non-CPIs</td>
<td>20475</td>
<td>8436</td>
<td>20145</td>
</tr>
<tr>
<td>Number of dual-class targets</td>
<td>98 (100%)</td>
<td>99 (99%)</td>
<td>82 (75%)</td>
</tr>
</tbody>
</table>
Active learning methodology

Active learning strategies based on random forest classification using the Python library scikit-learn [56] were applied to the interaction space of GPCRs and kinases as detailed in the previous section and identical to the protocol by Reker et al. [50]. The curious selection strategy steers instance picking per iteration towards the most "interesting" or controversial picks based on the largest disagreements in tree predictions (maximum prediction variance; uncertainty-based selection) [50].

The forest size, i.e. number of trees (ntrees), was varied through the scikit-learn constructor call RandomForestClassifier(n_estimators=X). Decision trees were not pruned for depth. The maximum number of features considered at a node split in a tree was the square root of the total number of features available, identical to the previous protocol [50]. Hence, each decision tree was constructed by considering as many as floor(sqrt(4096+400)) = 67 randomly selected features at a node. Each dataset was modeled using 500, 150, 100, 50, 25, 10, 5, 4, 3, 2, and 1 trees as a standard set for comparison. Additional AL models with larger forest sizes were calculated using ntrees of 1000 for GPCR SARfari 5, 2000 and 1000 for GPCR GLASS, and 1000, 300, 250, and 200 for Kinase SARfari 5.

In addition, control experiments using random picking per iteration as a subsampling-based selection strategy were performed, though for random sampling, the forest size parameter was left to the value of 500 as in accordance with Reker et al. [50]. In all experimental setups, 10 executions of modeling and evaluation are performed to assess aggregate performance.

Assessment of model performance and stopping criteria

The performance of actively trained classification models was assessed by calculating the Matthews correlation coefficient, MCC, per iteration. The MCC ranges between -1 (inverse classification) and 1 (perfect classification), and is defined as

\[
\text{MCC} = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN))},
\]

using confusion matrix counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Values above 0.6 signal moderate predictive ability, and values above 0.8 signal strong predictive ability. The MCC was chosen as the performance metric due to its superior reliability regarding imbalanced data sets compared to, for example, the accuracy metric \((TP + TN)/(TP + TN + FP + FN)\). Further, previous work demonstrated the potential deception in result evaluation for chemogenomic modeling that can occur as a result of evaluating only true positive or true negative prediction rates [50]. For comparison, the false positive (FPR), false negative (FNR), true positive (TPR), and true negative rate (TNR) were also calculated over the course of iterations.

In addition to MCC curves, the evolution of picked CPIs and non-CPIs per iteration was monitored, and the ratio of CPI vs. non-CPI picks was calculated in order to track the relationship between forest size, selection ratio, and model performance.

Addressing the question of when to stop model training, a previously-established stopping criterion [50] was employed that allows calculation of the number of iterations and corresponding MCC for a given local speed of learning, i.e. the slope (derivative) of MCC-iteration curve functions, which are retrieved through fitting MCC curves to exponential decay functions. Here, stopping criteria for \(d\text{MCC}/dT\text{er} = 1 = 0.8\) were assessed for all executions of active learning. Results from standardized experiments will be discussed below and are visualized in a dataset-wise fashion in Figures 2 to 4. Additional experiments with deviating numbers of trees are given as supplementary information.

Statistical analysis was performed by assessing distributions of MCC at stopping criteria for pairs of forest sizes. Though we have shown the normality of MCC distribution previously [50] for both curiosity and randomly-picked chemogenomic active learning, we nonetheless execute the statistical analyses by both parametric Welch t-test and non-parametric Wilcoxon rank-sum test procedures, as implemented in SciPy [57].

Hyperbola and sigmoidal curve fitting of prediction performance as a function of forest size and iteration was performed using GraphPad Prism version 7.0 (GraphPad Software, La Jolla California, USA, www.graphpad.com).

Execution times of active learning runs were assessed using the following computational host: Xeon E5-2697v4 18 core CPUx2, 16 threads per execution, ECFP-dipeptide descriptors for GPCR SARfari 3. Datasets were loaded into memory prior to learning, and the input time was subtracted from execution wall time.

3. Results

Rooted in the methodology established in the previous investigation [50], the initial number of trees was set to 500 (control experiments), and active learning with varying numbers of trees was performed. Due to strongly similar model performance for 500-, 300-, 250-, and 200-tree models for the Kinase SARfari 5 dataset (Supplementary Figure S3), active learning of the GPCR interaction spaces (GPCR SARfari 3 and GLASS) was
executed with a standardized set of forest sizes set to 150, 100, 50, 25, 10, 5, 4, 3, 2, and 1. Further experiments with larger numbers of trees are provided in Supplementary Figures S1 to S3.

**Influence of forest size on active learning performance**
Per pair of dataset and forest size, ten executions of model construction and prediction performance assessment were performed. Assessment of model performances was realized by calculating MCC values per iteration for all experiments (Figures 2a, 3a, and 4a). The overall shape of the MCC curves and average performances vary between the three datasets due to inherent dataset variabilities (e.g. number of CPIs and non-CPIs, complexity of the underlying structure-activity-relationship, and the potential for target bias). Overall, for forests of 25 or more trees, the curious selection strategy outperforms random selection and prediction using 500 trees. As can be seen in panel (a) of Figures 2-4, no practical loss of performance was incurred for ntrees of 500, 150 and 100. For the SARfari datasets, MCC values of 0.6 were reached within 1,000 iterations and MCCs of 0.8 within 2,000 and 3,000 iterations for Kinase SARfari 5 and GPCR SARfari 3, respectively (Figures 3a and 4a). The best-performing models for GPCR GLASS (ntrees = 500, 150, 100) reached MCC values of 0.6 within 3000 iterations (Figure 2a). Using the curious active learning strategy, prediction models for all datasets (ntrees = 500, 150, 100) achieved MCC values of 0.6 within 2.1%, 2.5%, and 4.3% of the GPCR SARfari 3, Kinase SARfari 5, and GPCR GLASS datasets, respectively. This fully corroborates the efficiency of
active learning for subset selection on chemogenomic datasets reported previously [50].

Assessing the influence of forest size on model performance, comparison of MCC curves indicates slower speed of learning (more shallow initial slopes) for decreased number of trees, and smaller forests generally converge on lower horizontal asymptotes of curve functions (in the range of 5,000 iterations). For nTrees ≤ 50 an increased number of iterations, i.e. model training, was required to reach reliable MCC values of 0.6 or 0.8. Across all datasets, MCC curves indicate that model training with 25 trees generally necessitates almost twice the amount of data to reach MCC values of 0.6 or 0.8 compared to models trained on 500 trees. For example, an MCC of 0.6 at 2500 iterations is achieved for the GLASS dataset with 500 trees, whereas a forest of 25 trees required approximately 5000 iterations (Figure 2a).

Though it is not difficult to conceive the hypothesis that more trees could boost predictive performance, actively trained models with 1,000 and 2,000 decision trees showed practically identical performance to models of 500 trees, suggesting that the per-dataset prediction

Figure 4. Active learning results for the Kinase SARfari 5 dataset. Panels are analogous to Figure 2. Critically, we see an effect from forest size with respect to model performance evolution in panel (a) despite balanced selection ratios as shown in panel (b).
performance is possibly already saturated at 500 trees and no further model complexity is necessary to fit the data. The same saturation effect can be found for all three datasets tested (Supplementary Figures S1 to S3).

From an alternative viewpoint, MCC increases over different dataset and different iteration thresholds seem to follow a sigmoidal development (Supplementary Figure S4). In addition, model performances were assessed by calculating the classification metrics of TNR (specificity), TPR (sensitivity), FPR, and FNR per training iteration for all three datasets (Supplementary Figure S5). For experiments with 500 trees, models based on the GPCR datasets showed high sensitivity (identification of true positives), while models based on kinases performed better at identifying true negatives (higher specificity or TNR). Our results support the hypothesis that prediction model performances should be evaluated holistically and that focusing on a single metric which does not incorporate the full confusion matrix for performance evaluation, such as the TPR, can be misleading. While MCC values give a balanced interpretation of the corresponding performance, the investigation of specificity and sensitivity can be insightful to understand whether a model over- or underestimates the interaction potential. For example, the GPCR datasets have high ratios of actives, and the calculated TPR of GPCR prediction models tended to over-estimate performance compared to the corresponding FPR and MCC.

To understand the influence of varying the numbers of decision trees, our results showed that forests trained on GPCR data with five or fewer trees tend to be weaker at improving TNR and FPR over training iterations. In contrast, forests with more than 25 trees showed significant improvements in identifying true negatives and reducing the number of false positives during training. Becoming better at identifying true negatives for the GPCR prediction models which were trained on inherently lower ratios of negatives (71% and 82% positives (CPs)) correlated with improved MCC performances. Overall, all of these classification metric curves for all three datasets indicated superior performance with forests of ntree > 25, suggesting that more complex models can improve model performance from different perspectives.
Stopping criteria for active learning with varying forest sizes

In order to compare different learning experiments and to determine the point at which sufficient prediction performance has been achieved or an increase in training rounds (iterations) would not significantly contribute to an increase in performance, a stopping criterion is needed. As previously shown [50], differentiating calculated MCC curve functions and solving them for given slopes ($\frac{dMCC}{dIter}$) of 1 and 0.8 provides such evaluation criterion. Solutions of given derivatives provide corresponding MCCs and iteration numbers at given rates of learning (Panels (c) and (d) in Figures 2-4).

For all three datasets, resulting MCC values decrease with decreasing number of trees. The corresponding stop iteration values decrease, slightly increase, and remain relatively equal for the datasets of GPCR GLASS, GPCR SARfari 3, and Kinase SARfari 5, respectively.

Analyzing the MCC results by the Welch and Wilcoxon tests, we find that the probabilities of equivalent means for each pair of forest sizes is almost always less than 0.05, which leads us to reject the null hypotheses of these tests that the means are equal. Thus, each increasingly large forest size contributes to some level of gain in prediction performance in a statistically reproducible manner.

While statistically significant, the difference between the MCC means of groups of 500 to 100 trees is relatively small, indicating very minor loss in model performance by reducing complexity to 100 trees. Reducing beyond 100 trees, the interval between means increases.

Balance in sample selection during active learning with random forest size

Another aspect influencing prediction model development is the ratio of positives (CPIs) and negatives (non-CPIs) picked per iteration. The sampled CPI and non-CPI picks per iteration were calculated for all experiments using the curious selection strategy, and the ratios of selected CPIs and non-CPIs were visualized in Figures 2b to 4b and Supplementary Figures S1b to S3b. In the preceding study [50], it was validated that the original training dataset biases (71%, 82%, and 48% CPIs for GPCR GLASS, GPCR SARfari 3, and Kinase SARfari 5, respectively) were converged upon when using the random picking strategy as a subsampling method.

The curious selection strategy, on the other hand, maintains balanced (1:1) picking ratios of CPI to non-CPI per iteration regardless of the original training dataset ratios [50]. Here we could show that a reduction in the number of decision trees influenced the selection ratio of sampled positives to negatives per iteration for the two imbalanced GPCR datasets (Figures 2b and 3b). Control experiments (ntrees = 500) and experiments with tree numbers of 150 and 100 resulted in pick ratios of approximately 50%, but further reduction of forest size led to progressive approximation of the underlying input ratios of GPCR GLASS and GPCR SARfari 3. In other words, the amount of selected CPIs per iteration changed towards the inherent dataset bias with decreasing numbers of decision trees. Apparently the curiosity selection function’s ability to sample skewed datasets in a balanced manner is dependent on sufficient model complexity.
Balanced selection during active learning is a major, but not the only, factor in efficient learning of chemogenomic datasets

We showed that forest size (and hence complexity) influences both the overall learning performance as well as the ability of active learning to sample interactions and non-interactions in a balanced manner. Taken together, this suggests that a balanced selection is key for efficient learning of the chemogenomic interaction space. Indeed, for the GPCR datasets, the model performances (MCC values) for different forest sizes (Figures 2a and 3a) seem to correlate with the ratio of selected positives (CPIs) per iteration (Figures 2b and 3b).

However, this trend was not detectable in the Kinase SARfari 5 dataset, which is inherently balanced, and therefore all model complexities, including naïve random subsampling, were selected with picking ratios remaining at about 50% (Figure 4b). Comparing these findings further supports the hypothesis that the superiority in resulting model performance of the curious active learning strategy (in comparison to random) originates from the quality of CPIs/non-CPIs picked for model training (e.g. sample position in the activity landscape) and not solely from the ratio of positives (CPIs) to negatives (non-CPIs).

Smaller random forest models require significantly less computational resources

Results suggest a sufficiently large number of trees in a random forest can ensure an optimal performance both in terms of maximal MCC values as well as a balanced selection of interactions and non-interactions. However, smaller models with reduced number of trees might be advantageous in terms of computational cost.

To test whether the models presented here indeed show a noticeable difference in terms of the required computational time and effort, we timed the active learning campaigns for selected model sizes (ntrees = 500, 150, 50). Not only did we see a markedly and statistically significant decrease in wall time for smaller random forests, the increase in computational time required for smaller forests at later iterations seemed to follow linear instead of exponential increases, at least for the practically relevant number of iterations (Figure 5). It is important to realize that the exact numbers and trends will depend on the execution architecture, e.g. available compute core/thread count versus threads used for computation, dataset size, and forest size.

Smaller models can still achieve good performance by drawing from more data

To better understand the relationship between model complexity and number of iterations necessary to achieve a certain performance, we visualized heatmaps of MCC values against those two parameters (Figure 6). These indicate that indeed a trade-off exists between data and model complexity used, such that simple models require more data while complex models can make sense out of smaller datasets.

In practical computational drug discovery and chemical biology applications, therefore, the tradeoff plots allow a team to identify the proper production-level parameters based on project-specific constraints (e.g., data availability, screening budget, permitted complexity, or compute time).

4. Discussion

The advantage of random forests compared to single decision trees is their superior prediction performance as an ensemble and their stability towards data perturbations. On the other hand, with increasing numbers of trees used for random forest development, computational costs increase. Therefore, the question of how many trees should be used in a random forest approach becomes worth considering. Despite the common procedure to set the number of trees to a default initial guess (generally between 100 and 500 [42,58-61]), a handful of investigatory studies have been reported that address the influence of tree numbers in a forest. One intuitive assumption follows the concept of "the larger the better" [62]. Nevertheless, Breiman's random forest introduction in 2001 suggests the existence of an asymptotic limit for the generalization error of a random forest which impedes model performance improvement by simply adding more trees [63], and we observed such limits for experiments with very large forests (Supplementary Figures).

To date, several non-chemogenomic studies have shown that the number of trees in a random forest could be reduced to a certain degree without losing model performance [64-68], and often rule-of-thumb suggestions are made and accepted within the different computational communities, which allow model training without previous parameter optimization. To probe whether parameter optimization might be an advantageous step instead of accepting rule-of-thumb guidelines, we investigated how chemogenomic active learning performance changed when reducing the number of trees. Our results suggest that model performance follows a sigmoidal development, with little change in performance for very small or very large forests.
A particularly novel finding in this report is our discovery that the number of trees influences the sample picking ratios of CPIs to non-CPIs during curiosity-driven sample selection. The curiosity strategy that was applied to actively train the prediction models generally steers the sample selection towards balanced picking ratios of CPIs to non-CPIs (1:1) for larger forests. As we gradually decreased the forest size, the picking ratios shifted towards inherent CPI to non-CPI ratios of the underlying datasets (i.e. the number of picked CPIs in relation to picked non-CPIs increased).

These findings lead us back to the question of how to set the number of trees prior to model development. Ideally, we would implement a minimized necessary number of trees to allow optimal performance with reduced computational complexity. In a study on random forest parameter sensitivity, it was pointed out that the choice for optimum parameters depends on the underlying dataset and that parameters should therefore be tuned data-set wise [69]. One way to determine parameters such as ntree prior to experimentation is the application of parameter estimators such as the grid search method in scikit-learn which has recently been applied to drug-target interaction prediction [56,70].

Alternatively, based on the concept of trial-and-error Boulesteix et al. suggest increasing the number of trees until convergence on the value of interest (e.g. prediction error) [60]. One could also consider a “parachuting” approach in which the initial guess is set extremely high, and in a systematic reduction method such as a base-2 logarithm, the number of trees is lowered to a value in which a minimal and acceptable impact in prediction performance is incurred. Another solution could be to use the sigmoidal shape of the MCC versus the number of trees (Figure S4) to predict an inflection point from a fitting of the data. Depending on the quality of the fit, this would then be indicative of the region of the number of trees necessary to achieve a certain predictive quality.

Overall, our findings from prediction model development showed that despite reducing the size of the ensemble (forest) used for prediction to 20% or less of our previous study and commonly used value of 500 trees [42,58,60,61], active learning strategies could still lead to reliable performances (MCC values of 0.6 or better) within the first 5,000 iterations. This MCC threshold is equivalent to roughly 7.2%, 12.6%, and 10.5% of data in the GPCR GLASS, GPCR SARfari 3, and Kinase SARfari 3 datasets, respectively. Thus, given that we have shown the feasibility of forest reduction for chemogenomic active learning, a reason for reducing the number of trees in prospective applied investigations will be to push toward increased model interpretability.

Another interesting approach towards forest reduction has been to not only quantitatively reduce the number of trees but doing so by considering the quality of trees and assessing their individual contribution to overall model performance [65,66,71]. Using backward deletion strategies during model training (e.g. deletion of trees with minimum contribution to overall prediction performance), it has been shown that smaller sub-forests can be capable of representing the "whole" random forest [65]. In a similar study on sub-forest development, authors pointed out based on their findings that the incorporation of specific trees may even diminish model performance of a larger random forest [66]. Incorporation of a strategy to change the underlying model as a way to adaptively improve predictive performance and sampling behavior represents a valuable extension for active learning approaches [26,72].

Regarding chemical interpretability, we can consider the implementation of the random forest classifier algorithm used. At each iteration, all descriptor values are presented for the CPIs subsampled, and up to floor(sqrt(4496))≈67 descriptors can be considered at each node split of a tree. We empirically found that few, if any, of the ECFP descriptors had zero variance; the bit fingerprints had a peak of variance distribution close to 0.1. We can therefore expect that the decision tree building algorithm will rapidly find discriminative descriptors, including the potential re-use of a discriminative descriptor at multiple nodes in a tree. Since the maximum number of descriptors to be included in a tree was not bounded in this study, small forests could still randomly sample and identify multiple discriminative chemical fingerprints. Therefore, we do not anticipate a change in chemical interpretability.

As another direction to further develop the reported active learning methodology, the algorithm could be expanded to address multi-class predictions. Consideration of the CPI data that were discarded in this study after applying to the definition of a separating threshold for active and inactive molecules would represent a suitable starting point for three-class predictions of low-, moderate- and high-affinity compound-protein interactions. It should be noted, however, that many classification challenges in proteochemometric modelling discard intermediate affinities, not only given their subjective nature but also because of potential differences in their class membership given by variance in experimental measurements.

Intermediate class modeling also induces an additional experimental design step to weight evaluation of prediction of instances close to the bioactivity thresholds used for creation of three classes. For example, if we consider the thresholds 100 nM and 10 μM, then CPIs in the intermediate class with bioactivities of 101 nM or 9
μM should not be penalized as strongly for misclassification, because of the subjectivity induced as noted above.

5. Conclusions

Chemogenomic active learning was applied to the compound-protein interaction spaces of kinases and GPCRs using random forests of reduced size, and resulted in robust prediction models and corresponding performances within 12.6% of the data (the first 5,000 iterations). Assessment of the influence of the number of decision trees on predictive performance indicates that forests can be safely cut down to 100 (20%) or 50 (10%) trees depending on the dataset. Further, model performances of experiments with varying numbers of trees indicate the existence of a threshold that cannot be overcome by increasing the quantity of trees. The way in which the pharmacological network was explored, meaning the number of CPIs and non-CPIs incorporated for model development, was shown to dynamically shift toward a bias in picking CPIs as the size of the forest was reduced. Prospective applications of chemogenomic active learning can benefit from understanding the behaviors uncovered in this work.

Acknowledgements

D.R. is grateful for support from the Swiss National Science Foundation (P2EZP3_168827 to D Reker). C.R. thanks Professor Stephan Irle and Assistant Professor Yuh Hijikata for supporting this study. The ITbM is supported by the World Premier International Research Initiative, Ministry of Education, Culture, Sports, Science and Technology of Japan. Support from the Japan Society for the Promotion of Science was used for computational resources under a Grant-in-Aid for Young Scientists (B) 25870336 and a Grant-in-Aid for Scientific Research (S) JP16H06306 (to J.B. Brown). We thank OpenEye Scientific Software (Santa Fe, NM) for providing an academic license and use of the OEChem chemical processing library.

References and Notes

Statistical Applications in Genetics and Molecular Biology, 10, 32 (2011).


Huang, B. F.; Boutros, P. C., BMC Bioinformatics, 17, 331 (2016).


Figure S 1. Active learning results for the GPCR GLASS dataset with tree numbers ≥ 500. a) Evolution of prediction model performances as MCC per iteration. No gain in performance is achieved for large forests. b) Ratio of counts of CPI to non-CPI samples picked per iteration during prediction model development with varying number of trees. Subfigure b) Relationship between counts of picked CPIs to non-CPIs during model development for experiments with ≥ 500 trees. Abbreviations: CPIs = compound-protein interactions; non-CPIs = compound-protein non-interactions; MCC = Matthews correlation coefficient.

Figure S 2. Active learning results for the GPCR SARfari 3 dataset with tree numbers ≥ 500. Panels are analogous to Figure S1.
Figure S 3. Active learning results for the Kinase SARfari 5 dataset with tree numbers $\geq$ 200. Panels are analogous to Figure S1. A limit on performance is reached as early as 200 trees.

Figure S 4. Active learning performance as a function of random forest complexity. MCC curves were calculated based on hyperbola (upper panel) and sigmoidal fit (lower panel).
Figure S 5. Evaluation of alternative predictive metrics. Evolution of prediction model performances as FNR, FPR, TNR, and TPR (from upper to lower panels) per iteration (10 executions per forest size) for all three datasets (columns). We observe that the GPCR datasets with higher ratios of actives have over-estimated True Positive Rates and could potentially mislead model performance interpretation. Performance curves are indexed according to the color key in the lower left corner. Abbreviations: FNR = false negative rate; FPR = false positive rate; TNR = true negative rate; TPR = true positive rate.
Figure S 6. Venn diagrams of identity calculations for the ChEMBL GPCR SARfari 3 (green) and GPCR GLASS datasets (red). a) Overlap of compounds based on full InChI string comparison. b) Overlap of molecular targets (GPCRs) based on FASTA sequences.
Figure S 7. Clustered pairwise comparison of FASTA sequences of targets in the GPCR datasets (ChEMBL GPCR SARfari 3 and GLASS). Scores represent normalized similarities of protein pairs as calculated by the Local Alignment Kernel. Deep colored clusters along the diagonal demonstrate cross-dataset homolog pairs.