JOURNAL OF CHEMICAL INFORMATION AND MODELING

Subscriber access provided by Universiteit Leiden / LUMC

Article

Data-driven Derivation of an "Informer Compound Set" for Improved Selection of Active Compounds in High-Throughput Screening

Shardul Paricharak, Adriaan P. IJzerman, Jeremy L. Jenkins, Andreas Bender, and Florian Nigsch

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.6b00244 • Publication Date (Web): 03 Aug 2016

Downloaded from http://pubs.acs.org on August 4, 2016

ACS Publications

Journal of Chemical Information and Modeling is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Data-driven Derivation of an “Informer Compound Set” for Improved Selection of Active Compounds in High-Throughput Screening

Shardul Paricharak,[1,2,3] Adriaan P. IJzerman,[2] Jeremy L. Jenkins,[4] Andreas Bender,[1§] and Florian Nigsch[3§]

[1]Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, Cambridge, United Kingdom

[2]Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands

[3]Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4056 Basel, Switzerland

[4]Developmental & Molecular Pathways, Novartis Institutes for BioMedical Research, Cambridge, MA 02139, United States of America

[§]Corresponding authors. E-mail: ab454@cam.ac.uk (AB) and florian.nigsch@novartis.com (FN)

ABSTRACT

Despite the usefulness of high-throughput screening in drug discovery, for some systems, low assay throughput or high screening cost can prohibit the screening of large numbers of compounds. In such cases, iterative cycles of screening involving active learning (AL) are employed, creating the need for smaller "informer sets" that can be routinely screened to build predictive models for selecting compounds from the screening collection for follow-up screens. Here, we present a data-driven derivation of an informer compound set with improved predictivity of active compounds in HTS, and validate its benefit over randomly selected training sets on 46 PubChem assays comprising at least 300,000 compounds and covering a wide range of assay biology. The informer compound set showed improvement in BEDROC(α=100), PRAUC and ROCAUC values averaged over all assays of 0.024, 0.014 and 0.016, respectively, compared to randomly selected training sets, all with paired *t*–test p-values < 10^{-15}. A per-assay assessment showed that the BEDROC(α=100), which is of particular relevance for early retrieval of actives, improved for 38 out of 46 assays, increasing the success rate of smaller follow-up screens. Overall, we showed that an informer set derived from historical HTS activity data can be employed for routine small-scale exploratory screening in an assay-agnostic fashion. This approach led to a consistent improvement in hit rates in follow up screens without compromising on scaffold retrieval. The informer set is adjustable in size depending on the number of compounds one intends to screen, as performance gains are realized for sets with more than 3,000 compounds, and this set is therefore applicable to a variety of situations. Finally, our results indicate that random sampling may not adequately cover descriptor space, drawing attention to the importance of the composition of the training set for predicting actives.

INTRODUCTION

Over the past three decades, high-throughput screening (HTS) has become a well-established method used during early drug discovery.[1–7] However, low assay throughput or high screening cost can at times prohibit the screening of large numbers of compounds.[8,9] Given this drawback, iterative cycles of design-screen-refine involving active learning (AL) strategies can be used when only a small number of compounds can or should be screened.[10–12] This, in combination with recent advances in machine learning, has recently prompted efforts to improve bioactivity modeling in order to identify active compounds *in silico*, with the aim of increasing the hit rates in compound screens.[11]

For this purpose, a high-throughput screening fingerprint (HTS-FP) was developed by Petrone *et al.*[13] and later by Dančik *et al.*,[14] which profiles compounds according to their bioactivity across a range of HTS assays. This work was based on the idea that such fingerprints are predictive of compound affinity on targets *not* covered in the fingerprint and showed the value of HTS-FP for virtual screening and biodiverse selection of actives. This concept has previously been explored computationally on smaller datasets,[15–18] but without large-scale experimental validation. More recently, Riniker *et al.*[19] benchmarked the predictive performance of chemical fingerprints and HTS-FP in conjunction with a variety of classification methods across a large number of assays performed in Novartis and those in the public domain (available in PubChem).[20] It was found that random forest (RF) methods with HTS-FP often outperformed machine learning methods developed on chemical descriptors.[19] On a related note, Maciejewski *et al.*[21] explored an experimental design strategy where AL was used to enhance the chemical diversity of large training sets comprising over 50,000 compounds, leading to improvement in model performance. While the mentioned studies addressed the dependence of the model on descriptor and classification method used, a comprehensive assessment of how the composition of the initially screened compound set (training set) affects model performance and early retrieval of actives from the remaining screening collection was not performed.

The effectiveness of HTS screening sets in identifying actives has been widely discussed.[22] Given the possible existence of over 10^{63} drug-like molecules,[7] it is remarkable that HTS campaigns comprising “only” 10^{6} compounds succeed in finding hits at all.[22–24] A plausible explanation for this is that screening libraries are not random, but rather biased towards biogenic compounds, likely to interact with the druggable proteome. This claim has been reinforced by studies showing the chemical similarity between metabolite space, natural product space and bioactive space.[25–27] A comprehensive analysis by Klekota *et al.*[28] showed that certain “privileged” chemical substructures, such as benzodiazepines,[29] enrich for bioactivity, creating further avenues for modeling the likelihood of compounds being bioactive in *any* therapeutically relevant setting (hereafter referred to as joint bioactivity modeling), rather than target- or phenotype-specific bioactivity modeling (also shown by Gillet *et al.*).[30]

In this study, we harnessed bioactivity information from a large number of PubChem[20] HTS assays to derive an assay-agnostic “informer compound set” that, once screened, predicts bioactivity better than randomly selected sets for almost all HTS assays, improving the efficiency of subsequent screens. We used AL to iteratively derive this set. Due to the difficulty in implementing AL under extreme class imbalance[31] as is the case for all HTS assays analyzed in this study, activities from multiple assays were combined to derive binary labels representing assay-agnostic bioactivity for each compound. This was based on the idea of joint bioactivity modeling[28,30] and led to a class-balanced dataset suitable for AL. HTS-FPs were used as descriptors, as they showed improved performance over chemical fingerprints.[19] Moreover, this informer set was constructed with the aim to facilitate routine screens, as pre-composed sets are easier to screen routinely from an infrastructure point of view.

Related studies by Young *et al.*[32] and Taylor[33] describe screening strategies aimed at increasing the chances of finding active compounds by predictive modeling using extreme value theory (validated on ~75k data points in a single cell-based assay) and intelligent sampling methods (validated on 2k data points), respectively. However, our study differs considerably, as we validate our method on over 10,000,000 HTS data points across a wide range of assay biology, and use descriptors based on a large amount of bioactivity data, hereby significantly increasing predictive power.

METHODS

HTS data

The public HTS data used by Riniker *et al.*[19] was used in this study (see Tables S1 and S2 of this reference for the list of assays used). HTS data from the NIH molecular libraries program (MLP) comprising at least 300,000 compounds per assay, and submitted by the NCGC, the Scripps Research Institute Molecular Screening Center, or the Burnham Center for Chemical Genomics were extracted from PubChem.[20] This resulted in a total of 141 cell-based and target-based assays (mainly using fluorescence readout technologies), covering a wide range of assay biology (kinases, proteases, ion channels, GPCRs and other target classes). Assay-specific z-scores were calculated for all compounds tested based on the activity measurement used to define the PubChem activity outcome. The set of assays was subsequently split into 2 groups: 95 “group 1 assays” (comprising over 338,000 compounds) and 46 “group 2 assays” (comprising 300,000–338,000 tested compounds, depending on operational turnover of the compound collection at the screening centers). Group 1 assays (referred to as “historical assays” by Riniker *et al.*)[19] were used exclusively for the construction of HTS-FP,[13] a fingerprint used as a descriptor for machine learning, profiling the activity of a compound across HTS assays based on z-scores (float version).[13] Group 2 assays (referred to as “test assays” by Riniker *et al.*)[19] were used for deriving labels and for model training and testing. This distinction between assay groups ensured that there was no overlap in targets between the two groups.[19]

HTS-FP

For each compound, an HTS-FP was computed, in which each element corresponds to the z-score (based on activity) of the compound in one of the group 1 assays. Missing z-scores (15% of all data points; not every compound is tested in each assay) were assumed to be 0 (the mean of z-scores), as implemented earlier by Riniker *et al.*[19]

Workflow

In this study we tested the performance of bioactivity models developed on an informer set derived with AL. As this set was iteratively augmented, the informer set is available at multiple sizes from 1,000 to the maximum size of the AL set (when AL is terminated). First, we evaluated the performance for predicting bioactivity independent of tested assay (Figure 1, joint bioactivity modeling).

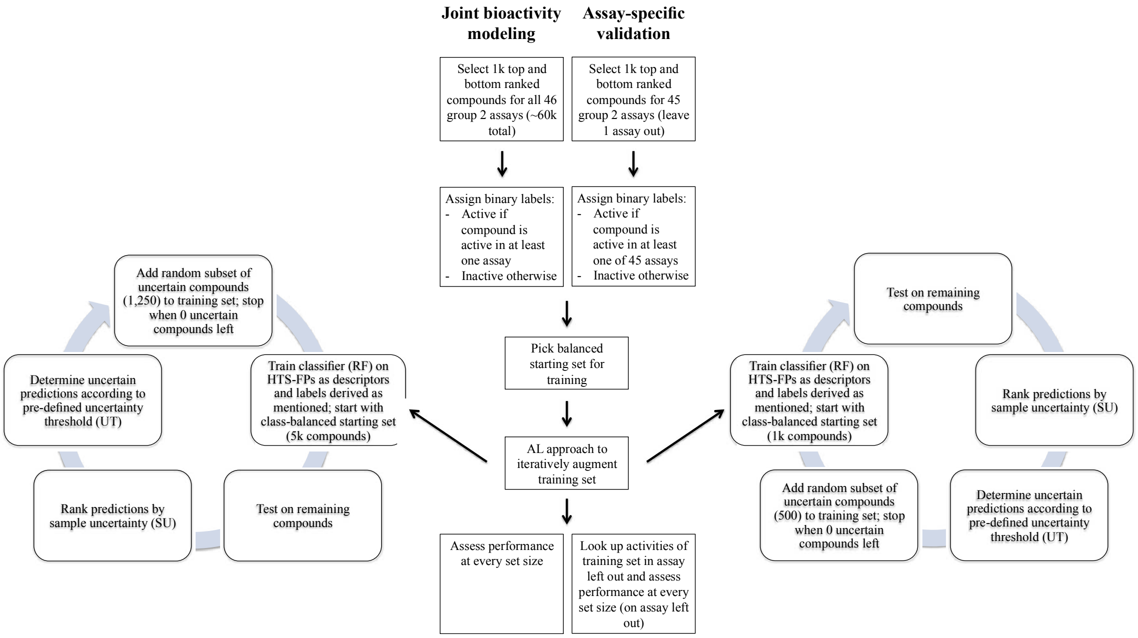

Joint bioactivity modeling
Assay-specific validation
Select 1k top and bottom ranked compounds for all 46 group 2 assays (~60k total)
Select 1k top and bottom ranked compounds for 45 group 2 assays (leave 1 assay out)
Assign binary labels: - Active if compound is active in at least one assay - Inactive otherwise
Assign binary labels: - Active if compound is active in at least one of 45 assays - Inactive otherwise
Pick balanced starting set for training
AL approach to iteratively augment training set
Assess performance at every set size
Look up activities of training set in assay left out and assess performance at every set size (on assay left out)
Add random subset of uncertain compounds (1,250) to training set; stop when 0 uncertain compounds left
Train classifier (RF) on HTS-FPs as descriptors and labels derived as mentioned; start with class-balanced starting set (5k compounds)
Test on remaining compounds
Rank predictions by sample uncertainty (SU)
Determine uncertain predictions according to pre-defined uncertainty threshold (UT)
Test on remaining compounds
Rank predictions by sample uncertainty (SU)
Determine uncertain predictions according to pre-defined uncertainty threshold (UT)
Add random subset of uncertain compounds (500) to training set; stop when 0 uncertain compounds left
Train classifier (RF) on HTS-FPs as descriptors and labels derived as mentioned; start with class-balanced starting set (1k compounds)

Figure 1. Overview of workflow. In this study, two analyses were performed. Firstly (left), a joint bioactivity model was developed on the 1,000 top and bottom ranked (based on z-scores) compounds. An AL approach was used to iteratively augment the training set, for which model performance (ROCAUC) was assessed at every set size. The second analysis (right) involved an assay-specific validation, where a joint bioactivity model was developed on all assays except the assay left out of training. The training set was iteratively augmented with uncertain samples using AL, and at every set size, activities of these compounds were looked up in the assay left out.

Subsequently, model performance (ROCAUC, PRAUC, BEDROC) for the training set was assessed on the assay left out, rather than on the joint activities dataset.

Here, activities from group 2 assays were combined to derive binary labels representing assay-agnostic bioactivity for each compound in order to construct a class-balanced dataset suitable for AL. Improved model performance at this step was considered a prerequisite for the more challenging task of predicting actives for individual assays. An assay-specific validation was performed to address the latter task: the informer set was derived from activity data from 45 group 2 assays and predictivity was assessed on the one assay remaining (Figure 1, assay-specific validation). This was repeated 46 times, effectively leaving each group 2 assay out once.

Joint bioactivity modeling

The 1,000 least and most active compounds (based on z-scores) were selected from each group 2 assay, resulting in a total of 58,768 compounds. A skewed distribution of the number of assays these compounds were active in was observed, with 45%, 33%, 12% and 10% of compounds active in 0, 1, 2 and more than 2 assays, respectively (Supplementary Figure S1). Each compound was labeled as “active” if it was active in *any* of the group 2 assays (as defined by the PubChem activity outcome) or “inactive” otherwise, resulting in a total of 32,171 actives and 26,597 inactives. This labeling was based on the concept of considering activities independent of the assay they were tested in (joint bioactivity). An RF model (scikit-learn)[34–36] was developed on a randomly selected class-balanced training set of 5,000 compounds (to initiate training), and the performance of the model was assessed on the remaining compounds. Using AL, this training set was iteratively augmented with up to 1,250 uncertain samples at each iteration, with the aim to improve model performance on the remaining compounds (see “Active learning” section for more details). The model for this training set, the informer set, was benchmarked against a model developed on a randomly selected set at each set size using the area under the receiver operating characteristic curve (ROCAUC).

Assay-specific validation

Here, the informer set was derived from activity data from 45 group 2 assays, and a model was trained on group 1 assay HTS-FPs and labels derived from the one assay left out. The performance of the model was assessed on the compounds in the assay left out minus those present in the informer set. The starting set for training initiation was a class-balanced set of 1,000 compounds comprising 500 actives and 500 inactives, both selected randomly from the compounds available in the assay left out. This set was iteratively augmented by up to 500 compounds using AL (see “Active learning” section for more details). The size of the training and augmentation set was kept smaller here than for the joint bioactivity modeling due to observed improvement in performance at the earlier stages of the algorithm. Performance on the assay left out was assessed at each set size using the ROCAUC, the area under the precision-recall curve (PRAUC),[37] Boltzmann-enhanced discrimination of ROC (BEDROC) (α=100),[38,39] and the retrieval of Murcko scaffolds[40] belonging to the active compounds. The BEDROC(α=100) was used due to its relevance in early retrieval of actives in imbalanced datasets and the PRAUC was used because it captures the effect of the large number of inactive compounds on the model’s performance.[37] Both these metrics were therefore considered more relevant than the ROCAUC for the assay-specific validation (by contrast, for the joint bioactivity modeling the ROCAUC was considered an adequate metric due to class balance).

The model was benchmarked against models developed on a randomly selected set and a set comprising compounds with the highest median z-scores across the 45 assays left in (the frequent hitter set). The randomly selected set was sampled across Murcko scaffolds:[40] Murcko scaffolds[40] were randomly selected, followed by the selection of one compound per scaffold. This was performed to avoid undersampling low-density areas of chemical space. The comparison with the frequent hitter set was included to ensure that the performance gain for the informer set was better than when simply more actives from other assays (including more frequent hitters) were trained on.

Machine learning

The RF parameters used were: 100 trees (maximum depth = 10), minimum samples to split = 4, and minimum samples for a leaf = 4, random state = 12345.

Active learning (AL)

The AL approach consisted of three iterative steps: (1) training of an RF model, (2) model testing on the remaining compounds and (3) augmenting the training set with a randomly selected subset of uncertain labeled samples (1,250 and 500 compounds for the joint bioactivity modeling and assay-specific validation, respectively); when the number of uncertain samples was smaller than the size of the subset, all uncertain samples were selected. The AL algorithm was terminated when the number of uncertain samples was zero. Sample uncertainty (*SU*) of a given compound *c* was defined as the absolute probability difference in active versus inactive class predictions:

$$SU_c = \left| p_c^{active} - p_c^{inactive} \right| \qquad \text{Equation 1}$$

with SU_c in the range of 0–1 where 0 and 1 represent the most uncertainty and complete certainty in prediction, respectively. Only samples with an *SU* value smaller than the uncertainty threshold

(*UT*) were considered uncertain. We investigated the effect of varying the *UT* from 0.5 (least stringent) to 0.01 (most stringent) for the joint bioactivity modeling, and used a *UT* of 0.1 for the assay-specific validation. The presence of uncertain samples suggests undersampling of bioactivity space. Including these samples could improve model performance over random sampling.[10]

Software used

The workflow comprised Python scripts for data analysis, using scikit-learn[36] for machine learning and RDKit[41] for scaffold derivation. Tableau[42] was used for data exploration and R[43] was used for the visualization of results.

RESULTS AND DISCUSSION

The development of an informer set for the prediction of joint bioactivity is presented first (see Figure 1 – left). Prediction of joint bioactivity allowed the identification of compounds more likely to be bioactive regardless of the assay used. This was followed by a performance assessment of the informer set on individual assays (assay-specific validation; see Figure 1 – right), and an analysis of scaffold retrieval and set composition. The assay-specific validation was performed in order to determine whether the informer set is more useful than a randomly selected set in predicting actives for novel assays one might perform.

Joint bioactivity modeling

The gap in ROCAUC between models developed on the AL sets and on randomly selected sets consistently widens from set sizes of ~5,000 onwards (see Figure 2 – top).

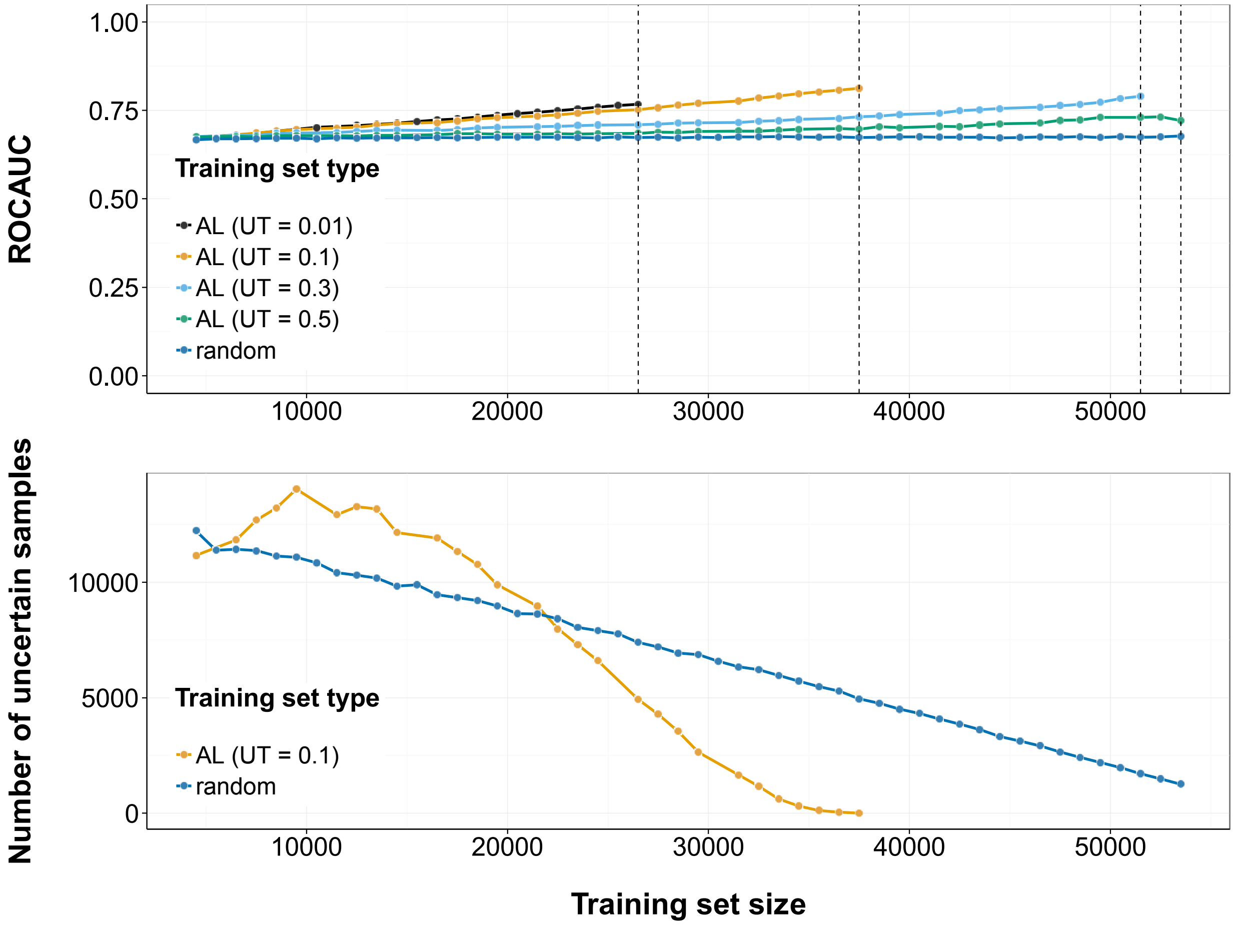


Figure 2. Comparison of model performance for the AL and randomly selected training sets. The ROCAUC (top) is shown for the models trained on AL and randomly selected sets. Performance across all set sizes is consistently better for all AL sets than it is for the randomly selected set. At a set size of 38,000 an average gain in performance of 0.08 is observed. In addition, lower *UT* values led to better performance than higher *UT* values. A *UT* value of 0.1 was chosen for the assay-specific validation on the basis of a trade-off between improvement in performance and maximum training set size. For the AL set (*UT* = 0.1), the number of uncertain reaches zero faster compared to the randomly selected set (bottom), indicating more efficient sampling of bioactivity space.

At a set size of 38,000 an average gain in performance of 0.08 is observed for the AL sets (average ROCAUC of 0.75 compared to 0.67 for randomly selected sets). Stringent *UT* values led to sets with a greater gain in performance at the cost of maximum set size, as fewer samples are classified as uncertain, and the number of uncertain samples reduces to zero earlier in the AL process. For set sizes between 10,000 and 20,000, the number of uncertain samples is larger for the AL (*UT* = 0.1) set than for the randomly selected set. However, for set sizes larger than 22,000, the number of uncertain samples declines faster for the AL (*UT* = 0.1) set than for the randomly selected set (Figure 2 – bottom), and reduces to zero earlier. For example, almost all uncertain samples were exhausted for a set size of approximately 35,000 using AL, whereas the random set did not exhaust the uncertain samples even at set sizes upwards of 50,000. In conjunction with an observed performance gain across all set sizes for the AL sets, this indicates the benefit of AL in sampling relevant bioactivity space more efficiently, hereby improving the identification of compounds bioactive in one or more group 2 assays. For further analysis, we chose a *UT* value of 0.1 on the basis of a trade-off between gain in performance and maximum training set size.

Predictive performance of informer set on individual assays

In an attempt to translate performance gain in predicting joint bioactivity (see previous section) to performance gain in individual large-scale assays, we performed an assay-specific validation for all group 2 assays. Improved predictive performance in this setting would corroborate the usefulness of an informer set, as no prior information about the assay left out would be required for its construction.

The BEDROC(α=100),[38,39] PRAUC and ROCAUC were calculated for an RF classifier trained on the informer set (AL), a randomly selected set, and the frequent hitter set. These values were averaged over all 46 assay-specific validation experiments and were binned by set size (see Figure 3).

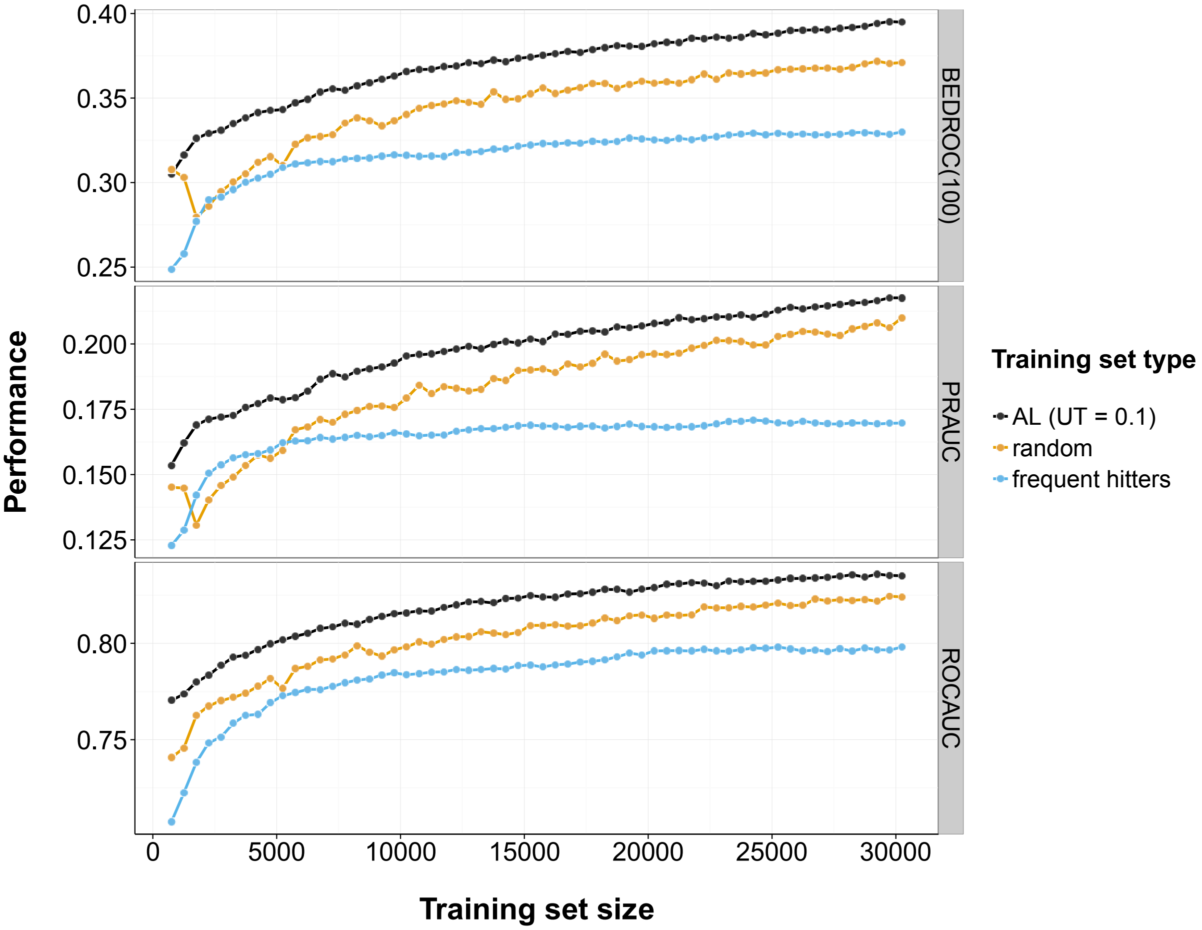


Figure 3. Comparison of model performance for the AL (*UT* = 0.1), random and frequent hitter training sets (assay-specific validation). The BEDROC(α=100)[38,39] (top), PRAUC (middle) and ROCAUC (bottom) binned by set size are shown for all three training sets (bin width=500). The assay-averaged performance for the AL set (all metrics) is consistently better than that for the randomly selected set. For the frequent hitter set, performance is consistently worse than both the AL set and the randomly selected set for training sets larger than 5,000 compounds. These results indicate that models trained on the AL set consistently retrieve more actives compared to models trained on the other sets.

The frequent hitter set was used as a benchmark, to ensure that the performance gain of the AL set was better than when simply more actives from other assays (including more frequent hitters) were trained on.

Overall, the performance for the AL set was enhanced compared to the randomly selected set, with an average increase of 0.024, 0.014 and 0.016 in average BEDROC, PRAUC and ROCAUC, respectively (all with paired *t*–test p-values < 10^{-15}). The apparent low values of the average BEDROC (0.25-0.40) can be explained by the Boltzmann enhancement, as early retrieval of actives is strongly preferred. Low values of the average PRAUC metric (0.10-0.25) can be explained by the extreme class imbalance: a random classifier would achieve a PRAUC of ~0.007 given the average fraction of actives is only ~0.7%.

For the frequent hitter set, performance is consistently worse for set sizes larger than 5,000, indicating that simply including more actives from other assays does not account for the performance gain observed for the informer set. This finding is in line with the results of the "weak reinforcement strategy" as described in the study by Maciejewski *et al.*[21] Here, training sets with a large number of actives similar in descriptor space (including frequent hitters[44,45] in our study, as the descriptor space is based on bioactivity profiles) were found to be poor at identifying the remaining small number of actives in the test set due to insufficient coverage of descriptor space. By contrast, training sets containing compounds outside the applicability domain, corresponding to uncertain samples in this study, were much better at identifying the remaining actives in the screening collection.

Next, the average improvement in performance over all set sizes of the informer set was calculated separately for each assay (see Figure 4).

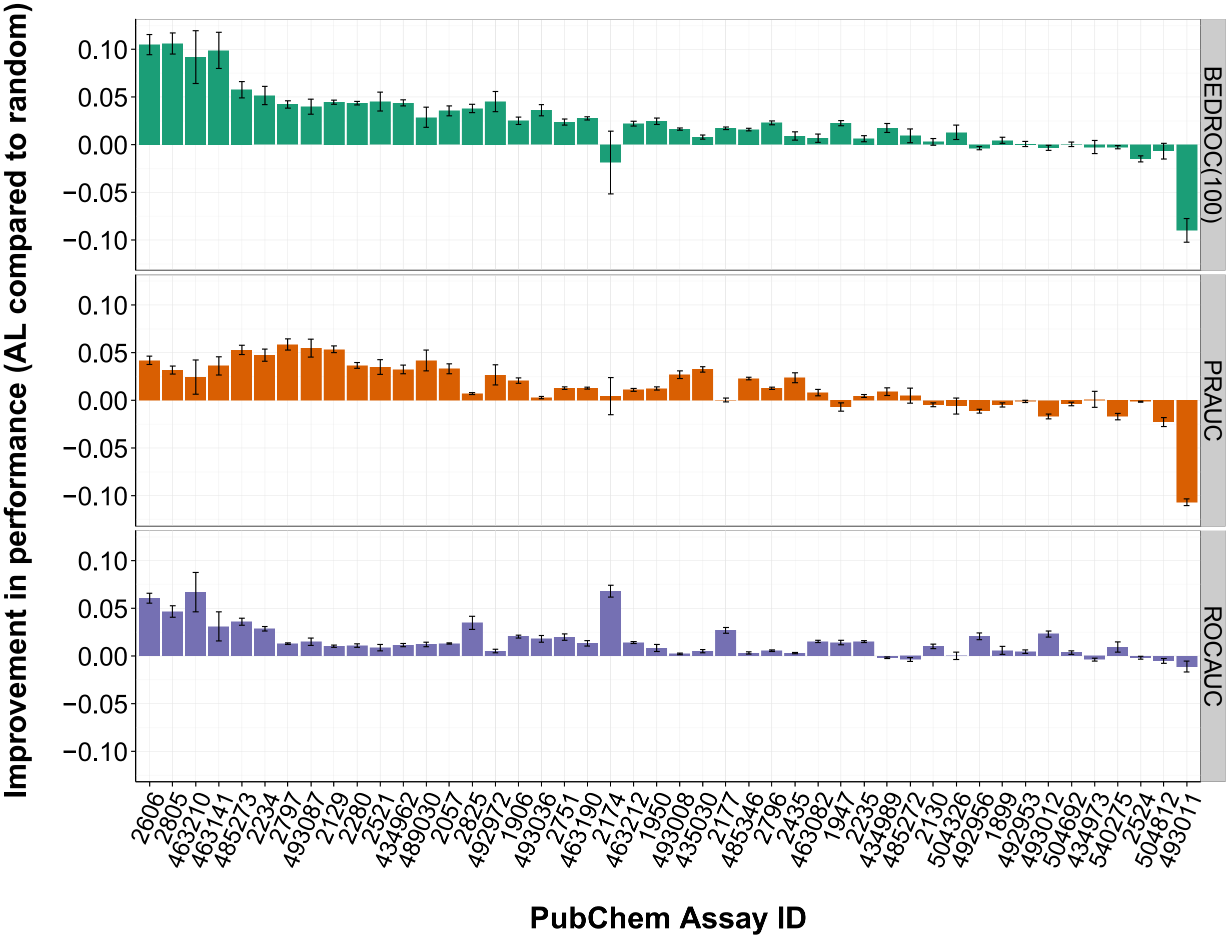


Figure 4. Improvement in model performance for the AL (*UT* = 0.1) set compared to the randomly selected set for separate assays. The average difference in BEDROC(α=100)[38,39] (top), PRAUC (middle) and ROCAUC (bottom) between the AL set and the randomly selected set is shown for separate assays. Error bars represent standard error of the mean. For 30 out of 46 assays all three metrics improved, whereas the BEDROC(α=100), which is of most relevance for early retrieval of actives,[38,39] improved for 38 out of 46 assays. In practice, the results indicate that if a subsequent screen were performed for each assay, more actives would be retrieved for 38 assays, compared to when random training sets would be used.

For 30 out of 46 assays, all three metrics improved by average 0.03 on average, whereas the BEDROC, which is of most relevance for early retrieval of actives,[38,39] improved for 38 out of 46 assays by 0.03 on average. The best increase in performance was observed for assays number 2606 (membrane-associated serine protease in *M. tuberculosis*), 2805 (intestinal alkaline phosphatase in mouse), 463210 (caspase 7) and 463141 (caspase 3), with BEDROC improvements of 0.11, 0.11, 0.09 and 0.06, respectively. By contrast, a significant drop of 0.09 in BEDROC was observed for assay number 493011 (DNA deaminase APOBEC-3A). While improvement was modest for most assays, it was consistent, as shown by the error bars representing the standard error of the mean difference in performance between the informer set and the randomly selected set across all sizes. Given the relatively small training sets, varying in size from ~0.3% to 10% of the entire screening collection, large improvements in predictive power over the remaining 90%-99.7% would be unrealistic. We attempted to investigate the cause for the performance loss for the remaining 8 assays, but could not find an explanation: there was no apparent relationship with the average performance for that assay, nor the number of actives in that assay.

Scaffold retrieval for individual assays

We analyzed the scaffold retrieval rate (defined as the retrieved percentage of unique scaffolds belonging to active compounds in the test set; see Figure 5 – top) and the median z-scores (see Figure 5 – bottom) of actives identified in the top 5% ranked compounds in order to assess whether these actives were enriched for frequent hitters.

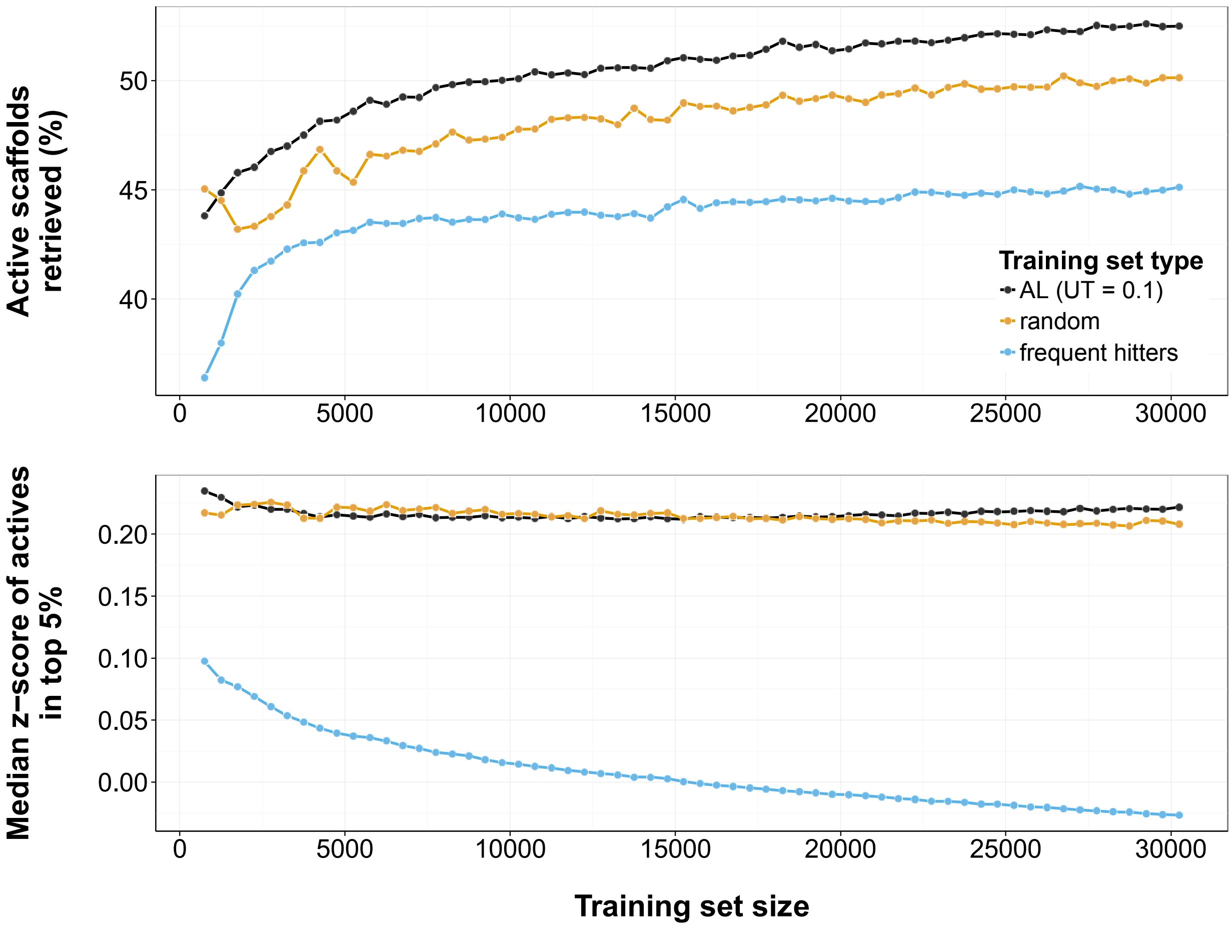


Figure 5. Active scaffold retrieval (%) and median z-scores of actives in top 5% (assay-specific validation). A consistently higher scaffold retrieval for the AL set, and similar median z-scores of actives in the top 5% ranked compounds (~0.20) for the AL set and the randomly selected set were observed. This indicates that the AL approach improves on the scaffold retrieval of active compounds, while not enriching for frequent hitters. For the frequent hitter set, scaffold retrieval is consistently reduced, hence showing that simply including active compounds from other assays in the training set does not improve the retrieval of diverse sets of actives.

A consistently higher scaffold retrieval for the AL set, and similar median z-scores (~0.20) for the AL set and the randomly selected set were observed. This indicates that the AL approach improves the retrieval of diverse sets of active compounds, while not enriching for frequent hitters. The frequent hitter set consistently shows worse performance than the other two sets in scaffold retrieval. In addition, the median z-score of the actives retrieved consistently drops from 0.09 to below 0 (Figure 5 – bottom). The latter drop is likely caused due to fewer compounds with high median z-scores remaining in the test set as training set size increases. Relative stability of the median z-score is observed for both the AL and random sets, indicating no enrichment for frequent hitters in the training set. In summary, we conclude that when the AL approach is used the scaffold retrieval is improved, frequent hitters are not enriched for and at the same time overall hit rates are improved.

Composition of informer set

In order to analyze the composition of the informer set in more detail, we calculated the fraction of the number of active compounds picked from the group 2 assays relative to the number of active compounds for each assay (see Figure 6).

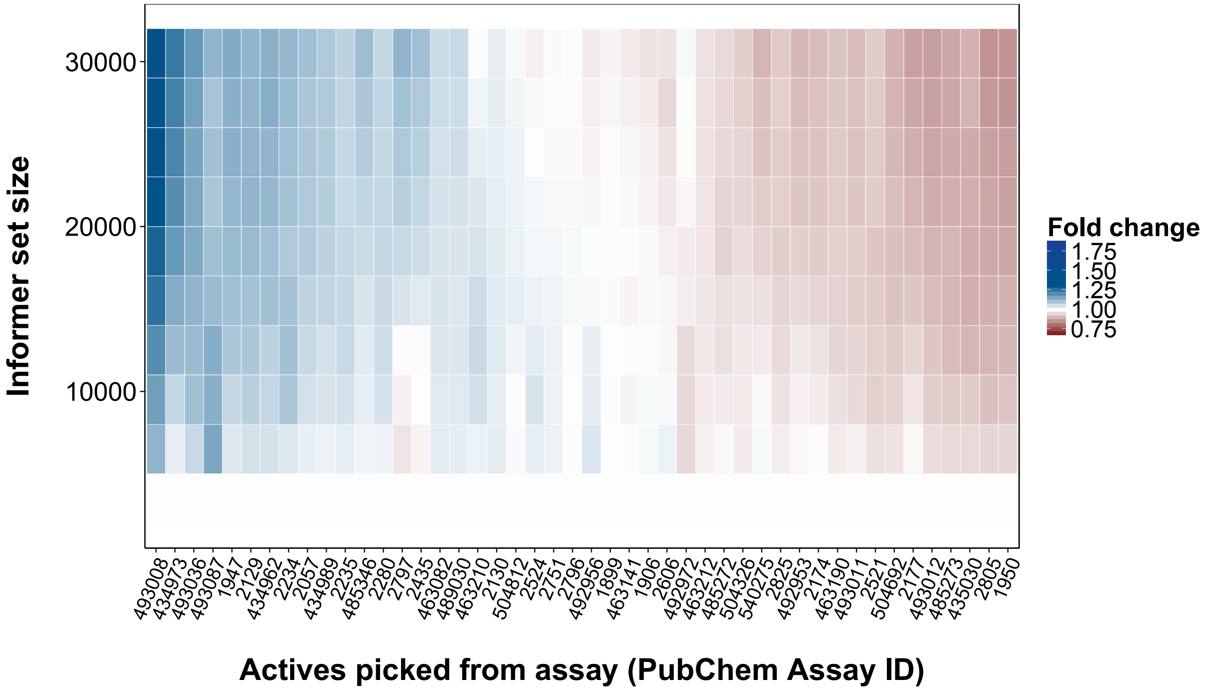


Figure 6. Composition of the informer set in terms of active compounds selected from group 2 assays. The heat map represents the composition of the informer set at varying sizes in terms of the fraction of the number of active compounds selected from group 2 assays relative to the number of active compounds for each assay. On the one hand, active compounds from assays number 493008 (troponin C type 1), 434973 (sentrin-specific protease 7), 493036 (neurotensin receptor type 1) and 493087 (insulin-degrading enzyme) are overrepresented (fold change > 1.1 at a set size of 30,000). On the other hand, active compounds from assays number 1950 (EBNA-1 protein), 2805 (intestinal alkaline phosphatase), 435030 (hypothetical protein HP1089) and 485273 (ubiquitin-conjugating enzyme E2N) are underrepresented (fold change < 0.9 at a set size of 30,000). While the AL approach improves performance for most assays, the average BEDROC(α=100) is higher for the assays with overrepresented actives (0.50) than for the assays with underrepresented actives (0.29).

On the one hand active compounds from assays number 493008 (troponin C type 1), 434973 (sentrin-specific protease 7), 493036 (neurotensin receptor type 1) and 493087 (insulin-degrading enzyme) are overrepresented in the informer set (maximum fold change > 1.15) while on the other hand active compounds from assays number 1950 (EBNA-1 protein), 2805 (intestinal alkaline phosphatase), 435030 (hypothetical protein HP1089) and 485273 (ubiquitin-conjugating enzyme E2N) are underrepresented (minimum fold change < 0.9). While the AL approach improves performance for most of the assays mentioned above (see Figure 4), interestingly, the average BEDROC is higher for those assays of which the active compounds are *overrepresented* (0.50) than for the assays of which the active compounds are *underrepresented* (0.29). This indicates that more actives are picked from assays already exhibiting good performance.

We attempted to investigate whether bias towards active compounds from particular assays in the informer set was related to improvement in performance over models trained on randomly selected sets for those assays, but could not find any link. We therefore conclude that this improvement in performance is due to better sampling of bioactivity space, as the AL approach iteratively augments the informer set with uncertain samples.

CONCLUSION

Strategies involving iterative cycles of feedback-driven compound selection and testing can be used when low assay throughput or high screening cost hinders the screening of large compound libraries. This creates the need for the exploratory screening of smaller informer sets to build predictive models for compound selection for follow-up testing. In this study, we performed a data-driven construction of an informer compound set with improved retrieval of actives in a subsequent selection round for apparently unrelated HTS assays. The benefit of this informer set was validated over randomly selected training sets on 46 PubChem[20] assays comprising at least 300,000 compounds. Overall, we highlight that such a set – of adjustable size, depending on the number of compounds one intends to screen – can be employed for routine exploratory screening in an assay-agnostic fashion for a gain in predictive power.

Averaged over all assays, an improvement in BEDROC, PRAUC and ROCAUC (of 0.024, 0.014 and 0.016, respectively) was observed with respect to random training sets, all with paired *t*–test p-values < 10^{-15}. The informer set improved the BEDROC for 38 out of 46 assays, indicating better early retrieval of actives. In addition, we found that our approach improved the retrieval of diverse sets of active compounds, while not enriching for frequent hitters, as scaffold retrieval was enhanced and the median z-score activity of the actives retrieved was unaffected. The informer set overrepresented actives from certain assays, and underrepresented actives from other assays. Interestingly, while the informer set increased performance for both groups of assays, the BEDROC was higher (0.50) for the assays of which the actives were overrepresented, than for assays with underrepresented actives (0.29).

We conclude that our AL approach is able to more effectively sample descriptor space, expected to improve the retrieval of active compounds in subsequent screens, thereby reducing the time and expense required to arrive at the same number of hits.

ASSOCIATED CONTENT

Supporting Information.

The following files are available free of charge.

Supplementary Figure S1 (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: florian.nigsch@novartis.com (FN)

*E-mail: ab454@cam.ac.uk (AB)

Notes

The authors declare no competing interests.

Author Contributions

SP designed the study, carried out the computational experiments and prepared the manuscript. AIJ, JJ, AB and FN participated in the study design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENT

SP thanks the Netherlands Organisation for Scientific Research (NWO, grant number NWO-017.009-065), Novartis Institutes for BioMedical Research (NIBR) and the Prins Bernhard Cultuurfonds for funding. AB thanks the European Research Commission for funding. The authors thank Oscar Méndez-Lucio and Avid M. Afzal for insightful discussions.

REFERENCES

(1) Macarron, R. Critical Review of the Role of HTS in Drug Discovery. *Drug Discov. Today* **2006**, *11* (7-8), 277–279.

(2) Mayr, L. M.; Fuerst, P. The Future of High-Throughput Screening. *J. Biomol. Screen.* **2008**, *13* (6), 443–448.

(3) Phatak, S. S.; Stephan, C. C.; Cavasotto, C. N. High-Throughput and in Silico Screenings in Drug Discovery. *Expert. Opin. Drug Discov.* **2009**, *4*, 947–959.

(4) Mayr, L. M.; Bojanic, D. Novel Trends in High-Throughput Screening. *Curr. Opin. Pharmacol.* **2009**, *9* (5), 580–588.

(5) Valler, M. J.; Green, D. Diversity Screening versus Focussed Screening in Drug Discovery. *Drug Discov. Today* **2000**, *5* (7), 286–293.

(6) Fox, S.; Farr-Jones, S.; Sopchak, L.; Boggs, A.; Nicely, H. W.; Khoury, R.; Biros, M. High-Throughput Screening: Update on Practices and Success. *J. Biomol. Screen.* **2006**, *11* (7), 864–869.

(7) Koutsoukas, A.; Paricharak, S.; Galloway, W. R. J. D.; Spring, D. R.; IJzerman, A. P.; Glen, R. C.; Marcus, D.; Bender, A. How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2014**, *54*, 230–242.

(8) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug. Discov.* **2002**, *1* (11), 882–894.

(9) Astashkina, A.; Mann, B.; Grainger, D. W. A Critical Evaluation of in Vitro Cell Culture Models for High-Throughput Drug Screening and Toxicity. *Pharmacol. Ther.* **2012**, *134* (1), 82–106.

(10) Settles, B. Active Learning Literature Survey. *Mach. Learn.* **2010**, *15* (2), 201–221.

(11) Reker, D.; Schneider, G. Active-Learning Strategies in Computer-Assisted Drug Discovery. *Drug Discov. Today* **2015**, *20* (4), 458–465.

(12) Paricharak, S.; IJzerman, A. P.; Bender, A.; Nigsch, F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-House HTS Data. *ACS Chem. Biol.* **2016**, *11* (5), 1255–1264.

(13) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* **2012**, *7*, 1399–1409.

(14) Dančík, V.; Carrel, H.; Bodycombe, N. E.; Seiler, K. P.; Fomina-Yadlin, D.; Kubicek, S. T.; Hartwell, K.; Shamji, A. F.; Wagner, B. K.; Clemons, P. A. Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. *J. Biomol. Screen.* **2014**, *19* (5), 771–781.

(15) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, Å.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding to Proteins by Affinity Fingerprinting. *Chem. Biol.* **1995**, *2*, 107–118.

(16) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. “Bayes

Affinity Fingerprints" improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When Are Multitarget Drugs a Feasible Concept? *J. Chem. Inf. Model.* **2006**, *46*, 2445–2456.

(17) Nguyen, H. P.; Koutsoukas, A.; Mohd Fauzi, F.; Drakakis, G.; Maciejewski, M.; Glen, R. C.; Bender, A. Diversity Selection of Compounds Based on "Protein Affinity Fingerprints" Improves Sampling of Bioactive Chemical Space. *Chem. Biol. Drug Des.* **2013**, *82*, 252–266.

(18) Givehchi, A.; Bender, A.; Glen, R. C. Analysis of Activity Space by Fragment Fingerprints, 2D Descriptors, and Multitarget Dependent Transformation of 2D Descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 1078–1083.

(19) Riniker, S.; Wang, Y.; Jenkins, J. L.; Landrum, G. A. Using Information from Historical High-Throughput Screens to Predict Active Compounds. *J. Chem. Inf. Model.* **2014**, *54*, 1880–1891.

(20) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucl. Acids Res.* **2012**, *40* (Database issue), D400–D412.

(21) Maciejewski, M.; Wassermann, A. M.; Glick, M.; Lounkine, E. An Experimental Design Strategy: Weak Reinforcement Leads to Increased Hit Rates and Enhanced Chemical Diversity. *J. Chem. Inf. Model.* **2015**, *55*, 956–962.

(22) Hert, J.; Irwin, J. J.; Laggner, C.; Keiser, M. J.; Shoichet, B. K. Quantifying Biogenic Bias in Screening Libraries. *Nat. Chem. Biol.* **2010**, *5* (7), 479–483.

(23) Fox, S.; Farr-Jones, S.; Sopchak, L.; Boggs, A.; Comley, J. High-Throughput Screening: Searching for Higher Productivity. *J. Biomol. Screen.* **2004**, *9* (4), 354–358.

(24) Pereira, D. A.; Williams, J. A. Origin and Evolution of High Throughput Screening. *Br. J. Pharmacol.* **2007**, *152* (1), 53–61.

(25) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-Likeness Score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, *48* (1), 68–74.

(26) Gupta, S.; Aires-de-Sousa, J. Comparing the Chemical Spaces of Metabolites and Available Chemicals: Models of Metabolite-Likeness. *Mol. Divers.* **2007**, *11* (1), 23–36.

(27) O'Hagan, S.; Swainston, N.; Handl, J.; Kell, D. B. A "Rule of 0.5" for the Metabolite-Likeness of Approved Pharmaceutical Drugs. *Metabolomics* **2015**, *11*, 323–339.

(28) Klekota, J.; Roth, F. P. Chemical Substructures That Enrich for Biological Activity. *Bioinformatics* **2008**, *24* (21), 2518–2525.

(29) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S. Methods for Drug Discovery: Devlopment of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988**, *31* (12), 2235–2246.

(30) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1997**, *38*, 165–179.

(31) Attenberg, J.; Ertekin, S. Class Imbalance and Active Learning. In *Imbalanced Learning:*

Foundations, Algorithms, and Applications, First Edition; He, H., Ma, Y., Eds.; John Wiley & Sons, Inc., 2013; pp 101–149.

(32) Young, S. S.; Sheffield, C. F.; Farmen, M. Optimum Utilization of a Compound Collection or Chemical Library for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 892–899.

(33) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59–67.

(34) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.

(35) Riniker, S.; Fechner, N.; Landrum, G. A. Heterogeneous Classifier Fusion for Ligand-Based Virtual Screening: Or, How Decision Making by Committee Can Be a Good Thing. *J. Chem. Inf. Model.* **2013**, *53*, 2829–2836.

(36) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn : Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(37) Davis, J.; Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine learning*; 2006; pp 233–240.

(38) Truchon, J.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics

for The "early Recognition" problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488–508.

(39) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminform.* **2013**, *5* (5), 26–42.

(40) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893.

(41) *RDKit: Cheminformatics and Machine Learning Software (Http://www.rdkit.org/); 2013*.

(42) *Tableau Desktop, Version 9.0.1; Tableau Software Inc., 2015*.

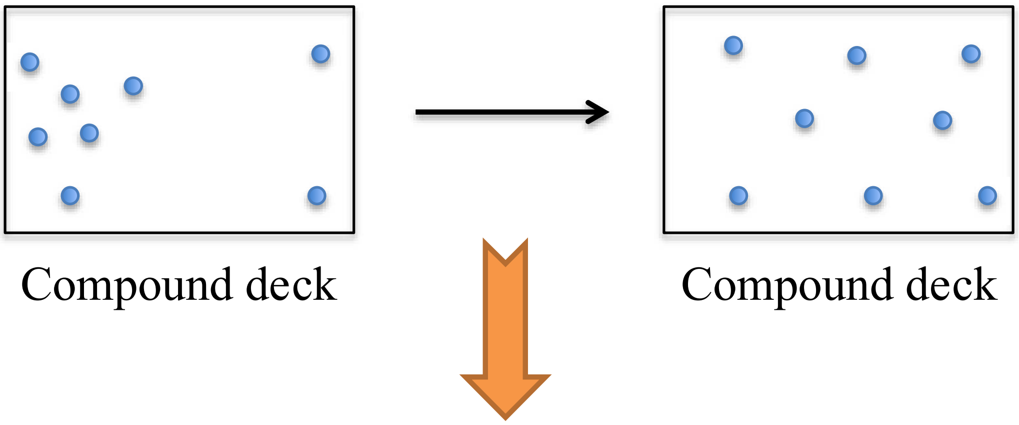
(43) Dessau, R. B.; Pipper, C. B. "R"--Project for Statistical Computing. *Ugeskr. Laeger.* **2008**, *170* (5), 328–330.

(44) Baell, J.; Walters, M. A. Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.

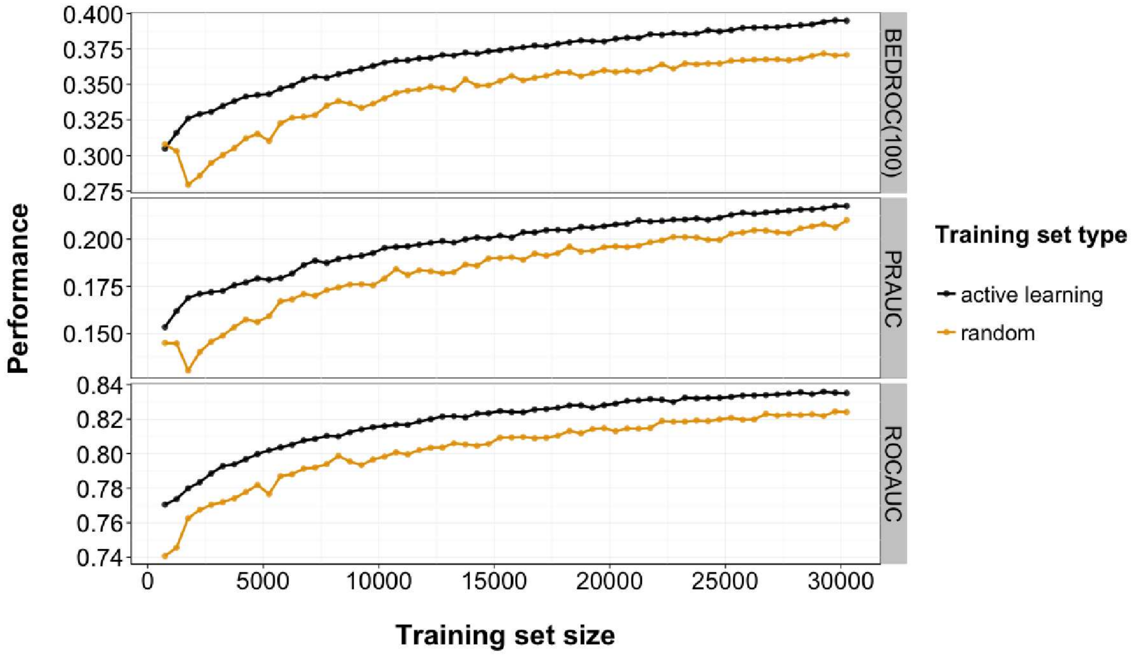
(45) Che, J.; King, F. J.; Zhou, B.; Zhou, Y. Chemical and Biological Properties of Frequent Screening Hits. *J. Chem. Inf. Model.* **2012**, *52*, 913–926.

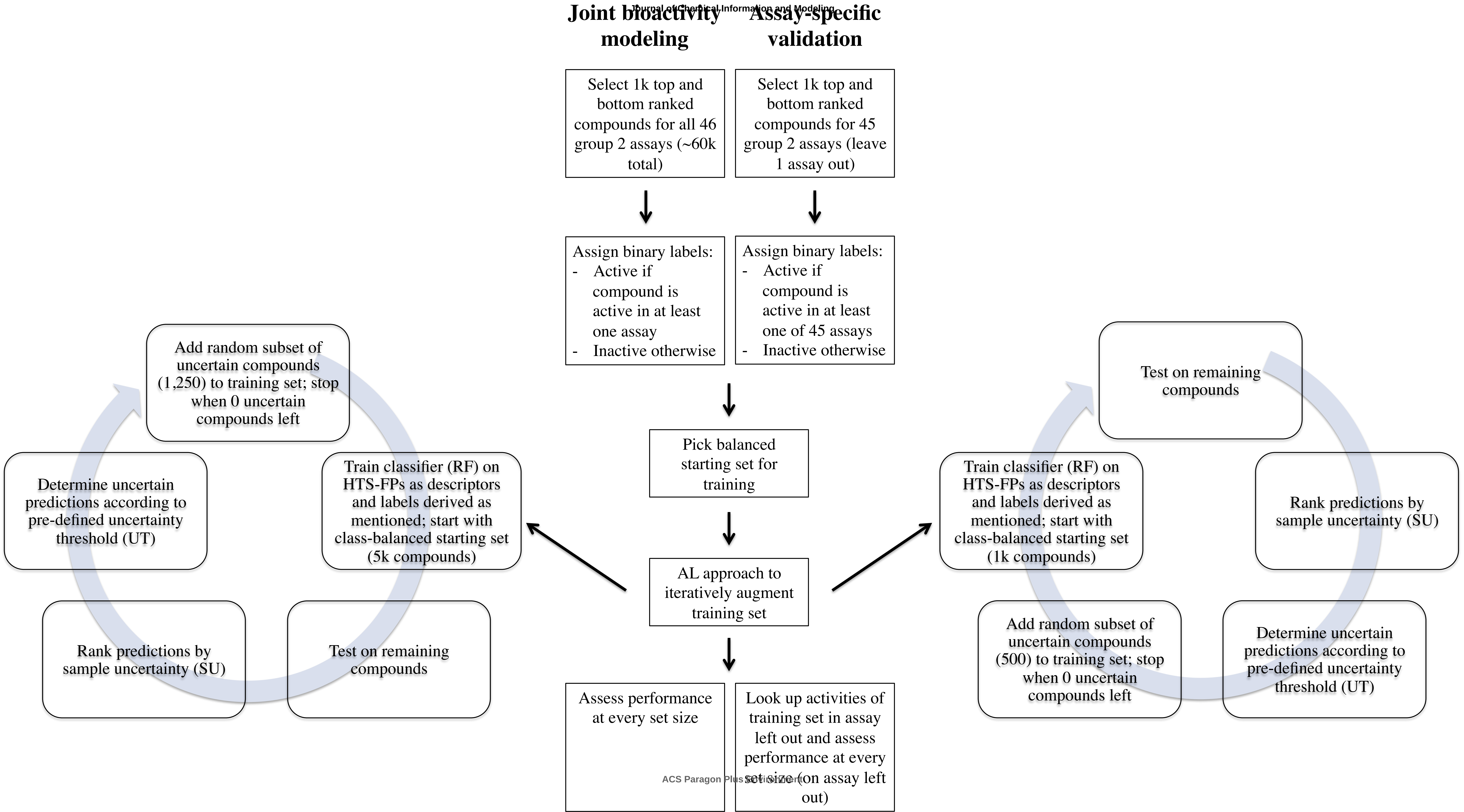
TABLE OF CONTENTS FIGURE

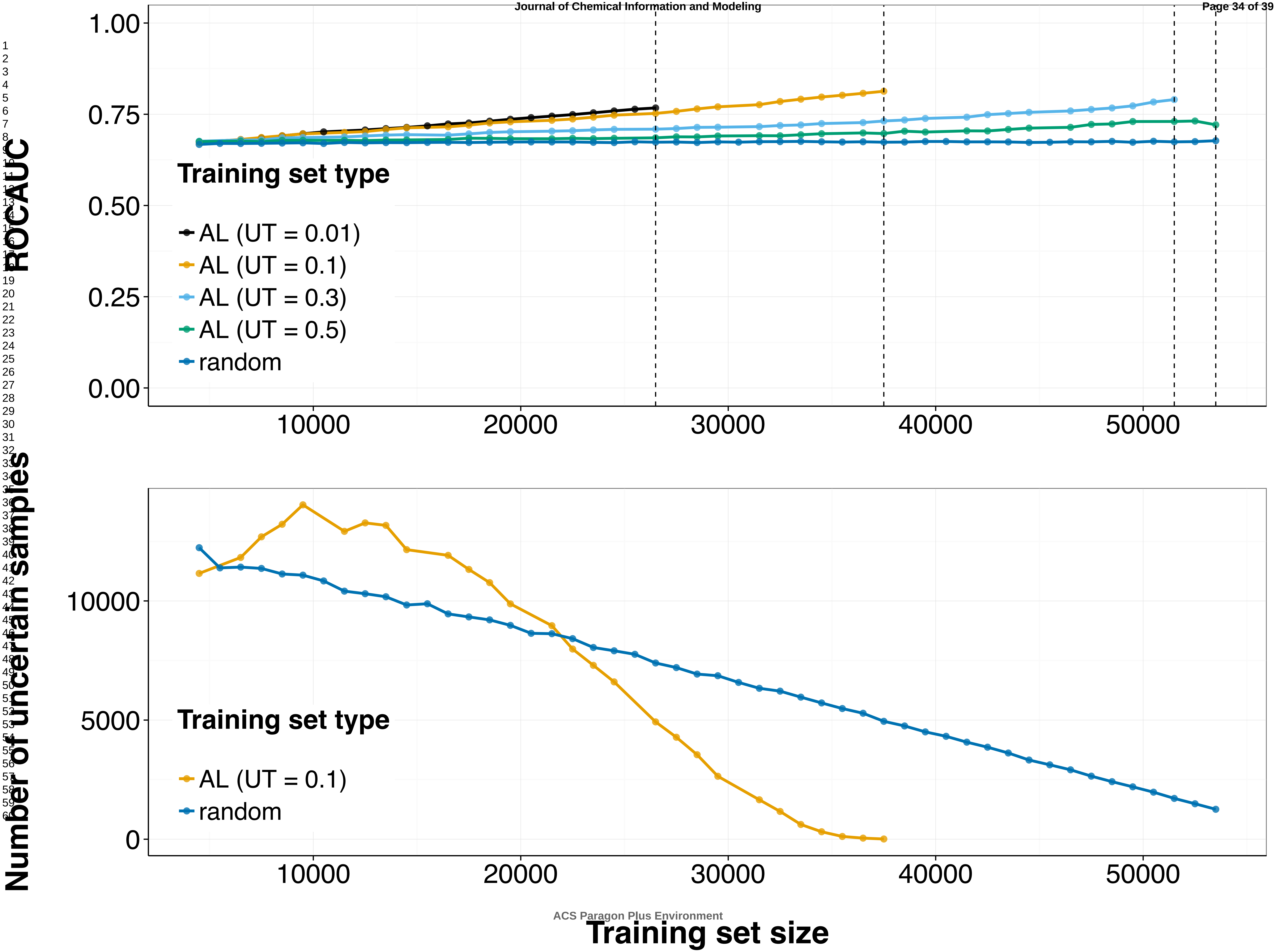
Improved sampling through active learning in high-throughput screening

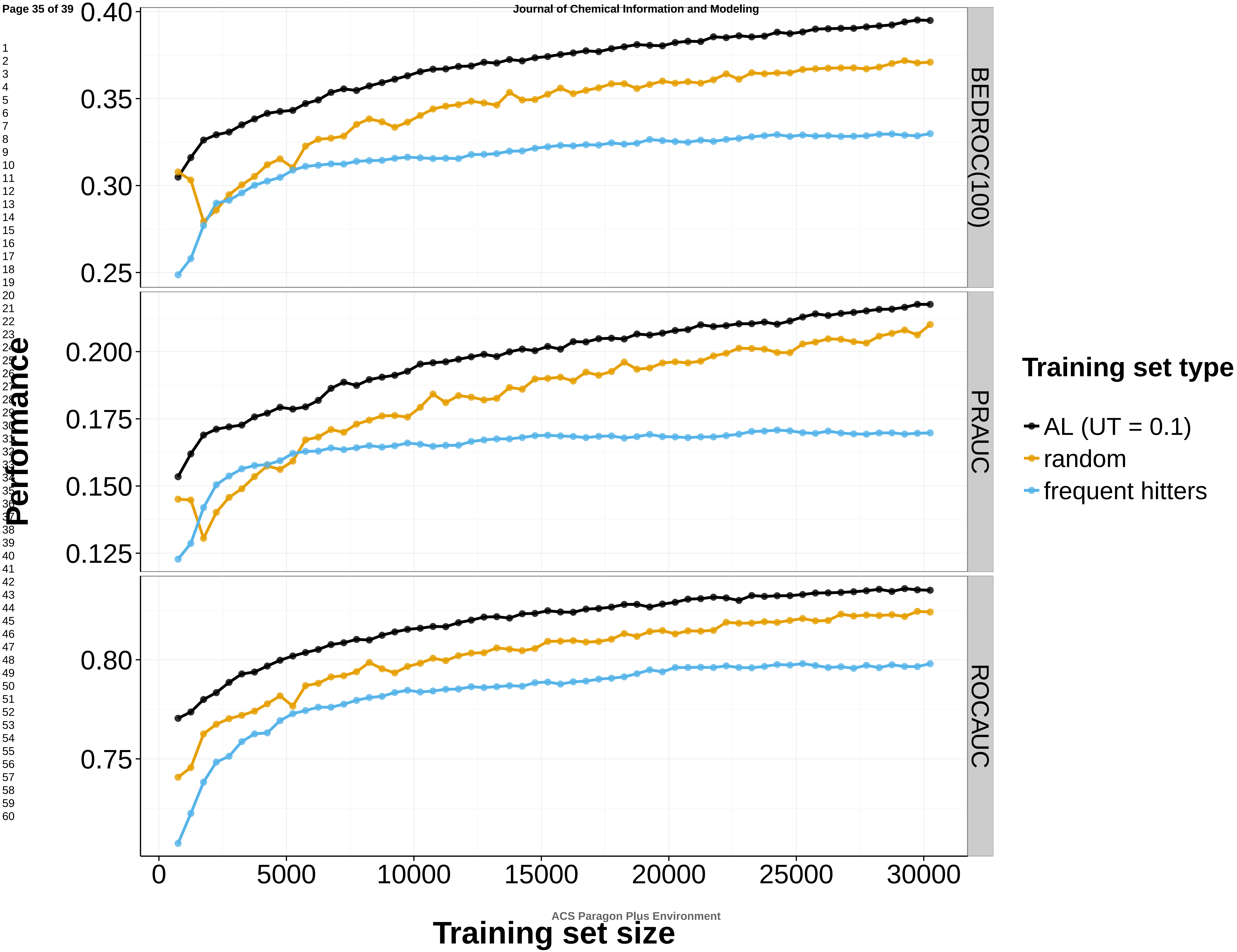


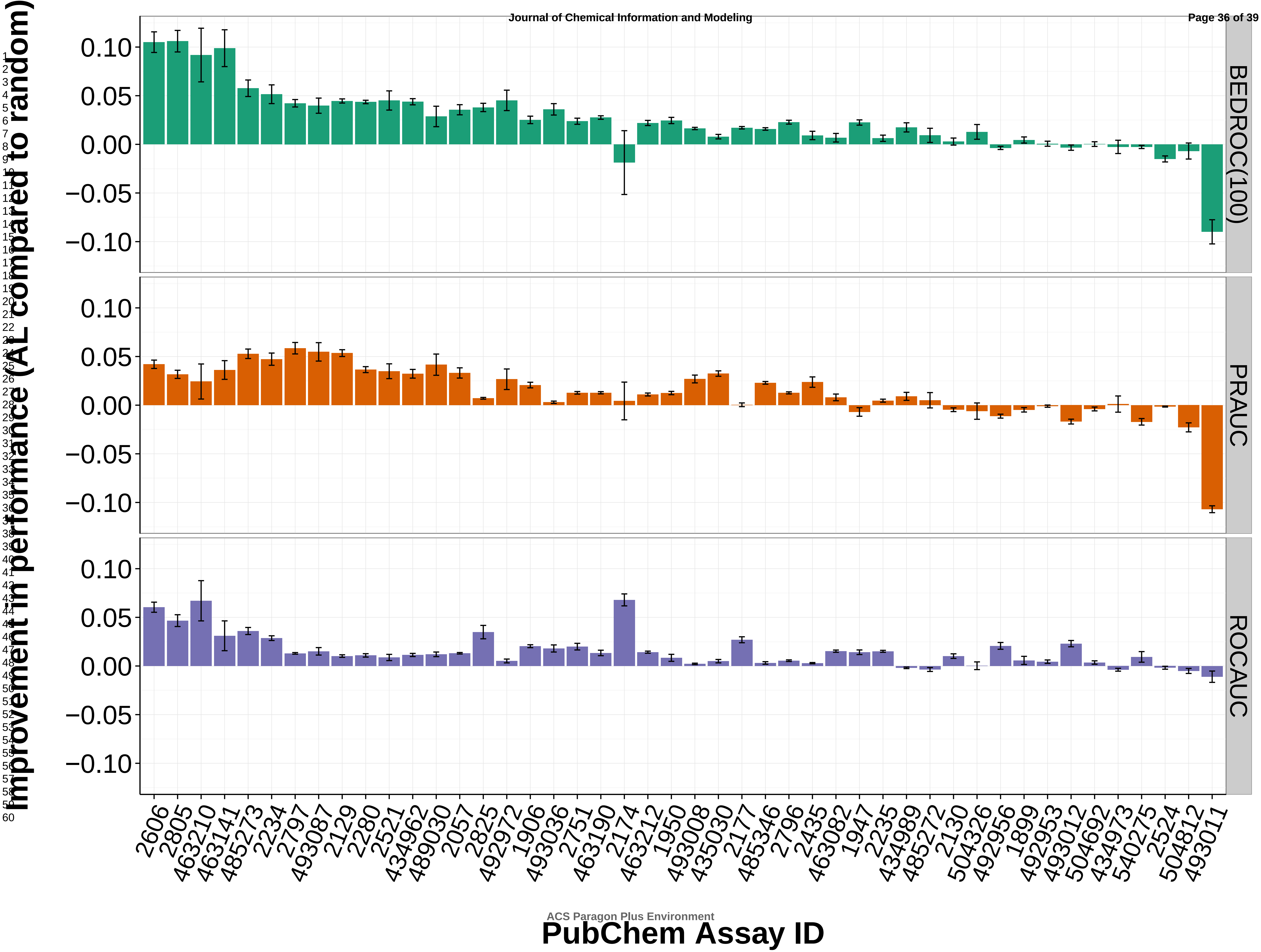
Consistently improved retrieval of actives

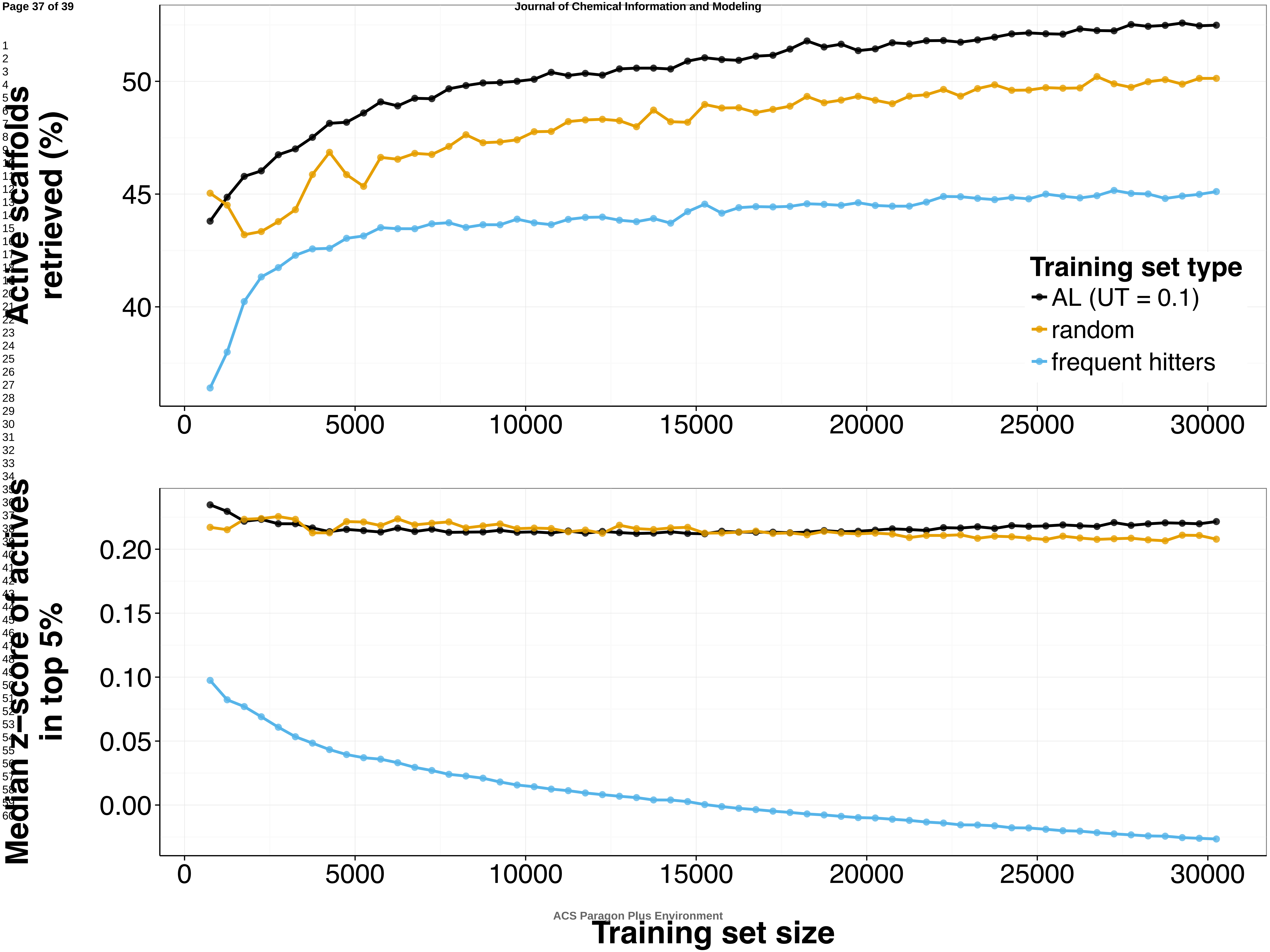


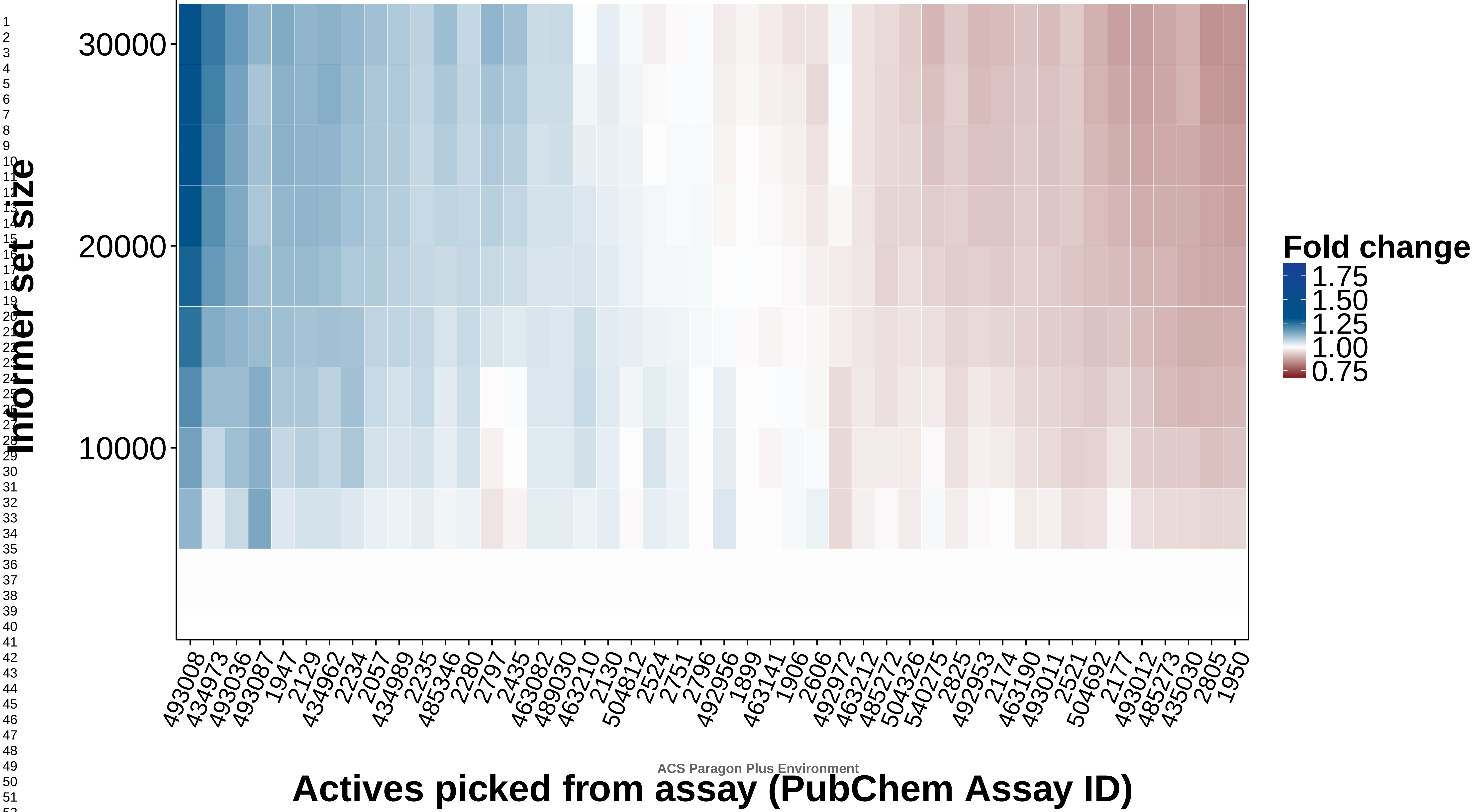

Joint bioactivity modeling
Assay-specific validation
Select 1k top and bottom ranked compounds for all 46 group 2 assays (~60k total)
Select 1k top and bottom ranked compounds for 45 group 2 assays (leave 1 assay out)
Assign binary labels:
- Active if compound is active in at least one assay
- Inactive otherwise
Assign binary labels:
- Active if compound is active in at least one of 45 assays
- Inactive otherwise
Pick balanced starting set for training
AL approach to iteratively augment training set
Train classifier (RF) on HTS-FPs as descriptors and labels derived as mentioned; start with class-balanced starting set (5k compounds)
Test on remaining compounds
Rank predictions by sample uncertainty (SU)
Determine uncertain predictions according to pre-defined uncertainty threshold (UT)
Add random subset of uncertain compounds (1,250) to training set; stop when 0 uncertain compounds left
Train classifier (RF) on HTS-FPs as descriptors and labels derived as mentioned; start with class-balanced starting set (1k compounds)
Test on remaining compounds
Rank predictions by sample uncertainty (SU)
Determine uncertain predictions according to pre-defined uncertainty threshold (UT)
Add random subset of uncertain compounds (500) to training set; stop when 0 uncertain compounds left
Assess performance at every set size
Look up activities of training set in assay left out and assess performance at every set size (on assay left out)


ROCAUC
1.00
0.75
0.50
0.25
0.00
Training set type
AL (UT = 0.01)
AL (UT = 0.1)
AL (UT = 0.3)
AL (UT = 0.5)
random
10000
20000
30000
40000
50000
Number of uncertain samples
10000
5000
0
Training set type
AL (UT = 0.1)
random
10000
20000
30000
40000
50000
Training set size

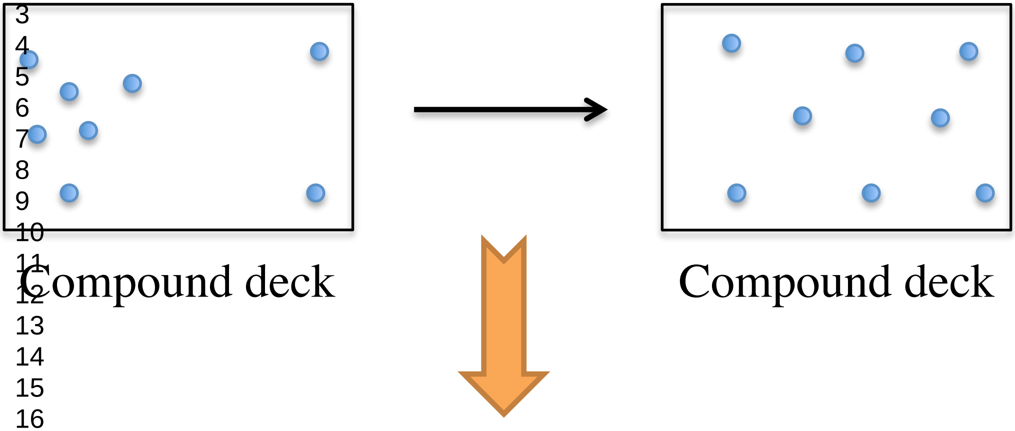

Performance
Training set size
0.40
0.35
0.30
0.25
BEDROC(100)
0.200
0.175
0.150
0.125
PRAUC
0.80
0.75
ROCAUC
0
5000
10000
15000
20000
25000
30000
Training set type
AL (UT = 0.1)
random
frequent hitters


Improvement in performance (AL compared to random)
0.10
0.05
0.00
−0.05
−0.10
BEDROC(100)
PRAUC
ROCAUC
2606
2805
463210
463141
485273
2234
2797
493087
2129
2280
2521
434962
489030
2057
2825
492972
1906
493036
2751
463190
2174
463212
1950
493008
435030
2177
485346
2796
2435
463082
1947
2235
434989
485272
2130
504326
492956
1899
492953
493012
504692
434973
540275
2524
504812
493011
PubChem Assay ID


Active scaffolds retrieved (%)
50
45
40
0
5000
10000
15000
20000
25000
30000
Training set type
AL (UT = 0.1)
random
frequent hitters
Median z−score of actives in top 5%
0.20
0.15
0.10
0.05
0.00
0
5000
10000
15000
20000
25000
30000
Training set size


Informer set size
30000
20000
10000
Fold change
1.75
1.50
1.25
1.00
0.75
493008
434973
493036
493087
1947
2129
434962
2234
2057
434989
2235
485346
2280
2797
2435
463082
489030
463210
2130
504812
2524
2751
2796
492956
1899
463141
1906
2606
492972
463212
485272
504326
540275
2825
492953
2174
463190
493011
2521
504692
2177
493012
485273
435030
2805
1950
Actives picked from assay (PubChem Assay ID)

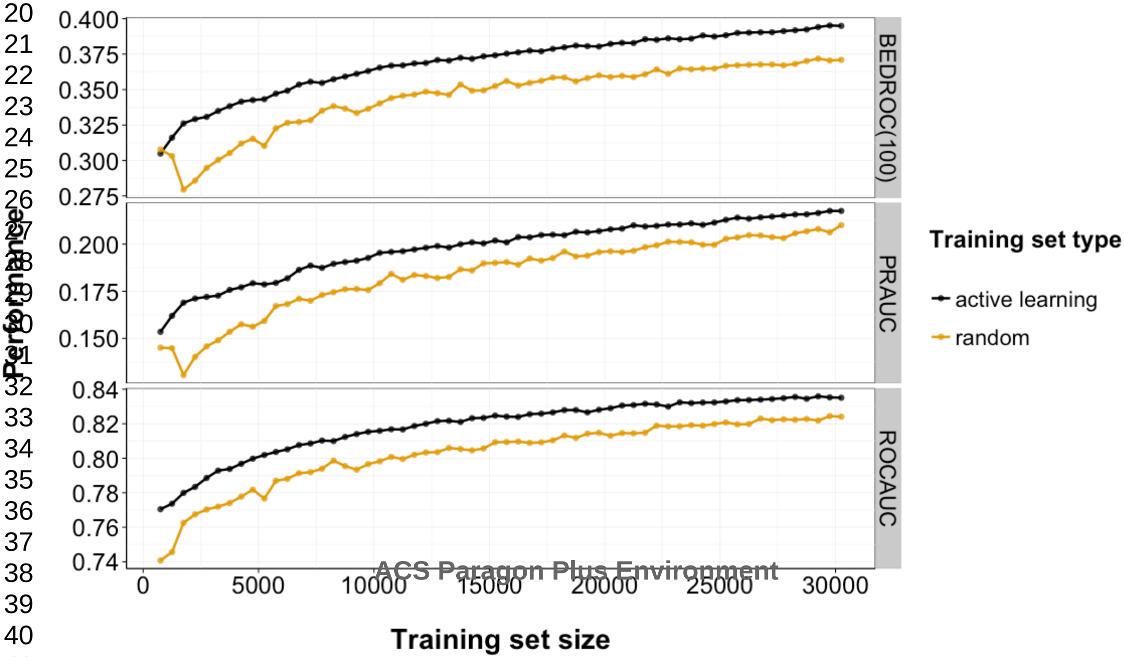
Improved sampling through active learning in high-throughput screening



Compound deck

Compound deck

Consistently improved retrieval of actives


BEDROC(100)
PRAUC
ROCAUC
Performance
Training set size
Training set type
active learning
random