Electronic Supplementary Material (ESI) for RSC Medicinal Chemistry.

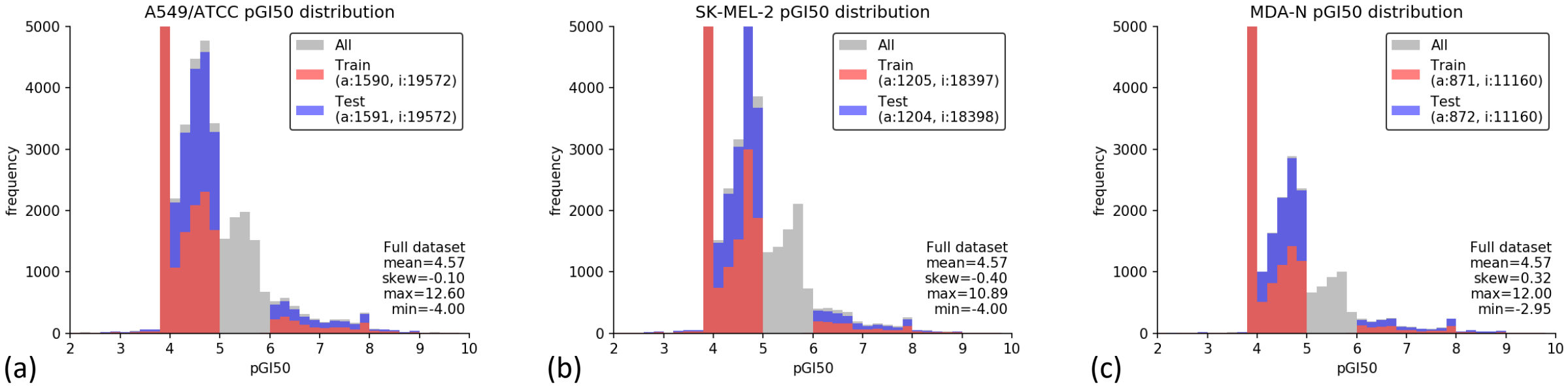
Active learning effectively identifies minimal set of maximally informative and asympotically performant cytotoxic structure-activity patterns in NCI-60 cell lines

Supplementary Figures and Tables

Takumi Nakano, Shunichi Takeda, J.B. Brown

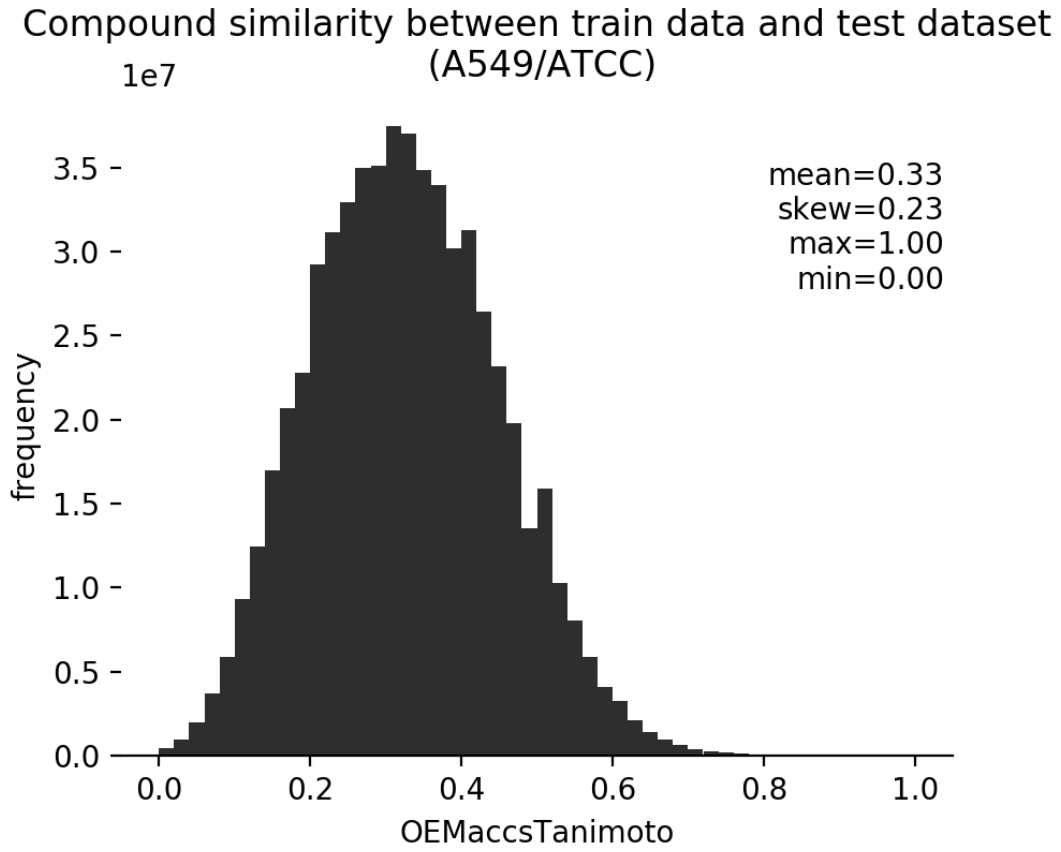
	All compounds	SK-MEL-2 pre-discretize	SK-MEL-2 post-discretize
Num. Compounds	274568	47305	40275
Mean	315	380	373
Stdev	144	159	157
Skew	3.2	3.1	3.0
Max	4031	3144	3144
10/90 percentile	173/472	231/548	226/539

Supplementary Table 1: NCI-60 dataset compound molecular weight distribution
Compound molecular weights were computed using the *oechem.OECalculateMolecularWeight* method in the OpenEye OEChem API.

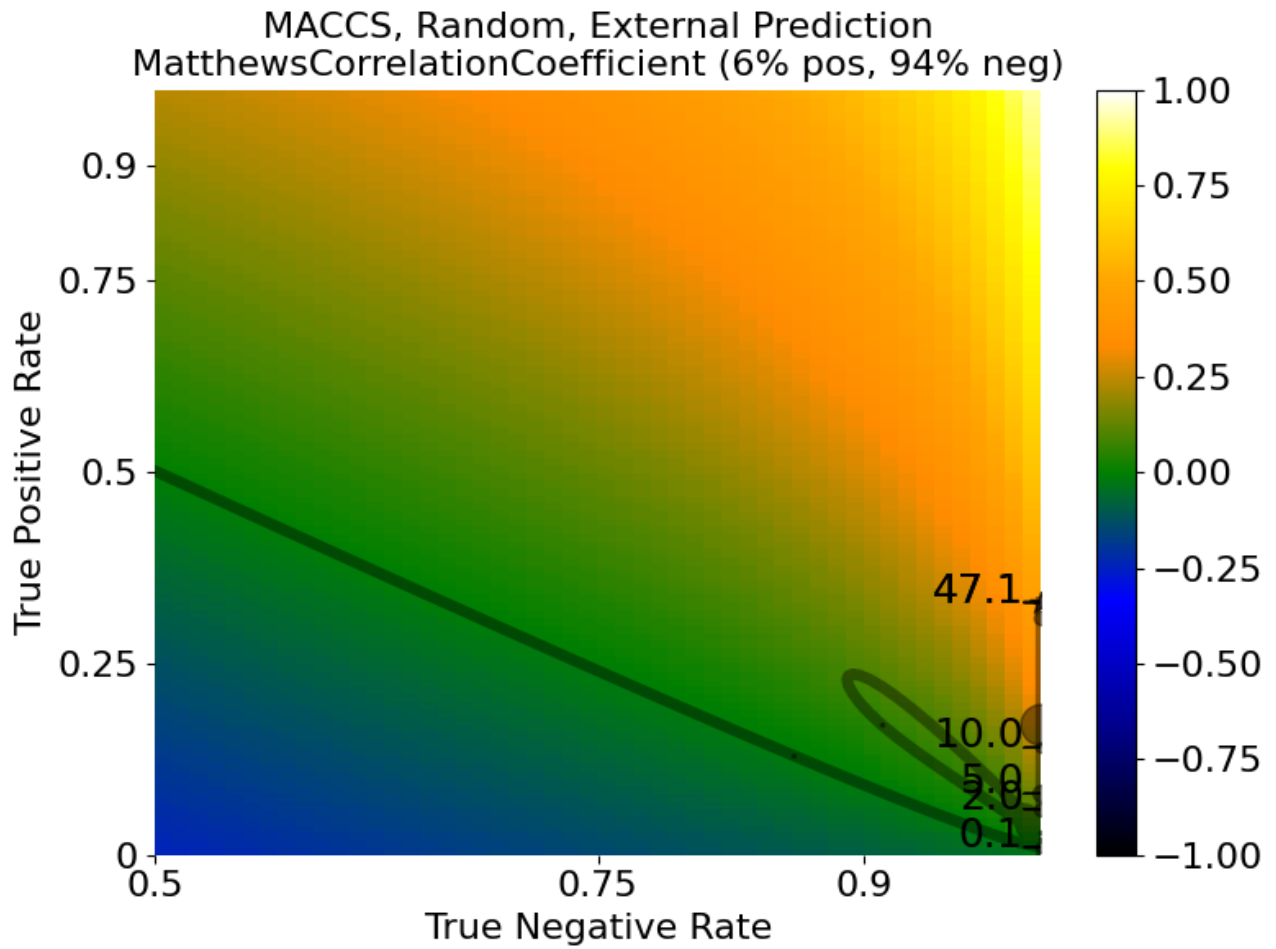


Supplementary Figure 1: pGI50 distributions of 3 cell lines.

Distributions of inhibitory concentrations are shown for each of the A549/ATCC lung (a), SK-MEL-2 melanoma (b), and MDA-N melanoma (c) cell lines. "a" and "i" represent the number of active/inactive compounds respectively. Gray histograms show the total distribution before dataset split, and red and blue histograms represent pGI50 distributions after one of the stratified dataset splits. The vertical axis was clipped to [0, 5000] and the horizontal axis was clipped to [2, 10] for better visualization.

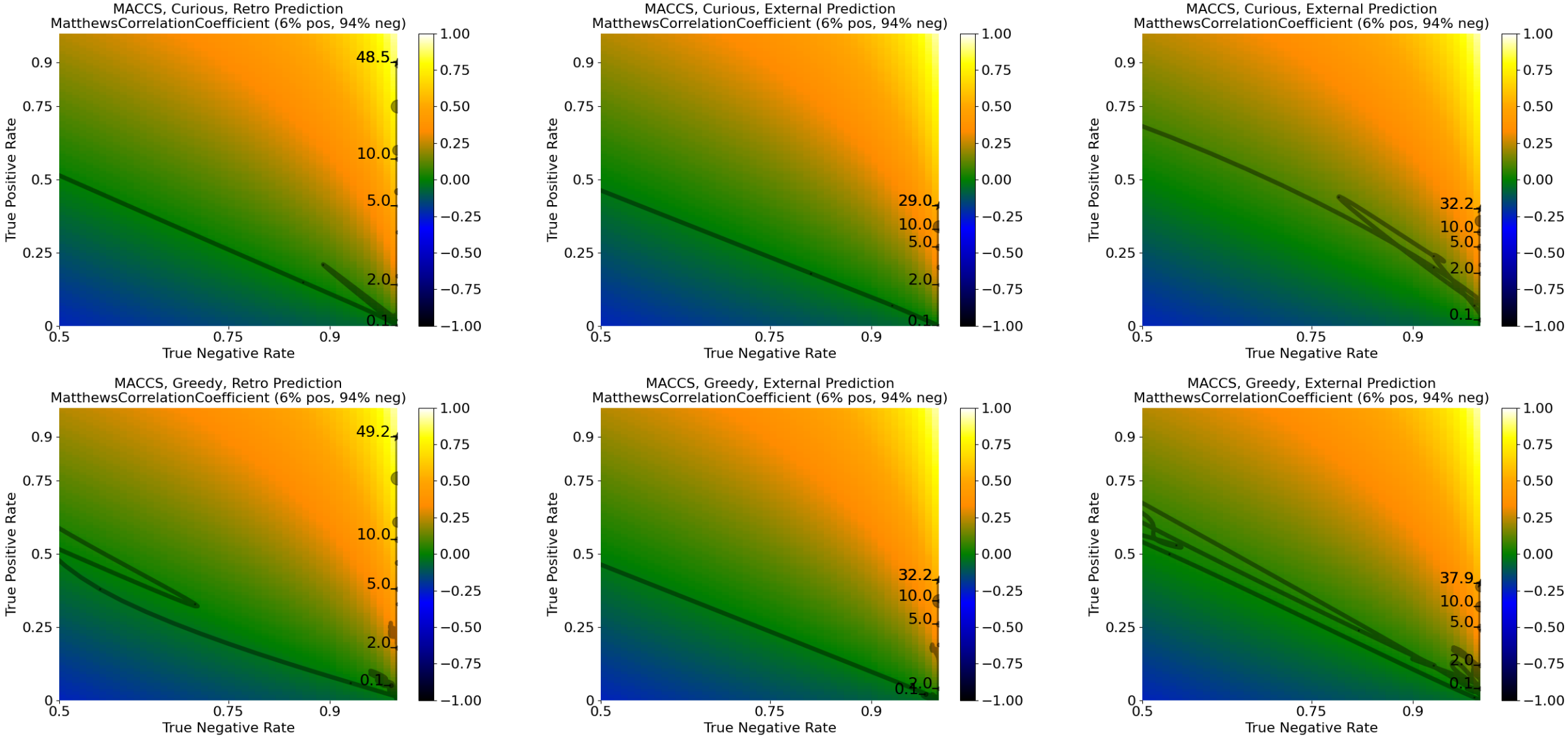


Supplementary Figure 2: MACCS-Tanimoto similarity distribution.
The distribution is generated by computing the similarities between each of approximately 20000 training and external compounds (400M pairs), repeated for five random splits (total of 2B MACCS-Tc values).



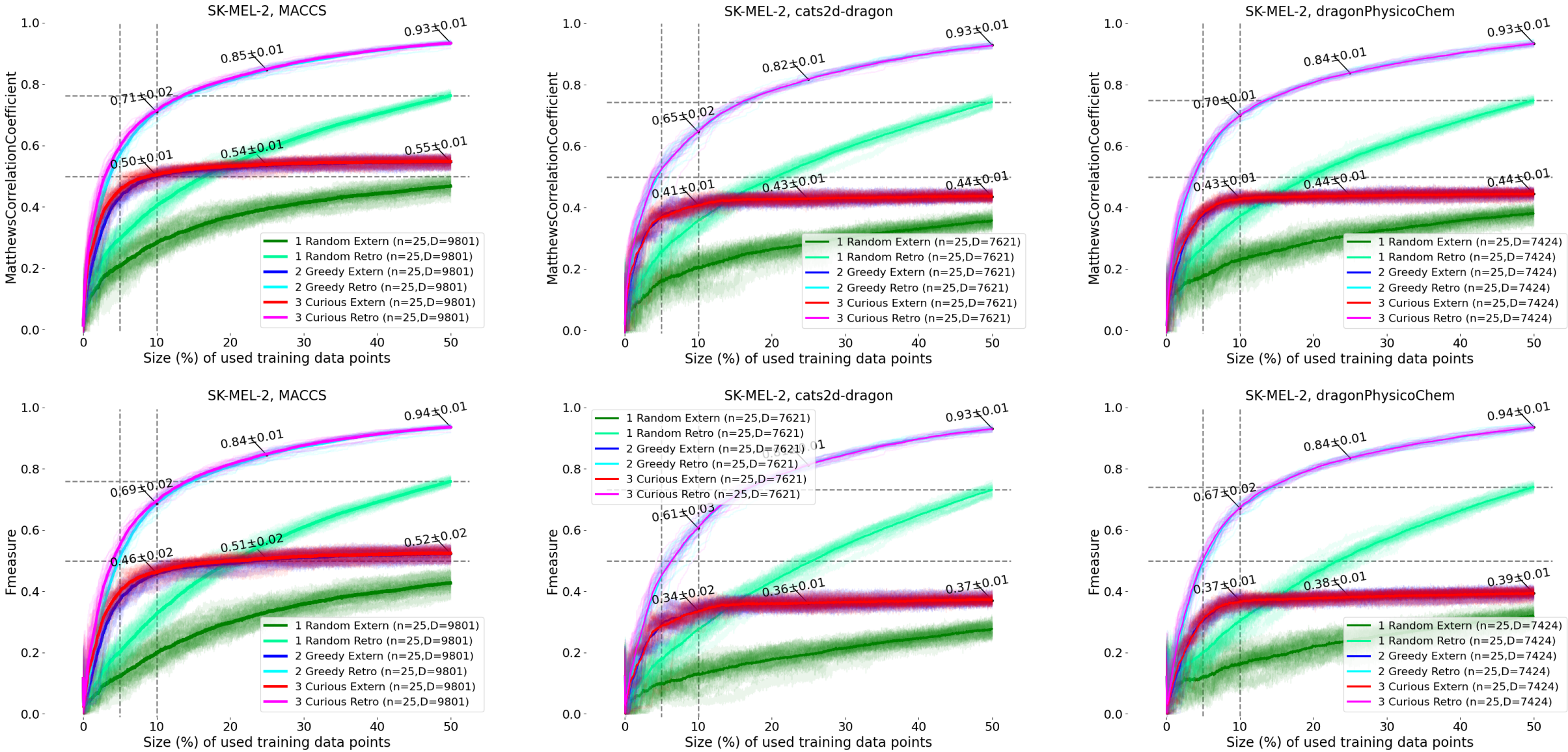
Supplementary Figure 3: Active projection for a run using random picking

Comparison to Figure 1 and Supplementary Figure 4 demonstrates similar TPR/TNR values at the model optimum, but with more data required when picking compound-response pairs randomly (melanoma SK-MEL-2).



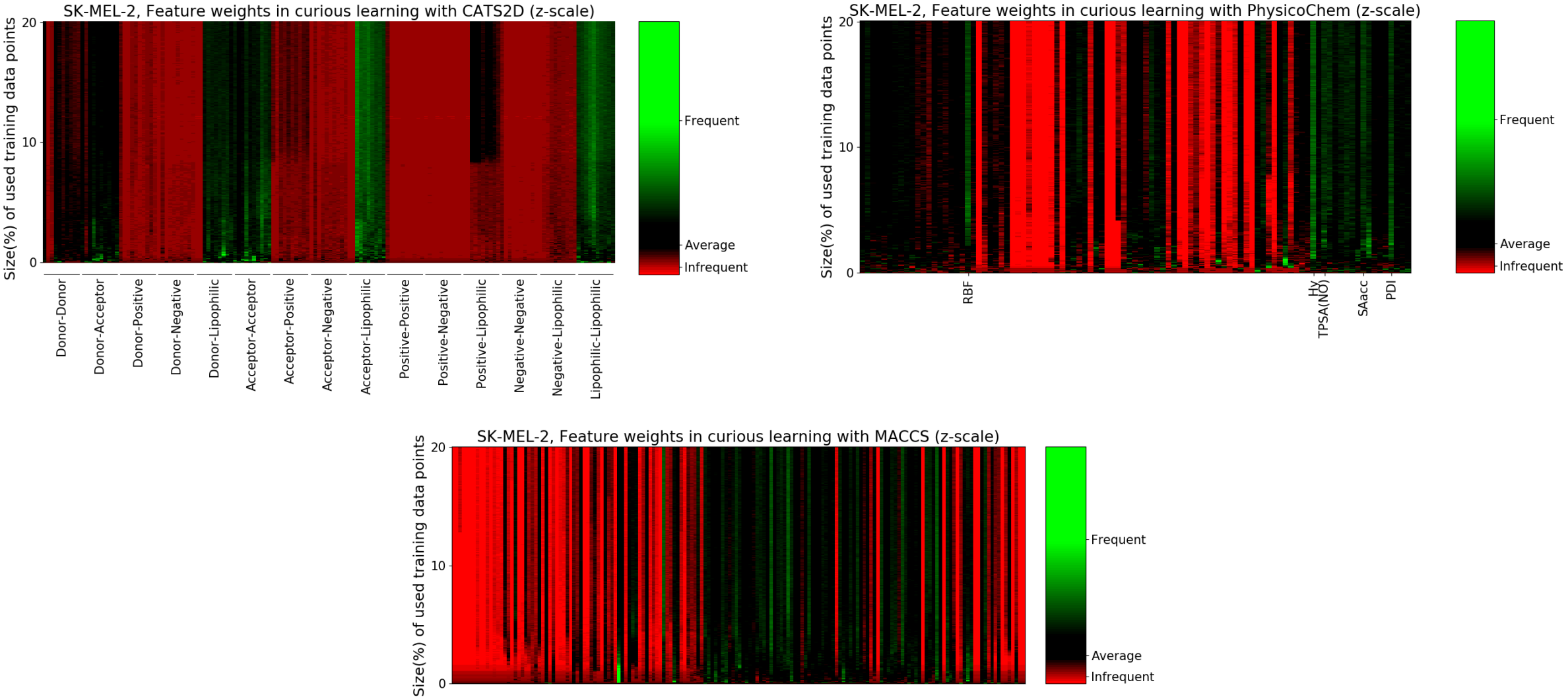
Supplementary Figure 4: Repeated runs of active learning and their active projections

Dynamics of actively learned model performance (melanoma SK-MEL-2) when picking compound-response data with the curiosity and greedy selection methods. Three independent executions are shown, where the left-most column demonstrates the active projection for the retrospective AL setting.

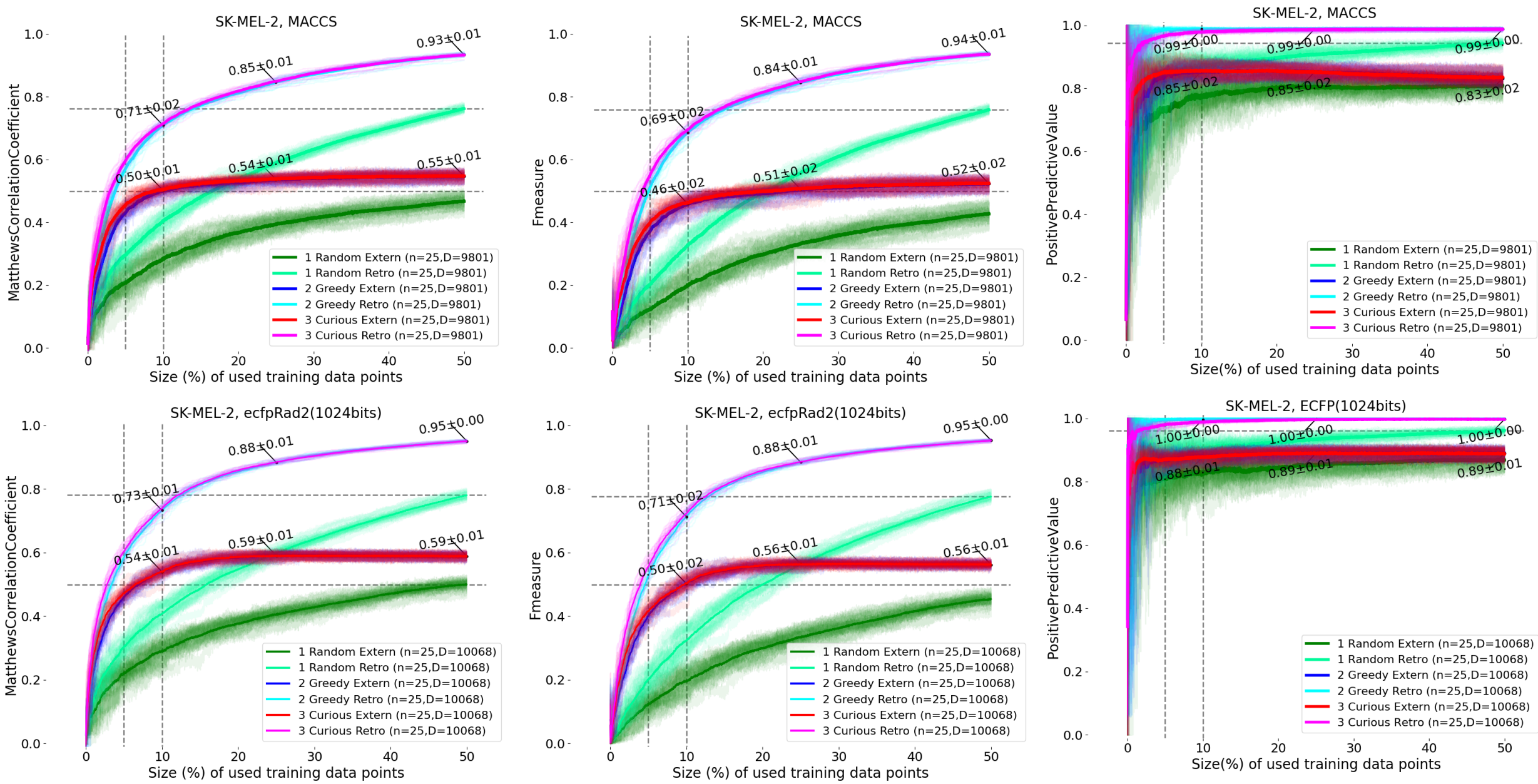


Supplementary Figure 5: Learning curves by additional compound descriptors.

In addition to the MACCS 166-bit representation (Figure 1 reproduced in left column here), pharmacophoric and physicochemical representations of compounds are tested. While retrospective performance on the training data is similar, the MACCS representation better extrapolates to the compound-response external data.

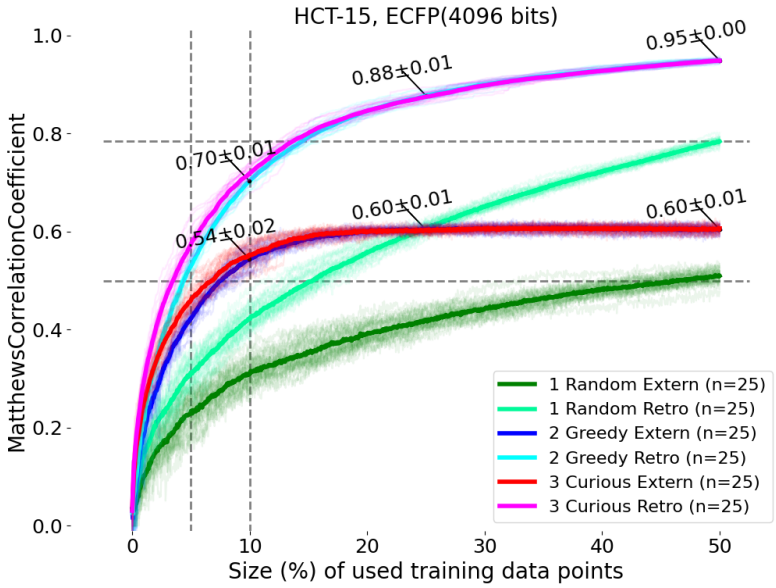


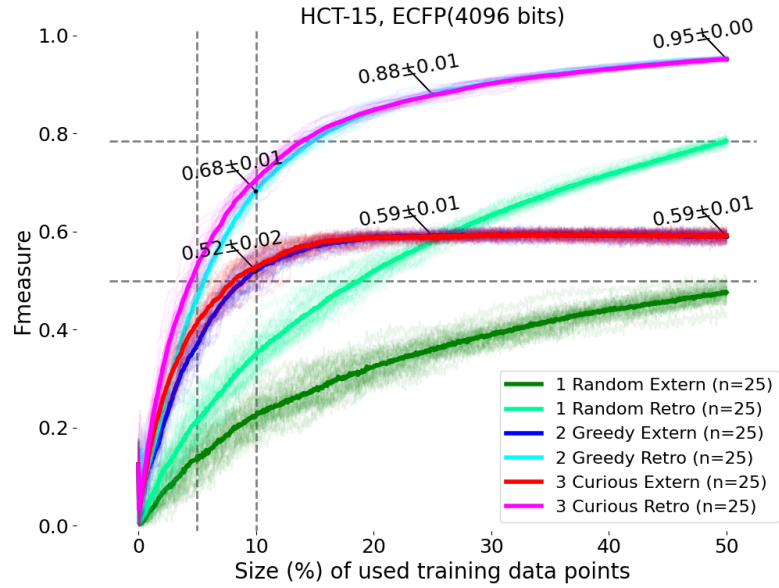
Supplementary Figure 6: Feature weight visualization for SK-MEL-2 with curious picking.
At each iteration of the fit-pick cycle, feature weights are calculated by their frequency of traversal in decision nodes, and then normalized for relative importance (z-scaling). Three different representations of compounds provide complementary views of chemical features frequent in (non-)cytotoxic decision making. Chemical features essential for penetratration into cellular cytoplasm, such as lipophilic character, are frequent. Physicochemical properties such as polarizable surface area and hydrophilic factor are consistently highly weighted. Whereas some features are initially (un)important, they shift during the course of the active learning cycle.

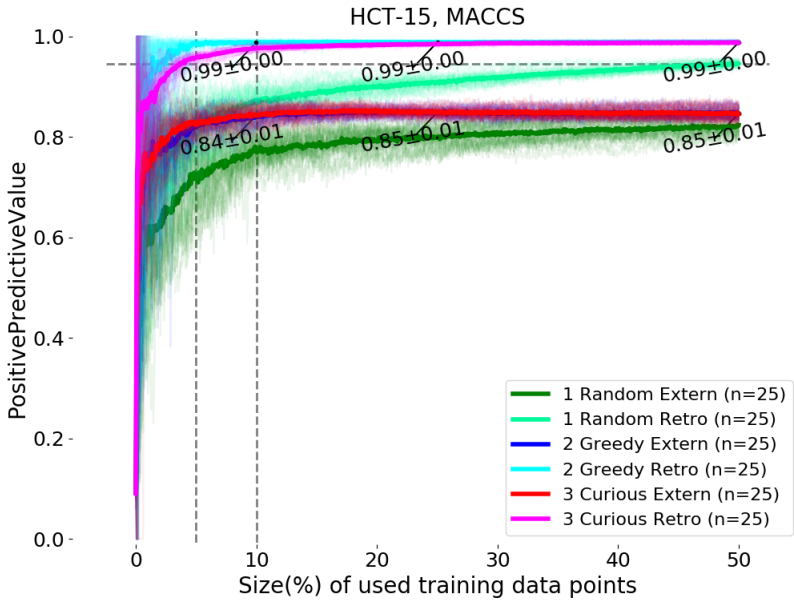


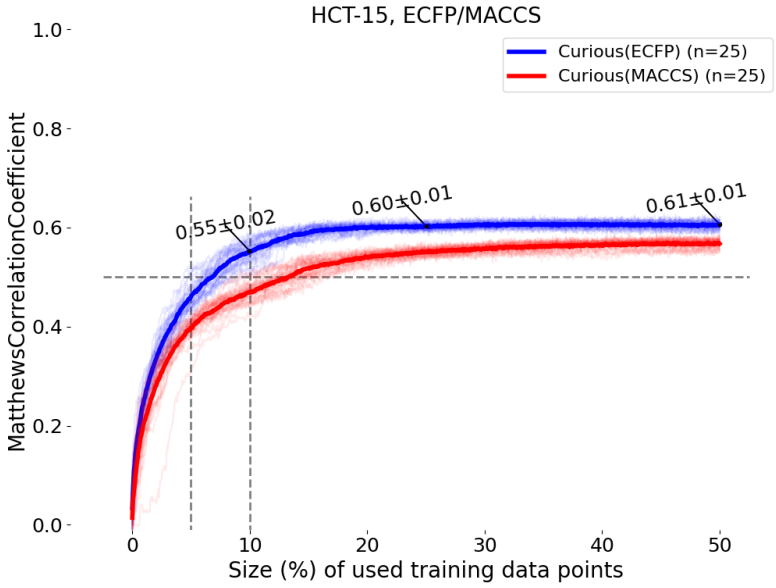
Supplementary Figure 7: Experiments modeling SK-MEL-2 by MACCS and ECFP descriptions.

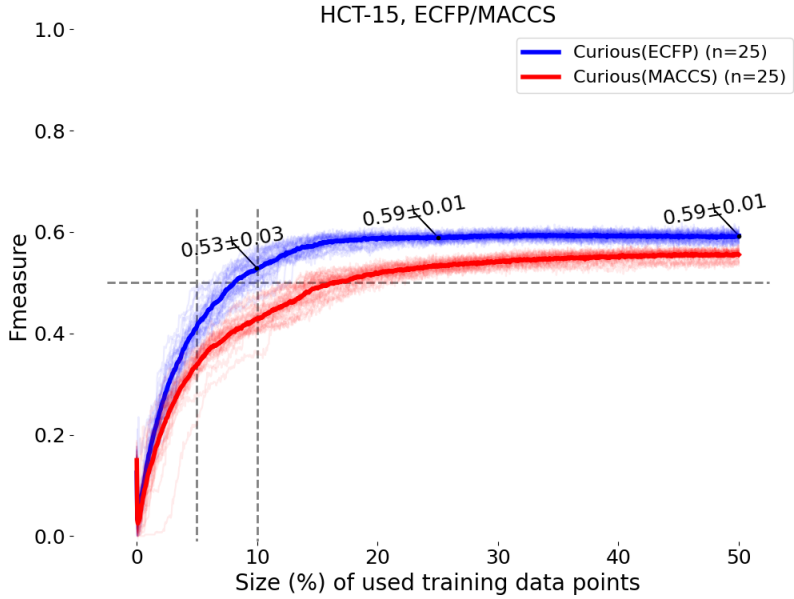
The melanoma cell line consistently analyzed in the main manuscript was used for a comparison of performance between MACCS and 1024-bit ECFPr2 descriptors.
[Top] Reproduced from main manuscript Figure 1, the MCC and F1 performances using the MACCS description of compounds.
[Bottom] The gain from replacing MACCS keys with ECFPs is approximately 0.05 for either metric.

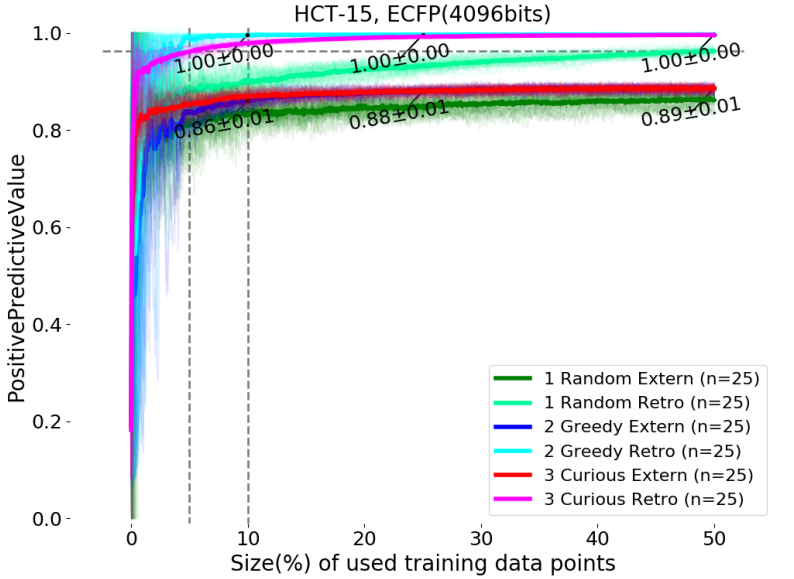






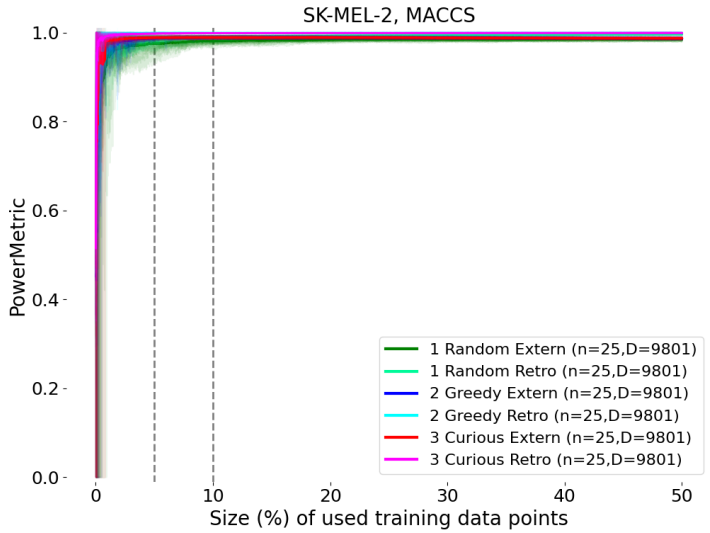


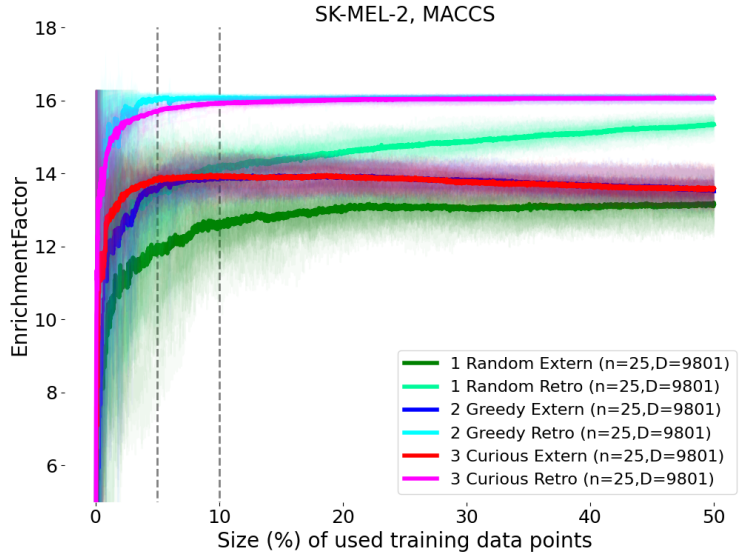


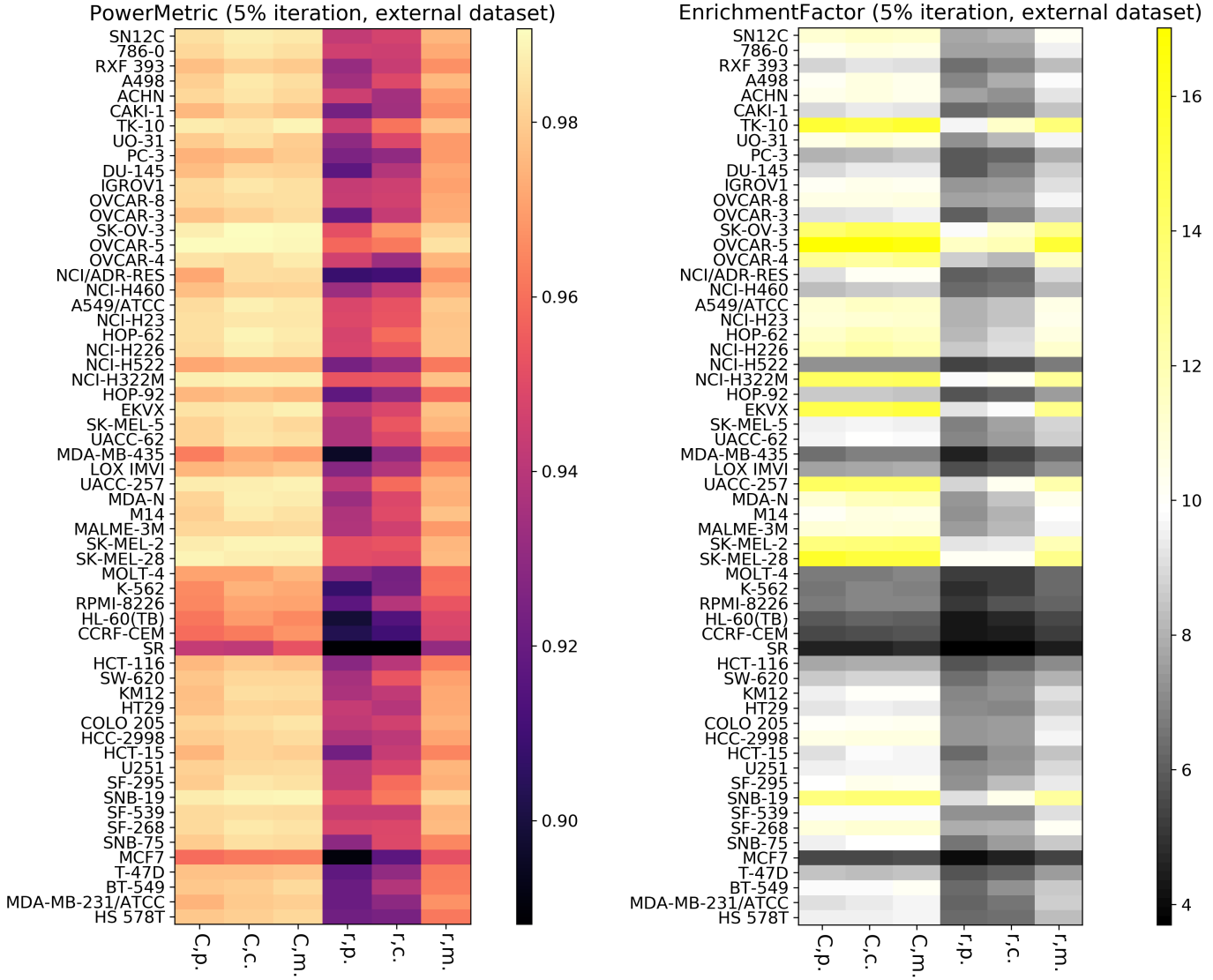


Supplementary Figure 8: Experiments modeling HCT-15 by MACCS and ECFP descriptions.

HCT-15 was also used for a comparison of performance between MACCS and 4096-bit ECFPr2 descriptors because it was one of the cell lines having less favorable predictability in terms of MCC or F1. [Top] Performance of active learning models on external data and training data for the MCC and F1 metrics using ECFP descriptors with each of the three pickers. [Bottom] Similar to the SK-MEL-2 line, the gain over MACCS is a few percentage points.

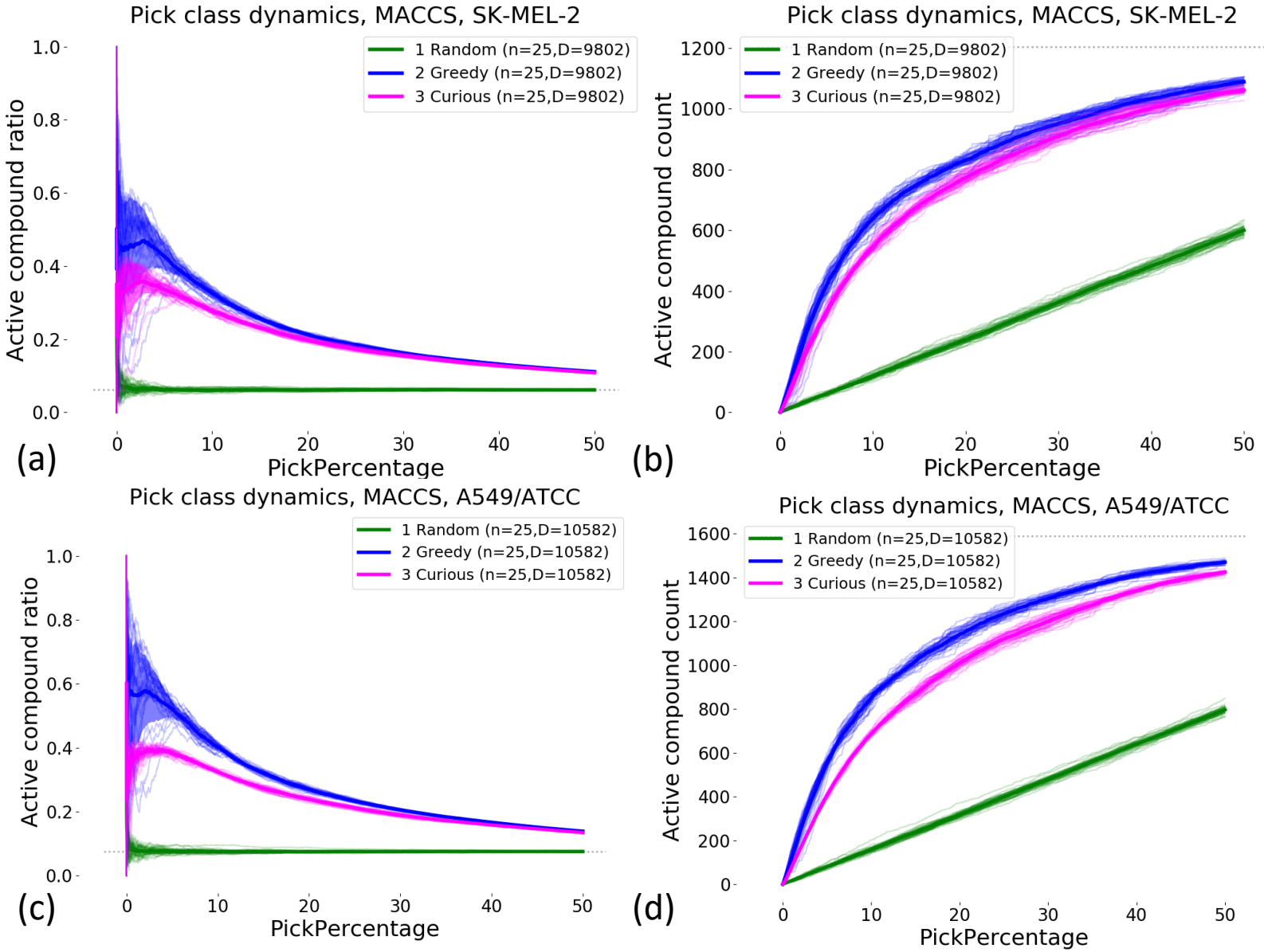






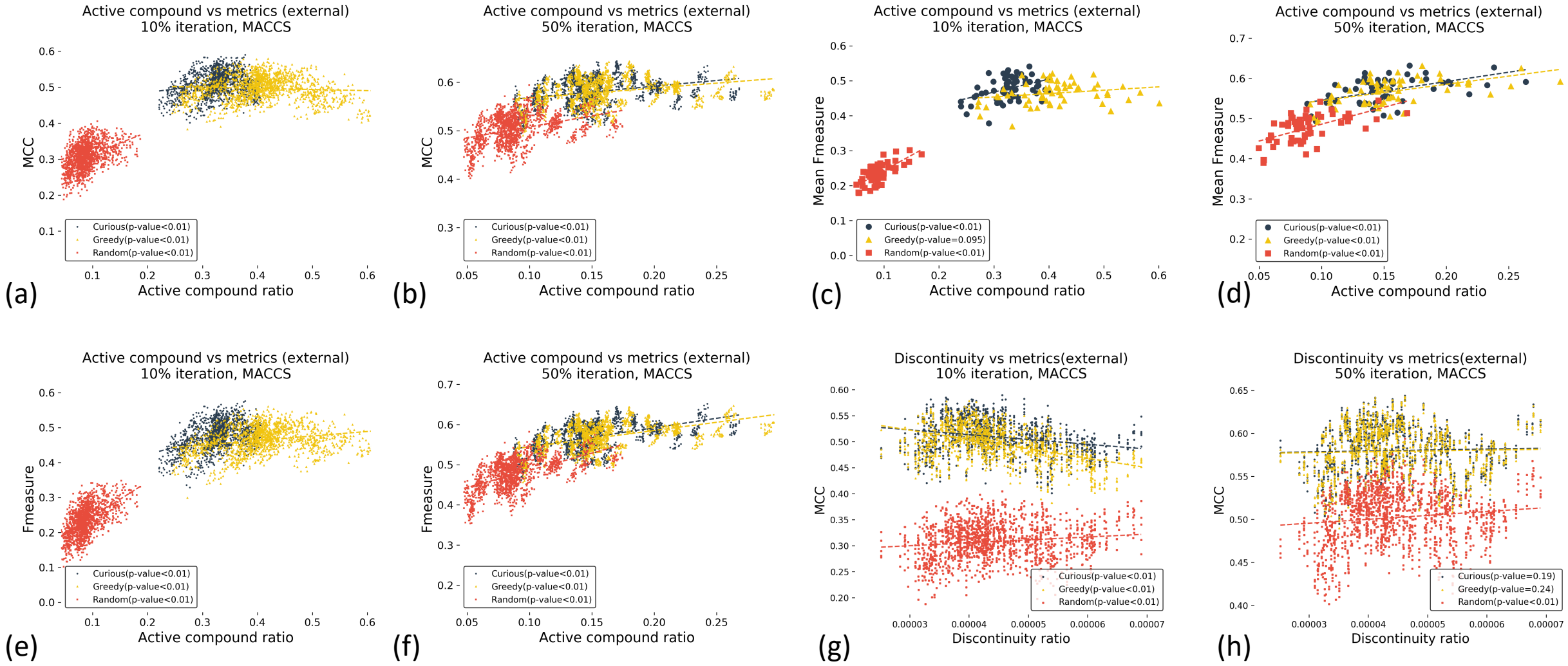
Supplementary Figure 9: Power Metric and Enrichment Factor evaluation of AL.

[Left] Power Metric and Enrichment Factor values when predicting the training and external datasets of SK-MEL-2 melanoma, as a function of data volume learned. Note that the maximum enrichment factor possible for the cell line is approximately 16.2. [Right] Evaluation of Power Metric and Enrichment Factor values for all 60 cell lines, after picking 5% of the learnable data (picking by the curiosity method).

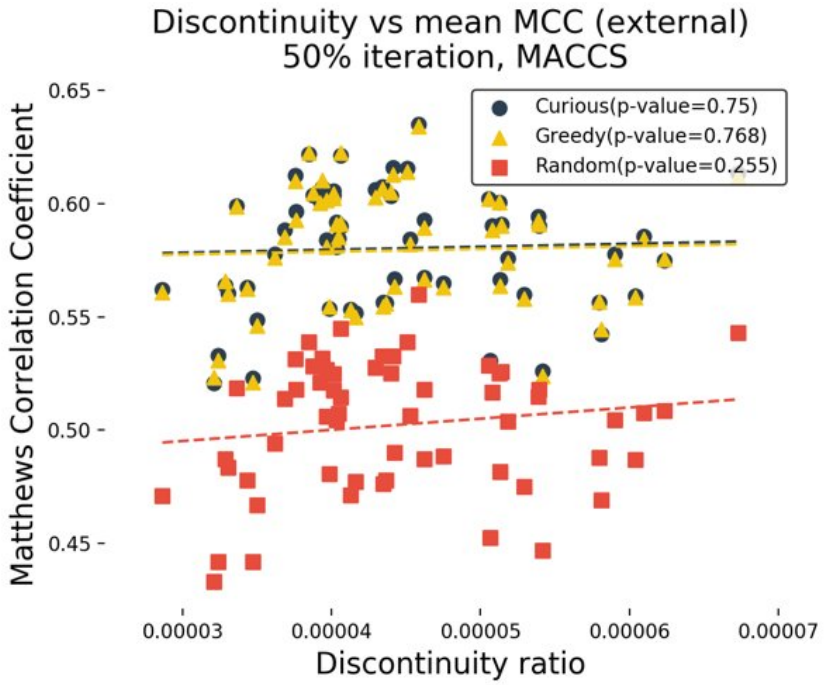


Supplementary Figure 10: Temporal dynamics of active compound ratio.

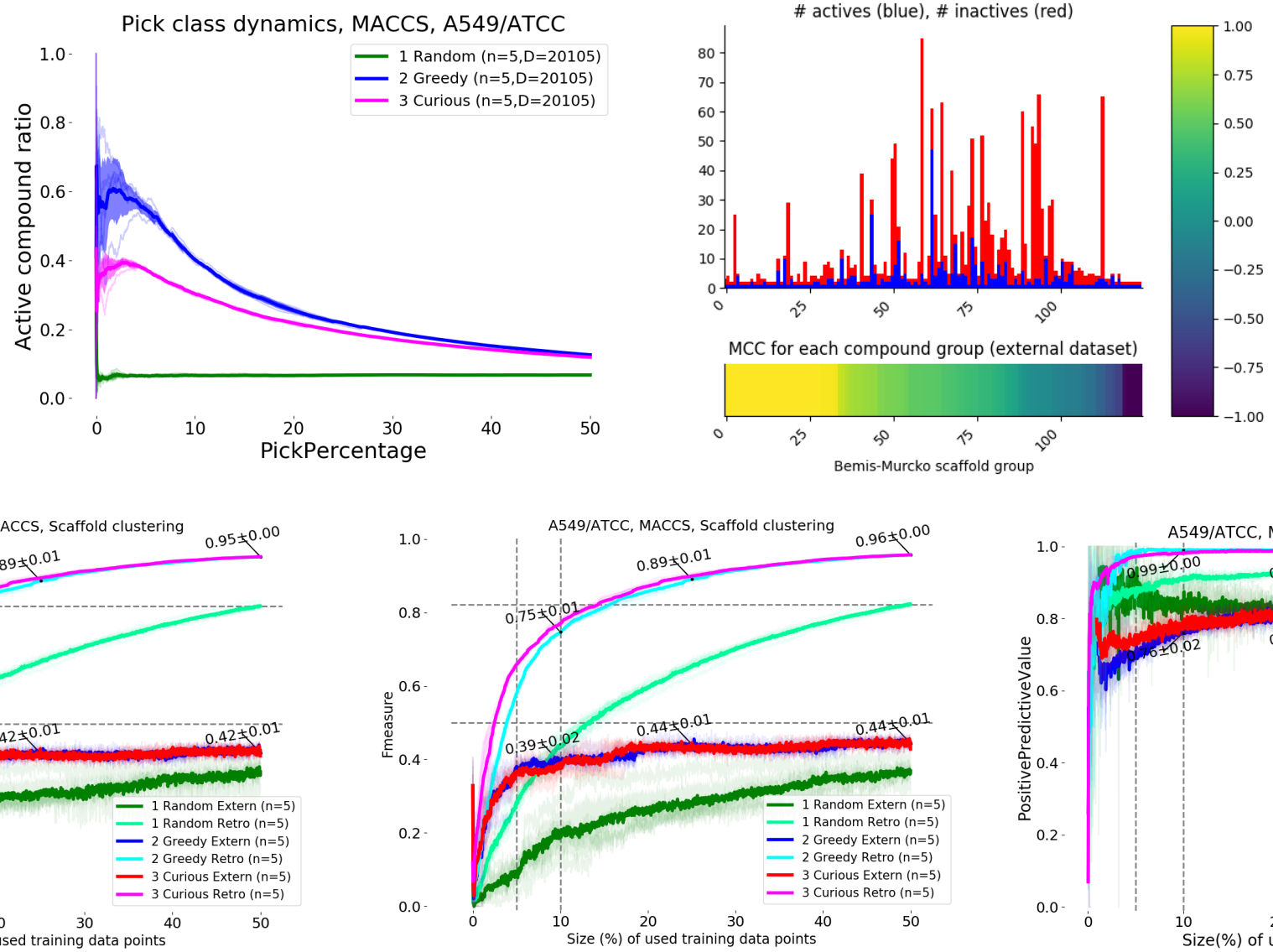
The fraction and actual number of active compounds picked as each active learning picking strategy progresses through iterations of the fit-pick cycle for cell lines SK-MEL-2 (top) and A549/ATCC (bottom). Underlying fractions and counts of actives are shown by gray dotted lines (c.f., Supplementary Figure 1). Early iterations using strategic picking result in models with 40%-60% actives in the model; after 5-10% picks, curiosity picking has added the most informative half of actives; algorithms dominantly add inactive compounds thereafter. Note that models do not make gains in MCC value beyond the 25% data volume (approximately 5000 compounds), for any of the MACCS, ECFP, physicochemical, or CATS2D representations of compounds.



Supplementary Figure 11: Extended analysis of ratio of active compounds selected and model performance.
Panels (a) and (b) show the raw data of Figure 4 for average active pick ratio versus MCC (5 random splits, 5 runs per split). Panels (c) and (d) are analogous versions using the F1 measure, where panels (e) and (f) show the raw data. Panels (g) and (h) show the raw data of Figure 5 for discontinuity versus MCC performance relationships.



Supplementary Figure 12: Discontinuity-performance relationship at 50% data picked.
Against the background of 6% actives in the data (e.g., Figure 1 or Supplementary Figure 1), the trend between cell line discontinuity ratio and MCC is diminished. At 50% picking, strategic pickers have exhaustively picked nearly all actives and incorporated dominant numbers of inactives (Figure 4 and Supplementary Figure 10). Fluctuation in performance is then attributable to the way each cell line responds to chemical treatment, and the resultant discontinuity ratio, though fluctuation appears small across 60 cell lines because of the dominance of inactives in all external datasets.



Supplementary figure 13: Active learning evaluation by scaffold-based data split.

Using the melanoma A549/ATCC cell line, an external dataset was formed by using the lower-performing half of BM frameworks (left panel, right-half of scaffold groups including scaffolds C,E, and F from Figure 6; total of 2116 compounds). The remainder of the compounds and bioactivities formed the training pool to be actively learned and used for predicting cytotoxicity of the external compounds. Evaluation using MCC (middle panel) and F1 (right panel) is shown.