



SCAFFOLD-AWARE PREDICTION OF ESTROGEN RECEPTOR ALPHA INHIBITION WITH UNCERTAINTY-GUIDED CANDIDATE PRIORITIZATION

Ene Abah¹, Anthony Agada Ameh²

¹LFC, Porthacourt, Rivers State, Nigeria. Email: glofahdinfo@yahoo.com

²Nigeria Security and Civil Defence Corps. Email: cimidocentre@gmail.com

Received: 28/11/2025 Revised: 10/01/2026 Acceptance: 12/02/2026 Published: 27/02/2026

ABSTRACT

An approach that relies on computational prioritization might be useful for identifying estrogen receptor alpha (ER α) antagonists, but false positive gains can be experienced when the same molecular scaffolds are found in both training and test sets. This study evaluated the efficacy of scaffold-aware prediction and uncertainty-based candidate prioritization with an ER α antagonist activity set. Following molecular structure standardization and removal of invalid and duplicate molecules, 5,845 unique molecules remain with 255 labelled as positive. The sets of compounds were divided into training (n = 4,465), validation (n = 561), and independent test set (n = 819) where no overlap occurred among these groups by scaffold. Class weighted logistic regression, random forest and extra-trees models were compared with morgan fingerprints. Random forest had the highest validation precision–recall area under the curve (PR-AUC; 0.354) and was then refitted on the combined development set. It performed well on the scaffold-separated test set: PR-AUC = 0.323 (95% bootstrap confidence interval: 0.203–0.473), and an ROC-AUC = 0.840 (0.787–0.887). Precision was 0.545 and recall was 0.150 at the threshold which was selected for validation. Overall test prevalence was 40/819; 6/10 highest ranked test compounds were known positives. Prediction difference between 5 random seeds was averaged and reported as model disagreement but not as calibrated uncertainty. The results also indicate that ranked selection of the candidates for experimental follow-up is appropriate and that classification based on thresholds did not correctly classify many of the known positives. Independent chemical-series validation is still required.

Keywords: estrogen receptor alpha; antagonist activity; scaffold split; molecular fingerprints; random forest; uncertainty; candidate prioritization



1. Introduction

The ESR1 gene encodes estrogen receptor alpha (ER α) which controls the transcription of programs involved in the growth of various hormone receptor positive breast cancers. Although the treatment has improved significantly with the use of endocrine therapies, it is still a major challenge to overcome resistance. Mutations in ESR1 can help maintain receptor activity in the estrogen-depleted state and change the way that receptors respond to therapies that are already in use (Herzog & Fuqua, 2022). ER α -targeted agents are thus clinically relevant for the continued development of ER α inhibitors, especially for ER α -positive disease that has progressed after endocrine therapy (Will et al. 2023).

Selective estrogen receptor modulators, antagonists and degraders are examples of compounds directed to ER α . Their effects are not equivalent; in a cellular assay, a compound can be an antagonist that does not give the same response as a compound that is a agonist. Attempts to find structures that could inhibit ER activity (Rej et al., 2023) persist in the field of medicinal chemistry. Clinical results with elacestrant confirm the therapeutic potential of new ER targeted agents, but also highlight the need for new and new agents to be identified and tested (Bidard et al., 2022). Experimental studies with camizestrant also demonstrate continued efforts to overcome mechanisms of endocrine resistance (Lawson et al., 2023).

Experimental testing of large chemical collections requires a significant amount of resources. A quantitative structure–activity relationship (QSAR) model provides a way of gaining knowledge of the relationships between the structure of a compound and its measured activity, which can then be used to rank compounds for testing. Machine learning has been used to classify ER agonism, antagonism and binding in numerous assays (Ciallella et al., 2021). This distinction becomes important when determining the prediction endpoint: antagonist activity as a measure of cellular signaling doesn't necessarily imply direct receptor binding or biochemical inhibition.

Molecular descriptors and fingerprints and more elaborate representations have been used in recent ER focused models. A machine-learning study was conducted to analyze the structural characteristics of estrogen receptor (ER) agonists, antagonists, and binders (Piir et al., 2024). The SMILES-derived descriptors and predictive models were used together to find potential ER α and ER β antagonists, dubbed as StackER (Schaduengrat et al., 2023). An ER α inhibitor prediction deep-learning algorithm was also developed using graphs (Pusparini et al., 2023). The results presented here reflect the breadth of modeling approaches available, but must be compared with caution, taking into account variations in assay end point, compound collection, class balance, and assay validation design.

One of the issues that may arise is whether a model can be able to rank molecules that are not included in the training set. Structural similarity of compounds in both the training set and the test set can make performance seem more robust than performance on structurally different compounds. Broad benchmarking has revealed that the composition of data, molecular representation, molecular activity cliffs, and evaluation choices all impact molecular property prediction (Deng et al., 2023). Further, evaluations on out-of-distribution compounds highlight the need to consider the chemical space of the compounds in which predictions are needed for selection of the appropriate model (Fooladi et al., 2025). Scaffold separation is one approach to test transfer to unseen molecular cores, but other splitting strategies can bring their own distinct and sometimes more challenging generalization problems (Guo et al., 2025).

Another difficulty in screening for antagonists is the class imbalance, which means that there are a few active compounds in a large population of inactive compounds. In such cases, the number of true antagonists in the top-ranked candidates cannot be described by the receiver operating characteristic performance. The objective of screening is directly met by precision–recall performance and retrieval amongst top ranked compounds. Trust in a ranking also must be nurtured. The scores obtained from different fitted models may differ for the same molecule, and the error bars may be unreliable when the predicted molecules cross over into new scaffolds (Guo, 2026). Since calibration research warns against using an uncalibrated model score as a confirmed probability of activity (Friesacher et al., 2025), the ability to use the model score as a probability of activity is dependent on calibrating the



model.

In the present study, a scaffold-aware classification framework for compounds classified by an ER α antagonist signaling assay is developed. Conventional machine-learning models are compared by their molecular fingerprint and the model chosen is tested on compounds not included in the scaffold groups. In addition to discrimination and threshold-based classification, the study analyzes retrospective recovery of known antagonists from compounds with high scores as well as the variation in scores for repeated model fits. This design compares the ability of molecular structure to support candidate prioritization while maintaining an interpretation anchored to the measured cellular antagonist endpoint without restricting the design to scaffold separation.

2. Materials and Methods

2.1 Dataset and molecular structure curation

The secondary data was used in the analysis, which is 6,329 compounds screened for estrogen receptor alpha (ER α) antagonist activity (Chavan, 2026). Molecular structures were included in the file in SMILES, binary activity, InChI string, and InChIKey formats. A 1 rating indicated antagonist activity in the assay, 0 indicated inactive.

The molecular structures were parsed and standardized using RDKit was used to parse and standardize molecular structures. For every valid molecule, the biggest fragment was kept, the molecule was neutralized, and a canonical isomeric SMILES was produced. Five records were dropped due to inability to obtain a usable standardized structure. Structures with conflicting activity labels were not included, and duplicate standardized structures were not included. After these steps, 479 additional records were removed leaving 5,845 unique compounds (255 of them labelled antagonists and 5,590 of them labelled inactive compounds).

2.2 Molecular representation and scaffold grouping

A 2,048-bit Morgan fingerprint of radius 2 was used to represent each curated structure. Standardized molecular structures were used to derive Bemis–Murcko scaffolds which were then used to define groups for partitioning. Those compounds for which a Bemis–Murcko scaffold was not generated were given just one acyclic group. This acyclic group was part of a curated dataset with 1,732 scaffold groups.

2.3 Scaffold-separated data partitioning

Compounds were scaffolded to ensure that compounds did not end up in different partitions. Based on a target test fraction of 20% a group shuffle procedure yielded 400 candidate partitions. The deterministic selection rule that returned partitions with a size and antagonist prevalence similar to the complete curated data set also maintained a wide array of scaffold groups. The separation of the development set to the training and validation sets and the separation of the development and held out test sets were first done with random-seed 42, then with random-seed 123.

This left 4,465, 561 and 819 compounds in the training, validation and test partitions, respectively. The corresponding numbers of positive compounds were 191, 24, and 40. The test partition consisted of 347 scaffold groups (without any overlaps in the scaffold-group sets).

2.4 Model development and selection

The class weighted logistic regression model, the random forest model and extra trees were compared based on the same fingerprint features and the same training partition. The logistic regression approach was performed with the liblinear solver and a maximum of 1,000 iterations. Both the random forest and extra-trees were fitted with 300 trees and a minimum of two samples per leaf, class balancing was applied during the fitting.

The selection criterion for models was area under the precision–recall curve (PR-AUC) and models were compared on the scaffold-separated validation partition. Once selected, the model selected was refitted with the combined training and validation data sets before being evaluated with the held-out



test partition.

2.5 Performance assessment and candidate ranking

The discrimination of the held-out test set was evaluated by the PR-AUC and the area under the receiver operating characteristic curve (ROC-AUC). A Brier score was also determined. A classification threshold was chosen that maximised the F1 on the validation partition and then a fixed threshold was used on the held-out predictions. The threshold was used to calculate precision, recall, F1, Matthews correlation coefficient, balanced accuracy, and confusion-matrix counts.

Retrospective candidate prioritisation was evaluated by determining the order of held out compounds by their model scores. The number of known antagonists in the top 10, top 20, top 50 and top 100 ranked compounds were calculated and compared with the prevalence of known antagonists in the test set. These rankings are based on compounds that have previously been labelled with assays; they are not a proof of the experimental validation of new compounds.

2.6 Prediction variability and confidence intervals

The chosen model was retrained five times with different random seeds on the same development partition. The standard deviation of the five held-out scores was determined for each held-out compound to describe the variation across model fits. This measure was interpreted as "disagreement of prediction, rather than a "calibrated uncertainty interval".

The 95% confidence intervals for held-out PR-AUC and ROC-AUC were produced using 2,000 stratified bootstrap resamples of paired test labels and fixed-model predictions. Such intervals are sampling variation within the held out set and do not contain variation resulting from other partitions of scaffolds or development choices of models.

2.7 Ethical considerations

This study was based on an already published publicly available chemical bioactivity database and did not require recruitment of human or animal subjects. Nothing personal or identifiable information was processed. The results are computational predictions and comparisons with the labels on the assays and are not evidence of clinical efficacy or experimental confirmation of new ER α inhibitors.

3. Results

3.1 Dataset curation and scaffold partitions

A total of 6329 records were provided, with 289 positive and 6040 negative labels. The entries were molecularly parsed and standardized, but five were not retained. After standardization to largest neutral fragment, removal of conflicting structural labels and deduplication, 5,845 compounds were left, of which 255 (4.36%) were ER α antagonists and 5,590 were inactive compounds. There were 1,732 Murcko scaffold groups in the curated set, of which one group was shared for the 1,566 acyclic structures. The numbers of partitions and classes are shown in Table 1 and Figure 1.

Table 1. Curation and scaffold-separated partitions of the ER α antagonist dataset

Partition	Compounds	Antagonists	Inactive	Antagonists (%)	Scaffold groups
Supplied file	6,329	289	6,040	4.57	—
Curated set	5,845	255	5,590	4.36	1,732
Training	4,465	191	4,274	4.28	1108



Partition	Compounds	Antagonists	Inactive	Antagonists (%)	Scaffold groups
Validation	561	24	537	4.28	277
Held-out test	819	40	779	4.88	347

Note. Training and validation together form the development set ($n = 5,026$). There were no scaffold groups shared between development and test. The 484 removed records include five invalid or empty standardized structures and 479 duplicate or conflicting structure records; 13 standardized structures had contradictory labels.

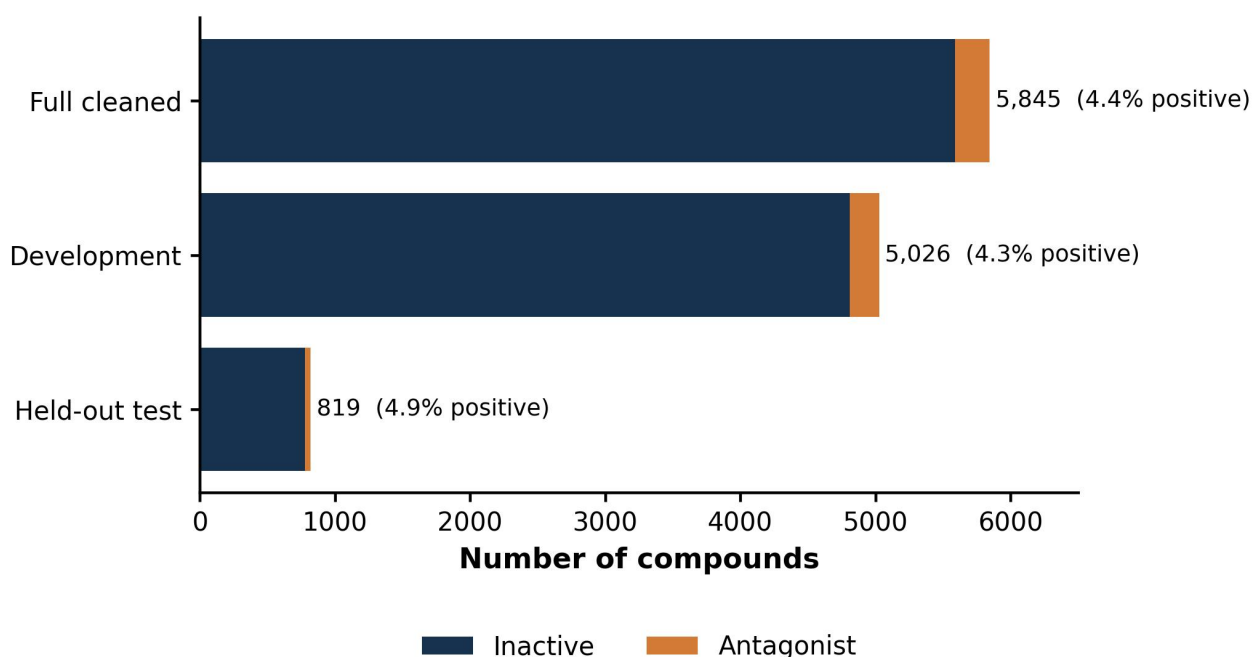


Figure 1. Inactive and antagonist counts, development partition, and scaffold-held-out test partition

The 40 antagonists and 347 scaffold groups were in the 819 held-out test partition. There were 561 compounds in the validation set and 24 of them were positives. There was no transfer across these partitions for scaffold groups, and the evaluation is based on transfer to unseen scaffold groups.

3.2 Validation model selection

Three models were trained on 2,048-bit radius-2 Morgan fingerprints and validated with the same training/validation partition that was used for the scaffold separation. Among these, the random forest model achieved the best validation average precision (PR-AUC), which was 0.354, just over the extra trees model (0.347), and logistic regression achieved 0.126. Table 2 and Figure 2 show the discrimination and Brier scores corresponding to the different values. Random forest was selected and then evaluated on the held-out set.

Table 2. Scaffold-separated validation results used to select the prediction model

Model	PR-AUC	ROC-AUC	Brier score
-------	--------	---------	-------------



Model	PR-AUC	ROC-AUC	Brier score
Logistic regression	0.126	0.675	0.086
Random forest	0.354	0.843	0.051
Extra trees	0.347	0.847	0.050

Note. A higher PR-AUC and ROC-AUC and a lower Brier score is desirable. PR-AUC was the previously stated selection metric and validation prevalence was 24/561 (4.28%).

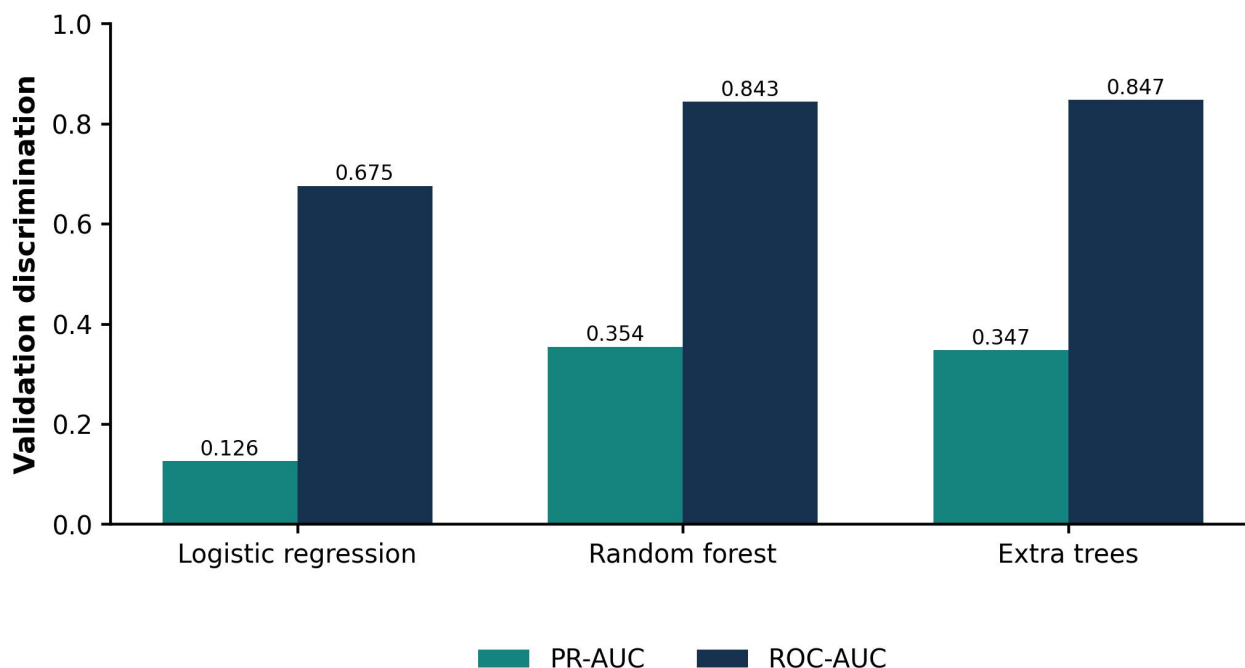


Figure 2. Precision–recall and ROC areas for the three models on the scaffold-separated validation partition

3.3 Held-out scaffold performance

After refitting the selected random forest on all 5,026 development compounds, the model obtained a test PR-AUC of 0.323 (95% bootstrap CI 0.203–0.473) and a ROC-AUC of 0.840 (95% CI 0.787–0.887). Prevalence in the test set was 0.049, and the precision–recall baseline and both holdout curves are shown in Figure 3. The Brier score was 0.051, while a constant test-prevalence score would have been 0.046. The model was not a probability calibrated model and the output is considered a ranking score, not a substantiated probability.

The test confusion matrix at the optimum threshold of 0.477 (which maximized F1 on the validation set) was: True Negative= 774, False Positive= 5, False Negative= 34, True Positive= 6. Precision was 0.545, recall 0.150, F1 0.235, and MCC 0.269 (Table 3). However, only 6 of 40 known antagonists were recovered thus, useful ranking discrimination was achieved.

Table 3. Performance of the selected random forest on the unseen-scaffold test partition

Metric	Held-out result
Test compounds (positive / negative)	819 (40 / 779)



Metric	Held-out result
PR-AUC (95% bootstrap CI)	0.323 (0.203–0.473)
ROC-AUC (95% bootstrap CI)	0.840 (0.787–0.887)
Brier score	0.051
Validation-selected threshold	0.477
Precision / recall / F1	0.545 / 0.150 / 0.235
MCC / balanced accuracy	0.269 / 0.572
TN / FP / FN / TP	774 / 5 / 34 / 6

Note. Confidence intervals are 2,000 resamples of fixed test predictions which do not account for uncertainty stemming from the choice of scaffold split or the training of the model.

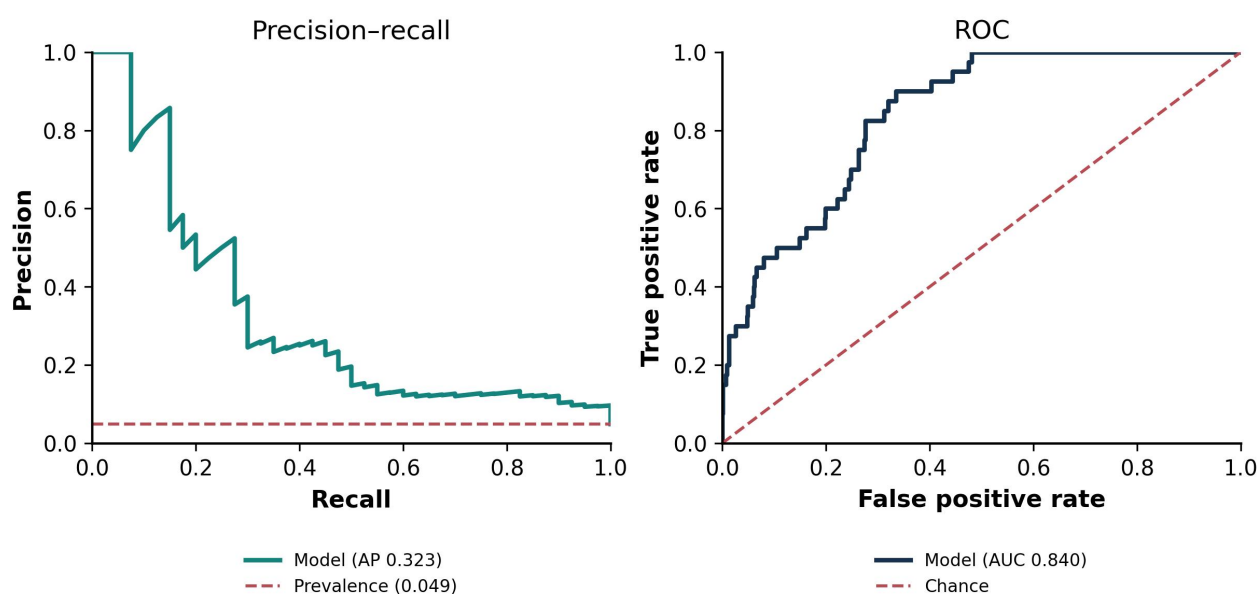


Figure 3. Held-out precision–recall and ROC curves for the random forest. Dashed lines indicate the positive prevalence and chance ROC reference, respectively

3.4 Retrospective candidate prioritization and uncertainty

Six known antagonists were ranked among the top 10 holds out, and 10 among the top 20, compared to an overall prevalence of 4.88%. Precision at 50 and 100 were 0.26 and 0.19, respectively (Table 4; Figure 4A). These are a selection of retrieval results of known compounds that have already been assayed, not new compounds. The accompanying prediction table provides the top-ranked records and identifiers.

Table 4. Retrospective antagonist enrichment among the highest-ranked unseen-scaffold compounds

Top K	Known antagonists	Precision at K	Enrichment over test prevalence	Mean ensemble SD
10	6	0.600	12.29×	0.024



Top K	Known antagonists	Precision at K	Enrichment over test prevalence	Mean ensemble SD
20	10	0.500	10.24×	0.026
50	13	0.260	5.32×	0.025
100	19	0.190	3.89×	0.023

Note. Across-seed SD: Standard deviation of five independently seeded random forests trained on the same development partition. Small SD indicates degree of agreement between these fits, but not a calibrated uncertainty interval or evidence of experimental activity.

The median across-seed score standard deviation for all 819 test compounds was 0.015 and the mean was 0.016 (Figure 4B). The top 10 mean SD was 0.024. These values measure how sensitive the random forest seed is to a particular dataset and split, and do not necessarily reflect assay noise, chemistry outside of the dataset or uncertainty in alternative partitions of the scaffold.

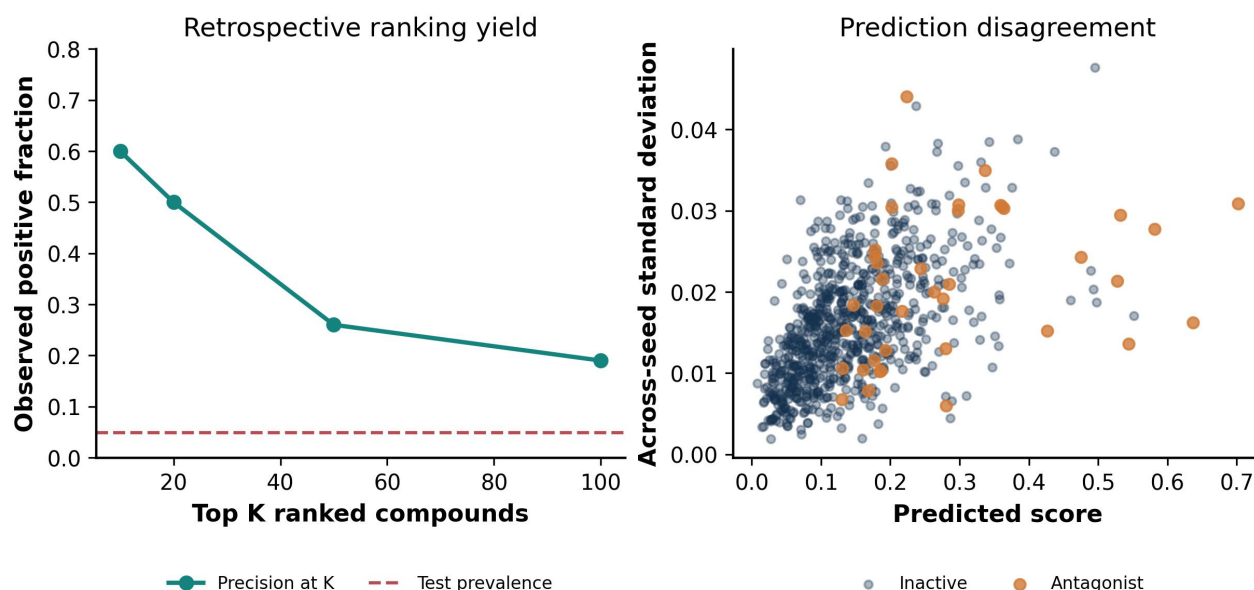


Figure 4. (A) Observed antagonist fraction among the top ranked test compounds, with the test prevalence as reference. (B) Model score versus across-seed prediction standard deviation for held-out compounds, colored by observed label

4. Discussion

The evaluation revealed that molecular fingerprints can be useful for retrospective ranking of the responses of ER α antagonist assays with significant potential for improvement, when assessed using the scaffold-separation approach. The random forest selected had a held-out PR-AUC of 0.323 for a positive prevalence of 0.049 and a ROC-AUC of 0.840. There were 6 of 10 of the top compounds that were known antagonists. These results can be used for ranking compounds for further screening, but do not prove activity in untested compounds. More previous work on ER α based on QSAR, pharmacophore and structure-based approaches has shown the utility of computational screening and also involved experimental evaluation to determine biological effects (Kurtanović et al., 2022; Akinola et al., 2023).

These results should be interpreted in the context of the endpoint selected. The labels do not differentiate all mechanisms that can decrease the assay signal, and do not indicate whether the interference is due to an inhibitor or an antagonist. It is not only the context of the assays that the



models are sensitive to, but also whether the responses are agonist or antagonist, as demonstrated in studies using single-cell measurements (Aghayev et al., 2023a). The same input type (cellular features) and different experimental conditions were used in an ANN study that also categorized responses to EDs (Aghayev et al., 2023b). Thus, the value of the present scores should not be taken as direct binding affinities or biochemical inhibition constants, nor as proof of anticancer effect.

While there was no overlap between the scaffold groups in the development and test partitions, scaffold separation does not necessarily imply prospective generalization. Closely related compounds can show vastly different responses with structurally different scaffolds while there are local features that can be preserved across structurally similar scaffolds. Van Tilborg et al. (2022) demonstrated the weaknesses in molecular prediction using activity cliffs and López-Pérez and Miranda-Quintana (2025) focused on the distribution of activity cliffs across partitions and its impact on the modeling of bioactivities. This one split also has a split which leaves people wondering about performance with other splits and chemical collections. Data definition and data validation design are important to take into consideration in the context of algorithm choice (Tropsha et al., 2024).

There was a clear difference between ranking and classification. When the threshold was set as validation, precision was 0.545, and recall was 0.150, meaning that 34 of 40 known test antagonists were missed. This can lead to a potentially low number of compounds to follow up on, but exclude many active compounds if the threshold set is high. The five-seed standard deviation showed variation in the fitted random forests; agreement among these fits does not mean that a score is reliable for a new scaffold. Model disagreement that is not formalized uncertainty is different from the calibrated coverage (Li et al., 2024; Lanini et al., 2024). Conformal prediction can be useful to present uncertainty from future screening, but its guarantees require the validity of some assumptions that can be checked by chemical distribution shift (Astigarraga et al., 2025).

The clinical importance of the advances in ER α targeting is still significant. Experimental evidence indicates that tumor state and expression of ER α can be altered by mutations of ESR1 and therapeutic pressure, which complicates the expectations that can be derived from a single assay response (Liang et al., 2025). In recent years, several hierarchical models have also considered various mechanisms of endocrine disruption, not only as a unique effect of estrogen (Zhang et al., 2026). The results validate the importance of testing prioritized molecules using orthogonal assays and testing their selectivity, mechanism and cellular effects before moving forward with an assertion of therapeutic activity.

The present model provides a replicable baseline for the retro-looking prioritization of candidates in unseen scaffold groups. However, its use is limited by low threshold recall, uncalibrated scores, the use of a single scaffold partition and a single cellular antagonist endpoint. To establish if the ranking advantage observed holds true for novel candidates, further independent chemical collections and follow-up experimental studies are required.

5. Conclusion

In this study, whether molecular structure could help in prioritization of compounds, when predictions are tested on a new set of scaffold groups was evaluated. Following structure standardization and elimination of invalid and duplicates, 5,845 unique compounds were analyzed, with 255 being labelled as active. To ensure that the same group was not found in both the model development set and the independent test set, the separation of the scaffold groups was used. Random forest was the model with the best validation precision–recall area under the curve (PR-AUC) and was chosen for final evaluation.

The model had a PR-AUC of 0.323, and a receiver operating characteristic area under the curve of 0.840, on the 819-compound test set. Of the top 10, 6 were known positives while 40 of the 819 test compounds were positives. This enrichment is used to aid the ordering of candidates for further investigation when using the model. Validation-selected classification threshold performance was more restricted, as evidenced by the recall of 0.150 and the fact that 34 of 40 positives were missed. The study therefore backs up the ranking of candidates, rather than reliable rejection of compounds



with low scores. The range of predictions from repeated model seeds was used to provide a measure of model disagreement but it is not calibrated as a measure of error. Prioritized compounds should not be considered as validated ER α antagonists without prospective testing on independent chemical series and experimental testing.

References

1. Aghayev, Z., Szafran, A. T., Tran, A., Ganesh, H. S., Stossi, F., Zhou, L., ... & Beykal, B. (2023a). Machine learning methods for endocrine disrupting potential identification based on single-cell data. *Chemical engineering science*, 281, 119086.
2. Aghayev, Z., Walker, G. F., Iseri, F., Ali, M., Szafran, A. T., Stossi, F., ... & Beykal, B. (2023b). Binary classification of the endocrine disrupting chemicals by artificial neural networks. In *Computer Aided Chemical Engineering* (Vol. 52, pp. 2631-2636). Elsevier.
3. Akinola, L. K., Uzairu, A., Shallangwa, G. A., & Abechi, S. E. (2023). Development and validation of predictive quantitative structure–activity relationship models for estrogenic activities of hydroxylated polychlorinated biphenyls. *Environmental Toxicology and Chemistry*, 42(4), 823-834.
4. Astigarraga, M., Sánchez-Ruiz, A., & Colmenarejo, G. (2025). Conformal prediction-based machine learning in cheminformatics: Current applications and new challenges. *Artificial Intelligence in the Life Sciences*, 7, 100127.
5. Bidard, F. C., Kaklamani, V. G., Neven, P., Streich, G., Montero, A. J., Forget, F., ... & Bardia, A. (2022). Elacestrant (oral selective estrogen receptor degrader) versus standard endocrine therapy for estrogen receptor–positive, human epidermal growth factor receptor 2–negative advanced breast cancer: results from the randomized phase III EMERALD trial. *Journal of Clinical Oncology*, 40(28), 3246-3256.
6. Chavan, S. (2026). Cleaned dataset for PubChem AID 743078: qHTS assay of estrogen receptor alpha (ER-alpha) antagonists (Version 1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.20271538>
7. Ciallella, H. L., Russo, D. P., Aleksunes, L. M., Grimm, F. A., & Zhu, H. (2021). Predictive modeling of estrogen receptor agonism, antagonism, and binding activities using machine-and deep-learning approaches. *Laboratory investigation*, 101(4), 490-502.
8. Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., & Wang, F. (2023). A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14(1), 6395.
9. Fooladi, H., Vu, T. N. L., Mathea, M., & Kirchmair, J. (2025). Evaluating machine learning models for molecular property prediction: performance and robustness on out-of-distribution data. *Journal of Chemical Information and Modeling*, 65(19), 9871.
10. Friesacher, H. R., Engkvist, O., Mervin, L., Moreau, Y., & Arany, A. (2025). Achieving well-informed decision-making in drug discovery: a comprehensive calibration study using neural network-based structure-activity models. *Journal of Cheminformatics*, 17(1), 29.
11. Guo, Q., Hernandez-Hernandez, S., & Ballester, P. J. (2025). UMAP-based clustering split for rigorous evaluation of AI models for virtual screening on cancer cell lines. *Journal of Cheminformatics*, 17(1), 94.
12. Guo, Y. (2026). Scaffold-shift uncertainty calibration in molecular activity prediction: A multi-target benchmark of coverage, efficiency, and risk-aware selection. *Computational Biology and Chemistry*, 109371.
13. Herzog, S. K., & Fuqua, S. A. (2022). ESR1 mutations and therapeutic resistance in metastatic breast cancer: progress and remaining challenges. *British journal of cancer*, 126(2), 174-186.
14. Kurtanović, N., Tomašević, N., Matic, S., Proia, E., Sabatino, M., Antonini, L., ... & Ragno, R. (2022). Human estrogen receptor alpha antagonists, part 3: 3-D pharmacophore and 3-D QSAR guided brefeldin A hit-to-lead optimization toward new breast cancer suppressants. *Molecules*, 27(9), 2823.
15. Lanini, J., Huynh, M. T. D., Scebbba, G., Schneider, N., & Rodríguez-Pérez, R. (2024). UNIQUE:



- A framework for uncertainty quantification benchmarking. *Journal of Chemical Information and Modeling*, 64(22), 8379.
16. Lawson, M., Cureton, N., Ros, S., Cheraghchi-Bashi, A., Urosevic, J., D'Arcy, S., ... & Carnevalli, L. S. (2023). The next-generation oral selective estrogen receptor degrader camizestrant (AZD9833) suppresses ER⁺ breast cancer growth and overcomes endocrine and CDK4/6 inhibitor resistance. *Cancer research*, 83(23), 3989-4004.
17. Li, P., Hua, L., Ma, Z., Hu, W., Liu, Y., & Zhu, J. (2024). Conformalized graph learning for molecular admet property prediction and reliable uncertainty quantification. *Journal of Chemical Information and Modeling*, 64(23), 8705-8717.
18. Liang, J., Yao, X., Aouad, P., Wang, B. E., Crocker, L., Chaudhuri, S., ... & Metcalfe, C. (2025). ER α dysfunction caused by ESR1 mutations and therapeutic pressure promotes lineage plasticity in ER⁺ breast cancer. *Nature Cancer*, 6(2), 357-371.
19. López-Pérez, K., & Miranda-Quintana, R. A. (2025). Extended Activity Cliffs-Driven Approaches on Data Splitting for the Study of Bioactivity Machine Learning Predictions. *Molecular informatics*, 44(1), e202400054.
20. Piir, G., Sild, S., & Maran, U. (2024). Interpretable machine learning for the identification of estrogen receptor agonists, antagonists, and binders. *Chemosphere*, 347, 140671.
21. Pusparini, R. T., Krisnadhi, A. A., & Firdayani. (2023). MATH: A deep learning approach in QSAR for estrogen receptor alpha inhibitors. *Molecules*, 28(15), 5843.
22. Rej, R. K., Thomas, J. E., Acharyya, R. K., Rae, J. M., & Wang, S. (2023). Targeting the estrogen receptor for the treatment of breast cancer: recent advances and challenges. *Journal of Medicinal Chemistry*, 66(13), 8339-8381.
23. Schaduangrat, N., Homdee, N., & Shoombuatong, W. (2023). StackER: a novel SMILES-based stacked approach for the accelerated and efficient discovery of ER α and ER β antagonists. *Scientific Reports*, 13(1), 22994.
24. Tropsha, A., Isayev, O., Varnek, A., Schneider, G., & Cherkasov, A. (2024). Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nature Reviews Drug Discovery*, 23(2), 141-155.
25. Van Tilborg, D., Alenicheva, A., & Grisoni, F. (2022). Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling*, 62(23), 5938.
26. Will, M., Liang, J., Metcalfe, C., & Chandarlapaty, S. (2023). Therapeutic resistance to anti-oestrogen therapy in breast cancer. *Nature Reviews Cancer*, 23(10), 673-685.
27. Zhang, R., Chang, B., Tan, H., Guo, J., Zhang, X., Yu, H., & Shi, W. (2026). Mechanism Based Hierarchical Machine Learning for High-Throughput Quantitative Prediction of Estrogenic, Androgenic, and Thyroid Disruption Activities. *Environmental Science & Technology*, 60(5), 4184-4200.