COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

BACHELOR'S DEGREE IN MATHEMATICAL ENGINEERING AND AI

FINAL DEGREE PROJECT

AI for biochemistry applications

Author: María Fernanda Marcos Gámez

Director: Simón Rodríguez Santana

Co-Director: Jaime Pizarroso Gonzalo

Madrid, June 2026

Declaration of originality

I declare under my responsibility that the Project presented with the title **AI for biochemistry applications** at the ICAI School of Engineering of the Comillas Pontifical University in the academic year 2025/2026 is of my authorship and has not been presented previously for other purposes. The Project is not plagiarised from any other, either totally or partially, and the information that has been taken from other documents is duly referenced.

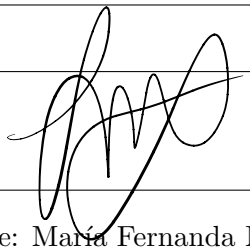
Use of Artificial Intelligence[1]

I declare under my responsibility (indicate the correct option):

☐ I have not used Artificial Intelligence in the preparation of this document.

☒ I have used Artificial Intelligence in the preparation of this document and/or Annex B under the conditions allowed by Comillas Pontifical University, i.e. applying Level 2 of the Perkins et al. (2024) Assessment Scale: *"AI can be used for pre-task activities such as brainstorming, description and initial research. This level focuses on the use of AI for planning, synthesising and generating ideas, but assessments should emphasise the ability to develop and refine these ideas independently"*. Specifically, Artificial Intelligence has been used to:

1. Conduct research and summarize literature
2. Generate code
3. Rephrase or summarize ideas to improve readability of this work



Signature: María Fernanda Marcos Gámez

Date: 15/06/2026

Authorisation for Project delivery

[1]This declaration refers to the use of generative Artificial Intelligence to carry out the Project documents (Annex B and Memory). It does not apply to Projects where, by their nature, artificial intelligence must be used as part of them (application of machine learning techniques, neural networks, data analysis...).

Thesis Supervisor	Thesis Deputy Supervisor (if applicable)
Firmado por RODRIGUEZ SANTANA SIMON - ***4141** el día 16/06/2026 con un certificado emitido por AC FNMT Usuarios	Jaime Pizarroso Gonzalo Firmado digitalmente por Jaime Pizarroso Gonzalo Fecha: 2026.06.16 11:02:25 +02'00'
Signature: Simón Rodríguez Santana	Signature: Jaime Pizarroso Gonzalo
Date: 15/06/2026	Date: 15/06/2026

Acknowledgments

To my parents for giving me this opportunity,

and to my friends for being there every step of the way.

INTELIGENCIA ARTIFICIAL PARA APLICACIONES BIOQUÍMICAS

Autor: María Fernanda Marcos Gámez

Director: Simón Rodríguez Santana

Co-Director: Jaime Pizarroso Gonzalo

Resumen

Este trabajo desarrolla un marco de aprendizaje profundo explicable para la predicción de toxicidad molecular en los datos de Tox21, y lo aplica para responder una pregunta concreta: ¿qué subestructuras moleculares impulsan las predicciones de toxicidad? Se entrena un conjunto de redes neuronales profundas de una sola tarea, utilizando descriptores moleculares y cinco semillas aleatorias, y se interpreta mediante permutation importance y SHAP. Las subestructuras de consenso obtenidas se contrastan con la literatura de química medicinal, recuperando varios toxicóforos documentados (furano, tiofeno, grupos nitro, azufre) y generando listas de candidatos a toxicóforos anotadas químicamente para cada endpoint.

Palabras clave: Predicción de toxicidad; XAI; SHAP; Tox21; fingerprints de Morgan; redes neuronales profundas; importancia por permutación de clústeres; sensibilidad neuronal.

Resumen ejecutivo

1 Introducción

La inteligencia artificial ya puede diseñar nuevas moléculas con actividad biológica muy afinada para un objetivo concreto. Sin embargo, la toxicología computacional no ha progresado al mismo ritmo. Esto crea un problema: una molécula que funciona bien para su diana principal puede también unirse a otros receptores no deseados o activar vías biológicas que provocan toxicidad. Este trabajo se centra en predecir la toxicidad usando el conjunto de datos Tox21 [1], un estándar de referencia en cribado toxicológico.

Los modelos de predicción de toxicidad más avanzados se comportan como "cajas negras": una predicción de tipo "tóxico" no ofrece información sobre qué características estructurales la han generado, ni si el modelo ha aprendido una regla químicamente coherente o una tendencia del conjunto de datos. Los toxicólogos necesitan saber *qué* subestructuras son señaladas y *si* corresponden a mecanismos de toxicidad conocidos, y los reguladores exigen cada vez más esta transparencia antes de que los modelos de IA puedan utilizarse en evaluación de

riesgo. Este trabajo está motivado por la necesidad de cerrar la brecha entre el rendimiento predictivo y la interpretabilidad en la predicción de toxicidad molecular.

2 Objetivos

La pregunta central es: **¿qué subestructuras moleculares impulsan las predicciones de toxicidad y son estas señales genuinas y reproducibles en lugar de artefactos de una única ejecución de entrenamiento?** Los objetivos son (1) desarrollar un predictor de toxicidad explicable para cada endpoint de Tox21 con explicaciones estables y químicamente significativas a través de distintas semillas, y (2) integrar la importancia por permutación a nivel de clúster [2], SHAP [3] y el análisis de sensibilidad [4] para identificar estructuras asociadas a la toxicidad, validados mediante enriquecimiento frente a datos reales y la literatura de toxicóforos.

3 Descripción del modelo

Las moléculas del conjunto de datos Tox21 se representan mediante vectores binarios de 2048 bits, construidos a partir de subestructuras circulares locales de radio 2 alrededor de cada átomo. Esta representación codifica la presencia o ausencia de patrones estructurales específicos dentro de la molécula. Este tipo de descriptor se conoce como *Morgan fingerprints* [5]. El conjunto de datos Tox21 está compuesto por doce variables objetivo, cada una de las cuales corresponde a un ensayo biológico distinto relacionado con distintos mecanismos de toxicidad, incluyendo receptores nucleares y vías de respuesta al estrés celular, cada uno de estos objetivos se denomina *endpoint*. Para cada *endpoint*, los datos se dividen 80/10/10 en train, validation y test. Tres modelos clásicos (regresión logística, random forest y gradient boosting) fueron entrenados por endpoint como puntos de referencia. El modelo principal es una red neuronal profunda de una sola tarea (tres capas ocultas de 1024, 512 y 256 unidades, con batch normalization, dropout y activaciones LeakyReLU), basada en una arquitectura de la literatura de predicción de toxicidad [6]. Se entrenó un modelo independiente para cada uno de los 12 endpoints a lo largo de 5 semillas aleatorias, obteniendo un total de 60 modelos, con el objetivo de evaluar la estabilidad tanto de las predicciones como de las explicaciones.

Trabajar directamente con 2048 bits produce explicaciones inestables entre semillas. Para resolverlo, los bits se agruparon en 300 clusters [7] mediante clustering jerárquico aglomerativo [8] sobre una matriz de distancia basada en correlación, de forma que bits que co-activan en moléculas se agrupan en conjuntos químicamente coherentes. Posteriormente se aplicaron tres métodos de explicación a nivel de cluster en todas las semillas y endpoints: permutation importance, SHAP y análisis de sensibilidad basado en gradientes. La estabilidad entre semillas se midió usando el índice de Jaccard (solapamiento de los top-k clusters más importantes)

y la correlación de Spearman (acuerdo global en el ranking).

Para cada endpoint, los *consensus clusters* se definieron como aquellos que aparecen en el top-50 de SHAP en magnitud para *todas* las semillas. Para cada cluster consenso se calculó un score de *directionality* como la media del valor SHAP con signo sobre las moléculas que activan el cluster (positivo si favorece toxicidad, negativo si tiene efecto protector), y un *enrichment factor* ($\mathrm{EF}_{\mathrm{actual}}$) calculado en el conjunto de test como el ratio entre la proporción de moléculas tóxicas entre las activadas por el cluster y la proporción base de toxicidad. Combinando ambas señales, un cluster se etiqueta como **toxic driver** si $\mathrm{EF}_{\mathrm{actual}} > 1$ y la dirección SHAP es positiva, o como **protective** si ambas son negativas.

Finalmente, cada cluster consenso fue anotado con sus grupos funcionales más enriquecidos (*e.g*), identificados mediante el matching de patrones SMARTS contra las moléculas que activan el cluster y comparando su frecuencia con la frecuencia base del dataset.

4 Resultados

Los modelos *baseline* clásicos lograron una precisión (*accuracy*) alta, pero una baja precisión balanceada (*balanced accuracy*) debido al fuerte desbalance de clases (solo $\sim$5 % de ejemplos positivos por criterio de toxicidad), con Random Forest y Boosting tendiendo a predecir mayoritariamente la clase dominante. El conjunto de redes neuronales profundas, entrenado con una función de pérdida ponderada y un umbral de decisión reducido (0.2 en lugar de 0.5), alcanzó una precisión balanceada media de 0.65 ± 0.09 y un ROC-AUC medio de 0.74 ± 0.09 en todos los criterios de toxicidad, con variabilidad significativa entre ellos y entre semillas aleatorias. Por ejemplo, SR-MMP y NR-AR-LBD muestran una separación fuerte y consistente (ROC-AUC > 0.8), mientras que NR-ER y SR-HSE se mantienen próximos al nivel de azar.

Agrupar los bits del fingerprint mejora sustancialmente la estabilidad de las explicaciones frente al nivel de bit individual. A nivel de cluster, SHAP muestra la mayor estabilidad entre semillas (Jaccard top-50 entre 0.73 y 0.81, correlación de Spearman > 0.93 en todos los endpoints), frente a permutation importance (Jaccard 0.37–0.59) y análisis de sensibilidad (Jaccard 0.41–0.60). Esto justifica el uso de SHAP como base del análisis de consenso y enriquecimiento.

La aplicación del filtro de consenso y enriquecimiento identificó entre 4 y 13 clusters driver por endpoint (y entre 2 y 14 clusters protectores), cada uno anotado con grupos funcionales. Se observan patrones recurrentes: el Cluster 43 (furano; indol; aldehído; tiofeno; piperidina) es el principal driver tóxico en tres endpoints de receptores nucleares (NR-AR-LBD, NR-ER, NR-ER-LBD), con factores de enriquecimiento entre 2.6 y 4.6; el Cluster 33 (indol; sulfoxido; halógeno(F); éter; éster) aparece como driver en dos ensayos de estrés (SR-HSE, SR-MMP), mientras que su contraparte protectora (Cluster 29) muestra EF < 1 en ambos casos.

Varias estructuras recuperados coinciden con alertas estructurales conocidas en la literatura de química medicinal: el furano y el tiofeno son toxicóforos clásicos bioactivados a electrófilos reactivos capaces de formar aductos covalentes con proteínas [9]; el grupo nitro, identificado como principal driver en SR-ATAD5, es uno de los toxicóforos mutagénicos más reconocidos; los motivos con azufre (sulfona, sulfoxido, tioéter) aparecen repetidamente en drivers relacionados con disrupción endocrina. Esta correspondencia proporciona validación externa de que las representaciones aprendidas por el modelo están parcialmente alineadas con reactividad química documentada, en lugar de ser únicamente correlaciones del dataset.

5 Conclusiones

Este trabajo demuestra que la combinación de SHAP a nivel de cluster con validación por enriquecimiento produce una lista interpretable y químicamente coherente de candidatos a toxicóforos por endpoint, varios de los cuales coinciden con mecanismos toxicológicos descritos en la literatura. Las principales contribuciones son: (1) un procedimiento de clustering que mejora sustancialmente la estabilidad de las explicaciones frente a bits individuales; (2) un filtro dual (direccionalidad SHAP y factor de enriquecimiento) que separa drivers reales de clusters sin soporte empírico; y (3) un análisis sistemático cross-endpoint que revela drivers estructurales compartidos entre rutas toxicológicas relacionadas.

Sin embargo, los resultados también muestran limitaciones importantes: el rendimiento predictivo varía considerablemente entre endpoints, el fuerte desbalance de clases limita el número de ejemplos positivos disponibles para validar cada cluster, y las conclusiones sobre clusters individuales dependen de pocas moléculas activadoras y no son generalizables fuera del test set. Los modelos también siguen siendo sensibles a la inicialización aleatoria, lo que indica que se necesitan datasets mayores o representaciones con mayor señal química antes de que este tipo de pipelines pueda utilizarse en evaluación toxicológica de grado regulatorio. Trabajos futuros deberían explorar representaciones moleculares alternativas (graph-based o embeddings aprendidos), incorporación de conocimiento químico como prior del modelo, y enfoques generativos o de búsqueda para validar experimentalmente los toxicóforos candidatos identificados.

6 Referencias

[1] A. M. Richard et al., "The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology", *Chemical Research in Toxicology*, vol. 34, n.º 2, págs. 189-216, feb. de 2021, ISSN: 0893-228X. DOI: 10.1021/acs.chemrestox.0c00264. dirección: https://doi.org/10.1021/acs.chemrestox.0c00264.

[2] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, n.º 1, págs. 5-32, 2001. DOI: 10.1023/A:1010933404324. dirección: https://doi.org/10.1023/A:1010933404324.

[3] S. Lundberg y S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*, 2017. arXiv: 1705.07874 [cs.AI]. dirección: https://arxiv.org/abs/1705.07874.

[4] J. Pizarroso, J. Portela y A. Muñoz, "NeuralSens: Sensitivity Analysis of Neural Networks", *Journal of Statistical Software*, vol. 102, n.º 7, págs. 1-36, 2022. DOI: 10.18637/jss.v102.i07. dirección: https://www.jstatsoft.org/index.php/jss/article/view/v102i07.

[5] D. Rogers y M. Hahn, "Extended-Connectivity Fingerprints", *Journal of Chemical Information and Modeling*, vol. 50, n.º 5, págs. 742-754, 2010, PMID: 20426451. DOI: 10.1021/ci100050t. eprint: https://doi.org/10.1021/ci100050t. dirección: https://doi.org/10.1021/ci100050t.

[6] B. Sharma et al., "Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations", *Scientific Reports*, vol. 13, n.º 1, pág. 4908, mar. de 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-31169-8. dirección: https://doi.org/10.1038/s41598-023-31169-8.

[7] B. Gregorutti, B. Michel y P. Saint-Pierre, "Grouped variable importance with random forests and application to multiple functional data analysis", *Computational Statistics Data Analysis*, vol. 90, págs. 15-35, 2015, ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2015.04.002. dirección: https://www.sciencedirect.com/science/article/pii/S0167947315000997.

[8] J. H. W. Jr., "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association*, vol. 58, n.º 301, págs. 236-244, 1963. DOI: 10.1080/01621459.1963.10500845. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500845. dirección: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

[9] N. Dang, T. B. Hughes, G. Miller y S. J. Swamidass, "Computational Approach to Structural Alerts: Furans, Phenols, Nitroaromatics, and Thiophenes", *Chemical research in toxicology*, vol. 30, págs. 1046-1059, 2017. DOI: 10.1021/acs.chemrestox.6b00336.

AI FOR BIOCHEMISTRY APPLICATIONS

Author: María Fernanda Marcos Gámez

Director: Simón Rodríguez Santana

Co-Director: Jaime Pizarroso Gonzalo

Abstract

This work develops an explainable deep learning framework for predicting molecular toxicity across the twelve Tox21 endpoints, and applies it to answer a specific question: which molecular substructures drive toxicity predictions? A single-task DNN ensemble, trained on Morgan fingerprints across five random seeds, is interpreted using cluster-level SHAP attribution combined with ground-truth enrichment validation. The resulting consensus substructures are cross-checked against the medicinal chemistry literature, recovering several documented toxicophores (furan, thiophene, nitro groups, sulfur-containing motifs) and producing chemically annotated, per-endpoint candidate toxicophore lists.

Keywords: Toxicity prediction; XAI; SHAP; Tox21; Morgan fingerprints; deep neural networks; cluster permutation importance; neural sensitivity

Executive Summary

1 Introduction

Artificial intelligence can now design novel molecules with optimised biological activity, but computational toxicology has not kept pace: a molecule optimised for one biological target may bind off-target receptors or trigger unrelated toxic mechanisms. This work addresses toxicity prediction on the Tox21 dataset [1], a screening benchmark covering twelve assays from two pathways: seven nuclear receptor (NR) endpoints (hormone signalling disruption) and five stress response (SR) endpoints (oxidative stress, heat shock, DNA damage response).

State-of-the-art toxicity predictors behave as "black boxes": a "toxic" prediction gives no insight into which structural features drove it, or whether the model learned a chemically plausible rule versus a dataset artefact. Toxicologists need to know *which* substructures are flagged and *whether* they correspond to known toxicity mechanisms, and regulators increasingly require this transparency before AI models can inform risk assessment. This thesis is motivated by the need to close the gap between predictive performance and interpretability in molecular toxicity prediction.

2 Objectives

The central question is: **which molecular substructures drive toxicity predictions, and are these signals genuine and reproducible rather than artefacts of a single training run?** The objectives are to (1) develop an explainable toxicity predictor for each Tox21 endpoint with stable, chemically meaningful explanations across seeds, and (2) integrate cluster-level permutation importance [2], SHAP [3], and sensitivity analysis [4] to identify structural motifs associated with toxicity, validated against ground-truth enrichment and the toxicophore literature.

3 Description of the Model

Molecules from the Tox21 dataset are encoded as fixed-length binary vectors in which each bit records the presence or absence of a local circular substructure centred on an atom and extending up to a given radius. This representation, known as Morgan fingerprints or extended-connectivity fingerprints (ECFP) [5], is computed here with a radius of 2 and a vector length of 2048 bits. For each endpoint, the data is split 80/10/10 into training, validation, and test sets, preserving class proportions.

Three classical baselines (logistic regression [6], random forest [2], and LightGBM [7] gradient boosting [8]) were trained per endpoint as reference points. The primary model is a fully-connected single-task DNN (three hidden layers of 1024, 512, and 256 units, with batch normalisation, dropout, and LeakyReLU activations), based on an architecture from the toxicity prediction literature [9]. An independent model was trained for each of the 12 endpoints across 5 random seeds, giving 60 trained models in total, to assess the stability of both predictions and explanations.

Working directly with 2048 individual fingerprint bits produces unstable explanations across seeds. To address this, the bits were grouped into 300 clusters [10] via hierarchical agglomerative clustering [11] on a correlation-based distance matrix, so that bits that co-activate across molecules are aggregated into chemically coherent groups. Three explanation methods were then applied at the cluster level across all seeds for each endpoint: permutation importance, SHAP, and sensitivity analysis. Explanation stability across the five seeds was measured using the Jaccard index (overlap of top-k important clusters) and Spearman correlation (global rank agreement).

For each endpoint, *consensus clusters* were defined as those ranked in the top-50 for *every* seed. For each consensus cluster, a *directionality* score was computed as the mean signed SHAP value over molecules that activate the cluster (positive if it is toxic driver, negative if it acts as protective), and an *enrichment factor* was computed on the held-out test set as the ratio between the proportion of true toxic molecules among cluster-active molecules and the

baseline toxic proportion. Combining these two signals, a cluster is labelled a **toxic driver** if EF > 1 and the SHAP direction is positive, or **protective** if both are negative. Finally, each consensus cluster was annotated with its most enriched named functional groups, identified by matching a library of SMARTS patterns against the molecules that activate the cluster and comparing their frequency to the dataset background.

4 Results

The classical baselines achieved high raw accuracy but generally poor balanced accuracy due to the strong class imbalance (only $\sim$5% positive examples per endpoint), with random forest and boosting models frequently defaulting toward the majority class. The DNN ensemble, trained with class-weighted loss and a lowered decision threshold (0.2), achieved a mean balanced accuracy of 0.651 ± 0.092 and a mean ROC-AUC of 0.741 ± 0.088 across endpoints, with substantial endpoint-to-endpoint and seed-to-seed variability — for example, SR-MMP and NR-AR-LBD show strong, consistent separation (ROC-AUC > 0.8), while NR-ER and SR-HSE remain close to chance.

Clustering the fingerprint bits substantially improved explanation stability relative to the raw bit level. At the cluster level, SHAP showed the highest stability across seeds (Jaccard top-50 between 0.73 and 0.81, Spearman correlation above 0.93 for all endpoints), compared to permutation importance (Jaccard 0.37–0.59) and sensitivity analysis (Jaccard 0.41–0.60). This motivated using SHAP as the basis for the consensus and enrichment analysis.

Applying the consensus-plus-enrichment filter identified between 4 and 13 toxic driver clusters per endpoint (and between 2 and 14 protective clusters), each annotated with functional group labels. Several patterns recurred across endpoints: Cluster 43 (Furan; Indole; Aldehyde; Thiophene; Piperidine) was the top toxic driver for three nuclear receptor endpoints (NR-AR-LBD, NR-ER, NR-ER-LBD), with enrichment factors between 2.6 and 4.6; Cluster 33 (Indole; Sulfoxide; Halogen(F); Ether; Ester) drove both stress-response membrane assays (SR-HSE, SR-MMP), with its protective counterpart (Cluster 29) showing EF < 1 in both cases.

Several recovered motifs match established structural alerts in the medicinal chemistry literature: furan and thiophene are canonical alerts bioactivated to reactive electrophiles capable of covalent protein adducts [12]; the nitro group, identified as the top driver for SR-ATAD5, is one of the most widely recognised mutagenic toxicophores; sulfur-containing motifs (sulfone, sulfoxide, thioether) recur across multiple endocrine-disruption-related drivers. This correspondence provides external validation that the model's learned representations are at least partially aligned with documented chemical reactivity, rather than purely dataset-specific correlations.

5 Conclusions

This work demonstrates that combining cluster-level SHAP attribution with ground-truth enrichment validation produces a chemically interpretable shortlist of candidate toxicophores per endpoint, several of which align with documented toxicophores. The main contributions are: (1) a clustering procedure that substantially improves explanation stability over raw fingerprint bits; (2) a two-source (SHAP direction + enrichment factor) filter that distinguishes genuine toxicity drivers from clusters the model merely relies on without empirical support; and (3) a systematic cross-endpoint analysis revealing shared structural drivers across mechanistically related toxicity pathways.

At the same time, the results highlight important limitations: predictive performance varies substantially across endpoints, the severe class imbalance limits the number of positive examples available to validate each cluster, and the conclusions for any individual cluster depend on a small number of activating molecules and cannot be generalised beyond the test set. The models also remain sensitive to random initialisation, underscoring that larger datasets or representations that capture more chemical signal relative to noise are needed before such pipelines could support regulatory-grade toxicity assessment. Future work should explore alternative molecular encodings, incorporation of chemical knowledge as a modelling prior, and generative or search-based approaches to experimentally validate the candidate toxicophores identified here.

6 References

[1] A. M. Richard et al., "The tox21 10k compound library: Collaborative chemistry advancing toxicology", *Chemical Research in Toxicology*, vol. 34, no. 2, pp. 189–216, Feb. 2021, ISSN: 0893-228X. DOI: 10.1021/acs.chemrestox.0c00264. [Online]. Available: https://doi.org/10.1021/acs.chemrestox.0c00264.

[2] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[3] S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, 2017. arXiv: 1705.07874 [cs.AI]. [Online]. Available: https://arxiv.org/abs/1705.07874.

[4] J. Pizarroso, J. Portela, and A. Muñoz, "Neuralsens: Sensitivity analysis of neural networks", *Journal of Statistical Software*, vol. 102, no. 7, pp. 1–36, 2022. DOI: 10.18637/jss.v102.i07. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v102i07.

[5] D. Rogers and M. Hahn, "Extended-connectivity fingerprints", *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010, PMID: 20426451. DOI: 10.1021/ci100050t. eprint: https://doi.org/10.1021/ci100050t. [Online]. Available: https://doi.org/10.1021/ci100050t.

[6] D. R. Cox, "The regression analysis of binary sequences", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958. DOI: 10.1111/j.2517-6161.1958.tb00292.x.

[7] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree", in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[8] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. DOI: 10.1214/aos/1013203451.

[9] B. Sharma et al., "Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations", *Scientific Reports*, vol. 13, no. 1, p. 4908, Mar. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-31169-8. [Online]. Available: https://doi.org/10.1038/s41598-023-31169-8.

[10] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Grouped variable importance with random forests and application to multiple functional data analysis", *Computational Statistics Data Analysis*, vol. 90, pp. 15–35, 2015, ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2015.04.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947315000997.

[11] J. H. W. Jr., "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963. DOI: 10.1080/01621459.1963.10500845. eprint: https://www.tandfonline.com/doi/pdf/10.

1080/01621459.1963.10500845. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

[12] N. Dang, T. B. Hughes, G. Miller, and S. J. Swamidass, "Computational approach to structural alerts: Furans, phenols, nitroaromatics, and thiophenes", *Chemical research in toxicology*, vol. 30, pp. 1046–1059, 2017. DOI: 10.1021/acs.chemrestox.6b00336.

Contents

List of Figures

List of Tables

Chapter 1 Introduction and Objectives

The accelerating pace of artificial intelligence research is reshaping how new molecules are designed and evaluated. Generative models, large chemical language models, and reinforcement-learning-based molecular optimisation frameworks can now propose compounds with finely tuned biological activity, physicochemical properties, and target specificity, often surpassing the throughput of traditional medicinal chemistry workflows. This generative capability is transformative for drug discovery, agrochemical development, and materials science, where the ability to explore vast regions of chemical space *in silico* dramatically reduces the time and cost associated with identifying promising candidates.

Yet this same capability exposes a structural asymmetry in the field: while molecular generation has advanced at a remarkable rate, the computational tools used to assess the safety of these generated molecules have not kept pace. A model that can design a compound optimised to bind a biological target with high affinity offers no guarantee that the same compound is safe — it may bind unintended off-target receptors, disrupt cellular signalling, or trigger toxic mechanisms entirely unrelated to its intended function. As generative design becomes more powerful and more autonomous, the absence of equally capable toxicity prediction tools becomes increasingly consequential: the question is no longer only *can we design molecules that do what we want*, but *can we reliably know what else they might do.*

This concern is particularly acute for nuclear receptor and stress-response pathways, the focus of the Tox21 initiative, which represent some of the most common mechanisms by which chemicals interfere with normal physiological function. Endocrine disruption, oxidative stress, and DNA damage response activation are not always apparent from a molecule's intended pharmacological profile, and historically have only been detectable through resource-intensive *in vitro* and *in vivo* assays. Machine learning models trained on large-scale screening data such as Tox21 offer the possibility of flagging these risks early and at scale, but only if their predictions can be trusted and understood.

Modern toxicity predictors increasingly rely on deep learning architectures — fully-connected networks operating on molecular fingerprints, graph neural networks operating directly on molecular graphs, and transformer-based models operating on SMILES strings — which can achieve strong predictive performance on benchmark datasets. However, these models are largely *black boxes*: a prediction of "toxic" or "non-toxic" provides no insight into *why* the model reached that conclusion, which substructures contributed to the decision, or whether the model has learned a chemically plausible rule or a dataset-specific artefact.

This opacity is a serious limitation in a safety-critical context. A toxicologist or medicinal chemist evaluating a new compound needs more than a binary flag; they need to know which structural features are driving the prediction, whether those features correspond to known mechanisms of toxicity, and how confident the model's reasoning is across different training conditions. Without this information, a model's output cannot be meaningfully incorporated

into a risk assessment workflow, regardless of its raw accuracy.

Explainable Artificial Intelligence (XAI) has emerged as the principal response to this problem. Broadly, XAI approaches to molecular toxicity prediction fall into two categories. *White-box, or intrinsically interpretable, models* (such as decision trees, logistic regression, and shallow random forests) provide direct access to feature importances and decision rules, at the cost of typically lower predictive performance on complex tasks. *Post-hoc explainability methods*, by contrast, are applied after a black-box model has been trained, and attempt to recover an explanation of its behaviour without modifying the model itself. Within this second category, several families of techniques have been developed specifically for molecular applications: SHAP-based feature attribution methods quantify the contribution of individual features (or groups of features) to a prediction, providing both global and local interpretability [1]; contrastive and counterfactual methods identify the minimal structural changes required to flip a toxicity prediction [2]; attention-based mechanisms, common in graph neural networks and transformer architectures, highlight the atoms or substructures the model "focuses on" when making a prediction [3]; structural alert and feature importance methods aim to recover known or novel toxicophores directly from model attributions [1]; and test-time augmentation approaches assess the robustness and consistency of explanations under small input perturbations.

Despite this growing toolkit, several open challenges remain. There is a lack of XAI methods specifically tailored to the structure of molecular data, as many attribution techniques were originally developed for image or tabular domains. Explanations are often inconsistent across different molecular representations — fingerprints, graphs, and SMILES strings can yield different attributions for the same underlying chemistry, undermining confidence in any single explanation. There is also a recognised need to integrate chemical domain knowledge into the explanation process, so that model attributions can be related back to known mechanisms of toxicity rather than treated as purely statistical artefacts [1]. Graph neural network architectures such as Chemprop, while widely used for molecular property prediction, still lack mature, architecture-specific explainability tools, relying instead on general GNN explainability methods that may not fully exploit the structure of the model [2], [3]. Finally, the field as a whole lacks standardised evaluation protocols for explanation quality, which is a prerequisite for any regulatory acceptance of XAI-based toxicity assessments.

1.1 Motivation

The asymmetry between the pace of molecular generation and the pace of toxicity assessment raises significant scientific and ethical concerns. Highly optimised molecules are not inherently safe, and without rigorous toxicity prediction, enhanced design capability risks unintentionally producing compounds with harmful or even weapon-like properties. A molecule designed for maximal interaction with a biological target may inadvertently bind off-target receptors, disrupt cellular pathways, or trigger toxic mechanisms that current models fail

to detect. The absence of robust predictive toxicology undermines the value of molecular design itself: designing compounds is meaningless if their safety cannot be ensured. As generative models continue to grow in scale and autonomy, the lack of aligned toxicology capabilities increases the likelihood of designing molecules with dangerous or unpredictable biological effects, highlighting the urgent need for toxicity prediction methods that match the sophistication and pace of modern molecular generation.

The opacity of state-of-the-art toxicity predictors compounds this challenge. Deep learning models can achieve remarkable performance, yet they often behave as "black boxes", providing little insight into the structural or mechanistic factors driving their predictions. In safety-critical domains, such opacity severely limits trust, accountability, and scientific utility. Experts require not only accurate predictions but also interpretable explanations that highlight toxicophores, mechanistic pathways, and structure–toxicity relationships.

Global regulatory agencies increasingly emphasise transparency, reproducibility, and mechanistic justification in chemical risk assessment. Models submitted for regulatory evaluation must be interpretable and robust; otherwise, they cannot be used in decision-making processes. As AI-driven molecule generation advances rapidly, the gap between what we can design and what we can confidently deem safe continues to widen. This thesis is motivated by the need to close this gap by developing toxicity prediction models that are not only accurate but interpretable, trustworthy, and aligned with real-world safety requirements.

1.2 Objectives

Main Objectives

- Develop an explainable toxicity prediction framework capable of accurately identifying toxic compounds across the 12 Tox21 endpoints, while simultaneously extracting interpretable molecular-level insights into the drivers of toxicity. This includes building predictive models whose explanations are stable, chemically meaningful, and consistent across multiple molecular representations and random initialisations.
- Integrate a set of explainable AI (XAI) techniques, specifically permutation importance, SHAP, and neural sensitivity analysis, to provide both global and local interpretations of model behaviour. These methods are used to identify relevant substructures, quantify feature importance, and highlight candidate toxicophores associated with toxicity.
- Systematically validate the discovered molecular patterns by comparing the identified toxic-driving substructures against established toxicophore knowledge from the medicinal and chemical literature. This contrast with prior chemical studies is used to assess the chemical plausibility of the model explanations and to distinguish learned mechanistic signals from dataset-specific artefacts.

Secondary Objectives

- Conduct a systematic comparison of classical machine learning models (e.g. logistic regression, random forests, gradient boosting) and deep learning architectures (MLPs, GNNs, message-passing networks), evaluating not only predictive accuracy but also explainability and robustness.
- Investigate the influence of molecular encodings (such as SMILES strings, molecular fingerprints, graph-based representations, and 3D spatial descriptors) on both model performance and the interpretability of explanations, identifying which representations yield chemically coherent insights.
- Develop user-friendly visualisation tools that enable chemists and toxicologists to intuitively interpret model outputs, such as atom-level heatmaps, substructure relevance diagrams, feature attribution plots, and comparative explanation dashboards.

1.3 SDG Alignment

This project contributes meaningfully to several United Nations Sustainable Development Goals by advancing safer chemical design, improving public health protections, and supporting ethical innovation in computational toxicology. The integration of explainable AI into toxicity prediction not only enhances model reliability but also strengthens global efforts to reduce chemical risks and transition toward more sustainable industrial and pharmaceutical practices.

SDG 3 – Good Health and Well-Being. The development of explainable toxicity prediction models directly supports SDG 3 by enhancing the safety and effectiveness of pharmaceutical and chemical products. Traditional toxicity assessments are increasingly being supplemented or replaced by AI-based toxicity predictors. However, black-box models introduce uncertainty regarding reliability and potential biases. By ensuring interpretability, this project improves the transparency and trustworthiness of toxicity predictions, enabling clinicians, toxicologists, and regulatory authorities to understand the precise chemical drivers of harmful effects. As a result, explainable models contribute to earlier detection of toxic liabilities, reduction of adverse drug reactions, and more informed decision-making throughout the drug development pipeline. In the broader public health context, accessible and interpretable tools allow institutions to evaluate chemical hazards more effectively, ultimately supporting safer consumer products and reducing exposure to harmful substances.

SDG 9 – Industry, Innovation, and Infrastructure. The integration of interpretable AI methodologies into chemical safety pipelines advances SDG 9 by promoting responsible innovation and strengthening digital research infrastructure. Modern industries are increasingly invested in deploying machine learning models to accelerate discovery and reduce development costs. However, the adoption of such systems often depends on the ability to explain and justify model outputs, especially in regulatory or safety-critical contexts. By developing explainable toxicity predictors, this project contributes to building trustworthy AI infrastructure that industries can rely on for high-stakes decision-making. Furthermore, explainability enables more efficient R&D cycles, as transparent models provide actionable insights that support rational molecular design rather than black-box predictions. This accelerates innovation and fosters the development of new products, tools, and research workflows aligned with ethical and sustainable technological progress.

SDG 12 – Responsible Consumption and Production. Ensuring the environmental and human safety of chemicals is central to SDG 12, which calls for more sustainable production systems and reduced ecological harm. Explainable toxicity predictors play a crucial role in identifying hazardous chemicals early in development, preventing harmful compounds from reaching manufacturing, supply chains, or ecosystems. Transparent models allow chemists and environmental scientists to understand the structural basis of toxicity and make informed decisions to design safer, more environmentally friendly alternatives.

SDG 15 – Life on Land. The adoption of explainable AI reduces dependence on animal testing, aligning with ethical imperatives and international commitments to alternative testing strategies. By enabling transparent, mechanistically interpretable toxicity predictions, these models support safer chemical assessment while reducing harm to wildlife and terrestrial ecosystems. Moreover, improved *in silico* toxicology helps prevent the release of hazardous compounds into the environment, contributing to the preservation of biodiversity and the protection of vulnerable species.

Chapter 2 State of the art

Predicting molecular toxicity is an inherently difficult problem for several reasons. First, the chemical space of potential compounds is astronomically large (estimated at 10^{23}–10^{60} drug-like molecules [4]) making exhaustive experimental screening infeasible and leaving machine learning models with sparse, incomplete coverage of this space. Second, activity relationships are highly non-linear: structurally similar compounds can exhibit drastically different toxicity profiles, while structurally diverse compounds may share the same adverse mechanism. Third, toxicity is a multi-endpoint phenomenon driven by distinct biological pathways (nuclear-receptor activation, cellular stress responses, DNA damage, etc.), each potentially governed by different structural determinants, so a single universal model rarely generalises across all endpoints. Fourth, available labelled datasets such as Tox21 are small by deep learning standards and severely class-imbalanced, with toxic compounds typically representing fewer than 10–20% of examples (see Section 4.1.1), amplifying overfitting risk and evaluation uncertainty. Taken together, these challenges motivate the need for carefully chosen molecular representations, robust modelling strategies, and explainability methods that can distinguish genuine structural signals from model artefacts.

2.1 Molecular representations

To feed molecular data into an AI model, it must be converted into a vectorised representation. Several approaches exist, each with different trade-offs for predictive performance and explainability.

2.1.1 SMILES representation.

Molecules are encoded as linear strings describing their structure, enabling sequence-based models such as Transformers, with no explicit feature engineering and the ability to benefit from large-scale pretraining [5]. However, interpretability is very low: tokens do not map cleanly to chemical meaning, multiple valid SMILES exist for the same molecule, and attention-based explanations are unreliable. This representation is shown in Figure 1

2.1.2 XYZ (3D Coordinates).

Molecules are represented as atoms with explicit 3D spatial coordinates, capturing geometric structure and interatomic distances [6]. This is physically meaningful and useful for quantum chemistry and binding-related tasks, but requires conformer generation (adding noise

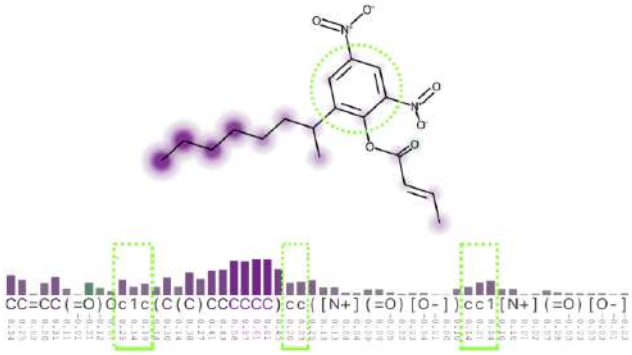


Figure 1: Smiles representation. [5]

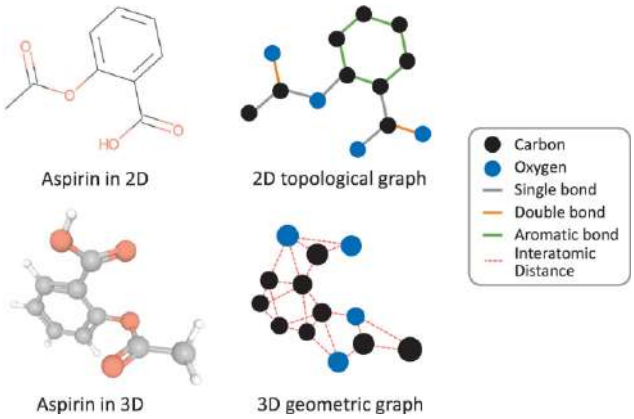


Figure 2: Graph representations [8]

and computational cost), produces ambiguity from multiple conformations, and yields explanations that focus on geometric relationships rather than recognisable chemical motifs — making it poorly suited for XAI.

2.1.3 Graph representation.

Molecules are modelled as graphs where atoms are nodes and bonds are edges, preserving structural relationships [7]. This is a natural and chemically meaningful representation that supports substructure-level explanations (nodes, edges, subgraphs) and achieves strong predictive performance with GNNs. However, explanations can be noisy and unstable across runs, and attribution methods applied to graphs often lack consistency, providing only moderate interpretability in practice. This representation is shown in Figure 2.

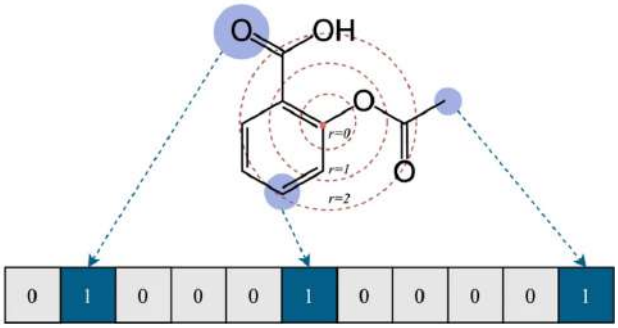


Figure 3: Morgan Fingerprints visualisation [10]

2.1.4 Morgan Fingerprints

Morgan fingerprints, also known as extended-connectivity fingerprints (ECFP), are a widely used molecular representation for machine learning in cheminformatics [9]. They encode molecular structure into a fixed-length binary vector, where each bit represents the presence or absence of a particular local substructure.

The construction of Morgan fingerprints begins by assigning an initial identifier to each atom based on its atomic properties (e.g. atom type, degree, valence, and charge). These identifiers are then iteratively updated by aggregating information from neighbouring atoms within a predefined radius. At each iteration, atom neighbourhoods are hashed into numerical identifiers, which are ultimately folded into a fixed-length bit vector of size n (e.g. 1024 or 2048 bits). If a substructure is encountered, the corresponding bit position is set to 1. A visual representation can be found in Figure 3.

A key parameter of Morgan fingerprints is the radius, which controls the size of the local chemical environment considered around each atom. A radius of $r = 2$, for example, encodes substructures up to two bonds away from each atom.

Morgan fingerprints are particularly attractive for machine learning applications due to their simplicity, efficiency, and strong empirical performance. Importantly, they are also highly interpretable: each activated bit can be traced back to a specific substructure, enabling direct chemical interpretation and supporting explainability methods such as feature importance or permutation-based analyses.

However, Morgan fingerprints also have limitations. Since they rely on hashing substructures into a fixed-length vector, collisions may occur, where different substructures map to the same bit. Additionally, they do not explicitly encode global molecular geometry or long-range dependencies, which may be relevant for certain tasks.

2.2 XAI in molecular toxicity prediction

Explainable Artificial Intelligence (XAI) has become essential in molecular toxicity prediction, where deep learning and machine learning models are increasingly used to assess chemical safety and drug toxicity. While these models offer high predictive accuracy, their "black-box" nature limits trust, regulatory acceptance, and scientific insight. XAI methods address this by providing human-understandable explanations for model predictions, which is crucial for regulatory compliance, risk assessment, and guiding molecular design in drug discovery and environmental safety. The need for transparency is further emphasised by regulatory bodies and the scientific community, especially as AI models increasingly complement or replace traditional animal-based toxicity testing.

XAI approaches in this field can be broadly categorised as follows

2.2.1 White-box (intrinsic) models.

These include interpretable algorithms such as decision trees, logistic regression, and simple random forests, which provide direct insight into feature importance and decision rules. Such models are favoured for their transparency, though they may sacrifice some predictive power compared to deep learning.

2.2.2 Post-hoc explainability for black-box models.

These methods interpret complex models after training:

1. **SHAP:** Quantifies feature contributions for predictions, enabling global and local interpretability [1]. König and Vellido apply SHAP and related explainable ML models to drug profile prediction, demonstrating how feature attributions can be related to known pharmacological and toxicological properties of compounds [11]. However, unlike this work, they do not use Morgan Fingerprints. SHAP is also used by Walter et al. [12] as a method for comparison.
2. **Contrastive and counterfactuals:** Identify minimal structural changes that alter toxicity. This is the XAI method used in [2], whose model is used as a starting point for this work.
3. **Attention mechanisms:** Highlight key atoms or substructures, especially in GNNs and transformers, used in [3], another considered starting point for this work.
4. **Structural alerts and feature importance:** Identify toxicophores or relevant motifs [1], [13]. Closely related to the approach in this thesis, Walter et al. extract learned chemical features directly from trained neural network toxicity predictors, recovering

structural motifs associated with toxicity from model internals rather than from predefined libraries [12].

5. **Test-time augmentation:** Evaluates robustness and consistency of explanations [14].

Chemprop and graph-based explainability. Chemprop (MPNN-based) is widely used for molecular property prediction. While Chemprop-specific XAI is still emerging, general GNN explainability methods apply [2], [3].

2.2.3 Permutation importance

Permutation importance is a model-agnostic technique used to quantify the contribution of input features to a model's predictions. The central idea is to measure how much the model's performance changes when the information content of a feature is disrupted. This method was first introduced in [15], where it was used to assess variable importance by measuring the decrease in predictive performance after permuting feature values.

Formally, let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the input feature matrix and $f(\cdot)$ the trained model. The baseline predictions are given by:

$$\hat{y} = f(\mathbf{X})$$

For a given feature j, a perturbed dataset $\mathbf{X}^{(j)}$ is constructed by randomly permuting the values of feature j across samples, while keeping all other features unchanged. The importance of feature j is then defined as the expected change in model output:

$$I(j) = \mathbb{E}\left[\left|f(\mathbf{X}) - f(\mathbf{X}^{(j)})\right|\right]$$

To reduce variance, this procedure is repeated multiple times and averaged. Finally, to account for feature scale effects, the importance scores may optionally be normalised. This approach provides a robust estimate of feature relevance by quantifying the sensitivity of model predictions to the disruption of individual input features.

2.2.4 SHAP (Shapley Additive Explanations)

SHAP values provide a game-theoretic framework for attributing a model's prediction to individual input features [16]. The method is based on Shapley values from cooperative game theory, where each feature is treated as a "player" contributing to the final prediction.

Given a trained model $f(\cdot)$ and an input $\mathbf{x} \in \mathbb{R}^d$, SHAP explains the deviation of the prediction from a baseline expectation $\mathbb{E}[f(\mathbf{X})]$ as an additive combination of feature contributions:

$$f(\mathbf{x}) = \mathbb{E}[f(\mathbf{X})] + \sum_{i=1}^{d} \phi_i$$

where ϕ_i denotes the SHAP value for feature i, representing its marginal contribution averaged over all possible feature subsets. Formally, each SHAP value is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \left[f_{S \cup \{i\}}(\mathbf{x}) - f_S(\mathbf{x}) \right]$$

where F is the full set of features and $f_S(\mathbf{x})$ denotes the model evaluated with only features in subset S present (with others marginalised or imputed).

2.2.5 NeuralSens (Neural Sensitivity Analysis)

NeuralSens is a gradient-based feature attribution method that quantifies the sensitivity of a model's output with respect to small perturbations in the input features [17]. Unlike permutation-based approaches, which rely on discrete feature corruption, NeuralSens evaluates local changes in the model function using derivatives.

Given a differentiable model $f(\mathbf{x})$, the sensitivity of the output with respect to feature x_i is defined as:

$$S_i(\mathbf{x}) = \left| \frac{\partial f(\mathbf{x})}{\partial x_i} \right|$$

This quantity measures how rapidly the prediction changes in response to infinitesimal variations in each input dimension. To obtain a global importance score, sensitivities are averaged over the dataset:

$$I_i = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\left| \frac{\partial f(\mathbf{x})}{\partial x_i} \right| \right]$$

2.3 Research gaps and future directions

Despite this growing toolkit, several open challenges remain:

- Need for integration with chemical domain knowledge [1].
- Limited GNN- and Chemprop-specific XAI tools [3].
- Need for standardisation to reach regulatory acceptance.
- Need for consistent explanations across random initialisations, to verify that explanations actually encode chemical information rather than noise or data artefacts.
- Discovery of substructures that encode toxicity on any given compound, rather than those on an specific subset.

To improve stability and chemical interpretability, all explainability methods in this work are applied in a grouped setting, where Morgan fingerprint bits are aggregated into chemically meaningful clusters prior to analysis. This design is motivated by the high redundancy and strong correlation between fingerprint bits encoding related substructures, which can lead to unstable or non-unique attributions at the individual feature level.

Grouped importance follows the framework of Gregorutti et al. [18], who demonstrate that aggregating correlated variables prior to importance estimation leads to more stable and interpretable feature rankings in high-dimensional settings. In this work, clustering is used to define groups of fingerprint bits corresponding to shared chemical motifs, and all three explainability methods (permutation importance, SHAP, and NeuralSens) are computed at the cluster level rather than at the bit level.

This approach ensures that explanations reflect broader structural patterns instead of arbitrary individual fingerprint activations, improving robustness across random initialisations and aligning the analysis with chemically meaningful substructures.

The methods used in this work are introduced above and are developed in detail in the following sections, with particular attention to the consistency gap identified here.

Chapter 3 Methodology

This chapter describes the full pipeline developed to predict molecular toxicity across the twelve Tox21 endpoints and to extract chemically interpretable explanations from the trained models. The pipeline proceeds in four stages. First, molecules are encoded as Morgan fingerprints, providing a compact, unambiguous, and interpretable numerical representation of each compound. Second, an exploratory data analysis is carried out to characterise the dataset, assess class imbalance across endpoints, and examine the global structure of the encoded chemical space via dimensionality reduction. Third, predictive models are trained independently for each endpoint: three classical baseline models (logistic regression, random forest, and gradient boosting) are first evaluated to establish reference performance benchmarks, after which a single-task deep neural network is trained taking the Morgan fingerprint vectors as input and outputting a toxicity probability. Fourth, and centrally to this work, three complementary explainability methods are applied to the trained models to attribute predictions to input features. To improve the stability and chemical interpretability of these attributions, all three methods are computed at the cluster level, where fingerprint bits are first grouped into chemically meaningful clusters, so that explanations reflect broader structural motifs rather than arbitrary individual bit activations. The following sections describe each stage in detail.

3.1 Dataset

The Tox21 dataset is a widely used benchmark for computational toxicology, originally released as part of the Toxicology in the 21st Century initiative [19]. It was designed to support the development of predictive models for assessing the toxicity of chemical compounds using high-throughput screening assays. The dataset contains molecular structures represented as fingerprints, together with binary labels indicating whether a compound is active or inactive with respect to a specific biological target or toxicity pathway.

This work considers 12 Tox21 endpoints, each corresponding to a distinct nuclear receptor or stress response pathway, formulated as binary classification tasks.

NR (Nuclear Receptor) endpoints measure whether a compound interacts with or disrupts nuclear receptor signalling pathways, primarily associated with endocrine-related activity, including androgen, estrogen, and other hormone receptors.

SR (Stress Response) endpoints measure whether a compound induces general cellular stress pathways, including oxidative stress, DNA damage, protein misfolding, and mitochondrial dysfunction, capturing broader indicators of cellular toxicity rather than specific receptor binding.

The endpoints are described as follows:

- **NR-AhR**: Aryl hydrocarbon receptor. Involved in the response to environmental toxins such as dioxins and polycyclic aromatic hydrocarbons, mediating xenobiotic metabolism and toxicological signalling.
- **NR-AR**: Androgen receptor. A nuclear receptor involved in male sexual development and endocrine signalling; disruption may indicate endocrine toxicity.
- **NR-AR-LBD**: Androgen receptor ligand-binding domain. A more specific assay targeting binding interactions with the ligand-binding domain of the androgen receptor.
- **NR-Aromatase**: Enzyme responsible for converting androgens to estrogens; inhibition or activation is associated with endocrine disruption.
- **NR-ER**: Estrogen receptor. A key receptor in hormone signalling, reproductive biology, and endocrine disruption.
- **NR-ER-LBD**: Estrogen receptor ligand-binding domain assay focusing specifically on ligand interactions.
- **NR-PPAR-gamma**: Peroxisome proliferator-activated receptor gamma. Involved in lipid metabolism, glucose homeostasis, and adipogenesis; disruption may indicate metabolic toxicity.
- **SR-ARE**: Antioxidant response element pathway. Measures activation of oxidative stress response via NRF2-mediated signalling.
- **SR-ATAD5**: DNA damage response pathway associated with genomic stability and DNA repair mechanisms.
- **SR-HSE**: Heat shock element pathway. Reflects cellular stress response due to protein misfolding or thermal stress.
- **SR-MMP**: Mitochondrial membrane potential disruption, indicating mitochondrial dysfunction and cytotoxic stress.
- **SR-p53**: Tumor protein p53 pathway, a key regulator of apoptosis, DNA repair, and cellular stress response.

Together, these endpoints span a diverse range of biological processes, including nuclear receptor signalling, oxidative stress response, DNA damage repair, and mitochondrial toxicity. This diversity makes Tox21 a challenging and informative benchmark for evaluating machine learning models in predictive toxicology.

This dataset was selected for its extended use, specifically because it was used in [3] and [2], which are the starting points for this work.

3.2 Exploratory Data Analysis (EDA)

Molecular data analysis presents several challenges due to its high-dimensional and sparse nature. In addition, datasets in computational toxicology are often characterized by class imbalance and incomplete annotations, which can affect downstream modeling performance.

To address these challenges, an exploratory data analysis (EDA) and feature representation pipeline was developed to better understand the structure of the Tox21 dataset prior to model training. This includes an examination of molecular representations, class distributions, and the structure of the chemical feature space.

The dataset used in this work consists of multiple toxicity endpoints, each formulated as a binary classification task. These endpoints exhibit varying degrees of class imbalance, with a higher proportion of inactive (*non-toxic*) compounds relative to active (*toxic*) compounds. Additionally, missing values are present across several endpoints and are handled during preprocessing.

3.3 Dimensionality reduction

To explore the structure of the molecular feature space, several dimensionality reduction techniques were applied to the Morgan fingerprint representation. Molecular structures were encoded using 2048-bit Morgan fingerprints (radius = 2), which capture the presence or absence of local chemical substructures in a high-dimensional binary space.

Three complementary techniques were used for analysis. Principal Component Analysis (PCA) was applied as an unsupervised linear projection method that maps the data onto directions of maximum variance. t-Distributed Stochastic Neighbor Embedding (t-SNE) was used to visualize potential nonlinear structure while preserving local neighborhood relationships in the data. Finally, Linear Discriminant Analysis (LDA) was applied as a supervised dimensionality reduction method that finds directions maximizing class separability based on toxicity labels.

For binary classification tasks, LDA produces at most one discriminant component, corresponding to (C - 1) dimensions where C is the number of classes. Since all Tox21 endpoints are binary, each LDA projection results in a single one-dimensional representation per endpoint.

3.4 Baseline models

To establish reference performance levels for toxicity prediction, three classical machine learning models were evaluated on each Tox21 endpoint independently: Logistic Regression [20], Random Forest [15], and Gradient Boosting [21] using LightGBM [22]. Molecular compounds

were represented using pre-computed molecular descriptors and fingerprints, with each toxicity endpoint formulated as a separate binary classification task.

For each endpoint, the dataset was split into training and test subsets using an 80:20 stratified partition to preserve the original class distribution. Given the strong class imbalance present in several Tox21 endpoints, class weighting strategies were incorporated during training. For Logistic Regression and Random Forest models, class weights were computed from the training set and applied during model fitting. For LightGBM, class imbalance was handled through the `scale_pos_weight` parameter, which increases the contribution of minority-class samples during training.

The Logistic Regression model served as a linear baseline and was trained using the `liblinear` solver. Random Forest was included as a non-linear ensemble method capable of capturing interactions between molecular features. LightGBM was selected as a boosting-based approach due to its strong performance on structured tabular data and its ability to model complex non-linear relationships efficiently. Prior to model fitting, input features were standardized using z-score normalization within a preprocessing pipeline.

All models produced probabilistic predictions for each compound. To ensure consistency with the final deep learning model developed in this work, a fixed classification threshold of 0.2 was applied when converting predicted probabilities into binary predictions. This threshold was selected to maintain a consistent evaluation protocol across all experiments and to better account for the class imbalance present in the Tox21 dataset.

Model performance was assessed using Accuracy and Balanced Accuracy. In addition, confusion matrices were generated for each endpoint and model, and probability distributions of model predictions were analyzed to provide further insight into classification behaviour and class separation.

3.5 Deep Neural Networks

In addition to the classical machine learning baselines, two deep learning approaches were investigated for molecular toxicity prediction: a fully connected Deep Neural Network (DNN) operating on molecular fingerprints, developed in the paper [2] and an Equivariant Graph Neural Network (EGNN) operating directly on molecular graph representations [3]. These architectures were selected because they represent two widely adopted paradigms in modern cheminformatics: descriptor-based learning and graph-based learning.

3.5.1 Equivariant Graph Neural Network

A graph-based approach was also investigated based on the methodology described in [3]. In this framework, molecules are represented as graphs in which atoms correspond to nodes

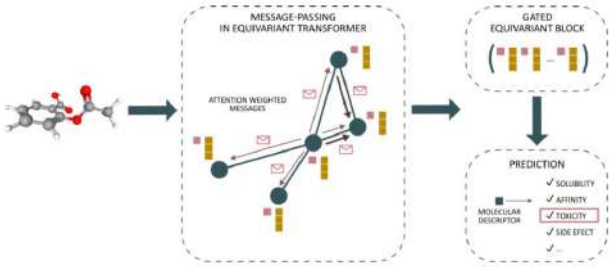


Figure 4: EGNN overview [3]

and chemical bonds correspond to edges. Unlike fingerprint-based methods, graph neural networks learn molecular representations directly from the molecular structure, potentially capturing more detailed topological information. An overview of this work can be found in Figure 4.

Equivariant Graph Neural Networks are specifically designed to preserve geometric symmetries during message passing, allowing molecular representations to remain consistent under rotations and translations. This property makes them particularly attractive for molecular modelling tasks where spatial information plays an important role.

Despite their theoretical advantages, the EGNN approach was not adopted as the final model for this project. First, the original study primarily reported performance metrics such as ROC-AUC, making direct comparison using Balanced Accuracy more difficult. Since the Tox21 dataset exhibits substantial class imbalance, Balanced Accuracy is a more informative evaluation metric. Second, reproducing the EGNN implementation proved challenging due to dependency and library compatibility issues commonly associated with graph deep learning frameworks.

Importantly, even if a fully functional EGNN implementation had been successfully reproduced, its use would have introduced additional methodological challenges. In particular, explainability for graph neural networks remains an active and non-trivial research area, with many XAI techniques still lacking robustness, standardisation, and chemical interpretability guarantees. Addressing these limitations would have significantly increased the scope and complexity of the interpretability analysis, extending beyond the objectives of this work.

Finally, the fingerprint-based DNN provided a significantly simpler and more transparent workflow while achieving competitive predictive performance. For these reasons, the DNN architecture was selected as the primary model for subsequent experiments and interpretability analyses.

3.5.2 Deep Neural Network

The primary deep learning architecture used in this work was based on the model proposed in [2]. Molecular structures were represented using 4096-bit Morgan fingerprints (radius = 2), which were provided as input to a fully connected neural network. For performance reasons, 2048-bit Morgan FPs were used in this project.

The architecture consisted of three hidden layers containing 1024, 512 and 256 neurons, respectively. Batch normalization was applied after the first two hidden layers to improve training stability, while dropout regularization ($p = 0.2$) was incorporated to reduce overfitting. LeakyReLU activation functions were used throughout the network, and a sigmoid output layer produced probabilities for binary toxicity classification.

The original publication evaluated both multi-task and single-task learning frameworks. For the Tox21 dataset using Morgan fingerprint representations, the single-task models reported stronger predictive performance than their multi-task counterparts. Consequently, the implementation used in this project followed a single-task approach, training an independent model for each toxicity endpoint. This choice not only aligned with the best-performing configuration reported in the original study but also facilitated a more direct comparison with the baseline machine learning models, which were trained separately for each endpoint.

The same model was trained on 5 different seeds to avoid local optima for each endpoint. In total $12 \times 5 = 60$ models were trained.

3.5.3 Explainable AI (XAI)

Once all the models were trained, different XAI techniques were applied to interpret the behaviour of the model. This is done aiming to answer the question: which chemical motifs cause toxicity?

Permutation importance, SHAP and sensitivity analysis were performed on the 12×5 models. To test the robustness of the explanations, two different measures were used.

3.5.4 Jaccard index (IoU)

The Jaccard index (Jaccard similarity coefficient or Intersection over Union) measures the overlap between two sets of data [23]. Mathematically, it is expressed as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The Jaccard index was used to asses the overlap of the top-k important features across different seeds. Jaccard index is expected to be high: the models should rely on similar

variables. If this value is extremely low, it is a sign that models are totally dependent on randomness and are not learning chemical signal, just noise.

Jaccard index can be better visualized in Figure 5.

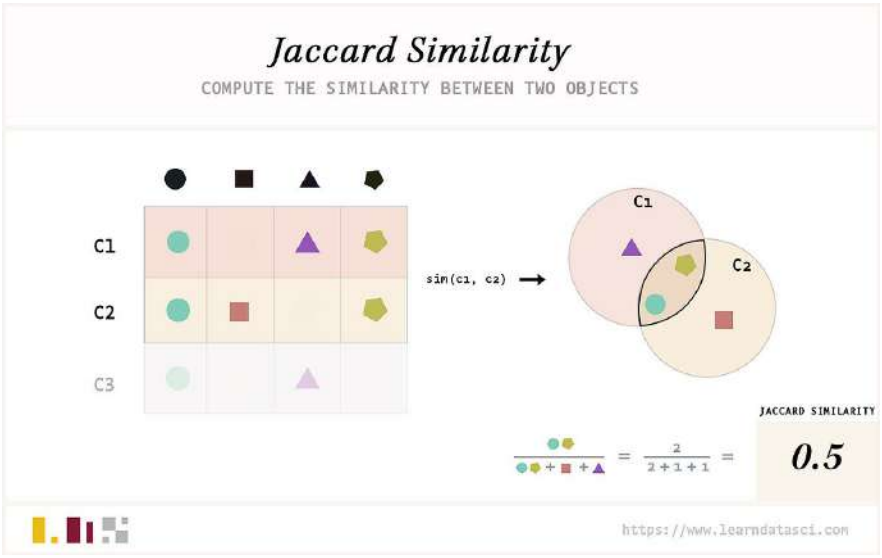


Figure 5: Jaccard index visualization

3.5.5 Spearman index

The Spearman rank correlation coefficient measures the degree of agreement between two ranked lists [24]. In this work, it was used to evaluate the consistency of feature importance rankings obtained from different random seeds. Unlike the Jaccard index, which focuses on the overlap between selected features, the Spearman coefficient assesses whether variables are ordered similarly according to their importance scores across different model instances.

Given two rankings, R_A and R_B, of n variables, the Spearman coefficient is defined as

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

where d_i is the difference between the ranks assigned to variable i in the two rankings. The coefficient takes values in the interval $[-1, 1]$, where $\rho = 1$ indicates identical rankings, $\rho = 0$ indicates no monotonic relationship, and $\rho = -1$ indicates completely reversed rankings.

In the context of XAI robustness, a high Spearman coefficient across different seeds indicates that the explanatory method consistently assigns similar relative importance to the variables, even if the exact importance values vary between model instances.

3.5.6 Input clustering

Morgan fingerprint representations are sparse and highly redundant. Multiple fingerprint bits may encode correlated or related substructures that frequently co-occur within molecules. As a result, models trained with different random seeds may rely on different fingerprint bits while still capturing the same underlying chemical motifs. This can lead to apparent disagreement in feature-level explanations, where importance is distributed across different but correlated bits. Consequently, low Jaccard or Spearman agreement between explanations does not necessarily indicate that models have learned different chemical patterns; instead, it may reflect alternative selections among redundant, highly correlated features.

To address this issue and obtain more reliable explanations, an input-clustering approach was adopted. Rather than computing explanations directly on individual fingerprint bits, features were first grouped into clusters of correlated inputs. This reduces redundancy in the representation and improves the stability and interpretability of importance estimates.

Clustering was performed exclusively on the training set to avoid information leakage. First, Morgan fingerprints were computed for all training molecules. A pairwise correlation matrix was then constructed across fingerprint bits, where each entry represents the Pearson correlation coefficient between two bits over the training samples. A distance metric was defined as

$$d_{ij} = 1 - |r_{ij}|,$$

where r_{ij} denotes the Pearson correlation between fingerprint bits i and j. The absolute correlation ensures that both positively and negatively correlated bits are grouped when they encode related structural information.

Agglomerative hierarchical clustering with average linkage was applied to this distance matrix, producing 300 clusters of fingerprint bits. Each cluster therefore represents a set of features that tend to co-activate across molecules and are likely associated with related chemical substructures.

Finally, explanations were computed at the cluster level rather than at the individual bit level. In particular, all fingerprint bits within a cluster were treated as a single unit to ensure that attribution methods capture higher-level chemical patterns rather than arbitrary correlations between individual bits. This aggregation strategy was applied consistently across permutation importance, SHAP, and gradient-based sensitivity analysis: for permutation importance, all fingerprint bits within a cluster were treated as a single unit during permutation-based importance estimation; for SHAP, bit-level attributions were first computed and then aggregated within clusters to reflect the overall contribution of each chemical motif, while for sensitivity analysis, gradient-based feature sensitivities were averaged and similarly pooled at the cluster level. This approach reduces the effect of feature redundancy and ensures that

importance scores reflect coherent molecular substructures rather than isolated or redundant fingerprint activations.

To obtain robust and chemically meaningful explanations, all feature attribution methods were adapted from the original 2048-dimensional Morgan fingerprint space to a clustered feature space. Each cluster contains a set of highly correlated fingerprint bits, and explanations are computed at the cluster level by aggregating contributions of all bits within each cluster. This follows the idea of grouped feature importance, where correlated variables are treated as a single explanatory unit rather than interpreted independently [18]. This ensures that importance is assigned to coherent chemical motifs rather than redundant individual bits.

3.5.7 Permutation importance (cluster-level)

Permutation importance is adapted by jointly permuting all fingerprint bits within a cluster while keeping all other clusters fixed. Specifically, for each cluster, the corresponding feature submatrix is randomly shuffled across molecules, breaking any association between that cluster and the model output while preserving marginal feature distributions.

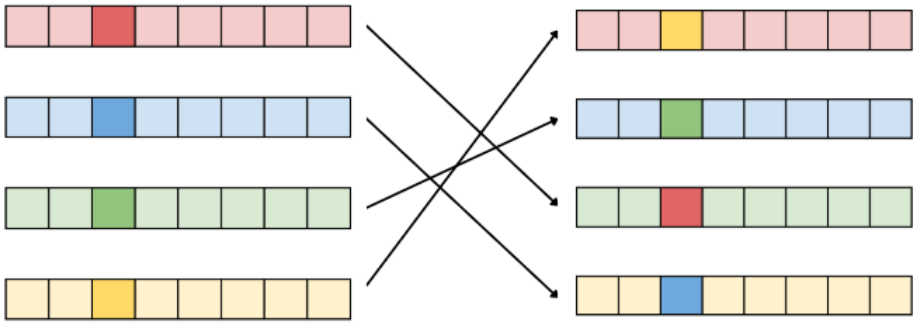
Let C_k denote the set of bit indices belonging to cluster k. The importance of cluster k is defined as the average change in model predictions after permutation:

$$I_k = \frac{1}{R} \sum_{r=1}^{R} \mathbb{E}\left[\left| f(X) - f(\tilde{X}^{(k,r)}) \right|\right],$$

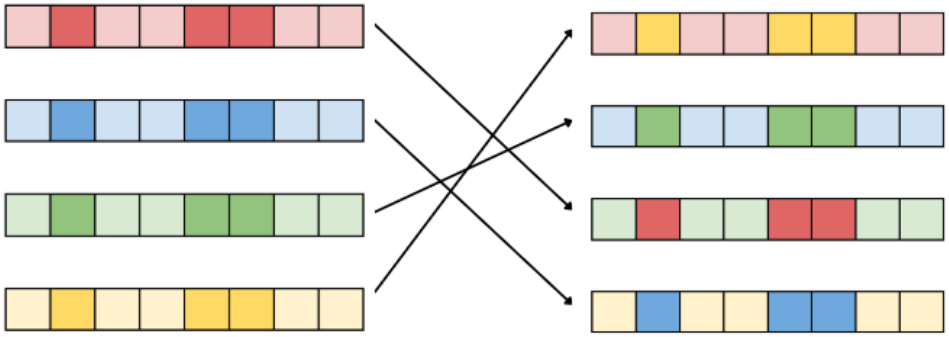
where $\tilde{X}^{(k,r)}$ denotes the dataset in which features in cluster k are permuted, and R is the number of repeats. To ensure comparability across clusters of different sizes, the resulting importance is normalised by $\sqrt{|C_k|}$.

This procedure attributes importance to groups of correlated bits rather than individual fingerprint features, reducing redundancy-induced instability.

In Figures 6a and 6b, a visual example is shown of raw permutation importance vs clustered permutation importance.



(a) Raw bit-level permutation importance



(b) Cluster-level permutation importance

Figure 6: Comparison between raw bit-level and clustered permutation importance. Clustering stabilises importance scores by aggregating correlated fingerprint bits into chemically meaningful groups, reducing noise and redundancy in the attribution signal.

3.5.8 SHAP (cluster-level)

SHAP values are computed using `GradientExplainer`, which estimates feature attributions based on expected gradients with respect to a background distribution. Let $\phi_{i,j}$ denote the SHAP value of feature j for sample i. Cluster-level importance is obtained by aggregating absolute SHAP values over all features in a cluster:

$$I_k = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in C_k} |\phi_{i,j}|.$$

This aggregation ensures that all correlated bits contributing to the same chemical pattern are jointly accounted for. Unlike permutation importance, SHAP retains local (per-sample) attribution structure before aggregation, but final explanations are still expressed at the cluster level for consistency.

3.5.9 Sensitivity analysis (gradient-based importance)

Gradient-based sensitivity analysis measures the local influence of each input feature on the model output. For each input X, gradients are computed via backpropagation and combined with the input values to capture signed feature contributions:

$$S_{i,j} = \left| \frac{\partial f(X_i)}{\partial X_{i,j}} \cdot X_{i,j} \right|.$$

To obtain cluster-level importance, bit-level sensitivities are first averaged across all samples and then aggregated within each cluster:

$$I_k = \frac{1}{\sqrt{|C_k|}} \sum_{j \in C_k} \mathbb{E}[S_j].$$

This produces a smooth sensitivity measure that reflects how strongly perturbations in a cluster influence the model output.

3.6 Reliability of explanations

The previous explanations answer three questions:

- **Permutation importance**: which clusters affect prediction error the most if permuted?
- **SHAP**: which clusters contribute most to the prediction on average across samples, based on their marginal contribution relative to a background distribution?
- **Sensitivity analysis**: which clusters most strongly influence the model output under local infinitesimal perturbations of the input?

However, the ideal question is: do these clusters actually cause toxicity? To connect the clusters labeled as "important" to toxic molecules, several approaches were taken.

First, the concept of a *molecule activating a certain cluster* needs to be defined.

Definition of cluster activation A molecule is considered to activate a cluster if its fingerprint exhibits stronger activation within that cluster than the average molecule in the dataset. Let C_k denote the set of fingerprint bit indices belonging to cluster k, and let $x_i \in \{0,1\}^D$ denote the binary Morgan fingerprint of molecule i.

The cluster activation score for molecule i and cluster k is defined as the fraction of active bits within the cluster:

$$a_{i,k} = \frac{1}{|C_k|} \sum_{j \in C_k} x_{i,j}.$$

The mean activation of cluster k over the dataset is:

$$\bar{a}_k = \frac{1}{N} \sum_{i=1}^{N} a_{i,k}.$$

A molecule i is considered to activate cluster k if:

$$a_{i,k} > \bar{a}_k.$$

Example: Assume each molecule is encoded by 6 bits. Consider cluster C_k, which consists on the bits in positions 1, 2, 3 and 5. So for each molecule $x = [x_1, x_2, x_3, x_4, x_5, x_6]$, the bits x_1, x_2, x_3, x_5 form a cluster.

Consider a dataset consisting on three molecules:

$$x_A = [1,0,0,0,0,1], \quad x_B = [1,0,0,0,1,0], \quad x_C = [0,0,1,1,0,0]$$

Their cluster activation scores are:

$$a_{A,k} = \frac{1}{4} = 0.25, \quad a_{B,k} = \frac{2}{4} = 0.5, \quad a_{C,k} = \frac{1}{4} = 0.25$$

The dataset activation mean for this cluster is

$$\frac{0.25 + 0.5 + 0.25}{3} = 0.3$$

- molecule A does not activate the cluster $(0.25 < 0.3)$,
- molecule B activates the cluster $(0.5 > 0.3)$.
- molecule C does not activate the cluster $(0.25 < 0.3)$,

3.6.1 Molecular cluster properties

To assess whether highly important clusters are associated with toxic activity, we analyze the distribution of toxic and non-toxic molecules among those molecules that activate the most important clusters. In particular, we examine whether clusters with high importance scores are preferentially activated by toxic compounds.

However, this analysis must be interpreted with caution due to the strong class imbalance in the dataset, where only approximately 5% of molecules are labeled as toxic. In this setting, it is expected that many of the most important clusters are associated with the majority (non-toxic) class. Therefore, observing that important clusters are predominantly activated by non-toxic molecules is not necessarily informative, as this simply reflects the underlying class distribution rather than model bias.

Moreover, because the baseline prevalence of toxic molecules is very low, it is statistically difficult for any individual cluster to contain a large absolute number of toxic compounds. Even clusters that are strongly associated with toxicity in a relative sense may still appear to have low toxic coverage when measured in raw counts or percentages. As a result, absolute frequency-based analyses tend to underestimate enrichment effects in imbalanced datasets.

3.6.2 Refined enrichment analysis of toxic activity

To address the limitations of raw distribution analysis, two modifications were introduced to better capture meaningful associations between cluster activation and toxicity.

First, instead of relying on raw percentages of toxic molecules within clusters, we compute an enrichment factor (EF), which compares the observed proportion of toxic molecules in a cluster to the expected proportion under the dataset background distribution. This allows us to quantify whether a cluster is over-represented in toxic molecules relative to chance, rather than relying on absolute counts.

Second, we extend the analysis beyond single-cluster activation and examine molecules that activate multiple important clusters simultaneously. Specifically, we compute enrichment factors for molecules that activate two or more high-importance clusters, rather than considering each cluster in isolation. This allows us to capture more complex structure-toxicity relationships that may emerge from the combination of multiple correlated chemical motifs.

Chapter 4 Experimental Results

This sections shows all the results obtained by following the previously described methodologies. Some aggregated Figures or examples for endpoints can be found here. For detailed results, consult the Annex 7.

4.1 Exploratory Data Analysis (EDA)

As stated in previous sections, molecular data is complex by nature, making EDA an indispensable step before training models.

4.1.1 Class balance

The Tox21 dataset exhibits a strong class imbalance across all endpoints, with a small proportion of active (toxic) compounds compared to inactive ones. This imbalance can significantly affect both model training and evaluation, particularly for metrics sensitive to class distribution. The class balance for each endpoint, as well as the mean balance is shown in Table 1. Effective positives are the percentage of positives only over labelled data. This same data visualised as bar plots per endpoint are found in Annex A.1

Endpoint	**% Positive**	**% Negative**	**% Missing**	**% Effective positive**
NR-AR	3.95	88.83	7.23	4.25
NR-AR-LBD	3.03	83.27	13.70	3.51
NR-AhR	9.81	73.82	16.37	11.73
NR-Aromatase	3.83	70.50	25.67	5.15
NR-ER	10.13	68.96	20.92	12.80
NR-ER-LBD	4.47	84.34	11.19	5.03
NR-PPAR-gamma	2.38	79.99	17.64	2.88
SR-ARE	12.03	62.44	25.53	16.15
SR-ATAD5	3.37	86.94	9.69	3.73
SR-HSE	4.75	77.83	17.42	5.75
SR-MMP	11.72	62.47	25.81	15.80
SR-p53	5.40	81.10	13.50	6.24
Mean	6.24	76.71	17.05	7.75

Table 1: Class distribution per Tox21 endpoint. The dataset exhibits strong imbalance and a non-negligible proportion of missing labels, with an effective positive rate (computed over labelled molecules only) ranging from 2.88% to 16.15%.

4.1.2 Dimensionality reduction

To gain insight into the structure of the molecular representation space, several dimensionality reduction techniques were applied to the Morgan fingerprint features, including Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Linear Discriminant Analysis (LDA). These methods were used to explore whether toxicity labels form meaningful structure in the high-dimensional fingerprint space.

PCA and t-SNE on Morgan fingerprints PCA was first applied to reduce the dimensionality of the 2048-bit Morgan fingerprint vectors. A two-dimensional projection was used for visualization, while a 50-dimensional PCA embedding was used as an intermediate step for t-SNE due to computational constraints. The first 50 principal components retained 41.42% of variance.

PCA does not capture non-linear structure in the data. For this reason, PCA was not pursued further as a standalone analysis tool beyond variance inspection and preprocessing.

t-SNE was then applied to the PCA-reduced space to capture potential non-linear structure in the data and to visualise possible class separation.

In this case, no clear separation between toxic and non-toxic molecules was observed, as both classes appear highly mixed in the embedding space (Figure 7). This suggests that, in the original Morgan fingerprint space, class structure is either weakly separable or distributed across multiple overlapping substructures rather than forming distinct clusters. Figures for all endpoints can be found in Annex A.2.

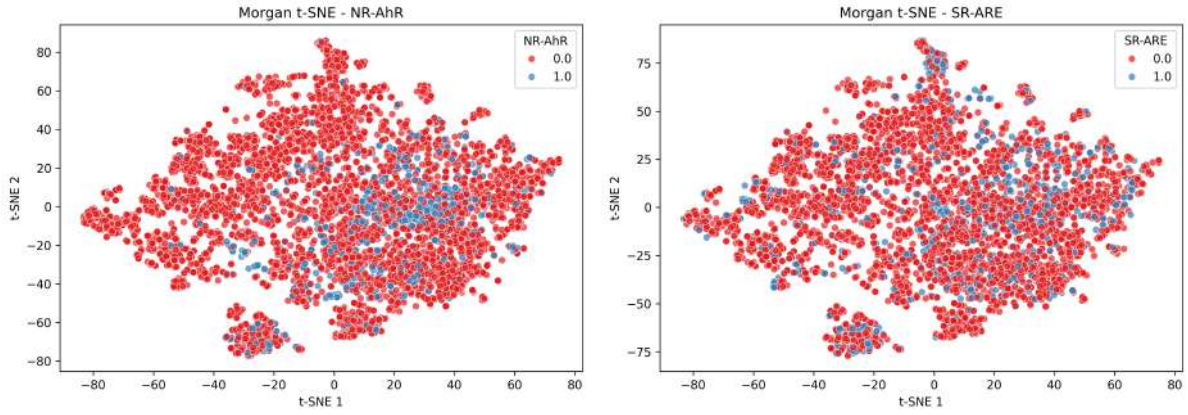


Figure 7: Scatter plot of the projection of the input data onto the t-SNE1 and t-SNE2 dimensions for the NR-AhR and SR-ARE endpoints. It can be observed that there is no clear separability in this space between positively and negatively toxic samples, indicating that the feature representations of both classes are very similar under this dimensionality reduction method.

Linear Discriminant Analysis (LDA) In addition to unsupervised dimensionality reduction, Linear Discriminant Analysis (LDA) was applied as a supervised projection method. For each toxicity endpoint, a separate one-dimensional LDA projection was computed to maximise class separation between toxic and non-toxic molecules.

The resulting LDA scores were visualised using kernel density estimates, providing an indication of how separable the two classes are along a single discriminative axis in the feature space.

LDA assumes that each class follows a Gaussian distribution with identical covariance matrices and that class separation can be captured through a linear combination of the input features. These assumptions are often violated in molecular fingerprint representations, which are highly sparse, binary, and non-Gaussian in nature. In particular, the distribution of Morgan fingerprint bits is strongly multimodal and influenced by discrete substructure presence rather than continuous variation. As a result, the linear decision boundary learned by LDA is limited in its ability to capture the underlying structure of the data.

This limitation is reflected in the observed results: class distributions overlap significantly along the LDA projection, indicating weak linear separability between toxic and non-toxic molecules.

Representative examples are shown for NR-AhR and NR-AR-LBD in Figure 8. The full set of results for all endpoints is provided in Annex A.3 .

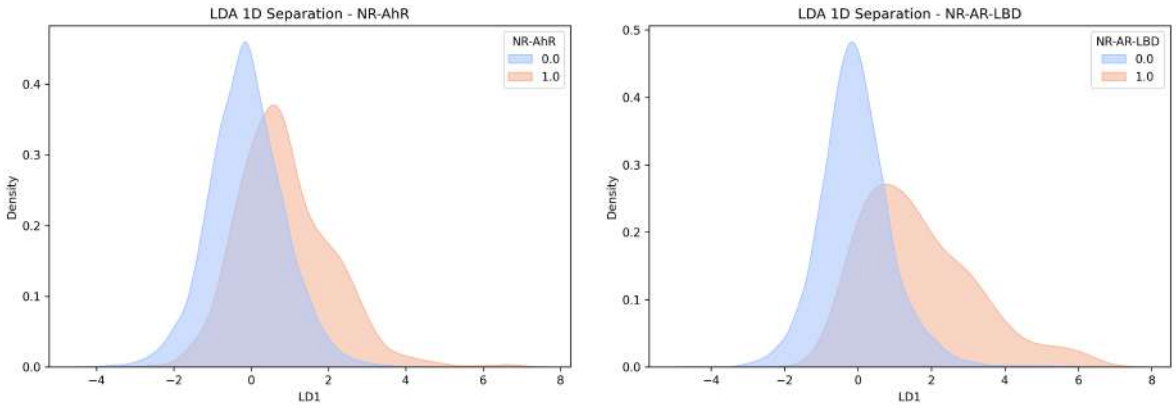


Figure 8: LDA projections for NR-AhR and NR-AR-LBD. Although there is a slight separation between the distributions, the classes are effectively overlapped and LDA is not useful for classs separation.

Despite applying both linear (PCA, LDA) and non-linear (t-SNE) dimensionality reduction techniques, no strong global separation between toxic and non-toxic compounds was observed in the reduced spaces. This is consistent with the nature of Morgan fingerprints, which are high-dimensional, sparse, and binary representations of local substructures.

In particular, PCA captures directions of maximum variance rather than class separation, while t-SNE preserves local neighbourhood structure but is sensitive to parameter choices and does not guarantee meaningful global clustering. LDA, although supervised, is limited by the strong class imbalance and the overlap between structural patterns associated with toxic and non-toxic molecules.

Overall, these results suggest that toxicity prediction in this setting does not correspond to simple geometric separation in fingerprint space, motivating the use of more expressive non-linear models and structured feature aggregation methods such as the cluster-based representations used in the subsequent analysis.

4.2 Baseline predictive models

To evaluate the predictive performance of the molecular representations, three standard machine learning models were trained independently for each Tox21 endpoint: logistic regression, random forest, and gradient boosting (LightGBM). These models provide complementary inductive biases, ranging from linear decision boundaries to non-linear tree-based interactions.

1. **Logistic regression** A logistic regression model [20] with $L2$ regularisation was trained as a linear baseline. Input features were standardised using z-score normalisation prior to training.
2. **Random forest** A random forest classifier [15] composed of 200 decision trees was trained using bootstrap aggregation. This model captures non-linear feature interactions and is robust to noise in high-dimensional sparse fingerprint representations. Class imbalance was handled using class weighting.
3. **Gradient boosting** A LightGBM [22] classifier was trained with 300 estimators, a learning rate of 0.05, and subsampling of both samples and features to reduce overfitting. Gradient boosting iteratively corrects errors of previous trees, enabling strong predictive performance on structured molecular descriptors.

4.2.1 Evaluation protocol

For each endpoint, the dataset was first split into training and test sets using an 80/20 stratified split in order to preserve the inherent class imbalance of the Tox21 dataset. Model evaluation was then performed using 5-fold stratified cross-validation on the training set to obtain robust performance estimates and to quantify variability across folds.

The models were assessed using two main metrics:

- **Accuracy**, which measures overall predictive correctness.

- **Balanced accuracy**, which accounts for the strong class imbalance by equally weighting both classes, ensuring that performance on the minority toxic class is properly reflected.

Predicted probabilities were converted into class labels using a decision threshold of 0.2 instead of the standard 0.5, increasing sensitivity toward the minority (toxic) class and reducing the number of false negatives.

Regarding the underlying optimisation criteria:

- Logistic regression optimises the logistic loss (binary cross-entropy).
- Gradient boosting models (LightGBM) optimise the binary log-loss objective.
- Random forests do not optimise an explicit differentiable loss; instead, they rely on Gini impurity as the splitting criterion in each decision tree.

Overall, the evaluation protocol combines a stratified 80/20 train-test split, internal 5-fold stratified cross-validation for stability assessment, the use of accuracy and balanced accuracy for imbalanced classification, and a reduced decision threshold of 0.2 to prioritise sensitivity to toxic compounds.

4.2.2 Classification performance

Overall, the ensemble-based models (Random Forest and Boosting) clearly outperform the linear model in terms of standard accuracy, reaching values close to 0.95–0.98 across most endpoints, while the linear model consistently remains much lower. However, this gap is significantly reduced when considering balanced accuracy, where the effect of the strong class imbalance becomes evident. In this setting, there is no single dominant method: Random Forest and Boosting alternate as the best-performing models depending on the endpoint, with each showing local advantages for specific toxicity tasks. The linear model, despite its lower overall accuracy, remains comparatively competitive in certain endpoints where class separation is weaker or noisier. Overall, these results indicate that nonlinear ensemble methods are better at capturing complex structure in the data, but meaningful gains in minority-class detection are highly endpoint-dependent. This reinforces the importance of using balanced accuracy rather than accuracy alone when evaluating models under strong class imbalance. Tables 2 and 3, as well as Figure 9 show these results.

Endpoint	Linear	Random Forest	Boosting
NR-AR	0.38 ± 0.03	**0.97** ± 0.00	**0.97** ± 0.00
NR-AR-LBD	0.54 ± 0.03	**0.98** ± 0.00	**0.98** ± 0.00
NR-AhR	0.56 ± 0.02	**0.88** ± 0.01	**0.88** ± 0.01
NR-Aromatase	0.40 ± 0.02	**0.95** ± 0.01	0.95 ± 0.01
NR-ER	0.21 ± 0.01	**0.84** ± 0.01	0.83 ± 0.01
NR-ER-LBD	0.35 ± 0.03	**0.95** ± 0.01	0.95 ± 0.01
NR-PPAR-gamma	0.33 ± 0.02	**0.97** ± 0.00	**0.97** ± 0.00
SR-ARE	0.26 ± 0.01	**0.79** ± 0.01	0.78 ± 0.00
SR-ATAD5	0.42 ± 0.02	**0.96** ± 0.00	**0.96** ± 0.00
SR-HSE	0.20 ± 0.01	**0.93** ± 0.00	**0.93** ± 0.01
SR-MMP	0.50 ± 0.01	0.83 ± 0.01	**0.85** ± 0.01
SR-p53	0.32 ± 0.02	**0.94** ± 0.00	0.93 ± 0.00

Table 2: Accuracy (ACC) across Tox21 endpoints. Values are mean ± standard deviation over cross-validation folds.

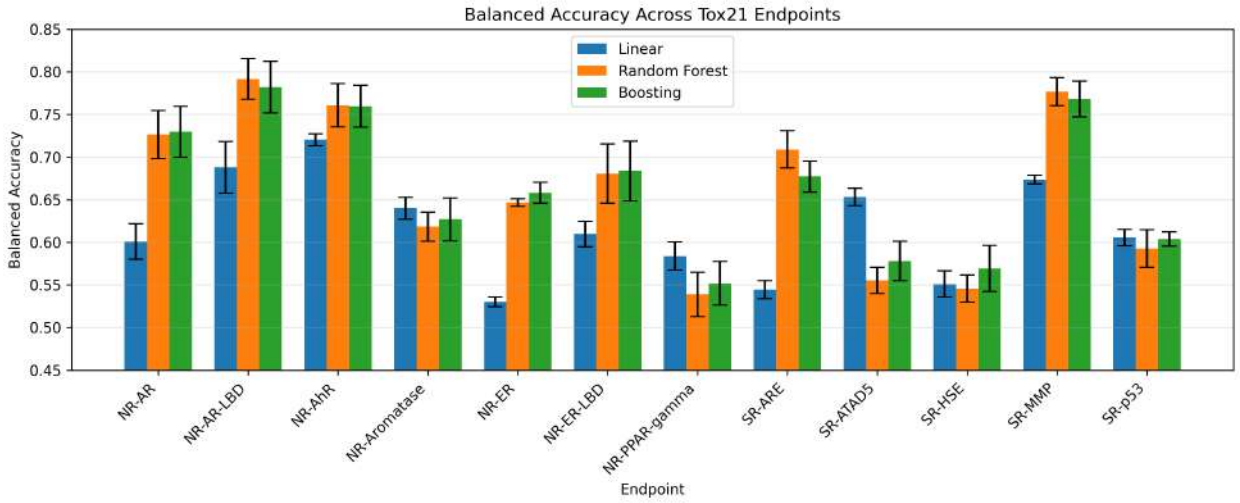


Figure 9: BACC comparison across endpoints- Random Forest and Boosting show consistent better performance than Linear, with no model exhibiting consistent good predictive performance across all endpoints.

Endpoint	Linear	Random Forest	Boosting
NR-AR	0.60 ± 0.02	0.73 ± 0.03	**0.73** ± 0.03
NR-AR-LBD	0.69 ± 0.03	**0.79** ± 0.02	0.78 ± 0.03
NR-AhR	0.72 ± 0.01	**0.76** ± 0.03	0.76 ± 0.03
NR-Aromatase	**0.64** ± 0.01	0.62 ± 0.02	0.63 ± 0.03
NR-ER	0.53 ± 0.01	0.65 ± 0.00	**0.66** ± 0.01
NR-ER-LBD	0.61 ± 0.02	0.68 ± 0.04	**0.68** ± 0.04
NR-PPAR-gamma	**0.58** ± 0.02	0.54 ± 0.03	0.55 ± 0.03
SR-ARE	0.54 ± 0.01	**0.71** ± 0.02	0.68 ± 0.02
SR-ATAD5	**0.65** ± 0.01	0.56 ± 0.02	0.58 ± 0.02
SR-HSE	0.55 ± 0.02	0.55 ± 0.02	**0.57** ± 0.03
SR-MMP	0.67 ± 0.01	**0.78** ± 0.02	0.77 ± 0.02
SR-p53	**0.61** ± 0.01	0.59 ± 0.02	0.60 ± 0.01

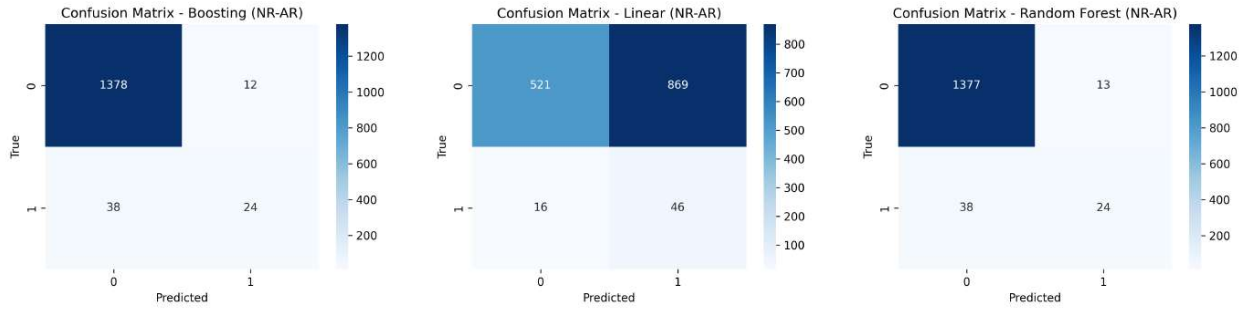
Table 3: Balanced accuracy (BACC) across Tox21 endpoints. Values are mean ± standard deviation over cross-validation folds.

4.2.3 Confusion matrices

Confusion matrices were computed for each model and endpoint to analyse classification errors in more detail. Note that both the Boosting and Random Forest models tend to predict the negative class for a large proportion of samples. This behaviour leads to a high false negative rate (FNR), which is particularly problematic in this context, as toxic compounds being incorrectly classified as non-toxic poses a significant safety risk.

In contrast, the Linear model exhibits the opposite behaviour, tending to classify a substantial number of non-toxic molecules as toxic. While this results in a lower false negative rate and is therefore safer from a toxicity screening perspective, it also produces a high false positive rate (FPR), which limits its practical utility due to excessive over-prediction of toxicity.

All models exhibit similar behaviours for all endpoints. The full set of confusion matrices is available in Annex B, an example for endpoint NR-AR is shown in Figures 10a, 10b and 10c.



(a) Boosting: significant bias towards the negative class, failing to classify correctly almost all positives. Effectively unusable.

(b) Logistic Regression: opposite behaviour to Boosting, with an extremely high FPR while still misclassifying ∼30% of positives.

(c) Random Forest: near-identical behaviour to Boosting, classifying almost every compound as non-toxic, with the FNR nearly doubling the TPR.

Figure 10: Confusion matrices for the three baseline models on endpoint NR-AR. All three models exhibit a strong majority-class bias consistent with the class imbalance of the dataset.

These confusion matrices suggest that the Random Forest and Boosting models produce highly confident (near-binary) probability outputs, which makes their predictions largely insensitive to threshold adjustments. In contrast, the linear model tends to generate more calibrated and smoothly distributed probability scores. As a result, changes in the decision threshold have a greater impact on its classification behavior, leading to more noticeable variations in the confusion matrix. This hypothesis is furtherly explored in the next section.

4.2.4 Effect of decision threshold and probability calibration

During threshold tuning, a marked difference in model behaviour was observed across classifiers. Although all models were trained to output probabilistic scores, their score distributions and sensitivity to threshold changes differed substantially.

The Linear model was highly sensitive to the decision threshold. Small changes in the threshold led to large variations in predicted labels. This behaviour is consistent with a relatively well-calibrated but weakly separated probability distribution, where toxic and non-toxic molecules show strong overlap in predicted probabilities. As a result, classification performance depends heavily on selecting an appropriate operating point.

In contrast, the Boosting model exhibited highly discrete probability outputs, with predictions concentrated near 0 and 1. However, most samples were pushed towards values close to 0, even for toxic molecules. This indicates strong confidence in the negative class and suggests an asymmetric decision boundary. Importantly, the model only changed its predictions under extreme threshold values (e.g., 0.05 or 0.95), reflecting low sensitivity in the mid-range due to saturated predictions.

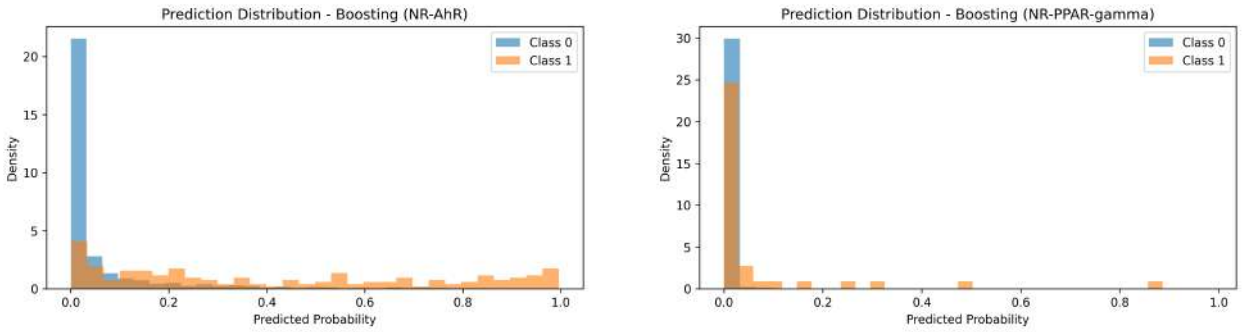


Figure 11: Predicted probability distributions of Boosting model for endpoints NR-AhR and NR-PPAR-gamma. The model outputs $P \approx 0$ for both endpoints. NR-AhR seems more separated, but still some overlap is found. NR-PPAR-gamma probabilities completely overlap between classes.

The Random Forest model behaved in an intermediate manner. While it produced more differentiated probability estimates than Boosting, its outputs were still heavily skewed toward low probabilities. Some separation between classes was visible, but high-confidence toxic predictions were relatively rare. This suggests partial discrimination capacity but also a conservative bias toward the majority or negative class.

These patterns highlight fundamental differences in probability calibration across models: the Linear model produces smooth but overlapping scores, Boosting produces overconfident and skewed predictions, and Random Forest lies between both regimes with moderate separation but limited high-probability toxic assignments.

Moreover, model behaviour varies substantially depending on the toxicity endpoint. Figures 11, 12 and 13 show the difference in the distributions of the predictions.

Clear differences can be observed between both cases. For NR-AhR, the predicted probabilities exhibit a higher degree of separation between toxic and non-toxic molecules, indicating that the models are able to extract a more structured signal from the input representation. In contrast, for NR-PPAR-gamma, the probability distributions are heavily overlapping and tend to concentrate near extreme values, with little meaningful separation between classes.

These results reinforce the idea that endpoint-specific difficulty plays a crucial role in model performance and calibration. Some endpoints contain more learnable signal, while others appear highly noisy or weakly separable in the Morgan fingerprint space.

4.3 DNN

The trained DNN is taken from the work of *Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations* [2]. It can be visualized in Table

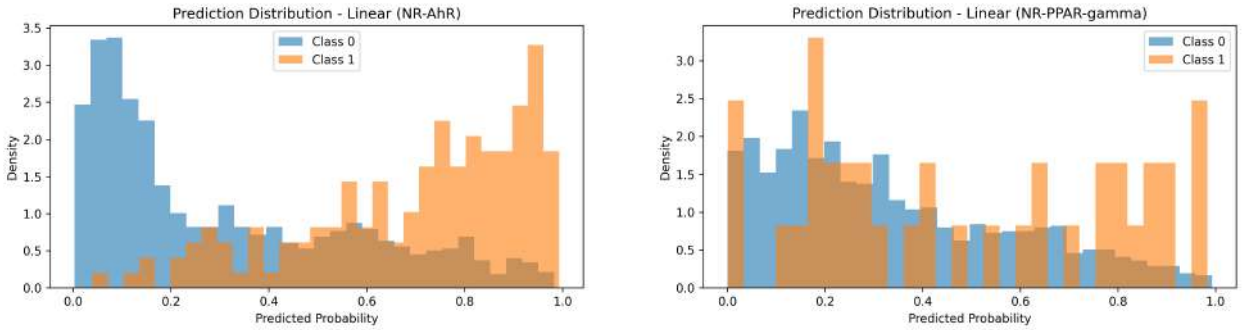


Figure 12: Predicted probability distributions of the Linear model for the NR-AhR and NR-PPAR-gamma endpoints. A substantial overlap between classes is observed in both cases. For NR-AhR, the non-toxic and toxic classes are correctly skewed towards $P = 0$ and $P = 1$, respectively, indicating a reasonable degree of separability. In contrast, NR-PPAR-gamma shows no clear separation between classes, with heavily overlapping distributions and limited discriminative power.

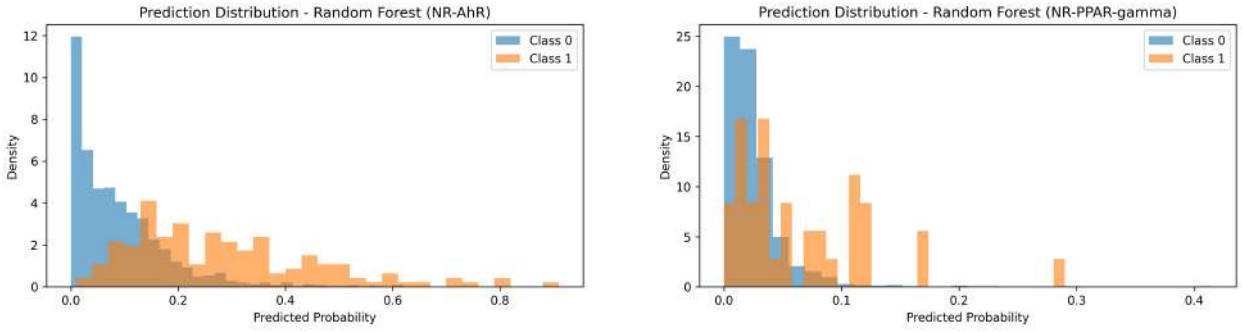


Figure 13: Predicted probability distributions of Random Forest model for endpoints NR-AhR and NR-PPAR-gamma. NR-AhR negatives are skewed towards $P = 0$, but there is significant overlap with the positive class. NR-PPAR-gamma is again unseparable.

4.

Layer	Output Dimension
Input Morgan fingerprint	2048
Linear + BatchNorm + LeakyReLU	1024
Dropout ($p = 0.2$)	1024
Linear + BatchNorm + LeakyReLU	512
Dropout ($p = 0.2$)	512
Linear + LeakyReLU	256
Output Linear Layer	1
Sigmoid	1

Table 4: Architecture of the single-task deep neural network used as the main predictive model in this study. The network takes as input 2048-dimensional Morgan fingerprints and progressively reduces dimensionality through three fully connected hidden layers (1024, 512, and 256 units), each incorporating non-linear activations and regularisation techniques such as batch normalization and dropout. The final layer produces a single output passed through a sigmoid activation to estimate the probability of toxicity for each Tox21 endpoint independently.

The deep neural network baseline was evaluated on all twelve Tox21 endpoints using five independent random seeds (122–126). For each seed, a separate train/validation/test split was generated, and the model achieving the highest validation ROC-AUC during training was selected for final evaluation on the corresponding test set. Performance metrics were subsequently aggregated across all seeds, reporting the mean and standard deviation for each endpoint as seen in Table 5.

To address the pronounced class imbalance present in several Tox21 tasks, a weighted binary cross-entropy loss was employed through the `BCEWithLogitsLoss` function. The positive class weight was computed as the ratio between negative and positive samples for each endpoint. However, to avoid excessively large class weights dominating the optimization process and causing unstable training behaviour, the positive class weight was capped at a maximum value of 10. This limitation provided a balance between compensating for class imbalance and maintaining stable gradient updates throughout training. Additionally, the threshold for negatives was set on 0.2.

More refined tuning of the decision threshold and class-weighting scheme could potentially yield additional performance gains. However, the observed variability across random seeds suggests that model behavior is already sensitive to data splits and initialization effects. In this context, identifying a single set of hyperparameters that consistently generalizes well across all seeds appears inherently challenging. As a result, extensive hyperparameter optimization was not prioritized, since improvements on one seed often did not translate reliably

to others, making such tuning somewhat ambitious given the stability constraints of the setup.

Endpoint	ACC	BACC	ROC-AUC
NR-AR	0.81 ± 0.30	0.67 ± 0.10	0.73 ± 0.07
NR-AR-LBD	0.87 ± 0.31	0.78 ± 0.12	0.82 ± 0.06
NR-AhR	0.85 ± 0.14	0.66 ± 0.07	0.83 ± 0.05
NR-Aromatase	0.85 ± 0.15	0.65 ± 0.08	0.76 ± 0.08
NR-ER	0.58 ± 0.27	0.58 ± 0.05	0.66 ± 0.04
NR-ER-LBD	0.86 ± 0.19	0.66 ± 0.07	0.78 ± 0.05
NR-PPAR-gamma	0.86 ± 0.30	0.63 ± 0.06	0.75 ± 0.06
SR-ARE	0.77 ± 0.10	0.67 ± 0.05	0.76 ± 0.04
SR-ATAD5	0.94 ± 0.05	0.65 ± 0.09	0.77 ± 0.05
SR-HSE	0.92 ± 0.03	0.58 ± 0.06	0.67 ± 0.04
SR-MMP	0.78 ± 0.19	0.74 ± 0.08	0.83 ± 0.04
SR-p53	0.92 ± 0.04	0.68 ± 0.03	0.79 ± 0.03
Mean	0.72 ± 0.30	0.65 ± 0.09	0.74 ± 0.09

Table 5: Classification performance of the deep neural network across the 12 Tox21 endpoints, reported as mean $\pm$ std over five seeds. Results show a consistent pattern of moderate predictive performance, with mean ROC-AUC of 0.741 ± 0.088 and mean balanced accuracy of 0.651 ± 0.092. Performance varies substantially across endpoints. The relatively high standard deviations in some endpoints, particularly in accuracy, highlight sensitivity to random initialization and the underlying class imbalance of the dataset.

The results show a generally fair and consistent performance of the deep neural network baseline across the different Tox21 endpoints, with competitive ROC-AUC and balanced accuracy values on several tasks. However, the performance is not uniformly strong across all endpoints, and there is still noticeable variability depending on the task and the random seed used during training. In particular, some endpoints exhibit relatively high standard deviations, indicating that model stability is sensitive to the specific data split and initialization, especially under the strong class imbalance present in certain tasks. This highlights that, while the model is capable of learning meaningful predictive patterns, its reliability is not fully uniform across the dataset.

An important aspect of this evaluation is that performance is reported per endpoint and across multiple random seeds, which is not always the case in related literature. Many previous studies report only aggregated performance metrics or a single train/test split, without providing a detailed breakdown across individual tasks or quantifying variability across seeds. As a result, the observed instability in certain endpoints and the spread in performance metrics may be underreported in other works. By explicitly including both

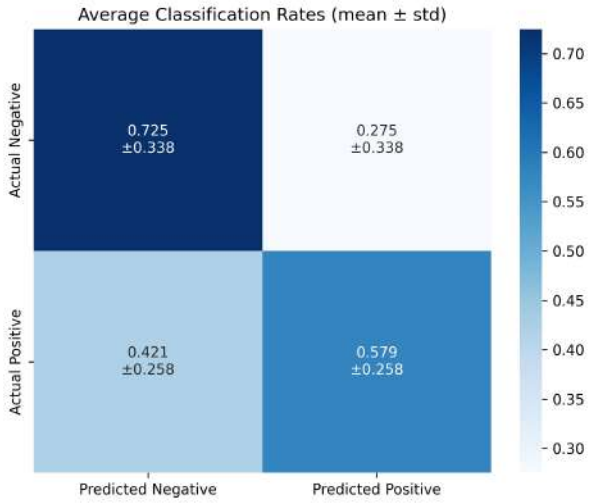


Figure 14: Averaged confusion matrix for the DNN across all Tox21 endpoints and random seeds.The model achieves a TNR of 0.725 and a TPR of 0.579, reflecting a bias towards the non-toxic majority class that is consistent with the class imbalance of the dataset. The higher FNR (0.421) relative to the FPR (0.275) indicates that the model is more prone to missing toxic compounds than to incorrectly flagging non-toxic ones.

per-task results and variability across seeds, this evaluation provides a more transparent and realistic assessment of model robustness under different training conditions.

To gain a better general insight into the model performance, the averaged confusion matrix for the model is shown in Figure 14.

Fairly high positive-class weights, together with the lowered decision threshold, increase the number of predicted positive cases. However, there remains a substantial proportion of false negatives (42%). These correspond to compounds that the model fails to correctly identify, likely due to strong structural similarities between toxic and non-toxic molecules, which limit the separability of the feature space.

4.4 XAI

4.4.1 Raw fingerprints

The values reported in Table 6 correspond to permutation importance computed directly on the raw 2048-bit Morgan fingerprint representation. Stability is evaluated across seed pairs to assess how consistent the learned feature attributions are under different random initializations. The Jaccard index is computed over the top-100 most important bits for each model, providing a measure of agreement in the highest-ranked features. In contrast,

Endpoint	Jaccard Top-k	Spearman Correlation
NR-AR	0.37 ± 0.05	0.62 ± 0.06
NR-Aromatase	0.29 ± 0.13	0.48 ± 0.09
NR-PPAR-gamma	0.38 ± 0.05	0.52 ± 0.04
SR-HSE	0.43 ± 0.05	0.60 ± 0.03
NR-AR-LBD	0.29 ± 0.06	0.52 ± 0.07
NR-ER	0.37 ± 0.06	0.61 ± 0.06
SR-ARE	0.41 ± 0.05	0.66 ± 0.03
SR-MMP	0.48 ± 0.05	0.64 ± 0.03
NR-AhR	0.45 ± 0.04	0.65 ± 0.04
NR-ER-LBD	0.40 ± 0.03	0.60 ± 0.03
SR-ATAD5	0.45 ± 0.08	0.59 ± 0.09
SR-p53	0.29 ± 0.16	0.48 ± 0.14

Table 6: Stability of explanations across random seeds for raw fingerprint-bit explanations using permutation importance. Stability is measured using top-k Jaccard similarity and Spearman correlation. Lower values (highlighted in red) indicate weak reproducibility of feature importance rankings across different initialisations, reflecting high sensitivity to noise at the individual bit level.

Spearman correlation is computed over all 2048 bits, offering a global view of rank consistency across the entire feature space.

Overall, the results are moderately encouraging. Although the full-rank agreement is relatively weak, indicating that importance rankings vary substantially across seeds, there is still a noticeable overlap in the most influential features: roughly within the top 5% of bits, models tend to agree at a reasonable level. However, the low Spearman correlations suggest that the global ordering of feature importance is not stable. This discrepancy motivated the shift away from raw-bit explanations toward a clustered representation of the input space, where stability and interpretability can be improved by aggregating correlated features into higher-level structural groups.

4.4.2 Clustered bits

The input space was structured by clustering the Morgan fingerprint features rather than the molecules themselves. Each molecule was first encoded as a 2048-bit Morgan fingerprint (radius 2), producing a binary vector that captures the presence of local substructures. To measure relationships between features, the fingerprint matrix was transposed so that each bit was treated as a variable, and pairwise Pearson correlations were computed across the training set. These correlations were converted into a distance matrix using

$$d = 1 - |\text{corr}|\,,$$

so that both positively and negatively correlated bits were considered similar in magnitude. Hierarchical agglomerative clustering with average linkage was then applied to this precomputed distance matrix, grouping bits that exhibited similar activation patterns across molecules. This procedure yields clusters of structurally and statistically related fingerprint features, effectively reducing the original high-dimensional representation into a set of 300 interpretable feature groups that can be analyzed collectively in downstream permutation importance experiments.

Permutation importance The clustered permutation importance results show a clear improvement in stability compared to the raw bit-level analysis. Across endpoints, Jaccard top-50 values are consistently higher and more homogeneous, ranging from approximately 0.37 to 0.59, while Spearman correlations remain moderate to strong. This indicates that when features are aggregated into chemically meaningful clusters, the ranking of important groups becomes more consistent across seeds. In particular, SR tasks such as SR-ARE and SR-MMP show strong agreement, suggesting that higher-level structural representations reduce sensitivity to initialization noise. Overall, clustering substantially improves interpretability stability, especially in the most influential feature subsets.

SHAP SHAP-based explanations computed on clustered features exhibit the highest stability among all evaluated methods. Jaccard top-50 values are consistently high, typically in the range of 0.72 to 0.81, and Spearman correlations are extremely strong, generally above 0.93 across all endpoints. This reflects a high degree of agreement in both the top-ranked clusters and the global ordering of feature importance across seeds. The low variance further suggests that SHAP at the cluster level provides a robust and reliable attribution method, with minimal sensitivity to model initialization or data splitting effects.

Sensitivity analysis The neural sensitivity analysis applied to clustered inputs produces intermediate stability results. Jaccard top-50 values fall roughly between 0.41 and 0.60, while Spearman correlations range from about 0.71 to 0.81. This indicates that while there is reasonable agreement on the most influential clusters, the full ranking of features still shows noticeable variation across seeds. Compared to permutation-based clustering, sensitivity analysis tends to preserve a slightly more coherent global ordering, but still exhibits variability that reflects dependence on gradient-based attributions and model training dynamics.

Endpoint	Jaccard Top-k	Spearman Correlation
NR-AR	0.46 ± 0.11	0.75 ± 0.06
NR-AR-LBD	0.41 ± 0.10	0.65 ± 0.06
NR-AhR	0.50 ± 0.08	0.73 ± 0.05
NR-Aromatase	0.37 ± 0.07	0.63 ± 0.08
NR-ER	0.45 ± 0.07	0.71 ± 0.07
NR-ER-LBD	0.48 ± 0.07	0.71 ± 0.04
NR-PPAR-gamma	0.43 ± 0.06	0.69 ± 0.02
SR-ARE	0.59 ± 0.03	0.82 ± 0.02
SR-ATAD5	0.54 ± 0.10	0.74 ± 0.08
SR-HSE	0.51 ± 0.06	0.71 ± 0.05
SR-MMP	0.53 ± 0.05	0.78 ± 0.04
SR-p53	0.46 ± 0.13	0.63 ± 0.12

Table 7: Explanation stability across random seeds for permutation importance computed at the cluster level. Clustering substantially improves robustness, yielding higher and more consistent agreement between seed-specific rankings. Spearman correlations are consistently high across all endpoints (0.65 – 0.82), indicating strong preservation of global importance ordering. Jaccard scores are lower but notably improved relative to the bit-level case, reflecting residual variability in the exact composition of top-ranked clusters.

Endpoint	Jaccard Top-k	Spearman Correlation
NR-AR	0.74 ± 0.04	0.94 ± 0.01
NR-AR-LBD	0.73 ± 0.03	0.94 ± 0.01
NR-AhR	0.77 ± 0.04	0.95 ± 0.01
NR-Aromatase	0.73 ± 0.05	0.93 ± 0.01
NR-ER	0.79 ± 0.05	0.94 ± 0.01
NR-ER-LBD	0.77 ± 0.04	0.94 ± 0.01
NR-PPAR-gamma	0.73 ± 0.05	0.94 ± 0.01
SR-ARE	0.75 ± 0.03	0.95 ± 0.01
SR-ATAD5	0.78 ± 0.02	$0.95 < 0.00$
SR-HSE	0.75 ± 0.03	0.94 ± 0.01
SR-MMP	0.81 ± 0.03	0.96 ± 0.01
SR-p53	0.75 ± 0.04	0.95 ± 0.01

Table 8: Explanation stability for SHAP values at the cluster level. SHAP exhibits the highest overall stability among the evaluated methods, with consistently high Spearman correlations (0.93 – 0.96), indicating strong agreement in global cluster importance rankings across seeds. Jaccard scores are also substantially higher than in other methods. SHAP provides more robust and reproducible attributions at the cluster level, making it the most reliable method for identifying stable molecular drivers of toxicity across model initialisations.

Endpoint	**Jaccard Top-k**	**Spearman Correlation**
NR-AR	0.57 ± 0.05	0.78 ± 0.02
NR-AR-LBD	0.48 ± 0.04	0.75 ± 0.03
NR-AhR	0.60 ± 0.05	0.78 ± 0.02
NR-Aromatase	0.41 ± 0.05	0.72 ± 0.03
NR-ER	0.48 ± 0.04	0.79 ± 0.01
NR-ER-LBD	0.50 ± 0.04	0.75 ± 0.02
NR-PPAR-gamma	0.50 ± 0.04	0.74 ± 0.03
SR-ARE	0.54 ± 0.05	0.79 ± 0.01
SR-ATAD5	0.47 ± 0.05	0.74 ± 0.02
SR-HSE	0.57 ± 0.06	0.75 ± 0.02
SR-MMP	0.56 ± 0.04	0.81 ± 0.01
SR-p53	0.50 ± 0.04	0.73 ± 0.03

Table 9: Explanation stability across seeds for sensitivity analysis at the cluster level. Overall, sensitivity analysis shows moderate-to-high stability, with Spearman correlations consistently above 0.70 across all endpoints, indicating reasonably stable global ranking of cluster importance. However, Jaccard scores remain lower than those obtained with SHAP, reflecting greater variability in the exact composition of top-ranked clusters. While gradient-based attributions capture a consistent global signal, they are more sensitive to local fluctuations in model parameters, leading to reduced reproducibility at the level of specific explanatory clusters.

4.5 Analysis of explanations

Figure 15 shows the aggregated distribution of active versus inactive molecules for the top permutation-importance clusters across endpoints (separate histograms can be found in Annex C) For the majority of clusters, the number of inactive molecules exceeds the number of active ones. This is consistent with two properties of the dataset and the permutation importance metric itself. First, with only around 5% of molecules being toxic per endpoint, any cluster activation pattern reflecting the overall chemical space will naturally be dominated by non-toxic compounds. Second, permutation importance measures the change in prediction caused by shuffling a cluster's features; since the model must correctly classify the majority (non-toxic) class to achieve low error, clusters that define the structural signature of non-toxic molecules can produce large prediction changes when permuted, simply because they affect a larger fraction of the dataset. High permutation importance therefore does not necessarily imply a cluster is a *toxic* driver — a distinction that motivates the SHAP-based directionality and enrichment analysis introduced below.

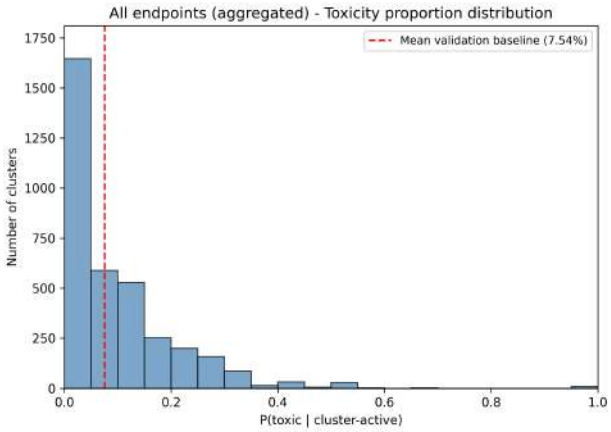


Figure 15: Aggregated distribution of toxic molecules for the top permutation-importance clusters across all endpoints. The distribution is strongly skewed toward inactive molecules, consistent with the low baseline toxicity rate (7.54%). A clear peak appears at low active fractions, followed by a gradual decrease up to approximately 0.5, with only a few outlier clusters near 1 corresponding to very small, highly specific clusters (often involving only a handful of molecules). Overall, the results reflect both the class imbalance and the tendency of permutation importance to highlight clusters dominated by the majority class.

4.5.1 Top enriched clusters by permutation consensus

To further illustrate the limitation discussed above — that high permutation importance does not necessarily imply a cluster is a toxic driver — the same ground-truth enrichment methodology used later in this chapter was applied directly to the permutation-importance

consensus clusters (clusters ranked in the top-50 across all five seeds). Table 10 reports the highest-enrichment single cluster per endpoint.

Endpoint	Cluster	n_{active}	% positive	EF actual
NR-AR	Cluster 114	126	6.60	1.47
NR-AR-LBD	Cluster 265	274	3.73	0.91
NR-AhR	Cluster 265	274	21.49	1.82
NR-Aromatase	Cluster 121	90	14.49	2.99
NR-ER	Cluster 234	53	37.50	**3.85**
NR-ER-LBD	Cluster 234	53	23.26	**6.02**
NR-PPAR-gamma	Cluster 265	274	3.17	1.12
SR-ARE	Cluster 265	274	21.28	1.26
SR-ATAD5	Cluster 265	274	7.94	2.27
SR-HSE	Cluster 265	274	8.26	1.43
SR-MMP	Cluster 299	29	29.17	1.86
SR-p53	Cluster 121	90	15.19	2.21

Table 10: Top single cluster by ground-truth enrichment factor per endpoint, among permutation-consensus clusters with $n_{\text{active}} \geq 10$. Cluster 265 appears frequently across multiple endpoints but does not consistently show enrichment above the baseline toxicity rate (7.54%), suggesting it likely captures broadly prevalent structural patterns rather than true toxicological signals. In contrast, Cluster 234 exhibits strong enrichment in NR-ER and NR-ER-LBD, two mechanistically related endpoints, indicating a more specific and potentially meaningful toxic driver. This consistency across related assays makes Cluster 234 a particularly promising candidate for a shared endocrine disruption-related structural motif.

Recurrence of Cluster 265. Cluster 265 is the top enriched single cluster for five of the twelve endpoints (NR-AR-LBD, NR-AhR, NR-PPAR-gamma, SR-ARE, and SR-HSE), reflecting its large size and broad activation across the test set ($n_{\text{active}} = 274$, the largest single-cluster activation observed). However, its enrichment factor varies substantially across these endpoints, from 0.91 for NR-AR-LBD — effectively no enrichment over baseline — to 1.82 for NR-AhR. This is a direct illustration of the issue raised above: Cluster 265 is consistently flagged as permutation-important, most likely because it affects a large fraction of the dataset and therefore produces large prediction changes when shuffled, but this importance does not translate into a consistent or even meaningfully positive enrichment in true positives. A cluster can thus be permutation-important for an endpoint without being a toxic driver for that endpoint.

Cluster 234 and the estrogen receptor pathway. The most notable result in Table 10 is Cluster 234, which achieves the highest enrichment factors of any combination in this

analysis: $\text{EF}_{\text{actual}} = 3.85$ for NR-ER and $\text{EF}_{\text{actual}} = 6.02$ for NR-ER-LBD. Both endpoints correspond to the estrogen receptor pathway (full receptor and ligand-binding domain assays respectively), and Cluster 234 has adequate support in both cases ($n_{\text{active}} = 53$). This cross-endpoint consistency — the same cluster strongly enriched across two mechanistically related assays — is a promising signal, and motivates checking whether this same cluster also emerges from the SHAP-based consensus analysis introduced next.

Implications. Taken together, these results show that permutation importance, even when restricted to clusters that are consensus across all five seeds, is not a reliable indicator of which clusters drive toxicity: enrichment factors for the top permutation-consensus cluster range from below 1 (NR-AR-LBD) to over 6 (NR-ER-LBD), with no consistent pattern relating importance rank to enrichment. This motivates the SHAP-based directionality and enrichment analysis presented in the remainder of this chapter, which combines a signed attribution score with ground-truth enrichment to explicitly distinguish toxic drivers from clusters the model relies on for other reasons.

4.5.2 SHAP analysis

Given that SHAP at the cluster level demonstrated the highest stability across seeds (Jaccard top-50 ranging from 0.73 to 0.81, Spearman > 0.93 across all endpoints, Table 8), it was selected as the basis for the structural interpretation of toxicity predictions. SHAP was also specially useful because it provides a direction of importance, gaining insight into what the model uses to distinguish toxic from non-toxic.

For each endpoint, a consensus set of clusters was defined as those appearing in the top-50 by SHAP magnitude for *every* seed. This intersection requirement ensures that only clusters whose importance is robust to random initialisation are retained, filtering out attributions driven by a single model instance. Beyond magnitude, a directionality score $\bar{\phi}^{+}$ was computed for each consensus cluster as the mean signed SHAP value over molecules where the cluster is active, summed across the cluster's features:

$$\bar{\phi}^{+} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \sum_{j \in C} \phi_{ij} \tag{4.1}$$

where $\mathcal{A}$ is the set of molecules whose mean bit rate across cluster C exceeds the cluster's test-set mean, and ϕ_{ij} is the SHAP value of feature j for molecule i. A positive $\bar{\phi}^{+}$ indicates that the cluster pushes the model toward a toxic prediction when active; a negative value indicates a protective effect.

To validate that the SHAP signal reflects a genuine structural pattern rather than a model

artefact, the actual enrichment factor was computed for each consensus cluster on the held-out test set:

$$\mathrm{EF}_{\mathrm{actual}} = \frac{\%\text{ toxic among cluster-active molecules}}{\%\text{ toxic in the validation set}} \tag{4.2}$$

Combining the two signals yields a two-source classification of each consensus cluster. A cluster is labelled a **toxic driver** if $\mathrm{EF}_{\mathrm{actual}} > 1$ and $\bar{\phi}^{+} > 0$, meaning it is enriched in ground-truth positives and the model assigns it a toxicity-increasing role. Conversely, a cluster is labelled **protective** if $\mathrm{EF}_{\mathrm{actual}} < 1$ and $\bar{\phi}^{+} < 0$. Clusters where the two signals disagree are excluded from interpretation, as they indicate model–data inconsistency rather than a learnable structural rule.

Table 11 summarises the results. The number of toxic drivers ranges from 4 (NR-AR-LBD, SR-ATAD5) to 13 (NR-Aromatase, NR-ER, NR-ER-LBD, SR-p53), reflecting differences in how structurally specific each endpoint's toxicity signal is.

Endpoint	**Toxic drivers**	**Protective**	**Top toxic cluster**
NR-AR	7	10	Cluster 8
NR-Aromatase	13	3	Cluster 121
NR-PPAR-gamma	6	10	Cluster 226
SR-HSE	8	10	Cluster 33
NR-AR-LBD	4	13	Cluster 43
NR-ER	13	9	Cluster 43
SR-ARE	9	13	Cluster 64
SR-MMP	11	9	Cluster 33
NR-AhR	7	14	Cluster 121
NR-ER-LBD	13	2	Cluster 43
SR-ATAD5	4	13	Cluster 265
SR-p53	13	5	Cluster 92

Table 11: Number of consensus clusters classified as toxic drivers ($\mathrm{EF}_{\mathrm{actual}} > 1$ and $\bar{\phi}^{+} > 0$) and protective ($\mathrm{EF}_{\mathrm{actual}} < 1$ and $\bar{\phi}^{+} < 0$) for each Tox21 endpoint, together with the highest-enrichment toxic driver cluster per endpoint. Toxic-driver counts vary considerably across endpoints, from 4 to 13, reflecting differences in the structural specificity of each toxicity signal. Several clusters recur as the top toxic driver across multiple endpoints—Cluster 43 leads three nuclear-receptor endpoints (NR-AR-LBD, NR-ER, NR-ER-LBD) and Cluster 121 leads two (NR-Aromatase, NR-AhR)—suggesting shared structural motifs that broadly modulate receptor-mediated toxicity. Endpoints with a high protective count relative to toxic drivers (e.g. NR-AhR: 7 drivers, 14 protective; SR-ATAD5: 4 drivers, 13 protective) indicate that the model has also learned structural features whose presence actively suppresses the toxic prediction.

To translate consensus clusters into chemically interpretable labels, each cluster was annotated with its most characteristic functional groups via an enrichment-based procedure, applied only to clusters that passed the consensus filter described previously (top-50 by SHAP magnitude across all five seeds).

For each consensus cluster, the set of *active molecules* was first identified: training molecules whose mean bit rate over the cluster's features exceeds the cluster's mean activation across the training set. A library of 32 named functional groups, defined by SMARTS patterns (including amines, amides, halogens, sulfur-containing groups, aromatic heterocycles such as furan, thiophene, and indole, and reactive carbonyls such as aldehydes and ketones), was matched directly against each active molecule using RDKit substructure search.

For each functional group g, an enrichment score was computed as

$$E_g = \frac{f_g^{\text{active}}}{f_g^{\text{background}}} \tag{4.3}$$

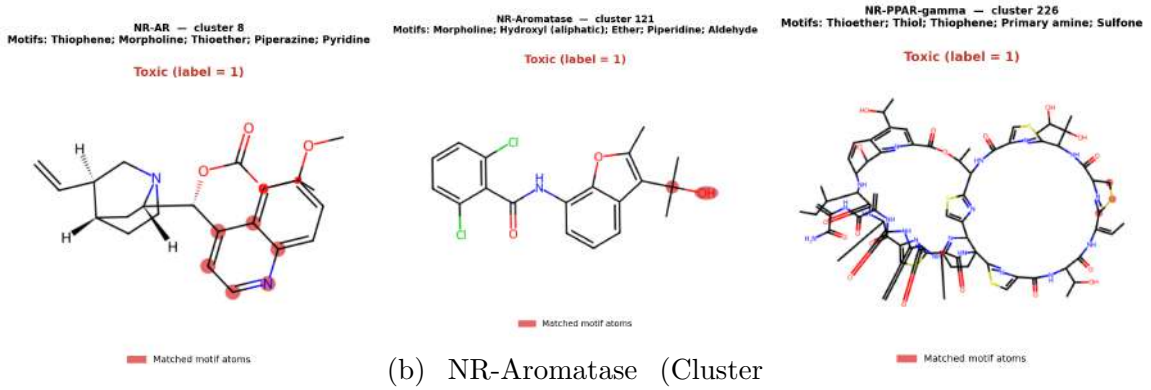
where f_g^{active} is the fraction of cluster-active molecules containing g, and $f_g^{\text{background}}$ is the fraction of all training molecules containing g. This enrichment score identifies functional groups that are disproportionately represented among the molecules activating a given cluster, relative to their overall prevalence in the dataset, rather than simply reporting the most common groups overall (which would trivially favour ubiquitous motifs such as the benzene ring). For each consensus cluster, the five functional groups with the highest enrichment scores were retained as its label, together with their corresponding enrichment values.

This procedure was applied once per unique consensus cluster across all endpoints, avoiding redundant computation for clusters that appear as consensus drivers for multiple tasks (e.g. Cluster 43, which is consensus for three nuclear receptor endpoints). The resulting functional group labels are reported alongside the SHAP importance and enrichment factor results in Table 11, and form the basis for the literature correspondence discussed below.

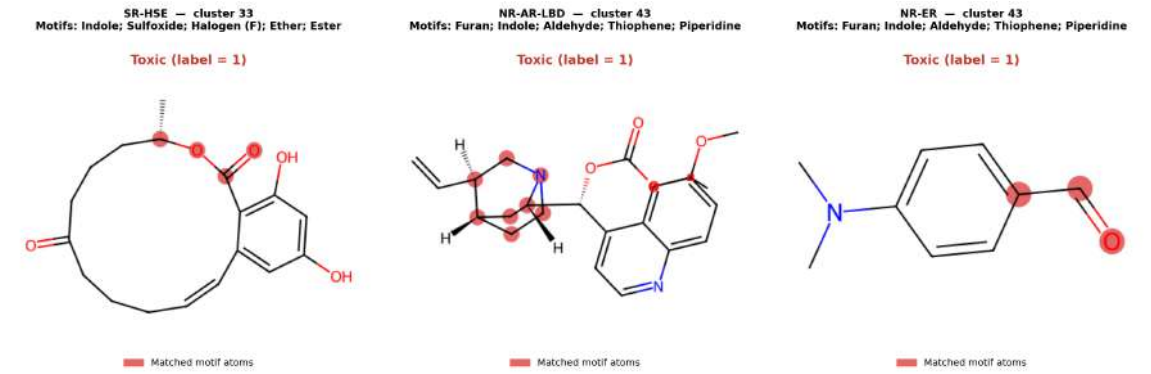
Several structural families recur across endpoints, suggesting shared toxicity mechanisms rather than endpoint-specific noise. Cluster 43 (Furan; Indole; Aldehyde; Thiophene; Piperidine) is the top toxic driver for NR-AR-LBD, NR-ER, and NR-ER-LBD, with enrichment factors of 3.80, 2.62, and 4.58 respectively. The consistent signal across three nuclear receptor endpoints suggests that the structural families captured by this cluster interact with a shared binding.

Cluster 33 (Indole; Sulfoxide; Halogen(F); Ether; Ester) is the top driver for both SR-HSE and SR-MMP. Notably, its protective counterpart at both endpoints is Cluster 29 (Aldehyde; Thiophene; Thioether; Sulfoxide; Piperazine), which shows EF < 1 and $\bar{\phi}^+ < 0$ in both cases. The fact that the same pair of clusters appears as toxic driver and protective cluster across two distinct stress-response endpoints suggests a structurally coherent signal rather than a statistical coincidence.

Cluster 121 (Morpholine; Hydroxyl; Ether; Piperidine; Aldehyde) is the top driver for both NR-Aromatase and NR-AhR. These are mechanistically distinct endpoints, so the shared structural signal may reflect a common scaffold preference in the dataset rather than a single underlying mechanism — a distinction that cannot be resolved from fingerprint-level analysis alone. Figures 16a–17f show a representative toxic molecule from the test set for each endpoint, drawn from the pool of cluster-active ground-truth positives. Atoms matching the top cluster's functional group motifs are highlighted in red.



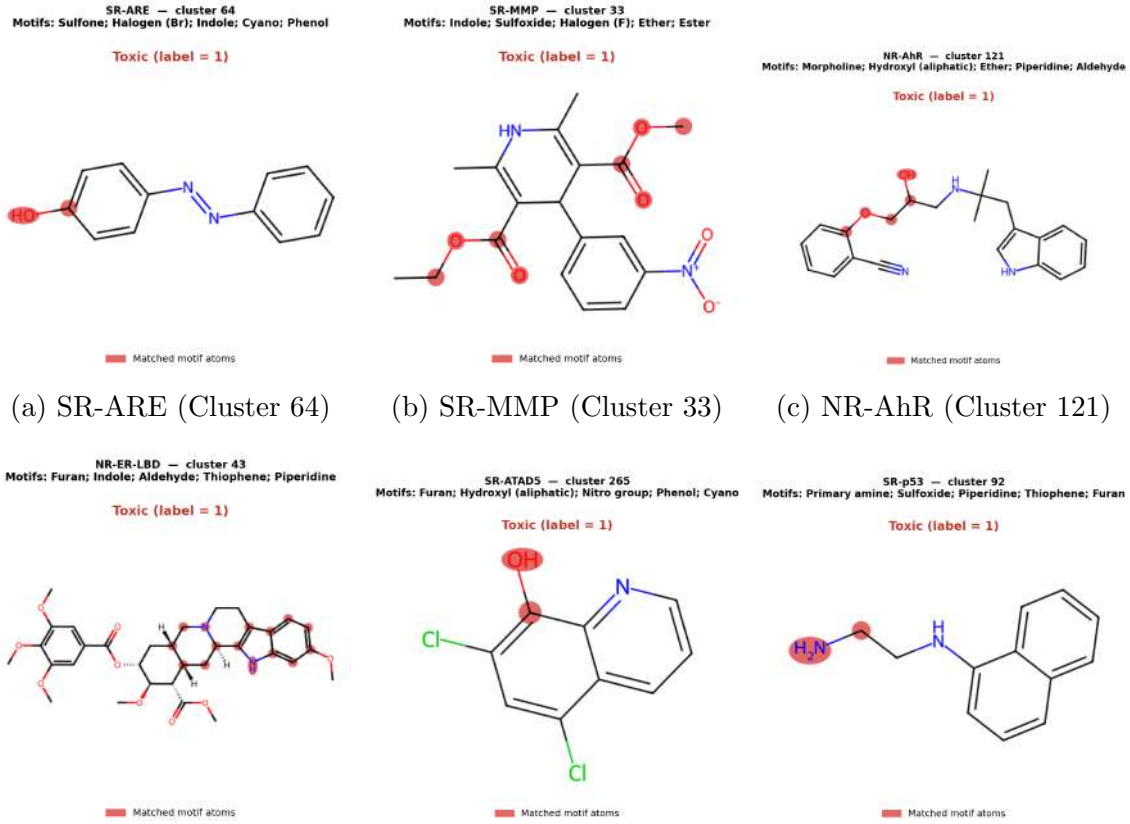
(a) NR-AR (Cluster 8) (b) NR-Aromatase (Cluster 121) (c) NR-PPAR-γ (Cluster 226)



(d) SR-HSE (Cluster 33) (e) NR-AR-LBD (Cluster 43) (f) NR-ER (Cluster 43)

Figure 16: Representative toxic molecules for six endpoints (part I). Atoms matching the top consensus cluster motifs are highlighted in red.

Two caveats apply to this analysis. First, cluster motifs are derived from co-activation patterns in Morgan fingerprint space rather than from direct atom-level SHAP attribution: a label such as "Furan" indicates that furan-containing molecules disproportionately activate the cluster, not that the furan atom itself received the highest score. Second, enrichment factors for small clusters are estimated from few molecules and carry substantial uncertainty; results should be interpreted as indicative structural hypotheses rather than definitive toxicophore assignments.



(a) SR-ARE (Cluster 64) (b) SR-MMP (Cluster 33) (c) NR-AhR (Cluster 121)

(d) NR-ER-LBD (Cluster 43) (e) SR-ATAD5 (Cluster 265) (f) SR-p53 (Cluster 92)

Figure 17: Representative toxic molecules for six endpoints (part II). Atoms matching the top consensus cluster motifs are highlighted in red.

4.6 Correspondence with known toxicophores

Several structural motifs identified by the SHAP consensus analysis correspond to functional groups with documented toxicological activity, providing external validation for the model's learned representations.

1. **Furan and thiophene.** Both furan and thiophene appear repeatedly as toxic driver motifs across endpoints, particularly in Cluster 43 (NR-AR-LBD, NR-ER, NR-ER-LBD) and Cluster 92 (SR-p53). These heterocycles are canonical structural alerts: furan is bioactivated by epoxidation to a reactive cis-enedione, and thiophene undergoes sulfur oxidation to form reactive sulfoxide and epoxide intermediates, both of which are known to form covalent protein adducts [25].

2. **Nitro group and phenol.** The nitro group appears as a discriminating motif in Cluster 265 (SR-ATAD5) and phenol in Cluster 64 (SR-ARE). Nitro-containing compounds are widely recognised as mutagenic toxicophores through bioreductive activation to

nitroso and hydroxylamine intermediates [25], [26]. Nitrophenols in particular are associated with strong cytotoxicity via quinone-forming oxidative pathways [27].

3. **Sulfur-containing motifs.** Sulfone, sulfoxide, thioether, and thiol appear across multiple toxic driver clusters, including Cluster 8 (NR-AR), Cluster 226 (NR-PPAR-gamma), and Cluster 33 (SR-HSE, SR-MMP). Sulfur-containing structural alerts are repeatedly identified as endocrine-disruption alerts and bioactivation sites, with thiophene, sulfamate, and thioether derivatives among the most commonly flagged motifs in toxicophore databases [25], [28].

4. **Halogens and substituted aromatics.** Halogen (F) appears in Cluster 33 (SR-HSE, SR-MMP) and halogen (Br) in Cluster 64 (SR-ARE). Halogenated aromatic fragments are consistently identified as toxic substructures across multiple endpoints in recent XAI-based analyses of Tox21 [27], [29], [30], with chloro- and nitro-substituted phenols among the most cytotoxic representatives.

5. **Electrophilic carbonyls.** Aldehyde appears as a discriminating motif in Cluster 43 and Cluster 121, both of which are top toxic drivers for nuclear receptor endpoints. Electrophilic carbonyl motifs, including aldehydes and Michael acceptor-like fragments, are broadly implicated in protein crosslinking and are identified by XAI-based toxicophore extraction methods as co-occurring with mutagenic substructures [12], [29].

Cluster motifs are derived from co-activation patterns in Morgan fingerprint space rather than atom-level attribution, so a motif label indicates that molecules containing that functional group disproportionately activate the cluster, not that the group itself is the causal toxicophore. These correspondences are hypothesis-generating observations consistent with the established literature, not mechanistic conclusions.

Chapter 5 Conclusions and Future work

This thesis set out to answer the question of which molecular substructures drive toxicity predictions in a DNN trained on the Tox21 dataset. By combining cluster-level SHAP attributions with ground-truth enrichment validation, a set of consensus substructures was identified for each of the twelve endpoints, several of which correspond to functional groups already documented as toxicophores in the medicinal chemistry literature — including furan and thiophene heterocycles, nitro groups, phenols, and sulfur-containing motifs. The recurrence of specific clusters across mechanistically related endpoints, such as Cluster 43 across three nuclear receptor assays and Cluster 33/Cluster 29 as a toxic/protective pair across both stress-response membrane assays, suggests that the model has learned structural patterns that are at least partially aligned with known chemical reactivity and biological pathways, rather than relying on spurious correlations alone. The two-source classification combining SHAP directionality with enrichment factors proved to be a useful filter, producing a manageable and chemically annotated shortlist of candidate toxicophores per endpoint rather than an unranked list of 300 clusters.

5.1 Conclusions

The results of this work lead to several interconnected conclusions. To begin with, model complexity did not translate into proportionally better predictive performance: the classical baseline models performed comparably to the single-task deep neural network across most endpoints, suggesting that for tabular Morgan fingerprint data the added complexity of a neural network is not always justified. More importantly, clustering Morgan fingerprint bits into 300 chemically motivated groups substantially improved the stability of all three explainability methods across random seeds, with SHAP at the cluster level achieving the highest consistency (Jaccard top-50 between 0.73 and 0.81, Spearman > 0.93 for all endpoints).

The analysis also revealed important differences between the explainability methods. Permutation importance alone proved to be an unreliable toxicophore filter: high importance often reflected a cluster's role in defining the non-toxic majority class rather than a genuine toxicity signal, with top-enriched clusters per endpoint ranging from enrichment factors of 0.91 (NR-AR-LBD, essentially no enrichment) to 6.02 (NR-ER-LBD) and no consistent relationship between importance rank and ground-truth enrichment. In contrast, combining SHAP directionality with empirical enrichment factors yielded a two-source filter that identified between 4 and 13 toxic driver clusters and between 2 and 14 protective clusters per endpoint, depending on how structurally specific each endpoint's toxicity signal is. Permutation and SHAP analyses were found to be complementary rather than redundant: Cluster 234, identified via permutation consensus, achieved the highest enrichment factors observed across either analysis (3.85 for NR-ER, 6.02 for NR-ER-LBD), yet did not appear among

the SHAP-based toxic drivers, illustrating that the two methods can independently surface different mechanistically relevant substructures.

Several structural families recurred as top toxic drivers across multiple endpoints, pointing to shared mechanisms. Cluster 43 (Furan; Indole; Aldehyde; Thiophene; Piperidine) was the highest-enrichment toxic driver for three nuclear receptor endpoints (NR-AR-LBD, NR-ER, NR-ER-LBD), with enrichment factors between 2.62 and 4.58, while Cluster 33 (Indole; Sulfoxide; Halogen(F); Ether; Ester) drove both stress-response endpoints SR-HSE and SR-MMP. Many of the recovered motifs correspond to documented toxicophores in the literature: furan and thiophene are canonical electrophilic alerts capable of forming covalent protein adducts; nitro groups are well-established mutagenic toxicophores; sulfur-containing motifs are associated with endocrine disruption; and aldehydes and halogenated aromatics are recognised electrophilic carbonyls and persistent environmental hazards, respectively. This correspondence with external knowledge provides literature-based validation of the model's learned representations. Nevertheless, these results should be interpreted as hypothesis-generating rather than definitive: cluster motifs reflect co-activation patterns in fingerprint space rather than atom-level attributions, and enrichment factors for smaller clusters carry substantial uncertainty. The structural correspondences identified here should be regarded as candidate toxicophores warranting further experimental investigation, not as confirmed mechanistic findings.

5.2 Limitations

Several limitations must be acknowledged. First, the predictive performance of the underlying single-task models varies considerably across endpoints and is not strong enough in several cases to fully trust the model's internal notion of toxicity: explanations derived from a poorly calibrated model reflect its errors as much as genuine structure–activity relationships. Second, the severe class imbalance characteristic of the Tox21 dataset means that for many endpoints the number of confirmed positive examples in the test set is very small, so the enrichment factors reported here are estimated from limited evidence and should be interpreted with caution. Third, the molecular examples and enrichment statistics presented for each consensus cluster depend on the specific set of compounds that activate that cluster in the test set, which in several cases amounts to only a handful of molecules. Conclusions drawn from such small, cluster-specific subsets cannot be generalised to the broader chemical space and should be read as localised observations about the test set rather than universal structure–toxicity rules.

A further limitation concerns initialisation sensitivity. Although SHAP importance rankings are reasonably stable at the cluster level, they still vary across seeds, and the consensus filter discards a non-trivial number of clusters that are important for only some seeds. Many of the identified clusters are moreover neither clearly enriched nor depleted with respect to toxic activity, suggesting weak or ambiguous association with the target endpoints and indicating

that the models are still partly capturing noise in addition to genuine signal. Finally, an attempt was made to evaluate graph neural networks under the hypothesis that they would outperform fingerprint-based models by directly exploiting molecular connectivity. This line of work could not be rigorously pursued due to library compatibility issues that prevented reproducibility of previously reported GNN results. A closer examination of the literature also revealed that many published GNN studies reported performance primarily using aggregate metrics such as AUROC, omitting class-balanced metrics more appropriate for highly imbalanced toxicity datasets, leaving the practical advantage of GNN-based approaches over the fingerprint-based pipeline used here unclear.

5.3 Future work

Several directions could extend and strengthen the findings presented in this thesis. On the representation side, SMILES-based sequence models, graph neural networks (once library compatibility issues are resolved), or learned embeddings from variational autoencoders could replace or complement Morgan fingerprints, potentially producing substructure representations that align more naturally with chemically meaningful units rather than hashed circular environments. Incorporating chemical domain knowledge as a prior—for instance by constraining or initialising the clustering procedure using known toxicophore libraries, reaction mechanism databases, or pharmacophore definitions—could further bridge the gap between statistical co-activation patterns and genuine mechanistic interpretation.

From a modelling perspective, replacing twelve independent binary classifiers with multi-task or multi-class architectures could allow information to be shared across endpoints, potentially improving performance on data-scarce endpoints and producing richer signals for the XAI pipeline. Alternatively, a simpler formulation predicting whether a compound is toxic for any endpoint would yield a more balanced dataset and likely higher overall accuracy, though at the cost of losing the endpoint-specific resolution that is central to the structural analysis presented here; future work should explicitly weigh this trade-off depending on the intended use case, screening versus mechanistic insight. Reducing initialisation sensitivity through larger training datasets or more robust architectures would also increase confidence in the resulting consensus clusters by decreasing seed-to-seed variability in the SHAP-based explanations.

Finally, generative and search-based methods could provide a form of *in silico* validation that goes beyond the observational analysis presented here. Variational autoencoders could be used to explore continuous latent chemical spaces and identify directions associated with toxicity, while genetic algorithms could be applied to optimise or perturb molecules along the substructure axes identified in this work, directly testing whether the candidate toxicophores recovered by the pipeline have the causal role in toxicity that the structural correspondences suggest.

Chapter 6 Code Reuse and Attribution

Part of the implementation developed in this project is based on code released by the authors of the original work under the Apache License 2.0 [2]. The original codebase served as a starting point for certain components of the experimental pipeline, particularly those related to data processing, model training, and evaluation procedures.

Significant modifications and extensions were introduced to adapt the framework to the objectives of this thesis. In particular, the original implementation was extended with a comprehensive explainability pipeline incorporating permutation importance, SHAP, and neural sensitivity analysis. Additional functionality was developed to cluster fingerprint features into chemically coherent groups, evaluate explanation stability across multiple random seeds, compute enrichment factors for important clusters, and compare the resulting candidate toxicophores against findings reported in the medicinal chemistry literature.

All reused code remains subject to the terms of the Apache License 2.0, and appropriate attribution is provided to the original authors. Any modifications, extensions, and analyses presented in this thesis are the responsibility of the author and do not necessarily reflect the views of the original contributors.

The code can be found in the repository `https://github.com/fernandamarcos/toxicity-filter.git`

Chapter 7 References

[1] M. V. Togo et al., "Tiresia: An explainable artificial intelligence platform for predicting developmental toxicity", *Journal of chemical information and modeling*, 2022. DOI: 10.1021/acs.jcim.2c01126.

[2] B. Sharma et al., "Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations", *Scientific Reports*, vol. 13, no. 1, p. 4908, Mar. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-31169-8. [Online]. Available: https://doi.org/10.1038/s41598-023-31169-8.

[3] J. Cremer, L. Medrano Sandonas, A. Tkatchenko, D.-A. Clevert, and G. De Fabritiis, "Equivariant graph neural networks for toxicity prediction", *Chemical Research in Toxicology*, vol. 36, no. 10, pp. 1561–1573, Oct. 2023, ISSN: 0893-228X. DOI: 10.1021/acs.chemrestox.3c00032. [Online]. Available: https://doi.org/10.1021/acs.chemrestox.3c00032.

[4] P. G. Polishchuk, T. I. Madzhidov, and A. Varnek, "Estimation of the size of drug-like chemical space based on GDB-17 data", *Journal of Computer-Aided Molecular Design*, vol. 27, no. 8, pp. 675–679, Aug. 2013. DOI: 10.1007/s10822-013-9672-4.

[5] H. Heberle, L. Zhao, S. Schmidt, T. Wolf, and J. Heinrich, "Xsmiles: Interactive visualization for molecules, smiles and xai attribution scores", *Journal of Cheminformatics*, vol. 15, no. 1, p. 2, Jan. 2023, ISSN: 1758-2946. DOI: 10.1186/s13321-022-00673-w. [Online]. Available: https://doi.org/10.1186/s13321-022-00673-w.

[6] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules", *Scientific Data*, vol. 1, no. 1, p. 140 022, Aug. 2014, ISSN: 2052-4463. DOI: 10.1038/sdata.2014.22. [Online]. Available: https://doi.org/10.1038/sdata.2014.22.

[7] A. T. Balaban, "Applications of graph theory in chemistry", *Journal of Chemical Information and Modeling*, vol. 25, pp. 334–343, 1985. DOI: 10.1021/ci00047a033.

[8] Y. Wang, Z. Li, and A. Barati Farimani, "Graph neural networks for molecules", in *Machine Learning in Molecular Sciences*, C. Qu and H. Liu, Eds. Cham: Springer International Publishing, 2023, pp. 21–66, ISBN: 978-3-031-37196-7. DOI: 10.1007/978-3-031-37196-7_2. [Online]. Available: https://doi.org/10.1007/978-3-031-37196-7_2.

[9] D. Rogers and M. Hahn, "Extended-connectivity fingerprints", *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010, PMID: 20426451. DOI: 10.1021/ci100050t. eprint: https://doi.org/10.1021/ci100050t. [Online]. Available: https://doi.org/10.1021/ci100050t.

[10] T.-H. Nguyen-Vo, P. Teesdale-Spittle, J. Harvey, and B. Nguyen, "Molecular representations in bio-cheminformatics", *Memetic Computing*, vol. 16, pp. 519–536, Jul. 2024. DOI: 10.1007/s12293-024-00414-6.

[11] C. König and A. Vellido, "Understanding predictions of drug profiles using explainable machine learning models", *BioData Mining*, vol. 17, no. 1, p. 25, Aug. 2024, ISSN: 1756-

0381. DOI: 10.1186/s13040-024-00378-w. [Online]. Available: https://doi.org/10.1186/s13040-024-00378-w.

[12] M. Walter, S. J. Webb, and V. J. Gillet, "Interpreting neural network models for toxicity prediction by extracting learned chemical features", *Journal of Chemical Information and Modeling*, vol. 64, pp. 3670–3688, 2024. DOI: 10.1021/acs.jcim.4c00127.

[13] A. Setiya, V. Jani, U. Sonavane, and R. Joshi, "Moltoxpred: Small molecule toxicity prediction using machine learning approach", *RSC Advances*, vol. 14, pp. 4201–4220, 2024. DOI: 10.1039/d3ra07322j.

[14] P. B. R. Hartog, F. Krüger, S. Genheden, and I. V. Tetko, "Using test-time augmentation to investigate explainable ai: Inconsistencies between method, model and human intuition", *Journal of Cheminformatics*, vol. 16, 2024. DOI: 10.1186/s13321-024-00824-1.

[15] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[16] S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, 2017. arXiv: 1705.07874 [cs.AI]. [Online]. Available: https://arxiv.org/abs/1705.07874.

[17] J. Pizarroso, J. Portela, and A. Muñoz, "Neuralsens: Sensitivity analysis of neural networks", *Journal of Statistical Software*, vol. 102, no. 7, pp. 1–36, 2022. DOI: 10.18637/jss.v102.i07. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v102i07.

[18] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Grouped variable importance with random forests and application to multiple functional data analysis", *Computational Statistics Data Analysis*, vol. 90, pp. 15–35, 2015, ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2015.04.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947315000997.

[19] A. M. Richard et al., "The tox21 10k compound library: Collaborative chemistry advancing toxicology", *Chemical Research in Toxicology*, vol. 34, no. 2, pp. 189–216, Feb. 2021, ISSN: 0893-228X. DOI: 10.1021/acs.chemrestox.0c00264. [Online]. Available: https://doi.org/10.1021/acs.chemrestox.0c00264.

[20] D. R. Cox, "The regression analysis of binary sequences", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958. DOI: 10.1111/j.2517-6161.1958.tb00292.x.

[21] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. DOI: 10.1214/aos/1013203451.

[22] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree", in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[23] P. Jaccard, "The distribution of the flora in the alpine zone", *The New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912, ISSN: 0028646X, 14698137. Accessed: Jun. 15, 2026. [Online]. Available: http://www.jstor.org/stable/2427226.

[24] C. Spearman, "The proof and measurement of association between two things", *International Journal of Epidemiology*, vol. 39, no. 5, pp. 1137–1150, Oct. 2010, ISSN: 0300-5771. DOI: 10.1093/ije/dyq191. eprint: https://academic.oup.com/ije/article-pdf/39/5/1137/18481215/dyq191.pdf. [Online]. Available: https://doi.org/10.1093/ije/dyq191.

[25] N. Dang, T. B. Hughes, G. Miller, and S. J. Swamidass, "Computational approach to structural alerts: Furans, phenols, nitroaromatics, and thiophenes", *Chemical research in toxicology*, vol. 30, pp. 1046–1059, 2017. DOI: 10.1021/acs.chemrestox.6b00336.

[26] K. Nepali, H.-Y. Lee, and J. Liou, "Nitro-group-containing drugs.", *Journal of medicinal chemistry*, vol. 62 6, pp. 2851–2893, 2018. DOI: 10.1021/acs.jmedchem.8b00147.

[27] C. Chen, Y. Du, Y. Zhou, Q. Wu, S. Zheng, and J. Fang, "Formation of nitro(so) and chlorinated products and toxicity alteration during the uv/monochloramine treatment of phenol.", *Water research*, vol. 194, p. 116 914, 2021. DOI: 10.1016/j.watres.2021.116914.

[28] L. C. S. Rosa, M. Sarhan, and A. S. Pimentel, "Toxic alerts of endocrine disruption revealed by explainable artificial intelligence", *Environment Health*, vol. 3, pp. 321–333, 2025. DOI: 10.1021/envhealth.4c00218.

[29] Y. Zhou, Y. He, W. Zhou, Z. Hua, Y.-J. Wang, and C. Chen, "Enhancing toxicity prediction of synthetic chemicals via novel smiles fragmentation and interpretable deep learning", *Journal of chemical information and modeling*, 2025. DOI: 10.1021/acs.jcim.5c01042.

[30] S. Li, M. Zhang, and P. Sun, "Prediction of acute toxicity of organic contaminants to fish: Model development and a novel approach to identify reactive substructures.", *Journal of hazardous materials*, vol. 491, p. 137 917, 2025. DOI: 10.1016/j.jhazmat.2025.137917.

Chapter A EDA results

A.1 Class balance

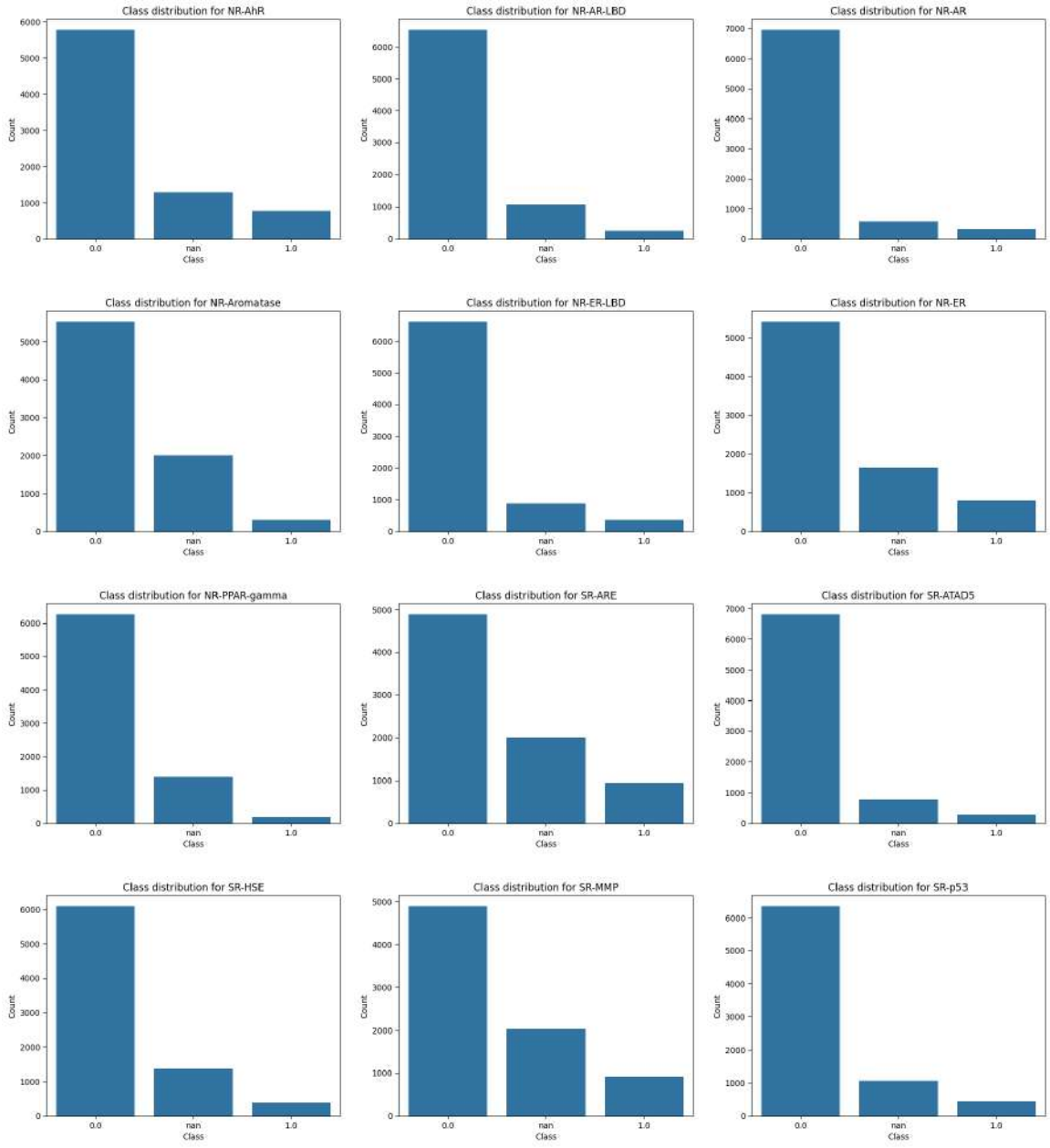


Figure 18: Class balance distribution across all Tox21 endpoints.

A.2 t-SNE visualisation

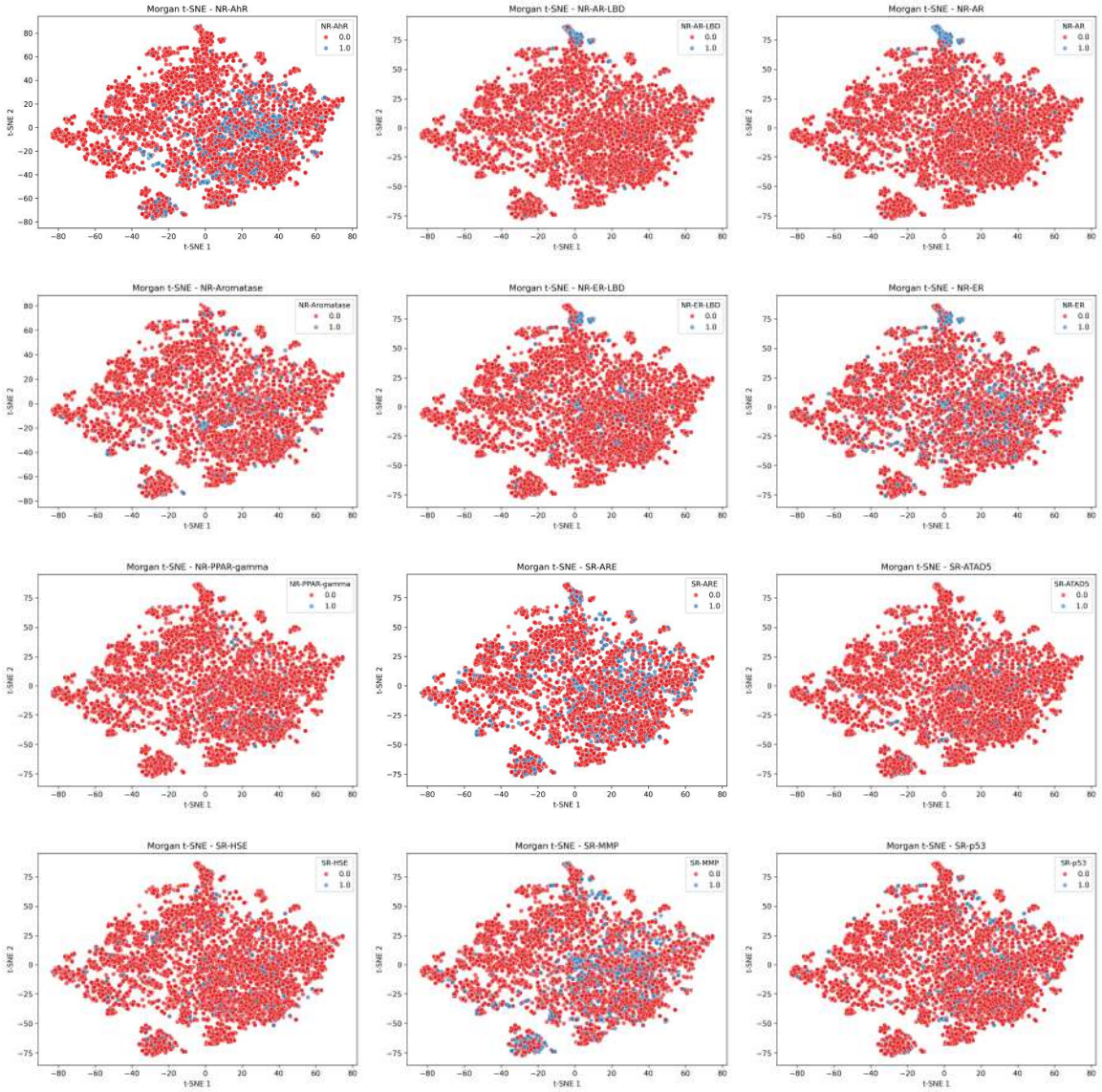


Figure 19: t-SNE visualisation across all endpoints.

A.3 LDA visualisation

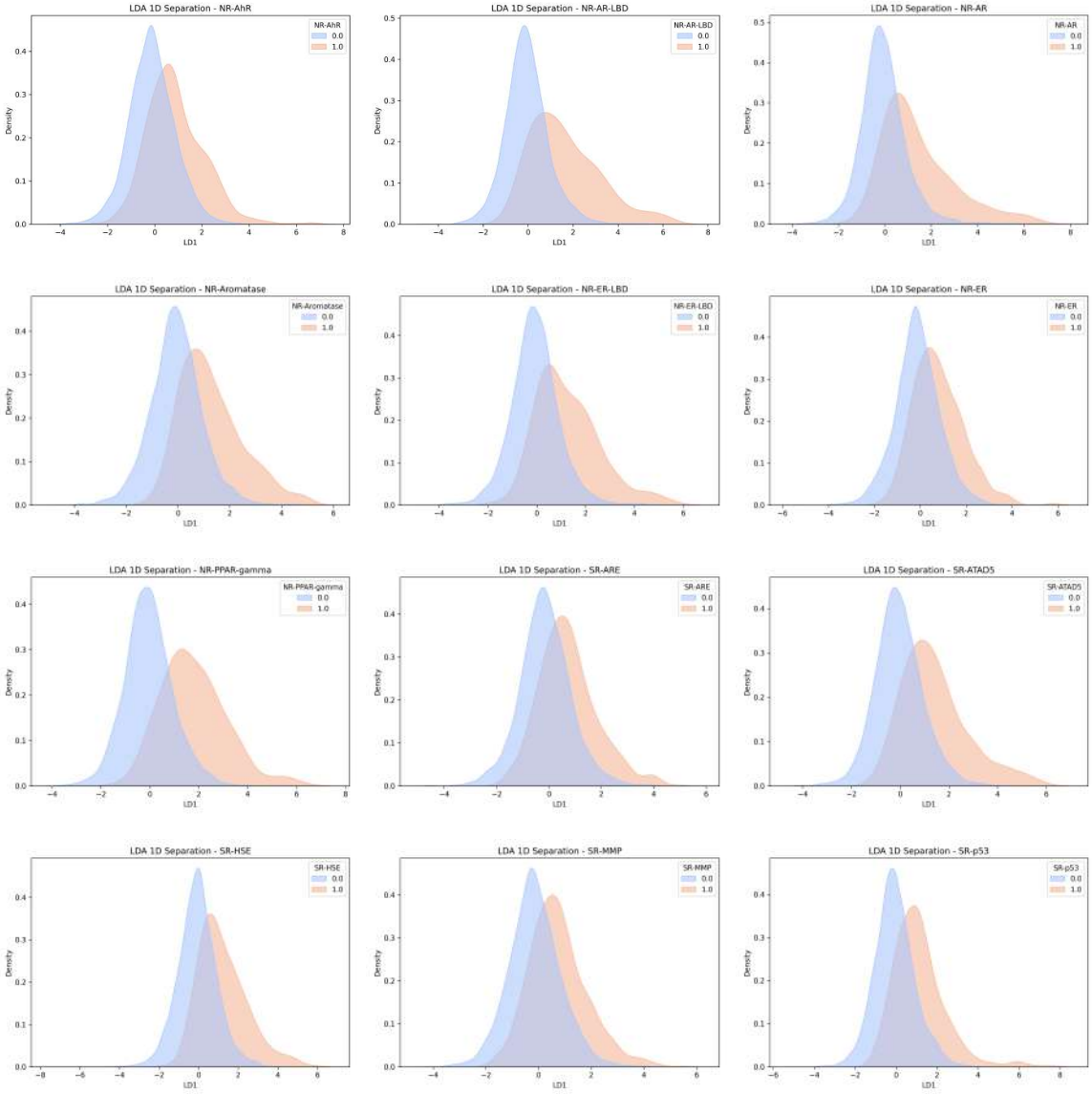


Figure 20: LDA visualisation across all endpoints.

Chapter B Baseline models

B.1 Confusion matrices

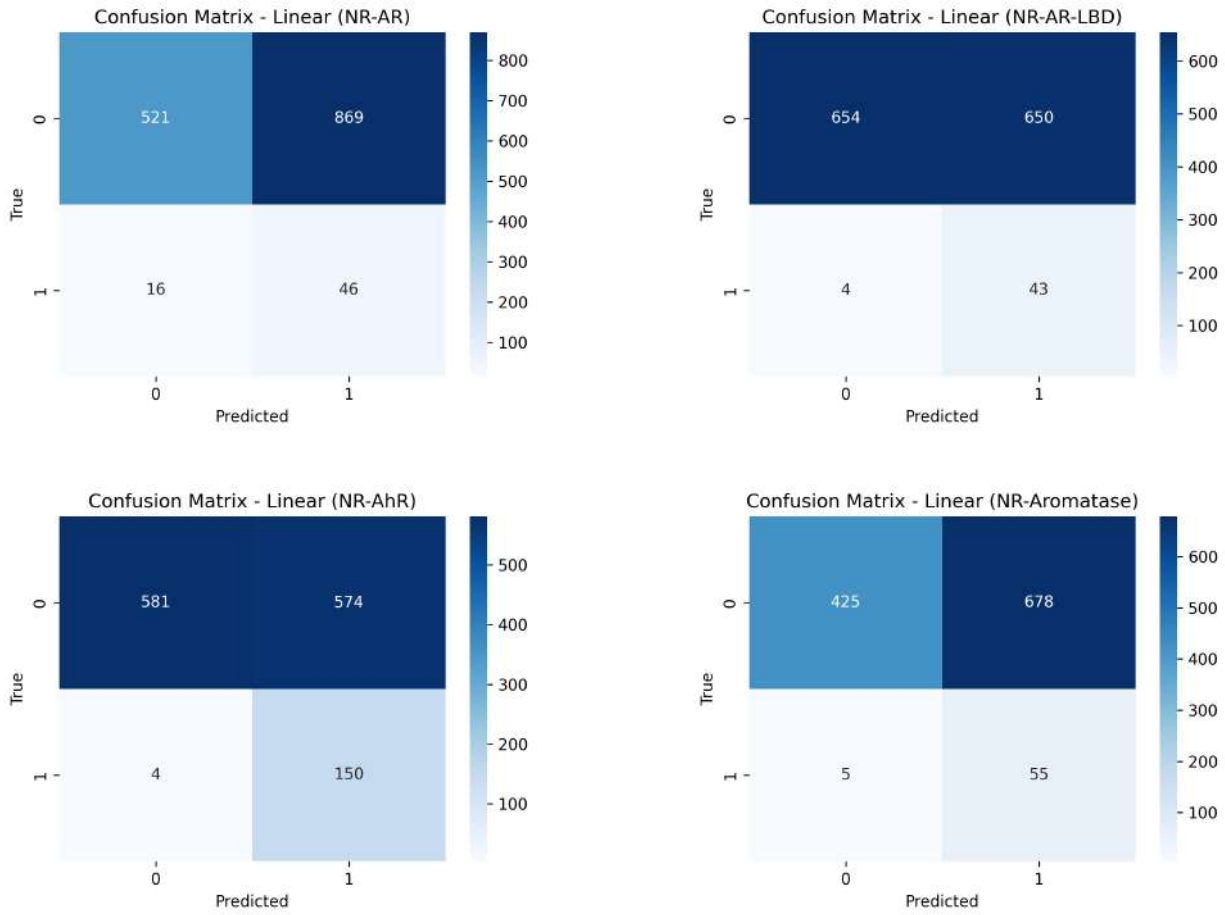


Figure 21: Linear model confusion matrices (part I).

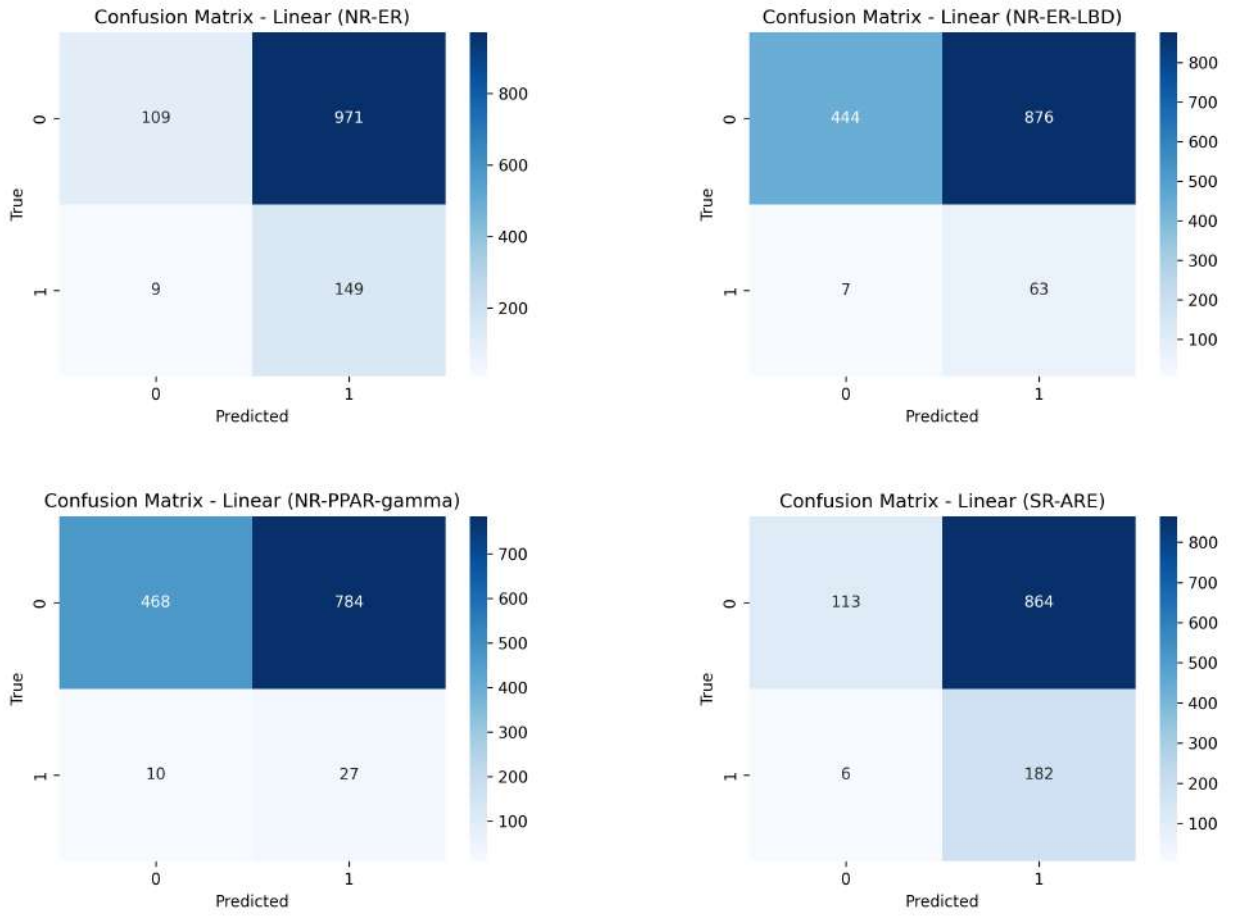


Figure 22: Linear model confusion matrices (part II).

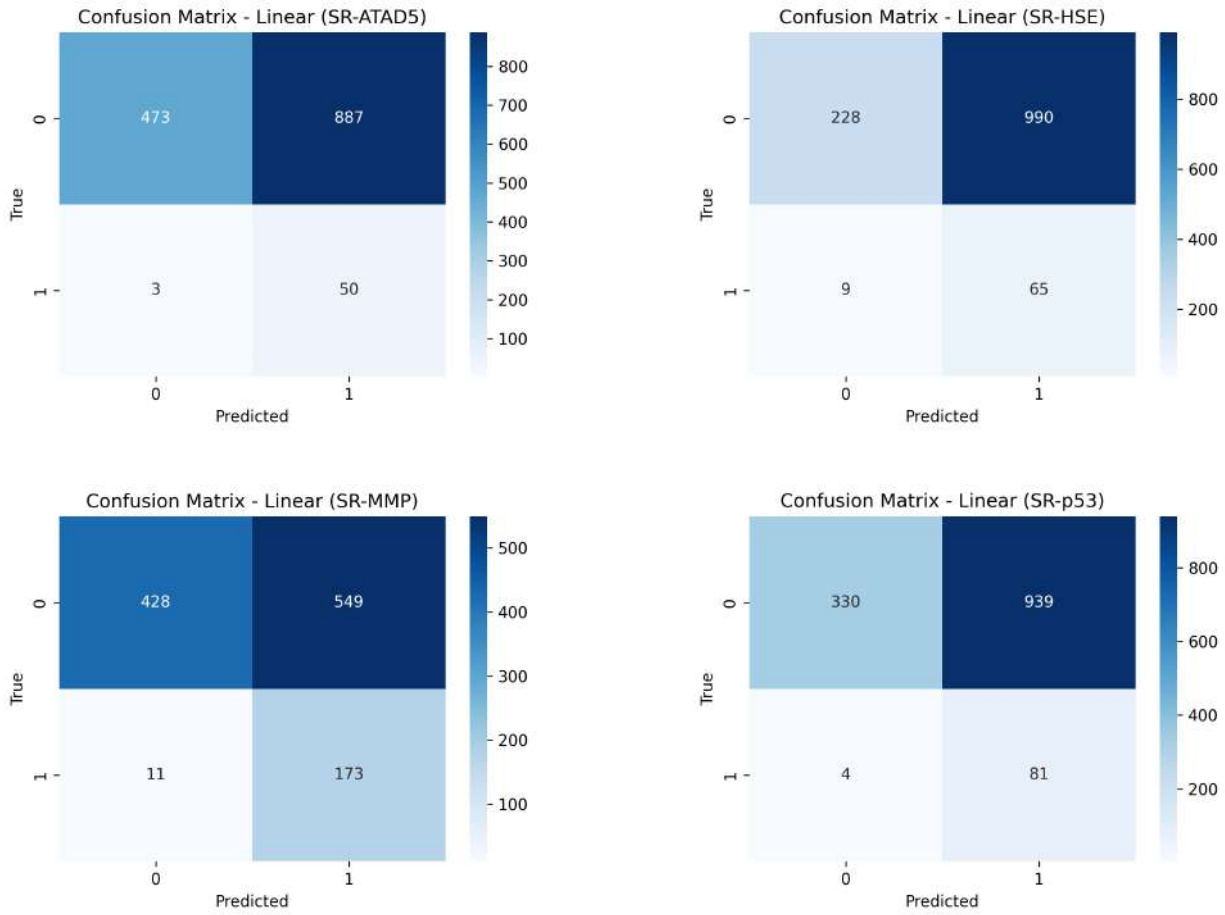


Figure 23: Linear model confusion matrices (part III).

B.1.1 Random Forest

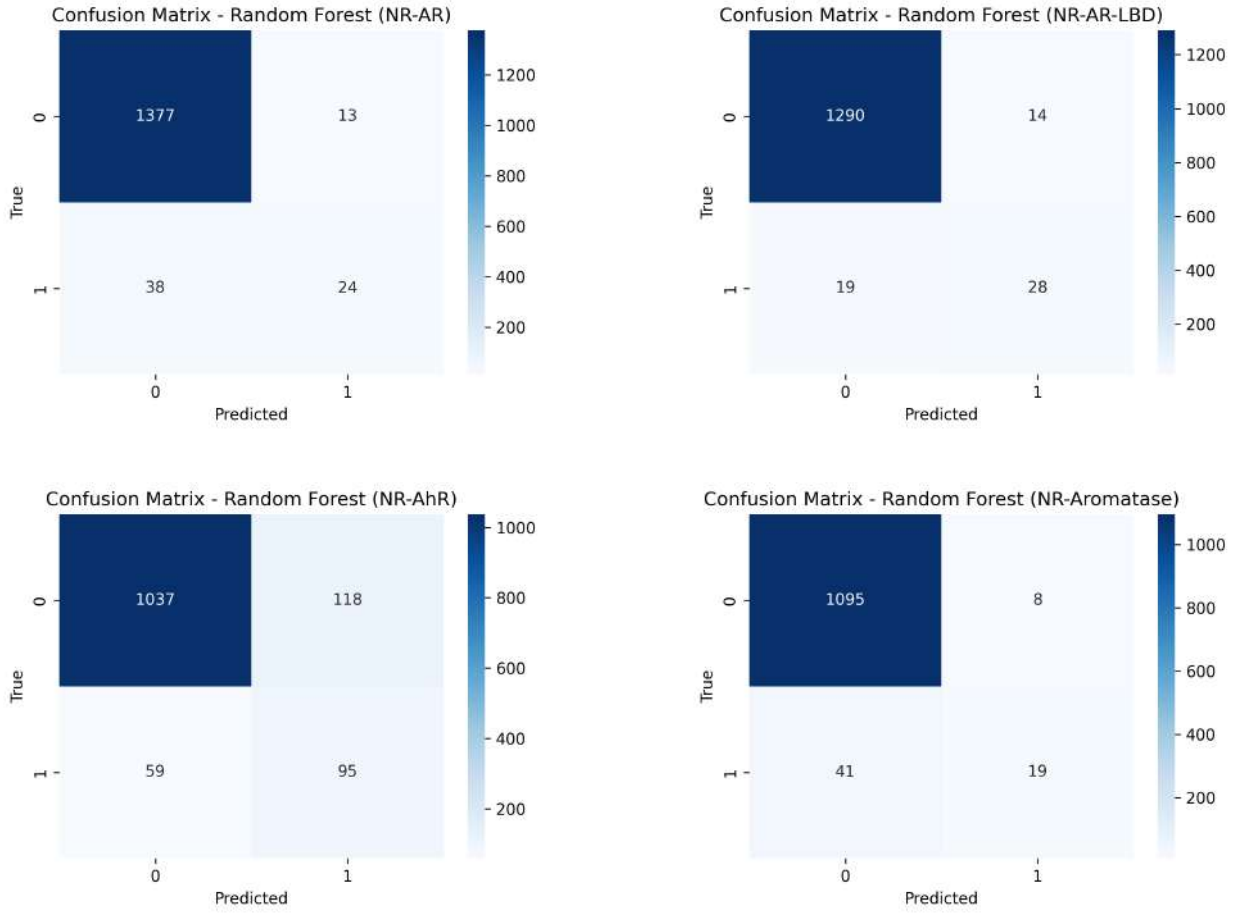


Figure 24: Random Forest confusion matrices (part I).

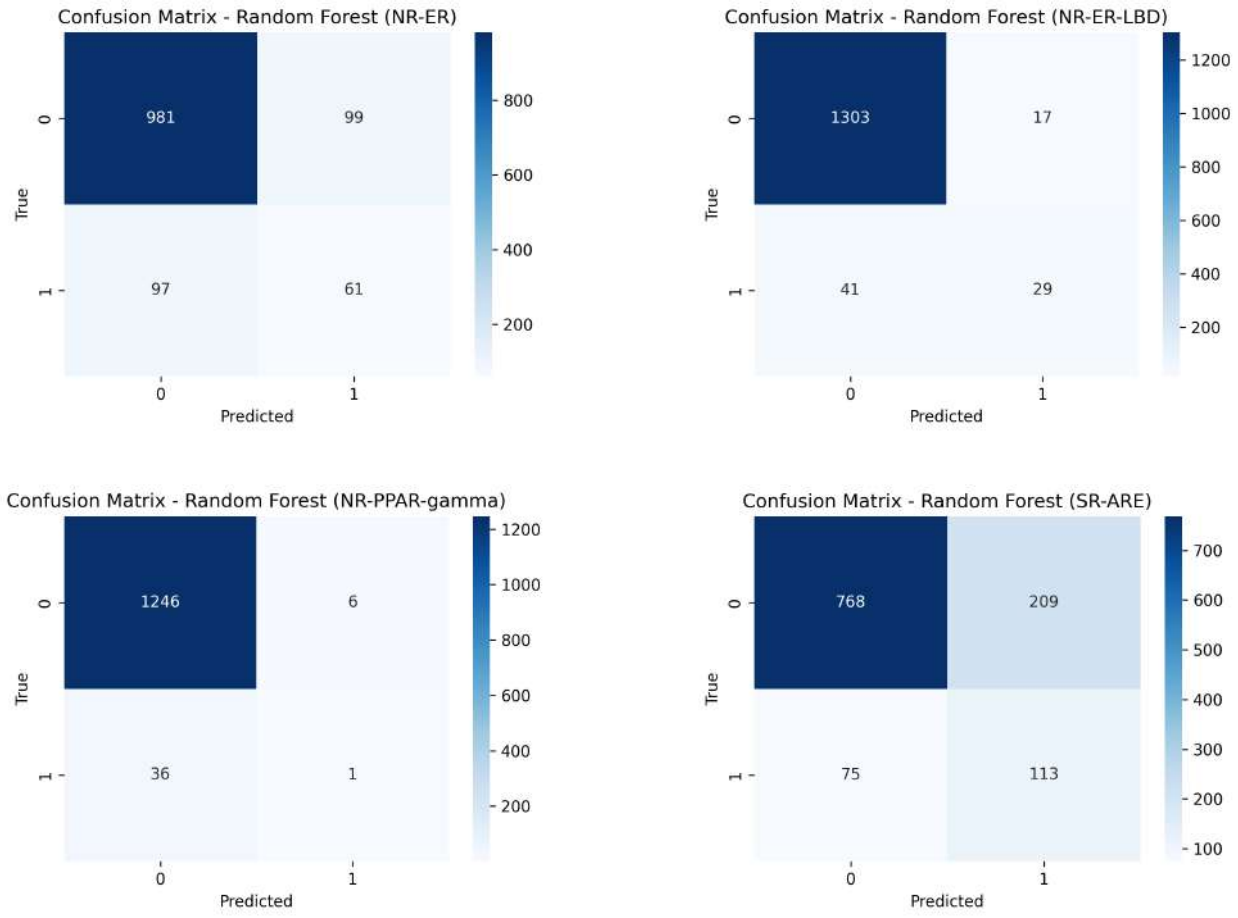


Figure 25: Random Forest confusion matrices (part II).

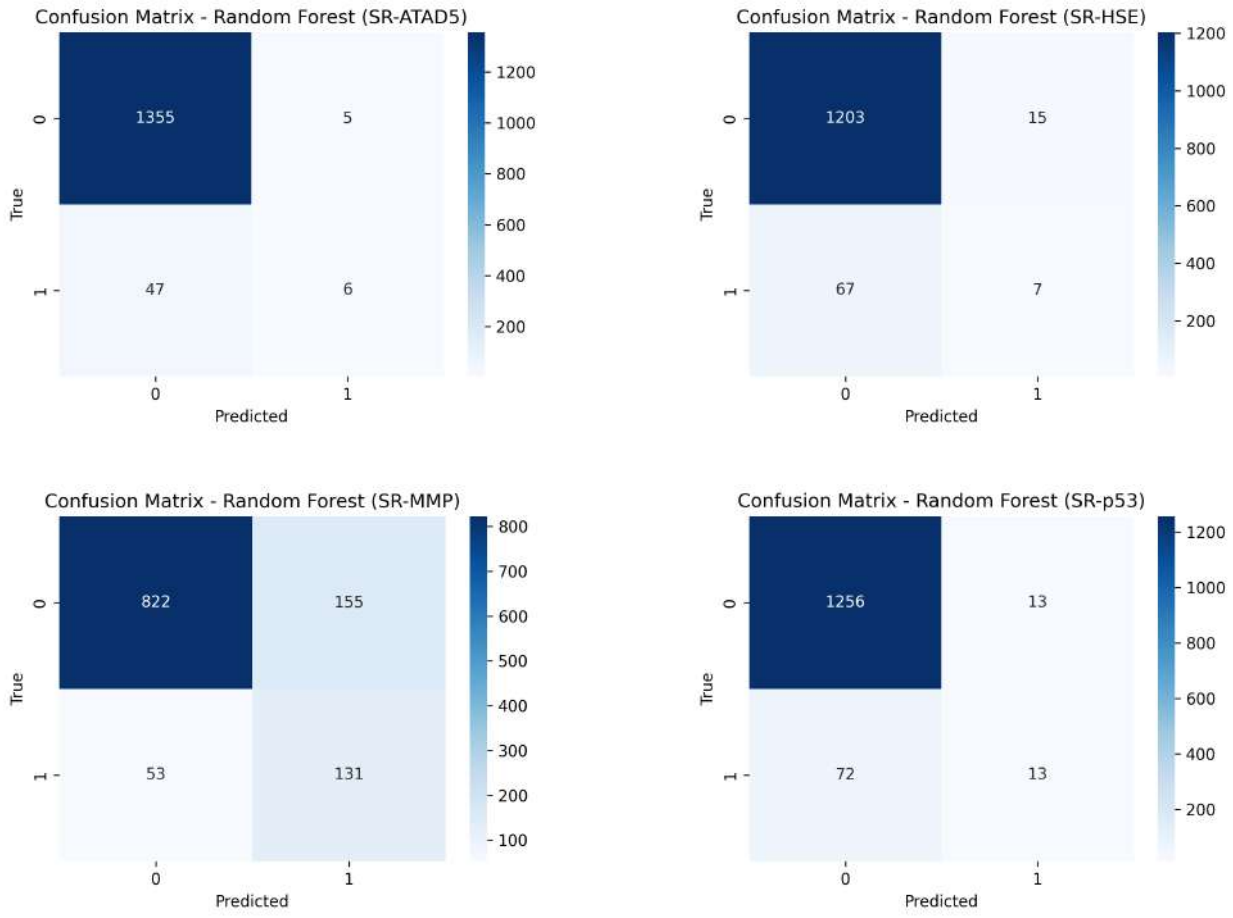


Figure 26: Random Forest confusion matrices (part III).

B.1.2 Boosting

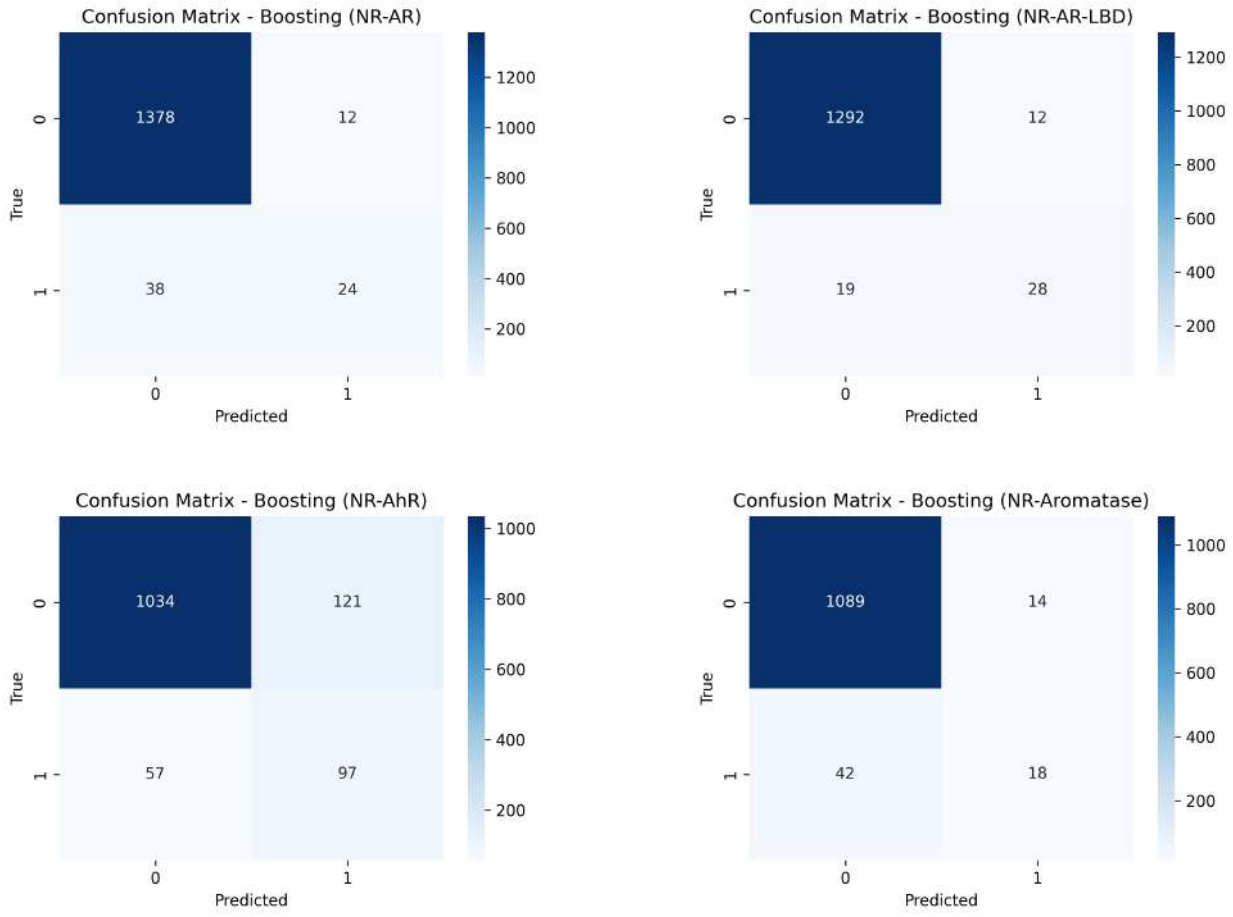


Figure 27: Boosting confusion matrices (part I).

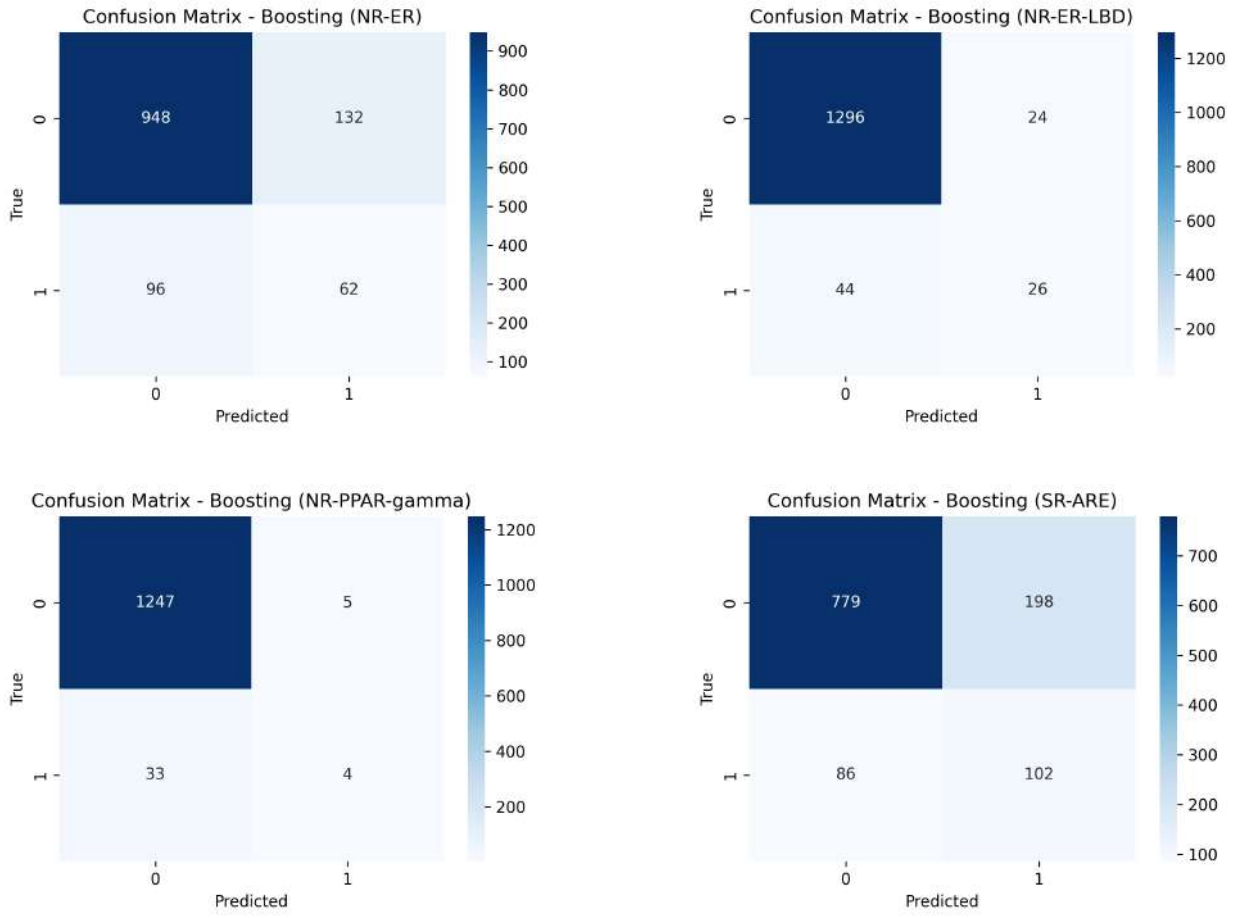


Figure 28: Boosting confusion matrices (part II).

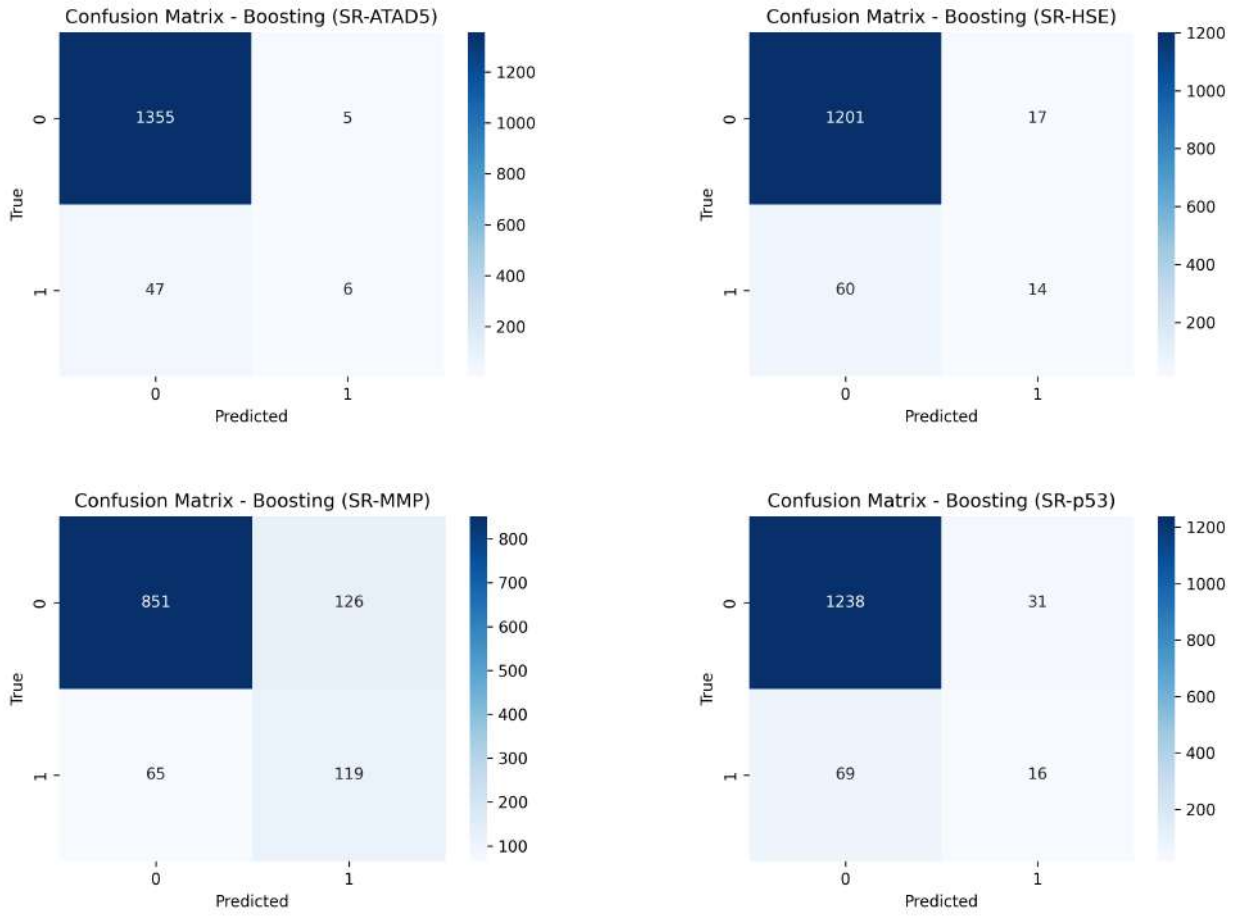


Figure 29: Boosting confusion matrices (part III).

B.2 Probability distributions

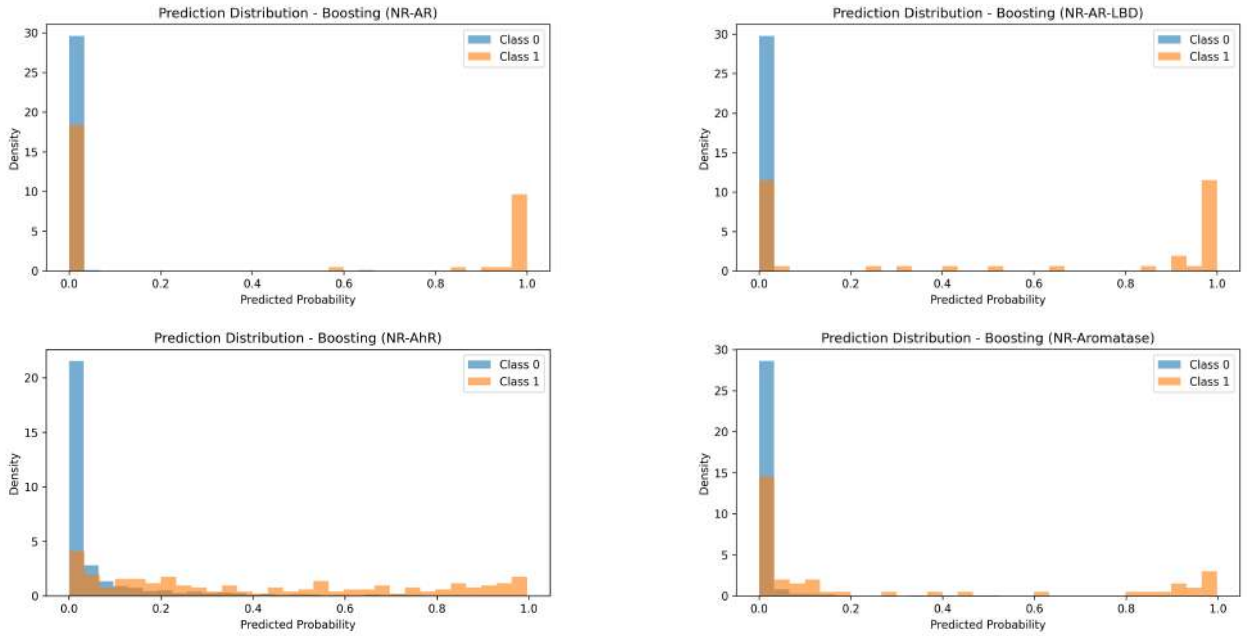


Figure 30: Boosting model prediction histograms (part I).

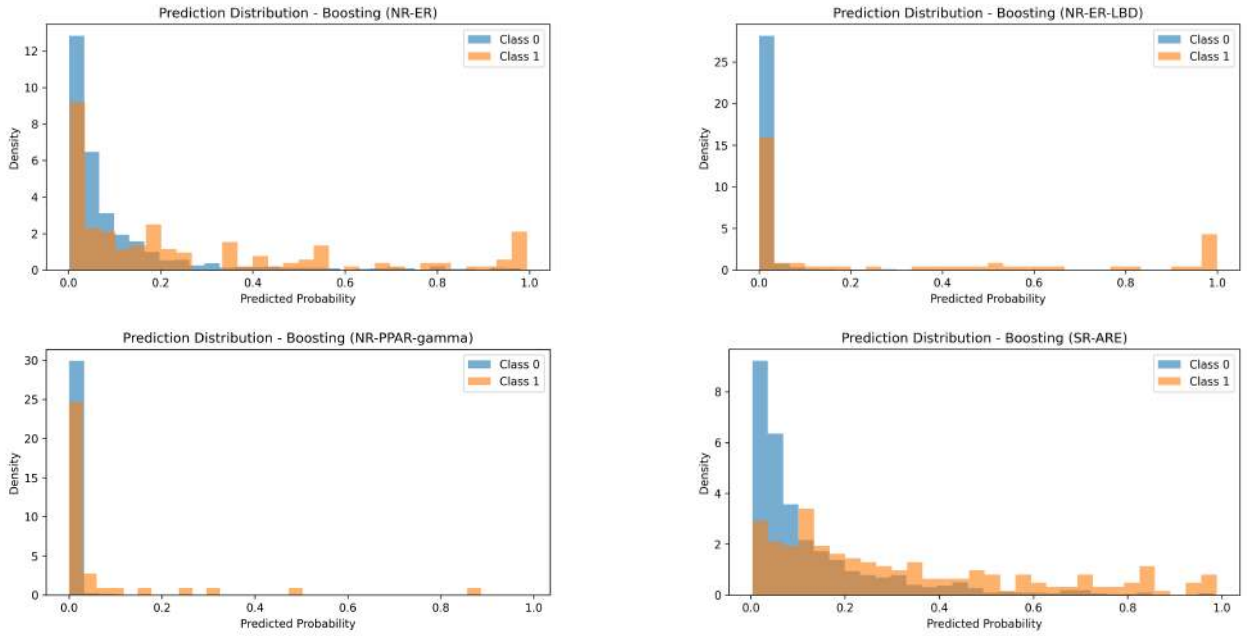


Figure 31: Boosting model prediction histograms (part II).

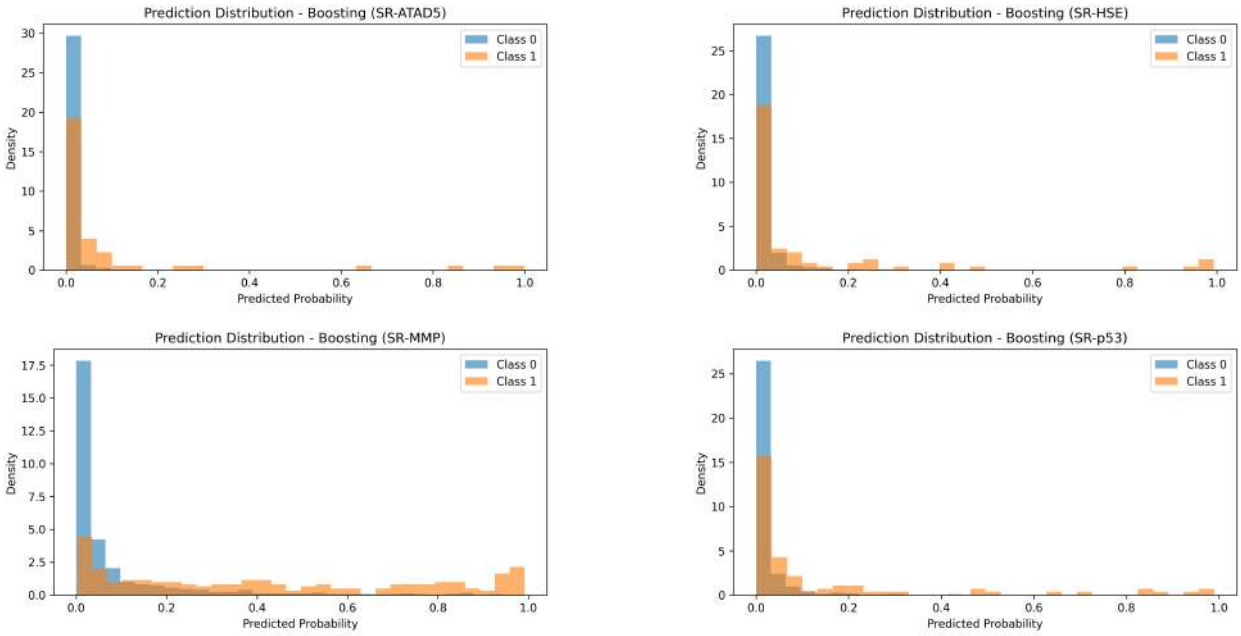


Figure 32: Boosting model prediction histograms (part III).

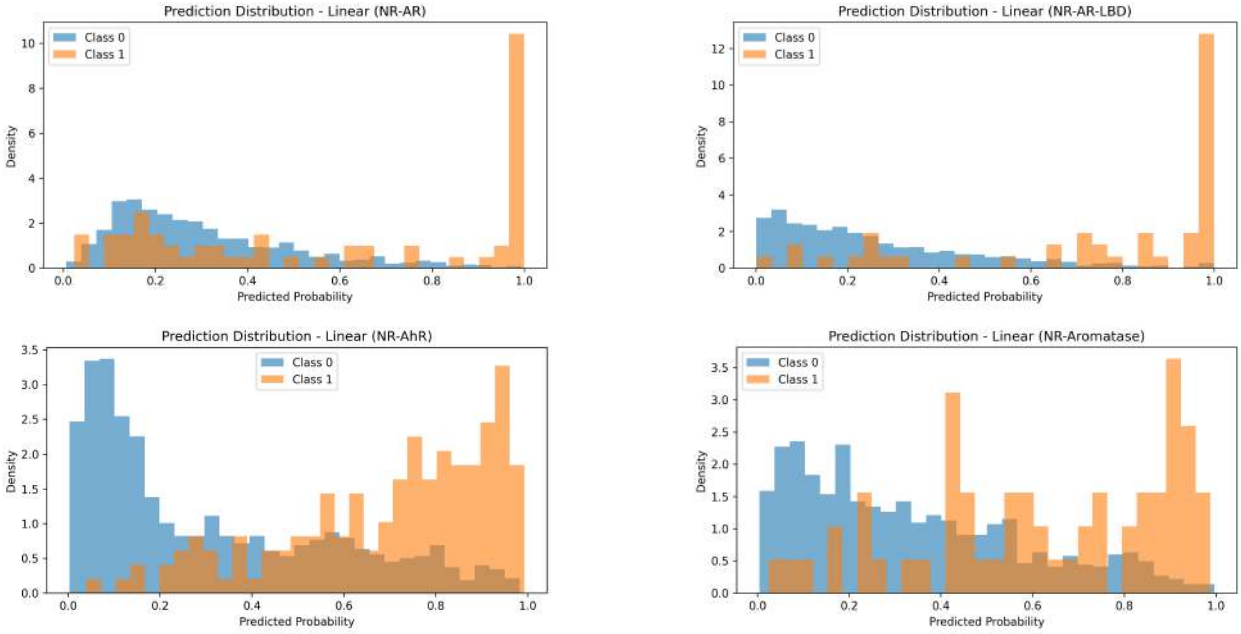


Figure 33: Linear model prediction histograms (part I).

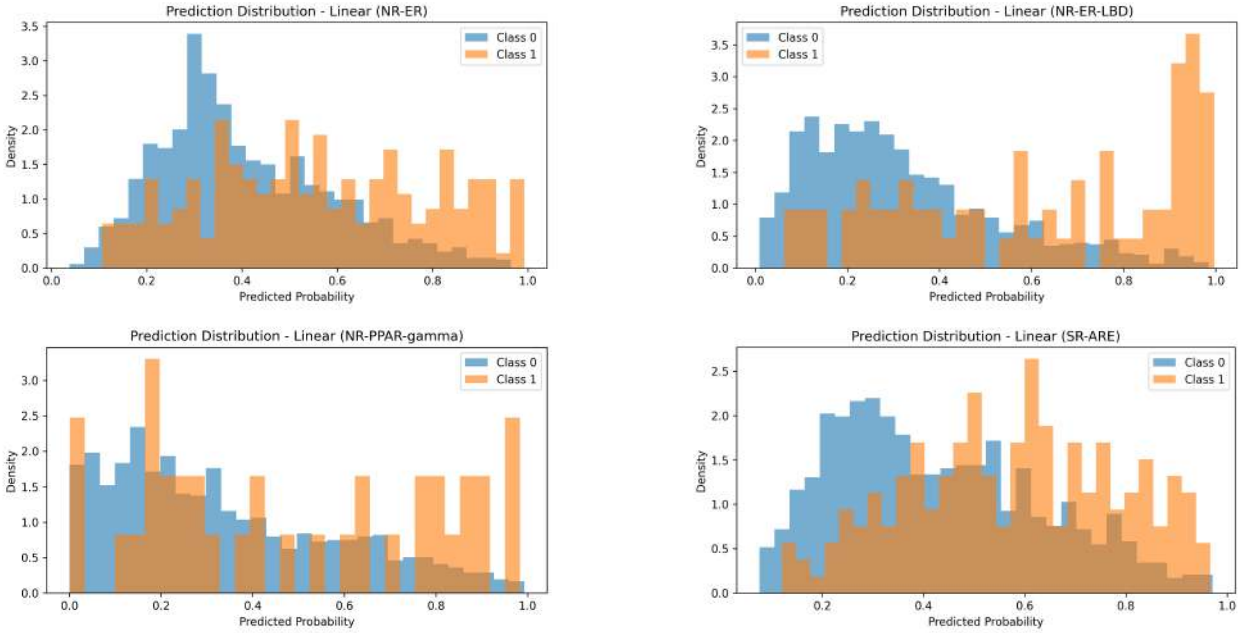


Figure 34: Linear model prediction histograms (part II).

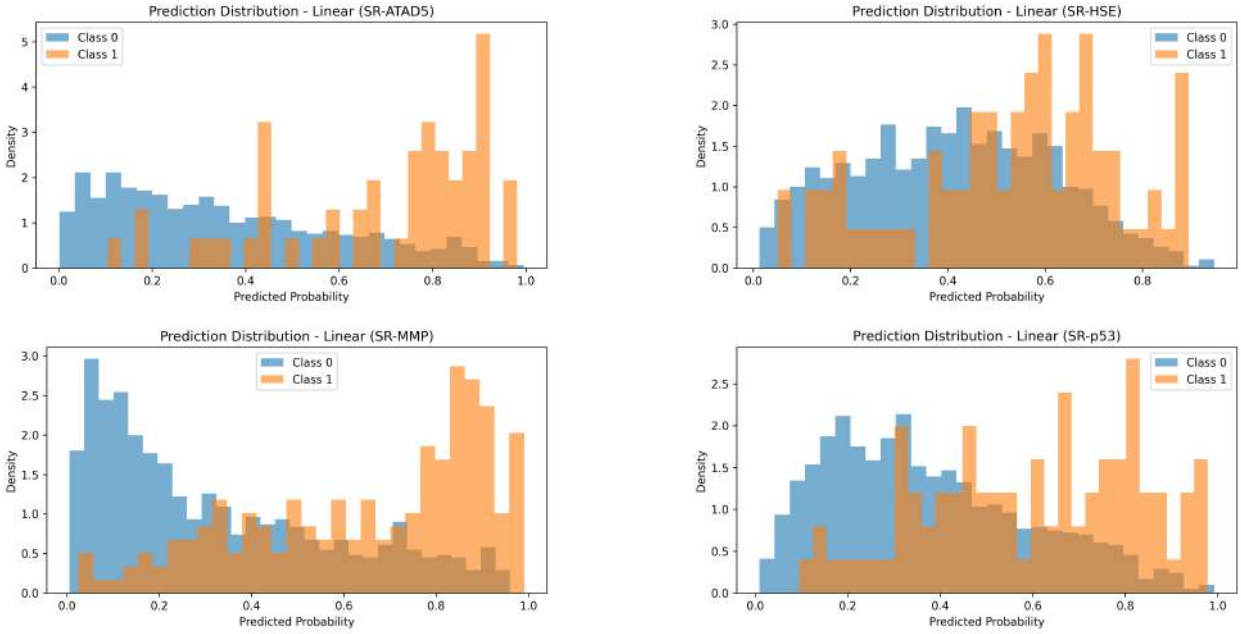


Figure 35: Linear model prediction histograms (part III).

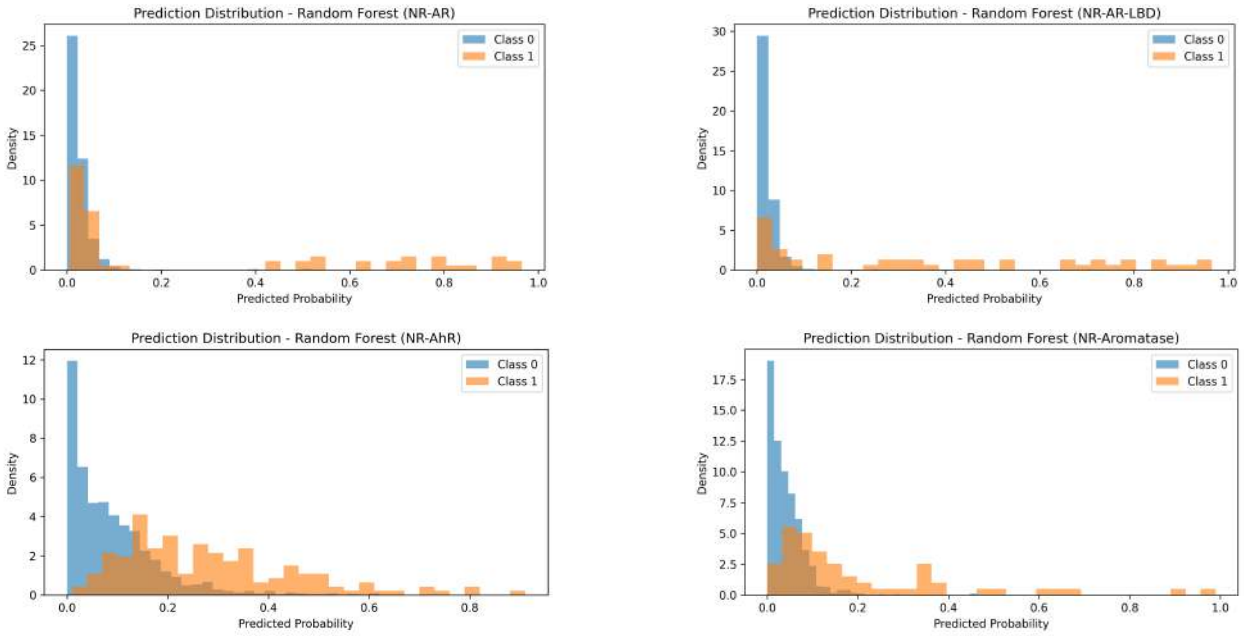


Figure 36: Random Forest model prediction histograms (part I).

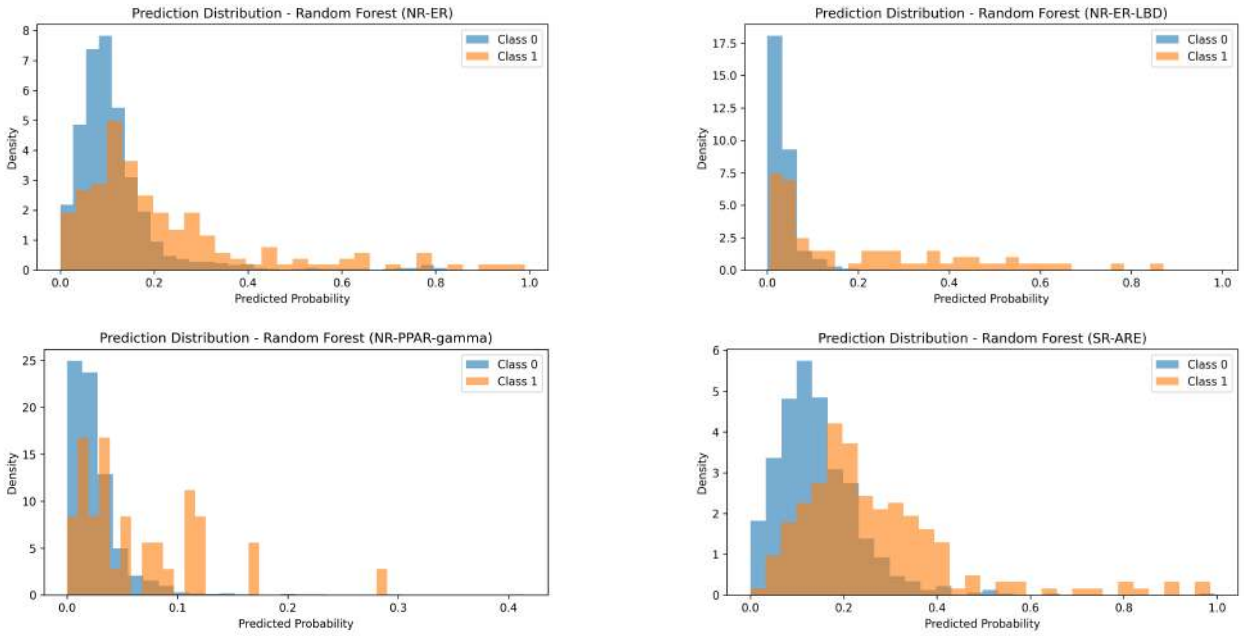


Figure 37: Random Forest model prediction histograms (part II).

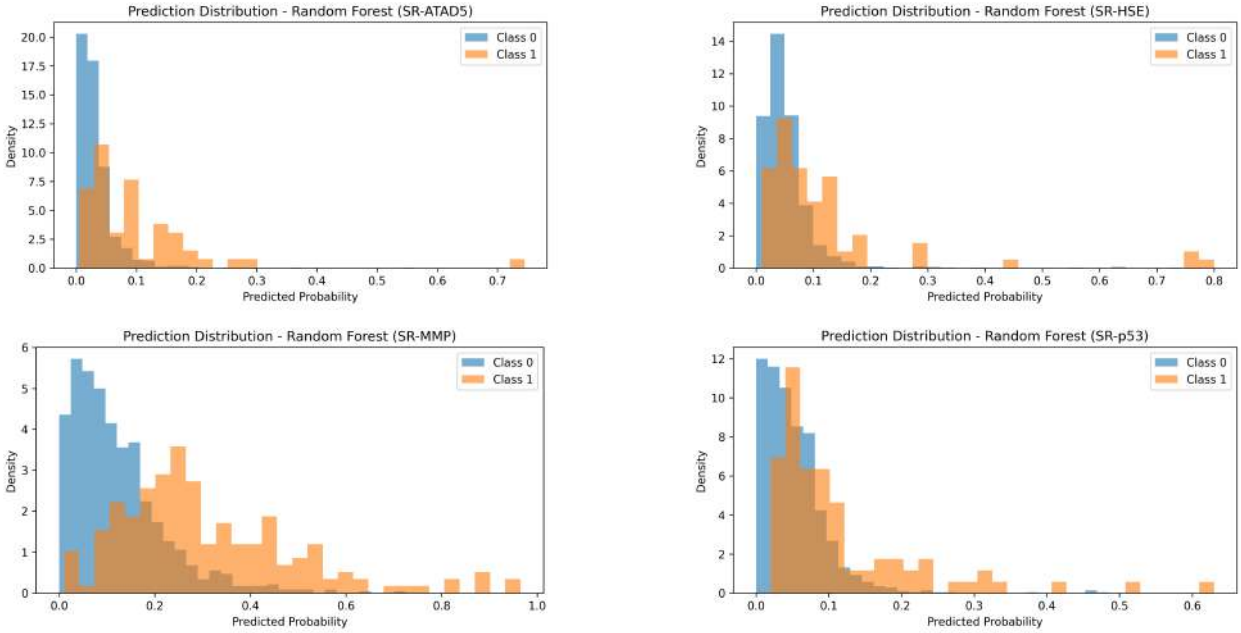


Figure 38: Random Forest model prediction histograms (part III).

Chapter C XAI

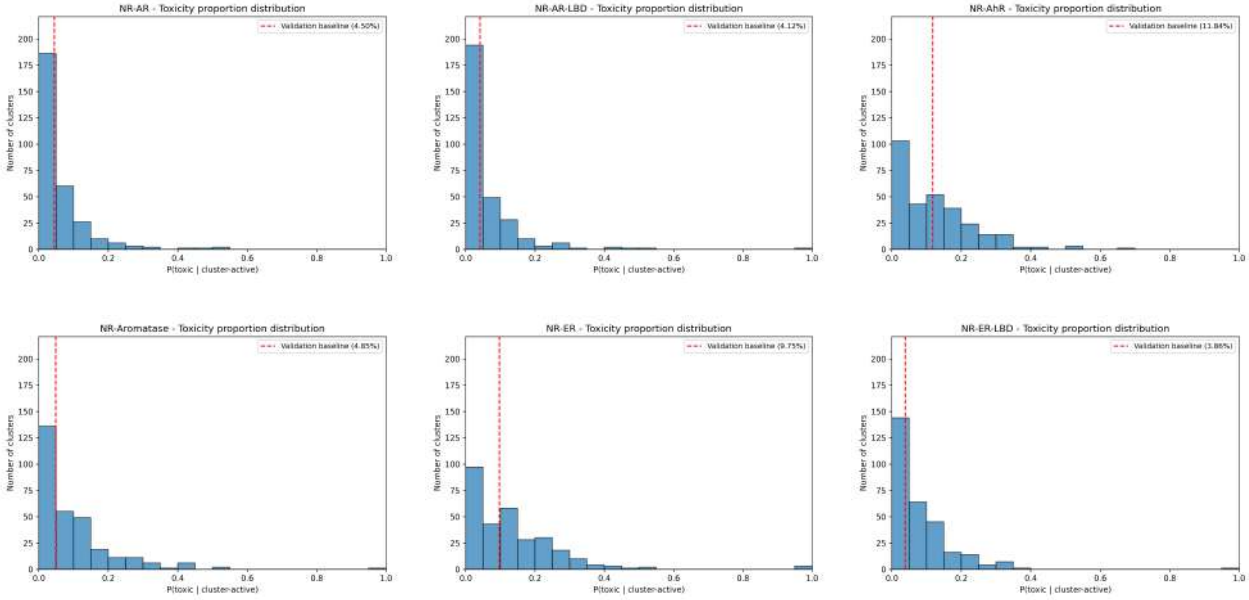


Figure 39: Distribution of $P(\text{toxic} \mid \text{cluster-active})$ (part I).

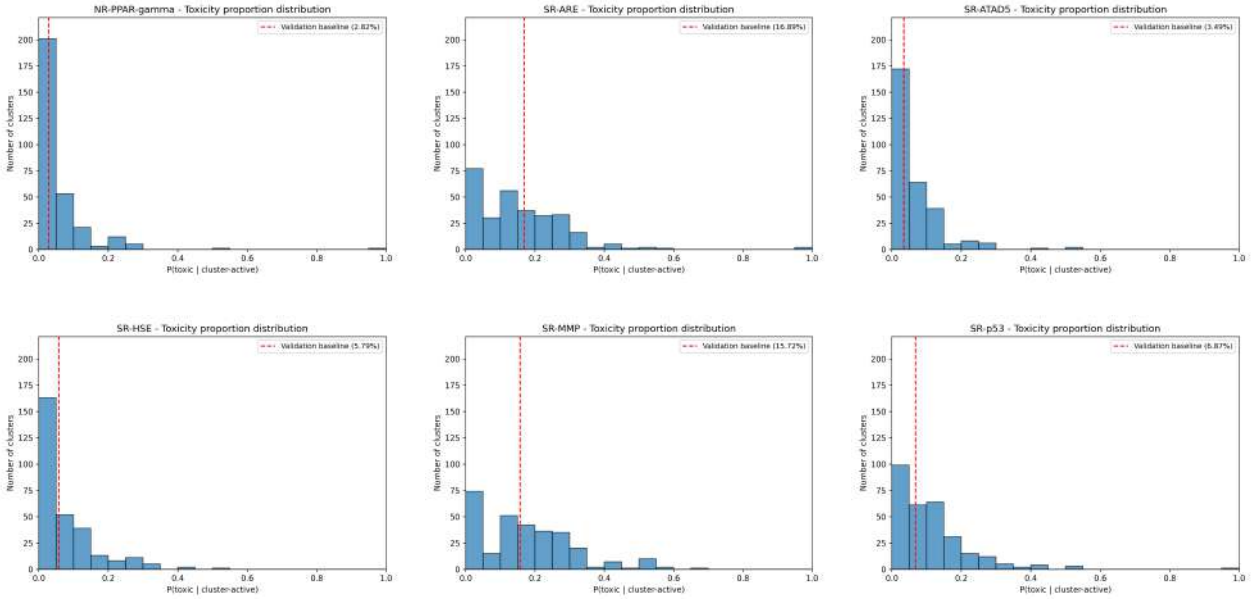


Figure 40: Distribution of $P(\text{toxic} \mid \text{cluster-active})$ (part II).