Supplementary Material

MolToxPred: Small molecule toxicity prediction using machine learning approach

Anjali Setiya[a], Vinod Jani [a], Uddhavesh Sonavane [a] and Rajendra Joshi[a*]

[a] HPC-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Innovation Park, Panchawati, Pashan, Pune, India-411008.

[*]Corresponding author email: rajendra@cdac.in

2.3.1 Feature Selection for molecular descriptors

- **Calculation of Mutual Information**

MI between feature X and target variable Y can be calculated as shown in eq. (1)

$$\text{MI(feature;target)} = \text{Entropy(feature)} - \text{Entropy(feature|target)} \tag{1}$$

Mathematically,

$$I(X,Y) = H(X) + H(Y) - H(X|Y) \tag{2}$$

Wherein eq. (2), H(X) and H(Y) is the marginal entropy and H(X) is related to how much information can be learned of the random descriptor/feature X. H(X|Y) is the joint entropy that measures the uncertainty when considering both features and the target variable together. When the features are continuous values, the mutual information is as follows:

$$I(X,Y) = -logy + logy|x \tag{3}$$

In eq. (3), consider X a continuous variable and Y as a discrete variable, drawn from probability density μ(x,y). μ(y) is the probability density for sampling y irrespective of the value of, and μ(y|x) = μ(x,y)/p(x) is the probability density for sampling y given a particular value of x.

- **Selected Molecular descriptors**

IPC, MaxEStateIndex, MinEStateIndex, MinAbsEStateIndex, qed,
MolWt, MaxPartialCharge, MinPartialCharge, MinAbsPartialCharge,

FpDensityMorgan1, FpDensityMorgan2, BCUT2D_MWHI, BCUT2D_MWLOW, BCUT2D_CHGHI, BCUT2D_MRHI, BCUT2D_MRLOW, BalabanJ, Chi2v, Chi3v, Chi4v, HallKierAlpha, Kappa1, Kappa2, Kappa3, PEOE_VSA1, PEOE_VSA10, PEOE_VSA11, PEOE_VSA12, PEOE_VSA14, PEOE_VSA2, PEOE_VSA3, PEOE_VSA6, PEOE_VSA7, PEOE_VSA8, PEOE_VSA9, SMR_VSA1, SMR_VSA10, SMR_VSA3, SMR_VSA5, SMR_VSA6, SMR_VSA7, SlogP_VSA1, SlogP_VSA10, SlogP_VSA12, SlogP_VSA2, SlogP_VSA3, SlogP_VSA5, SlogP_VSA8, TPSA, EState_VSA1, EState_VSA10, EState_VSA2, EState_VSA3, EState_VSA4, EState_VSA5, EState_VSA8, EState_VSA9, VSA_EState1, VSA_EState10, VSA_EState2, VSA_EState3, VSA_EState4, VSA_EState5, VSA_EState7, VSA_EState8, FractionCSP3, NHOHCount, NumAliphaticRings, NumAromaticHeterocycles, NumAromaticRings, NumRotatableBonds, RingCount, MolLogP, fr_Ar_N, fr_NH1, fr_bicyclic

2.3.2 Feature Selection for molecular fingerprints

Chi-square Test

Chi-square statistic (4) for a feature can be defined as:

$$\chi^2 = \sum_{i=1}^{C}\sum_{j=1}^{I}\frac{(N_{ij} - E_{ij})^2}{E_{ij}} \tag{4}$$

Where, N: Total number of samples i.e. compounds

I: Number of intervals

N_{ij}: Number of samples in class C_i within the j^{th} interval

E_{ij}: Expected frequency of N_{ij} (5)

$$E_{ij} = \frac{M_{Ij} \ * \ C_i}{N} \tag{5}$$

M_{Ij} : Number of samples in j^{th} interval

When two variables are not dependent, the observed count is close to the expected count, thus a smaller chi-square value. So high chi-square value indicates that the null hypothesis of independence is incorrect. Chi-square statistics are calculated using the scipy.stats[1] library in Python.

Bonferroni Corrections

The Bonferroni correction helps to control the family-wise error rate (FWER) i.e. the probability of making at least one Type I error among multiple statistical analyses[2,3]. FWER can be controlled below a certain threshold (generally <5%) by applying a Bonferroni correction and thus allowing very few occurrences of false positives[3].

In this analysis, the chi-square test is employed for feature selection, treating each fingerprint as an independent feature set. Each feature within a fingerprint is individually tested for its association with the target variable. The null hypothesis (H0) posits that there is no dependence between the feature and the target variable, while the alternate hypothesis (H1) asserts their dependence. Given that each bit within a fingerprint constitutes a separate feature, concurrently multiple hypotheses are being tested during feature selection.

To address the issue of multiple hypothesis testing, the Bonferroni correction has been applied to the chi-square test. This correction is executed independently for each fold in our five-fold cross-validation process. The p-value of each feature (pi) is multiplied by the number of performed statistical tests (npi): n i.e. here length of the fingerprint type to compute the corrected p-values3. The corrected p-values are then compared with the predefined significance level (0.05), and features with corrected p-values below this threshold are considered significant for that specific fold by rejecting the null hypothesis.

CramersV Test

The CramersV function in eq (6) is used from the association-metrics package[4] in python to compute this statistic. If X and Y are two categorical variables, such that X has m categories, labeled $X_1,\ldots\ldots\ldots\ldots,X_m$ and Y has n categories, labeled $Y_1,\ldots\ldots\ldots\ldots,Y_n$, and N is total pairs of observations of occurrences of both categories m and n, then Cramers V is defined as:

$$V = V(X,Y) = \sqrt{\frac{\chi^2}{N \min(m - 1, n - 1)}} \qquad (6)$$

4.4 Selection of optimal Hyperparameters

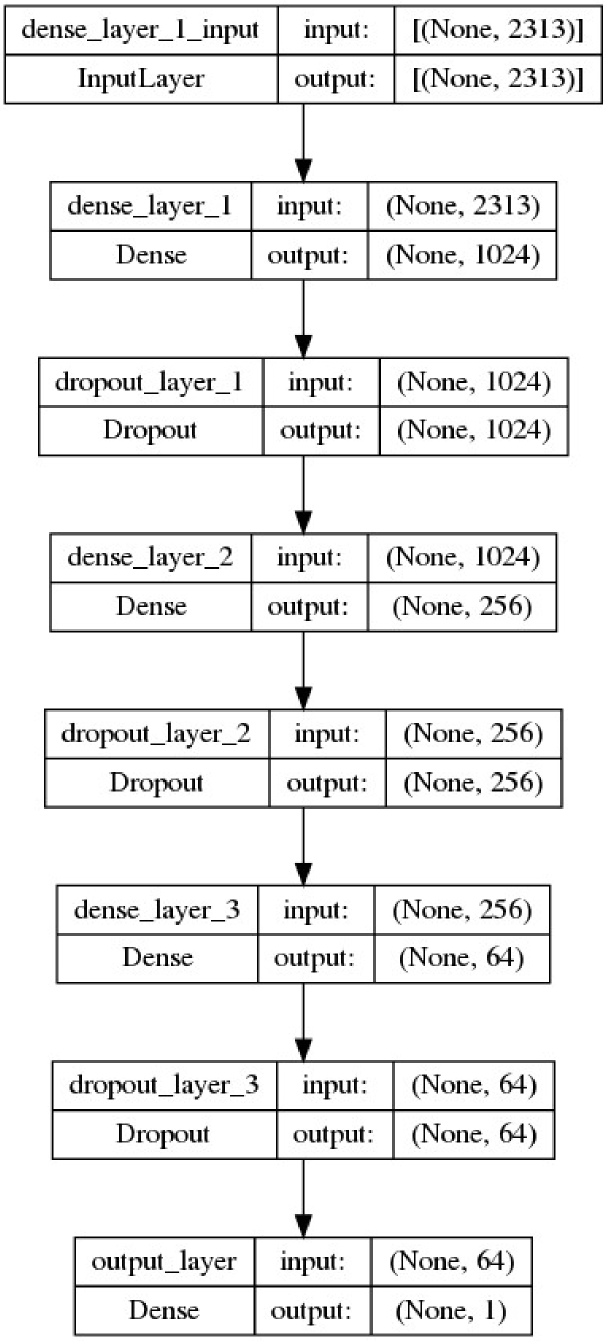

Fig S1: The Multi Layer Perceptron feed-forward architecture

4.5 Toxicity Label prediction

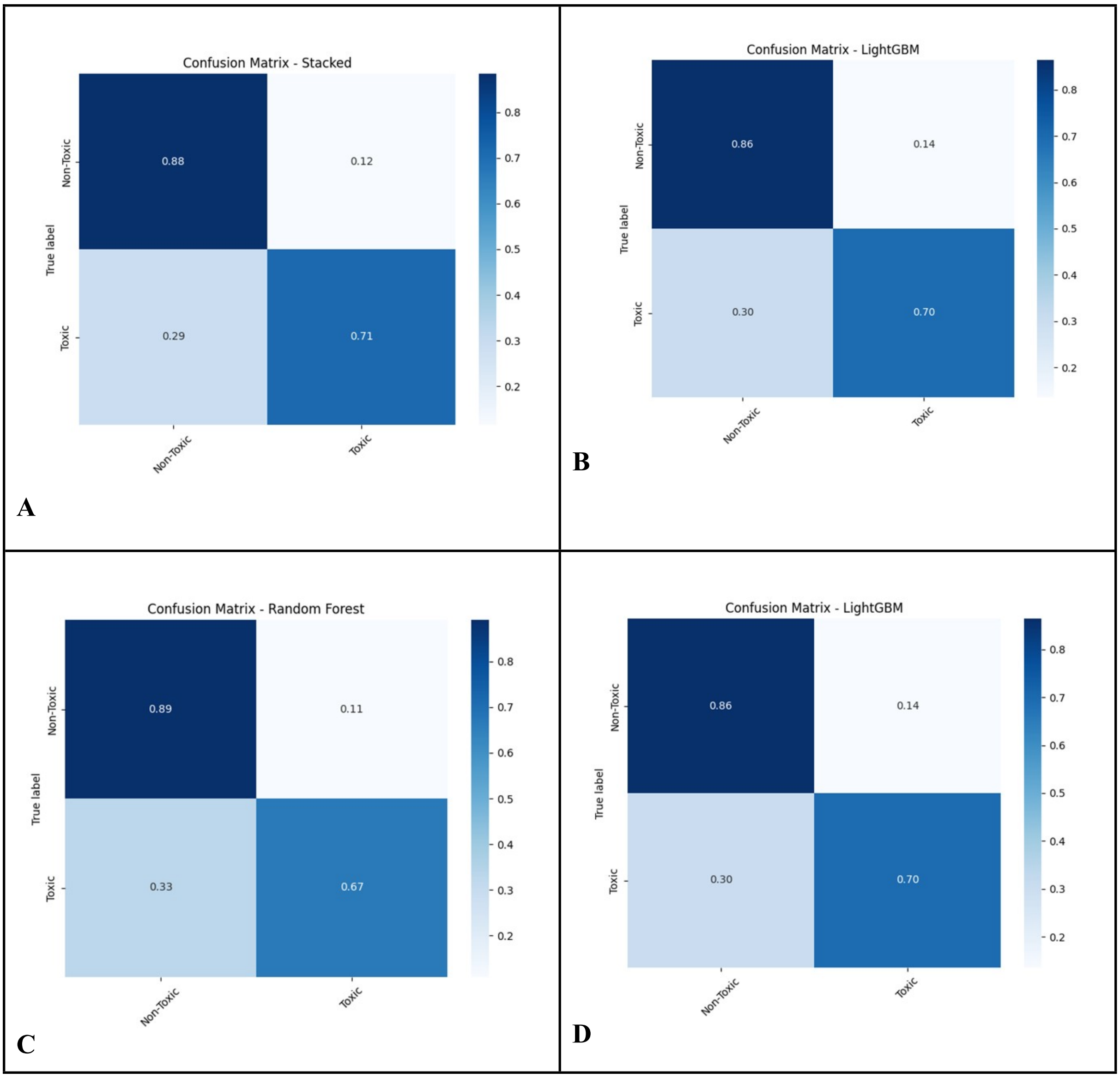

Fig S2: Confusion matrix on the test set for A) Stacked Model B)LightGBM C)Random Forest D)MLP

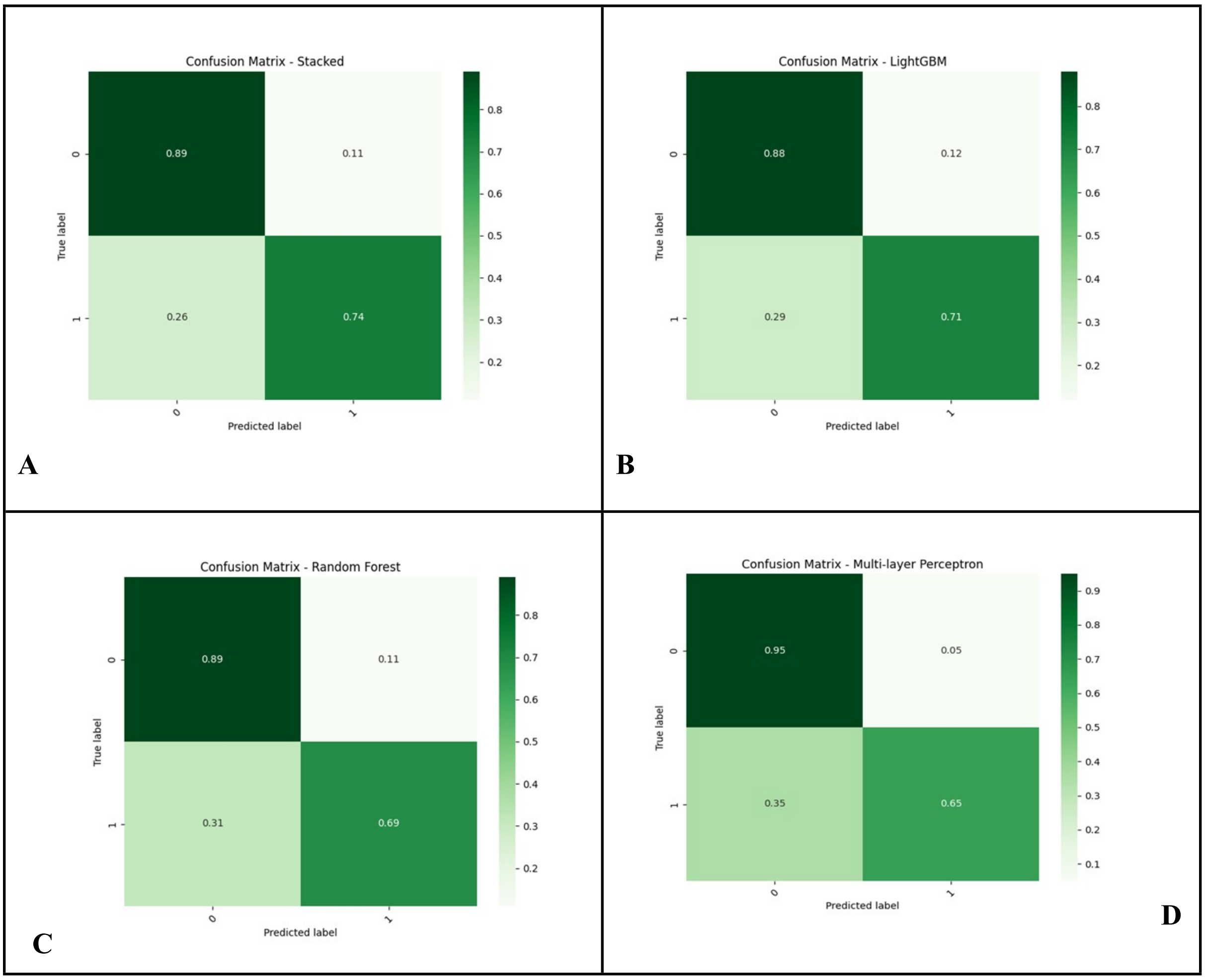

Fig S3: Confusion matrix on external validation set for A) Stacked Model B)LightGBM C)Random Forest D)MLP

Table S1: Five Fold Stratified Cross-Validation AUROC Scores Comparison for all base models

Model	AUC Train	AUC Test
Random Forest	0.930 ± 0.002	0.862 ± 0.002
Multi-layer Perceptron	0.892 ± 0.008	0.864 ± 0.006
LightGBM	0.906 ± 0.002	0.864 ± 0.002

Towards identifying the correlation between Structural alerts and biological pathways for toxicity

Table S2: Substructure matching result metrics on Tox21 test set

	Combined	nr-ahr	nr-ar	nr-ar-lbd	nr-aromatase	nr-er	nr-er-lbd	nr-ppar-gamma	sr-are	sr-atad5	sr-hse	sr-mmp	sr-p53
True Positives (TP)	116	29	1	1	12	11	8	5	30	8	4	33	17
False Positives (FP)	150	146	51	62	68	83	105	89	69	55	41	106	102
True Negatives (TN)	21	95	238	187	128	155	172	163	117	192	216	94	139
False Negatives (FN)	9	2	2	3	6	16	2	10	18	17	6	5	11

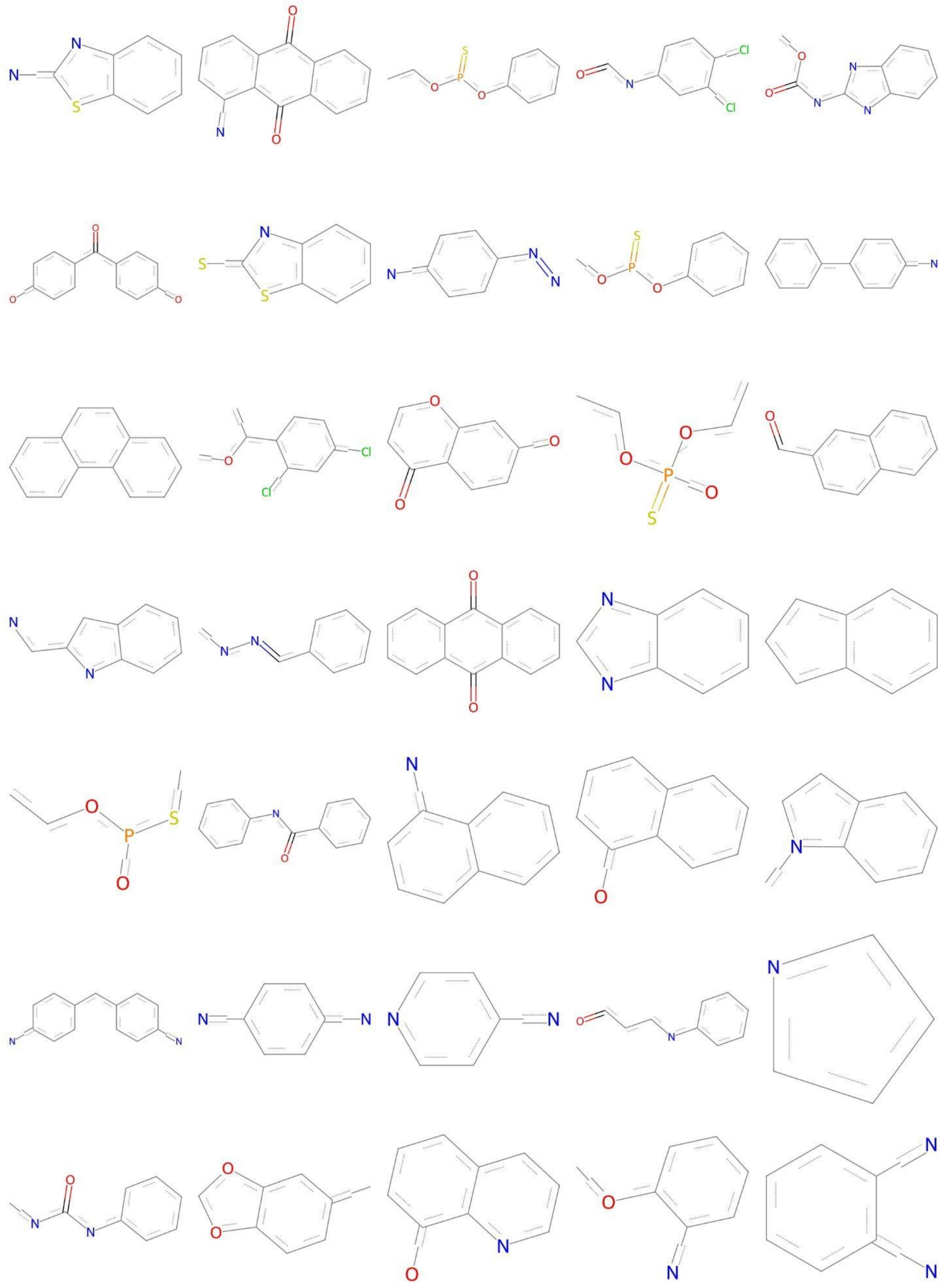

Fig S4A: Structural alerts that influenced the toxicity of nr-ahr data of Tox21. The order does not represent the toxic influence of each fragment.

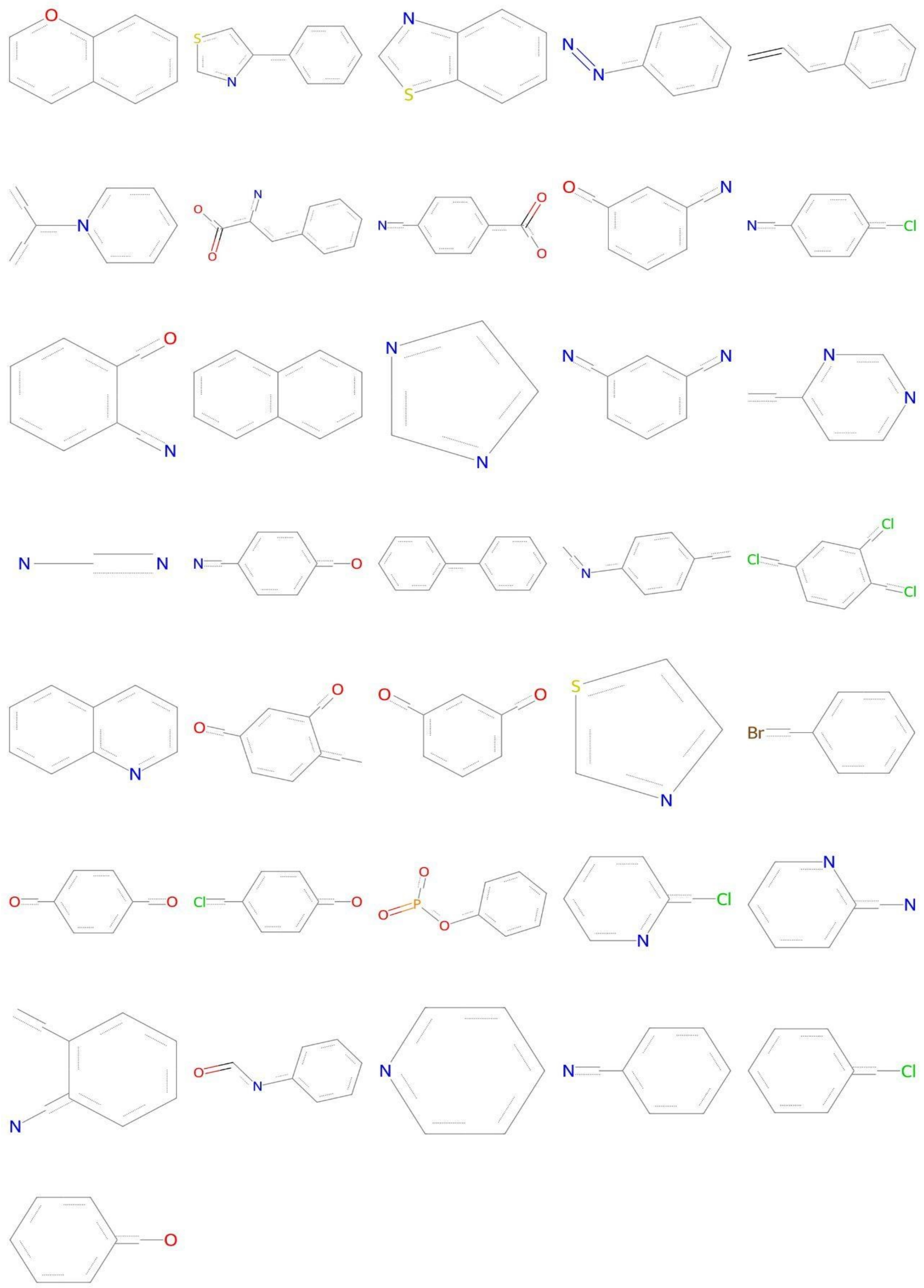

Fig S4B: Structural alerts that influenced the toxicity of nr-ahr data of Tox21. The order does not represent the toxic influence of each fragment.

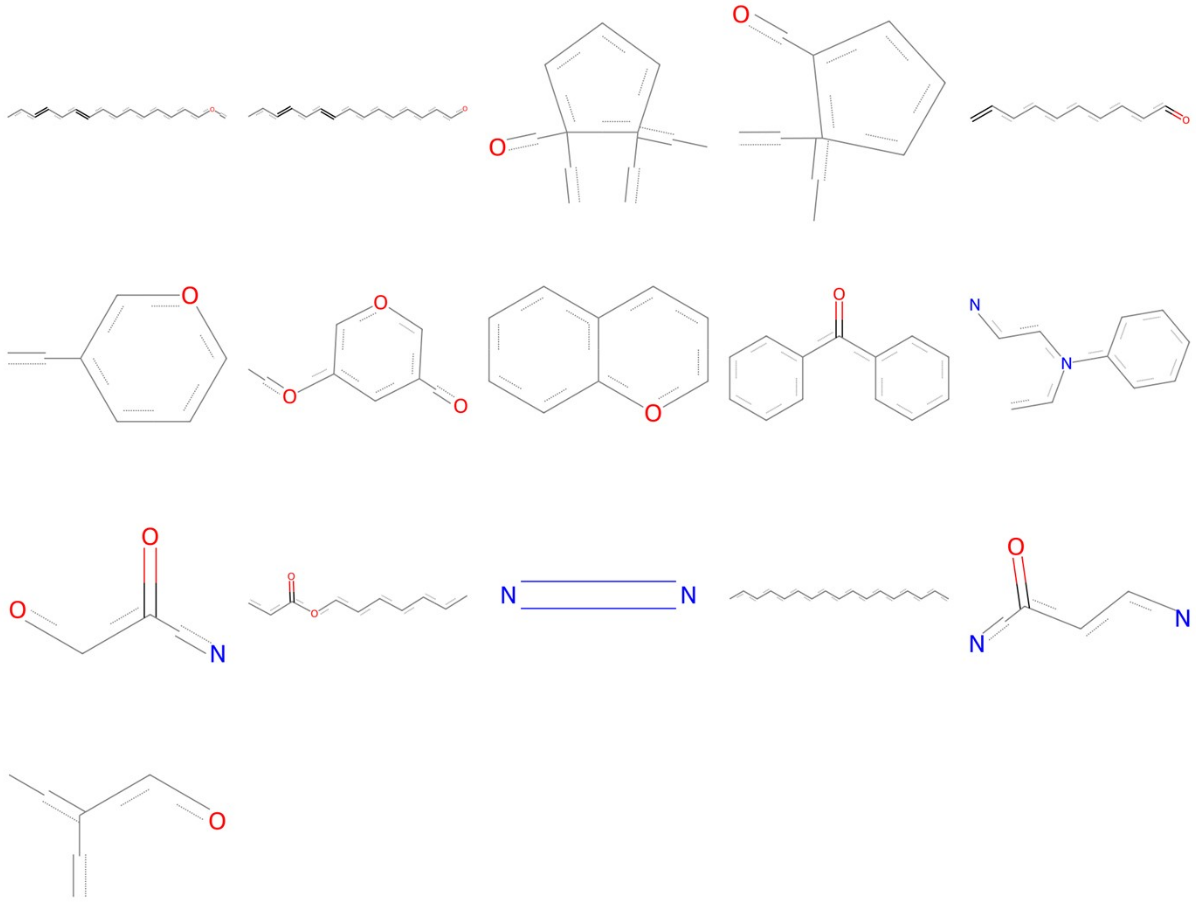

Fig S5: Structural alerts that influenced the toxicity of nr-ar data of Tox21. The order does not represent the toxic influence of each fragment.

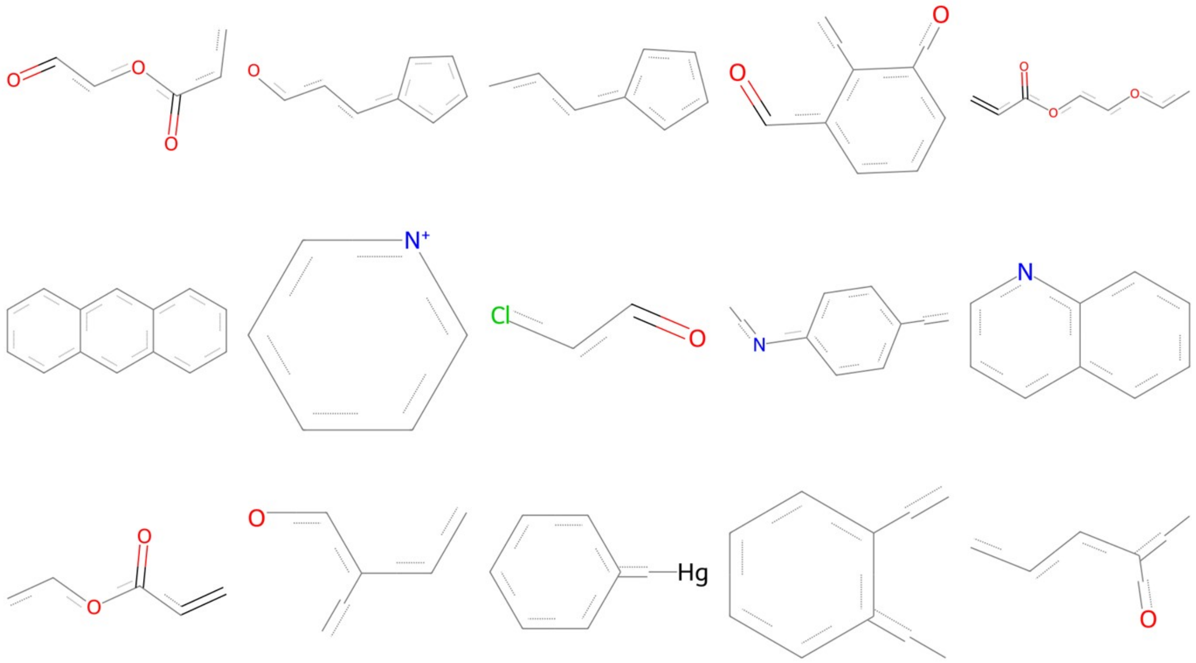

Fig S6: Structural alerts that influenced the toxicity of nr-ar-lbd data of Tox21. The order does not represent the toxic influence of each fragment.

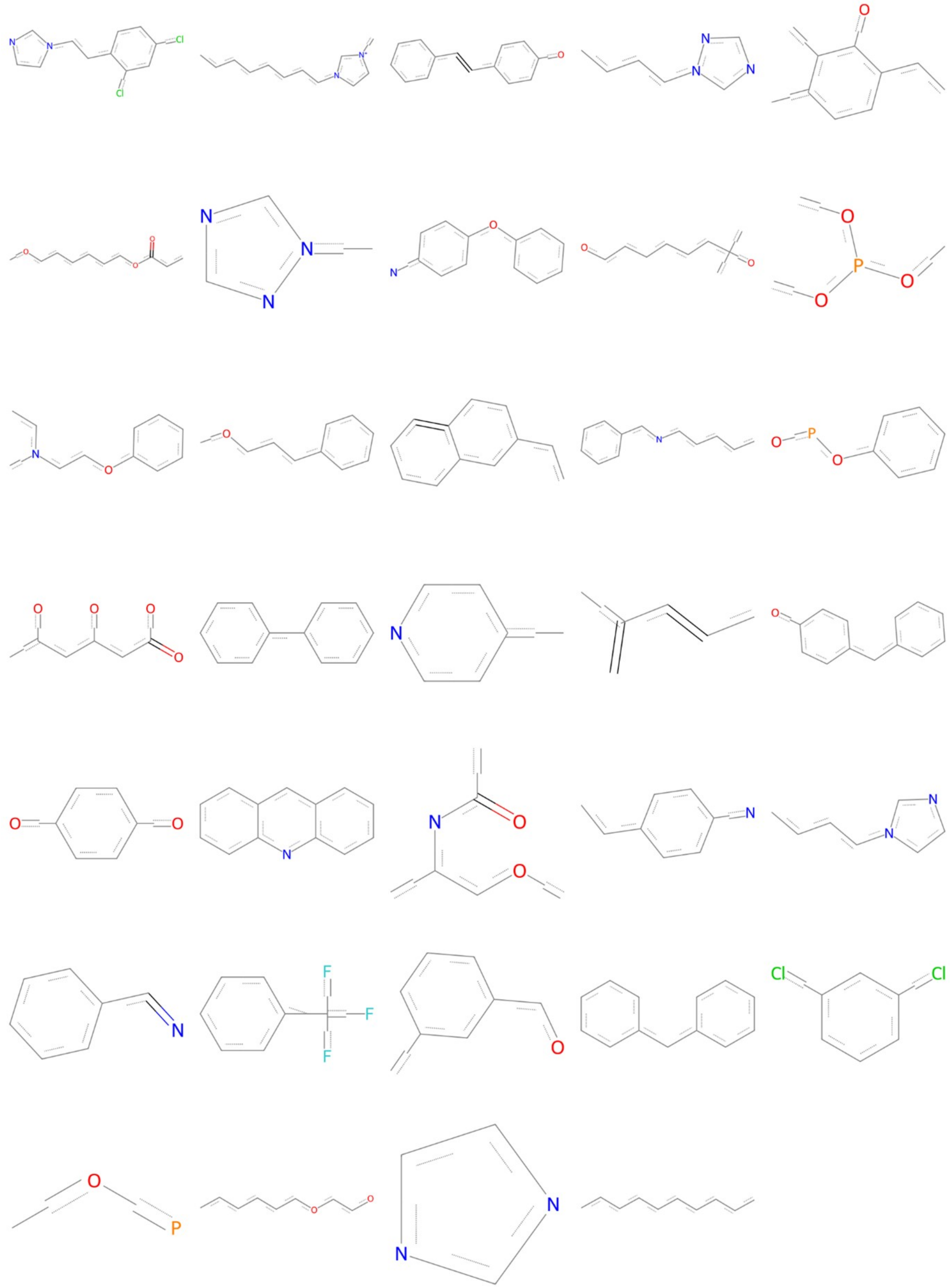

Fig S7: Structural alerts that influenced the toxicity of nr-aromatase data of Tox21. The order does not represent the toxic influence of each fragment.

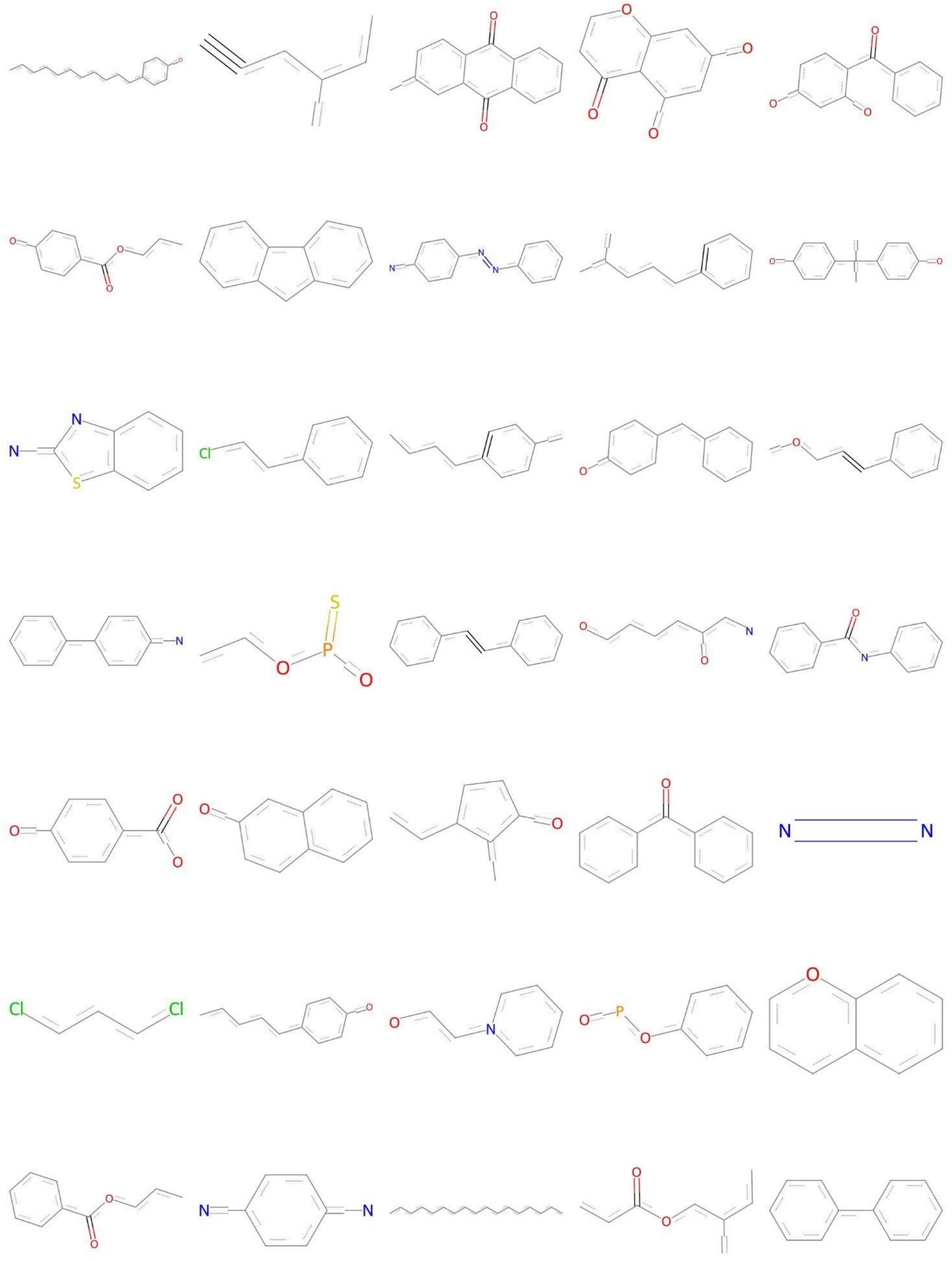

Fig S8A: Structural alerts that influenced the toxicity of nr-er data of Tox21. The order does not represent the toxic influence of each fragment.

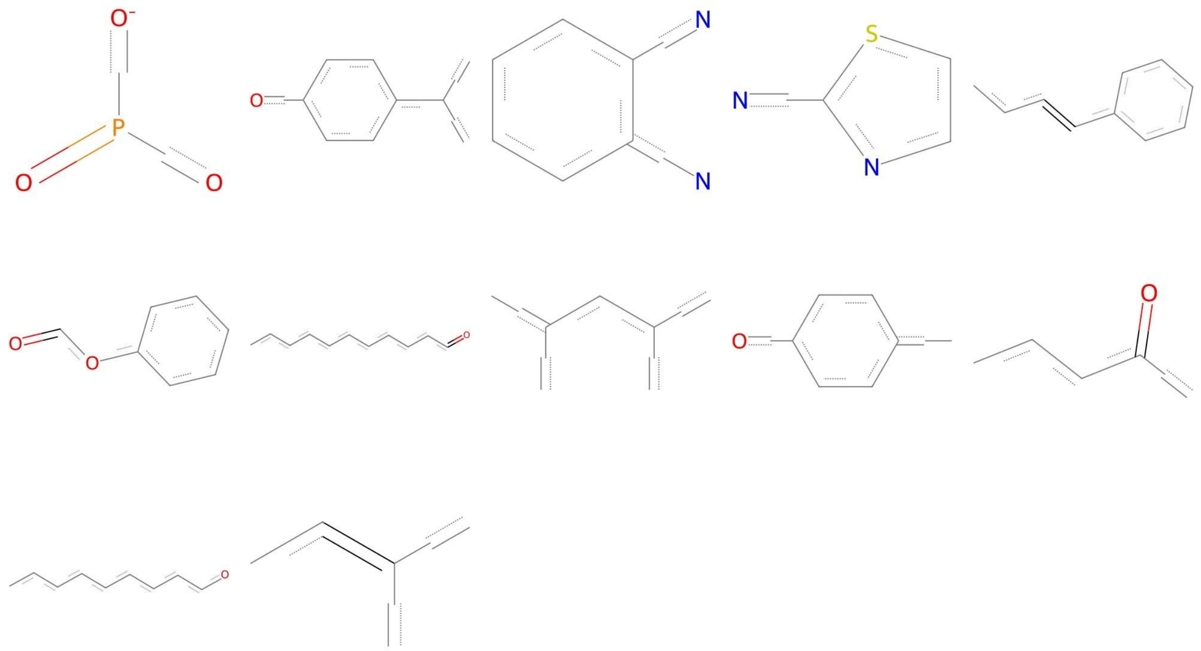

Fig S8B: Structural alerts that influenced the toxicity of nr-er data of Tox21. The order does not represent the toxic influence of each fragment.

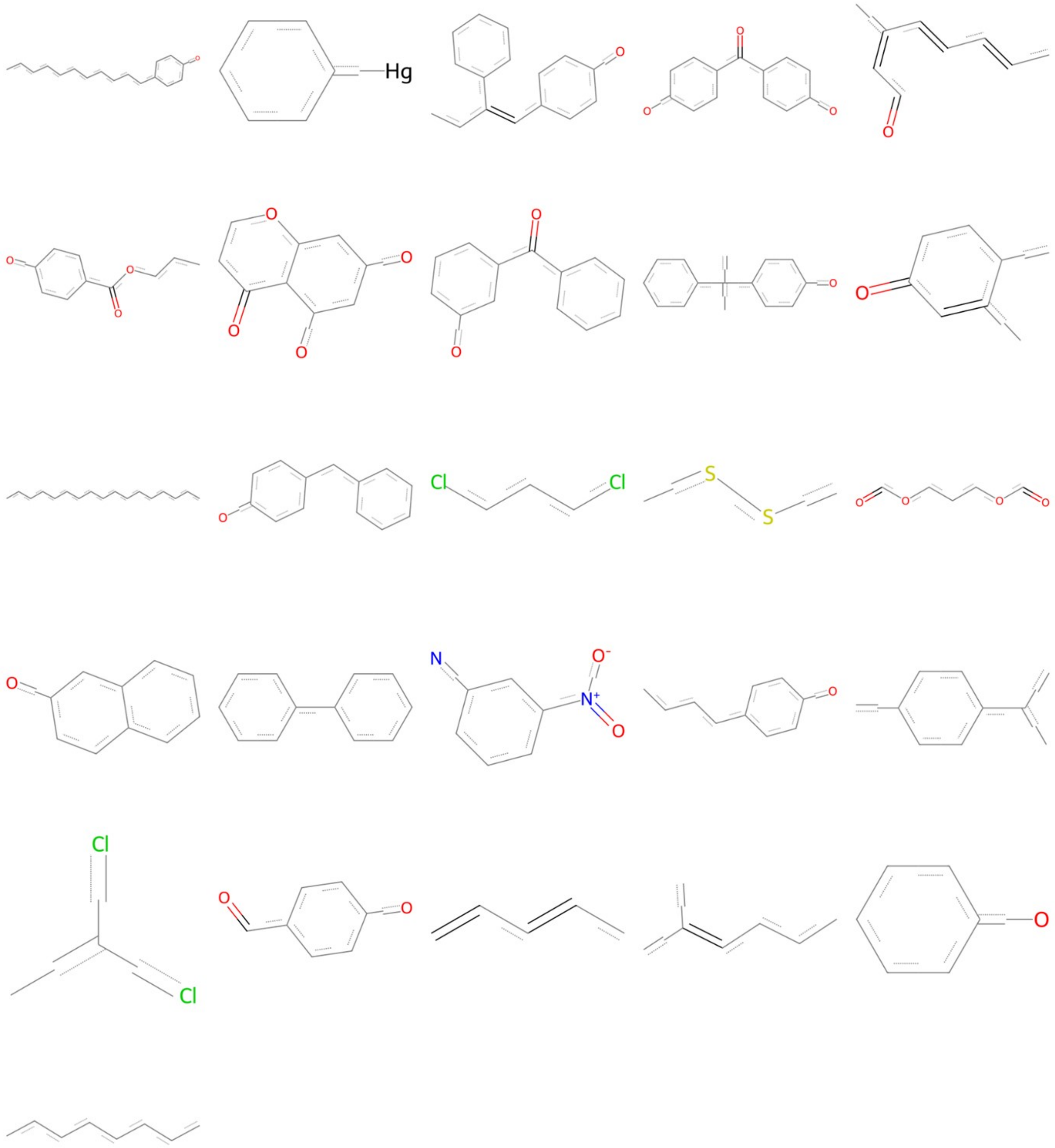

Fig S9: Structural alerts that influenced the toxicity of nr-er-lbd data of Tox21. The order does not represent the toxic influence of each fragment.

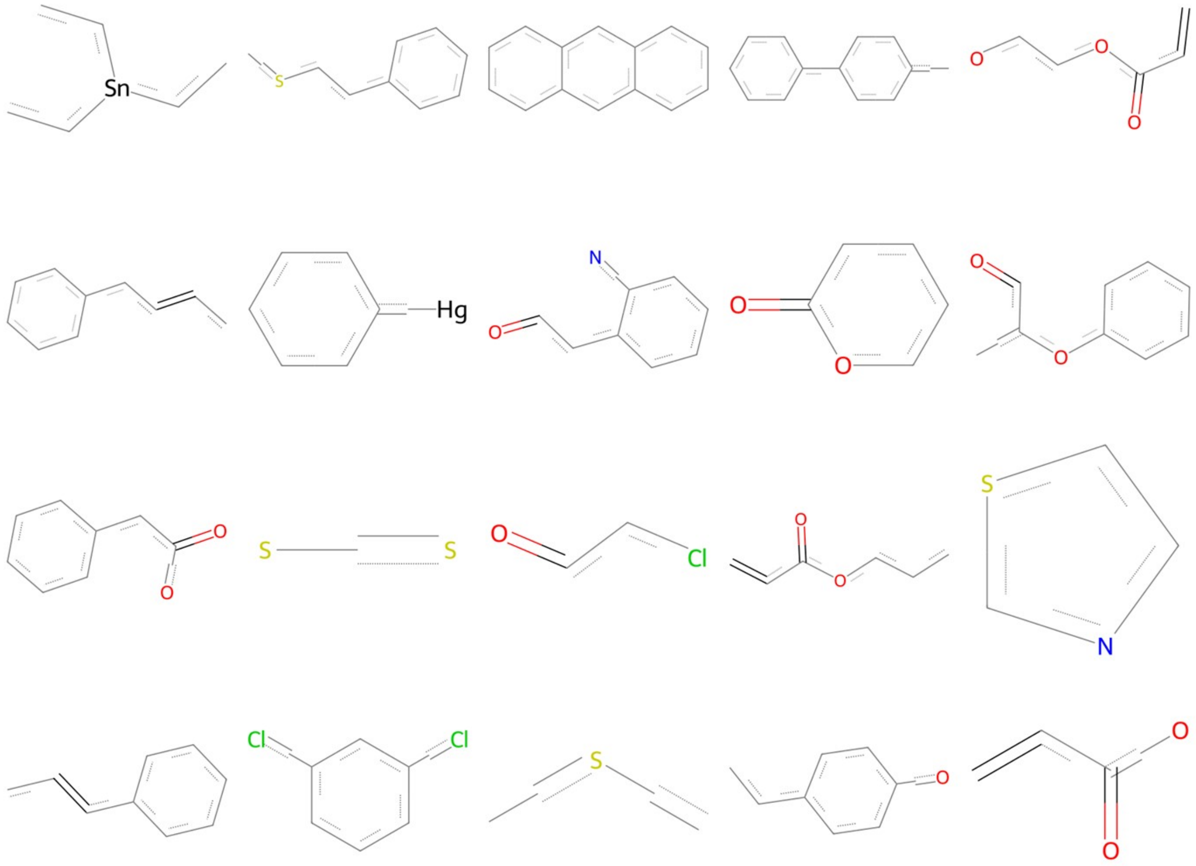

Fig S10: Structural alerts that influenced the toxicity of nr-ppar gamma data of Tox21. The order does not represent the toxic influence of each fragment.

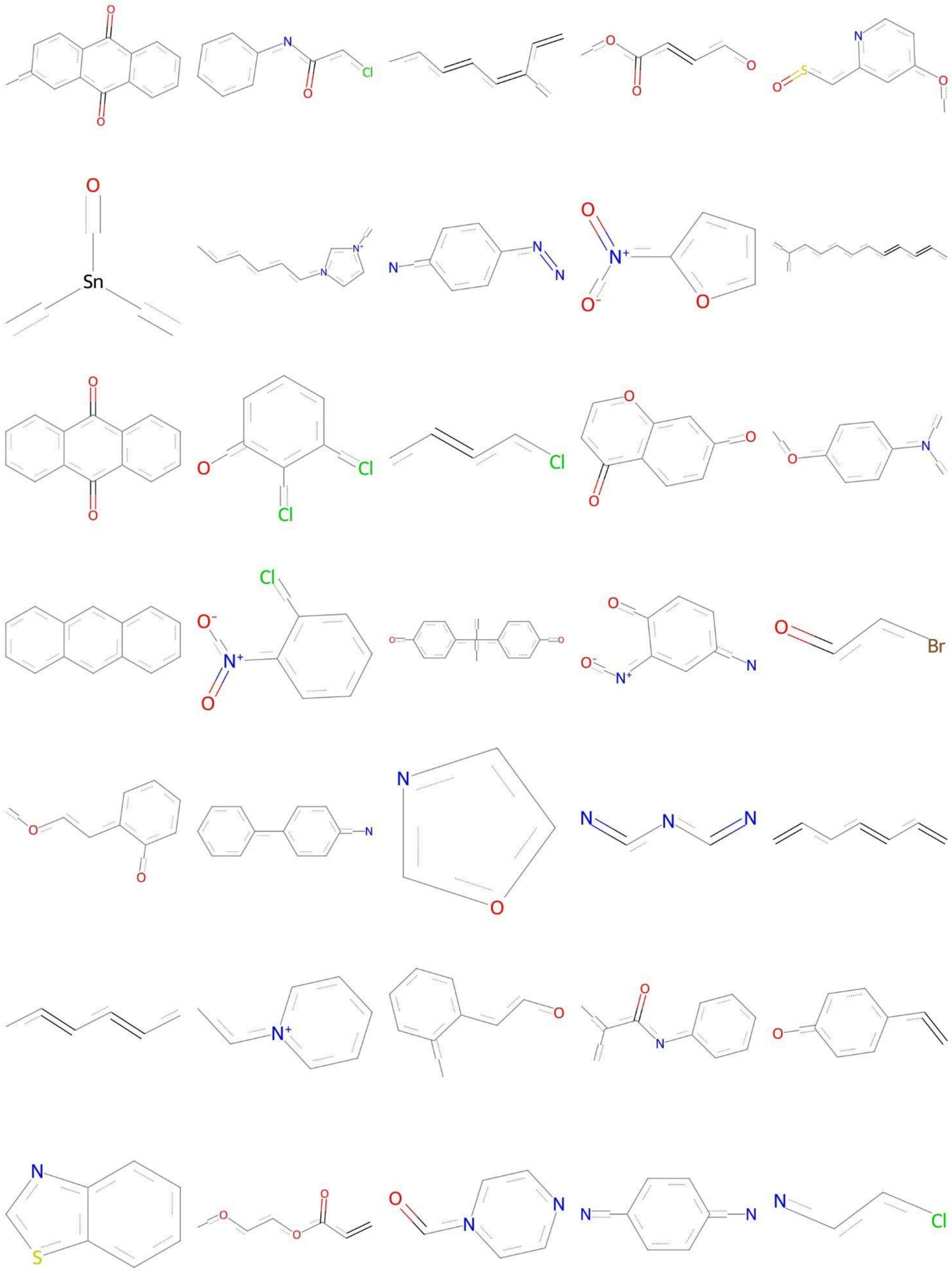

Fig S11A: Structural alerts that influenced the toxicity of sr-are data of Tox21. The order does not represent the toxic influence of each fragment.

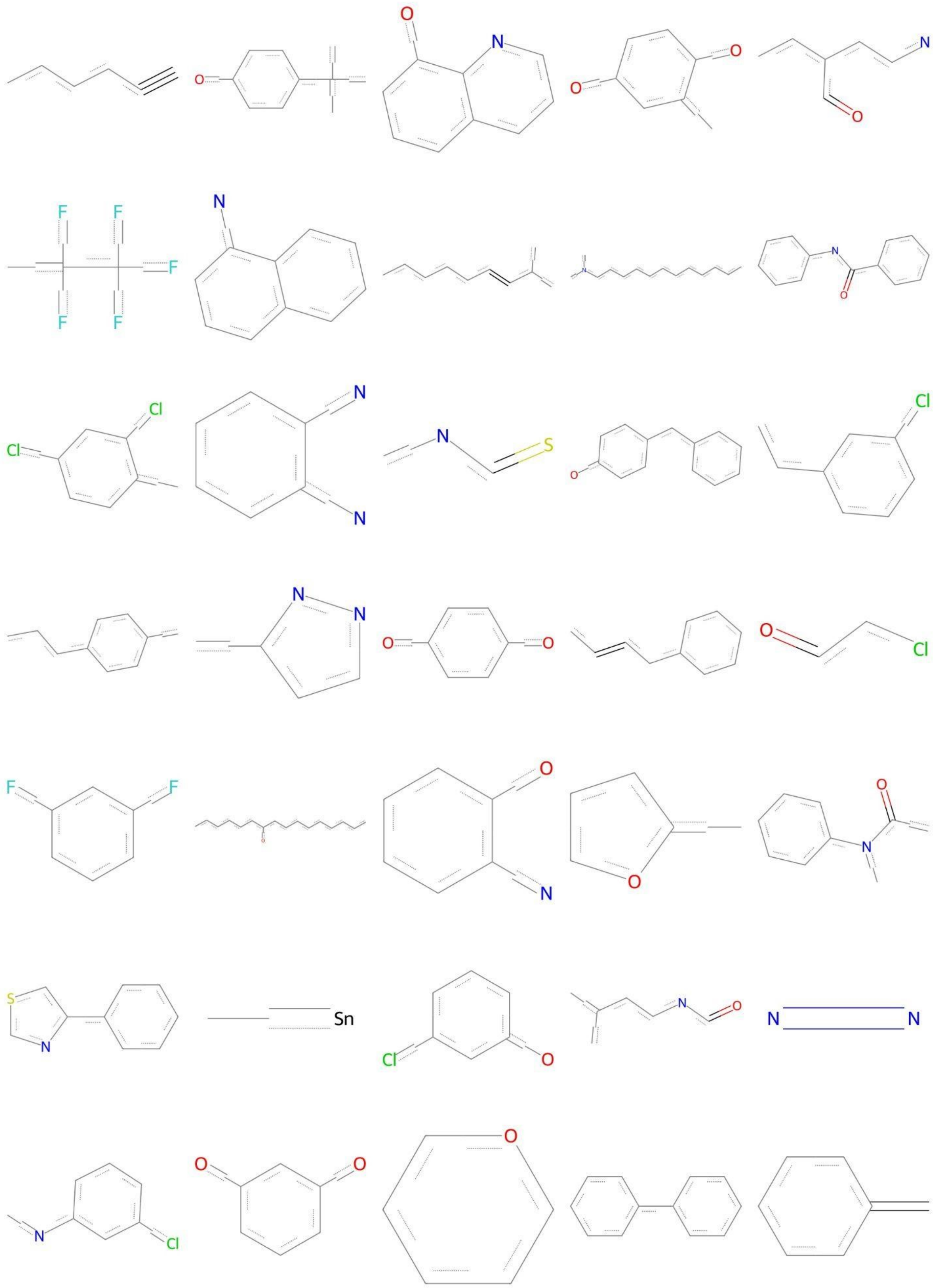

Fig S11B: Structural alerts that influenced the toxicity of sr-are data of Tox21. The order does not represent the toxic influence of each fragment.

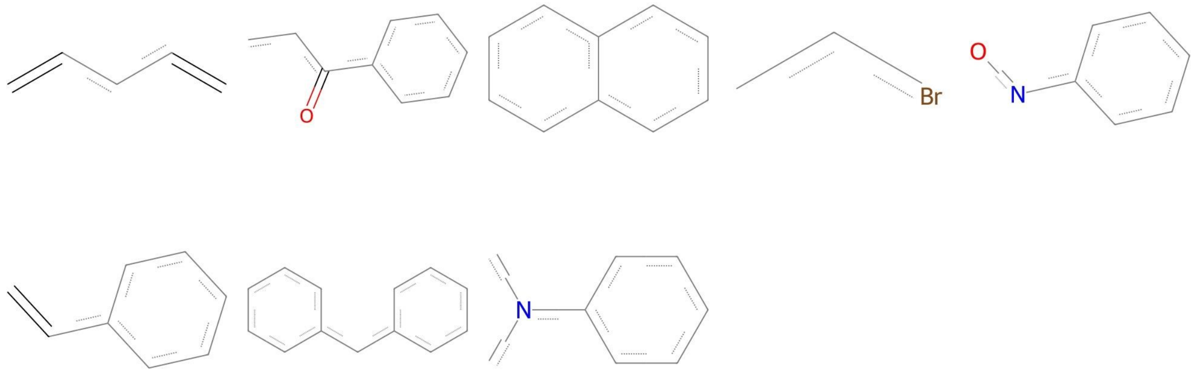

Fig S11C: Structural alerts that influenced the toxicity of sr-are data of Tox21. The order does not represent the toxic influence of each fragment.

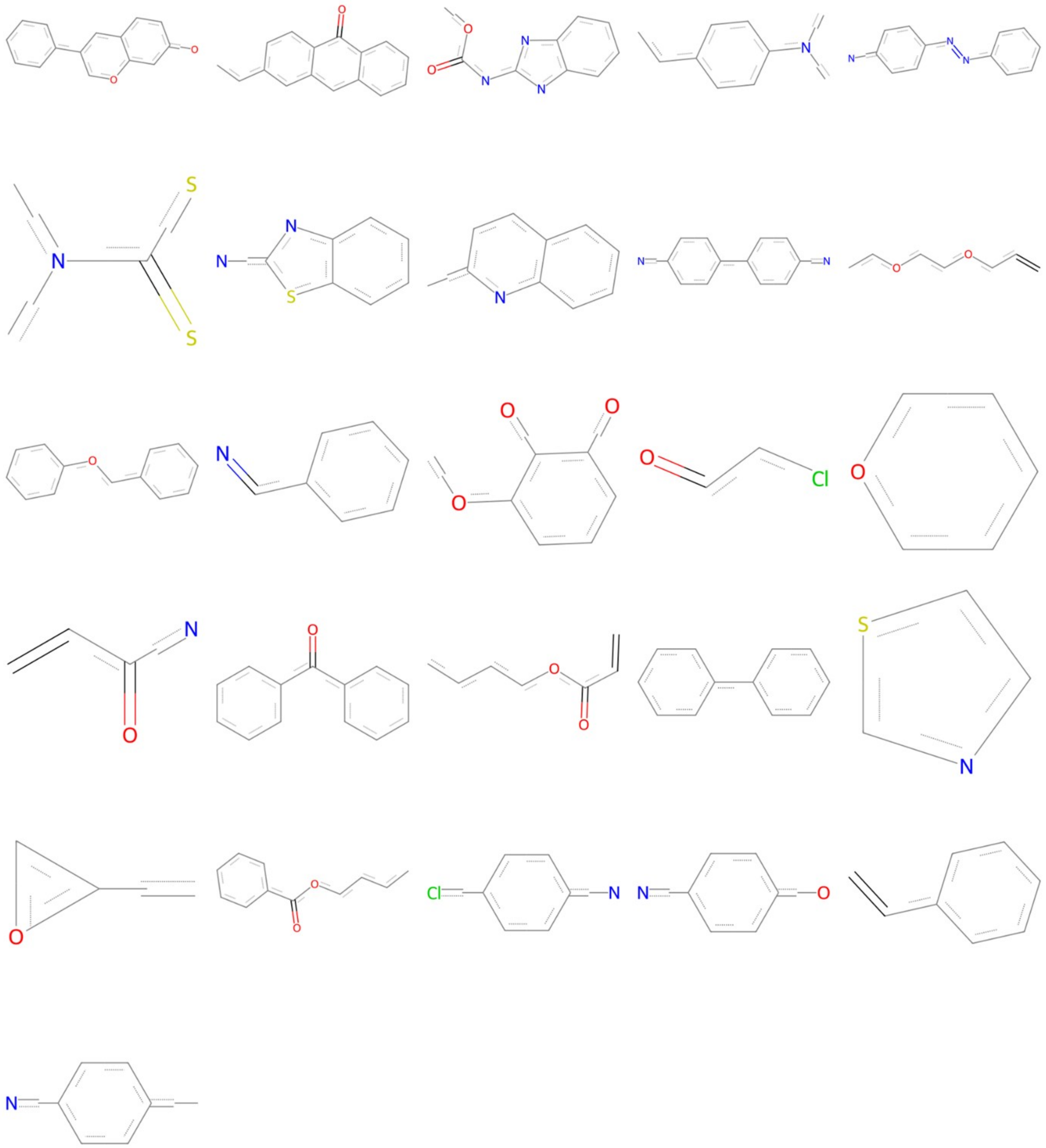

Fig S12: Structural alerts that influenced the toxicity of sr-atad5 data of Tox21. The order does not represent the toxic influence of each fragment.

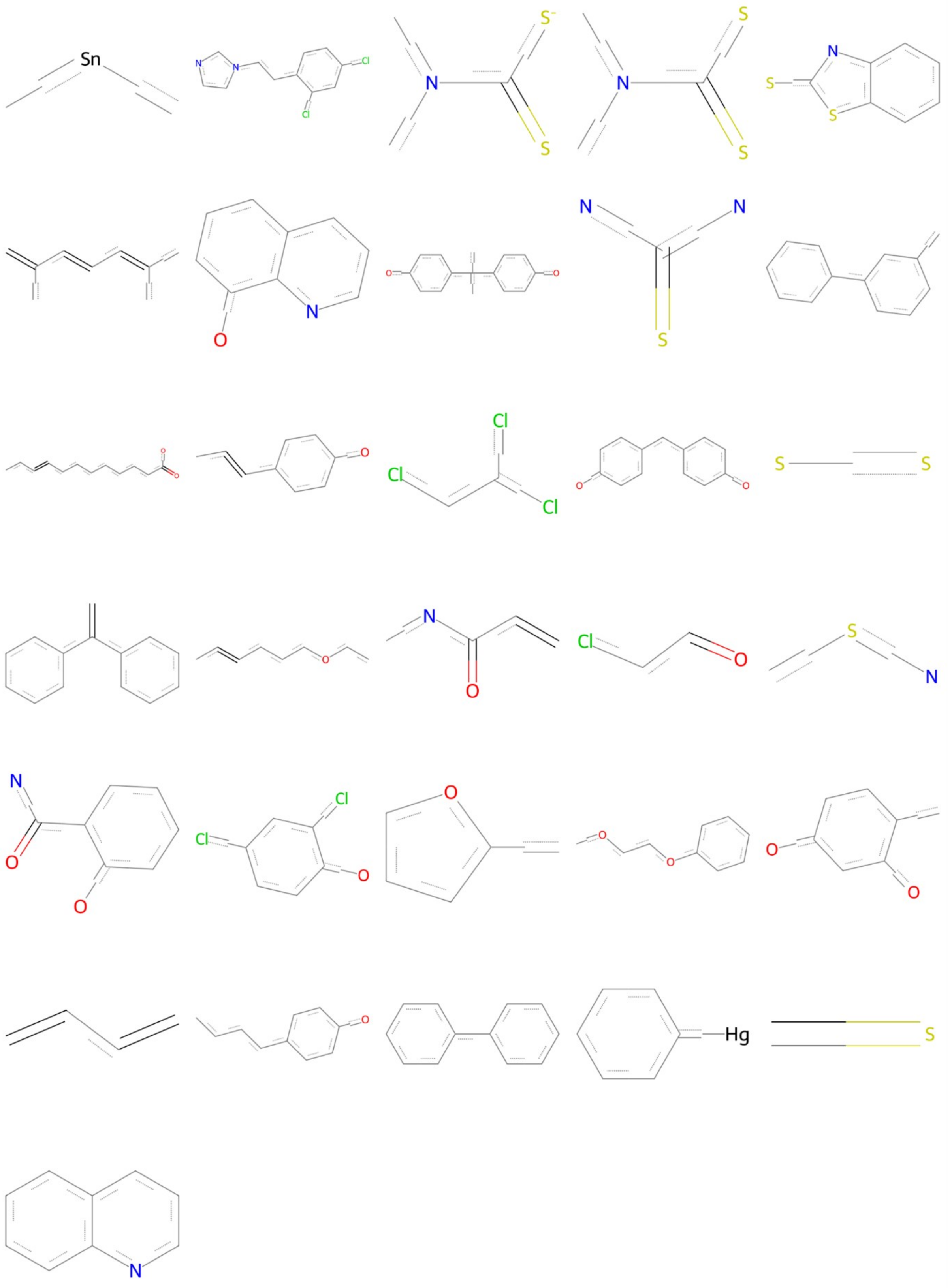

Fig S13: Structural alerts that influenced the toxicity of sr-hse data of Tox21. The order does not represent the toxic influence of each fragment.

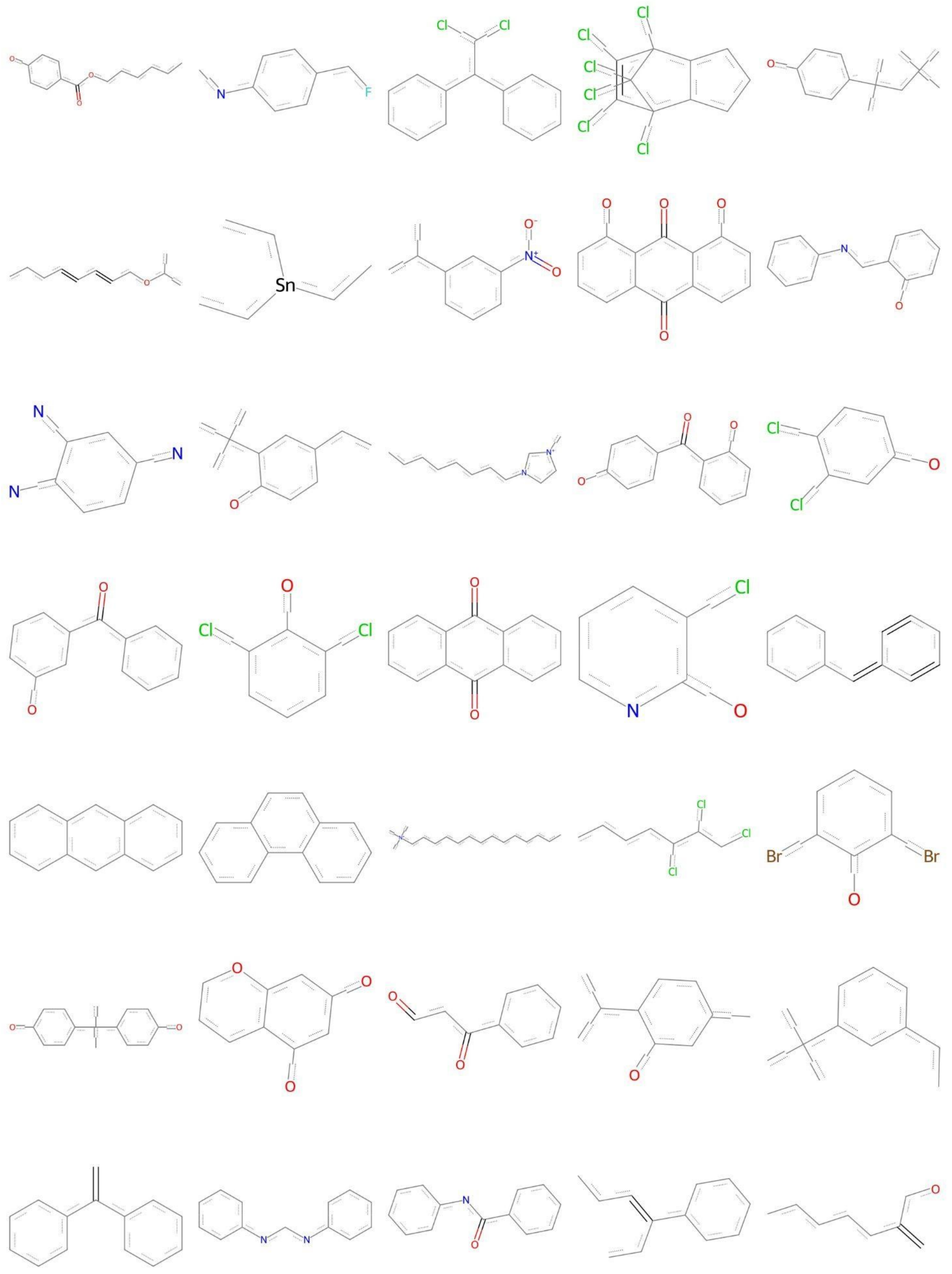

Fig S14A: Structural alerts that influenced the toxicity of sr-mmp data of Tox21. The order does not represent the toxic influence of each fragment.

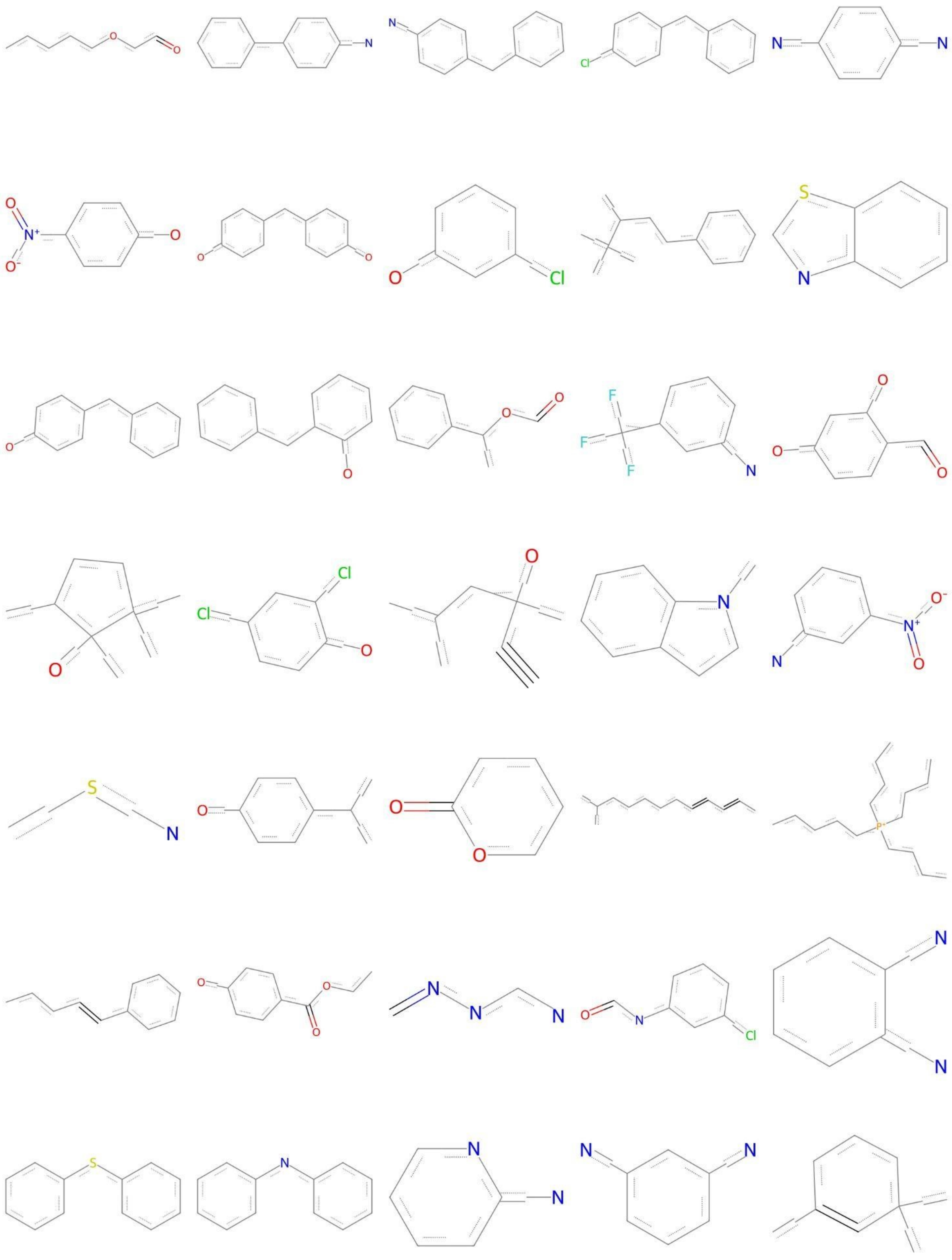

Fig S14B: Structural alerts that influenced the toxicity of sr-mmp data of Tox21. The order does not represent the toxic influence of each fragment.

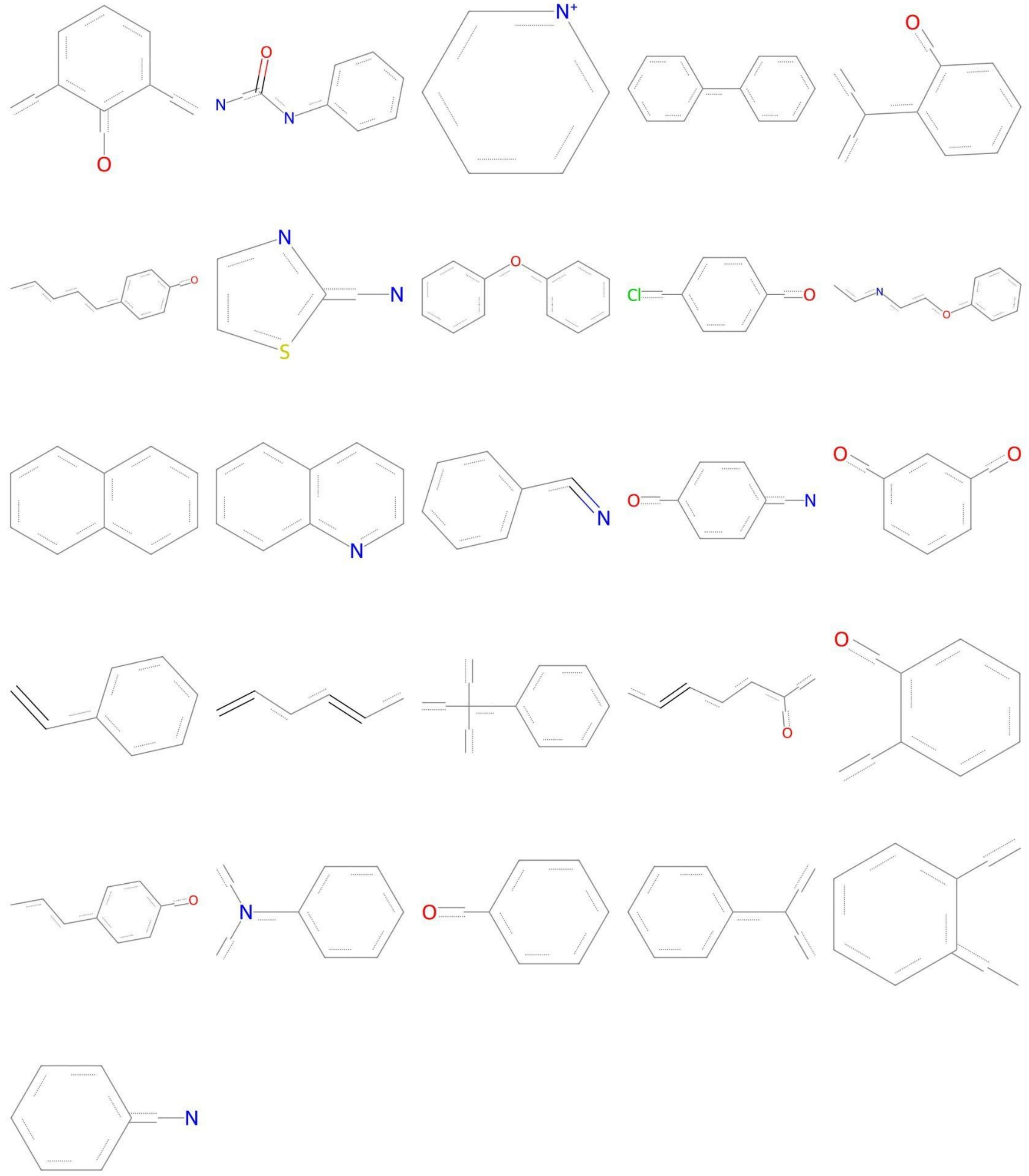

Fig S14C: Structural alerts that influenced the toxicity of sr-mmp data of Tox21. The order does not represent the toxic influence of each fragment.

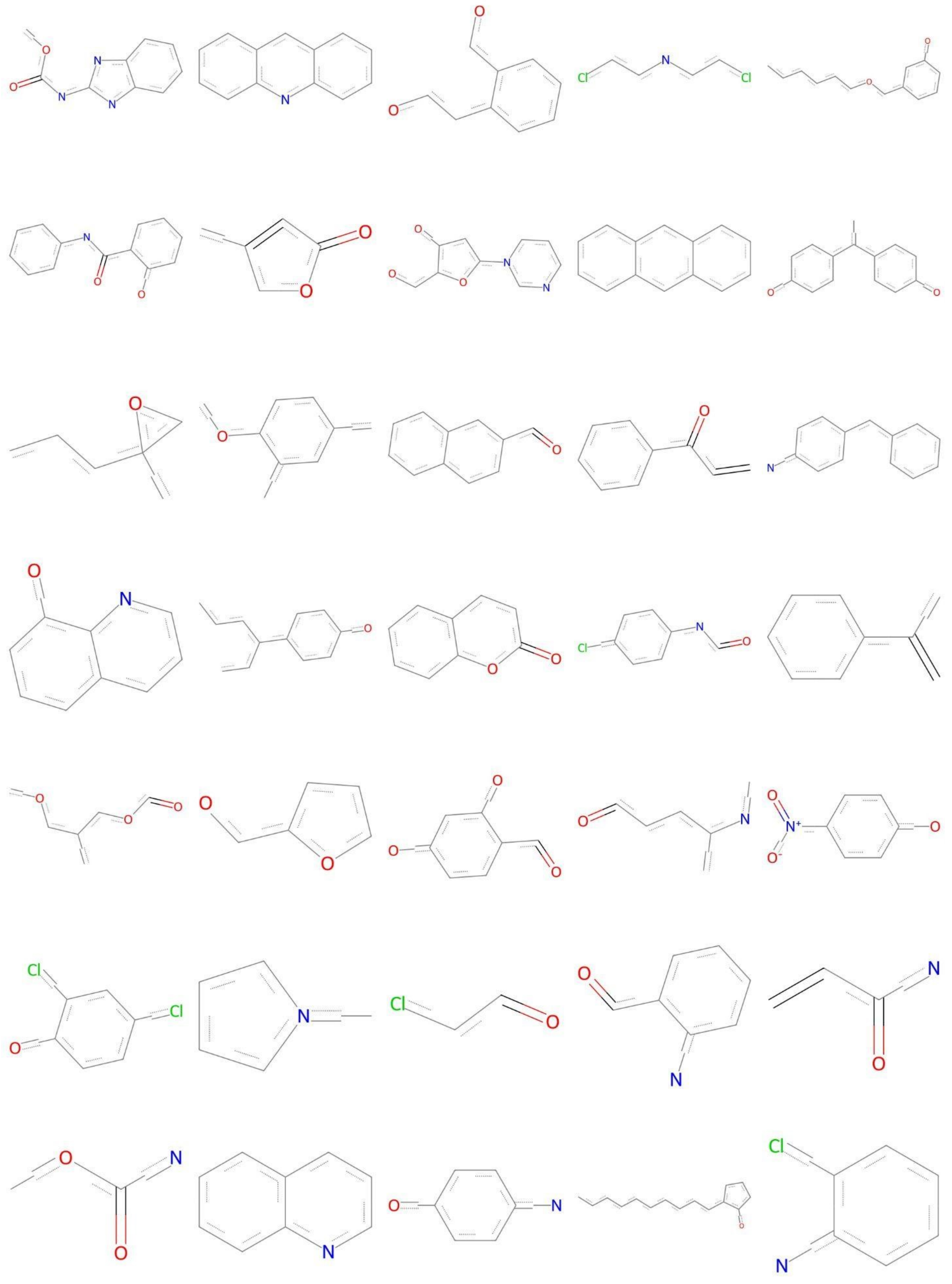

Fig S15A: Structural alerts that influenced the toxicity of sr-p53 data of Tox21. The order does not represent the toxic influence of each fragment.

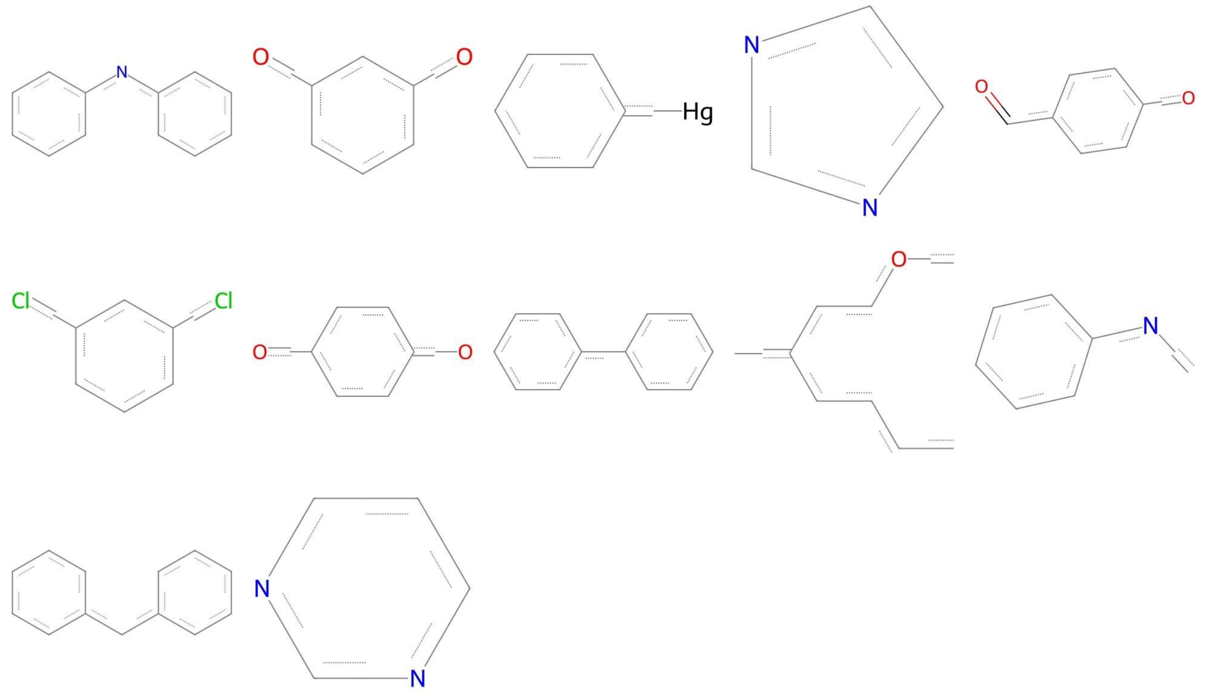

Fig S15B: Structural alerts that influenced the toxicity of sr-p53 data of Tox21. The order does not represent the toxic influence of each fragment.

4.3. Feature Importance

Optimal Estimator from Associated Uncertainties

In the permutation feature importance calculation for each of the three base models , we have K independent observations or measurements of permutation importance score (X) across all repeats, denoted as x_1,x_2,......x_k with corresponding squared uncertainty $\delta^2 x_k$ defined by:

$$\delta^2 x_k \equiv \langle (x_k - \langle x_k \rangle)^2 \rangle = \langle x_k^2 \rangle - \langle x_k \rangle^2 \tag{9}$$

And the optimal estimate of uncertainty in permutation scores is given by $\hat{X}$, the optimal estimator for $\langle X \rangle$ in the sense of minimizing $\delta^2 x_k$, by a weighted sum of the individual estimates. So the observations with smaller uncertainties get greater weight.

$$\hat{X} = \frac{\sum_{k=1}^{K} \frac{x_k}{\delta^2 x_k}}{\sum_{k=1}^{K} \frac{1}{\delta^2 x_k}} \tag{10}$$

4.5 Toxicity Label Prediction

Table S3: Contingency table for Stacked Model vs Random Forest

	Random Forest **correct**	Random Forest **wrong**
Stacked Model **correct**	1618	**73**
Stacked Model **wrong**	**43**	356

Table S4: Contingency table for Stacked Model vs Multi-layer Perceptron

	Multi-layer Perceptron **correct**	Multi-layer Perceptron **wrong**
Stacked Model **correct**	1589	**102**
Stacked Model **wrong**	**54**	345

Table 56: Contingency table for Stacked Model vs LightGBM

	LightGBM **correct**	LightGBM **wrong**
Stacked Model **correct**	1602	**89**

Stacked Model **wrong**	**52**	347

References

1 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko and Y. Vázquez-Baeza, *Nat. Methods*, 2020, **17**, 261–272.

2 O. Menyhart, B. Weltz and B. Győrffy, *PLOS ONE*, 2021, **16**, e0245824.

3 O. J. Dunn, *J. Am. Stat. Assoc.*, 1961, **56**, 52–64.

4 Association-metrics Package, https://pypi.org/project/association-metrics/.